Statistical Analysis with Missing Data

Second Edition

WILEY SERIES IN PROBABILITY AND STATISTICS

Established by WALTER A. SHEWHART and SAMUEL S. WILKS

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

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

Statistical Analysis with Missing Data

Second Edition

RODERICK J. A. LITTLE

DONALD B. RUBIN

A JOHN WILEY & SONS, INC., PUBLICATION

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

Library of Congress Cataloging-in-Publication Data

Little, Roderick J. A.
Statistical analysis with missing data / Roderick J Little, Donald B. Rubin. - - 2nd ed.
p. cm. - - (Wiley series in probability and statistics)
"A Wiley-Interscience publication."
Includes bibliographical references and index.
ISBN 0-471-18386-5 (acid-free paper)
1. Mathematical statistics. 2. Missing observations (Statistics) I Rubin, Donald B. II. Title. III. Series

QA276 .L57 2002
519.5- -dc21 2002027006

ISBN 0-471-18386-5

Printed in the United States of America.

20 19 18 17 16 15 14

Contents

Preface

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

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

of the main statistical issues. We have managed to cover about three-quarters of the material in the book in a 40-hour graduate statistics course.

When the first edition of this book was written in the mid-l980s, a weakness in the literature was that missing-data methods were mainly confined to the derivation of point estimates of parameters and approximate standard errors, with interval estimation and testing based on large-sample theory. Since that time, Bayesian methods for simulating posterior distributions have received extensive development, and these developments are reflected in the second edition. The closely related technique of multiple imputation also receives greater emphasis than in the first edition, in recognition of its increasing role in the theory and practice of handling missing data, including commercial software. The first part of the book has been reorganized to improve the flow of the material. Part II includes extensions of the EM algorithm, not available at the time of the first edition, and more Bayesian theory and computation, which have become standard tools in many areas of statistics. Applications of the likelihood approach have been assembled in a new Part III. Work on diagnostic tests of model assumptions when data are incomplete remains somewhat sketchy.

Because the second edition has some major additions and revisions, we provide a map showing where to locate the material originally appearing in Edition 1.

First Edition	**Second Edition**
1. Introduction	1. Introduction
2. Missing Data in Experiments	2. Missing Data in Experiments
3.2. Complete-Case Analysis	3.2. Complete-Case Analysis
3.3. Available-Case Analysis	3.4. Available-Case Analysis
3.4. Filling in the Missing Values	4.2. Imputing Means from a Predictive Distribution
4.2., 4.3. Randomization Inference with and without Missing Data	Omitted
4.4. Weighting Methods	3.3. Weighted Complete-Case Analysis
4.5. Imputation Procedures	4. Imputation
4.6. Estimation of Sampling Variance with Nonresponse	5. Estimation of Imputation Uncertainty
5. Theory of Inference Based on the Likelihood Function	6. Theory of Inference Based on the Likelihood Function
6. Factored Likelihood Methods	7. Factored Likelihood Methods, Ignoring the Missing-Data Mechanism
7. Maximum Likelihood for General Patterns of Missing Data	8. Maximum Likelihood for General Patterns 9.1. Standard Errors Based on the Information Matrix.

First Edition	Second Edition
8. ML for Normal Examples	11. Multivariate Normal Examples
9. Partially Classified Contingency Tables	13. Partially Classified Contingency Tables
10.2 The General Location Model	14. Mixed Normal and Categorical Data with Missing Values
10.3., 10.4. Extensions	
10.5. Robust Estimation	12. Models for Robust Estimation
11. Nonignorable Models	15. Nonignorable Models
12.1., 12.2. Survey Nonresponse	3.3. Weighted Complete-Case Analysis
12.3. Ignorable Nonresponse Models	4. Imputation
12.4. Multiple Imputation	5.4. Introduction to Multiple Imputation
	10. Bayes and Multiple Imputation
12.5. Nonignorable Nonresponse	15.5. Normal Pattern-Mixture Models

The statistical literature on missing data has expanded greatly since the first edition, in terms of scope of applications and methodological developments. Thus, we have not found it possible to survey all the statistical work and still keep the book of tolerable length. We have tended to confine discussion to applications in our own range of experience, and we have focused methodologically on Bayesian and likelihood-based methods, which we believe provide a strong theoretical foundation for applications. We leave it to others to describe other approaches, such as that based on generalized estimating equations.

Many individuals are due thanks for their help in producing this book. NSF and NIMH (through grants NSF-SES-83-11428, NSF-SES-84-11804, NIMH-MH-37188, DMS-9803720, and NSF-0106914) helped support some aspects of the research reported here. For the first edition, Mark Schluchter helped with computations, Leisa Weld and T. E. Raghunathan carefully read the final manuscript and made helpful suggestions, and our students in Biomathematics M232 at UCLA and Statistics 220r at Harvard University also made helpful suggestions. Judy Siesen typed and retyped our many drafts, and Bea Shube provided kind support and encouragement. For the second edition, we particularly thank Chuanhai Liu for help with computation, and Mingyao Li, Fang Liu, and Ying Yuan for help with examples. Many readers have helped by finding typographical and other errors, and we particularly thank Adi Andrei, Samantha Cook, Shane Jensen, Elizabeth Stuart, and Daohai Yu for their help on this aspect.

In closing, we continue to find that many statistical problems can be usefully viewed as missing-value problems even when the data set is fully recorded, and moreover, that missing-data research can be an excellent springboard for learning about statistics in general. We hope our readers will agree with us and find the book stimulating.

Ann Arbor, Michigan R. J. A. LITTLE
Cambridge, Massachusetts D. B. RUBIN

PART I

Overview and Basic Approaches

CHAPTER 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 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 breakdowns 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 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.

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 packages typically exclude units that have missing value codes for any of the variables involved in an analysis. This strategy, which we term a "complete-case analysis," is generally inappropriate, since the investigator is usually interested in making inferences about the entire target population, rather than 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 underlying values.

1.2. MISSING-DATA PATTERNS

We find it useful to distinguish the missing-data *pattern*, which describes which values are observed in the data matrix and which values are missing, and the missing-data *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 missing-data 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 pattern of missing data to see if an orderly pattern emerges. In this section we discuss some important patterns, and in the next section we formalize the idea of missing-data mechanisms.

Let $Y = (y_{ij})$ denote an $(n \times K)$ rectangular data set without missing values, with ith row $y_i = (y_{i1}, \ldots, y_{iK})$ where y_{ij} is the value of variable Y_j for subject i. With missing data, define the *missing-data indicator matrix* $M = (m_{ij})$, such that $m_{ij} = 1$ if y_{ij} is missing and $m_{ij} = 0$ if y_{ij} is present. The matrix M then defines the pattern of missing data. Figure 1.1 shows some examples of missing-data patterns. Some methods for handling missing data apply to any pattern of missing data, whereas other methods are restricted to a special pattern.

EXAMPLE 1.1. *Univariate Missing Data*. Figure 1.1*a* 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 1.1*a*, namely, the problem of missing data in designed experiments. In the context of agricultural trials this 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 $Y_1, \ldots, Y_{K-1}$, 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 and hence a simple analysis. However, sometimes 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 pattern with Y_K incomplete and $Y_1, \ldots, Y_{K-1}$ fully observed. Missing-data techniques fill in the missing values of Y_K in order 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.

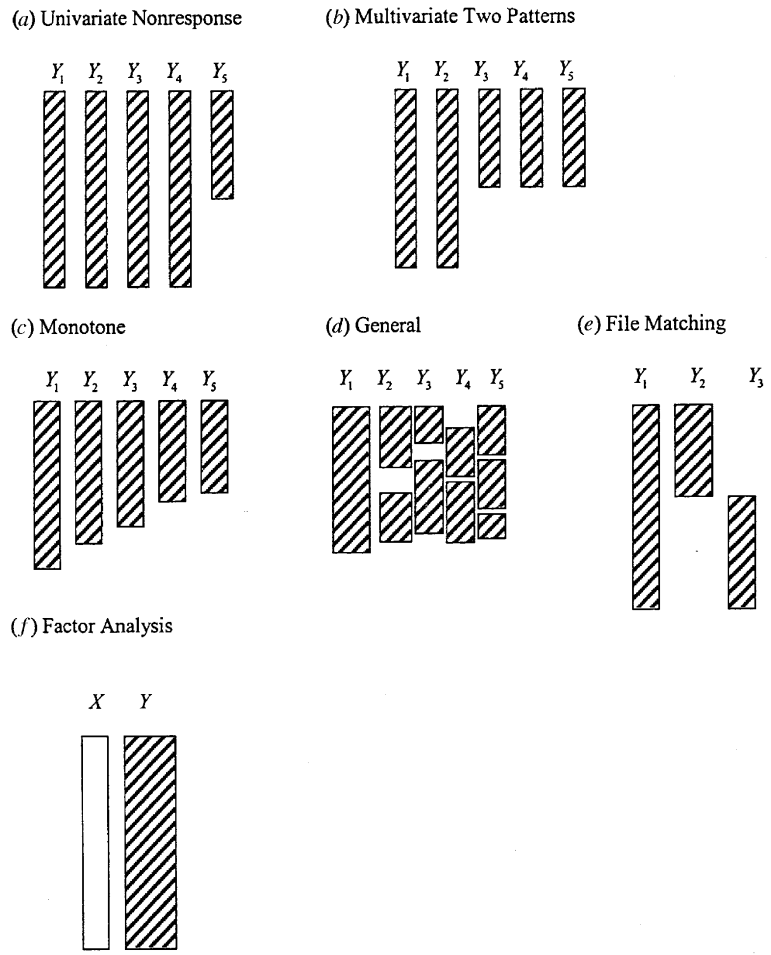


Figure 1.1. Examples of missing-data patterns. Rows correspond to observations, columns to variables.

EXAMPLE 1.2. *Unit and Item Nonresponse in Surveys*. Another common pattern is obtained when the single incomplete variable Y_K in Figure 1.1*a* is replaced by a set of variables $Y_{J+1}, \ldots, Y_K$, all observed or missing on the same set of cases (see Figure 1.1*b*, 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 do 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 respondents and nonrespondents, such as household location or characteristics measured in a listing operation 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 1.1*d*. 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, Nisselson, and Olkin (1983), Madow, Olkin, and Rubin (1983), Rubin (1987a) and Groves et al. (2002).

EXAMPLE 1.3. *Attrition in Longitudinal Studies*. Longitudinal studies collect information on a set of cases repeatedly over time. A common missing-data problem is attrition, where subjects 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 because they move to a location that is inaccessible to the researchers, or, in a clinical trial, some subjects drop out of the study 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 $Y_{j+1}, \ldots, Y_K$ are missing for cases where Y_j is missing, for all $J = 1, \ldots, K-1$ (see Figure 1.1*c* for $K = 5$). Methods for handling monotone missing data can be easier than methods for general patterns, as shown 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 1.1, which was obtained from the results of a panel study of students in 10 Illinois schools, analyzed by Marini, Olsen, and Rubin (1980). The first block of variables was recorded for all

Table 1.1 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 Cases	Percentage of Cases
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

Source: Marini, Olsen, and Rubin 1980.

[a]Observations 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).

[b]Observations 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).

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 were recorded for a subset of those responding to block 2 variables. Thus, block 2 variables are more observed than block 3 variables. Blocks 1, 2, and 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 below), the creation of a monotone pattern involves the loss of a substantial amount of information.

EXAMPLE 1.4. *The File-Matching Problem, with Two Sets of Variables Never Jointly Observed*. With large amounts of missing data, the possibility that variables are never observed together arises. When this happens, it is important to be aware of the problem since it implies that some parameters relating to the association between these variables are not estimable from the data, and attempts to estimate them may yield misleading results. Figure 1.1*e* illustrates an extreme version of this problem that arises in the context of combining data from two sources. In this pattern, Y_1

Table 1.2 Data Matrix for Children in a Survey Summarized by the Pattern of Missing Data: 0 = observed, 1 = missing.

	Variables					
Pattern	Age	Gender	Weight 1	Weight 2	Weight 3	N of Children with Pattern
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

represents a set of variables that is common to both data sources and fully-observed, Y_2 a set of variables observed for the first data source but not the second, and Y_3 a set of variables observed for 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 1.5. *Latent-Variable Patterns with 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 1.1*f*, where X represents a set of latent variables that are completely missing, and Y 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 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 missing values of Y as missing. This example is examined in more detail in Section 11.3.

We make the following key assumption throughout the book:

Assumption 1.1: missingness indicators hide true values that are meaningful for analysis.

Assumption 1.1 may seem innocuous, but it has important implications for the analysis. When the assumption applies, it makes sense to consider analyses that effectively predict, or "impute" (that is, fill in) the unobserved values. If, on the other hand, Assumption 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 missingness indicator is more appropriate. Example 1.6 describes a situation with longitudinal data on obesity where Assumption 1.1 clearly makes sense. Example 1.7 describes the case of a randomized experiment where it makes sense for one outcome variable (survival) but not for another (quality of life). Example 1.8 describes a situation in opinion polling where Assumption 1.1 may or may not make sense, depending on the specific setting.

EXAMPLE 1.6. *Nonresponse in a Binary Outcome Measured at Three Time Points*. Woolson and Clarke (1984) analyze data from the Muscatine Coronary Risk Factor Study, a longitudinal study of coronary risk factors in schoolchildren. Table 1.2 summarizes the data matrix by its pattern of missing data. Five variables (gender, age, and obesity for three rounds of the survey) are recorded for 4856 cases—gender and age are completely recorded, but the three obesity variables are sometimes missing with six patterns of missingness. Since age is recorded in five

categories and the obesity variables are binary, the data can be displayed as counts in a contingency table. Table 1.3 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 other patterns are defined similarly.

Woolson and Clarke analyze these data by fitting multinomial distributions over the $3^3 - 1 = 26$ response categories for each column in Table 1.3. That is, missingness is regarded as defining strata of the population. We suspect that for these data it

Table 1.3 Number of Children Classified by Population and Relative Weight Category in Three Rounds of a Survey

Response Category[a]	Males Age Group					Females Age Group				
	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

Source: Woolson and Clarke (1984).

[a]NNN indicates not obese in 1977, 1979, and 1981; O indicates obese, and M indicates missing in a given year.

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, and are described in Chapter 12. The methods involve quite straightforward modifications of existing algorithms for categorical data analysis currently available in statistical software packages. For an analysis of these data that averages over patterns of missing data, see Ekholm and Skinner (1998).

EXAMPLE 1.7. *Causal Effects of Treatments with Survival and Quality of Life Outcomes*. Consider a randomized experiment with two drug treatment conditions, $T = 0$ or 1, and suppose that a primary outcome of the study is survival ($D = 0$) or death ($D = 1$) at one year after randomization to treatment. For participant i, let $D_i(0)$ denote the one-year survival status if assigned treatment 0, and $D_i(1)$ survival status if assigned treatment 1. 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, so $D_i(1)$ is unobserved ("missing") for participants assigned treatment 0, and $D_i(1)$ is unobserved ("missing") for participants assigned treatment 1. Individual causal effects are unobserved, but randomization allows for unbiased estimation of average causal effects for a sample or population (Rubin, 1978a), 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 Assumption 1.1, since one can consider what the survival outcome would have been under the treatment not assigned, even though this outcome is never observed. For more applications of this "potential outcome" formulation to inference about causal effects, see, for example, Angrist, Imbens, and Rubin (1996), Barnard et al. (1998), Hirano et al. (2000), and Frangakis and Rubin (1999, 2001, 2002).

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 in some sense or "censored" due to death—we think it usually makes little sense to treat these outcomes as missing values as in Assumption 1.1, given that quality of life is a meaningless concept for people who are not alive. More specifically, let $D_i(T)$ denote the potential one-year survival outcome for participant i under treatment T, as before. The potential outcomes on D can be used to classify the patients into four groups:

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

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

For the *LL* patients, there is a bivariate distribution of individual potential outcomes of Y under treatment and control, with one of these outcomes being observed and one missing. For the *DD* patients, there is no information on Y, and it is dubious to treat these values as missing. For *LD* patients there is a distribution of Y under the control condition, but not under the treatment condition, and for *DL* patients there is a distribution of Y under the new treatment condition but not under the control condition. Causal inference about Y can be conceptualized within this framework as imputing the survival status of participants under the treatment not received, and then imputing quality of life of participants under the treatment not received *within the subpopulation of LL patients*.

EXAMPLE 1.8. *Nonresponse in Opinion Polls*. Consider the situation where individuals 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. Assumption 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. Assumption 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, Stern and Vehovar (1996) consider a situation where there is a complete list of eligible voters, and those who do not vote were counted as "no's" in the referendum. Here Assumption 1.1 applies to all the unobserved values in the initial poll. Consequently, Rubin, Stern and Vehovar (1996) consider methods that effectively impute the missing responses under a variety of modeling assumptions, as discussed in Example 15.14.

1.3. MECHANISMS THAT LEAD TO MISSING DATA

In the previous section we considered various patterns of missing data. A different issue concerns the mechanisms that lead to missing data, 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. Missing-data mechanisms are crucial since 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 missing-data indicators as random variables and assigning them a distribution. We now review this theory, using a notation and terminology that differs slightly from that of the

original paper but has come into common use in the modern statistics literature on missing data.

Define the complete data $Y = (y_{ij})$ and the missing-data indicator matrix $M = (M_{ij})$ as in the previous section. The missing-data mechanism is characterized by the conditional distribution of M given Y, say $f(M|Y, \phi)$, where ϕ denotes unknown parameters. If missingness does not depend on the values of the data Y, missing or observed, that is, if

$$f(M|Y, \phi) = f(M|\phi) \text{ for all } Y, \phi, \tag{1.1}$$

the data are called missing completely at random (MCAR)—note that this assumption does not mean that the pattern itself is random, but rather that missingness does not depend on the data values. Let Y_{obs} denote the observed components or entries of Y, and Y_{mis} the missing components. An assumption less restrictive than MCAR is that missingness depends only on the components Y_{obs} of Y that are observed, and not on the components that are missing. That is,

$$f(M|Y, \phi) = f(M|Y_{\text{obs}}, \phi) \text{ for all } Y_{\text{mis}}, \phi. \tag{1.2}$$

The missing-data mechanism is then called missing at random (MAR). The mechanism is called not missing at random (NMAR) if the distribution of M depends on the missing values in the data matrix Y.

Perhaps the simplest data structure is a univariate random sample for which some units are missing. Let $Y = (y_1, \ldots, y_n)^{\mathrm{T}}$ where y_i denotes the value of a random variable for unit i, and let $M = (M_1, \ldots, M_n)$ where $M_i = 0$ for units that are observed and $M_i = 1$ for units that are missing. Suppose the joint distribution of (y_i, M_i) is independent across units, so in particular the probability that a unit is observed does not depend on the values of Y or M for other units. Then,

$$f(Y, M|\theta, \phi) = f(Y|\theta) f(M|Y, \phi) = \prod_{i=1}^{n} f(y_i|\theta) \prod_{i=1}^{n} f(M_i|\, y_i, \phi), \tag{1.3}$$

where $f(y_i|\theta)$ denotes the density of y_i indexed by unknown parameters θ, and $f(M_i|\, y_i, \phi)$ is the density of a Bernoulli distribution for the binary 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 missing-data mechanism is MCAR (or in this case equivalently MAR). If the mechanism depends on y_i the mechanism is NMAR since it depends on y_i that are missing, assuming that there are some.

Let r denote the number of responding units ($M_i = 0$). An obvious consequence of the missing values in this example is the reduction in sample size from n to r. We might contemplate carrying out the same analyses on the reduced sample as we intended for the size-n sample. For example, if we assume the values are normally distributed and wish to make inferences about the mean, we might estimate the mean by the sample mean of the responding units with 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, since then the observed cases are a random subsample of all the cases. However, if the data are NMAR, the analysis based on the responding subsample is generally biased for the parameters of the distribution of Y.

EXAMPLE 1.9. *Artificially-Created Missing Data in a Univariate Normal Sample*. The data in Figure 1.2 provide a concrete illustration of this situation. Figure 1.2*a* 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 has the value -0.03. Figure 1.2*b* presents a subsample of data obtained from the original sample in Figure 1.2*a* by deleting units by the MCAR mechanism:

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

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.

Figures 1.2*c* and *d* illustrate NMAR mechanisms. In Figure 1.2*c*, negative values from the original sample have been retained and positive values have been deleted, that is,

$$\Pr(M_i = 1 | \, y_i, \phi) = \begin{cases} 1, & \text{if } y_i > 0 \\ 0, & \text{if } y_i \leq 0. \end{cases} \tag{1.5}$$

This mechanism is clearly NMAR, and the standard complete-data analysis that ignores the missing-data mechanism is biased. In particular, the sample mean, -0.89, obviously 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 1.2*d* are the respondents from the original sample with:

$$\Pr(M_i = 1 | \, y_i, \phi) = \Phi(-2.05 y_i), \tag{1.6}$$

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 missing-data mechanism is again NMAR, 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 1.2*c*, and we wish to estimate the population mean. If the censoring mechanism is *known*, then methods are available that correct for the selection bias of the sample mean, as discussed in Section 15.3. If the censoring mechanism is *unknown*, the problem is much more difficult. The principal evidence that the response mechanism is not MAR lies in the fact that the observed samples are *asymmetric*, which contradicts

−3.5	7		7	7
−3				
−2.5	8		8	8
−2				
−1.5	57889	578	57889	57889
−0.5	001111222233	1112233	001111222233	001111222233
−0.5	5556666778989999999	566788899999	5556666778888899999	555666778888899999
−0	0112222223344	011234	0112222223344	0112222234
+0	00112222222233344444	0122222234		012224
+0.5	56777778899	677789		
+1	0011113444	11144		
+1.5	56778	6		
+2	023	02		
+2.5				
+3	3			
Sample mean = −0.03		Sample mean = 0.11	Sample mean = −0.89	Sample mean = −0.81
(a) Uncensored Sample, $n = 100$		(b) Ignorable Censoring Mechanism, $\Pr(y_i)$ observed) $= 0.5$, $r = 52$	(c) Pure Censoring from above at 0.0, $r = 51$	(d) Nonignorable Censoring Mechanism, $\Pr(y_i$ observed) $= \Phi(-2.05y_i)$ $r = 53$

Figure 1.2. Stem and leaf displays of distribution of standard normal sample with stochastic censoring.

the assumption that the original data have a (symmetric) normal distribution. If we are confident that the uncensored sample has a symmetric distribution, we can use this information to adjust for selection bias. On the other hand, if we have little knowledge about the form of the uncensored 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 case, the sample mean is biased for the population mean; in the latter case it is not.

EXAMPLE 1.10. *Historical Heights*. Wachter and Trussell (1982) present an interesting illustration of stochastic censoring, involving the estimation of historical 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, since minimal height standards were often in operation and were enforced with variable 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, adapted from Wachter and Trussell, 1982. The shaded area in the figure represents the heights of men excluded from the recruit sample and is drawn under the assumption that heights are normally distributed in the 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, so the inferences from the stochastically

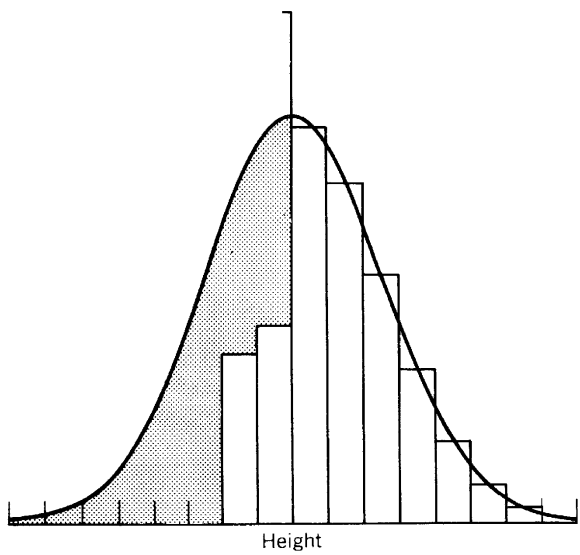


Figure 1.3. Observed and population distributions of historical heights. Population distribution is normal, observed distribution is represented by the histogram, and the shaded area represents missing data.

censored data under the assumption of normality have some validity. In many other problems involving missing data, such information is not available or is highly tenuous in nature. As discussed in Chapter 15, the sensitivity of answers from an incomplete sample to unjustifiable or tenuous assumptions is a basic problem in the analysis of data subject to unknown missing-data mechanisms, such as can occur in survey data subject to nonresponse.

EXAMPLE 1.11. *Mechanisms of Univariate Nonresponse (Example 1.1. continued)*. Suppose the data consist of an incomplete variable Y_K and a set of fully observed variables $Y_1, \ldots, Y_{K-1}$, yielding the pattern of Figure 1.1*a*. As discussed in Examples 1.1 and 1.2, a wide variety of situations lead to the pattern in this figure. Since $Y_1, \ldots, Y_{K-1}$ are fully observed, it is sufficient to define a single missing-data indicator variable M that takes the value 1 if Y_K is missing and 0 if Y_K is observed. Suppose that observations on Y and M are independent. The data are then MCAR if:

$$\Pr(M_i = 1 | \, y_{i1}, \ldots, y_{iK}; \phi) = \phi,$$

a constant that does not depend on any of the variables. The complete cases are then a random subsample of all the cases. The MCAR assumption is often too strong when the data on Y_K are missing because of uncontrolled events in the course of the data collection, such as nonresponse, or errors in recording the data, since these events are often associated with the study variables. The assumption may be more plausible if the missing data are *missing by design*. For example, if Y_K is the variable of interest but is expensive to measure, and $Y_1, \ldots, Y_{K-1}$ are inexpensive surrogate measures for Y_K, then the pattern of Figure 1.1*a* can be obtained by a planned design where $Y_1, \ldots, Y_{K-1}$ are recorded for a large sample and Y_K is recorded for a subsample. The technique of *double sampling* in survey methodology provides another instance of planned missing data. A large sample is selected, and certain basic characteristics are recorded. Then a random subsample is selected from the original sample, and more variables are measured. The resulting data form the pattern of this example, with Y_K replaced by a vector of measures (Fig. 1.1*b*).

The data are MAR if:

$$\Pr(M_i = 1 | \, y_{i1}, \ldots, y_{iK}; \phi) = \Pr(M_i = 1 | \, y_{i1}, \ldots, y_{i,K-1}; \phi),$$

so that missingness may depend on the fully observed variables $Y_1, \ldots, Y_{K-1}$ but does not depend on the incomplete variable Y_K. If the probability that Y_K is missing depends on Y_K after conditioning on the other variables, then the mechanism is NMAR.

For example, suppose $K = 2$, $Y_1 =$ age, and $Y_2 =$ income. 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 respondents 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, then the data are

NMAR. This latter case is hardest to deal with analytically, which is unfortunate, since it may be the most likely case in this application. When missing data are not under control of the sampler, the MAR assumption is made more plausible by collecting data $Y_1, \ldots, Y_{K-1}$ on respondents and nonrespondents that are predictive both of Y_K and the probability of being missing. Including these data in the analysis then reduces the association between M and Y_K, and helps to justify the MAR assumption. In the controlled missing-data environment of double sampling, the missing data are MAR, even if the probability of inclusion at the second stage is allowed to depend on the values of variables recorded at the first stage, a useful design strategy for improving efficiency in some applications.

The case of *censoring* illustrates a situation where the mechanism is NMAR but may be understood. 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 censored because the event had not occurred before the termination of data collection. If the time to censoring is known, then we have the partial information that the failure time exceeds the time to censoring. The analysis of the data needs to take account of this information to avoid biased conclusions.

The significance of these assumptions about the missing-data mechanism depends somewhat on the objective of the analysis. For example, if interest lies in the marginal distribution of $Y_1, \ldots, Y_{K-1}$, 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 typically of theoretical rather than practical importance). If interest lies in the conditional distribution of Y_K given $Y_1, \ldots, Y_{K-1}$, 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 if 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 $Y_1, \ldots, Y_{K-1}$ and Y_K, the data on $Y_1, \ldots, Y_{K-1}$ are typically not useful in estimating the mean of Y_K; however, when data on Y_K are missing, the data on $Y_1, \ldots, Y_{K-1}$ 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 selection bias when the data are not MCAR. These points will be examined in more detail in subsequent chapters.

EXAMPLE 1.12. *Mechanisms of Attrition in Longitudinal Data (Example 1.3. continued)*. For the monotone pattern of attrition in longitudinal data (Fig. 1.1*c* for $K = 5$) the notation can again be simplified by defining a single missing indicator M that now takes the value j if $Y_1, \ldots, Y_{j-1}$ are observed and $Y_j, \ldots, Y_K$ are missing (that is, dropout occurs between times $(j-1)$ and j, and the value $K+1$ for complete cases.) The missing data (dropout, attrition) mechanism is then MCAR if:

$$\Pr(M_i = j \mid y_{i1}, \ldots, y_{iK}; \phi) = \phi \text{ for all } y_{i1}, \ldots, y_{iK},$$

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

$$\Pr(M_i = j \mid y_{i1}, \ldots, y_{iK}; \phi) = \Pr(M_i = j \mid y_{i1}, \ldots, y_{i,j-1}, \phi) \text{ for all } y_{i1}, \ldots, y_{iK}.$$

Murray and Findlay (1988) provided an instructive example of MAR for longitudinal data from a study of hypertensive drugs where the outcome was diastolic blood pressure. By protocol, the subject was no longer included in the study when the observed diastolic blood pressure rose too high. This mechanism is not MCAR, since it depends on the values of blood pressure. But blood pressure at the time of dropout was observed before the subject dropped out. Hence the mechanism is MAR, because dropout only depends on the observed part of Y. Many methods for handling missing data assume the mechanism is MAR, and yield biased estimates when the data are not MAR.

EXAMPLE 1.13. *Missing at Random for a General Bivariate Pattern*. The most general missing data pattern for two variables has four kinds of units: complete cases, cases with only Y_1 observed, cases with only Y_2 observed, and cases with both variables missing. If we model as in Eq. (1.3) so that the pattern for case i depends only on the outcomes y_{i1}, y_{i2} for that case, we can write:

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

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 $g_{10}(y_{i1}; y_{i2}; \phi) = g_{10}(y_{i2}; \phi)$, since for this pattern y_{i2} is observed and y_{i1} is missing. Applying this logic to all four patterns, MAR implies that:

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

This assumption seems somewhat 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 Y_2 is 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 NMAR. The MAR assumption, although perhaps unrealistic, is generally a better approximation to reality than the MCAR assumption, which assumes that all four probabilities are unrelated to the outcomes. Moreover, 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 NMAR mechanism above (Rubin, Stern and Vehovar, 1996).

1.4. A TAXONOMY OF MISSING-DATA METHODS

The literature on the analysis of partially missing data is comparatively recent. Review papers include Afifi and Elashoff (1966), Hartley and Hocking (1971), Orchard and Woodbury (1972), Dempster, Laird, and Rubin (1977), Little and Rubin (1983a), Little and Schenker (1994), and Little (1997). Methods proposed in this 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 is discussed in Chapter 3. It is generally easy to carry out 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 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–Thompson (1952) estimator:

$$
\left(\sum_{i=1}^{n} \pi_i^{-1} y_i \right) \left(\sum_{i=1}^{n} \pi_i^{-1} \right)^{-1}, \tag{1.7}
$$

where the sums are over sampled units, 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

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

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, usually the proportion of responding units in a subclass of the sample. Weighting methods are described further in Chapter 3.

3. *Imputation-Based Procedures*. 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 recorded units in the sample are used to substitute values; *mean* imputation, where means from sets of 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 to 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 Procedures*. A broad class of procedures is generated by defining a model for the observed 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 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 1.14. *Estimating the Mean and Covariance Matrix with Monotone Missing Data*. 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. The question of how to estimate these quantities from incomplete data is, therefore, an important one. 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. However, this method of analysis 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 are therefore biased. A more generally successful strategy is to assume 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 observations, and inferences about the conditional distribution of Y_2 given Y_1 can be based on the subset of cases 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 other 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. The ML analysis is thus related to classical regression prediction.

EXAMPLE 1.15. *Estimating the Mean and Covariance Matrix with General Missing-Data Patterns*. Many data sets with missing values do not exhibit monotone patterns or convenient close approximations such as that displayed in Table 1.1. 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 the previous example, 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 developed in Chapter 8 is an important general technique for finding maximum likelihood estimates from incomplete data. It is applied to the case of multivariate normal data in Chapter 11. The resulting algorithm is particularly instructive, because it 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 pragmatic 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 1.16. *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 the variables are all categorical. In that case the data can be arranged as a contingency table with partially classified margins as in Example 1.6. Methods for analyzing such data are discussed in Chapter 13. More generally, Chapter 14 considers multivariate data where some of the variables are continuous and some are categorical. Again, a problem not usually thought of as related to missing data, the estimation of finite mixtures, is considered from a missing-data perspective.

EXAMPLE 1.17. *Estimation when the Data May not be Missing at Random*. Essentially all the literature on multivariate incomplete data assumes that the data are MAR, and much of it also assumes that the data are MCAR. Chapter 15 deals explicitly with the case when the data are not MAR, and models are needed for the missing-data mechanism. Since 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 missing-data mechanism.

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 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 missing-data 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 a data matrix and let $M = (m_{ij})$ be the corresponding missing-data indicator matrix, where $m_{ij} = 1$ indicates missing and $m_{ij} = 0$ indicates present.

(a) Propose situations where two values of m_{ij} are not sufficient. (Hint: see Heitjan and Rubin, 1991).

(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) Suppose $m_{ij} = 1$ or $m_{ij} = 0$. When attention is focused only on the units that fully respond, the conditional distribution of y_i given $m_i = (0, 0, \ldots, 0)$ is being estimated, where y_i and m_i are the ith rows of Y and M,

respectively. Propose situations where it makes sense to define the conditional distribution of y_i given other missing-data 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 missing-data patterns and their probabilities.

1.6. One way to understand missing data mechanisms is to generate hypothetical complete data and then create missing values by specific mechanisms. This is common in simulation studies of missing-data methods, since the deleted values can be retained and used to compare methods; missing-data methods are hard to assess on real data sets for the obvious reason that the missing values are rarely known. Consider 100 trivariate normal observations $\{(y_{i1}, y_{i2}, u_i), i = 1, \ldots, 100\}$ on (Y_1, Y_2, U) generated as follows:

$$\begin{aligned} y_{i1} &= 1 + z_{i1} \\ y_{i2} &= 5 + 2^* z_{i1} + z_{i2} \\ u_i &= a^*(y_{i1} - 1) + b^*(y_{i2} - 5) + z_{i3}, \end{aligned}$$

where $\{(z_{i1}, z_{i2}, z_{i3}), i = 1, \ldots, 100\}$ are independent standard normal (that is, mean 0, variance 1) deviates. Suppose the hypothetical complete data set consist of observations $\{(y_{i1}, y_{i2}), i = 1, \ldots, 100\}$, which have a bivariate normal distribution with means (1, 5), variances (1, 5) and correlation $2/\sqrt{5} = 0.89$. The observed (incomplete) data are $\{y_{i1}, y_{i2}), i = 1, \ldots, 100\}$, where the values $\{y_{i1}\}$ of Y_1 are all observed, but some values $\{y_{i2}\}$ of Y_2 are missing. The latent variable U determines missingness of Y_2 as follows:

$$y_{i2} \text{ is missing if } u_i < 0.$$

The missing indicator variable M_2 for missingness of Y_2 using this mechanism is thus: $m_{i2} = 1$ if $u_i < 0$ and $m_{i2} = 0$ if $u_i \geq 0$. Since U as defined above has mean zero this mechanism should create missing values of Y_2 for about 50% of the observations.

(a) Generate a data set of 100 observations using the above process with $a = b = 0$. Display the marginal distributions of Y_1 and Y_2 for complete and incomplete cases. Is this mechanism MCAR, MAR, or NMAR?

(b) Carry out a t-test comparing the means of Y_1 for complete and incomplete cases. Is there evidence from this test that the data are not (a) MCAR; (b) MAR; (c) NMAR?

(c) Repeat parts (a) and (b) with (i) $a = 2$, $b = 0$ and (ii) $a = 0$, $b = 2$.

CHAPTER 2

Missing Data in Experiments

2.1. INTRODUCTION

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 design. For example, with two factors being studied, the analysis is particularly simple when the same number of observations are taken at each combination of factor levels. Textbooks on experimental design catalog many examples of specialized analyses (Box, Hunter and Hunter, 1985; Cochran and Cox, 1957; Davies, 1960; Kempthorne, 1952; Winer, 1962; Wu and Hamada, 2000).

Since the levels of the design factors in an experiment are fixed by the experimenter, missing values, if they occur, do so far more frequently in the outcome variable, Y, than in the design factors, X. Consequently, we restrict attention to missing values in Y, and the data have the pattern of Figure 1.1*a* with $(Y_1, \ldots, Y_4)$ representing the fully observed covariates (X), and Y_5 representing the incomplete outcome (Y). With this pattern, and assuming MCAR as discussed in Example 1.11, the cases with Y missing actually 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. 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 the standard analysis 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: (1) It is easier to specify the data structure using the terminology of experimental design (e.g., as a balanced incomplete block), (2) it is easier to compute necessary statistical summaries, and (3)

it is easier to interpret the results of analyses since 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 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.

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 testing 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. Further discussion is given, for example, in Anderson (1946) and Jarrett (1978).

2.2. THE EXACT LEAST SQUARES SOLUTION WITH COMPLETE DATA

Let X be an $n \times p$ matrix whose ith row, $x_i = (x_{i1}, \ldots, x_{ip})$, provides the values of the fixed factors for the ith unit. For example, in a 2×2 design with two observations per cell and the levels of the factors labeled 0, 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 $Y = (y_1, \ldots, y_n)^{\mathrm{T}}$ is assumed to satisfy the linear model

$$Y = X\beta + e, \tag{2.1}$$

where $e = (e_1, \ldots, e_n)^{\mathrm{T}}$, 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

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

if $X^{\mathrm{T}}X$ is full rank and is undefined otherwise. If $X^{\mathrm{T}}X$ 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 is normally distributed with mean β and variance $(X^{\mathrm{T}}X)^{-1}\sigma^2$.

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

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

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

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

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

Tests of the hypothesis that some set of linear combinations of β are all zero are carried out 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

$$S = (C^{\mathrm{T}}\hat{\beta})^{\mathrm{T}}[C^{\mathrm{T}}(X^{\mathrm{T}}X)^{-1}C]^{-1}(C^{\mathrm{T}}\hat{\beta}). \tag{2.5}$$

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

$$F = \frac{S/w}{\hat{\sigma}^2}. \tag{2.6}$$

If the e_i are normal, then F in Eq. (2.6) is the likelihood ratio test for $C^{\mathrm{T}}\beta = 0$; if in addition $C^{\mathrm{T}}\beta = 0$, then F is distributed as Snedecor's F distribution with w and $n-p$ degrees of freedom, since S/σ^2 and $(n-p)\hat{\sigma}^2/\sigma^2$ are 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 easy and precise. In particular, the matrix $X^{\mathrm{T}}X$ 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, such as treatment effects and block effects, are easy to calculate. In fact, these summaries of the data are usually calculated by simple averaging of the observations and their squares. This simplicity can be important in experiments with several factors and many parameters to be estimated, because then $X^{\mathrm{T}}X$ 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 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 (i.e., these two sets of parameters lie in disjoint parameter spaces), the incomplete cases carry no information for the ANOVA model, and the correct analysis is to simply ignore the rows of X corresponding to missing y_i and carry out 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 cases may not be positive definite, so the least squares estimates may not be uniquely identifiable from the data. Dodge (1985) provides a detailed discussion of this problem, including procedures for detecting the problem and for determining which treatment effects remain estimable. We assume in what follows that the design matrix based on the complete cases 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 units.

The computational problem is that the specialized formulas and computing routines used with complete Y cannot be used, since 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 Eq. (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 at first seem to be circular and of little practical help since 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 (1) the correct least squares estimates of β, $\hat{\beta}_*$, and (2) the resultant residual sum of squares equals $(r-p)\hat{\sigma}_*^2$, so division by $(r-p)$ rather than $(n-p)$ yields the correct least squares estimate $\hat{\sigma}_*^2$ of σ^2. It is quite easy to prove these two facts. For the results in this chapter, a convenient notation lets $i = 1, \ldots, m$ index the m missing values and $i = m+1, \ldots, n$ the $r = n - m$ observed values. Let $\hat{y}_i = x_i\hat{\beta}_*$, $i = 1, \ldots, m$ denote the least squares estimates of the m missing values. Complete-data methods applied to the filled-in data minimize the quantity

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

with respect to β. By definition, $\beta = \hat{\beta}_*$ minimizes the second summation in $\mathrm{SS}(\beta)$ but also by definition, $\beta = \hat{\beta}_*$ minimizes the first summation in $\mathrm{SS}(\beta)$, setting it equal to zero. Consequently, with least squares estimates of missing values filled in, (1) $\mathrm{SS}(\beta)$ is minimized at $\beta = \hat{\beta}_*$ and (2) $\mathrm{SS}(\hat{\beta}_*)$ equals the minimal residual sum of squares over the r observed values of y_i. Hence (1) the correct least squares estimate of β, $\hat{\beta}_*$, equals the least squares estimate of β found by the complete-data ANOVA program, and (2) the correct least squares estimate of σ^2, $\hat{\sigma}_*^2$, is found from the complete-data ANOVA estimate of σ^2, $\hat{\sigma}^2$ by

$$\hat{\sigma}_*^2 = \hat{\sigma}^2\frac{n-p}{r-p}.$$

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

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

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, which 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. Since the residual sum of squares is quadratic in the missing value, the minimizing value of the one missing value can then be found. This method 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, (1) trial values are substituted for all missing values, (2) the complete-data analysis is performed, (3) predicted values are obtained for the missing values, (4) these predicted values are substituted for the missing values, (5) 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 11.5, 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 ith missing-value covariate is defined to be the indicator for the ith missing value, that is, zero everywhere except for the ith missing value where it equals one. The coefficient of the ith missing-value covariate, when subtracted from the initial guess of the ith missing value, yields the least squares estimate of the ith 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 implemented directly 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.

2.5.2. Notation

Suppose 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 $\tilde{y}_i$, $i = 1, \ldots, m$. Also, let Z be the $n \times m$ matrix of m missing-value covariates. The first row of Z equals $(1, 0, \ldots, 0)$, the mth row equals $(0, \ldots, 0, 1)$, and the last r rows of Z equal $(0, 0, \ldots, 0)$, since they correspond to observed y_i. The analysis of covariance uses both X and Z to predict Y.

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

$$Y = X\beta + Z\gamma + e, \tag{2.7}$$

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

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

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

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

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 Eq. (2.2) to the observed values, that is, to the last r rows of (Y, X); this minimizes the second summation in Eq. (2.8). But with $\beta = \hat{\beta}_*$, setting γ equal to $(\hat{\gamma}_1, \ldots, \hat{\gamma}_m)^{\mathrm{T}}$ where

$$\hat{\gamma}_i = \tilde{y}_i - x_i\hat{\beta}_*, \qquad i = 1, \ldots, m \tag{2.9}$$

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

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

Thus $(\hat{\beta}_*, \hat{\gamma})$ minimizes $\mathrm{SS}(\beta, \gamma)$ and gives the least squares estimate of (β, γ) obtained from the ANCOVA model in Eq. (2.7). Equation (2.9) also implies that the correct least squares estimate of the missing y_i, that is, $\hat{y}_i = x_i\hat{\beta}_*$, is given by $\tilde{y}_i - \hat{\gamma}_i$, or, in words:

(Correct least squares predicted value for ith missing value)
equals (initial guess for ith missing value)
minus (coefficient of ith 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 observations is computationally more attractive and 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 $\hat{\boldsymbol{\beta}}$

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. Conse-

quently, the residual mean square is correct and equal to $\hat{\sigma}_*^2$. If the covariance matrix of $\hat{\beta}_*$ from the ANCOVA equals V_* obtained by applying Eq. (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, $\hat{\sigma}_*^2$, times the upper left $p \times p$ submatrix of $[(X, Z)^{\mathrm{T}}(X, Z)]^{-1}$, say, U. Since 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,

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

By the definition of z_i,

$$X^{\mathrm{T}}Z = \sum_{i=1}^{m} x_i^{\mathrm{T}} z_i \tag{2.13}$$

and

$$Z^{\mathrm{T}}Z = \sum_{i=1}^{m} z_i^{\mathrm{T}} z_i = I_m, \tag{2.14}$$

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

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

But

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

whence (2.15) equals

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

and from (2.12)

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

so that $\hat{\sigma}_*^2 U = V_*$, the covariance matrix of $\hat{\beta}_*$ found by ignoring the missing observations, as required to complete the proof that the 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 7.4.3, 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

$$\hat{\gamma} = B^{-1}\rho, \tag{2.16}$$

where B is the $m \times m$ cross-products matrix for the residuals of the m missing-value covariates after adjusting for the design 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 design 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 observations on that treatment are missing. The method requires (1) the calculation of B and ρ using the complete-data ANOVA routine, (2) the inversion of B to obtain $\hat{\gamma}$ from Eq. (2.16), and (3) the calculation of the missing values from Eq. (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 missing values comprise the first row of B. Repeat the complete-data ANOVA on the jth missing-value covariate, $j = 2, \ldots, m$ (which is all zeros except for a one where the jth missing value occurs), and let the jth 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, $i = 1, \ldots, m$; 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 observed and fitted values for observation i from the ANOVA of the jth (kth) missing value covariate on X. Now $\sum_{i=1}^{n} x_{il}(z_{ik} - \hat{z}_{ik}) = 0$ for all X variables in the design matrix, by elementary properties of least squares

estimates. Hence $\sum_{i=1}^{n} \hat{z}_{ij}(z_{ik} - \hat{z}_{ik}) = 0$, since $\hat{z}_{ij}$ is a fixed linear combination of X variables $\{x_{il}: l = 1, \ldots, p\}$ for observation i. Consequently,

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

the residual for the jth missing data covariate for the kth missing value, since $z_{ij} = 1$ when $i = j$ and 0 otherwise. Similarly, the jth component of ρ is the sum over all n observations of the residual for Y (with initial values filled in) times the residual for the jth missing-value covariate. By an argument completely analogous to that just given, this is simply the residual for the jth 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 observations, 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, $\rho = -(0.0798, 0.1105)^{\mathrm{T}}$. Also, we obtain the correct total sum of squares, $\mathrm{TSS}_* = 1.1679$.

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

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

Table 2.1 Strength Index of Cotton in a Randomized Block Experiment

Treatments (pounds of potassium oxide per acre)	Blocks			
	1	2	3	Totals
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
Totals	30.53	31.40	38.55	100.48

The least squares estimates of the missing values are

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

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 $\hat{\sigma}_*^2$, is obtained when the number of missing values m is subtracted from the residual degrees of freedom, $n - p$. However, sums of squares and standard errors generally will be incorrect.

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 $\lambda = C^{\mathrm{T}}\beta$, where C is a vector of p constants, be a linear combination of β with estimate $\hat{\lambda} = C^{\mathrm{T}}\hat{\beta}$ from the ANOVA of the data filled-in by least squares estimates. Since 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

$$\mathrm{SE} = \hat{\sigma}\sqrt{C^{\mathrm{T}}(X^{\mathrm{T}}X)^{-1}C}, \tag{2.17}$$

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

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

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

$$\mathrm{SE}_* = \hat{\sigma}_*\sqrt{C^{\mathrm{T}}UC}, \tag{2.19}$$

and the correct sum of squares attributable to λ is

$$\mathrm{SS}_* = \hat{\lambda}_*^2 / C^{\mathrm{T}}UC. \tag{2.20}$$

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

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

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

$$C^{\mathrm{T}}UC = C^{\mathrm{T}}(X^{\mathrm{T}}X)^{-1}C + H^{\mathrm{T}}B^{-1}H. \tag{2.22}$$

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

$$\hat{\sigma}_*^2 = \hat{\sigma}^2(n-p)/(r-p)$$

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

$$\mathrm{SE}_* = \sqrt{\frac{n-p}{r-p}(\mathrm{SE}^2 + \hat{\sigma}^2 H^{\mathrm{T}}B^{-1}H)}. \tag{2.23}$$

Similarly, Eqs. (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:

$$\mathrm{SS}_* = \mathrm{SS}/[1 + (\mathrm{SS}/\hat{\lambda}^2)H^{\mathrm{T}}B^{-1}H]. \tag{2.24}$$

EXAMPLE 2.2. *Adjusting Standard Errors for Filled-In Missing Values (Example 2.1 continued)*. To apply the method just described, complete-data ANOVAs are required: an initial one on initial 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 corresponds to contrasting treatment 1 and treatment 2. In terms of the parameterization of Example 2.1, $C^{\mathrm{T}} = (1, -1, 0, 0, 0, 0, 0)$ and $X^{\mathrm{T}}X$ 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 of the three observations of treatment 1 minus the mean of the three observations of treatment 2 with associated complete-data standard error $\hat{\sigma}\sqrt{2/3}$ 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, $\mathrm{TSS}_* = 1.1679$. For $i = 1, 2, \ldots, m$, fill in one for the ith 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 $H^{\mathrm{T}}B^{-1}H$ 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 $\hat{\sigma}^2 = 0.0368$, SE $= 0.1567$, and SS $= 0.0235$. From Eq. (2.23), the correct standard error of $\hat{\lambda}$ is

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

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

$$\mathrm{SS}_* = 0.0235/(1 + 1.5 \times 0.2116) = 0.0178.$$

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). Related earlier work includes Tocher (1952) and Wilkinson (1958b), and later work includes Jarrett (1978).

Let $\lambda = C^{\mathrm{T}}\beta$, 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 $\hat{\lambda}_* = C^{\mathrm{T}}\hat{\beta}_*$ 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

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

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

$$\mathrm{SS} = \hat{\lambda}^{\mathrm{T}}\hat{\lambda}. \tag{2.26}$$

The correct sum of squares to attribute to λ is

$$\mathrm{SS}_* = \hat{\lambda}_*^{\mathrm{T}}(C^{\mathrm{T}}UC)^{-1}\hat{\lambda}_*. \tag{2.27}$$

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 Eq. (2.22) holds in general; hence, since the components of $\hat{\lambda}$ are orthonormal, and $\hat{\lambda} = \hat{\lambda}_*$ with least squares estimates for missing values,

$$\mathrm{SS}_* = \hat{\lambda}^{\mathrm{T}}(I + H^{\mathrm{T}}B^{-1}H)^{-1}\hat{\lambda}, \tag{2.28}$$

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

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

Equation (2.28) involves the inversion of a $w \times w$ symmetric matrix, whereas Eq. (2.29) involves the inversion of an $m \times m$ matrix. Consequently, Eq. (2.28) is 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:

$$\begin{aligned}
&\sqrt{\tfrac{3}{20}}(4, -1, -1, -1, -1, 0, 0),\\
&\sqrt{\tfrac{1}{4}}(0, 3, -1, -1, -1, 0, 0),\\
&\sqrt{\tfrac{1}{2}}(0, 0, 2, -1, -1, 0, 0),\\
&\sqrt{\tfrac{3}{2}}(0, 0, 0, 1, -1, 0, 0).
\end{aligned}$$

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 give 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 $\mathrm{SS} = 0.8191$, $\hat{\lambda}^{\mathrm{T}} = (0.3446, 0.6949, 0.4600, 0.0775)$. From Eq. (2.29), $\mathrm{SS}_* = 0.7755$.

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

Table 2.2 Corrected Analysis of Variance on the Filled-in Data

Source of variation	d.f.	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 $= 0.2077$.

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 Eq. (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 $\hat{\sigma}^2$ possess, and why?

2.3. Outline the distributional results leading to Eq. (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 Eq. (2.12) follows from the definition of U^{-1}.

2.6. Provide intermediate steps leading to Eqs. (2.13), (2.14), and (2.15).

2.7. Using the notation and results of Section 2.5.4, justify Eq. (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 Eqs. (2.17)–(2.20).

2.10. Show Eq. (2.22) and then Eq. (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).

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

[a]Spacings (in.): A, 2; B, 4; C, 6; D, 8; E, 10.

CHAPTER 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, which is related to completely observed predictor variables by 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 analysis, which confines the analysis to the set of cases 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 the methods appear in statistical computing software and are widely used, we do not generally recommend any of them except in special cases where the amount of missing information is limited. The procedures in Part II provide sounder solutions in more general circumstances.

3.2. COMPLETE-CASE ANALYSIS

Complete-case analysis confines attention to cases where all the variables are present. Advantages of this approach are (1) simplicity, since standard complete-data statistical analyses can be applied without modifications, and (2) comparability of univariate statistics, since these are all calculated on a common sample base of cases. Disadvantages stem from the potential loss of information in discarding incomplete cases. This loss of information has two aspects: loss of precision, and bias when the missing-data mechanism is not MCAR, and the complete cases are not a random sample of all the cases. Complete-case analysis may be justified in terms of simplicity when the loss of precision and the bias is minimal, so that the pay-off of

exploiting the information in the incomplete cases will be minimal. This is more likely when the fraction of complete cases is high, but it is difficult to formulate general rules of thumb, since the degree of bias and loss of precision depends not only on the fraction of complete cases and pattern of missing data, but also on the extent to which complete and incomplete cases differ, and on the parameters of interest.

Let $\hat{\theta}_{\mathrm{CC}}$ be an estimate of a scalar parameter θ from the complete cases. One might measure the increase in variance of $\hat{\theta}_{\mathrm{CC}}$ relative to the estimate $\hat{\theta}_{\mathrm{NM}}$ that would have been obtained in the absence of missing values, that is:

$$\mathrm{Var}(\hat{\theta}_{\mathrm{CC}}) = \mathrm{Var}(\hat{\theta}_{\mathrm{NM}})(1 + \Delta^*_{\mathrm{CC}}),$$

where Δ^*_{CC} is the proportional increase in variance from the loss of information. A more practically relevant measure of the loss of efficiency is Δ_{CC}, where

$$\mathrm{Var}(\hat{\theta}_{\mathrm{CC}}) = \mathrm{Var}(\hat{\theta}_{\mathrm{EFF}})(1 + \Delta_{\mathrm{CC}}) \tag{3.1}$$

and $\hat{\theta}_{\mathrm{EFF}}$ is an efficient estimate of θ based on all the available data, for example the ML estimate under a particular model.

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. Assume that the data are MCAR, so bias of CC analysis is not an issue. Suppose that the mean of Y_j is estimated by the complete-case mean $\bar{y}_j^{\mathrm{CC}}$. Then dropping the incomplete cases on Y_1 translates directly to a loss in sample size for estimating the mean of Y_1:

$$\Delta^*_{\mathrm{CC}} = \Delta_{\mathrm{CC}} = \frac{n - r}{r},$$

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

$$\Delta^*_{\mathrm{CC}} = \frac{n - r}{r}, \qquad \Delta_{\mathrm{CC}} \approx \frac{(n - r)\rho^2}{n(1 - \rho^2) + r\rho^2}. \tag{3.2}$$

(The derivation of this expression is in Section 7.2.) Thus Δ_{CC} ranges from zero when Y_1 and Y_2 are uncorrelated, to Δ^*_{CC} as $\rho^2 \to 1$. For the coefficient of the regression of Y_2 on Y_1:

$$\Delta^*_{\mathrm{CC}} \approx \frac{n - r}{r}, \qquad \Delta_{\mathrm{CC}} = 0.$$

Thus CC analysis is fully efficient, since the incomplete observations on Y_1 provide no information for the regression of Y_2 on Y_1.

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

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

the expected fraction of incomplete cases multiplied by the differences in the means for complete and incomplete cases. If the mechanism is MCAR then $\mu_{\mathrm{CC}} = \mu_{\mathrm{IC}}$ and the bias is zero.

EXAMPLE 3.3. *Bias of Complete-Case Inferences for Regression Coefficients*. Consider estimation of the regression of Y on $X_1, \ldots, X_p$ from data with missing values on Y and/or the Xs, 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 case depends on $X_1, \ldots, X_p$ but not on Y, since the analysis conditions on the values of the covariates (Glynn and Laird, 1986). This class of mechanisms includes NMAR mechanisms where the probability that a covariate is missing depends on the value of that covariate. The CC estimates of the regression coefficients are biased if the probability of being complete depends on Y after conditioning on the covariates.

EXAMPLE 3.4. *Bias 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 complete-case analysis is not subject to selection bias if the logarithm of the probability of response is an additive function of Y_1 and Y_2 (Kleinbaum, Morgenstern and Kupper, 1981). This result underpins the validity of case-control studies for estimating odds ratios from observational studies.

The discarded information from incomplete cases can be used to study whether the complete cases 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 cases with the distribution of Y_j based on incomplete cases for which Y_j is recorded. Significant differences indicate that the MCAR assumption is invalid, and the complete-case analysis yields potentially biased estimates. Such tests are useful but have limited power when the sample of incomplete cases is small. Also the tests can offer no direct evidence on the validity of the MAR assumption.

A strategy for adjusting for the bias in the selection of complete cases is to assign them case weights for use in subsequent analyses. This strategy is found most commonly in sample surveys, and is used particularly to handle the problem of

unit nonresponse, where all the survey items are missing for cases 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 complete-case analysis that differentially weights the complete cases 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 survey 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 $y_i = (y_{i1}, \ldots, y_{iK})$ represents a vector of K items for unit i, $i = 1, \ldots, N$. 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 $I = (I_1, \ldots, I_N)$. 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, $f(I|Y, Z) = f(I|Z)$, since 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 $\pi_i = E(I_i|Y, Z) = \Pr(I_i = 1|Y, Z)$, we require $\pi_i > 0$ 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 $Z = j$, $j = 1, \ldots, J$. The sampling distribution is then

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

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 can be found in sampling theory texts such as Cochran (1977) or Hansen, Hurwitz and Madow (1953). 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:

1. 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 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.
2. The choice of a statistic $v(Y_{\text{inc}})$ that is (approximately) unbiased for the variance of $t(Y_{\text{inc}})$ in repeated sampling. For example, under stratified random sampling it can be shown that the variance of $\bar{y}_{\text{st}}$ is

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

   where S_{yj}^2 is the population variance of the values of Y in stratum j. The statistic

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

   where s_{yj}^2 is the variance of the sampled values of Y in stratum j, is an unbiased estimate of $\text{Var}(\bar{y}_{\text{st}})$ under stratified random sampling.
3. The calculation of interval estimates of T, assuming t has an approximate normal sampling distribution over all stratified random samples. For example, a 95 percent confidence interval for $\bar{Y}$ under stratified random sampling is given by $C_{95}(\bar{Y}) = \bar{y}_{\text{st}} \pm 1.96\sqrt{v(Y_{\text{inc}})}$, 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).

Note that throughout this process, 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 values, although the confidence interval in step 3 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 sampling distribution, a model for Y, often in the form of a density $f(Y|Z, \theta)$ indexed by unknown parameters θ. This model is then used to predict the nonsampled values of Y, and hence 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 the weight π_i^{-1} in estimates of 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 estimator (Horvitz and Thompson, 1952). The stratified mean can be written in the form

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

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

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

$$\Pr(\text{selection and response}) = \Pr(\text{selection}) \times \Pr(\text{response} \mid \text{selection}) = \pi_{\text{i}} \phi_{\text{i}}$$

and Eq. (3.3) can be replaced by:

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

where the sum is now over responding units i and $w_i = r(\pi_i \phi_i)^{-1} / \sum_{k=1}^{r} (\pi_k \phi_k)^{-1}$. 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 provided in the next example, and a more general approach is provided in Example 3.7.

EXAMPLE 3.6. *Weighting Class Estimator of the Mean*. Suppose we divide 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 the number of respondents in weighting class $C = j$, and $r = \sum_{j=1}^{J} r_j$, a simple estimate of the response probability for units in class j is r_j/n_j. Then

responding units in weighting class j receive a weight

$$w_i = r(\pi_i \hat{\phi}_i)^{-1} / \sum_{k=1}^{r} (\pi_k \hat{\phi}_k)^{-1}, \tag{3.5}$$

where $\hat{\phi}_i = r_j / n_j$ for units i in class j.

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 a better approach is to incorporate design information in the formation of weighting classes (Little and Vartivarian, 2002). The weighting class estimator of the mean is then given by Eq. (3.4) with weights given by Eq. (3.5). For equal probability designs where π_i is constant, this estimator can be written in the simpler form:

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

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 estimator 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*).

Oh and Scheuren (1983) derive the variance of (3.6) under simple random sampling:

$$\text{Var}(\bar{y}_{\text{wc}}) = \sum_{j=1}^{J} \left(\frac{n_j}{n}\right)^2 f_j S_j^2,$$

where S_j^2 is the population variance of Y in class j, N_j is the population size in weighting class j and $f_j = r_j^{-1} - N_j^{-1}$ is a finite population correction, which can be ignored if the sampling fraction is small. Oh and Scheuren (1983) propose the following estimate of the mean squared error of $\bar{y}_{\text{wc}}$:

$$\hat{mse}(\bar{y}_{\text{wc}}) = \sum_{j=1}^{J} \left(\frac{n_j}{n}\right)^2 \left(1 - \frac{r_j n}{n_j N}\right) \frac{s_{jR}^2}{r_j} + \frac{N-n}{(N-1)n^2} \sum_{j=1}^{J} n_j (\bar{y}_{jR} - \bar{y}_{\text{wc}})^2, \tag{3.7}$$

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 in the form $\bar{y}_{\text{wc}} \pm z_{1-\alpha/2}[\hat{mse}(\bar{y}_{\text{wc}})]^{1/2}$, where $z_{1-\alpha/2}$ is the $100(1-\alpha/2)$ percentile of the standard normal distribution, but their performance in applications is not well studied.

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. A precise theory for the formation of adjustment classes is not provided, but some general comments can be offered. Adjustment classes should be chosen so that (1) Assumption 3.1 relating to the response distribution is satisfied, and (2) the mean squared error of estimates such as $\bar{y}_{\text{wc}}$ under Assumption 3.1 is minimized. Bias is limited by choosing adjustment classes that are predictive of response. Variance is limited by choosing adjustment classes that are predictive of Y, so that the within-class variance of Y is reduced, and that avoid small respondent sample sizes, r_j.

EXAMPLE 3.7. *Propensity Weighting*. Let X denote the set of variables observed for both respondents and nonrespondents. Weighting class estimators 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 since the number of weighting classes becomes too large, and includes cells with nonrespondents and 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 Little (1986), 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:

$$\Pr(M|X, Y) = \Pr(M|X), \tag{3.8}$$

so that Assumption 3.1 is satisfied when C is chosen to be X. Define the response propensity for unit i

$$p(x_i) = \Pr(m_i = 0|x_i),$$

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

$$\begin{aligned}
\Pr(m_i = 0|y_i, p(x_i)) &= E[\Pr(m_i = 0|y_i, x_i)|y_i, p(x_i)] \\
&= E[\Pr(m_i = 0|x_i)|y_i, p(x_i)], \text{ by Eq. (3.8)} \\
&= E[p(x_i)|y_i, p(x_i)] \text{ by definition of } p(x_i) \\
&= p(x_i), \text{ for all } x_i.
\end{aligned}$$

Hence

$$\Pr[M|p(X), Y] = \Pr[M|p(X)],$$

so that respondents are a random subsample within strata defined by the propensity score $p(X)$.

In practice, the transformed variable $p(X)$ is unknown and needs to be estimated from sample data. A practical procedure is (1) to estimate $p(X)$ as $\hat{p}(X)$ by logistic or probit regression of the missing-data indicator M on X, based on respondent and

nonrespondent data; (2) to form a grouped variable by coarsening the estimated $p(X)$ into five or six values; and (3) to 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 $[\hat{p}(x_i)]^{-1}$ (Cassel, Sarndal and Wretman, 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 $\Pr(M|X)$, this method removes nonresponse bias, but it may yield estimators with extremely high variance because respondents with very low estimated response propensity receive large nonresponse weights and may be unduly influential in estimates of means and totals. Also, weighting directly by $[\hat{p}(x_i)]^{-1}$ may place more reliance on correct model specification of the regression of M on X than response propensity stratification, which uses $\hat{p}(X)$ only to form adjustment classes.

EXAMPLE 3.8. *Weighted Generalized Estimating Equations*. More generally, let $y_i = (y_{i1}, \ldots, y_{iK})$ denote a vector of variables for unit i subject to missing values, and suppose y_i is fully observed for $i = 1, \ldots, r$ and y_i is missing or partially observed for $i = r+1, \ldots, n$. Define $m_i = 1$ if y_i is incomplete and $m_i = 0$ if y_i is complete. Let $x_i = (x_{i1}, \ldots, x_{ip})^{\mathrm{T}}$ 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. Further, let $z_i = (z_{i1}, \ldots, z_{iq})^{\mathrm{T}}$ 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 of interest. If there were no missing values, the solution of the generalized estimating equation (GEE),

$$\sum_{i=1}^{n} D_i(x_i, \beta)[y_i - g(x_i, \beta)] = 0, \tag{3.9}$$

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. (3.9) by

$$\sum_{i=1}^{r} D_i(x_i, \beta)[y_i - g(x_i, \beta)] = 0, \tag{3.10}$$

which yields consistent estimates provided

$$\Pr(m_i = 1|x_i, y_i, z_i) = \Pr(m_i = 1|x_i), \tag{3.11}$$

so that missingness does not depend on y_i or z_i after conditioning on x_i. Weighted GEE (Robins, Rotnitsky and Zhao, 1995) replaces Eq. (3.10) by:

$$\sum_{i=1}^{r} w_i(\hat{\alpha}) D_i(x_i, \beta)[y_i - g(x_i, \beta)] = 0, \tag{3.12}$$

where $w_i(\hat{\alpha}) = 1/p(x_i, z_i; \hat{\alpha})$ and $p(x_i, z_i; \hat{\alpha})$ is an estimate of the probability of being a complete case, obtained for example by a logistic regression of m_i on x_i and z_i. Here α are the parameters of the logistic regression, computed for example by maximum likelihood. If this logistic regression is correctly specified, Eq. (3.12) yields a consistent estimate of β provided

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

which is less restrictive than Eq. (3.11) in that missingness is allowed to depend on z_i as well as x_i. Thus weighted GEE can correct for the bias of unweighted GEE attributable to dependence of the missingness mechanism on z_i. Robins, Rotnitzky and Zhao (1995) discuss variance estimation for weighted GEE estimates, and extensions to monotone and nonmonotone patterns. Additional references to this approach include Manski and Lerman, 1977; Zhao and Lipsitz, 1992; Park, 1993; Robins and Rotnitzky, 1995; and Lipsitz, Ibrahim and Zhao, 1999. Other work considers extensions that recover information in the incomplete cases, and to non-MAR models. See in particular Scharfstein, Rotnitzky and Robins, 1999, and the discussion of that paper. As discussed in Chapter 15, there is never any direct empirical evidence against MAR without assumptions, so it is wise to consider such assumptions carefully.

3.3.2. Added Variance from Nonresponse Weighting

Weighting class adjustments are simple since the same weights are obtained regardless of the survey outcome Y to which they are applied. Thus in large surveys with ignorable nonresponse and hundreds of Ys, 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 variance. A simple formula for the increase in variance of a sample mean is obtained by assuming random sampling within weighting classes, ignoring sampling variation in the weights, and assuming an outcome Y has constant variance σ^2. Then if the weights are scaled to average one:

$$\mathrm{Var}\left(\frac{1}{r}\sum_{i=1}^{r} w_i y_i\right) = \frac{\sigma^2}{r^2}\left(\sum_{i=1}^{r} w_i^2\right) = \frac{\sigma^2}{r}[1 + cv^2(w_i)], \tag{3.13}$$

where $cv(w_i)$ is the coefficient of variation of the weights. Thus the squared coefficient of variation of the weights is a rough measure of the proportion increase in variance due to weighting. This increase in variance may be justified by the reduction in bias for variables that are strongly associated with nonresponse, but is not for variables that are weakly associated with nonresponse. One might contemplate elaborations that tailor the weights according to the degree of association between the outcome and nonresponse, but the simplicity of the method is then lost.

3.3.3. Post-Stratification and Raking to Known Margins

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

EXAMPLE 3.9. *Post-Stratification*. Suppose the population proportions N_j/N are known from external sources. In that case an alternative to the weighting class estimator is the post-stratified mean

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

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

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

An estimate of Eq. (3.15) is obtained by substituting the sample variance among respondents in class j, s_{jR}^2 for the population variance S_{jR}^2. In most circumstances $\bar{y}_{\text{ps}}$ has lower mean squared error than $\bar{y}_{\text{wc}}$, 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 post-stratification and extensions see Little (1993a); Lazzeroni and Little, 1998; Bethlehem, 2002; and Gelman and Carlin, 2002.

EXAMPLE 3.10. *Raking Ratio Estimation*. Suppose now 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. Suppose that n_{jl} units of N_{jl} in the population are sampled in the class with $X_1 = j$, $X_2 = l$, for $j = 1, \ldots, J$, $l = 1, \ldots, L$. The value of a variable Y is recorded for r_{jl} out of the n_{jl} sampled units in class (j, l). The post-stratified and weighting class estimators take the form

$$\bar{y}_{\text{ps}} = \frac{1}{N}\sum_{j=1}^{J}\sum_{l=1}^{L} N_{jl}\bar{y}_{jlR}$$

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

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

The method of *raking* applied to the class counts $\{n_{jl}\}$ consists in calculating estimates $\{N^*_{jl}\}$ of $\{N_{jl}\}$ that satisfy the marginal constraints

$$N^*_{j+} = \sum_{l=1}^{L} N^*_{jl} = N_{j+}, \qquad j = 1, \ldots, J;$$
$$N^*_{+l} = \sum_{j=1}^{J} N^*_{jl} = N_{+l}, \qquad l = 1, \ldots, L$$

and that differ from the observed counts $\{n_{jl}\}$ by row and column factors, that is, can be expressed in the form

$$N^*_{jl} = a_j b_l n_{jl}, \qquad j = 1, \ldots, J; l = 1, \ldots, L$$

for certain row constants $\{a_j, j = 1, \ldots, J\}$ and column constants $\{b_l, l = 1, \ldots, L\}$. The $\{N^*_{jl}\}$ table has margins equal to the known margins $\{N_{j+}\}$ and $\{N_{+l}\}$ but associations equal to those in the $\{n_{jl}\}$ table. The raked class counts $\{N^*_{jl}\}$ can be calculated by an iterative proportional fitting procedure (Bishop, Fienberg and Holland, 1975), where current estimates are scaled by row or column factors to match the marginal totals $\{N_{j+}\}$ and $\{N_{+l}\}$, respectively. That is, at the first step the estimators

$$N^{(1)}_{jl} = n_{jl}(N_{j+}/n_{j+}),$$

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

$$N^{(2)}_{jl} = N^{(1)}_{jl}(N_{+l}/N^{(1)}_{+l}),$$

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

$$N^{(3)}_{jl} = N^{(2)}_{jl}(N_{j+}/N^{(2)}_{j+}),$$

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 $\{N^*_{jl}/N\}$ of the class proportions are best asymptotically normal estimates under a multinomial assumption for the class counts $\{n_{jl}\}$, and as such are asymptotically equivalent to the (harder to calculate) maximum likelihood estimates under the multinomial model.

Combining the raked sample counts $\{N^*_{jl}\}$ with the respondent means $\{\bar{y}_{jl}\}$ yields the raked estimator of $\bar{Y}$:

$$\bar{y}_{\text{rake}} = \frac{1}{N} \sum_{j=1}^{J} \sum_{l=1}^{L} N^*_{jl} \bar{y}_{jlR}, \tag{3.16}$$

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

3.3.4. Inference from Weighted Data

Weighted complete-case estimators 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, formulas are available for estimating the standard errors. For more complex situations, methods based on Taylor Series expansions (Robins, Rotnitsky and Zhao, 1995), balanced repeated replication or jackknifing can be applied. Statistical packages are available for computing asymptotic standard errors of estimators 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. 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. Chapter 5 discusses one version of this approach, the bootstrap, in the context of imputation methods, and the same idea can be applied to weighted estimators.

3.3.5. Summary of Weighting Methods

Weighting is a relatively simple device for reducing bias from complete-case analysis. The methods are simple in that they yield the same weight for all variables measured for each case. Since the methods discard the incomplete cases and do not provide a built-in control of 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 variance.

3.4. AVAILABLE-CASE ANALYSIS

Complete-case analysis is potentially wasteful for univariate analyses such as estimation of means and marginal frequency distributions, since values of a particular variable are discarded when they belong to cases that are missing other variables. The loss in efficiency can be particularly large for datasets 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 cases 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 cases where the variable of interest is present, an option that has been termed *available-case* analysis. The method uses all the available values; its disadvantage is that the sample base changes from variable to variable according to the pattern of missing data. As many processors of large data sets know, this variability in the sample base creates practical problems, particularly when tables are computed for various conceptual sample bases (e.g., all women, ever-married women, currently married women in a demographic fertility survey). The analyst wishes to associate a fixed sample size to each base as a check that the tables are correctly defined. The changes in the sample bases in available-case analysis prevent such simple checks. They also yield problems of comparability across variables if the missing-data mechanism is a function of the variables under study, that is, if the data are not MCAR.

Estimates of means and variances can be calculated under MCAR, using the available-case procedure we have described, but modifications are required to estimate measures of covariation such as covariances or correlations. A natural extension is to *pairwise* available-case methods, where measures of covariation for Y_j and Y_k are based on cases i for which both y_{ij} and y_{ik} are present. In particular, one might compute pairwise covariances:

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

where I_{jk} is the set of $n^{(jk)}$ cases 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 cases. Let $s_{jj}^{(j)}$ denote the sample variance of Y_j over the set I_j of cases with Y_j observed. Combining these variance estimates with the covariances estimated from Eq. (3.17) yields the following estimate of the correlation:

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

A criticism of Eq. (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:

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

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

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

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

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

an estimator originally discussed in Wilks, 1932.

Pairwise available-case estimates such as Eqs. (3.17)–(3.20) attempt to recover some of the information in partially recorded units that is lost by complete-case analysis. Under MCAR, Eqs. (3.17)–(3.20) yield consistent estimates of the covariances and correlations being estimated. When considered collectively, however, the estimates have deficiencies that can severely limit their utility in practical problems.

For example, Eq. (3.18) can yield correlations outside the acceptable range. On the other hand, Eq. (3.19) yields correlations that always lie between ± 1. For $K > 3$ variables, both Eqs. (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 with 12 observations on three 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, since $\text{Corr}(Y_1, Y_2) = \text{Corr}(Y_1, Y_3)$ implies $\text{Corr}(Y_2, Y_3) = 1$, not -1. In the same way, covariance matrices based on Eqs. (3.17) or (3.20) are not necessarily positive definite. Since 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.

Since available-case methods apparently make use of all the data, one might expect them to be more efficient than complete-case 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 complete-case analysis when correlations are large (Haitovsky, 1968; Azen and Van Guilder, 1981). Neither method, however, is generally satisfactory. Although available-case estimates are easy to compute, even asymptotic standard errors are more complex (Van Praag, Dijkstra and Van Velzen, 1985).

PROBLEMS

3.1. List some standard multivariate statistical analyses that are based on the 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 cases yields unbiased estimates of the regression parameters.

3.3. Show that for dichotomous Y_1 and Y_2, the odds ratio based on complete cases is a consistent estimate of the population odds ratio if the log probability of response 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 are interviewed for a health survey, yielding the following data:

			Cholesterol	
Age Group	Sample Size	Number of Respondents	Mean	S.D.
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. Can this interval be applied to 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). Hence construct an approximate 95% confidence interval for the population mean and compare it with the result of Problem 3.4. What assumptions are made about the nonresponse 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%; 50–60: 10%. Calculate the post-stratified estimate of mean cholesterol, its associated standard error, and a 95% confidence interval for the population mean.

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

3.7. Calculate Horvitz–Thompson and weighting class estimators 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 known for the Horvitz–Thompson estimator but unknown for the weighting class estimators. Note that various weighting class estimators could be created.

3.8. Apply the Cassell, Sarndal and Wretman (1983) estimator 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 estimator.

3.9. The following table shows respondent means of an incomplete variable Y (say, 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 gender (male, female). Note that weighting classes cannot be based on age, marital status, and gender, since there is one class with four units sampled, none of which respond. Calculate the following estimates of the mean of Y, both for the whole population and for the subpopulation of males:

(a) The unadjusted mean based on complete cases.

(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 for (b). Explain why adjusted estimates are higher than the unadjusted estimates.

Respondent Means and Response Rates, Classified by Age, Marital Status, and Gender

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

3.10. Generalize the response propensity method in Example 3.7 to a monotone pattern of missing data. (See Little, 1986; Robins, Rotnitsky and Zhao 1995).

3.11. Oh and Scheuren (1983) propose an alternative to the raked estimate $\bar{y}_{\text{rake}}$ in Section 3.3.3, where the estimated counts N_{jl}^{*} are found by raking the respondent sample sizes $\{r_{jl}\}$ instead of $\{n_{jl}\}$. Show that (i) unlike $\bar{y}_{rl}$, this estimator exists when $r_{jl} = 0$, $n_{jl} \neq 0$ for some j, l, and (ii) the estimator is biased unless the expectation of r_{jl}/n_{jl} 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 the previous example) yields the same answer if and only if

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

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 $\{n_{jl}\}$			(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 (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 (3.17) are replaced by estimates from all available cases. Is the resulting estimated correlation (3.19) always in the range $(-1, 1)$? Either prove that it is or provide a counterexample.

3.17. Consider the relative merits of complete-case analysis and available-case analysis for estimating (a) means, (b) correlations, and (c) regression coefficients when the data are not MCAR.

3.18. Review the results of Haitovsky (1968), Kim and Curry (1977), and Azen and Van Guilder (1981). Describe situations where complete-cases analysis is more sensible than available-case analysis, and vice versa.

CHAPTER 4

Single Imputation Methods

4.1. INTRODUCTION

Both complete-case and available-case analysis make no use of cases with Y_j missing when estimating either the marginal distribution of Y_j or measures of covariation between Y_j and other variables. Is this a mistake? Suppose a case 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 items that are missing. These methods can be applied to impute one value for each missing item (single imputation) or, in some cases, to impute more than one value, to allow appropriate assessment of imputation uncertainty (multiple imputation).

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 are means or draws from a predictive distribution of the missing values, and require a method of 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 4.2–4.4), but receive more systematic attention in Parts II and III of the book.

Implicit modeling: the focus is on an algorithm, which implies 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 4.7–4.14.

Explicit modeling methods include:

(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 by predicted values from a regression of the missing item on items observed for the unit, usually calculated from units with both observed and missing variables present. 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 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 versus 0, and the imputed valued 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:

(d) *Hot deck imputation*, which involves substituting individual values drawn from "similar" responding units. Hot deck imputation is common in survey practice and can involve very elaborate schemes for selecting units that are similar for imputation. Despite their popularity in practice, the literature on the theoretical properties of the various methods is very sparse; see for example Ernst (1980), Kalton and Kish (1981), Ford (1983), and David et al. (1986). For a more recent discussion of hot deck applications, see Marker, Judkins and Winglee (2002).

(e) *Substitution*, a method for dealing with unit nonresponse at the fieldwork stage of a survey, replaces nonresponding units with alternative units not 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, since 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 of a particular type.

(f) *Cold deck imputation* replaces a missing value of an item by a constant value from an external source, such as a value from a previous realization of the same 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.

(e) *Composite methods* can also be defined 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), David et. al (1986).

We now discuss some of these methods in more detail. We consider methods that impute the mean of a predictive distribution in Section 4.2, and methods that impute a draw from a predictive distribution in Section 4.3. An important limitation of the single imputation methods described here is that standard variance formulas applied to the filled-in data systematically underestimate the variance of estimates, even if the model used to generate the imputations is correct. Methods that allow valid estimates of the 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 particularly 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 cases. Under MCAR, $s_{jj}^{(j)}$ is a consistent estimate of the true variance, so the sample variance from the filled-in data set underestimates the 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. Since mean imputation distorts the empirical distribution of the sampled Y-values, estimates of quantities that are not linear in the data, such as variances, percentiles or measures of shape, are not estimated consistently using standard complete-data methods applied to the completed data. A related problem occurs if the values of Y_j are grouped into subclasses for cross-tabulation, since 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 cases with Y_j and Y_k observed and $\tilde{s}_{jk}^{(jk)}$ is given by Eq. (3.21). Since under MCAR $\tilde{s}_{jk}^{(jk)}$ is a consistent estimate of the covariance, the estimate from 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 semi-definite, 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 Chapter 3, the resulting covariance matrix is not generally positive definite and tends to be unsatisfactory, particularly when the variables are highly correlated. This method cannot be recommended.

4.2.2. Conditional Mean Imputation

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

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 impute the respondent mean for nonrespondents in the same 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

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

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 post-stratified estimator $\bar{y}_{\mathrm{ps}}$, Equation (3.14), 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 $Y_1, \ldots, Y_{K-1}$ fully observed and Y_K observed for the first r observations and missing for the last $n-r$ observations. Regression imputation computes the regression of Y_K on $Y_1, \ldots, Y_{K-1}$ based on the r complete cases, and then fills in the missing values as predictions from the regression, an approach similar to that in Chapter 2. Specifically, suppose case i has y_{iK} missing and $y_{i1}, \ldots, y_{i,K-1}$ present. The missing value is imputed using the regression equation:

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

where $\tilde{\beta}_{K0\cdot 12\ldots K-1}$ is the intercept and $\tilde{\beta}_{Kj\cdot 12\ldots K-1}$ is the coefficient of Y_j in the regression of Y_K on $Y_1, \ldots, Y_{K-1}$ based on the r complete cases. If the observed variables are dummies for a categorical variable, the predictions (4.1) are respondent means within classes defined by that variable, and the method reduces to that of

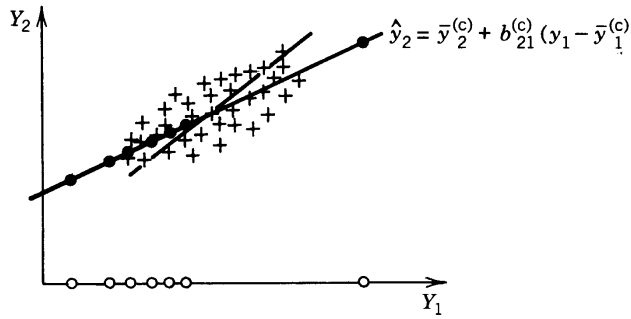


Figure 4.1. Regression imputation for $K = 2$ variables.

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 cases 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, $\hat{y}_{i2} = \tilde{\beta}_{20\cdot1} + \tilde{\beta}_{21\cdot1}y_{i1}$. Cases 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. Cases with Y_2 observed and Y_1 missing would be imputed on the regression line of Y_1 on Y_2, the other line in the diagram.

EXAMPLE 4.3. *Buck's Method.* Buck's method (Buck, 1960) extends regression imputation to a general pattern of missing values, for the case 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 based on the complete cases, and then uses these estimates to calculate the least squares linear regressions of the missing variables on the present 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 cases with each pattern of missing data may appear formidable, but in fact is relatively simple using the sweep operator discussed in Section 7.5.

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 missing-data 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 MAR holds, even though the distribution of Y_1 for complete and incomplete cases is different. Buck's method projects the incomplete cases 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 cases 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 normality assumptions are plausible. The sample covariance matrix from the filled-in data underestimates the sizes of variances and covariances, although the extent of underestimation is less than that obtained when unconditional means are substituted. Specifically, the sample variance of Y_j from data filled in by Buck's method underestimates σ_{jj} by the quantity $(n-1)^{-1}\sum_{i=1}^{n}\sigma_{jj\cdot\text{obs},i}$, where $\sigma_{jj\cdot\text{obs},i}$ is the residual variance from regressing Y_j on the variables present (or observed) in case 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 in case 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 observations, 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

If we assume MCAR and ignore sampling variability of the estimates of the mean and covariance matrix based on complete cases, then the conditional means imputed in Section 4.2.2 are the best point estimates of the missing values in the sense of minimizing the expected squared error. We have seen, however, that adjustments to the sample variances of the filled-in data are required to yield consistent estimates of variances, even under MCAR. More generally, marginal distributions and measures of covariation of the completed data are distorted by mean imputation. This distortion is particularly disturbing when the tails of the distribution or standard errors of estimates are being studied. For example, an imputation method that imputes conditional means for missing incomes tends to underestimate the percentage of cases 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*:

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

where z_{ik} is a random normal deviate with mean 0 and variance $\tilde{\sigma}_{KK\cdot 12\ldots K-1}$, the residual variance from the regression of Y_K on $Y_1, \ldots, Y_{K-1}$ based on the complete cases. 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 data with Y_1 fully observed, Y_2 missing for a fraction $\lambda = (n - r)/n$ cases, 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.
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 cases.
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 all four imputation methods yield consistent estimates of μ_2 under MCAR, 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 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 variance of the Cmean estimate of μ_2, namely $(1 - \lambda\rho^2)\sigma_{22}/r$. Secondly, the standard errors of the Cdraw parameter estimates from the filled-in data are too small, since they do not incorporate imputation

Table 4.1 Bivariate Normal Monotone MCAR Data: Large Sample Bias of Four Imputation Methods as a Function of the Fraction of Missing Data, λ.

	Parameter			
Method	μ_2	σ_{22}	$\beta_{21\cdot1}$	$\beta_{12\cdot2}$
Umean	0*	$-\lambda\sigma_{22}$	$-\lambda\beta_{21\cdot1}$	0*
Udraw	0	0	$-\lambda\beta_{21\cdot1}$	$-\lambda\beta_{12\cdot2}$
Cmean	0	$-\lambda(1-\rho^2)\sigma_{22}$	0*	$\frac{\lambda(1-\rho^2)}{1-\lambda(1-\rho^2)}\beta_{12\cdot2}$
Cdraw	0	0	0	0

*Estimator is same as CC estimate.

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 cases 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.

4.3.2. Draws Based on Implicit Models

With most hot-deck procedures (and here we cannot be precise since the term does not have a well-defined common usage), missing values are replaced by values from similar responding units in the sample. The hot deck literally refers to the deck of matching computer 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 let us label the first n units, $i = 1, \ldots, n$ as sampled, and the first $r < n$ units as respondents. Given an equal probability sampling scheme, the

mean Y may be estimated as the mean of the responding and the imputed units. This may be written in the form

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

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 $\bar{y}_{\mathrm{HD}}$ depend on the procedure used to generate the numbers $\{H_1, \ldots, H_r\}$. 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 $\{H_1, \ldots, H_r\}$ in repeated applications of the hot-deck method is known. The mean and variance of $\bar{y}_{\mathrm{HD}}$ can then be written as:

$$E(\bar{y}_{\mathrm{HD}}) = E[E(\bar{y}_{\mathrm{HD}}|Y_{\mathrm{obs}})], \tag{4.4}$$

$$\mathrm{Var}(\bar{y}_{\mathrm{HD}}) = \mathrm{Var}[E(\bar{y}_{\mathrm{HD}}|Y_{\mathrm{obs}})] + E[\mathrm{Var}(\bar{y}_{\mathrm{HD}}|Y_{\mathrm{obs}})], \tag{4.5}$$

where the inner expectations and variances are over the distribution of $\{H_1, \ldots, H_r\}$ given the observed data Y_{obs}, 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 Eq. (4.5) represents the added 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 4.10–4.13.

EXAMPLE 4.7. *The Hot Deck by Simple Random Sampling with Replacement.* Let $\bar{y}_{\mathrm{HD1}}$ denote the hot-deck estimator (4.3) when the $\{H_i\}$ are obtained by random sampling with replacement from the recorded values of Y. Conditioning on the sampled and recorded values, the distribution of $\{H_1, \ldots, H_r\}$ in repetitions of the hot deck is multinomial with sample size $n - r$ and probabilities $(1/r, \ldots, 1/r)$. (See Cochran, 1977, Section 2.8). Hence the moments of the distribution of $\{H_1, \ldots, H_r\}$ given the observed data Y are:

$$E(H_i|Y_{\mathrm{obs}}) = (n-r)/r,$$
$$\mathrm{Var}(H_i|Y_{\mathrm{obs}}) = (n-r)(1-1/r)/r,$$
$$\mathrm{Cov}(H_i, H_j|Y_{\mathrm{obs}}) = -(n-r)/r^2.$$

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

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

and

$$\text{Var}(\bar{y}_{\text{HD1}}|Y_{\text{obs}}) = (1 - r^{-1})(1 - r/n)s^2_{yR}/n. \tag{4.7}$$

In particular, assuming simple random sampling from a finite population of size N and missing data that are MCAR, Eqs. (4.4) and (4.5) yield

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

where 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 the imputed values do not distort the distribution of the sampled values of Y the way mean imputation does.

The added variance (4.7) from sampling imputations with replacement is a non-negligible quantity. Specifically, the proportionate variance increase of $\bar{y}_{\text{HD1}}$ 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 4.12). However, we prefer multiple imputation, as discussed in Chapter 5, to these approaches, since it not only reduces the increase in variance to negligible levels, but also provides valid standard errors that take into account imputation uncertainty.

The hot-deck estimators in Example 4.7 are unbiased only under the generally unrealistic MCAR assumption. If covariate information is available for responding and nonresponding units, then this information may be used to reduce nonresponse bias. Two approaches are worthy of mention:

EXAMPLE 4.8. *Hot Deck Within Adjustment Cells.* Adjustment cells may be formed, and missing values within each cell replaced by recorded 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. 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. Since adjustment cells are formed from the joint levels of categorical variables, they are not ideal for interval-scaled variables. The Census Bureau uses this method for imputing earnings items in the Income Supplement of the Current Population Survey (CPS) (Hanson, 1978). For each nonrespondent on one or more income items, the CPS hot deck finds a

matching respondent based on variables that are observed for both; the missing items 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 giant matrix. When no match can be found for a nonrespondent based on all of the variables, 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 4.9. *Nearest Neighbor Hot Deck.* A more general approach is to define a metric to measure distance between units, based on the values of covariates, and then to choose imputed values that come from responding units close to the unit with the missing value. For example, let $x_i = (x_{i1}, \ldots, x_{iK})^{\mathrm{T}}$ be the values of K appropriately scaled covariates for a unit i for which y_i is missing. If these variables are used to form adjustment cells, 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 the previous example. Other possible metrics are:

Maximum deviation:	$d(i,j) = \max_k \lvert x_{ik} - x_{jk} \rvert$;
Mahalanobis:	$d(i,j) = (x_i - x_j)^{\mathrm{T}} S_{xx}^{-1} (x_i - x_j)$, where S_{xx} is an estimate of the covariance matrix of x_i.

The metric need not be full rank, in the sense of only giving zero distance when (i,j) have $x_i = x_j$. For example, consider the following metric:

Predictive Mean: $d(i,j) = [\hat{y}(x_i) - \hat{y}(x_j)]^2$,

where $\hat{y}(x_i)$ is the predicted value of Y from the regression of Y on the x's computed using the complete cases. We choose an imputed value for y_i from those units j that are such that (1) $y_j, x_{j1}, \ldots, x_{jK}$ are observed, and (2) $d(i,j)$ is less than some value d_0. Varying the value of d_0 can control the number of available candidates j. A substantial statistical literature on matching methods 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). Since imputed values are relatively complex functions of the responding items, quasi-randomization properties of estimates derived from such matching procedures remain largely unexplored.

EXAMPLE 4.10. *A Sequential Hot Deck Ordered by a Covariate.* Colledge et al. (1978) present a case study where the hot-deck method is used extensively in a

Canadian survey of construction firms. The survey involved 50,538 firms, of which 41,432 were retained for the analysis. The survey items are divided into four groups: (a) Fully observed *key fields* from the tax return, including region, standard industrial classification (SIC), gross business income (GBI), net business income (NBI), and salary and wages indicator (SWI); (b) *basic financial* variables from the tax return, which were sometimes missing; (c) secondary financial variables, which were missing relatively frequently; and (d) *survey* variables collected for different but overlapping subsamples, and sometimes missing. Only 908 of the 41,432 records provided information in all four variable groups: most of the records, 34,181, had only key fields observed; 2316 records had key fields and basic financial variables observed; and 4027 had key fields and survey variables observed. Hot-deck imputation was conducted in several phases, where each phase replaced missing items in a particular group by the values of donor records for which all the items in that group were observed. This method assumes implicitly that the groups of variables are conditionally independent, given the matching variables. In order to match donors with candidates, the field of all records was post-stratified by province (or region), SIC, and SWI. The collection of donors (i.e., the hot deck) and the collection of candidates were identified for the particular phase. Within each post-stratum, the records were ordered by GBI. In order to impute values for a particular candidate record i, only the nearest five potential donors j on either side were considered, yielding ten possible donors with approximately the same value of GBI. From these ten donors, one was chosen to minimize a distance of the basic form

$$d(i, j) = |\ln TEXP_i - \ln TEXP_j|,$$

where $TEXP = GBI - NBI =$ total expenses was used because many of the fields to be imputed were detailed expense breakdowns or were highly correlated with expenses. The matching of donor to candidate was based entirely on key fields, which are all observed. More generally, the distance was expanded to depend on fields other than *TEXP*, and modified to spread donor usage by making it an increasing function of the number of times the potential donor j had already been used as an actual donor in the phase.

After a donor had been identified, the candidate's missing fields for the group of variables in this phase were replaced by the corresponding fields from the donor record. Some adjustment or transformation was sometimes necessary to ensure that certain edit constraints were satisfied. For example, suppose three fields x, y, and z have to satisfy $x + y \leq z$, with x, y, and z non-negative. The donor's values for these fields are x_j, y_j and z_j, whereas the candidate has only z_i present. If the values x_j and y_j are simply written into the corresponding candidate's fields, we may find that $x_j + y_j > z_i$, which violates the edit constraint. In this situation x_j and y_j were prorated to satisfy the edit constraint, by substituting $x_i = (x_j/z_j)z_i$ and $y_i = (y_j/z_j)z_i$. In other words, the proportions x_j/z_j and y_j/z_j were transferred to the candidate rather than the values x_j, y_j.

EXAMPLE 4.11. *Imputation Methods for Repeated Measures with Dropouts.* Longitudinal data are often subject to attrition when subjects leave the study

prematurely. Let $y_i = (y_{i1}, \ldots, y_{ik})$ be a $(K \times 1)$ complete-data vector of outcomes for subject i, possibly incompletely observed. Write $y_i = (y_{\text{obs},i}, y_{\text{mis},i})$, where $y_{\text{obs},i}$ = observed part of y_i, $y_{\text{mis},i}$ = missing part of y_i. Let M_i denote a missing-data indicator, with $M_i = 0$ for complete cases and $M_i = k$ if a subject drops out between the $(k-1)$th and kth observation time, that is, $y_{i1}, \ldots, y_{i,k-1}$ are observed and $y_{ik}, \ldots, y_{iK}$ are missing.

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

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

Although simple, this approach makes the strong assumption that the value of the outcome remains unchanged after dropout, which seems likely to be unrealistic in many settings. Alternative imputation methods retain the advantage of simplicity but provide for subject and time effects. For example, imputes might be based on the result of a row + column fit. For case i with $M_i = k$, let $\bar{y}_{\text{obs},i} = (k-1)^{-1} \sum_{t=1}^{k-1} y_{it}$ denote the mean for the available measurements for subject i, and let $\bar{y}^{(\text{cc})}_{\text{obs},+} = r^{-1} \sum_{l=1}^{r} \bar{y}_{\text{obs},l}$ be the corresponding mean averaged over the set of r complete cases. Let $\bar{y}^{(\text{cc})}_{+t} = r^{-1} \sum_{l=1}^{r} y_{lt}$ be the complete-case mean for time point t. The prediction from a row + column fit to the complete cases is:

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

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

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

which simplifies to:

$$\hat{y}_{it} = y_{lt} + (\bar{y}_{\text{obs},i} - \bar{y}_{\text{obs},l}, \qquad t = k, \ldots, K,$$

the simple hot-deck draw y_{lt} modified by a subject effect $(\bar{y}_{\text{obs},i} - \bar{y}_{\text{obs},l})$. 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:

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

These methods extend simply to a general pattern of missing data. For an application to a panel survey of income, see Little and Su (1989). See also Lavori, Dawson and Shera (1995) for multiple imputation of longitudinal missing data based on the propensity stratification methods discussed in Example 3.7.

4.4. CONCLUSIONS

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 the approaches discussed in the chapter is that inferences about parameters 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 small and confidence intervals are too narrow. In the next chapter we consider two approaches to this problem that have relatively general applicability, replication methods and multiple imputation. Multiple imputation has the added bonus of largely correcting the disadvantage of imputing draws from the predictive distribution, namely the loss of precision.

PROBLEMS

4.1. Consider a bivariate sample with $n = 45$, $r = 20$ complete cases, 15 cases with only Y_1 recorded, and 10 cases 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}); (c) the slope of the regression of Y_2 on Y_1 (σ_{12}/σ_{11}). You can ignore bias terms of order $1/n$.

4.2. Repeat the previous example 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 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 $\theta = (\mu_1, \sigma_{11}, \beta_{20\cdot13}, \beta_{21\cdot13}, \beta_{23\cdot13}, \sigma_{22\cdot13})$ is $N(\mu_1, \sigma_{11})$ and Y_2 given Y_1 and θ is $N(\beta_{20\cdot13} + \beta_{21\cdot13}Y_1 + \beta_{23.13}Y_1^2, \sigma_{22\cdot13})$. The data are MCAR, the first r cases are complete, the next r_1 cases record Y_1 only, and the last r_2 cases record 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 and the distribution of the variables has finite fourth moments.

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 large-sample variance of the Cmean and Cdraw estimates of μ_2 in the discussion of Example 4.5.

4.10. Derive the expressions (4.6)–(4.8) for the simple hot deck where imputations are by simple random sampling with replacement. Show that the proportionate variance increase of $\bar{y}_{\mathrm{HD1}}$ 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 defined by a covariate?

4.11. Consider a hot deck like that of Example 4.7, 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 non-negative 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

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

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

$$E(\bar{y}_{\mathrm{HD2}}|Y_{\mathrm{obs}}) = \bar{y}_R$$

and

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

Show that the proportionate variance increase of $\bar{y}_{\text{HD2}}$ 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 present 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 necessary, perhaps chosen from records in a prior survey. This method formed the basis for early imputation schemes for the Census Bureau's Current Population Survey.

Suppose sampled units are regarded as randomly ordered, units are selected by simple random sampling, and a Bernoulli nonresponse mechanism is operating. Then show that the sequential hot-deck estimate of Y, say $\bar{y}_{\text{HD3}}$, is unbiased for $\bar{Y}$ with variance (for large r and n and ignoring finite population corrections) given by

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

Hence the proportionate increase in variance over $\bar{y}_R$ is $(n - r)/n$, the fraction of missing data. (See Bailar, Bailey and Corby, 1978, for details).

4.13. Which of the metrics in Example 4.9 give the best imputations for a particular outcome Y? Propose an extension of the predictive mean matching metric to handle a set of missing outcomes $Y_1, \ldots, Y_K$. (See Little, 1986, for details).

4.14. Outline a situation where the "Last Observation Carried Forward" method of Example 4.11 gives poor estimates. (See, for example, Little and Yau, 1996).

4.15. 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 cases;

(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 cases into quartiles based on the distribution of Y_1.
Suggest a situation where (d) might be a superior method to (c).

CHAPTER 5

Estimation of Imputation Uncertainty

5.1. INTRODUCTION

Most of our discussion of imputation methods in Chapter 4 concerned point estimation of population quantities in the presence of nonresponse. In this section we focus on the question of deriving estimates of uncertainty that incorporate the added variance due to nonresponse. The variance estimates presented here all effectively assume that the method of adjustment for nonresponse has succeeded in eliminating nonresponse bias. It is important to emphasize that in many applications the issue of nonresponse bias is often more crucial than that of variance. In fact, it can be argued that providing a valid estimate of sampling variance is worse than providing no estimate if the estimator has a large bias, which dominates the mean squared error.

We distinguish four general approaches to accounting for the additional uncertainty:

1. Apply explicit variance formulas that allow for nonresponse. For instance, 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, the explicit formula (3.7) for mean squared error can be applied to estimate the precision, with the corresponding confidence interval $\bar{y}_{\text{wc}} \pm z_{1-\alpha/2}\{\hat{mse}(\bar{y}_{\text{wc}})\}^{1/2}$. Equation (4.8) gives the variance of the hot deck estimator by simple random sampling with replacement, and this formula can be modified to yield an estimated variance when the hot deck is 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 4.10, 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 dataset. 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 modifications required to the imputations may compromise the quality of the 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, Chen and Chen, 1998; Shao, 2002; Lee, Rancourt and Sarndal, 2002). Uncertainty is estimated from variability of point estimates of parameters 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 approaches 1 or 2, but provides consistent 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 extensive than the resampling methods since the multiple imputes are only used to determine the added uncertainty from the incomplete data. The method is particularly useful for data base construction, where a single data base 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 consistent estimates of variance is not a simple task even with complete response. As a result, approximate methods have been developed that can be applied to estimates of functions of means and totals for broad classes of sample designs. The simplicity of these methods results from restricting calculations to quantities calculated for collections of the sampled units known as ultimate clusters (UCs). These 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 one, and some non-self-representing EAs, which are sampled from the population of EAs. The ultimate clusters then consist of the non-self-representing EAs and the sampling units that form the first stage of subsampling of self-representing EAs.

Estimates of variance calculated from estimates for UCs are based on the following lemma:

Lemma. Let $\hat{\theta}_1, \ldots, \hat{\theta}_k$ be random variables that are (1) uncorrelated and (2) have common mean μ. Let

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

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

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

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

But

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

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

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 random sampling with replacement. Let t_j denote the total for a variable Y in UC j, and suppose we estimate the population total

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

by the Horvitz–Thompson estimate

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

where the sum is over selected UCs (say $j = 1, \ldots, k$), $\hat{t}_j$ is an unbiased estimate of t_j, and π_j is the probability that UC j is selected. Then (1) $\hat{t}_{\mathrm{HT}}$ and $\{k\hat{t}_j/\pi_j, j = 1, \ldots, k\}$ are all unbiased estimates of T, and (2) the estimates $\{k\hat{t}_j/\pi_j, j = 1, \ldots, k\}$ are uncorrelated, by the method of random sampling with replacement. Hence by the lemma,

$$\hat{v}(\hat{t}_{\mathrm{HT}}|Y_{\mathrm{obs}}) = \sum_{j=1}^{k} \frac{(k\hat{t}_j/\pi_j - \hat{t})^2}{k(k-1)} \tag{5.2}$$

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

Suppose now we have missing data, 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 variance, provided:

Condition 5.1. The estimates $\hat{t}_j$ are unbiased for t_j, that is, the imputation or weighting procedure does not lead to nonresponse bias within UC j; and

Condition 5.2. The imputations or weighting adjustments are carried out *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 cut across the 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 variance conflicts with the need for an estimator with an acceptably small bias, at least using the techniques discussed thus far.

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 variance. This 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 overestimate, but in fact this leads to an underestimate. An unbiased estimate requires information from the second and higher stages of sampling. Thus simple variance estimates based on UCs require that the proportion of UCs sampled is small, so that the overestimation introduced by sampling without replacement can be ignored. This is often the case 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 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 for UC j in stratum h, for $h = 1, \ldots, H$, $j = 1, \ldots, K_h$. We can estimate t by

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

say, 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 variance of $\hat{t}$ is estimated by

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

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

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

Conditions for using these estimates with imputed data are the same as those for random sampling. That is, each $\hat{t}_{hl}$ must be unbiased for t_{hl}, and the imputations must be carried out independently in each UC. We now consider alternative methods that relax the severe restriction that the imputations across UC must be 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 $S = \{i: i = 1, \ldots, n\}$ of independent observations. Let $S^{(b)}$ be a sample of size n obtained from the original sample S by simple random sampling *with replacement*, 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 $(\hat{\theta}^{(1)}, \ldots, \hat{\theta}^{(B)})$ 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:

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

Large-sample precision can be estimated from the bootstrap distribution of $\hat{\theta}^{(b)}$, which is estimated by the histogram formed by the bootstrap estimates $(\hat{\theta}^{(1)}, \ldots, \hat{\theta}^{(B)})$. In particular, the bootstrap estimate of the variance of $\hat{\theta}$ or $\hat{\theta}_{\text{boot}}$ is

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

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

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

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

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

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 Eq. (5.7) require bootstrap samples of the order of $B = 200$. Intervals based on Eq. (5.8) require much larger samples, for example $B = 2000$ or more (Efron, 1994). Efron (1987, 1994) discusses refinements of Eqs. (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 observation i is included in the bth bootstrap sample, with $\sum_{i=1}^{n} m_i^{(b)} = n$. Then for simple random sampling with replacement,

$$(m_1^{(b)}, \ldots, m_n^{(b)}) \sim MNOM(n; (n^{-1}, n^{-1}, \ldots, n^{-1})), \tag{5.9}$$

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 observation 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 Imputed Incomplete Data.* Suppose the data are a sample $S = \{i: i = 1, \ldots, n\}$ of independent observations, but some observations i are incomplete. 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 $\hat{S} = \text{Imp}(S)$, and then estimating θ from the filled-in data $\hat{S}$. Bootstrap estimates $(\hat{\theta}^{(1)}, \ldots, \hat{\theta}^{(B)})$ can be computed as follows:

For $b = 1, \ldots B$:

(a) Generate a bootstrap sample $S^{(b)}$ from the original unimputed sample S, with weights as in (5.9).

(b) Fill in the missing data in $S^{(b)}$ by applying the imputation procedure to the bootstrap sample $S^{(b)}$, $\hat{S}^{(b)} = \text{Imp}(S^{(b)})$.

(c) Compute $\theta^{(b)}$ on the filled-in data $\hat{S}^{(b)}$ from (b).

Then Eq. (5.6) provides a consistent estimate of the variance of $\hat{\theta}$, and Eqs. (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 would be to apply the imputation procedure just once to yield an imputed data set $\hat{S}$, and then bootstrap the estimation method applied to the filled-in data. However, this approach clearly does not propagate the uncertainty in the imputations and hence does not provide valid inferences. 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 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 adjacent cells or some other modification applied.

5.3.2. Jackknife Standard Errors

Quenoulle'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

observation or set of observations 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 a parameter θ based on a sample $S = \{i: i = 1, \ldots, n\}$ of independent observations. Let $S^{(\setminus j)}$ be a sample of size $n-1$ obtained by dropping the jth observation from the original sample, and let $\hat{\theta}^{(\setminus j)}$ be the estimate of θ from this reduced sample. The quantity

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

is called a pseudovalue. The jackknife estimate of θ is the average of the pseudovalues:

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

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

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

Observe that the multiplier $(n-1)/n$ of $(\theta^{(\setminus j)} - \bar{\theta})^2$ in Eq. (5.12) is larger than the corresponding multiplier $1/(B-1)$ of $(\theta^{(b)} - \bar{\theta})^2$ 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, since they differ from the computation of $\hat{\theta}$ in the treatment of a single observation. It can be shown that under certain conditions Eqs. (5.11) and (5.12) have properties analogous to those of the bootstrap. That is, (a) the jackknife estimator $\hat{\theta}_{\text{jack}}$ is less biased than the original estimator $\hat{\theta}$, and under quite general conditions (b) $\hat{V}_{\text{jack}}$ is a consistent estimate of the variance of $\hat{\theta}$ or $\hat{\theta}_{\text{jack}}$ as n tends to infinity. From property (b), if the jackknife distribution is approximately normal, a $100(1-\alpha)\%$ confidence interval for a scalar θ can be computed as

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

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

EXAMPLE 5.6. *The Simple Jackknife Applied to Imputed Incomplete Data.* Suppose as in Example 5.4 the data are a sample $S = \{i: i = 1, \ldots, n\}$ of independent observations, but some observations i are incomplete. A consistent estimate $\hat{\theta}$ of a parameter 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 from the filled-in data $\hat{S}$. The jackknife can be implemented as follows:

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

(a) Delete observation j from S, yielding the sample $S^{(\setminus j)}$.

(b) Fill in the missing data in $S^{(\setminus j)}$ by applying the imputation procedure Imp, yielding $\hat{S}^{(\setminus)} = \text{Imp}(S^{(\setminus j)})$.

(c) Compute $\hat{\theta}^{(\setminus j)}$ on the filled-in data $\hat{S}^{(\setminus j)}$ from (b).

Equations (5.10)–(5.12) then provide a consistent estimate of the variance of $\hat{\theta}$, and Eq. (5.13) generates a confidence interval for a scalar estimand. 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 observations, where $n = JK$, and then blocks of approximately K observations 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* (Rao and Shao, 1992; Rao, 1996; Fay, 1996). (*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. The jackknife where individual sample observations are dropped does not yield valid standard errors since observations 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 $h = 1, \ldots, H, j = 1, \ldots, K_h$. With complete data $\mathscr{S}$ 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 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 vector of totals for stratum h. To apply the jackknife with no missing data, let $\hat{t}^{(\setminus hj)}$ be the estimate of T with UC hj deleted. That is:

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

where

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

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 variance of $\hat{\theta} = \theta(\hat{t})$ is

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

where

$$\hat{\theta}^{(\backslash hj)} = \theta(\hat{t}^{(\backslash hj)}) \tag{5.15}$$

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 estimate (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 $\hat{S} = \hat{S}(I)$ 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 the variance as follows:

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

(a) Delete UC hj from S, yielding the depleted sample $S^{(\backslash hj)}$.

(b) Fill in the missing data in $S^{(\backslash hj)}$ by applying the imputation procedure Imp to $S^{(\backslash hj)}$, yielding $\hat{S}^{(\backslash hj)}$).

(c) Compute $\hat{t}^{(\backslash hj)}$ on the filled-in data $\hat{S}^{(\backslash hj)}$ from (b). That is:

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

where (5.16)

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

(d) Estimate the variance 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 replicate. In particular, unlike the complete-data case, estimates of totals in strata other than h are potentially affected when UC hj is removed, since imputations in those strata may be affected by the

depletion of the donor pool. Rao and Shao (1992), and Fay (1996) provide formulas for the adjustments to simple imputation methods that result 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 data set, and so on. Standard complete-data methods are used to analyze each 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 nonresponse, the D complete-data inferences can be combined to form one inference that properly reflects uncertainty due to nonresponse under that model. When the imputations are from two or more models for nonresponse, the combined inferences under the models can be contrasted across models to display the sensitivity of inference to models for nonresponse, a particularly critical activity when nonignorable nonresponse is being entertained.

MI was first proposed in Rubin (1978b), and a comprehensive treatment is given in Rubin (1987a). Other references include Rubin (1986), Herzog and Rubin (1983), Rubin and Schenker (1986), and Rubin (1996). 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 nonresponse. 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, because 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 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 nonresponse or uncertainty about the correct model for nonresponse.

MI shares both advantages of single imputation and rectifies both disadvantages. Specifically, when the D imputations are repetitions under one model for nonresponse, the resulting D complete-data analyses can be easily combined to create an inference that validly reflects sampling variability because of the missing values, and 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. The extra work in analyzing the data, however, is really quite modest in today's computing environments, since 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 Y_{mis} are D repetitions from the posterior predictive distribution of Y_{mis}, each repetition corresponding to an independent drawing of the parameters and 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 (1) an explicit regression model and (2) 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, $d = 1, \ldots, D$ be D complete-data estimates and their associated variances for an estimated parameter θ, calculated from D repeated imputations under one model. The combined estimate is

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

Since 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 Eq. (5.17) increases the efficiency of estimate over that obtained from a single data set with conditional draws imputed.

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

$$\bar{W}_D = \frac{1}{D} \sum_{d=1}^{D} W_d, \tag{5.18}$$

and the between-imputation component,

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

[with vector θ, $(\cdot)^2$ replaced by $(\cdot)^{\mathrm{T}}(\cdot)$ in Eq. (5.19)]. The total variability associated with $\bar{\theta}_D$ is

$$T_D = \bar{W}_D + \frac{D+1}{D} B_D, \tag{5.20}$$

where $(1 + 1/D)$ is an adjustment for finite D. Hence

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

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,

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

where the degrees of freedom,

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

is based on a Satterthwaite approximation (Rubin and Schenker, 1986; Rubin, 1987a). An improved expression for the degrees of freedom for small data sets is:

$$v^* = (v^{-1} + \hat{v}_{\text{obs}}^{-1})^{-1},$$

where

$$\hat{v}_{\text{obs}} = (1 - \hat{\gamma}_D)\left(\frac{v_{\text{com}} + 1}{v_{\text{com}} + 3}\right)v_{\text{com}}, \tag{5.24}$$

and v_{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 Eqs. (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 with up to 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, $\bar{Y}$ can be estimated by the stratified mean

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

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

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

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 $\sum_{h=1}^{H}(n_h - r_h)$ missing 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 $s_{h(d)}^2$, $d = 1, \ldots, D$. The MI estimate (5.17) of $\bar{Y}$ is the average of the D complete-data estimates of $\bar{Y}$:

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

From Eqs. (5.18)–(5.20), the variability associated with $\hat{\bar{Y}}_{\mathrm{MI}}$ is the sum of the two components:

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

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

If the missing data mechanism within each stratum is missing completely at random, imputation is not needed; the best estimate of $\bar{Y}$ (in the absence of additional covariate information) is the stratified respondent mean

$$\hat{\bar{Y}}_{\mathrm{ST}} = \sum_{h=1}^{H} P_h \bar{y}_{hR}, \tag{5.28}$$

with associated variance:

$$\mathrm{Var}(\bar{y}_{\mathrm{ST}}) = \sum_{h=1}^{H} P_h^2 \left(1 - \frac{r_h}{N_h}\right) \frac{s_{hR}^2}{r_h}, \tag{5.29}$$

where $\bar{y}_{hR}$ and s_{hR}^2 are the respondent mean and variance in stratum h. A desirable property of a MI method (in the absence of supplemental information) is that it

reconstructs this estimate and associated variance as the D tends to infinity. Since 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, $\hat{\bar{Y}}_{\mathrm{MI}} \to \hat{\bar{Y}}_{\mathrm{ST}}$ as $D \to \infty$, so MI yields the appropriate estimate as the number of imputations tends to infinity. However, the MI variance given by Eq. (5.27) is less than the variance of the stratified respondent mean (5.29), even for infinite D. The source of the problem is that hot deck imputation does 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 post-stratified estimator, but also the correct associated variance.

First consider a method based on an implicit model, which is called the *approximate Bayesian Bootstrap* by Rubin and Schenker (1986). For $d = 1, \ldots, D$ 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 (1987a) 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 variance (5.29) in this case. Another MI approach that yields the appropriate estimator (5.28) and associated variance (5.29) is to impute using the Bayesian predictive distribution under a normal 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 flat 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 parameters.
2. In large samples where asymptotic arguments apply, resampling methods yield consistent estimates of variance with minimal modeling assumptions, whereas MI estimates of variance tend to be more closely tied to a particular model for the data and missing-data 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 since they are apparently less dependent on parametric modeling assumptions. This 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 when these are deemed appropriate. 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. In contrast, 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, Smith and Holt, 1989).
6. MI is more useful than resampling methods for multiple-user data base construction, since a data set with a relatively small set of MIs (say 10 or less) can allow users to derive excellent inferences for a broad range of estimands with 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 observations $\{(y_{i1}, y_{i2}), i = 1, \ldots, 100\}$ on (Y_1, Y_2) as follows:

$$y_{i1} = 1 + z_{i1}$$
$$y_{i2} = 5 + 2^{*}z_{i1} + z_{i2},$$

where $\{(z_{i1}, z^{i_2}),\ i = 1, \ldots, 100\}$ are independent standard normal (mean 0, variance 1) deviates. The observations (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 formulas (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 $u_i < 0$. Since U depends on Y_1 but not Y_2 this mechanism is MAR, and since U has mean zero, 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 cases. Compute s.e.'s 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 (i.e., for each replication of incomplete data, do entire imputation and estimation). Which of these methods yield 90% intervals that actually cover the true parameter? 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 zero and variance given by the estimated residual variance to the conditional mean imputations.

5.4. For the data in Problem 5.3, create 10 multiply-imputed data sets with different sets of conditional draws of the parameters, using the method of Problem 5.3. Compute 90% confidence 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, since 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. They should be a bit wider than 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 confidence intervals with the simpler intervals based on Eq. (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 confidence coverage of intervals—valid intervals should exclude the true parameter in no more than about 50 of the 500 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 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 dth 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 $\bar{\bar{y}}_*$. 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 $D^{-1}\text{Var}(\bar{Y}_*) + (1 - D^{-1})\text{Var}(\bar{y}_R)$, 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 N/r.

(d) Show that the 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 5.10, 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 dth multiply-imputed dataset, 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 Missing Data

CHAPTER 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, Carlin, Stern and Rubin (1995).

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 $f(Y|\theta)$, 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 $f(Y|\theta)$ 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, $f(Y|\theta)$ is a function of Y that gives the probabilities or densities of various Y values.

Definition 6.1. Given the data value Y, the *likelihood function* $L(\theta|Y)$ is any function of $\theta \in \Omega_\theta$ proportional to $f(Y|\theta)$; by definition, $L(\theta|Y) = 0$ 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 Y, whereas the probability or density 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, since 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* $\ell(\theta|Y)$ is the natural logarithm (ln) of the likelihood function $L(\theta|Y)$.

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 observations, $Y = (y_1, \ldots, y_n)^{\mathrm{T}}$, from a normal population with mean μ and variance σ^2 is

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

For given Y, the loglikelihood function is

$$\ell(\mu, \sigma^2|Y) = \ln[f(Y|\mu, \sigma^2)],$$

or ignoring the additive constant,

$$\ell(\mu, \sigma^2|Y) = -\frac{n}{2}\ln\sigma^2 - \frac{1}{2}\sum_{i=1}^{n}\frac{(y_i - \mu)^2}{\sigma^2}, \tag{6.1}$$

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 observations from the exponential distribution with mean $\theta > 0$ is

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

Hence the loglikelihood of θ is

$$\ell(\theta|Y) = \ln\left[\theta^{-n}\exp\left(-\sum_{i=1}^{n}\frac{y_i}{\theta}\right)\right] = -n\ln\theta - \sum_{i=1}^{n}\frac{y_i}{\theta}, \tag{6.2}$$

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

EXAMPLE 6.3. *Multinomial Sample.* Suppose $Y = (y_1, \ldots, y_n)^{\mathrm{T}}$ where y_i is categorical and takes one of C possible values $c = 1, \ldots, C$. Let n_c be the number of observations for which $y_i = c$, with $\sum_{c=1}^{C} n_c = n$. Conditional on n, the counts $(n_1, \ldots, n_C)$ have a multinomial distribution with index n and probabil-

ities $\theta = (\pi_1, \ldots, \pi_{C-1})$ and $\pi_C = 1 - \pi_1 - \cdots - \pi_{C-1}$. Then

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

The loglikelihood of θ is then:

$$\ell(\theta|Y) = \left(\sum_{c=1}^{C-1} n_c \ln \pi_c\right) + n_C \ln(1 - \pi_1 - \cdots - \pi_{C-1}). \tag{6.3}$$

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

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

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

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 values for observation i. The likelihood of $\theta = (\mu, \Sigma)$ is Eq. (6.4) considered as a function of (μ, Σ) for fixed observed Y.

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

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

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 $L(\theta'|Y) = 2L(\theta''|Y)$; 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 $L(\hat{\theta}|Y) \geq L(\theta|Y)$ 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 leads to interest in the value of θ that maximizes the likelihood function. Further and more formal motivation is given in Section 6.2.

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

This definition is phrased to allow the possibility of more than one ML estimate. For many standard and important models, 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,

$$D_\ell(\theta) = 0, \qquad \text{where } D_\ell(\theta) = \partial\ell(\theta|Y)/\partial\theta,$$

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, in fact, a set of d simultaneous equations, defined by differentiating $\ell(\theta|Y)$ 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 Eq. (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 Eq. (6.3). Differentiating with respect to π_c gives the likelihood equation

$$\frac{\partial\ell(\theta|Y)}{\partial\pi_c} = \frac{n_c}{\pi_c} - \frac{n_C}{1 - \pi_2 - \cdots - \pi_{C-1}} = 0,$$

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 Eq. (6.1), the loglikelihood for a normal sample of n observations is

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

where $s^2 = \sum_{i=1}^{n}(y_i - \bar{y})^2/(n-1)$, 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

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

which, since $\hat{\mu} = \bar{y}$, 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 for the mean. Thus we obtain the ML estimates

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

EXAMPLE 6.8. *Multivariate Normal Sample (Example 6.4 continued).* Standard calculations in multivariate analysis (cf. Wilks, 1963; Rao, 1972; Anderson, 1965) show that maximizing Eq. (6.4) with respect to μ and Σ yields

$$\hat{\mu} = \bar{y}, \qquad \hat{\Sigma} = S/n,$$

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

The following property of ML can be important 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 $L(g^{-1}(\phi)|Y)$, which is maximized when $\phi = g(\hat{\theta})$. If $g(\theta)$ is not one-to-one (for example, the first component of θ), the property follows by defining a new one-to-one function of θ, $g^*(\theta) = (g(\theta), h(\theta))$ say, and applying the above argument to g^*.

EXAMPLE 6.9. *A Conditional Distribution Derived from a Bivariate Normal Sample.* The data consist of n independent and identically distributed observations (y_{i1}, y_{i2}), $i = 1, \ldots, n$, 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

$$\hat{\mu}_j = \bar{y}_j, \qquad j = 1, 2,$$
$$\hat{\sigma} = s_{jk}/n, \qquad j, k = 1, 2,$$

where $\bar{y}_1$ and $\bar{y}_2$ are the sample means and $S = (s_{jk})$ is the sum of squares and cross-products 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 $\mu_2 + \beta_{21\cdot1}(y_{i1} - \mu_1)$ and variance $\sigma_{22\cdot1}$, where

$$\beta_{21\cdot1} = \sigma_{12}/\sigma_{11} \text{ and } \sigma_{22\cdot1} = \sigma_{22} - \sigma_{12}^2/\sigma_{11}$$

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

$$\hat{\beta}_{21\cdot1} = \hat{\sigma}_{12}/\hat{\sigma}_{11} = s_{12}/s_{11},$$

the least squares estimate of the slope, and

$$\hat{\sigma}_{22\cdot1} = \hat{\sigma}_{22} - \hat{\sigma}_{12}^2/\hat{\sigma}_{11} = \text{RSS}/n,$$

where $\text{RSS} = \sum_{i=1}^{n}[y_{i2} - \bar{y}_2 - \hat{\beta}_{21\cdot1}(y_{i1} - \bar{y}_1)]^2$ is the residual sum of squares from the regression based on the n sampled observations.

The ML estimates of $\beta_{21\cdot1}$ and $\sigma_{22\cdot1}$ 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 observations $(y_i, x_{i1}, \ldots, x_{ip})$, $i = 1, \ldots, n$ on an outcome variable Y and p predictor variables $X = (X_1, \ldots, X_p)$. We assume that, given $x_i = (x_{i1}, \ldots, x_{ip})$, the values of y_i are independent normal random variables with mean $\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}$ and variance σ^2. The loglikelihood of $\theta = (\beta_0, \beta_1, \ldots, \beta_p, \sigma^2)$ given observed data (y_i, x_i), $i = 1, \ldots, n$ is

$$\ell(\theta|Y) = -(n/2)\ln\sigma^2 - \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_p x_{ip})^2/(2\sigma^2).$$

Maximizing this expression with respect to θ, the ML estimates of $(\beta_0, \ldots, \beta_p)$ are found to be the least squares estimates of the intercept and the regression coefficients. Specifically, let

$$X = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix} \text{ and } Y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix},$$

the $n \times (p+1)$ matrix of predictors including the constant term and the vector of outcomes. Then

$$\hat{\beta} = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}Y. \tag{6.5}$$

The ML estimate of σ^2 is

$$\hat{\sigma}^2 = (Y - X\hat{\beta})^{\mathrm{T}}(Y - X\hat{\beta})/n \equiv \mathrm{RSS}/n, \tag{6.6}$$

where RSS is the residual sum of squares from the least squares regression, the generalization of RSS in Example 6.9. Since 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 in estimating the $p+1$ location parameters.

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. Thus, $(y_i - \mu_i)\sqrt{w_i}$ are iid $N(0, \sigma^2)$, and the loglikelihood is

$$\ell(\theta|Y) = -(n/2)\ln\sigma^2 - \sum_{i=1}^{n} w_i(y_i - \mu_i)^2/(2\sigma^2).$$

Maximizing this function yields ML estimates given by the weighted least squares estimates

$$\hat{\beta} = (X^{\mathrm{T}}WX)^{-1}(X^{\mathrm{T}}WY), \tag{6.7}$$

and

$$\hat{\sigma}^2 = (Y - X\hat{\beta})^{\mathrm{T}}W(Y - X\hat{\beta})/n, \tag{6.8}$$

where $W = \mathrm{Diag}(w_1, \ldots, w_n)$.

EXAMPLE 6.11. *Generalized Linear Models.* Suppose that the data again consist of n observations $(y_i, x_{i1}, \ldots, x_{ip})$, $i = 1, \ldots, n$ on an outcome variable Y and p predictor variables $X_1, \ldots, X_p$. A more general class of models is obtained by assuming that, given $x_i = (x_{i1}, \ldots, x_{ip})$, the values of y_i are independent with a distribution from the regular exponential family:

$$f(y_i|\, x_i, \beta, \phi) = \exp\{[y_i\delta(x_i, \beta) - b(\delta(x_i, \beta))]/\phi + c(y_i, \phi)\}, \tag{6.9}$$

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:

$$E(y_i|\, x_i, \beta, \phi) = g^{-1}(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}), \tag{6.10}$$

or

$$g(\mu_i) = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}, \tag{6.11}$$

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

$$g_c(\mu_i) = \delta(x_i, \beta) = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}, \tag{6.12}$$

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:

Normal linear regression: $g_c = \text{identity}$, $b(\delta) = \delta^2/2$, $\phi = \sigma^2$;
Poisson regression: $g_c = \log$, $b(\delta) = \exp(\delta)$, $\phi = 1$;
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 log-likelihood of $\theta = (\beta, \phi)$ based on Eq. (6.9) is

$$\ell(\theta|Y) = \sum_{i=1}^{n} \{y_i \delta(x_i, \beta) - b[\delta(x_i, \beta]\}/\phi + c(y_i, \phi),$$

which for non-normal cases does not generally have an explicit maximum. Numerical maximization can be achieved using the Fisher scoring algorithm (McCullagh and Nelder, 1989, Section 2.5; Firth, 1991, Section 3.4).

6.1.2. Rudiments of Bayes Estimation

The likelihood function also plays a central role in Bayesian inference. In the Bayesian approach, all unknowns, including the parameters θ, are treated as random variables, and uncertainty about them is quantified using probability distributions. In particular, 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:

$$p(\theta|Y) = \frac{p(\theta)L(Y|\theta)}{p(Y)}, \tag{6.13}$$

where $p(Y) = \int p(\theta)L(Y|\theta)d\theta$ 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 define $\hat{\theta}$ to be the mode of the posterior distribution $p(\theta|Y)$. This corresponds to the ML estimate when the prior distribution is uniform:

$$p(\theta) = \text{constant for all possible } \theta.$$

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 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 ML and Bayes inference. References to these results include Huber (1967), DeGroot (1970), Rao (1972), Cox and Hinkley (1974), White (1982), and Gelman et al. (1995).

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 can be applied.

Approximation 6.1.

$$(\theta - \hat{\theta}) \sim N(0, C), \tag{6.14}$$

where C is 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 Eq. (6.14) is then that, conditional on $f(\cdot|\cdot)$ and 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,

$$\ell(\theta|Y) = \ell(\hat{\theta}|Y) + (\theta - \hat{\theta})^{\mathrm{T}} D_{\ell}(\hat{\theta}|Y) - \tfrac{1}{2}(\theta - \hat{\theta})^{\mathrm{T}} I(\hat{\theta}|Y)(\theta - \hat{\theta}) + r(\theta|Y),$$

[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, $A^{\mathrm{T}}A = AA^{\mathrm{T}} = C$) and I_d is the $(d \times d)$ identity matrix.

where $D_\ell(\theta|Y)$ is the score function, and $I(\theta|Y)$ is the observed information:

$$I(\theta|Y) = -\frac{\partial^2 \ell(\theta|Y)}{\partial\theta\partial\theta}.$$

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 flat in the range of θ supported by the data, the posterior distribution of θ has density

$$f(\theta|Y) \propto \exp\Big[-\tfrac{1}{2}(\theta - \hat{\theta})^{\mathrm{T}} I(\hat{\theta}|Y)(\theta - \hat{\theta})\Big],$$

which is the normal distribution of Approximation 6.1 with covariance matrix

$$C = I^{-1}(\hat{\theta}|Y),$$

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. (1995) 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. (1995, Section 5.4).
- **S5.** Prior distributions that exclude the value of θ at convergence, or points of convergence on the edge of the parameter space. 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 asymptotic covariance matrix of $\theta - \hat{\theta}$, as in Approximation 6.1. Then the large-sample covariance matrix of $g(\theta) - g(\hat{\theta})$ is

$$D_g(\hat{\theta})CD_g(\hat{\theta})^{\mathrm{T}},$$

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.

$$g(\theta) - g(\hat{\theta}) \sim N[0, D_g(\hat{\theta})CD_g(\hat{\theta})^{\mathrm{T}}]. \tag{6.15}$$

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 6.8.

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:

$$0 = D_\ell(\hat{\theta}|Y) = D_\ell(\theta|Y) - I(\theta|Y)(\hat{\theta} - \theta) + r(\hat{\theta}|Y).$$

If the remainder term $r(\hat{\theta}|Y)$ is negligible, we have

$$D_\ell(\theta|Y) \approx I(\theta|Y)(\hat{\theta} - \theta).$$

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 zero and covariance matrix

$$J(\theta) = E(I(\theta|\, y)|\theta) = \int I(\theta|\, y)\, f(\, y|\theta)dy,$$

which is called the expected information matrix. A version of the law of large numbers implies that

$$J(\theta) \cong J(\hat{\theta}) \cong I(\hat{\theta}|Y).$$

Combining these facts leads to Approximation 6.1, with covariance matrix

$$C = J^{-1}(\hat{\theta}),$$

the inverse of the expected information evaluated at $\theta = \hat{\theta}$, or

$$C = I^{-1}(\hat{\theta}),$$

the inverse of the observed information evaluated at $\theta = \hat{\theta}$. Ancillarity arguments (Efron and Hinkley, 1978) suggest that the observed information provides a better estimate of precision than the expected information. 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 $f(Y|\theta_0)$ for some true value θ_0 of the parameter. If the model is misspecified, and Y is in fact sampled from the true density $f^*(Y)$, the posterior mode or the ML estimate converges to the value θ^* of θ that maximizes the Kullback–Liebler Information $E\{\log[f(Y|\theta)/f^*(Y)]\}$ of the model distribution $f(Y|\theta)$ with respect to the true distribution $f^*(Y)$. The frequentist version of Approximation 6.1 can then be replaced by:

Approximation 6.3.

$$(\hat{\theta}|\, f^*) \sim N(\theta^*, C^*), \tag{6.16}$$

where θ^* is defined above and

$$C^* = J^{-1}(\theta)K(\theta)J^{-1}(\theta),$$

where $K(\theta) = E(D_\ell(\theta)D_\ell(\theta)^{\mathrm{T}})$. 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^*:

$$\hat{C}^* = I^{-1}(\hat{\theta})\hat{K}(\hat{\theta})I^{-1}(\hat{\theta}), \tag{6.17}$$

where

$$\hat{K}(\hat{\theta}) = D_\ell(\hat{\theta})D_\ell(\hat{\theta})^{\mathrm{T}}.$$

This estimator is less precise than the observed or expected information but provides some protection against model misspecification. See Gelman et al. (1995) for more details.

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 Eq. (6.17). 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 $I^{-1}(\hat{\theta}|Y)$, $J^{-1}(\hat{\theta})$, the sandwich estimator, or some other approximation) can be used to provide interval estimates for θ. For example, 95% intervals for scalar θ are given by

$$\hat{\theta} \pm 1.96C^{1/2}, \tag{6.18}$$

where 1.96 can often be replaced by 2 in practice. For vector θ, 95% ellipsoids are given by the inequality

$$(\theta - \hat{\theta})^{\mathrm{T}} C^{-1} (\theta - \hat{\theta}) \leq \chi^2_{0.95,d}, \tag{6.19}$$

where $\chi^2_{0.95,d}$ 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

$$(\theta_{(1)} - \hat{\theta}_{(1)})^{\mathrm{T}} C^{-1}_{(11)} (\theta_{(1)} - \hat{\theta}_{(1)}) \leq \chi^2_{0.95,q}, \tag{6.20}$$

where $\hat{\theta}_{(1)}$ is the ML estimate of $\theta_{(1)}$ and $C_{(11)}$ is the submatrix of C corresponding to $\theta_{(1)}$.

Under $f(\cdot|\cdot)$ and assuming large enough samples to make Approximation 6.1 appropriate, inference based on Eq. (6.14) is not only appropriate, but 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 6.12. *Exponential Sample (Example 6.2 continued).* Differentiating (6.2) twice with respect to θ gives

$$I(\theta|Y) = -n/\theta^2 + 2\sum y_i/\theta^3.$$

Taking expectations over Y under the exponential specification for y_i gives

$$\begin{aligned} J(\theta|Y) &= -n/\theta^2 + 2E(\textstyle\sum y_i|\theta)/\theta^3 \\ &= -n/\theta^2 + 2n\theta/\theta^3 = n/\theta^2. \end{aligned}$$

Substituting the ML estimate $\hat{\theta} = \bar{y}$ for θ gives

$$I(\hat{\theta}|Y) = J(\hat{\theta}) = n/\bar{y}^2,$$

whence the large sample variance of $\theta - \hat{\theta}$ is $\bar{y}^2/n$.

EXAMPLE 6.13. *Univariate Normal Sample (Example 6.1 continued).* For the univariate normal model of Example 6.1, the asymptotic approximation (6.14) is more appropriately applied to $(\mu, \log\sigma^2)$ than to (μ, σ^2). Differentiating Eqs. (6.1) twice with respect to μ and $\log\sigma^2$, and substituting ML estimates of the parameters yields

$$I(\hat{\mu}, \log\hat{\sigma}^2|\ Y) = J(\hat{\mu}, \log\hat{\sigma}^2) = \begin{bmatrix} n/\hat{\sigma}^2 & 0 \\ 0 & n/2 \end{bmatrix}.$$

Inverting this matrix yields the large sample second moments $\mathrm{Var}(\mu - \hat{\mu}) = \hat{\sigma}^2/n$, $\mathrm{Cov}(\mu - \hat{\mu}, \log\sigma^2 - \log\hat{\sigma}^2) = 0$, $\mathrm{Var}(\log\sigma^2 - \log\hat{\sigma}^2) = 2/n$, 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 Eq. (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,

$$W(\theta_0, \hat{\theta}) = (\theta_0 - \hat{\theta})^{\mathrm{T}} C^{-1} (\theta_0 - \hat{\theta}),$$

which is the left side of Eq. (6.19) 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 θ_0: $\Pr[\theta \in \{\theta|\, f(\theta|Y) < f(\theta_0|Y)\}|Y]$;

see Box and Tiao (1973) and Rubin (1987a, 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

$$p_L = \Pr[\chi_d^2 > \mathrm{LR}(\theta_0, \hat{\theta})],$$

where

$$\mathrm{LR}(\theta_0, \hat{\theta}) = 2\ln[L(\hat{\theta}|Y)/L(\theta_0|Y)] = 2[l(\hat{\theta}|Y) - l(\theta_0|Y)].$$

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:

$$p_C(\theta_{(1)0}) = \Pr[\chi_q^2 > (\theta_{(1)0} - \hat{\theta}_{(1)})^{\mathrm{T}} C_{(11)}^{-1} (\theta_{(1)0} - \hat{\theta}_{(1)})],$$

where $C_{(11)}$ is the variance–covariance matrix of $\theta_{(1)}$ as in Eq. (6.20), and

$$p_L(\theta_{(1)0} = \Pr[\chi_q^2 > \mathrm{LR}(\hat{\theta}, \tilde{\theta})],$$

where

$$\mathrm{LR}(\hat{\theta}, \tilde{\theta}) = 2[\ell(\hat{\theta}|Y) - \ell(\tilde{\theta}|Y)],$$

and $\tilde{\theta}$ is the value of θ that maximizes $\ell(\theta|Y)$ 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 6.14. *Univariate Normal Sample (Example 6.1 continued).* Suppose $\theta = (\mu, \sigma^2)$, $\theta_{(1)} = \mu$, $\theta_{(2)} = \sigma^2$. To test H_0: $\mu = \mu_0$, the likelihood ratio test statistic is

$$\begin{aligned} \mathrm{LR} &= 2(-n/2\ln[(n-1)s^2/n] - n/2 + n/2\ln s_0^2 + n/2) \\ &= n\ln[ns_0^2/(n-1)s^2], \end{aligned}$$

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 $\mathrm{LR} = n\ln(1 + t^2/n)$, where $t^2 = n^2(\bar{y} - \mu_0)^2/[(n-1)s^2]$ is, from Example 6.13, the test statistic for H_0 based on the asymptotic variance of $(\mu - \mu_0)$. Asymptotically, $\mathrm{LR} = t^2$ and is chi-squared distributed with $q = 1$ degrees of freedom under H_0.

An exact test is obtained in this case by comparing 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.

6.1.4. Bayes Inference Based on the Full Posterior Distribution

The theory of the previous section assumes large samples, and can yield unsatisfactory inferences when the sample size is small. One approach to this limitation is to adopt a Bayesian perspective and base inferences on the exact posterior distribution for a particular choice of prior. To measure uncertainty about the point estimate, the posterior standard deviation replaces the frequentist standard error, 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) replace the frequentist confidence interval. 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 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 can result from formally incorporating this information, using the Bayesian machinery. In cases where prior knowledge is more limited, or where objective inferences are sought that are not strongly driven by prior information, the Bayesian approach with dispersed priors often yields better frequentist inferences than the large-sample approximations of Section 6.1.3. In particular, in the next two examples we show that in problems involving the normal distribution, the Bayesian approach with noninformative priors can recover the t corrections obtained in small-sample frequentist inference. For more complex problems, including those involving missing data, the Bayesian approach is attractive since it provides answers in situations where no exact frequentist solutions are available. In fact, the Bayesian approach is prescriptive in describing what we should do, no matter how complex the problem. The Bayesian answer, once derived, can be evaluated from a frequentist 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 6.15. *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 conjugate prior distribution is chosen:

$$p(\mu, \sigma^2) = p(\sigma^2)p(\mu|\sigma^2),$$

where

$$\begin{aligned} \sigma^2 &\sim \text{Inv-}\chi^2(\nu_0, \sigma_0^2), \\ (\mu|\sigma^2) &\sim N(\mu_0, \sigma^2/\kappa_0), \end{aligned} \tag{6.21}$$

for known v_0, σ_0^2, μ, and κ_0. Here, the prior distribution for σ^2 is a scaled inverse chi-squared distribution with degrees of freedom v_0 and scale σ_0^2 (Gelman et al., 1995, Appendix A), and the prior distribution for μ given σ^2 is normal. The joint prior density then 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., 1995, Section 3.3) that the posterior distribution of θ is then

$$p(\mu, \sigma^2|Y) = p(\sigma^2|Y)p(\mu|\sigma^2, Y),$$

where the posterior distribution for σ^2 given Y has the scaled inverse chi-squared distribution

$$\begin{aligned} \sigma^2|Y &\sim \text{Inv-}\chi^2(v_n, \sigma_n^2), \\ v_n &= v_0 + n, \\ v_n\sigma_n^2 &= v_0\sigma_0^2 + (n-1)s^2 + \frac{\kappa_0 n}{\kappa_0 + n}(\bar{y} - \mu_0)^2, \end{aligned} \tag{6.22}$$

and the posterior distribution for μ given σ^2 and Y has the normal distribution

$$\begin{aligned} (\mu|\sigma^2, Y) &\sim N(\mu_n, \sigma^2/\kappa_n), \\ \kappa_n &= \kappa_0 + n, \\ \mu_n &= \frac{\kappa_0}{\kappa_0 + n}\mu_0 + \frac{n}{\kappa_0 + n}\bar{y}. \end{aligned} \tag{6.23}$$

Integrating this conditional posterior distribution over the posterior distribution of σ^2 yields the marginal posterior distribution of μ given Y:

$$(\mu|Y) \sim t(\mu_n, \sigma_n^2/\kappa_n, v_n), \tag{6.24}$$

the Student's t distribution with mean 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 diffuse prior is the Jeffrey's prior:

$$p(\mu, \sigma^2) \propto 1/\sigma^2, \tag{6.25}$$

(for example, Box, and Tiao, 1973) which is a degenerate special case of Eq. (6.21) with $\kappa_0 = 0$, $\nu_0 = -1$ and $\sigma_0^2 = 0$. Substituting these values of the parameters in Eqs. (6.22)–(6.24) yields $\mu_n = \bar{y}$, $\kappa_n = n$, $\nu_n = n - 1$, $\sigma_n^2 = s^2$ and

$$\sigma^2|Y \sim \text{Inv-}\chi^2(n-1, s^2), \tag{6.26}$$

$$(\mu|\sigma^2, Y) \sim N(\bar{y}, \sigma^2/n), \tag{6.27}$$

$$(\mu|Y) \sim t(\bar{y}, s^2/n, n-1). \tag{6.28}$$

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 mean 0, scale 1 and degrees of freedom $n-1$. This interval is identical to the standard $100(1-\alpha)\%$ confidence interval for the mean. Thus the Bayesian analysis with prior distribution (6.25) recovers the degrees of freedom correction and the t reference distribution of the classical normal-sample analysis.

EXAMPLE 6.16. *Bayes Inference for Unweighted and Weighted Multiple Linear Regression (Example 6.10 continued).* The noninformative Jeffreys' prior distribution for the normal linear regression model of Example 6.10 is

$$p(\beta_0, \beta_1, \ldots, \beta_p, \sigma^2) \propto 1/\sigma^2, \tag{6.29}$$

which places a uniform prior on the location parameters. The corresponding posterior distribution has the form:

$$\sigma^2|Y \sim \text{Inv-}\chi^2(n-p-1, s^2), \tag{6.30}$$

$$(\beta|\sigma^2, Y) \sim N_{p+1}(\hat{\beta}, (X^{\mathrm{T}}X)^{-1}\sigma^2), \tag{6.31}$$

$$(\beta|Y) \sim t_{p+1}(\hat{\beta}, (X^{\mathrm{T}}X)^{-1}s^2, n-p-1), \tag{6.32}$$

the multivariate t distribution with mean $\hat{\beta}$, scale $(X^{\mathrm{T}}X)^{-1}s^2$ and degrees of freedom $n-p-1$. Here $s^2 = n\hat{\sigma}^2/(n-p-1)$, the residual mean square corrected for degrees of freedom. Equation (6.32) gives $100(1-\alpha)\%$ 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 Eqs. (6.30)–(6.32), the weighted estimates (6.7) and (6.8) replace the unweighted estimates (6.5) and (6.6), and $X^{\mathrm{T}}WX$ replaces $X^{\mathrm{T}}X$.

The following two examples are important when we consider incomplete data.

EXAMPLE 6.17. *Bayes Inference for a Multinomial Sample (Example 6.3 continued).* Suppose Y forms a set of counts with a multinomial distribution as in Example 6.3. The conjugate prior distribution is a multivariate generalization of the

beta distribution known as the Dirichlet:

$$p(\pi_1, \ldots, \pi_C) \propto \prod_{c=1}^{C} \pi_c^{\alpha_c - 1}, \qquad \pi_c > 0, \qquad \sum_{c=1}^{C} \pi_c = 1. \tag{6.33}$$

Combining this prior distribution with the likelihood (6.3) yields the posterior distribution as Dirichlet with parameters $\{n_c + \alpha_c\}$:

$$p(\pi_1, \ldots, \pi_C | Y) \propto \prod_{c=1}^{C} \pi_c^{n_c + \alpha_c - 1}, \qquad \pi_c > 0, \qquad \sum_{c=1}^{C} \pi_c = 1. \tag{6.34}$$

By properties of the Dirichlet distribution (Gelman et al., 1995, Appendix A), the posterior mean of π_c is $(n_c + \alpha_c)/(n_+ + \alpha_+)$, where $n_+ = \sum_{c=1}^{C} n_c$, $\alpha_+ = \sum_{c=1}^{C} \alpha_c$. This equals the ML estimate when $\alpha_c = 0$ for all c. Alternative diffuse 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 6.18. *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:

$$p(\mu, \Sigma) \propto |\Sigma|^{-(K+1)/2}, \tag{6.35}$$

which reduces to Eq. (6.25) when $K = 1$. The corresponding posterior distribution can be written as:

$$\begin{aligned} (\Sigma | Y) &\sim \text{Inv-Wishart}(S, n - 1) \\ (\mu | \Sigma, Y) &\sim N_K(\bar{y}, \Sigma/n), \end{aligned} \tag{6.36}$$

where Inv-Wishart $(S, n - 1)$ denotes the inverse Wishart distribution with $n - 1$ degrees of freedom and scale matrix S (see Gelman et al., 1995, Appendix A). Equation (6.36) implies that the marginal posterior distribution of μ given Y is multivariate t with mean $\bar{y}$, scale matrix S/n and degrees of freedom $n - 1$.

6.1.5. Simulating Draws from Posterior Distributions

The application of Bayes methods to more complex problems was until recently constrained by the numerical difficulties involved in computing the posterior distribution of particular parameters, particularly when θ is high dimensional. For example if $\theta = (\theta_1, \theta_2)$, then the posterior distribution of θ_1 is

$$p(\theta_1 | Y) = \int p(\theta) L(\theta | Y) d\theta_2 \Big/ \int p(\theta) L(\theta | Y) d\theta,$$

which involves high-dimensional integration when θ_2 has many components. These problems have been greatly reduced by stochastic simulation methods that aim at taking draws from the posterior distribution of θ, rather than at 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 ($\theta_1^{(d)}$, $d = 1, \ldots, D$). If the posterior distribution is far from normal, 95% 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 a draw from the posterior distribution of θ. Then $g(\theta^{(d)})$ is a draw from the posterior distribution of $g(\theta)$.

This property will prove useful in applications of Bayes simulation methods to incomplete data problems, as discussed in Section 7.3 and Chapter 10.

EXAMPLE 6.19. *Bayes Inference for Multiple Linear Regression (Example 6.16 continued).* Draws $\beta^{(d)}$, $\sigma^{(d)}$ from the posterior distribution of β, σ for the normal regression model data of Example 6.16 with prior (6.25) are readily obtained from Eqs. (6.30) and (6.31), as follows:

1. Draw χ^2_{n-p-1} from a chi-squared distribution with $n-p-1$ degrees of freedom, and set

$$\sigma^{(d)2} = (n-p-1)s^2/\chi^2_{n-p-1}. \tag{6.37}$$

2. Draw p standard normal deviates, $z = (z_1, \ldots, z_0)^{\mathrm{T}}$, $z_i \sim N(0,1)$, $i = 1, \ldots, p$, and set

$$\beta^{(d)} = \hat{\beta} + A^{\mathrm{T}} z \sigma^{(d)}, \tag{6.38}$$

where A is an upper triangular ($p \times p$) Cholesky factor of $(X^{\mathrm{T}}X)^{-1}$, such that $A^{\mathrm{T}}A = (X^{\mathrm{T}}X)^{-1}$. Draws are not needed for inference about the regression coefficients themselves, which have a t distribution. However, they are useful for simulating 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 (6.37) and (6.38) play an important role for simulations of posterior distributions for normal missing-data problems, discussed later.

EXAMPLE 6.20. *Bayes Inference for a Multinomial Sample (Example 6.17 continued).* Draws $\{\pi_c^{(d)}\}$ from the Dirichlet posterior distribution (6.34) of $\{\pi_c\}$

under the multinomial model of Example 6.17 can be obtained by generating independent chi-squared deviates $\{\chi^2_{2(n_c+\alpha_c)}\}$ for $c = 1, \ldots, C$, and setting

$$\pi_c^{(d)} = \chi^2_{2(n_c+\alpha_c)} / \sum_{j=1}^{C} \chi^2_{2(n_j+\alpha_j)}. \tag{6.39}$$

Often chi-squared random variables, and the associated t distribution, are defined with integer degrees of freedom. This is an unnecessary restriction, but a notationally more general form of Eq. (6.39) replaces the $\chi^2_{2(n_j+\alpha_j)}$ random variables by standard (scale parameter = 1) gamma random variables, g, with parameters $n_j + \alpha_j$, with density proportional to $g^{n_j+\alpha_j-1}\exp(-g)$. See, for example, Gelman et al. (1995, 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 6.21. *Bayes Inference for a Multivariate Normal Sample (Example 6.18 continued).* The posterior distribution of $\theta = (\mu, \Sigma)$ for a multivariate normal sample with prior distribution (6.35) is given by Eq. (6.36). A draw from this distribution is obtained by first drawing $\Sigma^{(d)}$ from Inv-Wishart $(S, n-1)$, and then drawing $\mu^{(d)} = \bar{y} + A^{(d)}z$, where $z = (z_1, \ldots, z_K)^{\mathrm{T}}$ is a vector of independent $N(0, 1)$ draws, and $A^{(d)}$ is an upper triangular Cholesky factor such that $A^{(d)\mathrm{T}}A^{(d)} = \Sigma^{(d)}/n$.

The inv-Wishart draw $\Sigma^{(d)}$ is obtained by forming an upper triangular matrix B with elements

$$b_{ij} \sim \sqrt{\chi^2_{n-j}}, \qquad b_{jk} \sim N(0, 1), \qquad j < k, \tag{6.40}$$

and drawing

$$\Sigma^{(d)} = (B^{\mathrm{T}})^{-1}A, \tag{6.41}$$

where A is the Cholesky factor of S^{-1} (that is, $A^{\mathrm{T}}A = S^{-1}$). 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 formal sense there is no 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, however, since the observed data do not generally constitute an iid (independent, identically distributed) 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.

As before, let Y denote the data that would occur in the absence of missing values. We write $Y = (Y_{\text{obs}}, Y_{\text{mis}})$, where Y_{obs} denotes the observed values and Y_{mis} denotes the missing values. Let $f(Y|\theta) \equiv f(Y_{\text{obs}}, Y_{\text{mis}}|\theta)$ denote the probability or density of the joint distribution of Y_{obs} and Y_{mis}. The marginal probability density of Y_{obs} is obtained by integrating out the missing data Y_{mis}:

$$f(Y_{\text{obs}}|\theta) = \int f(Y_{\text{obs}}, Y_{\text{mis}}|\theta) dY_{\text{mis}}.$$

We define the likelihood of θ based on data Y_{obs} *ignoring the missing-data mechanism* to be any function of θ proportional to $f(Y_{\text{obs}}|\theta)$:

$$\text{L}_{\text{ign}}(\theta|Y_{\text{obs}}) \propto f(Y_{\text{obs}}|\theta), \qquad \theta \in \Omega_\theta. \tag{6.42}$$

Inferences about θ can be based on this likelihood, $\text{L}_{\text{ign}}(\theta|Y_{\text{obs}})$, providing the mechanism leading to incomplete data can be ignored, in a sense discussed below. In particular, ignorable ML estimates are obtained by maximizing Eq. (6.42) with respect to θ. Ignorable Bayes inference for θ based on data Y_{obs} is obtained by incorporating a prior distribution $p(\theta)$ for θ and basing inference on the posterior distribution:

$$p(\theta|Y_{\text{obs}}) \propto p(\theta) \times \text{L}_{\text{ign}}(\theta|Y_{\text{obs}}). \tag{6.43}$$

More generally, we can include in the model the distribution of a variable indicating whether each component of Y is observed or missing. As in earlier chapters, we define for each component of Y a *missing-data indicator*, taking value 1 if the component is missing and 0 if it is observed. For example, if $Y = (y_{ij})$ an $(n \times K)$ matrix of n observations measured for K variables, then

$$M_{ij} = \begin{cases} 1, & y_{ij} \text{ missing,} \\ 0, & y_{ij} \text{ observed.} \end{cases}$$

The full model treats M as a random variable and specifies the joint distribution of M and Y. The density of this distribution can be specified as the product of the densities of the distribution of Y and the conditional distribution of M given Y, which is indexed by an unknown parameter ψ and is called the distribution of the missing-data mechanism. That is,

$$f(Y, M|\theta, \psi) = f(Y|\theta) f(M|Y, \psi), \qquad (\theta, \psi) \in \Omega_{\theta,\psi},$$

where $\Omega_{\theta,\psi}$ is the parameter space of (θ, ψ). In some situations the distribution of the missing-data mechanism is known, and ψ is superfluous. Other specifications of the joint distribution of Y and M are discussed in Chapter 15.

The actual observed data consist of the values of the variables (Y_{obs}, M). The distribution of the observed data is obtained by integrating Y_{mis} out of the joint density of $Y = (Y_{\text{obs}}, Y_{\text{mis}})$ and M. That is,

$$f(Y_{\text{obs}}, M|\theta, \psi) = \int f(Y_{\text{obs}}, Y_{\text{mis}}|\theta) f(M|Y_{\text{obs}}, Y_{\text{mis}}, \psi) dY_{\text{mis}}. \tag{6.44}$$

The full likelihood of θ and ψ is any function of θ and ψ proportional to Eq. (6.44):

$$\mathrm{L}_{\text{full}}(\theta, \psi|Y_{\text{obs}}, M) \propto f(Y_{\text{obs}}, M|\theta, \psi), \qquad (\theta, \psi) \in \Omega_{\theta,\psi}. \tag{6.45}$$

The question is now when inference for θ should be based on the full likelihood $\mathrm{L}_{\text{full}}(\theta, \psi|Y_{\text{obs}}, M)$ in Eq. (6.45), and when it can be based on the simpler likelihood that ignores the missing-data mechanism, $\mathrm{L}_{\text{ign}}(\theta|Y_{\text{obs}})$ in Eq. (6.42). Observe that if the distribution of the missing-data mechanism does not depend on the missing values Y_{mis}, that is, if

$$f(M|Y_{\text{obs}}, Y_{\text{mis}}, \psi) = f(M|Y_{\text{obs}}, \psi) \text{ for all } Y_{\text{mis}}, \tag{6.46}$$

then from Eq. (6.44),

$$\begin{aligned} f(Y_{\text{obs}}, M|\theta, \psi) &= f(M|Y_{\text{obs}}, \psi) \times \int f(Y_{\text{obs}}, Y_{\text{mis}}|\theta) dY_{\text{mis}} \\ &= f(M|Y_{\text{obs}}, \psi) f(Y_{\text{obs}}|\theta). \end{aligned} \tag{6.47}$$

As discussed in Section 1.3, the missing data are called missing at random (MAR) when Eq. (6.46) holds (Rubin, 1976a). In many important practical applications, the parameters θ and ψ are distinct, in the sense that the joint parameter space of (θ, ψ) is the product of the parameter space of θ and the parameter space of ψ, $\Omega_{\theta,\psi} = \Omega_\theta \times \Omega_\psi$. If the mechanism is MAR and θ and ψ are distinct, then likelihood-based inferences for θ from $\mathrm{L}_{\text{full}}(\theta, \psi|Y_{\text{obs}}, M)$ will be the same as likelihood-based inferences for θ from $\mathrm{L}_{\text{ign}}(\theta|Y_{\text{obs}})$, since the resulting likelihoods are proportional. This motivates the following definition:

Definition 6.4. The missing-data mechanism is ignorable for likelihood inference if:

- **(a)** MAR: the missing data are missing at random; and
- **(b)** Distinctness: the parameters θ and ψ are distinct, in the sense that the joint parameter space of (θ, ψ) is the product of the parameter space of θ and the parameter space of ψ.

MAR is typically regarded as the more important condition here, in the sense that if the data are MAR but distinctness does not hold, inference based on the ignorable likelihood is still valid from the frequency perspective, but not fully efficient. That is, there possibly is information about θ in the factor $f(M|Y_{\text{obs}}, \psi)$ in Eq. (6.47) that is being ignored, but doing so does not affect the sampling distribution of any statistic, which is conditioned on the fixed but unknown parameters (θ, ψ).

Bayes inference under the full model for Y and M is obtained by combining the full likelihood (6.45) with a prior distribution for θ and ψ:

$$p(\theta, \psi|Y_{\text{obs}}, M) \propto p(\theta, \psi) \times \mathrm{L}_{\text{full}}(\theta, \psi|Y_{\text{obs}}, M). \tag{6.48}$$

Suppose that the data are MAR, and

$$p(\theta, \psi) = p(\theta)p(\psi), \tag{6.49}$$

that is, θ and ψ are *a priori* independent. Then it follows that

$$\begin{aligned} p(\theta, \psi|Y_{\text{obs}}, M) &\propto [\, p(\theta)L(\theta|Y_{\text{obs}})][\, p(\psi)L(\psi|Y_{\text{obs}}, M)] \\ &\propto p(\theta|Y_{\text{obs}})p(\psi|Y_{\text{obs}}, M), \end{aligned}$$

that is, θ and ψ are *a posteriori* independent. Inference about θ can then be based on the posterior distribution $p(\theta|Y_{\text{obs}})$ ignoring the missing-data mechanism. This motivates another definition:

Definition 6.5. The missing-data mechanism is ignorable for Bayesian inference if:

- **(a)** MAR: the missing data are missing at random; and
- **(b)** The parameters θ and ψ are *a priori* independent, that is, the prior distribution has the form (6.49).

Definition 6.5 is stronger than Definition 6.4, since distinctness of the parameter spaces is required for θ and ψ to have independent priors.[2]

EXAMPLE 6.22. *Incomplete Exponential Sample.* Suppose we have an incomplete univariate exponential sample with $Y_{\text{obs}} = (y_1, \ldots, y_r)^{\mathrm{T}}$ observed and $Y_{\text{mis}} = (y_{r+1}, \ldots, y_n)^{\mathrm{T}}$ missing. As in Example 6.2,

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

[2]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".

The likelihood ignoring the missing-data mechanism is proportional to the distribution of Y_{obs} given θ, which is

$$f(Y_{\text{obs}}|\theta) = \theta^{-r} \exp\left(-\sum_{i=1}^{r} \frac{y_i}{\theta}\right). \tag{6.50}$$

Now $M = (M_1, \ldots, M_n)^{\text{T}}$, where $M_i = 0$, $i = 1, \ldots, r$ and $M_i = 1$, $i = r+1, \ldots, n$.

Suppose that each unit is observed with probability ψ independent of Y so that Eq. (6.46) holds. Then

$$f(M|Y, \psi) = \frac{n!}{r!(n-r)!} \psi^r (1-\psi)^{n-r},$$

and

$$f(Y_{\text{obs}}, M|\theta) = \frac{n!}{r!(n-r)!} \psi^r (1-\psi)^{n-r} \theta^{-r} \exp\left(-\sum_{i=1}^{r} \frac{y_i}{\theta}\right).$$

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 proportional to $f(Y_{\text{obs}}|\theta)$. In particular, the ML estimate of θ is simply $\sum_{i=1}^{r} y_i/r$, the mean of the responding 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 recorded. Then

$$f(M|Y, \psi) = \prod_{i=1}^{n} f(M_i|\, y_i, \psi),$$

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

$$\begin{aligned} f(Y_{\text{obs}}, M|\theta) &= \prod_{i=1}^{r} f(y_i, M_i|\theta) \prod_{i=r+1}^{n} f(M_i|\theta) \\ &= \prod_{i=1}^{r} f(y_i|\theta) \Pr(y_i < c \mid y_i) \prod_{i=r+1}^{n} \Pr(y_i \geq c \mid \theta) \\ &= \theta^{-r} \exp\left(-\sum_{i=1}^{r} \frac{y_i}{\theta}\right) \exp\left(-\frac{(n-r)c}{\theta}\right), \end{aligned} \tag{6.51}$$

since $\Pr(y_i < c \mid y_i) = 1$ for respondents and $\Pr(y_i \geq c \,|\theta) = \exp(-c/\theta)$ for non-respondents, using the properties of the exponential distribution. In this case the

missing-data mechanism is not ignorable, and the correct likelihood (6.51) differs from (6.50). Maximizing Eq. (6.51) 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 6.23. *Bivariate Normal Sample with One Variable Subject to Nonresponse.* Suppose we have a bivariate normal sample as in Example 6.9, but the values y_{i2} of the second variable are missing for $i = (r+1), \ldots, n$. We thus have a monotone pattern with two variables. The loglikelihood ignoring the missing-data mechanism is

$$\ell_{\text{ign}}(\mu, \Sigma | Y_{\text{obs}}) = \ln[\mathrm{L}_{\text{ign}}(\mu, \Sigma | Y_{\text{obs}}] = -\frac{1}{2} r \ln |\Sigma| - \frac{1}{2} \sum_{i=1}^{r} (y_i - \mu)\Sigma^{-1}(y_i - \mu)^{\mathrm{T}} - \frac{1}{2}(n-r) \ln \sigma_{11} - \frac{1}{2} \sum_{i=r+1}^{n} \frac{(y_{i1} - \mu_1)^2}{\sigma_{11}}. \tag{6.52}$$

This loglikelihood is appropriate for inferences provided the distribution of M (and in particular, the probability that y_{i2} is observed) 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 parameters ψ of the missing-data mechanism. Under these conditions, ML estimates of μ and Σ can be found by maximizing Eq. (6.52). For Bayes inference, if these conditions hold and the prior distribution has the form $p(\mu, \Sigma, \psi) = p(\mu, \Sigma)p(\psi)$, then the posterior distribution of μ and Σ is proportional to the product of $p(\mu, \Sigma)$ and $\mathrm{L}_{\text{ign}}(\mu, \Sigma | Y_{\text{obs}})$ from Eq. (6.52). 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 6.24. *One-Way ANOVA with Missing Values, when Missingness Depends on the Unobserved Group Means.* Consider a one-way random effects analysis of variance (ANOVA) with I groups, data $X = \{x_{ij}: i = 1, \ldots, I; j = 1, \ldots, n_i\}$, and model

$$(x_{ij} | \mu_i, \theta) \sim_{\text{ind}} N(\mu_i, \sigma^2), \tag{6.53}$$

$$(\mu_i | \theta) \sim_{\text{ind}} N(\mu, \tau^2), \tag{6.54}$$

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 (6.54), for $i = 1, \ldots, I$. Suppose that some data values x_{ij} are missing, and let X_{obs} denote the

observed values of X and X_{mis} the missing values. Suppose the missing-data mechanism depends on the unobserved random variables $\{\mu_i\}$:

$$\Pr(m_{ij} = 1 | X, \mu_i, \psi) \equiv \pi(\mu_i; \psi) = \exp(\psi_0 + \psi_1 \mu_i)/[1 + \exp(\psi_0 + \psi_1 \mu_i)]. \quad (6.55)$$

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:

$$Y = (X, \{\mu_i\}), \qquad Y_{\text{obs}} = X_{\text{obs}} \text{ and } Y_{\text{mis}} = (X_{\text{mis}}, \{\mu_i\}),$$

and the missing-data mechanism (6.55) is nonignorable because the missing data are not MAR. Likelihood inference based on Eqs. (6.53)–(6.55) must be based on the full likelihood (6.45). Suppose r_i values of X are observed in group i, $r = \sum_{i=1}^{I} r_i$, and let $\bar{x}_{\text{obs},i}$ denote the mean of these observed values. The full likelihood is then:

$$L_{\text{full}}(\theta, \psi | X_{\text{obs}}, M) = \sigma^{-r}\tau^{-1} \prod_{i=1}^{I} \int \pi(\mu_i; \psi)^{r_i}[1 - \pi(\mu_i; \psi)]^{n_i - r_i} \exp[-r_i(\bar{x}_{\text{obs},i} - \mu_i)^2/(2\sigma^2) - (\mu_i - \mu)^2/(2\tau^2)] d\mu_i.$$

A common alternative model to Eqs. (6.53)–(6.54) is the fixed-effects ANOVA model

$$x_{ij} | \mu_i, \sigma^2 \sim_{\text{ind}} N(\mu_i, \sigma^2), \quad (6.56)$$

where $\theta^* = (\{\mu_i\}, \sigma^2)$ are regarded as the fixed unknown parameters. The complete data then do not include the $\{\mu_i\}$ since they are fixed parameters, and are defined as:

$$Y = X, \qquad Y_{\text{obs}} = X_{\text{obs}} \text{ and } Y_{\text{mis}} = X_{\text{mis}}.$$

The full likelihood with missing-data mechanism (6.55) is then

$$L_{\text{full}}(\theta^*, \psi | X_{\text{obs}}, M) = \sigma^{-r} \prod_{i=1}^{I} \exp[-r_i(\bar{x}_{\text{obs},i} - \mu_i)^2/(2\sigma^2)] \pi(\mu_i; \psi)^{r_i}[1 - \pi(\mu_i; \psi)]^{n_i - r_i}, \quad (6.57)$$

and the likelihood ignoring the missing-data mechanism is then

$$L_{\text{ign}}(\theta^* | X_{\text{obs}}) = \sigma^{-r} \prod_{i=1}^{I} \exp[-r_i(\bar{x}_{\text{obs},i} - \mu_i)^2/(2\sigma^2)]. \quad (6.58)$$

The mechanism (6.55) is now MAR, since (unlike in the random-effects model) it does not depend on missing data. However, the distinctness condition in Definition 6.4 is violated, since the models (6.55) and (6.56) both involve the parameters $\{\mu_i\}$.

Hence the missing-data mechanism is nonignorable for likelihood inference about θ*. Likelihood or Bayes inference ignoring the mechanism, based on Eq. (6.58) is strictly speaking incorrect, since a relevant part of the full likelihood (6.57) is being ignored. From the frequentist perspective, estimators based on Eq. (6.58) are valid despite the violation of the distinctness condition, but they are not fully efficient.

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

$$L_{\text{mispar}}(\theta, Y_{\text{mis}}|Y_{\text{obs}}) = L(\theta|Y_{\text{obs}}, Y_{\text{mis}}) = f(Y_{\text{obs}}, Y_{\text{mis}}|\theta) \tag{6.59}$$

be regarded as a function of (θ, Y_{mis}) for fixed Y_{obs}, and estimate θ by maximizing $L_{\text{mispar}}(\theta, Y_{\text{mis}}|Y_{\text{obs}})$ over both θ and Y_{mis}. When the missing data are not MAR, or θ is not distinct from ψ, θ would be estimated by maximizing

$$\begin{aligned} L_{\text{mispar}}(\theta, \psi, Y_{\text{mis}}|Y_{\text{obs}}, M) &= L(\theta, \psi|Y_{\text{obs}}, Y_{\text{mis}}, M) \\ &= f(Y_{\text{obs}}, Y_{\text{mis}}|\theta) f(M|Y_{\text{obs}}, Y_{\text{mis}}, \psi) \end{aligned} \tag{6.60}$$

over $(\theta, \psi, Y_{\text{mis}})$. Although this approach can be useful in particular problems, it is not a generally valid approach to the analysis of incomplete data. In particular, Little and Rubin (1983b) show that it does not generally share the optimal properties of ML estimation, except under the trivial asymptotics in which the proportion of missing data goes to zero as the sample size increases.

6.3.2. Background

The classic example of this approach is the treatment of missing plots in analysis of variance where missing outcomes Y_{mis} 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_{mispar} in Eq. (6.59) over (θ, Y_{mis}). Box, Draper, and Hunter (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. More

recently, Lee and Nelder (1996) advocate this approach for the analysis of generalized linear mixed models, as discussed in Example 6.28 below.

Formally, $L_{full}(\theta, \psi|Y_{obs}, M)$ defined in Eq. (6.45), or $L_{ign}(\theta|Y_{obs})$ defined in Eq. (6.42) if the missing-data mechanism is ignorable, define the correct likelihood for inferences about θ based on the observed data; the functions L_{mispar} in Eqs. (6.59) or (6.60) are not likelihoods since their arguments include random variables Y_{mis}, which have a distribution under the model and hence should not be treated as fixed parameters. Thus, maximization of L_{mispar} with respect to θ and Y_{mis} is *not* an ML procedure.

A serious problem with treating both θ and Y_{mis} as parameters is that the number of parameters increases with the number of observations. The maximization of L_{mispar} only has the optimal properties of ML when the fraction of missing values tends to zero as the sample size increases. In contrast, the parameter θ does not depend on the amount of data, and hence, loosely speaking, standard likelihood asymptotics based on the maximization of $L_{full}(\theta, \psi|Y_{obs}, M)$ or $L_{ign}(\theta|Y_{obs})$ apply, provided the information increases with the sample size. This deficiency when treating Y_{mis} as a parameter is illustrated in the following examples.

6.3.3. Examples

EXAMPLE 6.25. *Univariate Normal Sample with Missing Data.* Suppose that $Y = (Y_{obs}, Y_{mis})$ consists of n realizations from a normal distribution with mean μ and variance σ^2, where Y_{obs} consists of r observed values and Y_{mis} represents $(n - r)$ missing values, which are assumed MAR; θ is (μ, σ^2), which we assume is distinct from the parameters of the missing-data mechanism. Since

$$f(Y|\theta) = \prod_{i=1}^{n} f(y_i|\theta) = \prod_{i=1}^{r} f(y_i|\theta) \times \prod_{i=r+1}^{n} f(y_i|\theta) \tag{6.61}$$

it follows that $f(Y_{obs}|\theta) = \prod_{i=1}^{r} f(y_i|\theta)$ and $f(Y_{mis}|\theta) = \prod_{i=r+1}^{n} f(y_i|\theta)$. Thus $L_{ign}(\theta|Y_{obs})$ is identical in form to the likelihood for a sample of size r from a normal distribution. By Example 6.7, maximizing $L_{ign}(\theta|Y_{obs})$ over θ thus leads to ML estimates

$$\hat{\mu} = \sum_{i=1}^{r} \frac{y_i}{r} \text{ and } \hat{\sigma}^2 = \sum_{i=1}^{r} \frac{(y_i - \hat{\mu})^2}{r}. \tag{6.62}$$

In contrast,

$$L_{mispar}(\theta, Y_{mis}|Y_{obs}) = f(Y_{obs}|\theta) f(Y_{mis}|\theta), \tag{6.63}$$

which is to be maximized over θ and Y_{mis}. The maximization over Y_{mis} in the second factor in Eq. (6.61) gives maximizing values

$$\tilde{y}_i = \tilde{\mu} \text{ for } i = r+1, \ldots, n, \tag{6.64}$$

where $\tilde{\mu}$ is the maximizing value of μ. From Example 6.7, the maximizing values for μ and σ^2 are

$$\hat{\mu}\left[\sum_{i=1}^{r} y_i + \sum_{i=r+1}^{n} \tilde{y}_i\right] \Big/ n \text{ and } \hat{\sigma}^2 = \left[\sum_{i=1}^{r} (y_i - \tilde{\mu})^2 + \sum_{i=r+1}^{n} (\tilde{y}_i - \tilde{\mu})^2\right] \Big/ n. \tag{6.65}$$

Substituting Eq. (6.64) into Eq. (6.65) and comparing with Eq. (6.62) gives

$$\tilde{\mu} = \hat{\mu} \text{ and } \tilde{\sigma}^2 = r\hat{\sigma}^2/n.$$

Thus the ML estimate of the mean is obtained, but the ML estimate of the variance is multiplied by the fraction of data observed. When the fraction of missing data is substantial (e.g., $(n-r)/n = 0.5$)), the estimated variance $\tilde{\sigma}^2$ is badly biased, and this bias does not disappear as $n \to \infty$ unless $r/n \to 1$; more relevant asymptotics would fix r/n as the sample size increases.

EXAMPLE 6.26. *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 value of Y for observation i with covariate values 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 $L_{\text{ign}}(\theta|Y_{\text{obs}})$ are obtained by applying least squares regression to the r observed data points. The estimates of β_0 and β obtained by maximizing L_{mispar} are the same as the ML estimates. As in Example 6.25, however, the estimate of variance is the ML estimate multiplied by the proportion of values observed.

EXAMPLE 6.27. *An Exponential Sample with Censored Values.* In Examples 6.25 and 6.26, estimation based on maximizing L_{mispar} at least yields reasonable estimates of the location parameters, even though estimates of the scale parameter need adjustment. In other examples, however, estimates of location also can be badly biased. For example, consider, as in Example 6.22, a censored sample from an exponential distribution with mean θ, where Y_{obs} represents the r observed values, which lie below a known censoring point c, and Y_{mis} represents the $n-r$ values beyond c, which are censored. The ML estimate of θ is $\hat{\theta} = \bar{y} + (n-r)c/r$. Maximization of L_{mispar} in Eq. (6.63) over θ and Y_{mis} (ψ is null) leads to estimating censored values of Y at the censoring point c and estimating θ by $(r/n)\hat{\theta}$. Thus in this case, the estimate of the mean is inconsistent unless the proportion of missing values tends to zero as the sample size increases. Biased estimates of location parameters from L_{mispar} can also occur in problems involving the normal distribution, as Little and Rubin (1983b) show.

EXAMPLE 6.28. *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 an unobserved random effect u_i for subject i, the outcome y_i has the distribution of Eq. (6.9), that is

$$f(y_i|\, x_i, u_i, \beta, \phi) = \exp[y_i\delta(x_i, u_i, \beta) - b(\delta(x_i, u_i, \beta))/\phi + c(y_i, \phi)]. \tag{6.66}$$

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 μ_i and the covariates x_i by the expression

$$\mu_i = g^{-1}\left(\beta_0 + \sum_{j=1}^{p}\beta_j x_{ij} + u_i\right), \tag{6.67}$$

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 loglikelihood is then:

$$\ell(\beta, \phi, \alpha|Y) = \sum_{i=1}^{n}\ln\left(\int f(y_i|\, x_i, u_i, \beta, \phi)f(u_i|\, x_i, \alpha)du_i\right), \tag{6.68}$$

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 the h-likelihood, defined as

$$h(\beta, \phi, \alpha, u|Y) = \sum_{i=1}^{n}\ln[\, f(y_i|\, x_i, u_i, \beta, \phi)] + \ln[\, f(u_i|\, x_i, \alpha)], \tag{6.69}$$

with respect to (β, ϕ, α) and $u = (u_1, \ldots, u_n)^{\mathrm{T}}$. This is the logarithm of $\mathrm{L}_{\mathrm{mispar}}$ in Eq. (6.59), with $\theta = (\beta, \phi, \alpha)$, $Y_{\mathrm{obs}} = (y_1, \ldots, y_n)$ and $Y_{\mathrm{mis}} = (u_1, \ldots, u_n)$. When both distributions are normal, the h-likelihood is the joint loglikelihood maximized by Henderson (1975). Unlike maximization of Eq. (6.68), maximization of Eq. (6.69) does not necessarily give consistent estimates of the parameters (Breslow and Lin, 1995), especially for problems involving nonconjugate distributions for u_i. Algorithms for maximizing the loglikelihood (6.68) are discussed in Breslow and Clayton (1993), McCulloch (1997), and Aitkin (1999).

6.4. LIKELIHOOD THEORY FOR COARSENED DATA

Missing values are a form of data coarsening. Heitjan and Rubin (1991) develop a more general theory for coarsened data that includes heaped, censored, and grouped data as well as missing data. Denote by Y the complete-data matrix in the absence of coarsening with sample space Ψ, and let $f(Y|\theta)$ denote the density of Y under a complete-data model with unknown parameters θ. The observed data Y_{obs} consist of a subset of the sample space Ψ in which Y is known to lie. This subset is a function

of Y and a coarsening variable G that determines the level of precision of Y_{obs}, that is $Y_{\text{obs}} = Y_{\text{obs}}(Y, G)$, subject to the condition that the coarse data contains the unobserved true data, that is $Y \in Y_{\text{obs}}(Y, G)$.

To describe missing data in this framework, suppose Y consists of a data matrix with entries y_{ij}, that are either observed or missing. Then G is simply the missing-data indicator matrix, and $Y_{\text{obs}} = (y_{\text{obs},ij})$ where

$$y_{\text{obs},ij} = \begin{cases} \{y_{ij}\}, & \text{the set consisting of the single true value, if } G_{ij} = 0 \\ \Psi, & \text{the sample space of } Y, \text{ if } G_{ij} = 1. \end{cases}$$

EXAMPLE 6.29. *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 let g_i denote a stochastic censoring time. The complete data are (y_i, g_i), $i = 1, \ldots, n$. The coarsened data for subject i are:

$$y_{\text{obs},i} = y_{\text{obs},i}(y_i, g_i) = \begin{cases} \{y_i\} & \text{if } y_i \le g_i, \\ (g_i, \infty) & \text{if } y_i > g_i. \end{cases}$$

That is, if the event occurs before the censoring point, then $y_{\text{obs},i}$ is the point set consisting of the actual event time y_i, and if the event occurs after the censoring time, then $y_{\text{obs},i}$ is the set of times beyond g_i. Note that in the former case g_i is not observed, unlike the missing-data case where the missing-data indicator is always observed.

Uncertainty in the degree of coarsening is modeled by assigning G a probability distribution with density given $Y = y$ equal to $f(g|\, y; \phi)$. Write $Y = (Y_{\text{obs}}, Y_{\text{mis}})$ and $G = (G_{\text{obs}}, G_{\text{mis}})$, where Y_{mis} is the unobserved part of Y, and G_{obs} and G_{mis} are the observed and missing components of G. The full coarsened-data likelihood is then

$$L_{\text{full}}(\theta, \phi | Y_{\text{obs}}, G_{\text{obs}}) = \int\int f(G|Y, \phi) f(Y|\theta) dY_{\text{mis}} dG_{\text{mis}}, \tag{6.70}$$

and the likelihood ignoring the coarsening mechanism is:

$$L_{\text{ign}}(\theta, \phi | Y_{\text{obs}}) = \int f(Y|\theta) dY_{\text{mis}}. \tag{6.71}$$

The following definitions and lemma generalize the ideas of MAR and ignorable missing-data mechanisms to coarsened data:

Definition 6.6. The data Y_{obs} are coarsened at random (CAR) if:

$$f(g|\, y_{\text{obs}}, y_{\text{mis}}, \phi) = f(g|\, y_{\text{obs}}, \phi) \text{ for all } y_{\text{mis}}.$$

Definition 6.7. The coarsening mechanism is ignorable for likelihood and Bayesian inference if inference for θ based on L_{ign} is equivalent to inference based on the full likelihood L_{full}.

Sufficient conditions for ignoring the coarsening mechanism for likelihood inference are that (a) the data are CAR, and (b) the parameters θ and ϕ are distinct. The coarsening mechanism is ignorable for Bayesian inference if (a) the data are CAR, and (b) the parameters θ and ϕ have independent prior distributions.

EXAMPLE 6.30. *Censoring Mechanisms (Example 6.29 continued).* For the case of censored data and (y_i, g_i) independent of (y_j, g_j) when $i \neq j$, the full likelihood (6.70) is:

$$L_{full}(\theta, \phi | Y_{obs}, G_{obs} = \prod_{i:g_i \geq y_i} \int_{g \geq y_i} f(y_i | x_i, \theta) f(g | x_i, y_i, \phi) dg \prod_{i:g_i < y_i} \int_{y > g_i} f(g_i | x_i, y, \phi) f(y | x_i, \theta) dy, \quad (6.72)$$

where x_i denotes a set of fully observed covariates for subject i, with $f(y_i | x_i, \theta)$ the density of y_i given x_i and $f(g_i | x_i, y_i, \phi)$ the density of g_i given (x_i, y_i). The likelihood (6.71) ignoring the coarsening mechanism is:

$$L_{ign}(\theta | y_{obs}) = \prod_{i:g_i \geq y_i} f(y_i | x_i, \theta) \prod_{i:g_i < y} \int_{y > g_i} f(y_i | x_i, \theta) dy. \quad (6.73)$$

The data are CAR if

$$f(g_i | y_i, x_i, \phi) = f(g_i | x_i, \phi),$$

since otherwise, in general, the integrals in Eq. (6.72) cannot get passed over the first factors to yield Eq. (6.73). 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; if it were MAR, the correct likelihood would simply involve the first product in Eq. (6.73). For more discussion, see Heitjan (1994).

PROBLEMS

6.1. Write the likelihood function for an iid sample from the (a) beta distribution; (b) Poisson distribution; (c) Cauchy distribution.

6.2. Find the score function for the distributions in Problem 6.1. Which have closed-form ML estimates? 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) Relate 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 $\mu(x_i) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$, then ML estimates of $\beta_0, \ldots, \beta_p$ 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 on 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 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), $\mathrm{Var}(y_i|\delta_i, \phi) = \phi b''(\delta_i)$, where $\delta_i = \delta(x_i, \beta)$, and double prime denotes differentiation twice with respect to the function 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 6.14, show that for large n, $\mathrm{LR} = t^2$.

6.12. Derive the posterior distributions in Eqs. (6.22–6.24) for Example 6.15.

6.13. Derive the posterior distributions in Eqs. (6.30–6.32) for Example 6.16.

6.14. Derive the modifications of the posterior distributions in Eqs. (6.30–6.32) for weighted linear regression, discussed at the end of Example 6.16. Show that for 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 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 θ: $Y = (-4.2, -3.2, -2.0, 0.5, 1.5, 1.5, 3.5)$. Compute and compare 90% intervals for θ using (i) the asymptotic distribution based on the observed information, (b) the asymptotic distribution based on the expected information, (c) the posterior distribution assuming a uniform prior on θ.

6.16. Find large-sample variance estimates for the two ML estimates in Example 6.22.

6.17. For a bivariate normal sample on (Y_1, Y_2) with parameters $\theta = (\mu_1, \mu_2, \sigma_{11}, \sigma_{12}, \sigma_{22})$ and values of Y_2 missing, state for the following missing-data mechanisms whether (i) the data are MAR, and (ii) the missing-data mechanism is ignorable for likelihood-based inference.

(a) $\Pr(y_2 \text{ missing} \mid y_1, y_2, \theta, \psi) = \exp(\psi_0 + \psi_1 y_1)/\{1 + \exp(\psi_0 + \psi_1 y_1)\}$, $\psi = (\psi_0, \psi_1)$ distinct from θ.

(b) $\Pr(y_2 \text{ missing} \mid y_1, y_2, \theta, \psi) = \exp(\psi_0 + \psi_1 y_2)/\{1 + \exp(\psi_0 + \psi_1 y_2)\}$, $\psi = (\psi_0, \psi_1)$ distinct from θ.

(d) $\Pr(y_2 \text{ missing} \mid y_1, y_2, \theta, \psi) = 0.5\exp(\mu_1 + \psi y_1)/\{1 + \exp(\mu_1 + \psi y_1)\}$ scalar ψ distinct from θ.

6.18. Suppose that given sets of (possibly overlapping) covariates X_1 and X_2, 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 observations i with x_{i1} and x_{i2} always present, y_{i2} always missing and y_{i1} present if and only if $y_{i2} > 0$. Show that, given the parameters

$$\Pr(y_{i1} \text{ observed} \mid y_{i1}, x_{i1}, x_{i2}) = 1 - \Phi\left(\frac{-x_{i2}\beta_2 - (\rho/\sigma_1)(y_{i1} - x_{i1}\beta_1)}{\sqrt{1-\rho^2}}\right),$$

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 missing-data mechanism is ignorable for likelihood-based inference. (This model is considered in detail in Example 15.7.)

6.19. Describe ML estimates of the parameters in Example 6.24 under (a) the fixed effects model of Eq. (6.56) with fixed parameters $(\{\mu_i\}, \sigma^2)$ and (b) the random effects model of Eqs. (6.53)–(6.55) with parameters $\theta = (\mu, \sigma^2, \tau^2)$, assuming ignorable nonresponse. Show that for the mechanism of Eq. (6.55), the ignorable ML estimates are consistent for (a) but inconsistent for (b), for the asymptotics where $\{n_i\}$ increase, holding I fixed.

6.20. The definition of MAR can depend on how the complete data are defined. Suppose that $X = (x_i, \ldots, x_n)$ is an iid random sample, $Z = (z_1, \ldots, z_n)$ are completely unobserved latent variables, and (x_i, z_i) are bivariate normal with means $(\mu_x, 0)$, variances $(\sigma_x^2, 1)$ and correlation 0 (so X and Z are independent). Suppose that some values x_i are missing, and

$$\Pr(x_i \text{ missing} \mid X, Z) = \exp(z_i)/[1 + \exp(z_i)].$$

Show that if the complete data are defined as X then the data are MCAR, but if the complete data are defined as (X, Z) then the data are not MAR. Which is the more sensible definition?

CHAPTER 7

Factored Likelihood Methods, Ignoring the Missing-Data Mechanism

7.1. INTRODUCTION

We now assume that the missing-data mechanism is ignorable, and for simplicity write $\ell(\theta|Y_{\text{obs}})$ for the ignorable loglikelihood $\ell_{\text{ign}}(\theta|Y_{\text{obs}})$ based on incomplete data Y_{obs}. This can be a complicated function with no obvious maximum and an apparently complicated form for the information matrix. For certain models and incomplete-data patterns, however, analyses based on $\ell(\theta|Y_{\text{obs}})$ can employ standard complete-data techniques. The general idea will be described here in Section 7.1, and specific examples will be given in the remainder of this chapter for normal data and in Section 13.2 for multinomial (that is, cross-classified) data.

For a variety of models and incomplete data patterns, an alternative parameterization $\phi = \phi(\theta)$, where ϕ is a one–one function of θ, can be found such that the loglikelihood decomposes into components

$$\ell(\phi|Y_{\text{obs}}) = \ell_1(\phi_1|Y_{\text{obs}}) + \ell_2(\phi_2|Y_{\text{obs}}) + \cdots + \ell_J(\phi_J|Y_{\text{obs}}), \tag{7.1}$$

where:

1. $\phi_1, \phi_2, \ldots, \phi_J$ are distinct parameters, in the sense that the joint parameter space of $\phi = (\phi_1, \phi_2, \ldots, \phi_J)$ is the product of the individual parameter spaces for $\phi_j, j = 1, \ldots, J$. If a prior distribution is specified, $\phi_1, \phi_2, \ldots, \phi_J$ are assumed mutually *a priori* independent.
2. The components $\ell_j(\phi_j|Y_{\text{obs}})$ correspond to loglikelihoods for complete data problems, or more generally, for easier incomplete-data problems.

If a decomposition with these properties can be found, then since $\phi_1, \ldots, \phi_J$ are distinct, $\ell(\phi|Y_{\text{obs}})$ can be maximized by maximizing $\ell_j(\phi_j|Y_{\text{obs}})$ separately for each j. If $\hat{\phi}$ is the resulting 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 conditions (1) and (2) hold, the posterior distribution of ϕ is the product of J independent posterior distributions of $\phi_1, \phi_2, \ldots, \phi_J$, 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 ϕ^* from the posterior distribution of ϕ and then computing $\theta^* = \theta(\phi^*)$, which is a draw 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 Eq. (7.1) twice with respect to $\phi_1, \ldots, \phi_J$ yields a block diagonal information matrix for ϕ of the form

$$I(\phi|Y_{\text{obs}}) = \begin{bmatrix} I(\phi_1|Y_{\text{obs}}) & & & 0 \\ & I(\phi_2|Y_{\text{obs}}) & & \\ & & \ddots & \\ 0 & & & I(\phi_J|Y_{\text{obs}}) \end{bmatrix}.$$

Hence the large-sample covariance matrix for $\phi - \hat{\phi}$ is also block diagonal, with the form

$$C(\phi - \hat{\phi}|Y_{\text{obs}}) = \begin{bmatrix} I^{-1}(\hat{\phi}_1|Y_{\text{obs}}) & & & 0 \\ & I^{-1}(\hat{\phi}_2|Y_{\text{obs}}) & & \\ & & \ddots & \\ 0 & & & I^{-1}(\hat{\phi}_J|Y_{\text{obs}}) \end{bmatrix}. \tag{7.2}$$

Since the components of this matrix correspond to complete-data problems, they are often relatively easy to calculate. By Property 6.2, the approximate covariance matrix of the ML estimate of a function $\theta = \theta(\phi)$ of ϕ can be found using the formula

$$C(\theta - \hat{\theta}|Y_{\text{obs}}) = D(\hat{\theta})C(\phi - \hat{\phi}|Y_{\text{obs}})D^{\text{T}}(\hat{\theta}), \tag{7.3}$$

where $D(\cdot)$ is the matrix of partial derivatives of θ with respect to ϕ:

$$D(\theta) = \{d_{jk}(\theta)\}, \qquad \text{where } d_{jk}(\theta) = \frac{\partial \theta_j}{\partial \phi_k},$$

and θ is expressed as a column vector.

7.2. BIVARIATE NORMAL DATA WITH ONE VARIABLE SUBJECT TO NONRESPONSE: ML ESTIMATION

7.2.1. ML Estimates

Anderson (1957) first introduced factored likelihoods for the normal data of Example 6.23.

EXAMPLE 7.1. *Bivariate Normal Sample with One Variable Subject to Nonresponse (Example 6.23 continued).* The loglikelihood for a bivariate normal sample with r complete bivariate observations $\{(y_{i1}, y_{i2}),\ i = 1, \ldots, r\}$ and $n - r$ univariate observations $\{y_{i1},\ i = r + 1, \ldots, n\}$ is given by Eq. (6.52). ML estimates of μ and Σ can be 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}:

$$f(y_{i1}, y_{i2}|\mu, \Sigma) = f(y_{i1}|\mu_1, \sigma_{11})\, f(y_{i2}|y_{i1}, \beta_{20\cdot1}, \beta_{21\cdot1}, \sigma_{22\cdot1}),$$

where, by properties of the bivariate normal distribution discussed in Example 6.9, $f(y_{i1}|\mu_1,\ \sigma_{11})$ is the normal distribution with mean μ_1, and variance σ_{11}, and $f(y_{i2}|y_{i1},\ \beta_{20\cdot1},\ \beta_{21\cdot1},\ \sigma_{22\cdot1})$ is the normal distribution with mean

$$\beta_{20\cdot1} + \beta_{21\cdot1} y_{i1}$$

and variance $\sigma_{22\cdot1}$. The parameter

$$\phi = (\mu_1, \sigma_{11}, \beta_{20\cdot1}, \beta_{21\cdot1}, \sigma_{22\cdot1})^{\mathrm{T}}$$

is a one to one function of the original parameter

$$\theta = (\mu_1, \mu_2, \sigma_{11}, \sigma_{12}, \sigma_{23})^{\mathrm{T}}$$

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:

$$\begin{aligned} \beta_{21\cdot1} &= \sigma_{12}/\sigma_{11}, \\ \beta_{20\cdot1} &= \mu_2 - \beta_{21\cdot1}\mu_1, \\ \sigma_{22\cdot1} &= \sigma_{22} - \sigma_{12}^2/\sigma_{11}. \end{aligned} \tag{7.4}$$

Similarly, the components of θ other than μ_1 and σ_{11} can be expressed as the following functions of the components of ϕ:

$$\begin{aligned} \mu_2 &= \beta_{20\cdot1} + \beta_{21\cdot1}\mu_1, \\ \sigma_{12} &= \beta_{21\cdot1}\sigma_{11}, \\ \sigma_{22} &= \sigma_{22\cdot1} + \beta_{21\cdot1}^2\sigma_{11}. \end{aligned} \tag{7.5}$$

The density of the data Y_{obs} factors in the following way:

$$\begin{aligned} f(Y_{\text{obs}}|\theta) &= \prod_{i=1}^{r} f(y_{i1}, y_{i2}|\theta) \sum_{i=r+1}^{n} f(y_{i1}|\theta) \\ &= \left[\prod_{i=1}^{r} f(y_{i1}|\theta) f(y_{i2}|y_{i1}, \theta)\right]\left[\prod_{i=r+1}^{n} f(y_{i1}|\theta)\right] \\ &= \left[\prod_{i=1}^{n} f(y_{i1}|\mu_1, \sigma_{11})\right]\left[\prod_{i=1}^{r} f(y_{i2}|y_{i1}, \beta_{20\cdot1}, \beta_{21\cdot1}, \sigma_{22\cdot1}\right]. \end{aligned} \tag{7.6}$$

The first bracketed factor in Eq. (7.6) is the density of an independent sample of size n from the normal distribution with mean μ_1 and variance σ_{11}. The second factor is the density for r observations from the conditional normal distribution with mean $\beta_{20\cdot1} + \beta_{21\cdot1}y_{i1}$ and variance $\sigma_{22\cdot1}$. Furthermore, if the parameter space for θ is the standard natural parameter space with no prior restrictions, then (μ_1, σ_{11}) and $(\beta_{20\cdot1}, \beta_{21\cdot1}, \sigma_{22\cdot1})$ are distinct, since knowledge of (μ_1, σ_{11}) does not imply any information about $(\beta_{20\cdot1}, \beta_{21\cdot1}, \sigma_{22\cdot1})$. Hence ML estimates of ϕ can be obtained by independently maximizing the likelihoods corresponding to these two components.

Maximizing the first factor yields

$$\begin{aligned} \hat{\mu}_1 &= n^{-1}\sum_{i=1}^{n} y_{i1}, \\ \hat{\sigma}_{11} &= n^{-1}\sum_{i=1}^{n}(y_{i1} - \hat{\mu}_1)^2, \end{aligned} \tag{7.7}$$

that is, the sample mean and sample variance of the n observations $y_{11}, \ldots, y_{n1}$.

Maximizing the second factor uses standard regression results and yields (cf. Example 6.9):

$$\begin{aligned} \hat{\beta}_{21\cdot1} &= s_{12}/s_{11}, \\ \hat{\beta}_{20\cdot1} &= \bar{y}_2 - \hat{\beta}_{21\cdot1}\bar{y}_1, \\ \hat{\sigma}_{22\cdot1} &= s_{22\cdot1}, \end{aligned} \tag{7.8}$$

where $\bar{y}_j = r^{-1}\sum_{i=1}^{r} y_{ij}$, $s_{jk} = r^{-1}\sum_{i=1}^{r}(y_{ij} - \bar{y}_j)(y_{ik} - \bar{y}_k)$ for $j, k = 1, 2$, and $s_{22\cdot1} = s_{22} - s_{12}^2/s_{11}$.

The ML estimates of other parameters can now be obtained using Property 6.1. In particular, from Eq. (7.5),

$$\hat{\mu}_2 = \hat{\beta}_{20\cdot 1} + \hat{\beta}_{21\cdot 1}\hat{\mu}_1,$$

or from Eqs. (7.7) and (7.8),

$$\hat{\mu}_2 = \bar{y}_2 + \hat{\beta}_{21\cdot 1}(\hat{\mu}_1 - \bar{y}_1); \tag{7.9}$$

from Eq. (7.5),

$$\hat{\sigma}_{22} = \hat{\sigma}_{22\cdot 1} + \hat{\beta}^2_{21\cdot 1}\hat{\sigma}_{11},$$

or from Eqs. (7.7) and (7.8),

$$\hat{\sigma}_{22} = s_{22} + \hat{\beta}^2_{21\cdot 1}(\hat{\sigma}_{11} - s_{11}). \tag{7.10}$$

Finally, from Eq. (7.5), we have for the correlation

$$\rho \equiv \sigma_{12}(\sigma_{11}\sigma_{22})^{-1/2} = \beta_{21\cdot 1}\sigma_{11}^{1/2}(\sigma_{22\cdot 1} + \beta^2_{21\cdot 1}\sigma_{11})^{-1/2},$$

so from Eqs. (7.7) and (7.8),

$$\hat{\rho} = [s_{12}(s_{11}s_{22})^{-1/2}](\hat{\sigma}_{11}/s_{11})^{1/2}(s_{22}/\hat{\sigma}_{22})^{1/2}. \tag{7.11}$$

The first terms on the right side of Eqs. (7.9) and (7.10), and the first factor on the right side of Eq. (7.11) are the ML estimates of μ_2, σ_{22} and ρ with the $n - r$ incomplete observations discarded. Thus the remaining terms and factors represent adjustments based on the additional information from the $n - r$ extra values of y_{i1}.

The ML estimate (7.9) of the mean y_{i2} of is of particular interest. It can be written in the form

$$\hat{\mu}_2 = n^{-1}\left(\sum_{i=1}^{r} y_{i2} + \sum_{i=r+1}^{n} \hat{y}_{i2}\right), \tag{7.12}$$

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 linear regression of observed y_{i2} on observed y_{i1}.

EXAMPLE 7.2. *Bivariate Normal Numerical Illustration.* The first $r = 12$ observations in Table 7.1, taken from Snedecor and Cochran (1967, Table 6.9.1), give measurements on the size of crop on apple trees, in hundreds of fruits (y_{i1}) and 100 times the percentage of wormy fruits (y_{i2}). These observations 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 for some of the trees with smaller crops, numbered 13 to 18 in the table, the value of y_{i2} was not determined. The sample mean, $\bar{y}_2 = 45$, underestimates the percentage of wormy fruits, since the percentage for the six omitted trees is expected to be larger than the percentage for the measured trees because the omitted trees tended to be smaller (i.e., the data may be MAR but do not appear to be 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 observations. 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 normality assumptions.

Table 7.1 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

$\bar{y}_1 = 19$; $\bar{y}_2 = 45$; $\hat{\mu}_2 = 49.333$; $\hat{\mu}_1 = 14.7222$
$s_{11} = 77.0$; $s_{12} = -78.0$; $s_{22} = 101/8333$; $\hat{\sigma}_{11} = 89.5340$
Source: Adapted from Snedecor and Cochran (1967, Table 6.9.1).

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),

$$\ell(\phi|Y_{\text{obs}}) = -(2\sigma_{22\cdot1})^{-1}\sum_{i=1}^{r}(y_{i2} - \beta_{20\cdot1} - \beta_{21\cdot1}y_{i1})^2 - \frac{1}{2}r\ln\sigma_{22\cdot1}$$
$$- (2\sigma_{11})^{-1}\sum_{i=1}^{n}(y_{i1} - \mu_1)^2 - \frac{1}{2}n\ln\sigma_{11}.$$

Differentiating twice with respect to ϕ gives

$$I(\hat{\phi}|Y_{\text{obs}}) = \begin{bmatrix} I(\hat{\mu}_1, \hat{\sigma}_{11}|Y_{\text{obs}}) & 0 \\ 0 & I(\hat{\beta}_{20\cdot1}, \hat{\beta}_{21\cdot1}, \hat{\sigma}_{22\cdot1}|Y_{\text{obs}}) \end{bmatrix},$$

where

$$I(\hat{\mu}_1, \hat{\sigma}_{11}|Y_{\text{obs}}) = \begin{bmatrix} n/\hat{\sigma}_{11} & 0 \\ 0 & n/(2\hat{\sigma}_{11}^2) \end{bmatrix},$$

and

$$I(\hat{\beta}_{20\cdot1}, \hat{\beta}_{21\cdot1}, \hat{\sigma}_{22\cdot1}|Y_{\text{obs}}) = \begin{bmatrix} r/\hat{\sigma}_{22\cdot1} & r\bar{y}_1/\hat{\sigma}_{22\cdot1} & 0 \\ r\bar{y}_1/\hat{\sigma}_{22\cdot1} & \sum_{i=1}^{r} y_{i1}^2/\hat{\sigma}_{22\cdot1} & 0 \\ 0 & 0 & r/(2\hat{\sigma}_{22\cdot1}^2) \end{bmatrix}.$$

Inverting these matrices yields the large-sample covariance matrix of $(\phi - \hat{\phi})$:

$$C(\phi - \hat{\phi}) = \begin{bmatrix} I^{-1}(\hat{\mu}_1, \hat{\sigma}_{11}|Y_{\text{obs}}) & 0 \\ 0 & I^{-1}(\hat{\beta}_{20\cdot1}, \hat{\beta}_{21\cdot1}, \hat{\sigma}_{22\cdot1}|Y_{\text{obs}}) \end{bmatrix},$$

where

$$I^{-1}(\hat{\mu}_1, \hat{\sigma}_{11}|Y_{\text{obs}}) = \begin{bmatrix} \hat{\sigma}_{11}/n & 0 \\ 0 & 2\hat{\sigma}_{11}^2/n \end{bmatrix}$$

and

$$I^{-1}(\hat{\beta}_{20\cdot1}, \hat{\beta}_{21\cdot1}, \hat{\sigma}_{22\cdot1}|Y_{\text{obs}}) = \begin{bmatrix} \hat{\sigma}_{22\cdot1}(1 + \bar{y}_1^2/s_{11})/r & -\bar{y}_1\hat{\sigma}_{22\cdot1}/(rs_{11} & 0 \\ -\bar{y}_1\hat{\sigma}_{22\cdot1}/(rs_{11}) & \hat{\sigma}_{22\cdot1}/(rs_{11}) & 0 \\ 0 & 0 & 2\hat{\sigma}_{22\cdot1}^2/r \end{bmatrix}.$$

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. Since $\mu_2 = \beta_{20\cdot1} + \beta_{21\cdot1}\mu_1$, 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

$$D(\hat{\mu}_2)C(\phi - \hat{\phi})D(\hat{\mu}_2)^{\mathrm{T}} = \hat{\sigma}_{22\cdot1}\left[\frac{1}{r} + \frac{\hat{\rho}^2}{n(1-\hat{\rho}^2)} + \frac{(\bar{y}_1 - \hat{\mu}_1)^2}{rs_{11}}\right]. \tag{7.13}$$

The third term in the brackets is $O(r^{-2})$ if the data are MCAR, because $(\bar{y}_1 - \hat{\mu}_1)^2$ is $O(r^{-1})$ in this case. Ignoring this term yields

$$\hat{\sigma}_{22\cdot1}\left[\frac{1}{r} + \frac{\hat{\rho}^2}{n(1-\hat{\rho}^2)}\right] = \frac{\hat{\sigma}_{22}}{r}\left(1 - \hat{\rho}^2\frac{n-r}{n}\right). \tag{7.14}$$

This expression can be compared with the variance of $\bar{y}_2$, namely σ_{22}/r. Thus in large samples, under MCAR, the proportionate reduction in variance obtained by including the $n - r$ observations on y_1 alone is ρ^2 times the fraction of incomplete observations, $(n - r)/n$.

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. (6.14)), as discussed in Section 6.1.3. In particular, a 95% interval for μ_2 takes the form

$$\hat{\mu}_2 \pm 1.96\sqrt{\mathrm{Var}(\hat{\mu}_2 - \mu_2)}, \tag{7.15}$$

where $\mathrm{Var}(\hat{\mu}_2 - \mu_2)$ is approximated by Eq. (7.13). For parameters other than means or regression coefficients, better intervals are obtained by applying a transformation to 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, the appropriate transformation

for a variance is the logarithm, so to compute a 95% interval for σ_{22}, a 95% interval for $\log \sigma_{22}$ is computed as

$$\log \hat{\sigma}_{22} \pm 1.96\sqrt{\text{Var}(\ln \hat{\sigma}_{22} - \ln \sigma_{22})}, \tag{7.16}$$

where in large samples, $\text{Var}(\ln \hat{\sigma}_{22} - \ln \sigma_{22}) = \text{Var}(\hat{\sigma}_{22} - \sigma_{22})/\hat{\sigma}_{22}^2$. The 95% interval for $\hat{\sigma}_{22}$ is then $(\exp(l), \exp(u))$ where (l, u) is the interval computed from (7.16).

Small-sample inference is problematic from a frequentist perspective. The quantity $(\hat{\mu}_2 - \mu_2)/\sqrt{\text{Var}(\hat{\mu}_2 - \mu_2)}$ obtained from Eq. (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 $m - 1$ 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 $(\hat{\mu}_2 - \hat{\mu}_1 - \mu_2 + \mu_1)/\sqrt{\text{Var}(\hat{\mu}_2 - \hat{\mu}_1 - \mu_2 + \mu_1)}$. Approximate small-sample confidence intervals for other parameters, such as ρ, do not appear to have been developed in missing-data situations.

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 posterior distribution. Specifically, suppose μ_1, σ_{11}, $\beta_{20\cdot1}$, $\beta_{21\cdot1}$, and $\sigma_{22\cdot1}$ are assumed *a priori* independent with reference prior

$$f(\mu_1, \sigma_{11}, \beta_{20\cdot1}, \beta_{21\cdot1}, \sigma_{22\cdot1}) \propto \sigma_{11}^{-a}\sigma_{22\cdot1}^{-c}. \tag{7.17}$$

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 $\{y_{i1}: i = 1, \ldots, n\}$, we have the following results: the posterior distribution of (μ_1, σ_{11}) is such that (1) $n\hat{\sigma}_{11}/\sigma_{11}$ has a chi-squared distribution with $n + 2a - 3$ degrees of freedom, and (2) 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 $\{(y_{i1}, y_{i2}): i = 1, \ldots, r\}$, the posterior distribution of $(\beta_{20\cdot1}, \beta_{21\cdot1}, \sigma_{22\cdot1})$ is such that (3) $r\hat{\sigma}_{22\cdot1}/\sigma_{22\cdot1}$ has a chi-squared distribution with $r + 2c - 4$ degrees of freedom, (4) the posterior distribution of $\beta_{21\cdot1}$ given $\sigma_{22\cdot1}$ is normal with mean $\hat{\beta}_{21\cdot1}$ and variance $\sigma_{22\cdot1}/(rs_{11})$, and (5) the posterior distribution of $\beta_{20\cdot1}$ given $\beta_{21\cdot1}$ and $\sigma_{22\cdot1}$ is normal with mean $\bar{y}_2 - \beta_{21\cdot1}\bar{y}_1$ and variance $\sigma_{22\cdot1}/r$; furthermore, (6) (μ_1, σ_{11}) and $(\beta_{20\cdot1}, \beta_{21\cdot1}, \sigma_{22\cdot1})$ and are *a posteriori* independent. For derivations of these results, see Lindley (1965).

Results 1–6 and Property 6.1B imply that the posterior distribution of any function $g(\phi)$ of the parameters ϕ can be simulated by creating D draws g_d, $d = 1, \ldots, D$, as follows:

1. Draw independently x_{1t}^2 and x_{2t}^2 from chi-squared distributions with, respectively, $n + 2a - 3$ and $r + 2c - 4$ degrees of freedom. Draw three independent standard normal deviates z_{1t}, z_{2t}, and z_{3t}.

2. Compute $\phi^{(d)} = (\sigma_{11}^{(d)}, \mu_1^{(d)}, \sigma_{22\cdot1}^{(d)}, \beta_{20\cdot1}^{(d)}, \beta_{21\cdot1}^{(d)})^{\mathrm{T}}$, where

$$\begin{aligned}
\sigma_{11}^{(d)} &= n\hat{\sigma}_{11}/\chi_{1t}^2 \\
\mu_1^{(d)} &= \hat{\mu}_1 + z_{1t}(\sigma_{11}^{(d)}/n)^{1/2} \\
\sigma_{22\cdot1}^{(d)} &= r\hat{\sigma}_{22\cdot1}/\chi_{2t}^2 \\
\beta_{21\cdot1}^{(d)} &= \hat{\beta}_{21\cdot1} + z_{2t}[\sigma_{22\cdot1}^{(d)}/(rs_{11})]^{1/2} \\
\beta_{20\cdot1}^{(d)} &= \bar{y}_2 - \beta_{21\cdot1}^{(d)}\bar{y}_1 + z_{3t}(\sigma_{22\cdot1}^{(d)}/r)^{1/2}.
\end{aligned}$$

3. Compute $g_d = g(\phi^{(d)})$. For example, if $g(\phi) \equiv \mu_2 = \beta_{20\cdot1} + \beta_{21\cdot1}\mu_1$, then $g_d = \beta_{20\cdot1}^{(d)} + \beta_{21\cdot1}^{(d)}\mu_1^{(d)}$.

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. Asymptotic intervals 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 normalizing transformation of the correlation. In addition, with this method, intervals are obtained using the complete-case degrees of freedom, that is, the normal percentile 1.96 is replaced by the 97.5th percentile of the t distribution on $r - 1 = 11$ degrees of freedom, namely, 2.201.
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

Table 7.2 95% Intervals for Parameters of Bivariate Normal Distribution, Based on Data in Table 7.1

	Parameters		
Method	μ_2	σ_{22}	ρ
(1) Asymptotic theory	(43.98, 54.69)	(30.68, 198.71)	(−1.002, −0.788)
(2) Asymptotic theory, with transform and t approximation	(43.38, 55.28)	(50.80, 258.94)	(−0.968, −0.689)
(3)(i) Bayes simulation A	(43.71, 54.42)	(60.06, 289.73)	(−0.964, −0.662)
(3)(ii) Bayes simulation B	(43.54, 55.68)	(59.18, 293.93)	(−0.964, −0.656)
(4)(i) Normal approximation to A	(43.50, 55.21)	(15.91, 256.68)	(−1.029, −0.717)
(4)(ii) Normal approximation to B	(43.42, 55.22)	(14.62, 257.19)	(−1.033, −0.710)
ML estimates	49.33	114.70	−0.895

with 9999 simulated values; and

4. Intervals obtained by fitting normal distributions to the simulated posterior distributions from method 3, using the simulated posterior mean and variance.

Methods 3 and 4 are repeated for two independent sets of random numbers (i and ii) to give some idea of the simulation variance.

The interval for μ_2 from method (1) is shorter than for the other methods, and presumably has lower than the stated 95% coverage since 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 require more extensive simulation; for example, Little (1988a) studies coverage properties of various t approximations to the posterior distribution of μ_2.

7.4 MONOTONE DATA WITH MORE THAN TWO VARIABLES

7.4.1. Multivariate Data With One Normal Variable Subject to Nonresponse

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 Nonresponse.* Suppose we replace y_{i1} by a set of K completely observed variables, as for the data pattern in Figure 1.1*a*. This results in a special case of monotone data with $J = 2$ and Y_i representing K variables. Suppose first (y_{i1}, y_{i2}) are iid $(K + 1)$-variate normally distributed, and the data are MAR. The ML estimates of μ_2 and σ_{22} are then

$$\hat{\mu}_2 = \bar{y}_2 + (\hat{\mu}_1 - \bar{y}_1)^{\mathrm{T}} \hat{\beta}_{21\cdot 1}, \tag{7.18}$$

and

$$\hat{\sigma}_{22} = s_{22} + \hat{\beta}_{21\cdot 1}^{\mathrm{T}} (\hat{\sigma}_{11} - s_{11}) \hat{\beta}_{21\cdot 1},$$

where $\hat{\mu}_1$, $\bar{y}_1$ are $(K \times 1)$ mean vectors, $\hat{\beta}_{21\cdot 1}$ is the $(K \times 1)$ vector of regression coefficients from the multiple regression of y_{i2} on y_{i1}, and $\hat{\sigma}_{11}$ and s_{11} are $(K \times K)$

covariance matrices, $\hat{\sigma}_{11}$ based on all n observations of y_{i1} and s_{11} based on the r observations 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 multiple regression of y_{i2} on y_{i1}.

More generally, Eq. (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 $(\beta_{20\cdot1} + y_{i1}\beta_{21\cdot1})$ and variance $\sigma_{22\cdot1}$.
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 $\beta_{20\cdot1}$, $\beta_{21\cdot1}$, $\sigma_{22\cdot1}$ are distinct from the parameters of this distribution.

An important special case arises with *dummy variable regression*, where y_{i1} represents K dummy variables indicating $K + 1$ groups. The kth component of y_{i1} is defined to be one if the ith observation belongs to group k and is zero otherwise. For an observation in group 1, $y_{i1} = (1, 0, 0, \ldots, 0)$; for an observation in group 2, $y_{i1} = (0, 1, 0, \ldots, 0)$; for an observation in group K, $y_{i1} = (0, 0, \ldots, 0, 1)$; and for an observation in group $K + 1$, $y_{i1} = (0, 0, \ldots, 0, 0)$; 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 observations in each of the first K groups, μ_1 is the corresponding vector of expected proportions, and condition 2 above 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 constant variance $\sigma_{22\cdot1}$.

By the properties of dummy variable regression, the predicted value of y_{i2} for an observation in group k is the mean of the observed values of y_{i2} in group k. Thus the ML estimate corresponds to imputing the subclass 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. Factorization of the 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 1.1*c*, where for each observation i, y_{ij} is recorded if $y_{i,\,j+1}$ is recorded $(J = 1, \ldots, J - 1)$, 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 Section 7.2.2 and 7.3.

The appropriate factorization for this pattern is

$$\prod_{i=1}^{n} f(y_{i1}, \ldots, y_{iJ}|\phi) =$$
$$\prod_{i=1}^{n} f(y_{i1}|\phi_1)$$
$$\prod_{i=1}^{r_2} f(y_{i2}|y_{i1}, \phi_2) \cdots \prod_{i=1}^{r_J} f(y_{iJ}|y_{i1}, \ldots, y_{i,J-1}, \phi_J),$$

where for $j = 1, \ldots, J$, $f(y_{ij}|y_{i1}, \ldots, y_{i,j-1}, \phi_j)$ is the conditional distribution of y_{ij} given $y_{i1}, \ldots, y_{i,j-1}$, indexed by the parameter ϕ_j. If $(y_{i1}, \ldots, y_{iJ})$ follows a multivariate normal distribution, then $f(y_{ij}|y_{i1}, \ldots, y_{i,j-1}, \phi_j)$ is a normal distribution with mean linear in $y_{i1}, \ldots, y_{i,j-1}$ 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 $y_{i1}, \ldots, y_{i,j-1}$, using the set of observations for which $y_{i1}, \ldots, y_{ij}$ are all observed.

EXAMPLE 7.5. *Multivariate Normal Monotone Data*. Marini, Olsen and Rubin (1980) provide a numerical illustration of ML estimation for monotone data with $J > 2$ patterns, for panel study data with 4352 cases. The data pattern, given in Table 1.1, does not have a monotone pattern. 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 observations.
2. Calculate the multivariate linear regression of the next most observed variables, block 2 on block 1, from observations with both block 1 and block 2 variables recorded.
3. Calculate the multivariate linear regression of block 3 on blocks 1 and 2, from observations with blocks 1–3 recorded.
4. Calculate the multivariate linear regression of block 4 on blocks 1–3 from observations 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 to 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 gives a description of the variables. The next two columns give ML estimates of the means μ_{ML} and the standard deviations (σ_{ML}) of each variable. The rest of the table compares estimates of two alternative methods of estimation. Estimates (μ_A, σ_A) from the available-case method are the sample means and standard deviations using all the observations available for each variable (cf. Section 3.4). The two columns after the estimates indicate the magnitude of differences between the ML and available-case methods, measured in percent standard deviations. Even though these estimates of marginal parameters are quite close to the ML estimates, the available-case method is not recommended for measures of association such as covariances or regression coefficients, as noted in Chapter 3.

The last four columns of the table present and compare estimates (μ_{cc}, σ_{cc}) based only on the 1594 complete observations, the complete-case method discussed in Chapter 3. Estimates of means from this procedure can differ markedly from the ML estimates. For example, the estimate for grade point average is 0.35 of a standard

Table 7.3 Maximum Likelihood Estimates of Means and Standard Deviations and Comparisons with Two Alternative Sets of Estimates

	ML Estimates for Total Sample		Estimates Based on Available Cases				Estimates Based on Complete Cases			
Variable	Mean	S.D.	Mean	S.D.	$\frac{(\mu_A - \mu_{ML})}{\sigma_{ML}} \times 100$	$\frac{(\sigma_A - \sigma_{ML})}{\sigma_{ML}} \times 100$	Mean	S.D.	$\frac{(\mu_C - \mu_{ML})}{\sigma_{ML}} \times 100$	$\frac{(\sigma_C - \sigma_{ML})}{\sigma_{ML}} \times 100$
Block I Variables: Measured during Adolescence										
Father's education	11.702	3.528	11.702	3.528	0.0	0.0	12.050	3.449	9.9	−1.6
Mother's education	11.508	2.947	11.508	2.947	0.0	0.0	11.864	2.865	12.1	−2.4
Father's occupation	6.115	2.904	6.115	2.904	0.0	0.0	6.407	2.868	10.1	−1.2
Intelligence	106.625	12.910	106.625	12.910	0.0	0.0	109.036	11.174	18.7	−13.4
College preparatory curriculum	0.411	0.492	0.411	0.492	0.0	0.0	0.528	0.499	13.4	1.4
Time spent on homework	1.589	0.814	1.589	0.814	0.0	0.0	1.633	0.795	5.4	−2.3
Grade point average	2.324	0.773	2.324	0.773	0.0	0.0	2.594	0.701	34.9	−9.3
College plans	0.488	0.500	0.488	0.500	0.0	0.0	0.595	0.491	21.4	−1.8
Friends' college plans	0.512	0.369	0.512	0.369	0.0	0.0	0.572	0.354	16.3	−4.1
Participation in extracurricular activities	0.413	0.492	0.413	0.492	0.0	0.0	0.492	0.500	15.8	1.4
Membership in top leading crowd	0.088	0.283	0.088	0.283	0.0	0.0	0.131	0.338	8.6	19.4
Membership in intermediate leading crowd	0.170	0.376	0.170	0.376	0.0	0.0	0.198	0.399	5.6	4.1
Cooking/drinking	0.570	1.032	0.570	1.032	0.0	0.0	0.483	0.835	−8.4	−9.4
Dating frequency at time of survey	4.030	4.802	4.030	4.802	0.0	0.0	3.701	4.523	−6.8	−5.8
Liking for self	2.366	0.525	2.366	0.525	0.0	0.0	2.364	0.515	−0.4	−1.9
Grade in school	2.432	1.048	2.432	1.048	0.0	0.0	2.496	1.064	6.1	1.5

Block 2 Variables: Measured for All Follow-up Respondents										
Educational attainment	13.625	2.295	13.274	2.262	5.5	−1.4	14.196	2.204	24.9	−4.0
Occupational prestige	44.405	13.008	45.085	12.893	5.2	−0.9	47.056	12.745	20.4	−2.0
Marital status	0.940	0.238	0.940	0.238	0.0	0.0	0.940	0.237	0.0	−0.4
Number of children	1.991	1.306	1.973	1.304	−1.4	−0.2	1.928	1.242	−4.8	−4.9
Age	30.629	1.221	30.655	1.225	2.1	0.3	30.726	1.152	7.9	−5.4
Father's occupational prestige	43.998	14.821	44.258	14.786	1.8	−0.2	44.782	14.333	5.3	−3.2
Block 3 Variables: Measured Only for Initial Questionnaire Respondents to the Follow-up										
Personal esteem	3.128	0.377	3.148	0.378	5.2	0.3	3.148	0.373	5.3	−1.1
Dating frequency during last two years of high school	4.374	3.408	4.202	3.261	−5.1	−1.4	4.213	3.352	−4.7	−1.6
Number of siblings	2.219	1.748	2.099	1.744	−6.9	−0.2	2.055	1.660	−9.4	−5.0
Block 4 Variables: Measured on Parents' Questionnaire										
Family income	4.092	1.530	4.075	1.538	−1.1	0.5	4.215	1.570	8.0	2.6
Parental encouragement to go to college	0.714	0.434	0.706	0.455	−1.6	4.8	0.754	0.431	8.0	−0.7
Number of children in family of origin	3.039	1.539	3.067	1.671	1.8	8.6	2.975	1.551	−4.2	0.8

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

In this section we review the use of the *sweep operator* (Beaton, 1964) in linear regression with complete observations and show how this operator provides a simple and convenient way of performing 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 11 when we consider ML estimation for normal data with a general pattern of incompleteness.

The sweep operator is defined for symmetric matrices as follows. A $p \times p$ 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:

$$\begin{aligned} h_{kk} &= -1/g_{kk} \\ h_{jk} &= h_{kj} = g_{jk}/g_{kk}, \qquad j \neq k, \\ h_{jl} &= g_{jl} - g_{jk}g_{kl}/g_{kk}, \qquad j \neq k, l \neq k. \end{aligned} \tag{7.19}$$

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 SWP$[k]G$ to denote the matrix H defined by (7.19). Also, the result of successively applying the operations SWP$[k_1]$, SWP$[k_2]$, ..., SWP$[k_t]$ to the matrix G will be denoted by SWP$[k_1, k_2, \ldots, k_t]G$. In actual computations, the sweep operation is most 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_{jl} 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,

$$\mathrm{SWP}[j,k]G = \mathrm{SWP}[k,j]G.$$

It follows more generally that

$$\mathrm{SWP}[j_1,\ldots,j_t]G = \mathrm{SWP}[k_1,\ldots,k_t]G,$$

where $j_1,\ldots,j_t$ is any permutation of the set $k_1,\ldots,k_t$. 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. If $H = \mathrm{SWP}[1]G$, then h_{12} is the regression coefficient of Y_1 from the regression of Y_2 on Y_1, and is the residual variance of Y_2. Furthermore, if G is a sample covariance matrix from n independent observations, then $-h_{11}h_{22}/n$ is the estimated variance of the sample regression coefficient, h_{12}.

More generally, suppose we have a sample of n observations on K variables $Y_1,\ldots,Y_K$. Let G denote the $(K+1)\times(K+1)$ matrix

$$G = \begin{bmatrix} 1 & \bar{y}_1 & \cdots & \bar{y}_j & \cdots & \bar{y}_K \\ \bar{y}_1 & n^{-1}\sum y_1^2 & & & \cdots & n^{-1}\sum y_K y_1 \\ \vdots & \vdots & \ddots & & & \vdots \\ \bar{y}_k & & & n^{-1}\sum y_j y_k & & \\ \vdots & & & & \ddots & \\ \bar{y}_K & n^{-1}\sum y_1 y_K & & & \cdots & n^{-1}\sum y_K^2 \end{bmatrix},$$

where $\bar{y}_1,\ldots,\bar{y}_K$ are the sample means, and summations are over the n observations. 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

$$\mathrm{SWP}[0]G = \begin{bmatrix} -1 & \bar{y}_1 & \cdots & \bar{y}_j & \cdots & \bar{y}_K \\ \bar{y}_1 & s_{11} & & & \cdots & s_{K1} \\ \vdots & & \ddots & & & \vdots \\ \bar{y}_k & & & s_{jk} & & \\ \vdots & & & & \ddots & \\ \bar{y}_K & s_{1K} & & & \cdots & s_{KK} \end{bmatrix}, \tag{7.20}$$

where s_{jk} is the sample covariance of Y_j and Y_k with factor n^{-1} rather than $(n-1)^{-1}$. This operation corresponds to correcting the scaled cross-products matrix of $Y_1,\ldots,Y_K$, G, for the means of $Y_1,\ldots,Y_K$ to create the covariance matrix. In terms of regression, the means in the first row and column of $\mathrm{SWP}[0]G$ are coefficients from the regression of $Y_1,\ldots,Y_K$ on the constant term $Y_0 \equiv 1$, and the

corrected, scaled cross-products matrix $\{s_{jk}\}$ is the residual covariance matrix from this regression. Thus we also call this process sweeping on the constant term. We call matrix (7.20) the *augmented covariance matrix* of the variables $Y_1, \ldots, Y_K$.

Sweeping matrix (7.20) on row and column 1, corresponding to Y_1, yields the symmetric matrix

$$\text{SWP}[0,1]G = \begin{bmatrix} -(1+\bar{y}_1^2/s_{11}) & \bar{y}_1/s_{11} & \bar{y}_2-(s_{12}/s_{11})\bar{y}_1 & \cdots & \bar{y}_K-(s_{1K}/s_{11})\bar{y}_1 \\ & -1/s_{11} & s_{12}/s_{11} & \cdots & s_{1K}s_{11} \\ \vdots & & s_{22}-s_{12}^2/s_{11} & \cdots & s_{2K}-s_{1K}s_{12}/s_{11} \\ & & & & \vdots \\ \bar{y}_K-(s_{1K}/s_{11})\bar{y}_1 & & \cdots & & s_{KK}-s_{1K}^2/s_{11} \end{bmatrix} = \begin{bmatrix} -A & B \\ B^{\mathrm{T}} & C \end{bmatrix},$$

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 $Y_2, \ldots, Y_K$ on Y_1. In particular, the jth column of B gives the intercept and slope for the regression of Y_{j+1} on Y_1 for $j = 1, \ldots, K-1$. The matrix C gives the residual covariance matrix of $Y_2, \ldots, Y_K$ given Y_1. Finally, the elements of A, when multiplied by the appropriate residual variance or covariance in C and divided by n, yield 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 $Y_{q+1}, \ldots, Y_K$ on $Y_1, \ldots, Y_q$. Specifically, letting

$$\text{SWP}[0,1,\ldots,q]G = \begin{bmatrix} -D & E \\ E^{\mathrm{T}} & F \end{bmatrix},$$

where D is $(q+1) \times (q+1)$, E is $(q+1) \times (K-q)$ and F is $(K-q) \times (K-q)$, the jth column of E gives the least squares intercept and slopes of the regression of Y_{j+q} on $Y_1, \ldots, Y_q$, for $j = 1, 2, \ldots, K-q$; the matrix F is the residual covariance matrix of $Y_{q+1}, \ldots, Y_K$; and the elements of D can be used as above to give variances and covariances of the estimated regression coefficients in E.

In summary, ML estimates for the multivariate linear regression of $Y_{q+1}, \ldots, Y_K$ on $Y_1, \ldots, Y_q$ can be found by sweeping the rows and columns corresponding to the constant term and the predictor variables $Y_1, \ldots, Y_q$ out of the scaled cross-products matrix G.

The operation of sweeping on 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

$$\begin{aligned} h_{kk} &= -1/g_{kk} \\ h_{jk} &= h_{kj} = -g_{jk}/g_{kk}, \qquad j \neq k, \\ h_{jl} &= g_{jk}g_{kl}/g_{kk}, \qquad j \neq k, l \neq k. \end{aligned} \tag{7.21}$$

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 Eqs. (7.4) and (7.5) can be compactly expressed using the SWP[ ] and RSW[ ] notation. Also important, numerical values of ML estimates can be simply computed using these operators.
Suppose we arrange $\theta = (\mu_1, \mu_2, \sigma_{11}, \sigma_{12}, \sigma_{22})^{\mathrm{T}}$ in the following symmetric matrix, which is the population analog of matrix (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 we obtain from Eq. (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 Eq. (7.4) reveals that row and column 2 of SWP[1]θ^* provide the intercept $(\mu_2 - \mu_1\sigma_{12}/\sigma_{11})$, the slope of the regression of Y_2 on Y_1 $(\sigma_{12}/\sigma_{11})$ and the residual variance $\sigma_{22\cdot 1} = \sigma_{22} - \sigma_{12}^2/\sigma_{11}$. Also, although not in a particularly familiar

form, the 2×2 submatrix formed by rows and columns 0 and 1 provides the parameters of the distribution of Y_1. To see this, write

$$\phi^* = \text{SWP}[1]\theta^* = \begin{bmatrix} \text{SWP}[1] & \begin{bmatrix} -1 & \mu_1 \\ \mu_1 & \sigma_{11} \end{bmatrix} & \begin{bmatrix} \beta_{20\cdot1} \\ \beta_{21\cdot1} \end{bmatrix} \\ & \begin{bmatrix} \beta_{20\cdot1} & \beta_{21\cdot1} \end{bmatrix} & \sigma_{22\cdot1} \end{bmatrix}, \tag{7.22}$$

where ϕ^* is a slightly modified version of $\phi = (\mu_1,\ \sigma_{11},\ \beta_{20\cdot1},\ \beta_{21\cdot1},\ \sigma_{22\cdot1})^{\mathrm{T}}$ displayed as a matrix. By Property 6.1, a similar expression relates the ML estimates of θ to the ML estimates of ϕ:

$$\hat{\phi}^* = \text{SWP}[1]\hat{\theta}^* = \begin{bmatrix} \text{SWP}[1] & \begin{bmatrix} -1 & \hat{\mu}_1 \\ \hat{\mu}_1 & \hat{\sigma}_{11} \end{bmatrix} & \begin{bmatrix} \hat{\beta}_{20\cdot1} \\ \hat{\beta}_{21\cdot1} \end{bmatrix} \\ & \begin{bmatrix} \hat{\beta}_{20\cdot1} & \hat{\beta}_{21\cdot1} \end{bmatrix} & \hat{\sigma}_{22\cdot1} \end{bmatrix}.$$

Applying the RSW[1] operator to both sides yields

$$\hat{\theta}^* = \text{RSW}[1]\begin{bmatrix} \text{SWP}[1] & \begin{bmatrix} -1 & \hat{\mu}_1 \\ \hat{\mu}_1 & \hat{\sigma}_{11} \end{bmatrix} & \begin{bmatrix} \hat{\beta}_{20\cdot1} \\ \hat{\beta}_{21\cdot1} \end{bmatrix} \\ & \begin{bmatrix} \hat{\beta}_{20\cdot1} & \hat{\beta}_{21\cdot1} \end{bmatrix} & \hat{\sigma}_{22\cdot1} \end{bmatrix}. \tag{7.23}$$

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 1.1*c*, 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 covariance matrix of Y_1 based on all the observations.

STEP 2. Find the ML estimates $\hat{\beta}_{20\cdot1}$, $\hat{\beta}_{21\cdot1}$, and $\hat{\Sigma}_{22\cdot1}$ 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 observations with Y_1 and Y_2 both observed.

STEP 3. Find the ML estimates $\hat{\beta}_{30\cdot12}$, $\hat{\beta}_{31\cdot12}$, $\hat{\beta}_{32\cdot12}$, and $\hat{\Sigma}_{33\cdot12}$ 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 complete observations with Y_1, Y_2, and Y_3 observed.

STEP 4. Calculate the matrix

$$A = \text{SWP}[1]\begin{bmatrix} -1 & \hat{\mu}_1^{\mathrm{T}} \\ \hat{\mu}_1 & \hat{\Sigma}_{11} \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix},$$

where SWP[1] is shorthand for sweeping on the set of variables Y_1.

STEP 5. Calculate the matrix

$$B = \text{SWP}[2]\begin{bmatrix} a_{11} & a_{12} & \hat{\beta}_{20\cdot1}^{\mathrm{T}} \\ a_{21} & a_{22} & \hat{\beta}_{21\cdot1}^{\mathrm{T}} \\ \hat{\beta}_{20\cdot1} & \hat{\beta}_{21\cdot1} & \hat{\Sigma}_{22\cdot1} \end{bmatrix} = \begin{bmatrix} c_{11} & c_{12} & c_{13} \\ c_{21} & c_{22} & c_{23} \\ c_{31} & c_{32} & c_{33} \end{bmatrix},$$

where 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

$$\begin{bmatrix} -1 & \hat{\mu}^{\mathrm{T}} \\ \hat{\mu} & \hat{\Sigma} \end{bmatrix} = \text{RSW}[1,2]\begin{bmatrix} c_{11} & c_{12} & c_{13} & \hat{\beta}_{20\cdot1}^{\mathrm{T}} \\ c_{21} & c_{22} & c_{23} & \hat{\beta}_{31\cdot12}^{\mathrm{T}} \\ c_{31} & c_{32} & c_{33} & \hat{\beta}_{32\cdot12}^{\mathrm{T}} \\ \hat{\beta}_{20\cdot1} & \hat{\beta}_{31\cdot12} & \hat{\beta}_{32\cdot12} & \hat{\Sigma}_{33\cdot12} \end{bmatrix}.$$

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

$$\begin{bmatrix} -1 & \hat{\mu}^{\mathrm{T}} \\ \hat{\mu} & \hat{\Sigma} \end{bmatrix} = \text{RSW}[1,2]\begin{bmatrix} \text{SWP}[2] & \begin{bmatrix} \text{SWP}[1] & \begin{bmatrix} -1 & \hat{\mu}_1^{\mathrm{T}} \\ \hat{\mu}_1 & \hat{\Sigma}_{11} \end{bmatrix} & \hat{\beta}_{20\cdot1}^{\mathrm{T}} & \hat{\beta}_{30\cdot12}^{\mathrm{T}} \\ & & \hat{\beta}_{21\cdot1}^{\mathrm{T}} & \hat{\beta}_{31\cdot12}^{\mathrm{T}} \\ & \hat{\beta}_{20\cdot1} \quad \hat{\beta}_{21\cdot1} & \hat{\Sigma}_{22\cdot1} & \hat{\beta}_{32\cdot12}^{\mathrm{T}} \\ & \hat{\beta}_{30\cdot12} \quad \hat{\beta}_{31\cdot12} & \hat{\beta}_{32\cdot12} & \hat{\Sigma}_{33\cdot12} \end{bmatrix} \end{bmatrix},$$

with obvious generalizations to more than three blocks of variables. This equation defines the transformation from $\hat{\phi}$ to $\hat{\theta}$ for this problem.

Estimates of the precision of the ML estimates based on the asymptotic covariance matrix are not as easily accomplished by these sweep 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 calculations described above for the data in Table 7.4, taken from Draper and Smith (1981). The original labeling of the variables is $X_1, \ldots, X_5$. The data have the pattern of Figure 1.1*c* 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

$$\begin{aligned} \hat{\mu}_3 &= 11.769, \quad \hat{\mu}_5 = 95.423, \quad \hat{\sigma}_{33} = 37.870, \\ \hat{\sigma}_{35} &= -47.566, \quad \hat{\sigma}_{55} = 208.905. \end{aligned}$$

Step 2 is based on observations 1–9 and yields regression coefficients from the regression of (X_1, X_2) on $(X_3. X_5)$

$$\begin{aligned} \hat{\beta}_{10\cdot 35} &= 2.802, \quad \hat{\beta}_{13\cdot 35} = -0.526, \quad \hat{\beta}_{15\cdot 35} = 0.105, \\ \hat{\beta}_{20\cdot 35} &= -74.938, \quad \hat{\beta}_{23\cdot 35} = 1.062, \quad \hat{\beta}_{25\cdot 35} = 1.178, \end{aligned}$$

and estimated residual covariance matrix

$$\hat{\Sigma}_{12\cdot 35} = \begin{array}{c} \\ X_1 \\ X_2 \end{array} \begin{array}{c} \begin{array}{cc} X_1 & X_2 \end{array} \\ \begin{bmatrix} 3.804 & -8.011 \\ -8.011 & 24.382 \end{bmatrix} \end{array}.$$

Table 7.4 Data for Example 7.8[a]

	Variables				
Units	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

[a]Values in parentheses are considered missing in the example.
Source: Draper and Smith, 1981.

Step 3 is based on observations 1–6 and yields the following estimated coefficients and residual variance for the regression of X_4 on the other variables:

$$\hat{\beta}_{40\cdot1235} = 85.753, \quad \hat{\beta}_{41\cdot1235} = -1.863, \quad \hat{\beta}_{42\cdot1235} = -1.324,$$
$$\hat{\beta}_{43\cdot1235} = -1.533, \quad \hat{\beta}_{45\cdot1235} = 0.397, \quad \hat{\sigma}_{44\cdot1235} = 0.046.$$

Steps 4–6 yield

$$\begin{bmatrix} -1 & \hat{\mu}^{\mathrm{T}} \\ \hat{\mu} & \hat{\Sigma} \end{bmatrix} = \mathrm{RSW}[1235]\begin{bmatrix} \mathrm{SWP}[12]\begin{bmatrix} \mathrm{SWP}[35]\begin{bmatrix} -1 & 11.769 & 95.423 \\ 11.769 & 37.870 & -47.566 \\ 95.423 & -47.566 & 208.905 \end{bmatrix} & \begin{matrix} 2.802 & -74.938 \\ -0.526 & 1.062 \\ 0.105 & 1.178 \end{matrix} & \begin{matrix} 85.753 \\ -1.533 \\ 0.397 \end{matrix} \\ \begin{matrix} 2.802 & -0.526 & 0.105 \\ -74.938 & 1.062 & 1.178 \\ 85.753 & -1.533 & 0.397 \end{matrix} & \begin{matrix} 3.804 & -8.011 \\ -8.011 & 24.382 \\ -1.863 & -1.324 \end{matrix} & \begin{matrix} -1.863 \\ -1.324 \\ 0.046 \end{matrix} \end{bmatrix} \end{bmatrix}.$$

Calculating the right side and reordering the variables gives ML estimates

$$\begin{matrix} & x_1 & x_2 & x_3 & x_4 & x_5 \end{matrix}$$
$$\hat{\mu}^{\mathrm{T}} = [6.655 \quad 49.965 \quad 11.769 \quad 27.047 \quad 95.423]$$
$$\hat{\Sigma} = \begin{bmatrix} 21.826 & 20.864 & -24.900 & -11.473 & 46.953 \\ 20.864 & 238.012 & -15.817 & -252.072 & 195.604 \\ -24.900 & -15.817 & 37.870 & -9.599 & -47.566 \\ -11.473 & -252.072 & -9.599 & 294.183 & -190.599 \\ 46.953 & 195.604 & -47.556 & -190.599 & 208.905 \end{bmatrix}.$$

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

$$\hat{\phi} = (\hat{\mu}_1, \hat{\Sigma}_{11}, \hat{\beta}_{20\cdot1}, \hat{\beta}_{21\cdot1}, \hat{\Sigma}_{22\cdot1}, \hat{\beta}_{30\cdot12}, \hat{\beta}_{31\cdot12}, \hat{\beta}_{32\cdot12}, \hat{\Sigma}_{33\cdot12}),$$

by D draws from the posterior distribution of ϕ:

$$\phi^{(d)} = (\mu_1^{(d)}, \Sigma_{11}^{(d)}, \beta_{20\cdot1}^{(d)}, \beta_{21\cdot1}^{(d)}, \Sigma_{22\cdot1}^{(d)}, \beta_{30\cdot12}^{(d)}, \beta_{31\cdot12}^{(d)}, \beta_{32\cdot12}^{(d)}, \Sigma_{33\cdot12}^{(d)}),$$

and then applying the sweep operations in steps 4–6 to obtain D draws $\theta^{(d)}$ from the posterior of θ, $d = 1, \ldots, D$. 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.* The bootstrap was applied to the data in Example 7.8, yielding the following means and standard errors over 1000 bootstrap samples:

	x_1	x_2	x_3	x_4	x_5
Bootstrap means =	[7.22	46.75	10.78	31.28	94.15]
Bootstrap s.e.s =	[1.10	3.14	1.35	3.42	2.97]

For comparison, the following are posterior means and standard deviations obtained as draws from the posterior distribution, assuming the Jeffreys' prior (6.35):

	x_1	x_2	x_3	x_4	x_5
Posterior means =	[6.73	49.93	11.66	27.14	95.51]
Posterior s.d.s =	[2.13	6.86	2.49	7.28	5.98]

We 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 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 cases and five variables. Indeed, only about 5% of the bootstrap samples (1000 out of 23,128) were included in the bootstrap calculation, because the other samples did not yield unique ML estimates of parameters. Thus, the simple bootstrap should not be used in such small-sample situations.

7.5 FACTORIZATIONS FOR SPECIAL NONMONOTONE PATTERNS

Nonmonotone patterns of incomplete data where factorizations of the likelihood are possible have been noted by Anderson (1957), where each factor is a complete-data likelihood and the data are normal, and Rubin (1974) more generally. The basic case is given by Figure 7.1, adapted from Rubin (1974). 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 completely missing, and vice versa.

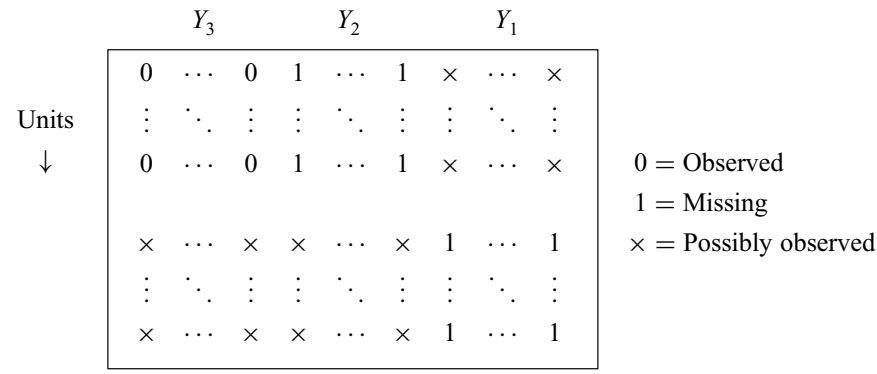


Figure 7.1. Data pattern where Y_3 is more observed than Y_1, and Y_1 and Y_2 are never jointly observed.

3. The rows of Y_1 are conditionally independent given Y_3 with the same set of 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 two terms: the first term, for the marginal distribution of Y_2 and Y_3 with parameter ϕ_{23}, is based on all the units; the second term for the conditional distribution of Y_1 given Y_3 with parameter $\phi_{1\cdot 3}$ is 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, as 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, in that the parameters of conditional association (e.g., partial correlation) between Y_1 and Y_2 given Y_3 are not included. 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 of 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 Y_1, Y_2, and Y_3 univariate, no complete observations, r_1 observations on Y_1 and Y_3, and r_2 observations on Y_2 and Y_3 with $n = r_1 + r_2$. Assuming the data are MAR, the likelihood factorizes into three components: (1) $r_1 + r_2$ observations on the marginal normal distribution of Y_3, with parameters μ_3 and σ_{33}; (2) r_1 observations on the conditional distribution of Y_1 given Y_3, with intercept $\beta_{10\cdot 3}$, slope $\beta_{13\cdot 3}$ and variance

$\sigma_{11\cdot3}$; and (3) r_2 observations on the conditional distribution of Y_2 given Y_3, with intercept $\beta_{20\cdot3}$, slope $\beta_{23\cdot3}$ and variance $\sigma_{22\cdot3}$. 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 missing 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 data.

Data having such a pattern of incompleteness are not uncommon. One context, where each Y_i is multivariate, is the *file-matching* problem, which arises when combining large government data bases. 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). 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 values from matching units. Such problems are discussed in Rubin (1986) and Raessler (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 $X = (X_1, \ldots, X_K)$ represent K standard tests given to all sampled subjects, 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 individual 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.

		Standard Tests			New Tests		
Subsample	Subject	$X_1, \ldots, X_K$			Y_1	Y_2	Y_3
1	1	0	$\cdots$	0	0	1	1
	$\vdots$	$\vdots$		$\vdots$	$\vdots$	$\vdots$	$\vdots$
	r_1	0	$\cdots$	0	0	1	1
2	$r_1 + 1$	0	$\cdots$	0	1	0	1
	$\vdots$	$\vdots$		$\vdots$	$\vdots$	$\vdots$	$\vdots$
	$r_1 + r_2$	0	$\cdots$	0	1	0	1
3	$r_1 + r_2 + 1$	0	$\cdots$	0	1	1	0
	$\vdots$	$\vdots$		$\vdots$	$\vdots$	$\vdots$	$\vdots$
	$r_1 + r_2 + r_3$	0	$\cdots$	0	1	1	0

Figure 7.2. Data structure with three new tests: 0 = score observed, 1 = score missing.

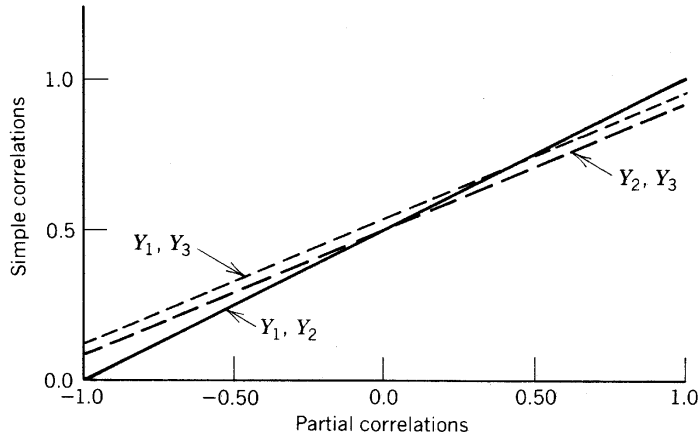


Figure 7.3. Simple correlations as a function of partial correlations. *Source:* Rubin and Thayer, 1978.

The partial correlations among the Y_j's given X are inestimable in the strict sense that there is no information about their values in the data. 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 values.

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 an extremely 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 observations, $\hat{\mu}_x$ and $\hat{\sigma}_{xx}$. This step yielded $\hat{\mu}_x^{\mathrm{T}} = (43.27, 26.79)$ and

$$\hat{\Sigma}_{xx} = \begin{bmatrix} 330.33 & 118.92 \\ 118.92 & 138.13 \end{bmatrix}.$$

STEP 2. Find the ML estimates $\hat{\beta}_{10\cdot x}$, $\hat{\beta}_{1x\cdot x}$ and $\hat{\sigma}_{11\cdot x}$ 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 observations with X and Y_1 both observed. This step yielded $(\hat{\beta}_{10\cdot x}, \hat{\beta}_{1x\cdot x}^{\mathrm{T}}) = (0.9925, 0.1010, 0.1718)$ and $\hat{\sigma}_{11\cdot x} = 11.0887$.

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 observations with X and Y_2 both observed. This step yielded $(\hat{\beta}_{20\cdot x}, \hat{\beta}_{2x\cdot x}^{\mathrm{T}}) = (-0.4444, 0.1760, 0.2278)$ and $\hat{\sigma}_{22\cdot x} = 27.3818$.

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 observations with X and Y_3 both observed. This step yielded $(\hat{\beta}_{30\cdot x}, \hat{\beta}_{3x\cdot x}^{\mathrm{T}}) = (0.3309, 0.2298, 0.5731)$ and $\hat{\sigma}_{33\cdot x} = 71.4943$.

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:

$$\begin{bmatrix} -1 & \hat{\mu}_{(0)}^{\mathrm{T}} \\ \hat{\mu}_{(0} & \hat{\Sigma}_{(0)} \end{bmatrix} = \mathrm{RSW}[x] \begin{bmatrix} \mathrm{SWP[x]} & \begin{bmatrix} -1 & \hat{\mu}_{xx}^{\mathrm{T}} \\ \hat{\mu}_{xx} & \hat{\Sigma}_{xx} \end{bmatrix} & \hat{\beta}_{10\cdot x} & \hat{\beta}_{20\cdot x} & \hat{\beta}_{30\cdot x} \\ & & \hat{\beta}_{1x\cdot x} & \hat{\beta}_{2x\cdot x} & \hat{\beta}_{3x\cdot x} \\ & \hat{\beta}_{10\cdot x} & \hat{\beta}_{1x\cdot x}^{\mathrm{T}} & \hat{\sigma}_{11\cdot x} & 0 & 0 \\ & \hat{\beta}_{20\cdot x} & \hat{\beta}_{2x\cdot x}^{\mathrm{T}} & 0 & \hat{\sigma}_{22\cdot x} & 0 \\ & \hat{\beta}_{30\cdot x} & \hat{\beta}_{3x\cdot x}^{\mathrm{T}} & 0 & 0 & \hat{\sigma}_{33\cdot x} \end{bmatrix}, \quad (7.24)$$

where the zeroes on the left side of (7.24) refer to the estimates being conditional on the zero partial correlations. Step 5 yielded $\hat{\mu}_y^{\mathrm{T}} = (9.96, 13.27, 25.63)$,

$$\hat{\Sigma}_{yy} = \begin{bmatrix} 22.66 & 17.61 & 32.84 \\ 17.61 & 54.31 & 49.61 \\ 32.84 & 49.61 & 165.64 \end{bmatrix},$$

$$\hat{\Sigma}_{xy} = \begin{bmatrix} 53.78 & 85.22 & 144.08 \\ 36.74 & 52.39 & 106.50 \end{bmatrix}.$$

STEP 6. Fix all inestimable partial correlations at 1 and find the corresponding ML estimates

$$\begin{bmatrix} -1 & \hat{\mu}_{(1)}^{\mathrm{T}} \\ \hat{\mu}_{(1)} & \hat{\Sigma}_{(1)} \end{bmatrix}.$$

These estimates are obtained by replacing the lower-right 3×3 submatrix on the right side of (7.23) with

$$\begin{bmatrix} \hat{\sigma}_{11\cdot x} & \sqrt{\hat{\sigma}_{11\cdot x}\hat{\sigma}_{22\cdot x}} & \sqrt{\hat{\sigma}_{11\cdot x}\hat{\sigma}_{33\cdot x}} \\ \sqrt{\hat{\sigma}_{11\cdot x}\hat{\sigma}_{22\cdot x}} & \hat{\sigma}_{22\cdot x} & \sqrt{\hat{\sigma}_{22\cdot x}\hat{\sigma}_{33\cdot x}} \\ \sqrt{\hat{\sigma}_{11\cdot x}\hat{\sigma}_{33\cdot x}} & \sqrt{\hat{\sigma}_{22\cdot x}\hat{\sigma}_{33\cdot x}} & \hat{\sigma}_{33\cdot x} \end{bmatrix}.$$

This step yielded 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 were 0.999, 0.996, and 0.990. The corresponding estimates from step 5 were 0.50, 0.54, 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 inestimable partial correlations, 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 posterior distribution of the inestimable parameters is the same as the prior distribution, which has to be proper in these cases. Details are omitted.

PROBLEMS

7.1. Assume the data in Example 7.1 are MAR. Show that given $(y_{11}, \ldots, y_{n1})$, $\hat{\beta}_{20\cdot 1}$ and $\hat{\beta}_{21\cdot 1}$ are unbiased for $\beta_{20\cdot 1}$ and $\beta_{21\cdot 1}$. Hence show that $\hat{\mu}_2$ is unbiased for μ_2.

7.2. Assume the data in Example 7.1 are MCAR. By first conditioning on $(y_{11}, \ldots, y_{n1})$, 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 complete cases.

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 after Eq. (7.17) in Section 7.3 (For help, see Chapter 2 of Box and Tiao (1973), and the material in Section 6.1.4.)

7.5. For the bivariate normal distribution, express the regression coefficient $\beta_{12\cdot 2}$ 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 MCAR.

7.7. Show that for the set-up 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 $n \to \infty$. (See Section 6.3; for help see Little and Rubin, 1983b.)

7.8. Show that the factorization 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. Using the computer or otherwise, generate a bivariate normal sample of 20 cases with parameters $\mu_1 = \mu_2 = 0$, $\sigma_{11} = \sigma_{12} = 1$, $\sigma_{22} = 2$, and delete values of Y_2 so that $\Pr(y_2 \text{ missing} \mid y_1, y_2)$ equals 0.2 if $y_1 < 0$ and 0.8 if $y_1 \geq 0$.

(a) Construct a test for whether the data are MCAR and carry out 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 cases; and (iii) the t-approximation in (2) of Table 7.2. Summarize the properties of these intervals for this missing-data mechanism.

7.10. Prove that SWP is commutative and conclude that the order in which a set of sweeps is taken is irrelevant algebraically. (However, it can be shown that the order can matter for computational ease and accuracy.)

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 was 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, then 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 $\Pr(X = 1) = 1 - \Pr(x = 0) = \pi$, and
(b) Y given $X = j$ is normal with mean μ_j, variance σ^2 ($j = 1, 0$).
Derive ML estimates of (π, μ_0, μ_1, σ^2) and the *marginal* mean and variance of Y for a complete random sample (x_i, y_i), $i = 1, \ldots, n$ on X and Y.

(ii) Suppose now that X is completely observed, but $n - r$ values of Y are missing, with 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 (π, μ_0, μ_1, σ^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.

CHAPTER 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 ML estimates to be calculated by exploiting factorizations of the likelihood. Furthermore, for some models a factorization exists, but the parameters ϕ_j 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. We write $Y = (Y_{\text{obs}}, Y_{\text{mis}})$ where Y_{obs} represents the observed part of Y and Y_{mis} denotes the missing part. In this chapter we assume for simplicity that the data are MAR and that the objective is to maximize the ignorable likelihood

$$L(\theta|Y_{\text{obs}}) = \int f(Y_{\text{obs}}, Y_{\text{mis}}|\theta)dY_{\text{mis}} \tag{8.1}$$

with respect to θ. Similar considerations apply to the more general case when the data are not MAR, and consequently a factor representing the missing-data 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

$$D_\ell(\theta|Y_{\text{obs}}) \equiv \frac{\partial \ln L(\theta|Y_{\text{obs}})}{\partial\theta} = 0. \tag{8.2}$$

When a closed-form solution of Eq. (8.2) cannot be found, iterative methods can be applied. Let $\theta^{(0)}$ be an initial estimate of θ, for example, an estimate based on the completely observed units. Let $\theta^{(t)}$ be the estimate at the tth iteration. The Newton–Raphson algorithm is defined by the equation

$$\theta^{(t+1)} = \theta^{(t)} + I^{-1}(\theta^{(t)}|Y_{\text{obs}})D_{\ell}(\theta^{(t)}|Y_{\text{obs}}), \tag{8.3}$$

where $I(\theta|Y_{\text{obs}})$ is the observed information, defined as:

$$I(\theta|Y_{\text{obs}}) = -\frac{\partial^2 \ell(\theta|Y_{\text{obs}})}{\partial\theta\partial\theta}.$$

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 Eq. (8.3) is replaced by the expected information:

$$\theta^{(t+1)} = \theta^{(t)} + J^{-1}(\theta^{(t)})D_{\ell}(\theta^{(t)}|Y_{\text{obs}}), \tag{8.4}$$

where the expected information is defined as:

$$J(\theta) = E\big\{I(\theta|Y_{\text{obs}})|\theta\big\} = -\int \frac{\partial^2 \ell(\theta|Y_{\text{obs}})}{\partial\theta\partial\theta} f(Y_{\text{obs}}|\theta)dY_{\text{obs}}.$$

Both 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 useful in practice the methods can require careful algebraic manipulations and efficient programming.

An alternative algorithm (Berndt et al., 1974) exploits the fact that, under the model, the sampling covariance matrix of the score $D_{\ell}(\theta|Y_{\text{obs}})$ is a consistent estimate of the information in the neighborhood of $\hat{\theta}$. The resulting iterative equation is

$$\theta^{(t+1)} = \theta^{(t)} + \lambda_t Q^{-1}(\theta^{(t)})D_{\ell}(\theta^{(t)}|Y_{\text{obs}})$$

where $Q(\theta) = \sum_{i=1}^{n}(\partial\ell_i/\partial\theta)(\partial\ell_i/\partial\theta)^{\mathrm{T}}$; ℓ_i is the loglikelihood of the ith observation; and λ_t is a positive step size designed to ensure convergence to a local maximum. This method avoids the need to calculate second derivatives of the loglikelihood. However, its performance in practice can be erratic because the accuracy of the approximation to the information depends on the validity of the model; thus, we do not recommend this method in general. Other variants of the Newton–Raphson algorithm approximate the derivatives of the loglikelihood numerically, by using first and second differences between successive iterates. These methods also require substantial care in computation.

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 $\ell(\theta|Y_{\text{obs}})$ to ML estimation based on the complete-data loglikelihood $\ell(\theta|Y)$. In many important cases, the EM algorithm is remarkably simple, both conceptually and computationally. Sections 8.2 to 8.4 of this chapter, as well as parts of other chapters, are devoted to the EM algorithm.

There can be two major drawbacks to EM, however: (1) in some cases, with large fractions of missing information, it can be very slow to converge; and (2) 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 problems.

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 the 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 loglikelihood or the actual loglikelihood. 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-stepping 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 more on missing-data applications.

8.2. INTRODUCTION TO THE EM ALGORITHM

The EM algorithm is a very 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, Yang and Lange, 1997).

The EM algorithm formalizes a relatively old *ad hoc* idea for handling missing data, already introduced in Chapter 2: (1) Replace missing values by estimated

values, (2) estimate parameters, (3) re-estimate the missing values assuming the new parameter estimates are correct, (4) re-estimate parameters, and so forth, iterating until convergence. Such methods are EM algorithms for models where the complete-data loglikelihood $\ell(\theta|Y_{\text{obs}}, Y_{\text{mis}}) = \ln \text{L}(\theta|Y_{\text{obs}}, Y_{\text{mis}})$ is linear in Y_{mis}; more generally, missing sufficient statistics rather than individual observations need to be estimated, 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 note 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 develop the theory for the normal model. The term EM was introduced by Dempster, Laird, and Rubin (1977), and this work exposed the full generality of the algorithm by (1) proving general results about its behavior, specifically that each iteration increases the likelihood $\ell(\theta|Y_{\text{obs}})$, and (2) 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 E step (expectation step) and an M step (maximization 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 $\ell(\theta|Y_{\text{obs}})$, and if $\ell(\theta|Y_{\text{obs}})$ is bounded, the sequence $\ell(\theta^{(t)}|Y_{\text{obs}})$ converges to a stationary value of $\ell(\theta|Y_{\text{obs}})$. Quite generally, if the sequence $\theta^{(t)}$ converges, it converges to a local maximum or saddle point of $\ell(\theta|Y_{\text{obs}})$. A disadvantage of EM is that its rate of convergence can be painfully slow when there is a large fraction of missing information: Dempster, Laird and Rubin (1977) show that convergence is linear with rate 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, EM does *not* share with Newton–Raphson or scoring algorithms the property of yielding estimates asymptotically equivalent to ML estimates after a single iteration.

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 $\ell(\theta|Y)$.

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 quotation marks 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 not Y_{mis} but the functions of Y_{mis} 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 $\theta^{(t)}$:

$$Q(\theta|\theta^{(t)}) = \int \ell(\theta|y) f(Y_{\text{mis}}|Y_{\text{obs}}, \theta = \theta^{(t)}) dY_{\text{mis}}.$$

The M step of EM determines $\theta^{(t+1)}$ by maximizing this expected complete-data loglikelihood:

$$Q(\theta^{(t+1)}|\theta^{(t)}) \geqslant Q(\theta|\theta^{(t)}), \qquad \text{for all } \theta.$$

EXAMPLE 8.1. *Univariate Normal Data*. Suppose y_i are iid $N(\mu, \sigma^2)$ where y_i for $i = 1, \ldots, r$ are observed, and y_i for $i = r+1, \ldots, n$ are missing, and assume the missing-data mechanism 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 $y_i, i = 1, \ldots, n$, 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

$$E\left(\sum_{i=1}^{n} y_i | \theta^{(t)}, Y_{\text{obs}}\right) = \sum_{i=1}^{r} y_i + (n-r)\mu^{(t)} \tag{8.5}$$

$$E\left(\sum_{i=1}^{n} y_i^2 | \theta^{(t)}, Y_{\text{obs}}\right) = \sum_{i=1}^{r} y_i^2 + (n-r)[(\mu^{(t)})^2 + (\sigma^{(t)})^2], \tag{8.6}$$

for current estimates $\theta^{(t)} = (\mu^{(t)}, \sigma^{(t)})$ of the parameters. Note that simple substitution of $\mu^{(t)}$ for missing values $y_{r+1}, \ldots, y_n$ would lead to the omission of the term $(n-r)(\sigma^{(t)})^2$ in Eq. (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 sufficient statistics. Thus the M step calculates

$$\mu^{(t+1)} = E\left(\sum_{i=1}^{n} y_i | \theta^{(t)}, Y_{\text{obs}}\right)/n \tag{8.7}$$

$$(\sigma^{(t+1)})^2 = E\left(\sum_{i=1}^{n} y_i^2 | \theta^{(t)}, Y_{\text{obs}}\right)/n - (\mu^{(t+1)})^2. \tag{8.8}$$

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_{obs} assuming MAR. Of course, the EM algorithm is unnecessary for this example since the explicit ML estimates $(\hat{\mu}, \hat{\sigma}^2)$ are available.

EXAMPLE 8.2. *A Multinomial Example*. This example is used in Dempster, Laird, and Rubin (1977) to introduce EM. Suppose that the data vector of observed counts $Y_{\text{obs}} = (38, 34, 125)$ 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 $Y_{\text{obs}} = (y_1, y_2, y_3 + y_4)$. Notice that if Y were observed, the ML estimate of θ would be immediate:

$$\hat{\theta}_{\text{complete}} = \frac{y_2 + y_3}{y_1 + y_2 + y_3}. \tag{8.9}$$

Also note that the loglikelihood $\ell(\theta|Y)$ is linear in Y, so finding the expectation of $\ell(\theta|Y)$ given θ and Y_{obs} involves the same calculation as finding the expectation of Y given θ and Y_{obs}, which in effect fills in estimates of the missing values:

$$\begin{aligned}
E(y_1|\theta, Y_{\text{obs}}) &= 38 \\
E(y_2|\theta, Y_{\text{obs}}) &= 34 \\
E(y_3|\theta, Y_{\text{obs}}) &= 125(\theta/4)/(1/2 + \theta/4) \\
E(y_4|\theta, Y_{\text{obs}}) &= 125(1/2)/(1/2 + \theta/4).
\end{aligned}$$

Thus at the tth iteration, with estimate $\theta^{(t)}$, we have for the E step

$$y_3^{(t)} = 125(\theta^{(t)}/4)/(1/2 + \theta^{(t)}/4), \tag{8.10}$$

and for the M step, from Eq. (8.9) we have

$$\theta^{(t+1)} = (34 + y_3^{(t)})/(72 + y_3^{(t)}). \tag{8.11}$$

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	0.500000000	0.126821498	0.1465
1	0.608247423	0.018574075	0.1346
2	0.624321051	0.002500447	0.1330
3	0.626488879	0.000332619	0.1328
4	0.626777323	0.000044176	0.1328
5	0.626815632	0.000005866	0.1328
6	0.626820719	0.000000779	—
7	0.626821395	0.000000104	—
8	0.626821484	0.000000014	—

Iterating between Eqs. (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 $\theta^{(0)} = 1/2$.

EXAMPLE 8.3. *Bivariate Normal Sample with Missing Data on Both Variables*. A simple but nontrivial application of EM is for a bivariate normal sample with a general pattern of missing data. The first group of units have both Y_1 and Y_2 observed, the second group of units have Y_1 observed but are missing Y_2, and the third group of units have Y_2 observed but are missing Y_1 (Fig. 8.1). We wish to calculate the ML estimates of μ and Σ, the mean and covariance matrix of Y_1 and Y_2.

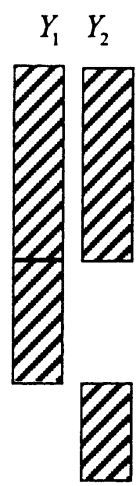


Figure 8.1. The missing-data pattern for Example 8.3.

As in 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:

$$s_1 = \sum_{i=1}^{n} y_{i1}, \quad s_2 = \sum_{i=1}^{n} y_{i2}, \quad s_{11} = \sum_{i=1}^{n} y_{i1}^2, \quad s_{22} = \sum_{i=1}^{n} y_{i2}^2, \quad s_{12} = \sum_{i=1}^{n} y_{i1} y_{i2}, \tag{8.12}$$

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 Eq. (8.12), given Y_{obs} and $\theta = (\mu, \Sigma)$. For the group of units with both y_{i1} and y_{i2} observed, the conditional expectations of the quantities in Eq. (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 y_{i2} on y_{i1}:

$$\begin{aligned}
E(y_{i2}|y_{i1}, \mu, \Sigma) &= \beta_{20\cdot1} + \beta_{21\cdot1} y_{i1} \\
E(y_{i2}^2|y_{i1}, \mu, \Sigma) &= (\beta_{20\cdot1} + \beta_{21\cdot1} y_{i1})^2 + \sigma_{22.1} \\
E(y_{i2}\, y_{i1}|y_{i1}, \mu, \Sigma) &= (\beta_{20\cdot1} + \beta_{21\cdot1} y_{i1}) y_{i1},
\end{aligned}$$

where $\beta_{20\cdot1}$, $\beta_{21\cdot1}$, and $\sigma_{22\cdot1}$ 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 Eq. (8.12) are found as the sums of these quantities over all n units. The M step calculates the usual moment-based estimators 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 for this problem consists of performing these steps iteratively. More details are provided in Chapter 9, where the EM algorithm is presented for the general 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 obligatory constraint that Σ is positive semi-definite. 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 to incomplete data, 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 expected complete-data sufficient statistics *subject to the parameter constraints*. If this step yields explicit estimates, 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, for example, in Examples 11.2–11.4.

8.4. THEORY OF THE EM ALGORITHM

8.4.1. Convergence Properties

The distribution of the complete data Y can be factored as

$$f(Y|\theta) = f(Y_{\text{obs}}, Y_{\text{mis}}|\theta) = f(Y_{\text{obs}}|\theta) f(Y_{\text{mis}}|Y_{\text{obs}}, \theta),$$

where $f(Y_{\text{obs}}|\theta)$ is the density of the observed data Y_{obs} and $f(Y_{\text{mis}}|Y_{\text{obs}}, \theta)$ is the density of the missing data given the observed data. The corresponding decomposition of the loglikelihood is

$$\ell(\theta|Y) = \ell(\theta|Y_{\text{obs}}, Y_{\text{mis}}) = \ell(\theta|Y_{\text{obs}}) + \ln f(Y_{\text{mis}}|Y_{\text{obs}}, \theta).$$

The objective is to estimate θ by maximizing the incomplete-data loglikelihood $\ell(\theta|Y_{\text{obs}})$ with respect to θ for fixed Y_{obs}; this task, however, can be difficult to accomplish directly.

First, write

$$\ell(\theta|Y_{\text{obs}}) = \ell(\theta|Y) - \ln f(Y_{\text{mis}}|Y_{\text{obs}}, \theta), \tag{8.13}$$

where $\ell(\theta|Y_{\text{obs}})$ is the observed loglikelihood to be maximized, $\ell(\theta|Y)$ is the complete-data loglikelihood, which we assume is relatively easy to maximize, and $\ln f(Y_{\text{mis}}|Y_{\text{obs}}, \theta)$ is the missing part of the complete-data loglikelihood.

The expectation of both sides of Eq. (8.13) over the distribution of the missing data Y_{mis}, given the observed data Y_{obs} and a current estimate of θ, say $\theta^{(t)}$, is

$$\ell(\theta|Y_{\text{obs}}) = Q(\theta|\theta^{(t)}) - H(\theta|\theta^{(t)}), \tag{8.14}$$

where

$$Q(\theta|\theta^{(t)}) = \int [\ell(\theta|Y_{\text{obs}}, Y_{\text{mis}})] f(Y_{\text{mis}}|Y_{\text{obs}}, \theta^{(t)}) dY_{\text{mis}} \tag{8.15}$$

and

$$H(\theta|\theta^{(t)}) = \int [\ln f(Y_{\text{mis}}|Y_{\text{obs}}, \theta)] f(Y_{\text{mis}}|Y_{\text{obs}}, \theta^{(t)}) dY_{\text{mis}}. \tag{8.16}$$

Note that

$$H(\theta|\theta^{(t)}) \leqslant H(\theta^{(t)}|\theta^{(t)}), \tag{8.17}$$

by Jensen's inequality (see Rao, 1972, p. 47).

Consider a sequence of iterates $\theta^{(0)}, \theta^{(1)}, \ldots$, where $\theta^{(t+1)} = M(\theta^{(t)})$ for some function $M(\cdot)$. The difference in values of $\ell(\theta|Y_{\text{obs}})$ at successive iterates is given by

$$\begin{aligned}\ell(\theta^{(t+1)}|Y_{\text{obs}}) - \ell(\theta^{(t)}|Y_{\text{obs}}) = [Q(\theta^{(t+1)}|\theta^{(t)}) - Q(\theta^{(t)}|\theta^{(t)})] \\ - [H(\theta^{(t+1)}|\theta^{(t)}) - H(\theta^{(t)}|\theta^{(t)})].\end{aligned} \tag{8.18}$$

An EM algorithm chooses $\theta^{(t+1)}$ to maximize $Q(\theta|\theta^{(t)})$ with respect to θ. More generally, a GEM (generalized EM) algorithm chooses $\theta^{(t+1)}$ so that $Q(\theta^{(t+1)}|\theta^{(t)})$ is greater than $Q(\theta^{(t)}|\theta^{(t)})$. Hence the difference of Q functions in Eq. (8.18) is positive for any EM or GEM algorithm. Furthermore, note that the difference in the H functions in Eq. (8.18) is negative by Eq. (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 result of Dempster, Laird, and Rubin (1977).

Theorem 8.1. Every GEM algorithm increases $\ell(\theta|Y_{\text{obs}})$ at each iteration, that is,

$$\ell(\theta^{(t+1)}|Y_{\text{obs}}) \geqslant \ell(\theta^{(t)}|Y_{\text{obs}}),$$

with equality if and only if

$$Q(\theta^{(t+1)}|\theta^{(t)}) = Q(\theta^{(t)}|\theta^{(t)}).$$

Corollary 1. Suppose that for some θ^* in the parameter space of θ, $\ell(\theta^*|Y_{\text{obs}}) \geqslant \ell(\theta|Y_{\text{obs}})$ for all θ. Then for every GEM algorithm,

$$\begin{aligned}\ell(M(\theta^*)|Y_{\text{obs}}) &= \ell(\theta^*|Y_{\text{obs}}), \\ Q(M(\theta^*)|\theta^*) &= Q(\theta^*|\theta^*),\end{aligned}$$

and

$$f(Y_{\text{mis}}|Y_{\text{obs}}, M(\theta^*)) = f(Y_{\text{mis}}|Y_{\text{obs}}, \theta^*),$$

almost everywhere.

Corollary 2. Suppose that for some θ^* in the parameter space of θ, $\ell(\theta^*|Y_{\text{obs}}) \geqslant \ell(\theta|Y_{\text{obs}})$ for all θ. Then for every GEM algorithm:

$$M(\theta^*) = \theta^*.$$

Theorem 8.1 implies that $\ell(\theta|Y_{\text{obs}})$ is nondecreasing after each iteration of a GEM algorithm, and is strictly increasing after any iteration such that Q increases; that is, whenever $Q(\theta^{(t+1)}|\theta^{(t)}, Y_{\text{obs}}) > Q(\theta^{(t)}|\theta^{(t)}, Y_{\text{obs}})$. 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 $Q(\theta|\theta^{(t)})$ is maximized by setting the first derivative equal to zero.

Theorem 8.2. Suppose a sequence of EM iterates is such that

(a) $D^{10}Q(\theta^{(t+1)}|\theta^{(t)}) = 0$, where D^{10} means the first derivative with respect to the first argument, that is,

$$D^{10}Q(\theta^{(t+1)}|\theta^{(t)}) = \frac{\partial}{\partial\theta}Q(\theta|\theta^{(t)})|_{\theta=\theta^{(t+1)}},$$

(b) $\theta^{(t)}$ converges to θ^*, and

(c) $f(Y_{\text{mis}}|Y_{\text{obs}}, \theta)$ is smooth in θ, where smooth is defined in the proof.

Then

$$D_\ell(\theta^*|Y_{\text{obs}}) = \frac{\partial}{\partial\theta}\ell(\theta|Y_{\text{obs}})|_{\theta=\theta^*} = 0,$$

so that if the $\theta^{(t)}$ converge, they converge to a stationary point.

PROOF.

$$\begin{aligned} D\ell(\theta^{(t+1)}|Y_{\text{obs}}) &= D^{10}Q(\theta^{(t+1)}|\theta^{(t)}) - D^{10}H(\theta^{(t+1)}|\theta^{(t)}) \\ &= -D^{10}H(\theta^{(t+1)}|\theta^{(t)}) \\ &= -\frac{\partial}{\partial\theta}\int[\ln f(Y_{\text{mis}}|Y_{\text{obs}}, \theta)]f(Y_{\text{mis}}|Y_{\text{obs}}, \theta^{(t)})dY_{\text{mis}}|_{\theta=\theta^{(t+1)}}, \end{aligned}$$

which, assuming sufficient smoothness to interchange the order of differentiation and integration, converges to

$$-\int\frac{\partial}{\partial\theta}f(Y_{\text{mis}}|Y_{\text{obs}}, \theta)dY_{\text{mis}}|_{\theta=\theta^{(t+1)}},$$

which equals zero after again interchanging the order of integration and differentiation.

Other EM results in Dempster, Laird, and Rubin (1977) and Wu (1983) regarding convergence include the following:

1. If $\ell(\theta|Y_{\text{obs}})$ is bounded, $\ell(\theta^{(t)}|Y_{\text{obs}})$ converges to some ℓ^*.
2. If $f(Y|\theta)$ is a general (curved) exponential family and $\ell(\theta|Y_{\text{obs}})$ is bounded, then $\ell(\theta^{(t)}|Y_{\text{obs}})$ converges to a stationary value ℓ^*.
3. If $f(Y|\theta)$ is a regular exponential family and $\ell(\theta|Y_{\text{obs}})$ is bounded, then $\theta^{(t)}$ converges to a stationary point θ^*.

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

$$f(Y|\theta) = b(Y)\exp[s(Y)\theta/a(\theta)], \tag{8.19}$$

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) consists in estimating the complete-data sufficient statistics $s(Y)$ by

$$s^{(t+1)} = E(s(Y)|Y_{\text{obs}}, \theta^{(t)}). \tag{8.20}$$

The M step determines the new estimate $\theta^{(t+1)}$ of θ as the solution of the likelihood equations

$$E(s(Y)|\theta) = s^{(t+1)}, \tag{8.21}$$

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 Eq. (8.20).

The following example, described in Dempster, Laird, and Rubin (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 embedding the 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_{obs}, consist of a random sample $X = (x_1, x_2, \ldots, x_n)$ from a Student's t distribution with center μ, scale parameter σ, and known degrees of freedom v, with density

$$f(x_i|\theta) \sim \frac{\Gamma(v/2 + 1/2)}{(\pi v\sigma^2)^{1/2}\Gamma(v/2)\left[1 + (x_i - \mu)^2/(v\sigma^2)\right]^{(v+1)/2}}. \tag{8.22}$$

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 data set $Y = (Y_{\text{obs}}, Y_{\text{mis}})$, where $Y_{\text{obs}} = X$ and $Y_{\text{mis}} = W = (w_1, w_2, \ldots, w_n)$ is a

vector of unobserved positive quantities, such that pairs (w_i, x_i) are independent across units i, with distribution specified by

$$(x_i|\theta, w_i) \sim_{\text{ind}} N(\mu, \sigma^2/w_i), \qquad (w_i|\theta) \sim \chi^2_v/v, \tag{8.23}$$

where χ^2_v denotes the chi-squared distribution with v degrees of freedom. This model leads to the marginal distribution of x_i given by Eq. (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

$$s_0 = \sum_{i=1}^{n} w_i, \qquad s_1 = \sum_{i=1}^{n} w_i x_i, \qquad s_2 = \sum_{i=1}^{n} w_i x_i^2.$$

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 $\chi^2_{v+1}\left(v + (x_i - \mu)^2/\sigma^2\right)^{-1}$. Hence

$$w_i^{(t)} = E(w_i|x_i, \theta^{(t)}) = \frac{v+1}{v + d_i^{(t)2}}, \tag{8.24}$$

where $d_i^{(t)} = (x_i - \mu^{(t)})/\sigma^{(t)}$ is the current estimate of the number of standard deviations of x_i 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:

$$\begin{aligned} \mu^{(t+1)} &= \frac{s_1^{(t)}}{s_0^{(t)}}; \\ \hat{\sigma}^{(t+1)2} &= \frac{1}{n}\sum_{i=1}^{n} w_i^{(t)}(x_i - \hat{\mu}^{(t+1)})^2 = \frac{s_2^{(t)} - s_1^{(t)2}/s_0^{(t)}}{n}, \end{aligned} \tag{8.25}$$

as displayed in general in Example 6.10. The EM algorithm defined by Eqs. (8.24) and (8.25) is iteratively reweighted least squares, with weights $w_i^{(t)}$ that downweight outlying values x_i far from the mean. Thus ML for this t model yields a form of robust estimation. Extensions of this example are given in the following two sections, and in Chapter 12.

8.4.3. Rate of Convergence of EM

Note that differentiating Eq. (8.13) twice with respect to θ yields, for any Y_{mis},

$$I(\theta|Y_{\text{obs}}) = I(\theta|Y_{\text{obs}}, Y_{\text{mis}}) + \partial^2 \ln f(Y_{\text{mis}}|Y_{\text{obs}}, \theta)/\partial\theta\partial\theta,$$

where $I(\theta|Y_{\text{obs}}, Y_{\text{mis}})$ is the observed information based on $Y = (Y_{\text{obs}}, Y_{\text{mis}})$, and the negative of the last term is the missing information from Y_{mis}. Taking expectations over the distribution of Y_{mis} given Y_{obs} and θ yields

$$I(\theta|Y_{\text{obs}}) = -D^{20}Q(\theta|\theta) + D^{20}H(\theta|\theta), \tag{8.26}$$

where Q and H are given by Eqs. (8.15) and (8.16), provided differentials with respect to θ can be passed through the integral signs. If we evaluate the functions in Eq. (8.26) at the converged value θ^* of θ, and call $i_{\text{com}} = -D^{20}Q(\theta|\theta)|_{\theta=\theta^*}$ the complete information, $i_{\text{obs}} = I(\theta|Y)_{\text{obs}})|_{\theta=\theta^*}$ the observed information, and $i_{\text{mis}} = -D^{20}H(\theta|\theta)|_{\theta=\theta^*}$ the missing information, then Eq. (8.26) leads to:

$$i_{\text{obs}} = i_{\text{com}} - i_{\text{mis}}, \tag{8.27}$$

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 Eq. (8.27) and multiplying both sides by i_{com}^{-1} yields:

$$DM = i_{\text{mis}}i_{\text{com}}^{-1} = I - i_{\text{obs}}i_{\text{com}}^{-1}, \tag{8.28}$$

where the matrix DM represents the fraction of missing information. Dempster, Laird and Rubin (1977) show that DM is the gradient of the EM mapping, and it controls the speed of convergence of EM: the greater the fraction of missing information, the slower the rate of convergence. Specifically, they show that in broad generality, for $\theta^{(t)}$ near θ^*,

$$|\theta^{(t+1)} - \theta^*| = \lambda|\theta^{(t)} - \theta^*|, \tag{8.29}$$

where $\lambda = DM$ for scalar θ or the largest eigenvalue of DM, for vector θ. For exceptions to Eq. (8.29), see Meng and Rubin (1994).

Louis (1982) rewrites the missing information in terms of complete-data quantities, and shows that

$$\begin{aligned} -D^{20}H(\theta|\theta) = E\big[&D_\ell(\theta|Y_{\text{obs}}, Y_{\text{mis}})D_\ell^{\text{T}}(\theta|Y_{\text{obs}}, Y_{\text{mis}})|Y_{\text{obs}}, \theta\big] \\ &- D_\ell(\theta|Y_{\text{obs}})D_\ell^{\text{T}}(\theta|Y_{\text{obs}}), \end{aligned}$$

where, as earlier, D_ℓ denotes the score function and T denotes matrix transpose. At the ML estimate $D_\ell(\hat{\theta}|Y_{\text{obs}}) = 0$, so the last term vanishes. Eq. (8.27) becomes

$$I(\hat{\theta}|Y_{\text{obs}}) = -D^{20}Q(\hat{\theta}|\hat{\theta}) - E\big[D_\ell(\theta|Y_{\text{obs}}, Y_{\text{mis}})D_\ell^{\text{T}}(\theta|Y_{\text{obs}}, Y_{\text{mis}})|Y_{\text{obs}}, \theta\big]|_{\theta=\hat{\theta}}, \quad (8.30)$$

which may be useful for computations.

An analogous expression to Eq. (8.27) for the expected information $J(\theta)$ is obtained by taking expectations of Eq. (8.26) over Y_{obs}. Specifically,

$$J(\theta) = J_c(\theta) + E\big[D^{20}H(\theta|\theta)\big], \quad (8.31)$$

where $J_c(\theta)$ is the expected complete information based on $Y = (Y_{\text{obs}}, Y_{\text{mis}})$. 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

$$y_1 \ln(1-\theta) + (y_2 + y_3)\ln\theta + \text{const.},$$

ignoring terms not involving θ. Differentiating with respect to θ yields

$$D_\ell(\theta|Y) = -\frac{y_1}{(1-\theta)} + \frac{(y_2+y_3)}{\theta};$$
$$I(\theta|Y) = \frac{y_1}{(1-\theta)^2} + \frac{(y_2+y_3)}{\theta^2}.$$

Hence

$$E\big[I(\theta|Y)|Y_{\text{obs}}, \theta\big] = \frac{y_1}{(1-\theta)^2} + \frac{y_2 + \hat{y}_3)}{\theta^2}$$
$$E\big[D_\ell^2(\theta|Y)|Y_{\text{obs}}, \theta\big] = \text{Var}\big[D_\ell(\theta|Y)|Y_{\text{obs}}, \theta\big] = \frac{V}{\theta^2},$$

where $\hat{y}_3 = E(y_3|Y_{\text{obs}}, \theta) = (y_3 + y_4)(0.25\theta)(0.25\theta + 0.5)^{-1}$, and $V = \text{Var}(y_3|Y_{\text{obs}}, \theta) = (y_3 + y_4)(0.5)(0.25\theta)(0.25\theta + 0.5)^{-2}$. Substituting $y_1 = 38$, $y_2 = 34$, $y_3 + y_4 = 125$ and $\hat{\theta} = 0.6268$ in these expressions yields

$$E\big[I(\theta|Y)|Y_{\text{obs}}, \theta\big]|_{\theta=\hat{\theta}} = 435.3,$$
$$E\big[D_\ell^2(\theta|Y)|Y_{\text{obs}}, \theta\big]|_{\theta=\hat{\theta}} = 57.8.$$

Hence $I(\hat{\theta}|Y_{\text{obs}}) = 435.3 - 57.8 = 377.5$, 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 shown 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 $\mathrm{Var}(s(Y)|\theta)$, and the missing information is $\mathrm{Var}(s(Y)|Y_{\mathrm{obs}}, \theta)$. Thus the observed information is

$$I(\theta|Y_{\mathrm{obs}}) = \mathrm{Var}(s(Y)|\theta) - \mathrm{Var}(s(Y)|Y_{\mathrm{obs}}, \theta), \tag{8.32}$$

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 (Dempster, Laird and Rubin, 1977). GEM algorithms increase the loglikelihood 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 $s = 1, \ldots, S$. 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 $\theta = (\theta_1, \ldots, \theta_s)$. In many applications, it is useful to take the sth 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 sth 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.

Since 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) establish 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 $\{g_s, s = 1, \ldots, S\}$, 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

$$y_i \sim_{\text{ind}} N_K(X_i\beta, \Sigma), i = 1, \ldots, n, \tag{8.33}$$

where X_i is a known $(K \times p)$ design matrix for the ith 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 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 simply the weighted least squares estimate:

$$\beta^{(t+1)} = \left(\sum_{i=1}^{n} X_i^{\mathrm{T}}(\Sigma^{(t)})^{-1}X_i\right)^{-1}\left(\sum_{i=1}^{n} X_i^{\mathrm{T}}(\Sigma^{(t)})^{-1}Y_i\right). \tag{8.34}$$

On the other hand, given $\beta = \beta^{(t+1)}$, the conditional ML estimate of Σ can be obtained directly from the cross-products of the residuals:

$$\Sigma^{(t+1)} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - X_i\beta^{(t+1)})(Y_i - X_i\beta^{(t+1)})^{\mathrm{T}}. \tag{8.35}$$

The complete-data loglikelihood function is increased by each conditional maximization, (8.34) and (8.35):

$$\text{CM1:} \qquad \ell(\beta^{(t+1)}, \Sigma^{(t)}|Y) \geqslant \ell(\beta^{(t)}, \Sigma^{(t)}|Y),$$

$$\text{CM2:} \qquad \ell(\beta^{(t+1)}, \Sigma^{(t+1)}|Y) \geqslant \ell(\beta^{(t+1)}, \Sigma^{(t)}|Y).$$

With missing data, the ECM algorithm replaces the original M step with the two CM steps given by Eqs. (8.34) and (8.35). More specifically, at the $(t+1)$st iteration of ECM, one first performs the same E step as with EM, that is, finds the conditional expectation of the complete-data sufficient statistics, in this example, $E(Y_i|Y_{\text{obs}}, \theta^{(t)})$ and $E(Y_i Y_i^{\text{T}}|Y_{\text{obs}}, \theta^{(t)})$, where $\theta^{(t)} = (\beta^{(t)}, \Sigma^{(t)})$. Details of this E step are deferred until Section 11.2.1, which discusses the EM algorithm for an incomplete multivariate normal sample in some detail. Then one performs the first CM step, which calculates $\beta^{(t+1)}$ using Eq. (8.34) with Y_i replaced by $E(Y_i|Y_{\text{obs}}, \theta^{(t)})$. Having obtained $\beta^{t+1)}$, one then performs the second CM step, which calculates $\Sigma^{(t+1)}$ using Eqs. (8.35), where Y_i and $Y_i Y_i^{\text{T}}$ on the right side are replaced with $E(Y_i|Y_{\text{obs}}, \theta^{(t)})$ and $E(Y_i Y_i^{\text{T}}|Y_{\text{obs}}, \theta^{(t)})$, respectively. Thus one iteration of ECM for this example consists of one E step and two CM steps, none of which requires numerical iteration. The ECM algorithm in this example can be viewed as a generalization of iteratively reweighted least squares, for example, Rubin, 1983a, in the presence of incomplete data.

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 secondly, 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 $s = 1, \ldots, S$.

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 no three-way association model for a three-way table. A well-known iterative algorithm for fitting these kinds of models is Iterative Proportional Fitting (IPF) (e.g., Bishop, Fienberg and Holland (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} be the count and θ_{ijk} be the 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, $\theta_{ijk}^{(0)} = \frac{1}{8}$ 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 consists of the following three sets of conditional maximizations:

$$\text{CM1:} \qquad \theta_{ijk}^{(t+1/3)} = \theta_{ij(k)}^{(t)}(y_{ij+}^{(t)}/n), \tag{8.36}$$

$$\text{CM2:} \qquad \theta_{ijk}^{(t+2/3)} = \theta_{i(j)k}^{(t+1/3)}(y_{i+k}^{(t)}/n), \tag{8.37}$$

$$\text{CM3:} \qquad \theta_{ijk}^{(t+3/3)} = \theta_{(i)jk}^{(t+2/3)}(y_{+jk}^{(t)}/n), \tag{8.38}$$

where n is the total count. It is easy to see that Eq. (8.36) corresponds to maximizing the loglikelihood $\ell(\theta|Y^{(t)})$ 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 loglikelihood $\ell(\theta|Y^{(t)})$ 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 (a) 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 $\theta_{ij+}, \theta_{i+k}$, and θ_{+jk} are simply the sample proportions, and (b) if $\theta^{(0)} \in \Omega_\theta$, then all $\theta^{(t)} \in \Omega_\theta$, so starting from a table of constant probabilities will yield 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 downweighting 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 Eq. (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:

$$\begin{aligned}\ell(\mu, \sigma^2, v|Y) = &-0.5n\log\sigma^2 - 0.5\sum_{i=1}^{n} w_i(x_i - \mu)^2/\sigma^2 \\ &+ 0.5nv\log(v/2) - n\log\Gamma(v/2) + (v/2 - 1)\sum_{i=1}^{n}\log w_i \\ &- 0.5v\sum_{i=1}^{n} w_i,\end{aligned}$$

so the Q-function is

$$Q(\mu, \sigma^2, \nu|\mu^{(t)}, \sigma^{(t)2}, \nu^{(t)}) = -0.5n \log \sigma^2 - 0.5 \sum_{i=1}^{n} w_i^{(t)}(x_i - \mu)^2/\sigma^2 + 0.5n\nu \log(\nu/2) - n \log \Gamma(\nu/2) + (\nu/2 - 1) \sum_{i=1}^{n} z_i^{(t)} - 0.5\nu \sum_{i=1}^{n} w_i^{(t)}, \quad (8.39)$$

where $w_i^{(t)}$ is given by Eq. (8.24) and $z_i^{(t)} = E(\log w_i|x_i, \mu^{(t)}, \sigma^{(t)2}, \nu^{(t)})$ involves digamma functions. Note that Eq. (8.39) is a complex function of ν, but the maximization is confined to a single scalar parameter, and $\nu^{(t+1)}$ can be found by an iterative one-dimensional search. A drawback is that convergence is considerably slowed by estimation of ν; the ECME algorithm described next provides a way to speed convergence without increasing the complexity of the algorithm.

8.5.2. ECME and AECM Algorithms

The ECME (Expectation/Conditional Maximization Either) 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. Secondly, 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 $\theta^{(t)} \to \theta^{(t+1)}$ 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 degrees of freedom 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, as before 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 various 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. PX-EM Algorithm

Parameter-Expanded EM (PX-EM) (Liu, Rubin and Wu, 1998) speeds EM by imbedding 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 (a) there is no information about α in the observed data Y_{obs}, that is:

$$f_x(Y_{\text{obs}}|\theta^*, \alpha) = f_x(Y_{\text{obs}}|\theta^*, \alpha') \text{ for all } \alpha, \alpha', \tag{8.40}$$

where f_x denotes the density of the expanded model, and (b) the parameters ϕ in the expanded model are identifiable from the complete data $Y = (Y_{\text{obs}}, Y_{\text{mis}})$. The PX-EM algorithm is simply EM applied to the expanded model; that is, for the tth iteration:

PX-E step: Compute $Q(\phi|\phi^{(t)}) = E(\log f_x(Y|\phi)|Y_{\text{obs}}, \phi^{(t)})$
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 $f_x(Y_{\text{obs}}|\theta^*, \alpha)$, which equals $f(Y_{\text{obs}}|\theta)$ when $\alpha = \alpha_0$. Hence each step of PX-EM increases the relevant likelihood $f(Y_{\text{obs}}|\theta)$, and convergence properties of PX-EM parallel that of standard EM.

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 degrees of freedom v by

imbedding 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:

$$(x_i|\mu_*, \sigma_*, \alpha, w_i) \sim_{\text{ind}} N(\mu_*, \sigma_*^2/w_i), \qquad (w_i|\mu_*, \sigma_*, \alpha) \sim_{\text{ind}} \alpha\chi_v^2/v, \tag{8.41}$$

where $\theta^* = (\mu_*, \sigma_*)$ and α is an additional scale parameter. The expanded model (8.41) reduces to the original model (8.23) when $\alpha = 1$. Since the marginal density (8.22) of the observed data X is unchanged and does not involve α, there is no information about α in $Y_{\text{obs}} = X$, but α can be identified 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:

$$w_i^{(t)} = E(w_i|x_i, \phi^{(t)}) = \alpha^{(t)} \frac{v + 1}{v + d_i^{(t)2}}, \tag{8.42}$$

where

$$d_i^{(t)} = \sqrt{\alpha^{(t)}}(x_i - \mu_*^{(t)})/\sigma_*^{(t)} = (x_i - \mu^{(t)})/\sigma^{(t)},$$

as in Eq. (8.24).

The PX-M step maximizes the expected complete-data loglikelihood 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

$$\mu_*^{(t+1)} = \frac{s_1^{(t)}}{s_0^{(t)}}; \qquad \sigma_*^{(t+1)2} = \frac{1}{n}\sum_{i=1}^{n} w_i^{(t)}(x_i - \mu_*^{(t+1)})^2 = \frac{s_2^{(t)} - s_1^{(t)2}/s_0^{(t)}}{n},$$

and

$$\alpha^{(t+1)} = \frac{1}{n}\sum_{i=1}^{n} w_i^{(t+1)}.$$

In the original parameter space, the PX-M step is:

$$\mu^{(t+1)} = \frac{s_1^{(t)}}{s_0^{(t)}}; \qquad \sigma^{(t+1)2} = \sum_{i=1}^{n} w_i^{(t)}(x_i - \mu_*^{(t+1)})^2 / \sum_{i=1}^{n} w_i^{(t)}. \tag{8.43}$$

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, Tyler and Vardi (1994) as a modification of EM to speed convergence.

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

$$I_{\text{obs}} = \begin{pmatrix} i_{\text{obs}} & 0 \\ 0 & 0 \end{pmatrix},$$

and

$$I_{\text{com}} = \begin{pmatrix} i_{\text{com}} & i_{\theta\alpha}^{\text{T}} \\ i_{\theta\alpha} & i_{\alpha\alpha} \end{pmatrix},$$

so the fraction of missing information (8.28) is the (θ, θ) submatrix of $I - I_{\text{obs}} I_{\text{com}}^{-1}$, namely

$$DM(\theta) = I - i_{\text{obs}}(i_{\text{com}} - i_{\theta\alpha}^{\text{T}} i_{\alpha\alpha}^{-1} i_{\theta\alpha})^{-1}.$$

This 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 seminegative definite. Since 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 $i_{\text{com}} - i_{\theta\alpha}^{\text{T}} i_{\alpha\alpha}^{-1} i_{\theta\alpha}$ 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 12.

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 of more EM step 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, since 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:

$$\theta^{(t+1)} = \theta^{(t)} + a^{(t)}\delta^{(t)}, \tag{8.44}$$

where

$$\delta^{(t)} = (-D^{20}Q(\theta;\ \theta^{(t)})|_{\theta=\theta^{(t)}})^{-1}D^{10}Q(\theta;\ \theta^{(t)})|_{\theta=\theta^{(t)}}, \tag{8.45}$$

where $a^{(t)}$ is a constant between 0 and 1 chosen so that

$$Q(\theta^{(t+1)};\ \theta^{(t)}) \geqslant Q(\theta^{(t)};\ \theta^{(t)}), \tag{8.46}$$

that is, so that Eq. (8.44) defines a GEM algorithm. Gradient EM is a special case of Eqs. (8.44) and (8.45) with $a^{(t)} = 1$. In many cases $-D^{20}Q(\theta;\ \theta)$ is positive definite, in which case Eq. (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 Eq. (8.45) replaced with

$$\delta^{(t)} = (-D^{20}Q(\theta;\ \theta^{(t)})|_{\theta=\theta^{(t)}} + B^{(t)})^{-1}D^{10}Q(\theta;\ \theta^{(t)})|_{\theta=\theta^{(t)}}, \tag{8.47}$$

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

$$\begin{aligned} B^{(t)} &= B^{(t+1)} - (v^{(t)}v^{(t)\mathrm{T}})/[v^{(t)\mathrm{T}}(\theta^{(t)} - \theta^{(t-1)})], \\ v^{(t)} &= h^{(t)} + B^{(t-1)}(\theta^{(t)} - \theta^{(t-1)}), \\ h^{(t)} &= D^{10}Q(\theta;\ \theta^{(t)})|_{\theta=\theta^{(t-1)}} - D^{10}Q(\theta;\ \theta^{(t-1)})|_{\theta=\theta^{(t-1)}}. \end{aligned}$$

If $-D^{20}Q(\theta;\ \theta^{(t)})|_{\theta=\theta^{(t)}} + B^{(t)}$ in Eq. (8.47) fails to be 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 Eq. (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 in 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 (AEM) 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 function 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 1 and 2 follow from Theorem 8.1.

8.5. Review results concerning the convergence of EM.

8.6. Show that Eqs. (8.20) and (8.21) are the E and M steps for the regular exponential family (8.19).

8.7. Suppose $Y = (y_1, \ldots, y_n)^{\mathrm{T}}$ are independent gamma random variables with unknown index k and mean $\mu_i = g(\sum_j \beta_j x_{ij})$, where g is a known function, $\beta = (\beta_1, \ldots, \beta_J)$ are unknown regression coefficients, and $x_{i1}, \ldots, x_{iJ}$ are the values of covariates $X_1, \ldots, X_J$ for case 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) $\beta_1, \ldots, \beta_J$ when k is assumed known; and (b) $\beta_1, \ldots, \beta_J$ and k, when k requires estimation.

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 2 of Theorem 8.1 with $\theta^{(t)} = \theta^*$.)

8.10. Write down the loglikelihood of θ for the observed data in Example 8.2. Show directly by differentiating this function that $I(\theta|Y_{\text{obs}}) = 435.3$, as found in Example 8.5.

8.11. Verify the E and M steps in Example 8.4.

8.12. Write down the large sample variance of the ML estimate of θ in Example 8.2, and compare it with the 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 $y_1, \ldots, y_r$ are observed and $y_{r+1}, \ldots, y_n$ 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 Eq. (8.32). Hence find the proportion of missing information from the censoring, and the large sample variances of $\hat{\phi} - \phi$ and $\hat{\theta} - \theta$.

8.14. Write down 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 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.

CHAPTER 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 ML inferences can be based on Approximation 6.1, namely that

$$(\theta - \hat{\theta}) \sim N(0,\ C), \tag{9.1}$$

where C is an estimate of the $d \times d$ covariance matrix of $(\theta - \hat{\theta})$, for example

$$C = I^{-1}(\hat{\theta}|Y_{\text{obs}}), \tag{9.2}$$

the inverse of the observed information evaluated at $\theta = \hat{\theta}$, or

$$C = J^{-1}(\hat{\theta}), \tag{9.3}$$

the inverse of the expected information evaluated at $\theta = \hat{\theta}$, or

$$\hat{C}^* = I^{-1}(\hat{\theta})\hat{K}(\hat{\theta})I^{-1}(\hat{\theta}), \tag{9.4}$$

where

$$\hat{K}(\hat{\theta}) = \frac{\partial \ell(\theta|Y_{\text{obs}})}{\partial \theta}\frac{\partial \ell(\theta|Y_{\text{obs}})^{\mathrm{T}}}{\partial \theta}|_{\theta=\hat{\theta}},$$

the sandwich estimator. The estimate (9.2) is computed as part of the Newton–Raphson algorithm for ML estimation, and Eq. (9.3) is computed as part of the scoring algorithm. When the 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 $I(\hat{\theta}|Y_{\text{obs}})$ in Eq. (9.2) can be found directly by differentiating the loglikelihood $\ell(\theta|Y_{\text{obs}})$ twice with respect to θ. Alternatively, it can be computed as the difference of the complete information and missing information using

$$I(\theta|Y_{\text{obs}}) = -D^{20}Q(\theta|\theta) + D^{20}H(\theta|\theta), \tag{9.5}$$

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 METHODS THAT DO NOT REQUIRE COMPUTING AND INVERTING AN ESTIMATE OF THE OBSERVED INFORMATION MATRIX

9.2.1. Supplemented EM Algorithm

Supplemented EM (SEM) (Meng and Rubin, 1991) is a way to calculate the large-sample covariance matrix associated with $\theta - \hat{\theta}$ using only (1) code for the E and M steps of EM, (2) code for the large-sample complete-data variance–covariance matrix, V_c, and (3) 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. Supplemented EM tends to be more computationally stable than numerically differentiating $\ell(\theta|Y_{\text{obs}})$, since the numerical approximations are applied only to the missing information, using analytical expressions for the complete-data information matrix.

Recall from Chapter 8 that

$$DM = i_{\text{mis}}i_{\text{com}}^{-1} = I - i_{\text{obs}}i_{\text{com}}^{-1} \tag{9.6}$$

where DM is the derivative of the EM mapping, $i_{\text{com}} = -D^{20}Q(\theta|\theta)|_{\theta=\theta^*}$ is the complete information, $i_{\text{obs}} = I(\theta|Y_{\text{obs}})|_{\theta=\theta^*}$ is the observed information, and $i_{\text{mis}} = -D^{20}H(\theta|\theta)|_{\theta=\theta^*}$ 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

$$V_{\text{obs}} = V_{\text{com}}(I - DM)^{-1} \tag{9.7}$$

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

$$V_{\text{obs}} = V_{\text{com}}(I - DM + DM)(I - DM)^{-1} = V_{\text{com}} + \Delta V \tag{9.8}$$

where

$$\Delta V = V_{\text{com}} DM(I - DM)^{-1} \tag{9.9}$$

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 $\theta^{(t+1)}$;

STEP 2. Fix $i = 1$. Calculate

$$\theta^{(t)}(i) = (\hat{\theta}_1, \ldots, \hat{\theta}_{i-1},\ \theta_i^{(t)},\ \hat{\theta}_{i+1}, \ldots, \hat{\theta}_d),$$

which is $\hat{\theta}$ except in the ith 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

$$r_{ij}^{(t)} = \frac{\tilde{\theta}_j^{(t+1)}(i) - \hat{\theta}_j}{\theta_i^{(t)} - \hat{\theta}_i}, \qquad \text{for } j = 1, \ldots, d.$$

STEP 5. Repeat steps 2 to 4 for $i = 2, \ldots, d$.

OUTPUT: $\theta^{(t+1)}$ and $\left\{ r_{ij}^{(t)} : \quad i,\ j = 1, \ldots, d \right\}$.

DM is the limiting matrix $\{r_{ij}\}$ as $t \to \infty$; the element r_{ij} is obtained when the sequence $r_{ij}^{(t^*)},\ r_{ij}^{(t^*+1)}, \ldots$ 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 ith row of DM have been obtained, there is no need to repeat the above steps 2 to 4 for that i in subsequent iterations.

EXAMPLE 9.1. *Standard Errors for Multinomial Example. (Example 8.5 continued).* Consider the multinomial data of Example 8.2, with observed counts $Y_{\text{obs}} = (38,\ 34,\ 125)$ from a multinomial distribution with cell probabilities $(1/2 - \theta/2,\ \theta/4,\ 1/2 + \theta/4)$. In Example 8.2 we applied EM, where the complete data are counts $Y_{\text{com}} = (Y_1,\ Y_2,\ Y_3,\ Y_4)^{\text{T}}$ from a multinomial distribution with parameters $(1/2 - \theta/2,\ \theta/4,\ \theta/4,\ 1/2)$ and $Y_{\text{obs}} = (y_1,\ y_2,\ y_3 + y_4)$. EM yielded $\hat{\theta} = 0.6268$ (see Table 8.1). In this case θ is a scalar, and SEM has a particularly simple form, since the standard error can be obtained directly from the EM

computations. The complete-data estimate of θ is $(y_2 + y_3)/(y_1 + y_2 + y_3)$ and the complete-data variance is

$$V_{\text{com}} = \hat{\theta}(1 - \hat{\theta})/(y_1 + y_2 + \hat{y}_3) = 0.6268(1 - 0.6268)/101.83 = 0.002297,$$

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

$$V_{\text{obs}} = V_{\text{com}}/(1 - DM) = 0.002297/(1 - 0.1328) = 0.00265.$$

The observed information by analytical calculation is $I_{\text{obs}} = 377.5$, as shown in Example 8.5. Inverting this quantity $V_{\text{obs}} = 1/377.5 = 0.00265$, which agrees with the SEM computation. When EM and SEM are applied to logit(θ), which should satisfy the assumptions of asymptotic normality better, we find 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

$$DM = \begin{array}{c} \\ d_1 \\ d_2 \end{array} \begin{array}{c} \begin{array}{cc} d_1 & d_2 \end{array} \\ \begin{pmatrix} 0 & A \\ 0 & DM^* \end{pmatrix} \end{array}, \qquad d_1 + d_2 = d, \tag{9.10}$$

and DM^* can be computed by running Steps 2 to 4 for $i = d_1 + 1, \ldots, d$. Writing

$$V_{\text{com}} = I_{\text{com}}^{-1} = \begin{array}{c} \\ d_1 \\ d_2 \end{array} \begin{array}{c} \begin{array}{cc} d_1 & d_2 \end{array} \\ \begin{pmatrix} G_1 & G_2 \\ G_2^{\mathrm{T}} & G_3 \end{pmatrix} \end{array}, \tag{9.11}$$

the large sample covariance matrix of $\hat{\theta}$ can be computed via the following generalizations of Eqs. (9.8) and (9.9):

$$V_{\text{obs}} = \begin{pmatrix} G_1 & G_2 \\ G_2^{\mathrm{T}} & G_3 + \Delta V^* \end{pmatrix}. \tag{9.12}$$

where

$$\Delta V^* = (G_3 - G_2^{\mathrm{T}} G_1^{-1} G_2) DM^* (I - DM^*)^{-1}. \tag{9.13}$$

EXAMPLE 9.2. *Standard Errors for a Bivariate Normal Sample with Monotone Missing Data.* 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 normal distribution with parameter $(\mu_1,\ \mu_2,\ \sigma_{11},\ \sigma_{22},\ \rho)$, 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. Since 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 $\theta_2{}^*$ 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). Since 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

$$I_{\text{com}}^{-1} = I^{-1}(\theta^*) = \begin{array}{c} \\ \mu_1 \\ \mu_2 \\ \ln\sigma_{11} \\ \ln\sigma_{22} \\ Z_\rho \end{array} \begin{array}{c} \begin{array}{ccccc} \mu_1 & \mu_2 & \ln\sigma_{11} & \ln\sigma_{22} & Z_\rho \end{array} \\ \begin{pmatrix} 4.9741 & -5.0387 & 0 & 0 & 0 \\ -5.0387 & 6.3719 & 0 & 0 & 0 \\ 0 & 0 & 0.1111 & 0.0890 & -0.0497 \\ 0 & 0 & 0.0890 & 0.1111 & -0.0497 \\ 0 & 0 & -0.0497 & -0.0497 & 0.0556 \end{pmatrix} \end{array}. \tag{9.14}$$

Table 9.1 ML Estimates for Data in Table 7.1, and Asymptotic Standard Errors (s.e.)

Parameter	μ_2	$\ln\sigma_{22}$	Z_ρ
ML estimate ($\theta_2^{(65)}$)	49.33	4.74	−1.45
s.e. from Table 7.2	2.73	0.37	0.274
s.e. from SEM	2.73	0.37	0.274

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

$$\begin{array}{c} \\ \mu_1 \\ \ln\sigma_{11} \\ \mu_2 \\ \ln\sigma_{22} \\ Z_\rho \end{array} \begin{array}{c} \begin{array}{ccccc} \mu_1 & \ln\sigma_{11} & \mu_2 & \ln\sigma_{22} & Z_\rho \end{array} \\ \begin{pmatrix} 4.9741 & 0 & -5.0387 & 0 & 0 \\ 0 & 0.1111 & 0 & 0.0890 & -0.0497 \\ -5.0387 & 0 & 6.3719 & 0 & 0 \\ 0 & 0.0890 & 0 & 0.1111 & -0.0497 \\ 0 & -0.0497 & 0 & -0.0497 & 0.0556 \end{pmatrix} \end{array} = \begin{pmatrix} G_1 & G_2 \\ G_2^{\mathrm{T}} & G_3 \end{pmatrix}, \tag{9.15}$$

where G_3 is the (3×3) lower right submatrix of (9.15). Applying formula (9.13), we obtain

$$\Delta V^* = \begin{array}{c} \\ \mu_2 \\ \ln\sigma_{22} \\ Z_\rho \end{array} \begin{array}{c} \begin{array}{ccc} \mu_2 & \ln\sigma_{22} & Z_\rho \end{array} \\ \begin{pmatrix} 1.0858 & 0.1671 & -0.0933 \\ 0.1671 & 0.0286 & -0.0098 \\ -0.0933 & -0.0098 & 0.0194 \end{pmatrix} \end{array}, \tag{9.16}$$

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 expression (9.15). For example, for the standard error of $\mu_2 - \mu_2^*$, we have from expressions (9.14) and (9.15) $(6.3719 + 1.0858)^{\frac{1}{2}} \approx 2.73$, as given in the third row of Table 9.1.

A highly 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 $V_{\mathrm{obs}} = V_{\mathrm{com}} + \Delta V$. 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 $S = \{i:\ i = 1, \ldots, n\}$ of independent (possibly incomplete) observations. Let $S^{(b)}$ be a sample of n observations 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 observation i, where

$$(m_1^{(b)}, \ldots, m_n^{(b)}) \sim MNOM[n;\ (n^{-1},\ n^{-1}, \ldots, n^{-1})],$$

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 observation i is included in the bth 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 $(\hat{\theta}^{(1)}, \ldots, \hat{\theta}^{(B)})$ 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:

$$\hat{\theta}_{\text{boot}} = \frac{1}{B}\sum_{b=1}^{B} \hat{\theta}^{(b)}. \tag{9.17}$$

and the bootstrap estimate of the variance of $\hat{\theta}$ or $\hat{\theta}_{\text{boot}}$ is

$$\hat{V}_{\text{boot}} = \frac{1}{B-1}\sum_{b=1}^{B} (\hat{\theta}^{(b)} - \hat{\theta}_{\text{boot}})^2. \tag{9.18}$$

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

$$I_{\text{norm}}(\theta) = \hat{\theta} \pm z_{1-\alpha/2}\sqrt{\hat{V}_{\text{boot}}}, \tag{9.19}$$

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

$$I_{\text{emp}}(\theta) = (\hat{\theta}^{(b,l)},\ \hat{\theta}^{(b,u)}), \tag{9.20}$$

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 Eq. (9.19) require bootstrap samples of the order of $B = 200$, whereas intervals based on Eq. (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 are not identified. These samples can be omitted in the calculation of Eqs. (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. It may be preferable to modify ML so that these samples can be included, but we are not aware of research that studies such procedures.

An advantage of the bootstrap is that it can provide a valid large-sample estimate of the standard errors of ML estimates even if the model is misspecified; asymptotically its properties parallel that of the sandwich estimator (9.4) of the covariance matrix, which has the same property. This characteristic suggests that the bootstrap standard errors may be less precisely determined when the model is correctly specified, and hence may yield less stable intervals when the sample size is modest. 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 missing-data mechanism is MAR. Although technically true, inferences based on ignorable ML with bootstrap standard errors are misleading if the ML estimate of θ is inconsistent: they yield incorrect P values and confidence intervals (9.19) or (9.20) with incorrect coverage, since they have the "right" width centered at the "wrong" value. Since model assumptions (including MAR) are needed to ensure consistency of the ML estimate, they remain crucial when the bootstrap is used to compute the sampling variance.

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,

Meng, and Pagano (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 loglikelihood. First, find some initial approximation (full rank) to the covariance matrix, for example, based on the complete cases when this is full rank. 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 out from the ML estimate. Then a quadratic response surface is fitted, with dependent variable $\ell(\theta^{(d)}|Y_{\text{obs}})$ 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.

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 precision without inverting an information matrix is to carry out a Bayesian analysis with a uniform prior, and use the posterior variance as the estimate of precision. 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 inference based directly on the posterior distribution without invoking large-sample normal approximations, which is likely to be 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 confidence 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. Are the iterates likely to converge more quickly or more slowly than EM? Further suppose that SEM is applied to the sequence of ECM iterates, assuming they were EM iterates. Would the resulting asymptotic covariance matrix more likely be an over or underestimate, and explain your reasoning. What about the asymmetry of the calculated matrix? (See Van Dyk, Meng and Rubin, 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, Meng and Rubin (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 were EM. Would the resulting asymptotic covariance matrix more likely be an over or underestimate? 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 $\ell(\theta|Y_{\text{obs}})$ twice.

CHAPTER 10

Bayes and Multiple Imputation

10.1. BAYESIAN ITERATIVE SIMULATION METHODS

10.1.1. Data Augmentation

When sample sizes are small, a useful alternative approach to ML is to add a 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 incomplete data through Example 7.3 and Section 7.4.4 in the special case of multivariate normal data with a monotone missing-data pattern.

The posterior distribution for a model with an ignorable missing-data mechanism is:

$$p(\theta|Y_{\text{obs}}, M) \equiv p(\theta|Y_{\text{obs}}) = \text{constant} \times p(\theta) \times f(Y_{\text{obs}}|\theta), \tag{10.1}$$

where $p(\theta)$ is the prior distribution and $f(Y_{\text{obs}}|\theta)$ is the density of the observed data. 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,

$$\text{L}(\phi|Y_{\text{obs}}) = \prod_{q=1}^{Q} \text{L}_q(\phi_q|Y_{\text{obs}}),$$

and, assuming that the parameters $\phi_1, \ldots, \phi_Q$ were also *a priori* independent, the posterior distribution factored in an analogous way, with $\phi_1, \ldots, \phi_Q$ *a posteriori* independent. Consequently, draws $\phi^{(d)} = (\phi_1^{(d)}, \ldots, \phi_Q^{(d)})$ 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. As with ML estimation with a general pattern of missing values, Bayes simulation requires iteration.

Data augmentation (Tanner and Wong, 1987)[1] is an iterative method of simulating the posterior distribution of θ that combines features of the 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 $Y_{\text{mis}}^{(t+1)}$ with density $p(Y_{\text{mis}}|Y_{\text{obs}}, \theta^{(t)})$;
P Step: Draw $\theta^{(t+1)}$ with density $p(\theta|Y_{\text{obs}}, Y_{\text{mis}}^{(t+1)})$.

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 $p(Y_{\text{mis}}|Y_{\text{obs}})$ and $p(\theta|Y_{\text{obs}})$, or the joint posterior distribution $p(\theta, Y_{\text{mis}}|Y_{\text{obs}})$. The iterative procedure can be shown eventually to yield a draw from the joint posterior distribution of Y_{mis}, θ given Y_{obs}, in the sense that as t tends to infinity, this sequence converges to a draw from the joint distribution of (θ, Y_{mis}) given Y_{obs}.

EXAMPLE 10.1. *Bivariate Normal Data with Ignorable Nonresponse and a General Pattern of Missing Data (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 the DA algorithm 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 independent given the parameters, each missing y_{i2} is drawn independently as

$$y_{i2}^{(t+1)} \sim_{\text{ind}} N(\beta_{20.1}^{(t)} + \beta_{21.1}^{(t)} y_{i1}, \sigma_{22.1}^{(t)}),$$

where $\beta_{20.1}^{(t)}$, $\beta_{21.1}^{(t)}$, and $\sigma_{22.1}^{(t)}$ are the tth iterates of the regression parameters of Y_2 on Y_1. Analogously, each missing y_{i1} is drawn independently as:

$$y_{i1}^{(t+1)} \sim_{\text{ind}} N(\beta_{10.2}^{(t)} + \beta_{12.2}^{(t)} y_{i2}, \sigma_{11.2}^{(t)}),$$

where $\beta_{10.2}^{(t)}$, $\beta_{12.2}^{(t)}$, and $\sigma_{11.2}^{(t)}$ are the tth 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

[1] The definition of data augmentation 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.

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_{mis} and a draw from the posterior distribution of θ. Data augmentation can be run independently D times to generate D iid draws from the approximate joint posterior distribution of θ and Y_{mis}. The values of Y_{mis} are multiple imputations of the missing values, drawn from their posterior predictive distribution.

Note that unlike EM, estimates of the covariance matrix from the filled-in data can be computed without adding corrections to the variances. The reason is that draws from the predictive distribution are imputed in the I step of DA, rather than conditional means 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 One-Parameter Multinomial Model (Example 9.1 continued).* Example 9.1 applied EM and 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 proportional to

$$\theta^{y_2 + y_3 + \alpha_1 - 1}(1-\theta)^{y_1 + \alpha_2 - 1},$$

which is Beta. 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 6.20. This P step is analogous to the M step of EM, Eq. (8.11).

Histograms of 90,000 draws from the posterior distribution of θ and logit(θ), for the Jeffreys' prior distribution with $\alpha_1 = \alpha_2 = 0.5$, are displayed in Figure 10.1. Note that the posterior distribution on the logit scale looks more normal, although even on the raw scale the normal approximation is not far off. Table 10.1

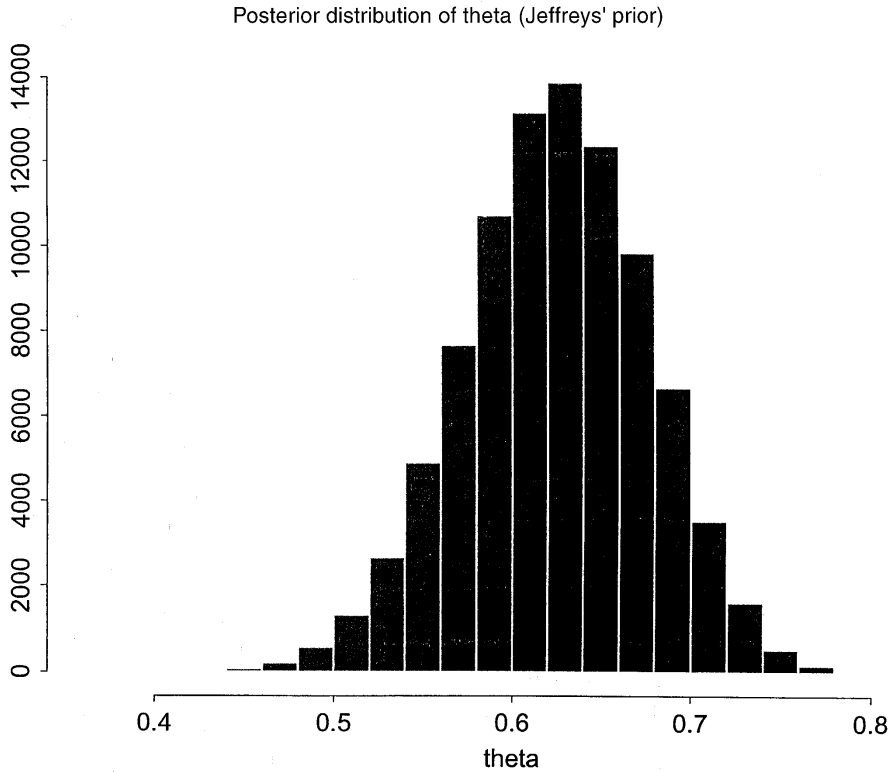


Figure 10.1. Posterior distribution of logit(θ) (Jeffreys' prior).

summarizes the posterior means and variances of θ and logit(θ) from this analysis, and the analysis based on the uniform prior for θ, $\alpha_1 = \alpha_2 = 1$. These are close to the ML estimate and asymptotic standard error from EM/SEM, displayed in the last column of the table.

10.1.2. The Gibbs' Sampler

The Gibbs' sampler is an iterative simulation method that eventually yields a draw from the joint distribution in the case of a general pattern of missing data, and

Table 10.1 Estimates from Bayesian and ML Analyses of Multinomial Example (Example 10.2)

Summary Quantity	Bayes, Jeffreys' Prior	Bayes, Uniform Prior	ML
Post. Mean/MLE of θ	0.624	0.623	0.626
Post. Var/Asympt SE of θ	0.00265	0.00258	0.00265
Post. Mean/MLE of logit θ	0.513	0.508	0.519
Post. Var/Asympt SE of logit θ	0.492	0.478	0.484

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.

The Gibbs' sampler eventually generates a draw from the distribution $P(x_1, \ldots, x_p)$ of a set of p random variables $X_1, \ldots, X_p$, in settings where draws from the joint distribution are hard to compute, but draws from conditional distributions $p(x_j|x_1, \ldots, x_{j-1}, x_{j+1}, \ldots x_p), j = 1, \ldots, p$, are relatively easy to compute. Initial values $x_1^{(0)}, \ldots, x_p^{(0)}$ are chosen in some way. Then given values $x_1^{(t)}, \ldots, x_p^{(t)}$ at iteration t, new values are found by drawing from the following sequence of p conditional distributions:

$$
\begin{aligned}
x_1^{(t+1)} &\sim p(x_1|x_2^{(t)}, x_3^{(t)}, \ldots, x_p^{(t)}) \\
x_2^{(t+1)} &\sim p(x_2|x_1^{(t+1)}, x_3^{(t)}, \ldots, x_p^{(t)}) \\
x_3^{(t+1)} &\sim p(x_3|x_1^{(t+1)}, x_2^{(t+1)}, x_4^{(t)}, \ldots, x_p^{(t)}) \\
&\vdots \\
x_p^{(t+1)} &\sim p(x_p|x_1^{(t+1)}, x_2^{(t+1)}, \ldots, x_{p-1}^{(t+1)}).
\end{aligned}
$$

It can be shown that, under quite general conditions, the sequence of iterates $x^{(t)} = (x_1^{(t)}, \ldots, x_p^{(t)})$ converges to a draw from the joint distribution of $X_1, \ldots, X_p$. In this method, the individual components X_j can be sets of variables, not just scalar variables.

When $p = 2$, the Gibbs' sampler is essentially the same as data augmentation if $X_1 = Y_{\text{mis}}, X_2 = \theta$, and distributions condition on Y_{obs}. Then we can, in the limit, obtain a draw from the joint distribution of $(Y_{\text{mis}}, \theta|Y_{\text{obs}})$ by applying the Gibbs' sampler, where at iteration t for the dth imputed data set:

$$
Y_{\text{mis}}^{(d,t+1)} \sim p(Y_{\text{mis}}|Y_{\text{obs}}, \theta^{(d,t)}); \qquad \theta^{(d,t+1)} \sim p(\theta|Y_{\text{mis}}^{(d,t)}, Y_{\text{obs}}).
$$

As with DA, one run of Gibbs' iterates to a draw from the posterior predictive distribution of Y_{mis} 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_{mis}. The values of Y_{mis} 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 helps 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:

$$
y_i \sim_{\text{ind}} N_K(X_i\beta, \Sigma), \qquad i = 1, \ldots, n, \tag{10.2}
$$

where X_i is a known $(K \times p)$ design matrix for the ith 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_{mis} and two conditional posterior steps (CP1 and CP2) for drawing the values of β and Σ. Let $(Y_{\text{mis}}^{(d,t)}, \beta^{(d,t)}, \Sigma^{(d,t)})$ 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_{mis} given Y_{obs}, $\mu^{(d,t)}$ and $\Sigma^{(d,t)}$ is multivariate normal. Let $y_{\text{mis},i}$ denote the set of missing values in observation i. Then $y_{\text{mis},i}$ given Y_{obs}, $\beta^{(d,t)}$ and $\Sigma^{(d,t)}$ is independent over i, and multivariate normal with mean and residual covariance matrix based on the linear regression of $y_{\text{mis},i}$ on $y_{\text{obs},i}$ and X_i. Draws 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_{obs}, $Y_{\text{mis}}^{(d,t)}$, and $\Sigma^{(d,t)}$ is normal with mean

$$\hat{\beta}^{(d,t+1)} = \left\{ \sum_{i=1}^{n} X_i^{\mathrm{T}} (\Sigma^{(d,t)})^{-1} X_i \right\}^{-1} \left\{ \sum_{i=1}^{n} X_i^{\mathrm{T}} (\Sigma^{(d,t)})^{-1} Y_i^{(d,t)} \right\}, \tag{10.3}$$

where $Y_i^{(d,t)} = (Y_{\text{obs},i}, Y_{\text{mis},i}^{(d,t)})$, and covariance matrix

$$\Sigma_{\beta}^{(d,t+1)} = \left\{ \sum_{i=1}^{n} X_i^{\mathrm{T}} (\Sigma^{(d,t)})^{-1} X_i \right\}^{-1}.$$

Hence $\beta^{(d,t+1)}$ is a random draw from this multivariate normal distribution.

CP2 Step: The conditional distribution of Σ given Y_{obs}, $Y_{\text{mis}}^{(d,t)}$, and $\beta^{(d,t+1)}$ is inverse Wishart with scale matrix given by the sum of squares and cross-products matrix of the residuals:

$$\Sigma^{(t+1)} = n^{-1} \sum_{i=1}^{n} (Y_i^{(d,t)} - X_i \beta^{(d,t+1)})(Y_i^{(d,t)} - X_i \beta^{(d,t+1)})^{\mathrm{T}} \tag{10.4}$$

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 ν, by

embedding the observed data X in a larger data set (X, W) from the expanded complete-data model:

$$(x_i|\mu_*, \sigma_*, \alpha, w_i) \sim_{\text{ind}} N(\mu_*, \sigma_*^2/w_i), \qquad (w_i|\mu_*, \sigma_*, \alpha) \sim_{\text{ind}} \alpha\chi_v^2/v, \tag{10.5}$$

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:

$$w_i^{(t)} \sim_{\text{ind}} \chi_{v+1}^2/(v + d_i^{(t)2}), \tag{10.6}$$

where

$$d_i^{(t)} = \sqrt{\alpha^{(t)}}(x_i - \mu_*^{(t)})/\sigma_*^{(t)} = (x_i - \mu^{(t)})/\sigma^{(t)},$$

as in Eq. (8.42).

The PX-M step maximizes the expected complete-data loglikelihood 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, since 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 data sets. 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 within variation roughly equals 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 $\psi_{d,t}$ $(d = 1, \ldots, D, t = 1, \ldots, T)$, and compute B and $\bar{V}$, the between and within sequence variances:

$$B = \frac{T}{D-1} \sum_{d=1}^{D} (\bar{\psi}_{d\cdot} - \bar{\psi}_{\cdot\cdot})^2,$$

where

$$\bar{\psi}_{d\cdot} = \frac{1}{T} \sum_{t=1}^{T} \psi_{d,t}, \qquad \bar{\psi}_{\cdot\cdot} = \frac{1}{D} \sum_{d=1}^{D} \bar{\psi}_{d\cdot}.$$

$$\bar{V} = \frac{1}{D} \sum_{d=1}^{D} s_d^2,$$

where

$$s_d^2 = \frac{1}{T-1} \sum_{t=1}^{T} (\psi_{d,t} - \bar{\psi}_{d\cdot})^2.$$

We can estimate $\mathrm{Var}(\psi|Y_{\mathrm{obs}})$, the marginal posterior variance of the estimand, by a weighted average of $\bar{V}$ and B, namely

$$\widehat{\mathrm{Var}}^{+}(\psi|Y_{\mathrm{obs}}) = \frac{T-1}{T}\bar{V} + \frac{1}{T}B,$$

which *overestimates* the marginal posterior variance assuming the starting distribution is appropriately overdispersed, 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 $\mathrm{Var}(\psi|Y_{\mathrm{obs}})$ 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 $T \to \infty$, the expectation of $\bar{V}$ approaches $\mathrm{Var}(\psi|Y_{\mathrm{obs}})$. 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 $T \to \infty$. This potential scale reduction is estimated by

$$\sqrt{\hat{R}} = \sqrt{\widehat{\mathrm{Var}}^{+}(\psi|Y_{\mathrm{obs}})/\bar{V}},$$

which declines to 1 as $T \to \infty$. 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 way the condition that $\sqrt{\hat{R}}$ is "near" 1 is implemented depends on the problem at hand; for most examples, values below 1.2 are acceptable, but for an important analysis or data set, 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, take logarithms of all-positive quantities and logits of quantities that lie between 0 and 1). Note that simulation inference from correlated draws is generally less precise than from the same number of independent draws, because of serial correlation within the run. If the simulation efficacy is unacceptably low (in the sense of requiring too much real time on any computer to obtain approximate convergence of posterior inference for quantities of interest), seek ways to alter the algorithm to speed convergence (Gelman et al., 1995, p. 330; Liu and Rubin, 1996, 2002).

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, Liu and Wong, 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 $\theta^*_1, \ldots, \theta^*_{D^*}$ 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 $\theta^*_1, \ldots, \theta^*_{D^*}$ with probability proportional to the "importance" ratios or weights, R_d, then in the limit as $D/D^* \to 0$, 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 $\theta_{d,1}, \theta_{d,2}, \ldots$ forms a Markov chain. Gelman et al. (1995, 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, path sampling, 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 reliably estimate the mean and variance of the posterior distribution, say a few hundred. 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 those cases where inference from the complete-data posterior distribution is based on multivariate normality (or the multivariate t), posterior moments of θ can be reliably estimated from a surprisingly small number, D, of draws of the missing data Y_{mis} (e.g., $D = 2$–10), if the fraction of missing information is not too large. This approach creates D draws of (θ, Y_{mis}) and applies 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_{mis}, namely:

$$p(\theta|Y_{\text{obs}}, Y_{\text{mis}}) \propto p(\theta)L(\theta|Y_{\text{obs}}, Y_{\text{mis}}). \tag{10.7}$$

Equations (10.1) and (10.7) can be related by standard probability theory as:

$$\begin{aligned} p(\theta|Y_{\text{obs}}) &= \int p(\theta, Y_{\text{mis}}|Y_{\text{obs}})dY_{\text{mis}} \\ &= \int p(\theta|Y_{\text{mis}}, Y_{\text{obs}})p(Y_{\text{mis}}|Y_{\text{obs}})dY_{\text{mis}}. \end{aligned} \tag{10.8}$$

Equation (10.8) implies that the posterior distribution of θ, $p(\theta|Y_{\text{obs}})$, can be simulated by first drawing the missing values, $Y_{\text{mis}}^{(d)}$, from their joint posterior distribution, $p(Y_{\text{mis}}|Y_{\text{obs}})$, imputing the drawn values to complete the data set, and then drawing θ from its completed-data posterior distribution, $p(\theta|Y_{\text{obs}}, Y_{\text{mis}}^{(d)})$. When the posterior mean and variance are adequate summaries of the posterior distribution, Eq. (10.8) can be effectively replaced by

$$E(\theta|Y_{\text{obs}}) = E[E(\theta|Y_{\text{mis}}, Y_{\text{obs}})|Y_{\text{obs}}], \tag{10.9}$$

and

$$\text{Var}(\theta|Y_{\text{obs}}) = E[\text{Var}(\theta|Y_{\text{mis}}, Y_{\text{obs}})|Y_{\text{obs}}] + \text{Var}[E(\theta|Y_{\text{mis}}, Y_{\text{obs}})|Y_{\text{obs}}]. \tag{10.10}$$

Multiple imputation effectively approximates the integral (10.8) over the missing values as the average:

$$p(\theta|Y_{\text{obs}}) \approx \frac{1}{D}\sum_{d=1}^{D} p(\theta|Y_{\text{mis}}^{(d)}, Y_{\text{obs}}), \tag{10.11}$$

where $Y_{\text{mis}}^{(d)} \sim p(Y_{\text{mis}}|Y_{\text{obs}})$ 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, for large D, using the simulated values of Y_{mis} as follows:

$$E(\theta|Y_{\text{obs}}) \approx \int \theta \frac{1}{D}\sum_{d=1}^{D} p(\theta|Y_{\text{mis}}^{(d)}, Y_{\text{obs}})d\theta = \bar{\theta}, \tag{10.12}$$

where $\bar{\theta} = \sum_{d=1}^{D} \hat{\theta}_d / D$, and $\hat{\theta}_d = E(\theta | Y_{\text{mis}}^{(d)}, Y_{\text{obs}})$ is the estimate of θ from the dth completed data set, and for scalar θ:

$$\text{Var}(\theta | Y_{\text{obs}}) \approx \frac{1}{D} \sum_{d=1}^{D} V_d + \frac{1}{D-1} \sum_{d=1}^{D} (\hat{\theta}_d - \hat{\theta})^2 = \bar{V} + B, \tag{10.13}$$

say, where V_d is the complete-data posterior variance of θ calculated for the dth data set $(Y_{\text{mis}}^{(d)}, Y_{\text{obs}})$, $\bar{V} = \sum_{d=1}^{D} V_d / D$ is the average of V_d over the MI 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 Eq. (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:

$$\text{Var}(\theta | Y_{\text{obs}}) \approx \bar{V} + (1 + D^{-1}) B. \tag{10.14}$$

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 $(\hat{\theta}_d - \bar{\theta})(\hat{\theta}_d - \bar{\theta})^{\text{T}}$.

A further refinement for small D is to replace the normal reference distribution by a t distribution with degrees of freedom given by

$$v = (D-1) \left(1 + \frac{D}{D+1} \frac{\bar{V}}{B} \right)^2. \tag{10.15}$$

When the completed data sets are based on limited degrees of freedom, say v_{com}, an additional refinement replaces v with:

$$v^* = (v^{-1} + \hat{v}_{\text{obs}}^{-1})^{-1},$$

where

$$\hat{v}_{\text{obs}} = (1 - \hat{\gamma}_D) \left(\frac{v_{\text{com}} + 1}{v_{\text{com}} + 3} \right) v_{\text{com}}. \tag{10.16}$$

The theoretical basis for Eq. (10.15) is given in Rubin and Schenker (1986), and for Eq. (10.16) is given in Barnard and Rubin (1999).

EXAMPLE 10.5. *Bivariate Normal Data with Ignorable Nonresponse and a General Pattern of Missing Data (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_{mis} can be quite 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 cases with Y_1 or Y_2 missing are limited. In that case, the draws of Y_{mis} yield five completed data sets, the dth with sample means, variances, and covariance that we denote $\{(\bar{y}_1^{(d)}, \bar{y}_2^{(d)}, s_{11}^{(d)}, s_{22}^{(d)}, s_{12}^{(d)}), d = 1, \ldots, 5\}$. 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)

$$\text{Var}(\mu_1) = (1/5) \sum_{d=1}^{5} (s_{11}^{(d)}/n) + (6/5)(1/4) \sum_{d=1}^{5} (\bar{y}_1^{(d)} - \tilde{\mu}_1)^2.$$

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

In additional to interval estimation, it is often of interest to summarize the posterior distribution for a multi-component 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 (1987a, Section 3.4). Meng and Rubin (1992) developed methods for likelihood ratio testing when the available information consists of point estimates and 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. (1991) 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. 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, $\hat{\theta}_d, d = 1, \ldots, D$, and variance–covariance matrices, $V_d, d = 1, \ldots, D$, using multivariate analogs of the previous expressions. First, let θ_0 be the null value of θ, and let

$$W(\theta_0, \bar{\theta}) = \frac{(\theta_0 - \bar{\theta})^{\text{T}} \bar{V}^{-1} (\theta_0 - \bar{\theta})}{(1 + r)k}, \tag{10.17}$$

where $r = (1 + D^{-1})\ \text{trace}(B\bar{V}^{-1})/k$, and $\text{trace}(B\bar{V}^{-1})/k$ is the average diagonal element of $B\bar{V}^{-1}$. Equation (10.17) is an estimated Wald statistic, as defined in Section 6.1.3. The P value is then

$$\Pr[F_{k,\ell} > W(\theta_0, \bar{\theta})], \tag{10.18}$$

where $F_{k,\ell}$ is an F random variable with k and ℓ degrees of freedom with

$$\ell = 4 + (k(D-1) - 4)\left(1 + \frac{a}{r}\right)^2, \qquad a = \left\{1 - \frac{2}{k(D-1)}\right\}; \tag{10.19}$$

if $k(D-1) \leqslant 4$, let $\ell = (k+1)v/2$. Rubin (1987a) and Li, Raghunathan, and Rubin (1991) provide motivation for this test statistic and its reference distribution.

With large data sets and large models, such as occur often with multiway contingency-table data, 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 $\theta = \theta_0$ than 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,

$$\text{P value} = \Pr(\chi^2_k > \text{LR}), \tag{10.20}$$

where $\text{LR} = \text{LR}[(\hat{\theta}, \hat{\phi}), (\theta_0, \hat{\phi}_0)]$, using the notation of Section 6.1.3, and χ^2_k 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 by $\bar{\theta}$, $\bar{\phi}$, $\bar{\phi}_0$, and $\overline{\text{LR}}$. Assume that the function LR can be evaluated for each of the D completed data sets at $\bar{\theta}, \bar{\phi}, \theta_0, \bar{\phi}_0$ to obtain D values of $\text{LR}[(\bar{\theta}, \bar{\phi}), (\theta_0, \bar{\phi}_0)]$ whose average across the D imputations is $\overline{\text{LR}}_0$. Then

$$\overline{\text{LR}}_0 \Big/ \left[k + \frac{(D+1)(\overline{\text{LR}} - \overline{\text{LR}}_0)}{(D-1)}\right] \tag{10.21}$$

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 cases, 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 multiple imputations result in D values $\text{LR}_1, \ldots, \text{LR}_D$. If so, the following

procedure due to Li et al. (1991) can be used. Let the repeated-imputation P value be $\Pr(F_{k,b} > \widetilde{\mathrm{LR}})$, where

$$\widetilde{\mathrm{LR}} = \frac{(\overline{\mathrm{LR}}/k) - (1 - D^{-1})\mathrm{v}}{1 + (1 + D^{-1})\mathrm{v}}, \tag{10.22}$$

and v is the sample variance of $(\mathrm{LR}_1^{1/2}, \ldots, \mathrm{LR}_D^{1/2})$, and

$$b = k^{-3/D}(D-1)\left\{1 + [(1 + D^{-1})\mathrm{v}]^{-1}\right\}^2. \tag{10.23}$$

10.2.3. Other Methods for Creating Multiple Imputations

We now return to the problem of creating the multiple imputations. The theory of the previous section suggests that we draw the missing values as

$$Y_{\mathrm{mis}}^{(d)} \sim p(Y_{\mathrm{mis}} | Y_{\mathrm{obs}}), \tag{10.24}$$

that is, from their joint posterior predictive distribution. Unfortunately, it is often difficult to draw from this predictive distribution in complicated problems, because of the implicit requirement in Eq. (10.24) 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 when 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. (10.24), 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 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 list some other alternatives:

1. **Improper MI.** An approximate method is to draw:

$$Y_{\mathrm{mis}}^{(d)} \sim p(Y_{\mathrm{mis}} | Y_{\mathrm{obs}}, \tilde{\theta}), \tag{10.25}$$

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 cases. 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, since 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 suggests discarding values to create a data set $Y_{\text{obs-mp}}$ with a monotone pattern, and then drawing θ from its posterior distribution given $Y_{\text{obs-mp}}$. Then draw $Y_{\text{mis}}^{(d)}$ as follows:

$$Y_{\text{mis}}^{(d)} \sim p(Y_{\text{mis}}|Y_{\text{obs}}, \tilde{\theta}^{(d)}), \tag{10.26}$$

where

$$\tilde{\theta}^{(d)} \sim p(\theta|Y_{\text{obs-mp}}).$$

An even simpler but less accurate example of this approach is to draw θ from its posterior distribution given the complete cases, that is,

$$Y_{\text{mis}}^{(d)} \sim p(Y_{\text{mis}}|Y_{\text{obs}}, \tilde{\theta}^{(d)}), \tag{10.27}$$

where

$$\tilde{\theta}^{(d)} \sim p(\theta|Y_{\text{obs-cc}}),$$

where $Y_{\text{obs}-\text{cc}}$ represents data from the complete cases. For the multivariate normal problem with missing values, Eq. (10.27) 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-case missing value restrictions, as discussed in Little (1993b).

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:

$$Y_{\text{mis}}^{(d)} \sim p(Y_{\text{mis}}|Y_{\text{obs}}, \tilde{\theta}^{(d)}),$$

where

$$\tilde{\theta}^{(d)} \sim p(\theta|Y_{\text{aug-mp}}),$$

where $Y_{\text{aug-mp}}$ is the observed data augmented to create a monotone pattern. This method could be combined with method 2 in various ways.

4. **Use the asymptotic distribution of the ML estimate.** Suppose the ML estimate $\hat{\theta}$ of θ is available, together with a consistent 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:

$$Y_{\text{mis}}^{(d)} \sim p(Y_{\text{mis}}|Y_{\text{obs}}, \tilde{\theta}^{(d)}),$$

where

$$\tilde{\theta}^{(d)} \sim N[\hat{\theta}, C(\hat{\theta})].$$

Draw 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, since it correctly propagates asymptotic uncertainty in the ML estimate of θ.

5. **Refining approximate draws using importance sampling.** Methods 2 to 4 draw pairs $(Y_{\text{mis}}^{(d)}, \tilde{\theta}^{(d)})$ from a joint distribution where the draw of $Y_{\text{mis}}^{(d)}$ 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 (e.g., 100–1000) of draws $Y_{\text{mis}}^{(d)}$, and then subsampling a smaller number (for example 2–10) from this set, with probability of selection of draw d proportional to $w_d \propto p(\tilde{\theta}^{(d)})\text{L}(\tilde{\theta}^{(d)}|Y_{\text{obs}})/g(\tilde{\theta}^{(d)})$. This is a version of sampling importance resampling (see Section 10.1.4). 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 $Y_{\text{obs}}^{(\text{boot},d)}$ of the cases (complete and incomplete), that is, a random sample with replacement of the same size as the observed sample. In symbols,

$$Y_{\text{mis}}^{(d)} \sim p(Y_{\text{mis}}|Y_{\text{obs}}, \tilde{\theta}^{(d)}),$$

where

$$\tilde{\theta}^{(d)} = \hat{\theta}(Y_{\text{obs}}^{(\text{boot},d)}).$$

This procedure is proper in 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, since 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. **Drawing from pragmatic conditional distributions.** With real multivariate data, it is often possible to formulate a set of conditional distributions relating each variable to a set of the other variables, which are reasonable when taken one at a time, but 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 as if they were coherent. For each of the variables, a draw of parameters and then missing data is made, the missing data are imputed for that variable, and the procedure cycles through the variables, replacing variables that are being conditioned in any regression by the observed or currently imputed values. A number of practical implementations of this idea include Kennickell (1991); MICE (Van Buuren and Oudshoorn, 1999); and IVEWARE (Raghunathan et al., 2001). Rubin (2002) proposes limiting the possibly incoherent draws to the creation of a monotone pattern, that is, to the creation of Yang-mp.

10.2.4. Use of 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, often a method can be chosen for creating multiple imputations without consideration of the precise model to be used for the analysis of the multiply-imputed data. When the model 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 potentially complicated complete-data analyses to the multiply-imputed data, and then use the combining rules and combined results even though the multiple imputations were created under a different model.

Somewhat surprisingly, this approach can be very successful, especially with relatively limited fractions of missing information, as suggested by theoretical results and as documented by empirical examples. A simple example can be used to illustrate 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 MAR (i.e., MCAR within adjustment cells), 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 to 4.4 in Rubin, 1987a).

When the imputation method uses more information than the complete-data analysis and this information is correct, the complete-data analyses will tend to be more efficient than anticipated: for instance, confidence intervals will have greater than the nominal convergence. This phenomenon was noted in Rubin and Schenker (1987) and Fay (1992, 1996), and termed "superefficiency" 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., anticonservative) inferences. The following example conveys some intuition; other examples are discussed by Meng (2002) and Robins and Wang (2002).

EXAMPLE 10.7. *Effects of a Misspecified Imputation Model.* Suppose we have a sample of values of (X, Y) where X is fully observed but Y is half missing due to a MAR process. In truth, Y is a monotone but nonlinear function of X, $Y = \exp(X)$. 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 on multiple imputation. First, the between-imputation variability (e.g., of the slope of the linear model for Y on X) will be greater than if the true model were being 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 being used. Thus, both between- and within-variability are exaggerated relative to when the correct model is being applied to create the imputations. 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, and thus, overcoverage of 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 extreme quartiles, this approximate validity may not hold.

In our 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 distributions in Figure 10.1, and compare the posterior mean and variance 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 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^2_{NR} 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

$$\bar{y}_* = \frac{(r\bar{y}_R + m\bar{y}_{NR})}{n}$$

and

$$s^2_* = \frac{[(r-1)s^2_R + (m-1)s^2_{NR} + rm(\bar{y}_R - \bar{y}_{NR})^2/n]}{(n-1)}.$$

10.3. Suppose in Problem 10.2, 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 $ms^2_R(1-r^{-1})/n^2$, and that the expectation of s^2_* is $s^2_R(1-r^{-1})[1+rn^{-1}(n-1)^{-1}]$.

(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.

10.4. Suppose multiple imputations are created using the method of Problem 10.3 D times, and let $\bar{y}^{(d)}_*$ and $U^{(d)}_*$ be the values of $\bar{y}_*$ and U_* for the dth 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 $\bar{\bar{y}}_*$. That is,

$$T_* = \bar{U}_* + (1+D^{-1})B_*,$$

where

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

(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 $D^{-1}\mathrm{Var}(\bar{Y}_*) + (1-D^{-1})\mathrm{Var}(\bar{y}_R)$, 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 N/r.

(d) Show that the 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 $s_R r^{-1/2} z_d$ to the imputed value for observation i, where the z_d are independent standard normal deviates.)

10.6. Consider the situation where the complete-data analysis is nonparametric, and produces no estimates but just a P value for a null hypothesis, for example, the P value for a Wilcoxon test in a randomized two-treatment experiment. Suppose that the missing data in this experiment have been multiply imputed with $D = 2$, and the two P values are p_1 and p_2. Let z_1 and z_2 be such that $\Pr(z < z_1) = p_1$, $\Pr(z < z_2) = p_2$, where z is standard normal, and $a = 3(z_1 - z_2)^2/4$. Show that a multiple-imputation combined P value can be found from treating

$$\sqrt{\frac{z_1 z_2}{1 + a}}$$

as a t random variable with $(1 + a^{-1})$ degrees of freedom. (Hint: consider Eq. (10.22) in this setting.)

PART III

Likelihood-Based Approaches to the Analysis of Incomplete Data: Some Examples

CHAPTER 11

Multivariate Normal Examples, Ignoring the Missing-Data 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 ANOVA and multivariate regression; repeated measures models, including random coefficient regression models where the coefficients themselves are regarded as missing data; and selected time series models. Robust estimation with missing data is discussed in Chapter 12, the analysis of categorical data with partially observed or missing 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 nonignorable missing data.

11.2. INFERENCE FOR A MEAN VECTOR AND COVARIANCE MATRIX WITH MISSING DATA UNDER NORMALITY

11.2.1. The EM Algorithm for Incomplete Multivariate Normal Samples

Many multivariate statistical analyses, including multiple linear regression, principal component 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 the efficient estimation of these quantities for an arbitrary pattern of missing values is a particularly important problem. In this section we discuss ML estimation of the mean and covariance matrix from an incomplete multivariate normal sample, assuming the missing-data mechanism is ignorable.

Although the assumption of multivariate normality may appear restrictive, the methods discussed here can provide consistent estimates under weaker assumptions about the underlying distribution. Furthermore, the normality will be relaxed somewhat when we consider linear regression in Section 11.4 and robust estimation in Chapter 12.

Suppose that $(Y_1, Y_2, \ldots, Y_K)$ have a K-variate normal distribution with mean $\mu = (\mu_1, \ldots, \mu_K)$ and covariance matrix $\Sigma = (\sigma_{jk})$. We write $Y = (Y_{\text{obs}}, Y_{\text{mis}})$, where Y represents a random sample of size n on $(Y_1, \ldots, Y_K)$, Y_{obs} the set of observed values, and Y_{mis} the missing data. Also, let

$$Y_{\text{obs}} = (y_{\text{obs},1}, y_{\text{obs},2}, \ldots, y_{\text{obs},n}),$$

where $y_{\text{obs},i}$ represents the set of variables observed for case $i, i = 1, \ldots, n$. The loglikelihood based on the observed data is then:

$$\ell(\mu, \Sigma | Y_{\text{obs}}) = \text{const} - \frac{1}{2}\sum_{i=1}^{n} \ln |\Sigma_{\text{obs},i}| - \frac{1}{2}\sum_{i=1}^{n} (y_{\text{obs},i} - \mu_{\text{obs},i})^{\mathrm{T}} \Sigma_{\text{obs},i}^{-1} (y_{\text{obs},i} - \mu_{\text{obs},i}), \tag{11.1}$$

where $\mu_{\text{obs},i}$ and $\Sigma_{\text{obs},i}$ are the mean and covariance matrix of the observed components of Y for observation i.

To derive the EM algorithm for maximizing Eq. (11.1), we note that the hypothetical complete data Y belong to the regular exponential family (8.19) with sufficient statistics

$$S = \left(\sum_{i=1}^{n} y_{ij},\ j = 1, \ldots, K; \qquad \sum_{i=1}^{n} y_{ij} y_{ik},\ j, k = 1, \ldots, K \right).$$

At the tth iteration of EM, let $\theta^{(t)} = \left(\mu^{(t)}, \Sigma^{(t)}\right)$ denote current estimates of the parameters. The E step of the algorithm consists in calculating

$$E\left(\sum_{i=1}^{n} y_{ij} | Y_{\text{obs}}, \theta^{(t)} \right) = \sum_{i=1}^{n} y_{ij}^{(t)}, \qquad j = 1, \ldots, K, \tag{11.2}$$

and

$$E\left(\sum_{i=1}^{n} y_{ij} y_{ik} | Y_{\text{obs}}, \theta^{(t)} \right) = \sum_{i=1}^{n} (y_{ij}^{(t)} y_{ik}^{(t)} + c_{jki}^{(t)}), \qquad j, k = 1, \ldots, K \tag{11.3}$$

where

$$y_{ij}^{(t)} = \begin{cases} y_{ij}, & \text{if } y_{ij} \text{ is observed,} \\ E(y_{ij} | y_{\text{obs},i}, \theta^{(t)}), & \text{if } y_{ij} \text{ is missing,} \end{cases} \tag{11.4}$$

and

$$c_{jki}^{(t)} = \begin{cases} 0 & \text{if } y_{ij} \text{ or } y_{ik} \text{ is observed,} \\ \text{Cov}(y_{ij}, y_{ik} | y_{\text{obs},i}, \theta^{(t)}, & \text{if } y_{ij} \text{ and } y_{ik} \text{ are missing.} \end{cases} \tag{11.5}$$

Missing values y_{ij} are thus replaced by the conditional mean of y_{ij} given the set of values $y_{\text{obs},i}$ observed for that case. 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 $y_{\text{obs},i}$ are predictors in the regression equation and the remaining variables are outcome variables. The sweep operator is described in Section 7.4.3.

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,

$$\begin{aligned} \mu_j^{(t+1)} &= n^{-1} \sum_{i=1}^{n} y_{ij}^{(t)}, \qquad j = 1, \ldots, K; \\ \sigma_{jk}^{(t+1)} &= n^{-1} E\left(\sum_{i=1}^{n} y_{ij} y_{ik} | Y_{\text{obs}} \right) - \mu_j^{(t+1)} \mu_k^{(t+1)} \\ &= n^{-1} \sum_{i=1}^{n} \left[(y_{ij}^{(t)} - \mu_j^{(t+1)})(y_{ik}^{(t)} - \mu_k^{(t+1)}) + c_{jki}^{(t)} \right], \qquad j, k = 1, \ldots, K. \end{aligned} \tag{11.6}$$

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: (1) to use the complete-case solution of Section 3.2; (2) to use one of the available-case solutions of Section 3.4; (3) to form the sample mean and covariance matrix of the data filled in by one of the imputation methods of Chapter 4; or (4) to form means and variances from observed values of each variable and set all correlations equal to zero. Option 1 provides consistent estimates of the parameters if the data are MCAR and there are at least $K + 1$ complete observations. Option 2 makes use of all the available data but can yield an estimated covariance matrix that is not positive definite, leading to problems in the first iteration. Options 3 and 4 generally yield inconsistent estimates of the covariance matrix, but estimates that are positive semidefinite and hence 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.

The link between ML estimation and an efficient form of imputation for the missing values is clear from the EM algorithm. The E step imputes the best linear predictors of the missing values, using current estimates of the parameters. It also calculates the adjustments c_{jki} to the estimated covariance matrix needed to allow for imputation of the missing values.

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 or obtained for 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 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 the EM algorithm 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, Laird, and Rubin, 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 otherwise.

11.2.2. Estimated Asymptotic Covariance Matrix of $(\theta - \hat{\theta})$

Let $\theta = (\mu, \Sigma)$, where Σ is represented as a row vector $(\sigma_{11}, \sigma_{12}, \sigma_{22}, \ldots, \sigma_{KK})$. 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

$$\psi_{jki} = \begin{cases} (j,k)\text{th element of } \Sigma_{\text{obs},i}^{-1}, & \text{if both } x_{ij} \text{ and } x_{ik} \text{ present,} \\ 0, & \text{otherwise,} \end{cases}$$

and $\Sigma_{\text{obs},i}$ is the covariance matrix of the variables present in observation i. The (lm, rs)th element of $J(\Sigma)$, corresponding to row σ_{lm}, column σ_{rs}, is

$$\frac{1}{4}(2 - \delta_{lm})(2 - \delta_{rs}) \sum_{i=1}^{n} (\psi_{lri}\psi_{msi} + \psi_{lsi}\psi_{mri}),$$

where $\delta_{lm} = 1$ if $l = m$, 0 if $l \neq m$. 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 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 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 are valid when the data are 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 compute an information matrix, so if either 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 is to compute the ML estimates on bootstrap samples, and apply the methods of Section 9.2.2.

11.2.3. Bayes Inference for the Normal Model via Data Augmentation

We now describe a Bayesian analysis of the multivariate normal model in Section 11.2.1. 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 $\theta = (\mu, \Sigma)$:

$$p(\mu, \Sigma | Y_{\text{obs}}) \propto |\Sigma|^{-(K+1)/2} \text{L}(\mu, \Sigma | Y_{\text{obs}}),$$

where $\text{L}(\mu, \Sigma | Y_{\text{obs}})$ is the exponent of the loglikelihood in Eq. (11.1). Let $\theta^{(t)} = (\mu^{(t)}, \Sigma^{(t)})$ and $Y^{(t)} = (Y_{\text{obs}}, Y_{\text{mis}}^{(t)})$ denote current draws of the parameters and filled-in data matrix at iteration t. The I step of the DA algorithm simulates

$$Y_{\text{mis}}^{(t+1)} \sim p(Y_{\text{mis}} | Y_{\text{obs}}, \theta^{(t)}).$$

Since the rows of the data matrix Y are conditionally independent given θ, this is equivalent to drawing

$$y_{\text{mis},i}^{(t+1)} \sim p(y_{\text{mis},i} | y_{\text{obs},i}, \theta^{(t)}) \tag{11.7}$$

independently for $i = 1, \ldots, n$. As noted in the discussion of EM, this distribution is multivariate normal with mean given by the linear regression of $y_{\text{mis},i}$ on $y_{\text{obs},i}$,

evaluated at current draws $\theta^{(t)}$ of the parameters. The regression parameters and residual covariance matrix of this normal distribution is obtained computationally by sweeping on the augmented covariance matrix

$$\Sigma^{*(t)} = \begin{pmatrix} -1 & \mu^{(t)\mathrm{T}} \\ \mu^{(t)} & \Sigma^{(t)} \end{pmatrix},$$

so that the observed variables are swept in and the missing variables are swept out. The draw $y_{\mathrm{mis},i}^{(t+1)}$ 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 covariance matrix $\Sigma_{\mathrm{mis},i\cdot\mathrm{obs},i}^{(t)}$.

The P step of DA draws

$$\theta^{(t+1)} \sim p(\theta|Y^{(t+1)}),$$

where $Y^{(t+1)} = (Y_{\mathrm{obs}}, Y_{\mathrm{mis}}^{(t+1)})$ is the filled-in data from the I step (11.7). The draw of $\theta^{(t+1)}$ can be carried out in two steps:

$$\begin{aligned} (\Sigma^{(t+1)}|Y^{(t+1)}) &\sim \mathrm{Inv-Wishart}(S^{(t+1)}, n-1) \\ (\mu^{(t+1)}|\Sigma^{(t+1)}, Y^{(t+1)}) &\sim N_K(\bar{y}^{(t+1)}, \Sigma^{(t+1)}/n), \end{aligned} \tag{11.8}$$

where $(\bar{y}^{(t+1)}, S^{(t+1)})$ is the sample mean and covariance matrix of Y from the filled-in data $Y^{(t+1)}$. For more computational details on the P step, see Example 6.21.

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. An alternative analysis is to create a smaller number of draws of the missing data based on Eq. (11.7), and then derive inferences using the MI formulas given in Section 10.2. The following example compares the results of these approaches with the ML/bootstrap analysis of Example 11.1.

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.

Table 11.1 St. Louis Risk Research Data for Example 11.1

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	—	1	—	—	—	—	—	—	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	—												

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 cth 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.

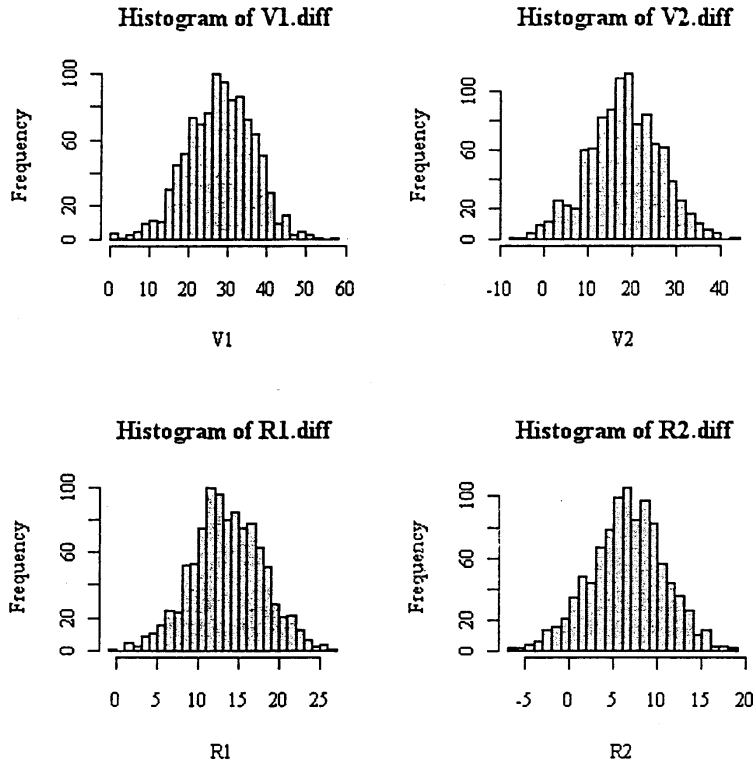
Table 11.2 displays estimates for the four continuous outcomes in the low-risk group and the combined moderate- and high-risk groups. The columns show estimates of the mean, standard error of the mean (s.e.m.), and the standard deviation from four methods: available-case (AC) analysis, ML with s.e.m. computed using

Table 11.2 Means and SDs of Continuous Outcomes in Low- and Moderate/High-Risk Groups, St. Louis Risk Research Data. Estimates from Available-Case Analysis (AC), Maximum Likelihood (ML), Data Augmentation (DA), and Multiple Imputation (MI), Under Normal Model. (Example 11.1.)

	Low Risk ($G = 1$)			Moderate/High Risk ($G = 2, 3$)		
	Mean	s.e.m.	SD	Mean	s.e.m.	SD
Variable: V_1						
(1) AC	146.12	4.76	19.65	105.46	6.56	30.76
(2) ML	143.37	5.51	19.53	115.68	5.89	31.83
(3) DA	143.72	5.40	22.70	115.60	6.31	34.30
(4) MI	143.77	5.42	22.50	115.59	6.31	34.38
Variable: V_2						
(5) AC	128.56	5.40	25.90	106.59	5.37	28.93
(6) ML	128.60	5.10	25.67	110.79	5.09	27.80
(7) DA	128.48	6.01	28.60	110.55	5.24	29.98
(8) MI	128.48	5.98	28.62	110.79	5.22	29.98
Variable: R_1						
(9) AC	117.88	2.27	9.35	102.74	3.18	17.72
(10) ML	116.83	2.88	9.96	103.36	3.29	18.14
(11) DA	116.80	2.82	12.18	103.41	3.39	19.51
(12) MI	116.82	2.79	12.21	103.33	3.33	19.52
Variable: R_2						
(13) AC	110.68	3.23	13.72	101.63	2.53	14.98
(14) ML	108.11	2.96	13.81	101.85	2.49	14.62
(15) DA	108.45	3.39	15.37	101.82	2.70	15.72
(16) MI	108.40	3.37	15.40	101.84	2.67	15.71

the bootstrap, data augmentation (DA) with estimates and standard errors based on 1000 draws from the posterior distribution, and multiple imputation (MI) based on 10 MIs and the formulas 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 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 based on the 2.5th to 97.5th percentiles are shown below the plots. The fact that three of these



95% Posterior Probability Intervals

V1	**(11.50, 44.26)**
V2	**(1.75, 33.37)**
R1	**(4.81, 22.28)**
R2	**(-2.24, 14.84)**

Figure 11.1. Posterior distributions of mean differences $\mu_{\text{low}} - \mu_{\text{med/high}}$, St. Louis Risk Research Data, based on 9000 draws. (Example 11.1.)

four intervals are entirely positive is evidence that reading and verbal means are higher in the low-risk group than in the moderate/high-risk group.

11.3. ESTIMATION 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. Some statistical models, however, place restrictions on θ. ML estimation with incomplete data from

such restricted models can be handled easily by EM, whenever the complete-data analysis is simple. The reason is that the E step of EM takes the same form whether θ is restricted or not; the only alteration in EM with a restriction on θ is to modify the M step to be ML for the restricted model. Similarly for Bayes inference via DA: the I step for the restricted model is the same as the I step for the unrestricted model, and the P step is replaced by a draw of the parameters for the restricted model, given the observed and current imputed data.

For some kinds of restrictions on θ, noniterative ML 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 observed variables.

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 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, consider the 3×3 stationary covariance pattern Σ_{11} and the 4×4 circular symmetry pattern Σ:

$$\Sigma_{11} = \begin{bmatrix} \theta_1 & \theta_2 & \theta_3 \\ \theta_2 & \theta_1 & \theta_2 \\ \theta_3 & \theta_2 & \theta_1 \end{bmatrix}, \qquad \Sigma = \begin{bmatrix} \theta_1 & \theta_2 & \theta_3 & \theta_2 \\ \theta_2 & \theta_1 & \theta_2 & \theta_3 \\ \theta_3 & \theta_2 & \theta_1 & \theta_2 \\ \theta_2 & \theta_3 & \theta_2 & \theta_1 \end{bmatrix} = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}.$$

Suppose that we have a random sample $y_1, \ldots, y_n$ from a multivariate normal distribution, $N_3(0, \Sigma_{11})$. These observations can be viewed as the first three of four components from a random sample $(y_1, z_1), \ldots, (y_n, z_n)$ from a multivariate normal distribution $N_4(0, \Sigma)$, where 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 $C = \sum (y_i, z_i)^{\mathrm{T}}(y_i, z_i)/n$ and $C_{11} = \sum (y_i^{\mathrm{T}} y_i)/n$. The matrix C is the complete-data sufficient statistic and C_{11} is the observed sufficient statistic. The ML estimate of Σ given the complete data, C, is explicit and obtained by simple averaging (Szatrowski, 1978), whence the M step of EM at iteration t is given by

$$\theta_1^{(t+1)} = \frac{1}{4}\left(\sum_{k=1}^{4} c_{kk}^{(t)}\right), \qquad \theta_2^{(t+1)} = \frac{1}{4}(c_{12}^{(t)} + c_{23}^{(t)} + c_{34}^{(t)} + c_{14}^{(t)}),$$
$$\theta_3^{(t+1)} = \frac{1}{2}(c_{13}^{(t)} + c_{24}^{(t)}), \tag{11.9}$$

where $c_{kj}^{(t)}$ is the (k,j)th element of $C^{(t)}$, the expected value of C from the E step at iteration t. These estimates of θ_1, θ_2, and θ_3 yield a new estimate of Σ for iteration $t+1$.

Since there is only one pattern of incomplete data (y_i observed and z_i missing), 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

$$\begin{bmatrix} \Sigma_{11}^{(t)-1} & \Sigma_{11}^{(t)-1}\Sigma_{12}^{(t)} \\ \Sigma_{21}^{(t)}\Sigma_{11}^{(t)-1} & \Sigma_{22}^{(t)} - \Sigma_{21}^{(t)}\Sigma_{11}^{(t)-1}\Sigma_{12}^{(t)} \end{bmatrix} = \mathrm{SWP}[1,2,3]\begin{bmatrix} \Sigma_{11}^{(t)} & \Sigma_{12}^{(t)} \\ \Sigma_{21}^{(t)} & \Sigma_{22}^{(t)} \end{bmatrix}.$$

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 $C_{12} = \sum_i y_i^{\mathrm{T}} z_i / n$ given C_{11} and $\Sigma^{(t)}$ is $C_{11}\Sigma_{11}^{(t)-1}\Sigma_{12}^{(t)}$. The expected value of $z_i^{\mathrm{T}} z_i$ given the observed data and $\Sigma = \Sigma^{(t)}$ is

$$(\Sigma_{21}^{(t)}\Sigma_{11}^{(t)-1}y_i^{\mathrm{T}})(y_i\Sigma_{11}^{(t)-1}\Sigma_{12}^{(t)}) + \Sigma_{22}^{(t)} - \Sigma_{21}^{(t)}\Sigma_{11}^{(t)-1}\Sigma_{12}^{(t)},$$

so that the expected value of $C_{22} = \sum_1^n \dfrac{z_i^{\mathrm{T}} z_i}{n}$ 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:

$$E(C|C_{11}, \Sigma^{(t)}) = \begin{bmatrix} C_{11} & C_{11}\Sigma_{11}^{(t)-1}\Sigma_{12}^{(t)} \\ \Sigma_{21}^{(t)}\Sigma_{11}^{(t)-1}C_{11} & \left\{\Sigma_{22}^{(t)} - \Sigma_{21}^{(t)}(\Sigma_{11}^{(t)-1} - \Sigma_{11}^{(t)-1}C_{11}\Sigma_{11}^{(t)-1})\Sigma_{12}^{(t)}\right\} \end{bmatrix}. \tag{11.10}$$

This new value of C is used in (11.9) 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 expression (11.10).

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 ith 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 β $(K\times q)$ is commonly called the factor-loading matrix, I_q is the $(q\times q)$ identity matrix, $\Sigma=\text{Diag}(\sigma_1^2,\ldots,\sigma_K^2)$ is called the uniqueness matrix, and $\theta=(\mu,\beta,\Sigma)$. Integrating out the unobserved factors z_i yields the exploratory factor analysis model:

$$(y_i|\theta)\sim_{\text{ind}} N_K(\mu,\beta\beta^{\text{T}}+\Sigma).$$

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, Rubin and Wu (1998) show to be an example of a 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 Eq. (11.11). Under model (11.12), $(y_i|\phi)\sim_{\text{ind}} N_K(\mu^*,\beta^*\Gamma\beta^{*\text{T}}+\Sigma^*)$, 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), if $\{(y_i,z_i),i=1,\ldots,n\}$ were fully observed, are

$$\bar{y}=\sum_{i=1}^n y_i/n,\qquad C_{yy}=\sum_{i=1}^n(y_i-\bar{y})(y_i-\bar{y})^{\text{T}}/n,\qquad C_{yz}=\sum_{i=1}^n(y_i-\bar{y})z_i^{\text{T}}/n,$$
$$C_{zz}=\sum_{i=1}^n z_i z_i^{\text{T}}/n.$$

Given current parameter estimates $\phi^{(t)}$, the E step of PX-EM consists of computing the expected complete-data sufficient statistics:

$$\begin{aligned}C_{yz}^{(t+1)} &= E(C_{yz}|Y_{\text{obs}},\phi^{(t)})=C_{yy}\gamma^{(t)},\\ C_{zz}^{(t+1)} &= E(C_{zz}|Y_{\text{obs}},\phi^{(t)})=\gamma^{(t)\text{T}}C_{yy}\gamma^{(t)}+C_{zz\cdot y}^{(t)},\end{aligned}$$

where $\gamma^{(t)}$ and $C^{(t)}_{zz\cdot y}$ are the regression coefficients and residual covariance matrix of Z on Y given $\phi^{(t)}$. Specifically, let

$$B^{(t)} = \begin{pmatrix} \beta^{*(t)\mathrm{T}}\beta^{*(t)} + \Sigma^{*(t)} & \beta^{*(t)\mathrm{T}} \\ \beta^{*(t)} & I_q \end{pmatrix}$$

be the current variance–covariance matrix of (Y, Z); then $\gamma^{(t)}$ and $C^{(t)}_{zz\cdot y}$ are obtained from the last q columns of SWP$[1, \ldots, K]B^{(t)}$.

The M step of PX-EM calculates the cross-products matrix

$$C^{(t+1)} = \begin{pmatrix} C_{yy} & C^{(t+1)}_{yz} \\ C^{(t+1)\mathrm{T}}_{yz} & C^{(t+1)}_{zz} \end{pmatrix}.$$

It then sets $\mu^{*(t+1)} = \bar{y}, \Gamma^{(t+1)} = C^{(t+1)}_{zz}$, and $\beta^{*(t+1)}$ and $\Sigma^{*(t+1)}$ from the last q columns of SWP$[1, \ldots, K]C^{(t+1)}$. Reduction to the original parameters θ gives $\mu^{(t+1)} = \mu^{*(t+1)}$, $\Sigma^{(t+1)} = \Sigma^{*(t+1)}$, and $\beta^{(t+1)} = \beta^{*(t+1)}$ Chol$(\Gamma^{(t+1)})$.

This EM algorithm for factor analysis can be extended to handle missing data Y_{mis} in the Y variables, by treating both Y_{mis} 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 analysis of variance models. The EM algorithm can be used to obtain ML estimates of variance components and more generally covariance components (Dempster, Laird, and Rubin, 1977; Dempster, Rubin, and Tsutakawa, 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 produce conceptions, 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

Table 11.3 Data for Example 11.4

Bull(i)	Percentages of Conception to Services for Successive Samples	n_i
1	46,31,37,62,30	5
2	70,59	2
3	52,44,57,40,67,64,70	7
4	47,21,70,46,14	5
5	42,64,50,69,77,81,87	7
6	35,68,59,38,57,76,57,29,60	9
Total		35

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 population. Thus, with the actual data, there is one component of variability due to sampling bulls from a population of bulls, which are of primary interest, and another due to sampling within each bull.

A common normal model for such data is

$$y_{ij} = \alpha_i + e_{ij}, \tag{11.13}$$

where $(\alpha_i|\theta) \sim_{\text{ind}} N(\mu, \sigma_\alpha^2)$ are the between-bull effects, $(e_{ij}|\theta) \sim_{\text{ind}} N(0, \sigma_e^2)$ 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 $\alpha_1, \ldots, \alpha_6$ 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:

$$\prod_{i,j} (2\pi\sigma_e^2)^{-1/2} \exp\left[-\frac{(y_{ij} - \alpha_i)^2}{(2\sigma_e^2)} \right] \prod_i (2\pi\sigma_\alpha^2)^{-1/2} \exp\left[-\frac{(\alpha_i - \mu)^2}{(2\sigma_\alpha^2)} \right].$$

The loglikelihood is linear in the following complete-data sufficient statistics:

$$\begin{aligned} T_1 &= \sum \alpha_i, \\ T_2 &= \sum \alpha_i^2, \\ T_3 &= \sum_{i,j} (y_{ij} - \alpha_i)^2 = \sum_{i,j} (y_{ij} - \bar{y}_i)^2 + \sum_i n_i (\bar{y}_i - \alpha_i)^2. \end{aligned}$$

The ML estimates based on complete data are:

$$\begin{aligned} \hat{\mu} &= \frac{T_1}{K}, \\ \hat{\sigma}_\alpha^2 &= \frac{T_2}{K} - \hat{\mu}^2, \\ \hat{\sigma}_e^2 &= \frac{T_3}{\sum_i n_i}. \end{aligned} \tag{11.14}$$

These equations define the M step of EM. The E step of EM is defined by taking the expectations of T_1, T_2, T_3 given current estimates of θ and the observed data $y_{ij}, i = 1, \ldots, K, j = 1, \ldots, n_i$. These follow by applying Bayes 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}:

$$\left(\alpha_i | \{y_{ij}\}, \theta\right) \overset{\text{ind}}{\sim} N(w_i\mu + (1 - w_i)\bar{y}_i, v_i),$$

where $w_i = \sigma_\alpha^{-2} v_i$ and $v_i = (\sigma_\alpha^{-2} + n_i\sigma_e^{-2})^{-1}$. Hence

$$\begin{aligned} T_1^{(t+1)} &= \sum [w_i^{(t)}\mu^{(t)} + (1 - w_i^{(t)})\bar{y}_i], \\ T_2^{(t+1)} &= \sum [w_i^{(t)}\mu^{(t)} + (1 - w_i^{(t)})\bar{y}_i]^2 + \sum v_i^{(t)}, \\ T_3^{(t+1)} &= \sum_{i,j} (y_{ij} - \bar{y}_i)^2 + \sum_i n_i \left[w_i^{(t)2} (\mu^{(t)} - \bar{y}_i)^2 + v_i^{(t)} \right]. \end{aligned} \tag{11.15}$$

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 analysis of variance (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 (e.g., see Dempster, Rubin, and Tsutakawa, 1981; 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 Missing Values Confined to the Dependent Variable

Suppose a scalar outcome variable Y is regressed on p predictor variables $X_1, \ldots, X_p$ and missing values are confined to Y. If the missing-data mechanism is ignorable, the incomplete observations do not contain information about the regression parameters, $\theta_{Y \cdot X} = (\beta_{Y \cdot X}, \sigma^2_{Y \cdot X})$. 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. In some cases 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 $(X_1, \ldots, X_p)$ is chosen to simplify the computation of least squares estimates. When Y given $(X_1, \ldots, X_p)$ is normal, least squares computations yield ML estimates. When values of Y, say $y_i, i = 1, \ldots, m$, are missing, the remaining complete observations no longer have the balance 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 given the current estimated parameters $\theta^{(t)}_{Y\cdot X} = (\beta^{(t)}_{Y\cdot X}, \sigma^{(t)2}_{Y\cdot X})$:

$$E(y_i|X, Y_{\text{obs}}, \theta^{(t)}_{Y\cdot X}) = \begin{cases} y_i, & \text{if } y_i \text{ is observed } (i = m+1, \ldots, n), \\ \beta^{(t)}_{Y\cdot X} x_i^{\mathrm{T}}, & \text{if } y_i \text{ is missing } (i = 1, \ldots, m). \end{cases}$$

$$E(y_i^2|X, Y_{\text{obs}}, \theta^{(t)}_{Y\cdot X}) = \begin{cases} y_i^2, & \text{if } y_i \text{ is observed,} \\ (\beta^{(t)}_{Y\cdot X} x_i^{\mathrm{T}})^2 + \sigma^{(t)2}_{Y\cdot X}), & \text{if } y_i \text{ is missing.} \end{cases}$$

where X is the $(n \times p)$ matrix of X values. Let Y be the $(n \times 1)$ vector of Y values, and $Y^{(t)}$ the vector Y with missing components y_i replaced by estimates from the E step at iteration t. The M step calculates

$$\beta^{(t+1)}_{Y\cdot X} = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}Y^{(t)}, \tag{11.16}$$

$$\sigma^{(t+1)^2}_{Y\cdot X} = n^{-1}\left[\sum_{i=m+1}^{n} (y_i - \beta^{(t)}_{Y\cdot X} x_i)^2 + m\sigma^{(t)^2}_{Y\cdot X}\right]. \tag{11.17}$$

The algorithm can be simplified by noting that Eq. (11.16) does not involve $\sigma^{(t)2}_{Y\cdot X}$, and that at convergence we have

$$\sigma^{(t+1)2}_{Y\cdot X} = \sigma^{(t)2}_{Y\cdot X} = \hat{\sigma}^2_{Y\cdot X},$$

so from Eq. (11.17)

$$\hat{\sigma}^2_{Y\cdot X} = \frac{1}{n}\sum_{m+1}^{n} (y_i - \hat{\beta}_{Y\cdot X} x_i)^2 + \frac{m}{n}\hat{\sigma}^2_{Y\cdot X},$$

or

$$\hat{\sigma}^2_{Y\cdot X} = \frac{1}{n-m}\sum_{m+1}^{n} (y_i - \hat{\beta}_{Y\cdot X} x_i)^2. \tag{11.18}$$

Consequently, the EM iterations can omit the M step estimation of $\sigma^2_{Y\cdot X}$ and the E step estimation of $E(y_i^2|\text{Data}, \theta^{(t)}_{Y\cdot X})$, and find $\hat{\beta}_{Y\cdot X}$ by iteration. After convergence, we can calculate $\hat{\sigma}^2_{Y\cdot X}$ directly from Eq. (11.18). These iterations, which fill in the missing data, reestimate 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 $\sigma^2_{Y\cdot X}$, obtained by replacing $n - m$ in Eq. (11.18) by $n - m - p$.

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 $(Y, X_1, \ldots, X_p)$. Then, applying Property 6.1, ML estimates for the regression of Y on $X_1, \ldots, X_p$ are standard functions of the ML estimates for multivariate normal data discussed in Section 11.2. Let

$$\theta = \begin{bmatrix} -1 & \mu_1 & \cdots & \mu_{p+1} \\ \mu_1 & \sigma_{11} & \cdots & \sigma_{1,p+1} \\ \vdots & \vdots & & \vdots \\ \mu_{p+1} & \sigma_{1,p+1} & & \sigma_{p+1,p+1} \end{bmatrix} \tag{11.19}$$

denote the augmented covariance matrix corresponding to the variables $X_1, \ldots, X_p$ and $X_{p+1} \equiv Y$. The intercept, slopes, and residual variance for the regression of Y on $X_1, \ldots, X_p$ are found in the last column of the matrix SWP$[1, \ldots, p]\theta$, 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 SWP$[1, \ldots, p]\hat{\theta}$.

Let $\hat{\beta}_{Y \cdot X}$ and $\hat{\sigma}_{Y \cdot X}$ 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 conditions more general than the multivariate normality of Y and $(X_1, \ldots, X_p)$. Specifically, suppose we partition $(X_1, \ldots, X_p)$ as $(X_{(1)}, X_{(0)})$ where the variables in $X_{(1)}$ are more observed than both Y and the variables in $X_{(0)}$, in the sense of Section 7.5 that any unit with any observation on Y or $X_{(0)}$ has all variables in $X_{(1)}$ observed. A particularly simple case occurs when $X = (X_1, \ldots, X_p)$ is fully observed so that $X_{(1)} = X$: see Figure 7.1 for the general case, where Y_1 corresponds to $(Y, X_{(0)})$, Y_3 corresponds to $X_{(1)}$ and Y_2 is null. Then if the conditional distribution of $(Y, X_{(0)})$ given $X_{(1)}$ is multivariate normal, $\hat{\mu}_{Y \cdot X}$ and $\hat{\sigma}_{YY \cdot X}$ are ML. See Chapter 7 for details.

This assumption is much less stringent than multivariate normality for $X_1, \ldots, X_{p+1}$ since it allows the predictors in $X_{(1)}$ 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. Methods for regression with a binary outcome, or incomplete categorical predictors, are considered in Chapter 14.

Unfortunately the $(p \times p)$ submatrix of the first p rows and columns of SWP$[1, \ldots, p]\hat{\theta}$ does not provide the 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. An approximate method (Beale and Little, 1975; Little, 1979) is to let

$$\text{Var}(\hat{\beta}_{Y\cdot X}) = S_W^{-1}\hat{\sigma}^2_{Y\cdot X},$$

where S_W is a $(p \times p)$ weighted sum of squares and cross-products matrix with (j, k)th element

$$S_{W_{jk}} = \sum_{i=1}^{n} w_i(\hat{x}_{ij} - \tilde{x}_j)(\hat{x}_{ik} - \tilde{x}_k),$$

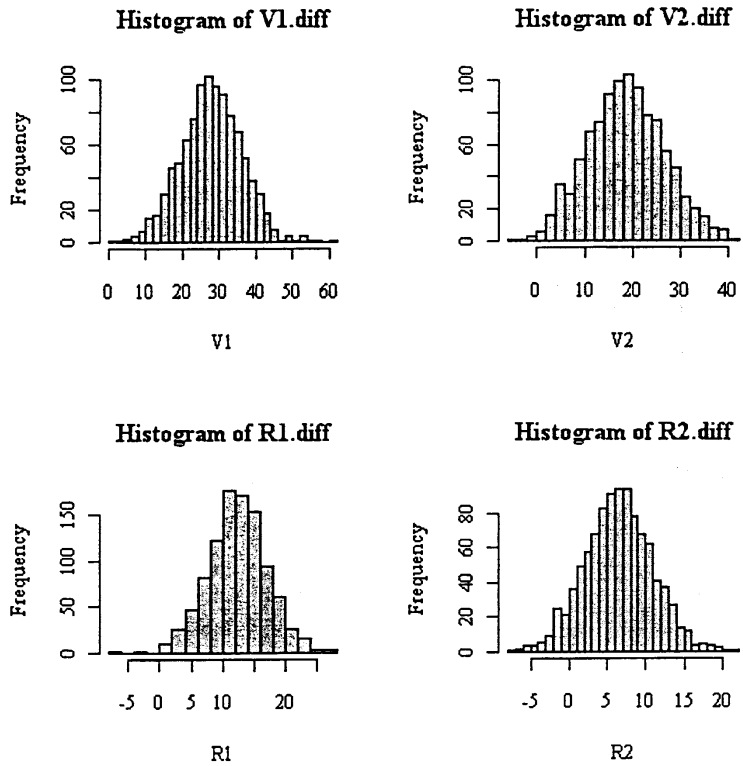
where $\hat{x}_{ij}$ and $\hat{x}_{ik}$ are observed or estimated values of x_{ij} and x_{ik}, respectively, from the last iteration of the EM algorithm, $\tilde{x}_j = \sum_i w_i\hat{x}_{ij} / \sum_i w_i$ is a weighted mean, and w_i is defined as

$$w_i = \begin{cases} \hat{\sigma}^2_{Y\cdot X}/\hat{\sigma}^2_{Y\cdot X_{\text{obs}},i}, & \text{if } y_i \text{ is present,} \\ 0, & \text{if } y_i \text{ is missing,} \end{cases}$$

where $\hat{\sigma}^2_{Y\cdot X_{\text{obs}},i}$ is the ML estimate of the variance of y given the independent variables observed in unit i. Alternatives are to apply the bootstrap, SEM, or to simulate the posterior distribution of the parameters. Specifically, suppose $\theta^{(d)}$ is a draw from the posterior distribution of θ in Eq. (11.19), computed using the data augmentation algorithm of Section 11.2.3; by Property 6.1B, SWP$[1, \ldots, p]\theta^{(d)}$ is a draw from the posterior distribution of SWP$[1, \ldots, p]\theta$. The last column of this matrix yields a draw of the regression parameters and residual variance.

ML estimation for multivariate linear regression can be achieved by applying the algorithm of Section 11.2.1, and then sweeping the independent variables in the resulting augmented covariance matrix. Specifically, if the dependent variables are $Y_1, \ldots, Y_K$ and the independent variables are $X_1, \ldots, X_p$, then the augmented covariance matrix of the combined set of variables $(X_1, \ldots, X_p, Y_1, \ldots, Y_K)$ is estimated using the multivariate normal EM algorithm, and then the variables $X_1, \ldots, X_p$ 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 Example 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 group. The regression coefficient of the group indicator measures the difference in mean outcome between the low- and medium/high-risk groups. Figure 11.2 displays histograms of 9000 draws of these regression coefficients from data augmentation. The 95% posterior probability intervals based on the 2.5th to 97.5th percentiles are shown below the plots.



95% Posterior Probability Intervals

V1	**(11.52, 43.34)**
V2	**(3.75, 34.54)**
R1	**(2.98, 21.48)**
R2	**(-2.25, 15.30)**

Figure 11.2. 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. (Example 11.6.)

Conclusions are similar to Example 11.1, namely, reading and verbal means appear higher in the low-risk group than in the moderate/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 BMDP (Dixon, 1988), SAS (1992), and S-Plus (Schafer, 1998a; Pinheiro and Bates, 2000).

Suppose that the hypothetical complete data for case i consists of K measurements $y_i = (y_{i1}, \ldots, y_{iK})$ on an outcome variable Y, with

$$y_i \sim_{\text{ind}} N_K[X_i\beta, \Sigma(\psi)], \tag{11.20}$$

where X_i is a known $(K \times m)$ design matrix for case 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 $\{y_{\text{obs},i}, i = 1, \ldots, n\}$ where $y_{\text{obs},i}$ is the observed part of the vector y_i. Missing values of y_i are assumed to be MAR. The complete-data loglikelihood is linear in the quantities $\{y_i, y_i^{\mathrm{T}} y_i, i = 1, \ldots, n\}$. Hence the E step consists in calculating the means of y_i and $y_i^{\mathrm{T}} y_i$ given $y_{\text{obs},i}$, 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, simplicity of the M step, is lost. Jennrich and Schluchter (1986) present a GEM 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. Hybrid computational strategies are discussed in Schafer (1998b).

A large number of situations can be modeled by combining different choices of mean and covariance structures, for example:

Independence: $\Sigma = \text{Diag}_K(\psi_1, \ldots, \psi_K)$, 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: $\Sigma = (\sigma_{jk})$, $\sigma_{jk} = \psi_1 \psi_2^{|j-k|}$, ψ_1, ψ_2 scalars,

Banded: $\sigma = (\sigma_{jk})$, $\sigma_{jk} = \psi_r$, where $r = |j - k| + 1, r = 1, \ldots, K$,

Factor analytic: $\Sigma = \Gamma\Gamma^{\mathrm{T}} + \psi$, $\Gamma = (K \times q)$ matrix of unknown factor loadings, and $\psi = (K \times K)$ diagonal matrix of specific variances,

Random effects: $\Sigma = Z\psi Z^{\mathrm{T}} + \sigma^2 I_K$, $Z = (K \times q)$ known matrix, $\psi = (q \times q)$ unknown dispersion matrix, σ^2 scalar, I_K the $K \times K$ identity matrix,

Unstructured: $\Sigma = (\sigma_{jk})$, $\psi_1 = \sigma_{11}, \ldots, \psi_K = \sigma_{1K}$ $\psi_{K+1} = \sigma_{22}, \ldots, \psi_v = \sigma_{KK}$, $v = K(K + 1)/2$.

The mean structure is also very flexible. If $X_i = I_K$, the $(K \times K)$ identity matrix, 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 years. Jennrich and Schluchter (1986) fit eight repeated measures models to these data. We fit the same models to the data obtained by deleting the ten values in parentheses in Table 11.4. The deletion mechanism is designed to be MAR but not MCAR. Specifically, for each gender, values at age 10 years are deleted for cases with low values at age 8 years. 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 ith 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 gender by age groups, and assumes that the

Table 11.4 Growth Data for 11 Girls and 16 Boys

Individual Girl	Age (years)				Individual Boy	Age (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)

Source: Potthoff and Roy (1964) as reported by Jennrich and Schluchter (1986). Values in parentheses are treated as missing in Example 11.7.

Table 11.5 Summary of Models Fit in Example 11.7

Model Number	Description	Number of Parameters	-2λ	Comparison Model	χ^2	df	Complete Data[a] 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]

[a] *Source:* Jennrich and Schluchter (1986).

(4×4) covariance matrix is unstructured. The design matrix for subject i can be written as

$$X_i = \begin{bmatrix} 1 & x_i & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & x_i & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & x_i & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & x_i \end{bmatrix}.$$

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

$$\mu_i^{\mathrm{T}} = X_i\beta = \begin{bmatrix} 1 & x_i & -3 & -3x_i \\ 1 & x_i & -1 & -x_i \\ 1 & x_i & 1 & x_i \\ 1 & x_i & 3 & 3x_i \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{bmatrix}, \tag{11.21}$$

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 and an unstructured Σ.

The likelihood ratio statistic comparing model 2 with model 1 is $\chi^2 = 6.33$ on four 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 gender groups. Compared with model 2, model 3 yields a likelihood ratio of 4.11 on one 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 gender group and a random intercept that varies across subjects about common gender means. 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, since 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 $Y = (y_0, y_1, \ldots, y_T)$ denote a completely observed univariate time series with $T + 1$ observations. The autoregressive model of lag p (ARp) assumes that y_i, the value at time i, is related to values at p previous time points by the model

$$(y_i | y_1, y_2, \ldots, y_{i-1}, \theta) \sim N(\alpha + \beta_1 y_{i-1} + \cdots + \beta_p y_{i-p}, \sigma^2), \tag{11.22}$$

where $\theta = (\alpha, \beta_1, \beta_2, \ldots, \beta_p, \sigma^2)$, α is a constant term, $\beta_1, \beta_2, \ldots, \beta_p$ are unknown regression coefficients, and σ^2 is an unknown error variance. Least squares estimates of $\alpha, \beta_1, \beta_2, \ldots, \beta_p$ and σ^2 can be found by regressing y_i on $x_i = (y_{i-1}, y_{i-2}, \ldots, y_{i-p})$, using observations $i = p, p + 1, \ldots, T$. These estimates are only approximately ML because the contribution of the marginal distribution of $y_0, y_1, \ldots, y_{p-1}$ 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 $y_0, y_1, \ldots, y_{p-1}$ can be ignored, since (1) missing values $y_i (i \geqslant p)$ appear as dependent and independent variables in the regressions, and (2) 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 ARp 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 AR1 Model for Time Series with Missing Values.* Setting $p = 1$ in Eq. (11.22), we obtain the model

$$(y_i|y_1, \ldots, y_{i-1}, \theta) \sim_{\text{ind}} N(\alpha + \beta y_{i-1}, \sigma^2). \tag{11.23}$$

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 $k \geqslant 1$. Ignoring the contribution of the marginal distribution of y_0, the complete-data loglikelihood for Y is $\ell(\alpha, \beta, \sigma^2|y) = -\sum_{i=1}^{T}(y_i - \alpha - \beta y_{i-1})^2/(20^2) - T\log\sigma^2/2$, which is equivalent to the loglikelihood for the normal linear regression of y_i on $x_i = y_{i-1}$, with data $\{(y_i, x_i), i = 1, \ldots, T\}$. The complete-data sufficient statistics are $S = (s_1, s_2, s_3, s_4, s_5)$, where

$$s_1 = \sum_{i=1}^{T} y_i, \qquad s_2 = \sum_{i=1}^{T} y_{i-1}, \qquad s_3 = \sum_{i=1}^{T} y_i^2, \qquad s_4 = \sum_{i=1}^{T} y_{i-1}^2, \qquad s_5 = \sum_{i=1}^{T} y_i y_{i-1}.$$

ML estimates of $\theta = (\alpha, \beta, \sigma)$ are $\hat{\theta} = (\hat{\alpha}, \hat{\beta}, \hat{\sigma})$, where

$$\begin{aligned} \hat{\alpha} &= (s_1 - \hat{\beta} s_2) T^{-1}, \\ \hat{\beta} &= (s_5 - T^{-1} s_1 s_2)(s_4 - T^{-1} s_2^2)^{-1}, \\ \hat{\sigma}^2 &= \left[s_3 - s_1^2 T^{-1} - \hat{\beta}^2 (s_4 - s_2^2 T^{-1}) \right] / T. \end{aligned} \tag{11.24}$$

Now suppose some observations are missing, and the data are MAR. 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

$$\begin{aligned} s_1^{(t)} &= \sum_{i=1}^{T} \hat{y}_i^{(t)}, \qquad s_2^{(t)} = \sum_{i=1}^{T} \hat{y}_{i-1}^{(t)}, \qquad s_3^{(t)} = \sum_{i=1}^{T} \left[(\hat{y}_i^{(t)})^2 + c_{ii}^{(t)} \right], \\ s_4^{(t)} &= \sum_{i=1}^{T} \left[(\hat{y}_{i-1}^{(t)})^2 + c_{i-1,i-1}^{(t)} \right], \qquad s_5^{(t)} = \sum_{i=1}^{T} \left[\hat{y}_{i-1}^{(t)} \hat{y}_i^{(t)} + c_{i-1,i}^{(t)} \right], \end{aligned}$$

and

$$\hat{y}_i^{(t)} = \begin{cases} y_i, & \text{if } y_i \text{ is present,} \\ E\left\{ y_i | Y_{\text{obs}}, \theta^{(t)} \right\}, & \text{if } y_i \text{ is missing,} \end{cases}$$

$$c_{ij}^{(t)} = \begin{cases} 0, & \text{if } y_i \text{ or } y_j \text{ is present,} \\ \text{Cov}\left\{ y_i, y_j | Y_{\text{obs}}, \theta^{(t)} \right\}, & \text{if } y_i \text{ and } y_j \text{ are missing.} \end{cases}$$

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 $Y_{\text{mis}}{}^* = (y_{j+1}, y_{j+2}, \ldots, y_{k-1})$ is a sequence of missing values, bounded by present observations, y_j and y_k. Then (1) Y^*_{mis} is independent of the other missing values, given Y_{obs} and θ, and (2) the distribution of Y^*_{mis} given Y_{obs} and θ depends on Y_{obs} 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 $(y_j, y_{j+1}, \ldots, y_k)$ 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

$$\text{SWP}[j, j+2]A = \frac{1}{1+\beta^2}\begin{bmatrix} -\sigma^{-2} & \beta & -\beta^2\sigma^{-2} \\ \beta & \sigma^2 & \beta \\ -\beta^2\sigma^{-2} & \beta & -\sigma^{-2} \end{bmatrix}. \tag{11.25}$$

Hence from stationarity and expression (11.25),

$$\begin{aligned} E\{y_{j+1} | y_j, y_{j+2}, \theta\} &= \mu + \beta(1+\beta^2)^{-1}(y_{j+2} - \mu) + \beta(1+\beta^2)^{-1}(y_j - \mu) \\ &= \mu\left\{1 - \frac{2\beta}{1+\beta^2}\right\} + \frac{\beta}{1+\beta^2}\{y_j + y_{j+2}\}, \\ \text{Var}(y_{j+1} | y_j, y_{j+2}, \theta) &= \sigma^2(1+\beta^2)^{-1}. \end{aligned}$$

Substituting $\theta = \theta^{(t)}$ in these expressions yields $\hat{y}^{(t)}_{j+1}$ and $\hat{c}^{(t)}_{j+1,j+1}$ for the E step. Note that $\hat{y}^{(t)}_{j+1}$ supplies a prediction for the missing value at the final iteration of the algorithm.

11.6.3. Kalman Filter Models

Shumway and Stoffer (1982) consider the Kalman filter model

$$\begin{aligned} (y_i | A_i, z_i, \theta) &\sim_{\text{ind}} N(z_i A_i, B), \\ (z_0 | \theta) &\sim N(\mu, \Sigma), \\ (z_i | z_1, \ldots, z_{i-1}, \theta) &\sim N(z_{i-1}\phi, Q), \; i \geqslant 1, \end{aligned} \tag{11.26}$$

where y_i is a $(1 \times q)$ vector of observed variables at time i, A_i is a known $(p \times q)$ design 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 $i = 1, 2, \ldots, n$ (smoothing) and for $i = n+1, n+2, \ldots$ (forecasting), using the observed series $y_1, y_2, \ldots, y_n$. If the parameter θ were known, the optimal 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_{obs}. 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 component Y_{mis} of Y *and* $z_1, z_2, \ldots, z_n$ 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_{obs} 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

$$\sum_{i=1}^{n} z_i, \qquad \sum_{i=1}^{n} z_i^{\mathrm{T}} z_i, \qquad \sum_{i=1}^{n} z_{i-1}, \qquad \sum_{i=1}^{n} z_{i-1}^{\mathrm{T}} z_{i-1}, \qquad \text{and} \qquad \sum_{i=1}^{n} z_{i-1}^{\mathrm{T}} z_i$$

from the E step; B is estimated by the expected value of the residual covariance matrix $n^{-1} \sum_{i=1}^{n} (y_i - z_i A_i)^{\mathrm{T}} (y_i - z_i A_i)$. 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 Social Security Administration (SSA), yielding Y_1, and Health Care Financing Administration (HCFA), yielding Y_2. Shumway and Stoffer (1982) analyze the data using the model

$$\begin{aligned} (y_{ij} | z_i, \theta) &\overset{\text{ind}}{\sim} N(z_i, B_j), \qquad i = 1949, \ldots, 1981, \\ (z_i | z_1, \ldots, z_{i-1}, \theta) &\sim N(z_{i-1}\phi, Q), \qquad i = 1950, \ldots, 1981, \end{aligned}$$

Table 11.6 Data Set for Example 11.9 and Predictions from EM Algorithm—Physician Expenditures (in Millions)

			Predictions from EM Algorithm	
Year, i	SSA y_{i1}	HCFA y_{i2}	$E(z_i\|\text{data}, \theta)$	$\text{Var}^{1/2}(z_i\|\text{data}, \theta)$
1949	2633	—	2541	178
1950	2747	—	2711	185
1951	2868	—	2864	186
1952	3042	—	3045	186
1953	3278	—	3269	186
1954	3574	—	3519	186
1955	3689	—	3736	186
1956	4067	—	4063	186
1957	4419	—	4433	186
1958	4910	—	4876	186
1959	5481	—	5331	186
1960	5684	—	5644	186
1961	5895	—	5972	186
1962	6498	—	6477	186
1963	6891	—	7032	185
1964	8065	—	7866	179
1965	8745	8474	8521	110
1966	9156	9175	9198	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.

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 expo-

nential 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.

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 multivariate normal with unrestricted means and variances and zero correlations, with ignorable nonresponse. Infer situations where the available cases method is appropriate for multivariate data with missing values.

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 means (μ_1, μ_2) and variances (σ_1^2, σ_2^2) assumed known, 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, 1983) 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 $\tau^2 \to \infty$, 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 value 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.

11.14. Review the GEM algorithm in Jennrich and Schluchter (1986) for ML estimation for the model of Eq. (11.20). Under what circumstances is the GEM algorithm an ECM algorithm, as defined in Section 8.5.1?

CHAPTER 12

Robust Estimation

12.1. INTRODUCTION

Examples 8.4, 8.8, 8.9, 8.10, and 10.4 concerned robust inference for a single sample based on the t distribution, with degrees of freedom fixed *a priori* or estimated from the data. In this chapter we develop this idea by considering alternative distributions to the t for robust inference, and robust inference for multivariate data sets with missing values. As in Chapter 11, we assume that the missing-data 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 12.2 describes a general mixture model for robust estimation of a univariate sample that includes the t and contaminated normal distributions as special cases. The case of multivariate data with unstructured mean and covariance matrix is described in Section 12.3.1, and extended to handle missing data and structured mean and covariance matrices in Section 12.3.2. Section 12.4 outlines extensions of the t model.

12.2. ROBUST ESTIMATION FOR A UNIVARIATE SAMPLE

Dempster, Laird and Rubin (1977, 1980) consider ML estimation for the following model. Let $X = (x_1, \ldots, x_n)^{\mathrm{T}}$ be an independent random sample from a population such that

$$(x_i|\theta, w_i) \sim_{\text{ind}} N(\mu, \sigma^2/w_i),$$

where the w_i are unobserved iid positive scalar random variables with known density $h(w_i)$. Inferences about $\theta = (\mu, \sigma)^{\mathrm{T}}$ can be based on the incomplete data methods, treating $W = (w_1, \ldots, w_n)^{\mathrm{T}}$ as missing data.

In particular, if X and W were observed, then ML estimates of (μ, σ) would be found by weighted least squares:

$$\hat{\mu} = \sum_{i=1}^{n} w_i x_i / \sum_{i=1}^{n} w_i = s_1/s_0, \tag{12.1}$$

$$\hat{\sigma}^2 = \sum_{i=1}^{n} w_i (x_i - \hat{\mu})^2 / n = (s_2 - s_1^2/s_0)/n, \tag{12.2}$$

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 as defined in Section 8.4.2. 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 $\theta^{(t)} = (\mu^{(t)}, \sigma^{(t)2})$ of θ. Since s_0, s_1, and s_2 are linear in the $\{w_i\}$, the E step reduces to finding estimated weights

$$w_i^{(t)} = E(w_i | x_i, \mu^{(t)}, \sigma^{(t)2}). \tag{12.3}$$

M Step:
Compute new estimates $(\mu^{(t+1)}, \sigma^{(t+1)2})$ from Eqs. (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 Eq. (12.3). Convergence can be speeded by replacing the denominator n in Eq. (12.2) by $\sum_{i=1}^{n} w_i^{(t)}$, as in the PX-EM algorithm of Example 8.10.

The form of the weights (12.3) in this iteratively reweighted least squares algorithm depends on the assumed distribution for w_i. Examples 8.4, 8.8, 8.9, 8.10, and 10.4 described the case where w_i is a scaled chi-squared distribution. Another choice, useful when the data contain extreme outliers, is given in the following example:

EXAMPLE 12.1. *The Univariate Contaminated Normal Model.* Suppose that $h(w_i)$ is nonzero at two values of w_i, 1 and $\lambda < 1$, such that

$$h(w_i) = \begin{cases} 1 - \pi & \text{if } w_i = 1, \\ \pi & \text{if } w_i = \lambda, \text{ known}, \\ 0 & \text{otherwise}, \end{cases} \tag{12.4}$$

where $0 < \pi < 1$. Then the marginal distribution of x_i is a mixture of $N(\mu, \sigma^2)$ and $N(\mu, \sigma^2/\lambda)$. This is a contaminated normal model, with probability of contamination π. For example, $\lambda = 0.1$ if the contamination is assumed to inflate the variance by a factor of 10.

A simple application of Bayes theorem yields

$$E(w_i | x_i, \mu, \sigma^2) = \frac{1 - \pi + \pi\lambda^{3/2} \exp[(1-\lambda)d_i^2/2]}{1 - \pi + \pi\lambda^{1/2} \exp[(1-\lambda)d_i^2/2]}, \tag{12.5}$$

and

$$\Pr(w_i = \lambda) = 1 - \Pr(w_i = 1) = \frac{\pi\lambda^{1/2}\exp[(1-\lambda)d_i^2/2]}{1-\pi+\lambda^{1/2}\exp[(1-\lambda)d_i^2/2]}, \tag{12.6}$$

where

$$d_i^2 = (x_i - \mu)^2/\sigma^2. \tag{12.7}$$

The weights $w_i^{(t)}$ for EM are obtained by substituting current estimates $\mu^{(t)}$ and $\sigma^{(t)}$ in Eqs. (12.5)–(12.7). Observe that values x_i far from the mean have large values of d_i^2 and (for $\lambda < 1$) reduced weights in the M step. Thus the algorithm leads to a robust estimate of μ that downweights outlying cases.

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 Eq. (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 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 Example 8.4 and Example 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.

12.3. ROBUST ESTIMATION OF THE MEAN AND COVARIANCE MATRIX

12.3.1. Multivariate 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 $X_1, \ldots, X_K$. We suppose that x_i has the K-variate normal distribution

$$(x_i|\theta, w_i) \sim_{\text{ind}} N_K(\mu, \Psi/w_i) \tag{12.8}$$

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 $W = (w_1, \ldots, w_n)^{\mathrm{T}}$ as missing data.

If W were observed, ML estimates of μ and Ψ would be the multivariate analogs of Eqs. (12.1) and (12.2). The complete-data sufficient statistics are $s_0 = \sum_{i=1}^n w_i$, $s_1 = \sum_{i=1}^n w_i x_i$ and $s_2 = \sum_{i=1}^n w_i x_i^{\mathrm{T}} x_i$, and:

$$\hat{\mu} = s_1/s_0 = \sum_{i=1}^n w_i x_i / \sum_{i=1}^n w_i, \tag{12.9}$$

$$\hat{\Psi} = \frac{s_2 - s_1^{\mathrm{T}} s_1/s_0}{n} = \sum_{i=1}^n \frac{w_i (x_i - \hat{\mu})^{\mathrm{T}} (w_i - \hat{\mu})}{n}, \tag{12.10}$$

Hence when W is not observed, the $(t+1)$st 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. Since s_0, s_1, and s_2 are linear in $\{w_i\}$, the E step again 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 Eqs. (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 Eq. (12.9) by the sum of the current weights, $\sum_{i=1}^n w_i^{(t)}$.

If $\{w_i\}$ are distributed as Eq. (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 Eqs. (12.5)–(12.7):

$$E(w_i | x_i, \mu, \Psi) = \frac{1 - \pi + \pi\lambda^{K/2+1} \exp[(1-\lambda) d_i^2/2]}{1 - \pi + \pi\lambda^{K/2} \exp[(1-\lambda) d_i^2/2]}, \tag{12.11}$$

where d_i^2 is now the squared Mahalanobis distance for case i:

$$d_i^2 = (x_i - \mu)\Psi^{-1}(x_i - \mu)^{\mathrm{T}}. \tag{12.12}$$

The model downweights cases with large values of d_i^2.

If, on the other hand, the distribution of $w_i \sim_{\text{ind}} \chi_\nu^2/\nu$, the weights are given by the following generalization of Eq. (8.24):

$$E(w_i | x_i, \mu, \Psi) = (\nu + K)/(\nu + d_i^2), \tag{12.13}$$

where d_i^2 is again given by Eq. (12.12).

Rubin (1983a) also considers extensions of these models to multivariate regression.

12.3.2. Robust Estimation of the Mean and Covariance Matrix from Data with Missing Values

Little (1988b) extends these algorithms to situations where some values of X are missing. Let $x_{\text{obs},i}$ denote the set of variables observed in case i, let $x_{\text{mis},i}$ denote the missing variables, and write $X_{\text{obs}} = \{x_{\text{obs},i} : i = 1, \ldots, n\}$ and $X_{\text{mis}} = \{x_{\text{mis},i} : i = 1, \ldots, n\}$. We assume that (1) x_i given w_i has the distribution given by Eq. (12.8) and (2) the missing data are MAR. ML estimates of μ and Ψ are found by applying the EM algorithm, treating the values of X_{mis} and W as missing data.

The M step 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_{obs} and current estimate $\theta^{(t)} = (\mu^{(t)}, \Psi^{(t)})$ of the parameter θ. We find

$$E(s_0|\theta^{(t)}, X_{\text{obs}}) = E\left(\sum_{i=1}^{n} w_i|\theta^{(t)}, x_{\text{obs}}\right) = \sum_{i=1}^{n} w_i^{(t)},$$

where $w_i^{(t)} = E(w_i|\theta^{(t)}, x_{\text{obs},i})$; the jth component of $E(s_1|\theta^{(t)}, X_{\text{obs}})$ is

$$\begin{aligned} E\left(\sum_{i=1}^{n} w_i x_{ij}|\theta^{(t)}, X_{\text{obs}}\right) &= \sum_{i=1}^{n} E[w_i E(x_{ij}|\theta^{(t)}, x_{\text{obs},i}, w_i)|\theta^{(t)}, x_{\text{obs},i}] \\ &= \sum_{i=1}^{n} w_i^{(t)} \hat{x}_{ij}^{(t)}, \end{aligned}$$

where $\hat{x}_{ij}^{(t)} = E(x_{ij}|\theta^{(t)}, x_{\text{obs},i})$, since the conditional mean of x_{ij} given $\theta^{(t)}, x_{\text{obs},i}$, and w_i does not depend on w_i. Finally, the (j, k)th element of $E(s_2|\theta^{(t)}, X_{\text{obs}})$ is

$$\begin{aligned} E\left(\sum_{i=1}^{n} w_i x_{ij} x_{ik}|\theta^{(t)}, X_{\text{obs}}\right) &= \sum_{i=1}^{n} E[w_i E(x_{ij} x_{ik}|\theta^{(t)}, x_{\text{obs},i}, w_i)|\theta^{(t)}, x_{\text{obs},i}] \\ &= \sum_{i=1}^{n} (w_i^{(t)} \hat{x}_{ij}^{(t)} \hat{x}_{ik}^{(t)} + \psi_{jk\cdot\text{obs},i}^{(t)}), \end{aligned}$$

where the adjustment $\psi_{jk\cdot\text{obs},i}^{(t)}$ is zero if x_{ij} or x_{ik} are observed, and w_i times the residual covariance of x_{ij} and x_{ik} given $x_{\text{obs},i}$, if x_{ij} and x_{ik} are both missing. The quantities $\hat{x}_{ij}^{(t)}$ and $\psi_{jk\cdot\text{obs},i}^{(t)}$ are found by sweeping on the current estimate $\Psi^{(t)}$ of Ψ to make $x_{\text{obs},i}$ 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 passed to the 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 in case i, and (ii) the squared Mahalanobis distance (12.12) computed using only the observed variables in case i.

Both the multivariate t and contaminated normal models downweight cases 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.* As an illustration, 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 an artificially generated multivariate t_4 data set with $K = 4$ variables, $n = 80$ cases, and 72 of the

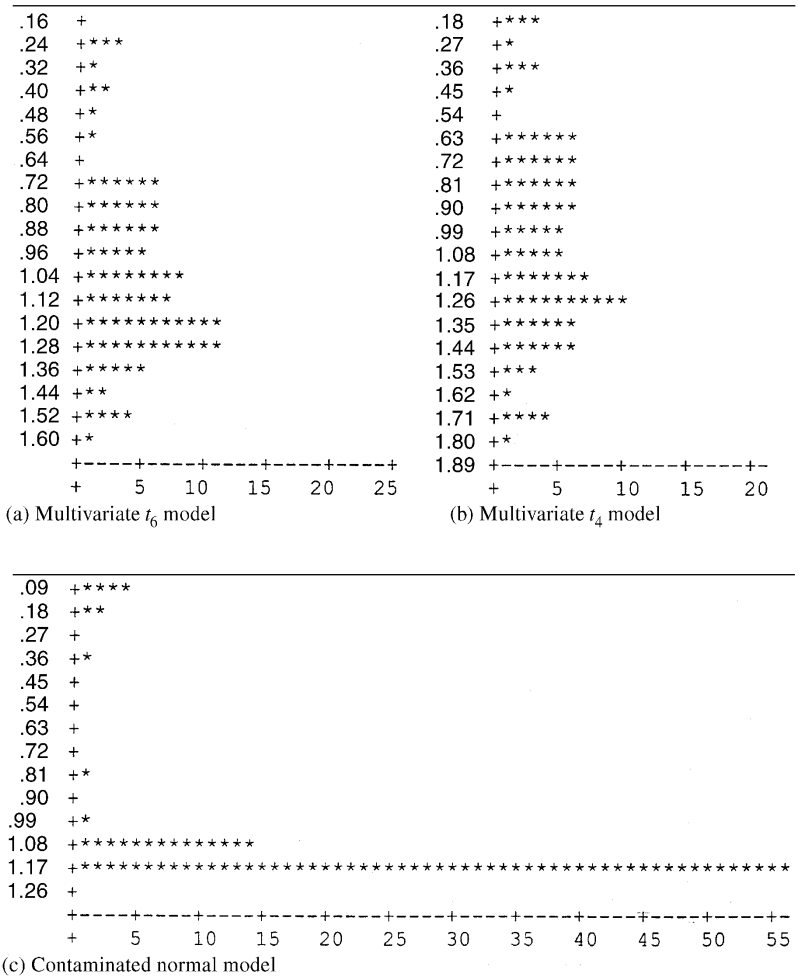


Figure 12.1. Distributions of weights* from robust ML methods, applied to multivariate t_4 data.

320 values randomly deleted. Observe that the t weights are more dispersed for $\nu = 4$ than for $\nu = 6$, and the downweighting for the contaminated normal model tends to be concentrated in a few outlying cases.

Little (1988b) shows by a simulation study that ML for these models can produce estimates of means, slopes, and correlations that are protected against outliers when the data are non-normal, with minor sacrifices of efficiency when the data are, in fact, normal. A graphical procedure for assessing normality is also presented.

12.3.3. Adaptive Robust Multivariate Estimation

The methods discussed so far assume that the parameters ν of the t model or (π, λ) of the contaminated normal model are known. A more flexible approach is to estimate these parameters in addition to the mean and scale parameters, yielding a form of adaptive robust estimation. ML estimation of ν 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 12.3.1 and 12.3.2. Iteration t of an ECME algorithm for the multivariate incomplete data of Section 12.3.2 is as follows:

E Step:
As in Section 12.3.2, with ν replaced by current estimate $\nu^{(t)}$.

M Step:
Compute new estimates $(\mu^{(t+1)}, \psi^{(t+1)})$ as in Sections 12.3.1 and 12.3.2. Compute $\nu^{(t+1)}$ to maximize the observed loglikelihood $\ell(\mu^{(t+1)}, \Psi^{(t+1)}, \nu|X_{\text{obs}})$ with respect to ν. This is a one-dimensional maximization, and can be achieved by a grid search or a Newton step.

12.3.4. Bayes Inferences for the t Model

The methods of ML estimation in Sections 12.3.1 to 12.3.3 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, and assume a noninformative prior:

$$p[\mu, \Psi, \ln(1/\nu)] \propto |\Psi|^{-(K+1)/2}, \qquad -10 < \ln(1/\nu) < 10. \tag{12.14}$$

The prior on ν is that used in Liu and Rubin (1998). Let $(\mu^{(t)}, \Psi^{(t)}, \nu^{(t)})$ and $(x_{\text{mis},i}^{(t)}, w_i^{(t)}, i = 1, \ldots, n)$ be draws of the parameters and missing values at iteration t. Iteration $t + 1$ consists of the following computations:

(a) For $i = 1, \ldots, n$, draw new weights $w_i^{(t+1)} \sim u_i/(\nu^{(t)} + d_{\text{obs},i}^{2(t)})$, where $u_i \sim \chi^2_{\nu^{(t)}+K_i}$.

(b) For $i = 1, \ldots, n$, draw missing data $x_{\text{mis},i}^{(t+1)} = \hat{x}_{\text{mis},i}^{(t)} + z_i$, where $\hat{x}_{\text{mis},i}^{(t)}$ is the predicted mean from the (normal) linear regression of $x_{\text{mis},i}$ on $x_{\text{obs},i}$ given

current parameter estimates, and z_i is normal with mean 0 and covariance matrix $(w_i^{(t)})^{-1}\Psi^{(t)}_{\text{mis·obs},i}$, where $\Psi^{(t)}_{\text{mis·obs},i}$ is obtained by sweeping $\Psi^{(t)}$ on the observed variables.

(c) Draw $(\mu^{(t+1)}, \Psi^{(t+1)})$ from the posterior distribution of these parameters given filled-in data $(X_{\text{obs}}, X^{(t+1)}_{\text{mis}})$ and weights $W^{(t+1)}$. This is a standard complete-data problem, with $\Psi^{(t+1)}$ being drawn from a scaled inv-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 since 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 (e.g., see Tanner, 1996, Section 6.4) was employed, with 400 equally spaced cut-points on the interval $(-10, 10)$.

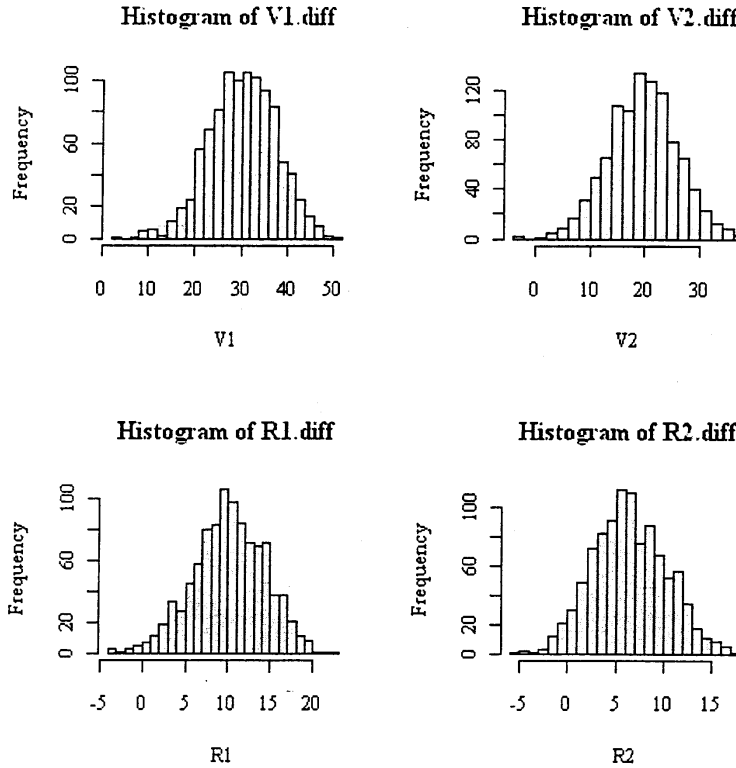
Convergence of the algorithm can be speeded by defining the missing values to yield a monotone pattern, and then developing the P step for a monotone missing-data pattern, using the ideas of Chapter 7. This approach is described in Liu (1995). A straightforward extension is robust multivariate regression 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. *Bayesian 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 similar. The posterior distribution of v is fairly dispersed and sensitive to the choice of prior for v, but inferences for the differences in means are less sensitive to this modeling choice.

12.4. 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, Little and Taylor (1989) analyze data reported by LaVange and Helms (1983) from a longitudinal study of lung function conducted on 72 children aged 3 to 12 years at a child development center. The variables consist of



95% Posterior Probability Intervals

V1	**(15.38, 43.99)**
V2	**(7.08, 31.42)**
R1	**(1.68, 17.94)**
R2	**(-0.60, 14.04)**

Figure 12.2. 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. (Example 12.3.)

race (black or white), gender, and $\log(\text{Vmax}_{75})$ for single-year ages from 3 to 12, where Vmax_{75} is the maximum expiratory flow rate after 75% of the forced vital capacity has been exhaled. Of the 10 possible measurements of Vmax_{75} for each child, 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 combinations of early and late ages are never observed together, so the covariance matrix is not estimable without placing restrictions on the parameters.

The results in Table 12.1 show whether there are differences in the growth curves of $\log(\text{Vmax}_{75})$ over time between males and females. Let y_{ij} denote the value of

Table 12.1 Normal and *t* Models of Lung Function Data. Summary of Fits for 12 Models. (Example 12.4.)

Model	Mean Constraints	Covariance Constraints	No. of Parameters	Maximum Loglikelihood	LR χ^2 (df)
A: Normal Models					
1N	None	None	9	164.4	0 (0)
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*(1)
B: *t* Models					
1T	None	None	10	187.1	0 (0)
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* (1)
6T	None	$\psi_2 = 0$	9	181.5	9.2* (1)

* Significantly worse fit than Model 1N (A) or 1T (B) at the 1% level (asymptotic LR chi-squared test).

$\log(\text{Vmax}_{75})$ for individual i at age $j+2$, for $1 \leqslant j \leqslant 10$. Table 12.1A shows the maximized loglikelihoods for normal repeated-measures models of the form (11.20), namely:

$$y_i \sim_{\text{ind}} N_{10}[\mu_i(\beta), \Sigma(\psi)]. \tag{12.15}$$

The covariance matrix Σ is modeled as $\sigma_{jk} = \psi_1(\psi_2 + (1-\psi_2)\psi_3^{|k-j|})$, where ψ_1 is the total dispersion, ψ_2 is a heritability parameter, and ψ_3 is an environmental decay constant. The jth component of the mean $\mu_i(\beta)$ has the form:

$$\mu_{ij} = \begin{cases} \beta_0 + \beta_1 \text{age}_j + \beta_2 \text{age}_j^2 & \text{if sex}_i = \text{male}, \\ \beta_3 + \beta_4 \text{age}_j + \beta_5 \text{age}_j^2 & \text{if sex}_i = \text{female}, \end{cases} \tag{12.16}$$

modeling separate 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 12.1A 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 12.1B shows fits for the same set of models with the normal distribution in Eq. (12.15) replaced by the t, with degrees of freedom ν estimated from the data by ML. Note that these models fit much better than the normal counterparts, with maximized loglikelihoods 15–23 units higher than the maximized loglikelihoods for the 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 zero does not fit well, unlike the corresponding normal model 5N. It appears that for the latter 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 degrees of freedom and increases the size and statistical significance of the ψ_3 parameter, as expected from the model comparisons in Table 12.1. The slopes of

Table 12.2. Normal and t Models of Lung Function Data. Parameter Estimates and Standard Errors for Models 4N and 4T. (Example 12.4.)

Model	β_0	β_1	ψ_1	ψ_2	ψ_3	ν
4N	−0.365	0.0637	0.167	0.362	0.175	∞
	(0.075)	(0.0102)	(0.017)	(0.076)	(0.092)	(—)
4T	−0.286	0.0608	0.109	0.406	0.304	4.4
	(0.069)	(0.0090)	(0.017)	(0.092)	(0.102)	(1.2)

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 w_i applied to model 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 $\Sigma\beta_j z_{ij}$ and covariance matrix Ψ/w_i, conditional on the unknown scaling quantity w_i defined as before. Thus, suppose the data can be arranged in a *monotone* missing-data pattern, with the variables arranged in blocks $X_1, \ldots, X_k$ such that for $j = 1, \ldots, K-1$, X_j is observed for all cases where X_{j+1} is observed. Then the joint distribution of $X_1, \ldots, X_k$ can be expressed as the product of distributions

$$f(X_1, \ldots, X_K|\phi) = f(X_1|\phi)f(X_2|X_1, \phi)\cdots f(X_K|X_1, \ldots, X_{K-1}, \phi),$$

as discussed in Chapter 7. The conditional distribution $f(X_j|X_1, \ldots, X_{j-1})$ in this factorization can then be modeled as multivariate normal with mean $\sum_{u=1}^{j-1}\beta_u X_u$ and covariance matrix $\Psi_{j\cdot 1\ldots j-1}/w_{ij}$, where now the scaling factor w_{ij} can be allowed to 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 $X_1, \ldots, X_k$ are found by transformation, as discussed in Chapter 7.

PROBLEMS

12.1. Derive the weighting function (12.5) for the model of Example 12.1.

12.2. Write down the data augmentation algorithm for Example 12.1.

12.3. Write a program to compute ML estimates for the contaminated normal model of Example 12.1.

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.

12.5. 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 seem to involve too much missing information to be practical?)

12.6. Extend the contaminated normal and t models with known df to the case of simple linear regression of X on a fixed covariate Z. Derive EM algorithms for these models. Do cases with Z observed and X missing contribute information to the regressions in these cases? (Hint: review Section 11.4.1.)

12.7. Derive the weighting functions (12.11) and (12.13) for the models of Section 12.3.

12.8. Derive the E step equations in Section 12.3.2.

CHAPTER 13

Models for Partially Classified Contingency Tables, Ignoring the Missing-Data Mechanism

13.1. INTRODUCTION

This chapter concerns the analysis of incomplete data when the 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 observations on V categorical variables $Y_1, \ldots, Y_V$ 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 $\{n_{jk\cdots t}\}$, where $n_{jk\cdots t}$ is the number of sampled cases in the cell with $Y_1 = j$, $Y_2 = k, \ldots, Y_V = t$. If the data matrix has missing items, some of the cases in the preceding contingency table are partially classified. The completely classified cases yield a V-dimensional table of counts $\{r_{jk\cdots t}\}$, and the incompletely classified cases form supplemental lower-dimensional subtables defined by the subset of variables $(Y_1, \ldots, Y_V)$ that are observed. For example, the first eight rows of Table 1.3 provide data from the complete cases in a five-way contingency table with variables gender, 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 via 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. Nonignorable nonresponse models for categorical data are deferred until Chapter 15.

A more general type of incomplete data occurs when level j of a particular item, Y_1, say, is not known, but it is known that the case 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 problems considered here should be carefully distinguished from the problem of structural zeros, where certain cells contain zero counts because the model assigns them zero probability of containing entries. For example, if $Y_1 =$ year of birth and $Y_2 =$ year of first marriage, and marriages below the age of 10 are ruled out, then cells with $Y_2 \leqslant Y_1 + 9$ are structural zeros in the joint distribution of Y_1 and Y_2. Instances of no data (zero counts) are not considered here as instances of missing data. For discussion of the structural zero problem, see, for example, Bishop, Fienberg and Holland (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 $\{n_{jk\cdots t}\}$ have a multinomial distribution with index (that is, total count) n and probabilities $\theta = \{\pi_{jk\cdots t}\}$. We also assume that the missing-data mechanism is ignorable, 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

$$\mathrm{L}(\theta|\{n_{jk\cdots t}\}) = \prod_{j,k,\ldots,t} \pi_{jk\cdots t}^{n_{jk\cdots t}}, \quad \sum_{j,k,\ldots t} \pi_{jk\cdots t} = 1, \tag{13.1}$$

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 $\{n_{jk\cdots t}\}$ are independent Poisson random variables with means $\{\mu_{jk\cdots t}\}$ and cell probabilities $\pi^*_{jk\cdots t} = \mu_{jk\cdots t} / \sum_{j,k,\ldots,t} \mu_{jk\cdots t}$. If the nonresponse mechanism is ignorable, likelihood inferences for $\{\pi^*_{jk\cdots t}\}$ are the same as those for $\{\pi_{jk\cdots t}\}$ under the multinomial model. This fact follows from arguments analogous to those for the complete data case (Bishop, Fienberg and Holland, 1975). We restrict attention to the multinomial model since it seems more common than the Poisson model in practical situations.

For complete data, the likelihood (13.1) yields the ML estimate

$$\hat{\pi}_{jk\cdots t} = n_{jk\cdots t}/n$$

with large-sample variance

$$\text{Var}(\pi_{jk\cdots t} - \hat{\pi}_{jk\cdots t}) = \hat{\pi}_{jk\cdots t}(1 - \hat{\pi}_{jk\cdots t})/n.$$

For Bayes inference, we multiply the complete-data likelihood by the conjugate Dirichlet prior:

$$p(\{\pi_{jk\cdots t}\}) \propto \prod_{j,k,\ldots,t} \pi_{jk\cdots t}^{\alpha_{jk\cdots t}-1}, \pi_{jk\cdots t} > 0, \qquad \sum_{jk,\ldots,t} \pi_{jk\cdots t} = 1. \tag{13.2}$$

Combining this prior distribution with the likelihood yields the Dirichlet posterior distribution

$$p(\{\pi_{jk\cdots t}\}|\{n_{jk\cdots t}\}) \propto \prod_{j,k,\ldots,t} \pi_{jk\cdots t}^{\alpha_{jk\cdots t}+n_{jk\cdots t}-1}, \qquad \sum_{jk,\ldots,t} \pi_{jk\cdots t} = 1; \tag{13.3}$$

(see Example 6.17).

Our objective is to obtain analogous answers from incomplete data. In this section we discuss special patterns of incomplete data that yield explicit results.

13.2.2. ML Estimation for Monotone Patterns

We first consider ML estimates for the simple case of a two-way contingency table with one supplemental one-way margin.

EXAMPLE 13.1. *Two-Way Contingency Table with One Supplemental One-Way Margin.* Consider two categorical variables Y_1, with levels $j = 1, \ldots, J$ and Y_2, with levels $k = 1, \ldots, K$. The data consist of r observations $\{(y_{i1}, y_{i2}), i = 1, \ldots, r\}$ with y_{i1} and y_{i2} recorded and $m = n - r$ observations $\{y_{i1}, i = r+1, \ldots, n\}$ with y_{i1} recorded and y_{i2} missing. The data pattern is identical to Example 7.1, but the variables are now categorical.

The r completely classified observations can be displayed in a $(J \times K)$ contingency table, with r_{jk} observations in the cell with $y_{i1} = j, y_{i2} = k$. The $n - r$ remaining observations form a supplemental $(J \times 1)$ margin, with m_j units in the cell with $y_{i1} = j$ (see Figure 13.1).

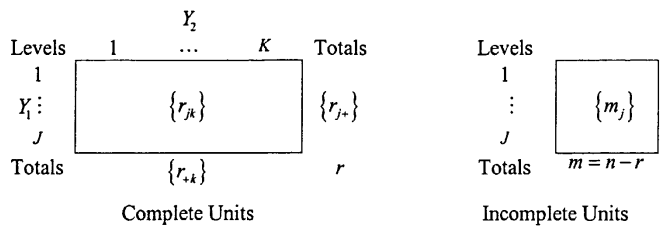


Figure 13.1. The data in Example 13.1.

We shall use the standard plus notation for summation over subscripts j and k. For this problem

$$\theta = (\pi_{11}, \pi_{12}, \ldots, \pi_{JK}) \quad \text{and} \quad \sum_{j=1}^{J}\sum_{k=1}^{K} \pi_{jk} \equiv \pi_{++} = 1.$$

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

$$\mathrm{L}(\phi|\{r_{jk}, m_j\}) = \left(\prod_{j=1}^{J} \pi_{j+}^{r_{j+}+m_j}\right) \times \left(\prod_{j=1}^{J}\prod_{k=1}^{K} \pi_{k\cdot j}^{r_{jk}}\right), \tag{13.4}$$

where the first term is the likelihood for the multinomial distribution of the marginal counts $r_{j+} + m_j$, with index n and probabilities π_{j+}, and the second term is the likelihood for the product of J conditional multinomial distributions of $\{r_{jk}\}$ given r_{j+}, with index r_{j+} and probabilities

$$\pi_{k\cdot j} = \Pr(Y_2 = k|Y_1 = j) = \pi_{jk}/\pi_{j+}, \qquad k = 1, \ldots, K.$$

Note that Eq. (13.4) is a factorization of the likelihood as discussed in Section 7.1, with distinct parameters

$$\phi_1 = \{\pi_{j+}, j = 1, \ldots, J\} \quad \text{and} \quad \phi_2 = \{\pi_{k\cdot j}, j = 1, \ldots, J; k = 1, \ldots, K\}.$$

Maximizing each component separately, we obtain ML estimates

$$\hat{\pi}_{j+} = \frac{r_{j+} + m_j}{n}, \qquad \hat{\pi}_{k\cdot j} = \frac{r_{jk}}{r_{j+}},$$

and so

$$\hat{\pi}_{jk} = \hat{\pi}_{j+}\hat{\pi}_{k\cdot j} = \frac{r_{jk} + (r_{jk}/r_{j+})m_j}{n}. \tag{13.5}$$

Thus the ML estimates effectively distribute a proportion r_{jk}/r_{j+} of the unclassified observations m_j into the (j, k)th cell.

For a Bayesian analysis, suppose independent Dirichlet prior distributions are specified for $\{\pi_{j+}\}$ and $\{\pi_{k\cdot j}\}$ corresponding to the factored likelihood in Eq. (13.2), that is:

$$p(\phi) \propto \left(\prod_{j=1}^{J} \pi_{j+}^{n_{j0}-1}\right) \times \left(\prod_{j=1}^{J}\prod_{k=1}^{K} \pi_{k\cdot j}^{r_{jk0}-1}\right).$$

Then the posterior distribution is a product of independent posterior distributions for $\{\pi_{j+}\}$ and $\{\pi_{k\cdot j}\}$, namely:

$$p(\phi|\text{data}) \propto \left(\prod_{j=1}^{J} \pi_{j+}^{n_{j0}+r_{j+}+m_j-1}\right) \times \left(\prod_{j=1}^{J}\prod_{k=1}^{K} \pi_{k\cdot j}^{r_{jk}+r_{jk0}-1}\right). \tag{13.6}$$

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 posterior distributions in Eq. (13.6), and then setting $\pi_{jk}^{(d)} = \pi_{j+}^{(d)}\pi_{k\cdot j}^{(d)}$, the analog of Eq. (13.5). Drawing from a Dirichlet distribution is easily accomplished using chi-squared or gamma random variables, as described in Example 6.17.

EXAMPLE 13.2. *Numerical Illustration of ML and Bayes for Monotone Bivariate Counted 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:

$$\hat{\pi}_{1+} = 190/410, \qquad \hat{\pi}_{2+} = 220/410.$$

The conditional probabilities of classification for Y_2 given Y_1 are estimated from the completely classified units:

$$\begin{array}{lll} \hat{\pi}_{1\cdot 1} = 20/90, & \hat{\pi}_{2\cdot 1} = 30/90, & \hat{\pi}_{3\cdot 1} = 40/90, \\ \hat{\pi}_{1\cdot 2} = 50/130, & \hat{\pi}_{2\cdot 2} = 60/130, & \hat{\pi}_{3\cdot 2} = 20/130. \end{array}$$

Hence the estimated probabilities (13.5) are:

$$\begin{array}{ll} \hat{\pi}_{11} = (20/90)(190/410) = 0.1030, & \hat{\pi}_{21} = (50/130)(220/410) = 0.2064 \\ \hat{\pi}_{12} = (30/90)(190/410) = 0.1545 & \hat{\pi}_{22} = (60/130)(220/410) = 0.2377 \\ \hat{\pi}_{13} = (40/90)(190/410) = 0.2060, & \hat{\pi}_{23} = (20/130)(220/410) = 0.0826. \end{array}$$

Table 13.1 Numerical Example of the Data Pattern of Figure 13.1

Complete Units						Incomplete Units		
		Y_2						
		1	2	3	Total			
Y_1	1	20	30	40	90	Y_1	1	100
	2	50	60	20	130		2	90
Total		70	90	60	220	Total		190

In contrast, estimates based on the completely-classified cases are as follows:

$$\begin{aligned} \tilde{\pi}_{11} &= 20/220 = 0.0909, & \tilde{\pi}_{21} &= 50/220 = 0.2273, \\ \tilde{\pi}_{12} &= 30/220 = 0.1364, & \tilde{\pi}_{22} &= 60/220 = 0.2727, \\ \tilde{\pi}_{13} &= 40/220 = 0.1818, & \tilde{\pi}_{23} &= 20/220 = 0.0909. \end{aligned}$$

The estimates $\{\tilde{\pi}_{jk}\}$ are less efficient than the ML estimates $\{\hat{\pi}_{jk}\}$. 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 not missing completely at random. 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 is independent of Y_1 and Y_2 Since the marginal distribution of Y_1 appears to be different for the completely and incompletely classified samples (a chi-squared test yields $\chi_1^2 = 5.23$, with associated p value < 0.01), the MCAR assumption is contradicted by these data.

For a Bayesian analysis of this example, suppose we assume the following independent Jeffreys' priors for $\{\pi_{j+}\}$ and $\{\pi_{k\cdot j}\}$:

$$p(\phi) \propto (\pi_{1+}^{-1/2}\pi_{2+}^{-1/2})(\pi_{1\cdot 1}^{-1/2}\pi_{2\cdot 1}^{-1/2}\pi_{3\cdot 1}^{-1/2})(\pi_{1\cdot 2}^{-1/2}\pi_{2\cdot 2}^{-1/2}\pi_{3\cdot 2}^{-1/2}).$$

Then the posterior distribution of these parameters is also a product of Dirichlet distributions:

$$p(\phi|\text{data}) \propto (\pi_{1+}^{189.5}\pi_{2+}^{219.5})(\pi_{1\cdot 1}^{19.5}\pi_{2\cdot 1}^{29.5}\pi_{3\cdot 1}^{39.5})(\pi_{1\cdot 2}^{49.5}\pi_{2\cdot 2}^{59.5}\pi_{3\cdot 2}^{19.5}).$$

Drawing from these distributions and setting $\pi_{jk}^{(d)} = \pi_{j+}^{(d)}\pi_{k\cdot j}^{(d)}$ 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:

$$\begin{aligned} E(\pi_{11}|\text{data}) &= 0.1044, & E(\pi_{21}|\text{data}) &= 0.2058, \\ E(\pi_{12}|\text{data}) &= 0.1549, & E(\pi_{22}|\text{data}) &= 0.2467, \\ E(\pi_{13}|\text{data}) &= 0.2045, & E(\pi_{23}|\text{data}) &= 0.0838. \end{aligned}$$

A useful feature of the Bayesian analysis is that it yields estimates of precision, for example, via the plots of the 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.

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 years, over 75 years),

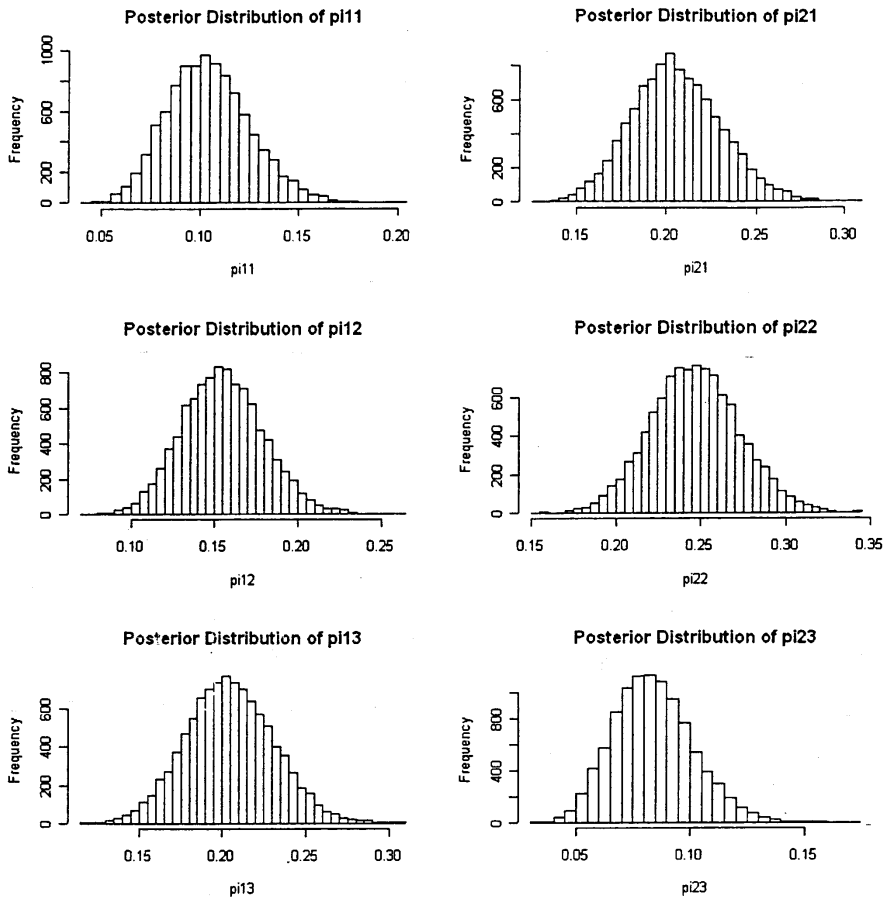


Figure 13.2. Plots of posterior distributions of the cell probabilities in Example 13.2.

P = physical status (poor, good), M = mental status (poor, good). The data on all the variables were available on 101 participants (Table 13.2a). Physical status was missing for one participant (Table 13.2b). Mental status was missing for 33 participants (Table 13.2c). Finally, physical and mental status were both missing on 29 participants (Table 13.2d).

When the information on mental status for the single observation in Table 13.2b is ignored, the data have a monotone pattern, and ML estimates of the cell probabilities can be derived using the factorization

$$\Pr(D, G, A, S, P, M) = \Pr(D, G, A, S)\Pr(P|D, G, A, S)\Pr(M|D, G, A, S, P).$$

The observed counts for estimating the three distributions on the right side are displayed in Table 13.3a. The resultant expected cell counts, which are the estimated

Table 13.2 Partially Classified Contingency Table for Example 13.3

			Male				Female			
			<75 years		>75 years		<75 years		>75 years	
Mental	Physical	Survival	E[a]	C[a]	E	C	E	C	E	C
			(a) *Fully Categorized*							
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
			(b) *Missing Physical Status*							
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
			(c) *Missing Mental Status*							
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
			(d) *Missing Both Physical and Mental Status*							
Missing	Missing	Deceased	0	1	2	2	1	0	3	1
		Survived	2	8	1	2	1	1	2	2

[a] E denotes experimental; C denotes control.

cell probabilities multiplied by the total sample size, 164, are displayed in Table 13.3b; 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

$$164\left(\frac{21}{164}\right)\left(\frac{6}{19}\right)\left(\frac{4}{6}\right) = 4.4211.$$

Table 13.3 ML Estimation for Monotone Data[a] in Table 13.2a, c, and d, Using Factored Likelihood Method

			Male				Female			
			<75 years		>75 years		<75 years		>75 years	
Mental	Physical	Survival	Exper.	Control	Exper.	Control	Exper.	Control	Exper.	Control
(a) *Partitioned Table for Monotone Patterns*										
			(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	3
		Survived	5	2	1	5	0	1	1	2
	Good	Deceased	2	1	6	8	3	1	2	4
		Survived	6	15	10	9	4	6	2	4
			(iii) *Available Information on All Variables (D, G. A. S, P, M) Given in Table 13.2a*							
(b) *Expected Cell Frequencies*										
Poor	Poor	Deceased	1.00	2.67	8.98	5.95	0.63	0.50	4.76	1.14
		Survived	1.48	0.00	0.00	0.00	0.00	0.00	0.00	2.67
	Good	Deceased	0.00	0.00	2.21	2.27	1.25	1.00	2.86	0.00
		Survived	0.00	3.68	2.95	0.00	0.00	0.00	0.00	0.00
Good	Poor	Deceased	1.00	0.00	5.39	1.98	0.63	0.50	2.38	2.28
		Survived	4.43	2.94	1.18	5.71	0.00	1.14	1.67	0.00
	Good	Deceased	2.00	1.33	4.42	6.80	2.50	0.00	0.00	4.57
		Survived	7.09	18.38	8.86	10.29	5.00	6.86	3.33	5.33

[a] The information on the mental status of the individual in Table 13.2b is ignored. *Source:* Fuchs (1982), with minor corrections.

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 cases 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 Oxspring (1974), illustrate another situation where ML estimates can be found by factoring the likelihood. Table 13.4a 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 13.4b on 400 different 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 subjects, and coarsely classified data can be collected inexpensively for a larger group of individuals.

The likelihood factorizes according to the joint distribution of the combined cell counts from the two tables, classified coarsely as in Table 13.4b, based on all 596 patients, and the conditional distribution of degree of improvement (marked, moderate, or slight) given improvement and degree of infiltration, based on the 196 patients. Resulting ML estimates of the cell probabilities are displayed in Table 13.4c, 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 $\hat{\pi}_{jk} = \hat{\pi}_{j+}\hat{\pi}_{k\cdot j}$ in Example 13.1, we write

$$\operatorname{Var} \hat{\pi}_{jk} = E(\operatorname{Var} \hat{\pi}_{jk} | \{n_{j+}\}) + \operatorname{Var}(E(\hat{\pi}_{jk} | \{n_{j+}\})),$$

where $\{n_{j+}\}$ is the set of marginal counts of Y_1. Hence

$$\begin{aligned} \operatorname{Var} \hat{\pi}_{jk} &= E\{\hat{\pi}_{j+}^2 \pi_{k\cdot j}(1 - \pi_{k\cdot j})/r_{j+}\} + \operatorname{Var}\{\hat{\pi}_{j+}\pi_{k\cdot j}\} \\ &= \pi_{j+}^2 \pi_{k\cdot j}(1 - \pi_{k\cdot j})/r_{j+} + \pi_{k\cdot j}^2 \pi_{j+}(1 - \pi_{j+})/n, \end{aligned}$$

Table 13.4 Patients Classified by Degree of Infiltration and Change of Condition

(a) *Finely Classified Data*

	Clinical Change					
	Improvement					
Degree of Infiltration	Marked	Moderate	Slight	Stationary	Worse	Totals
Little	11	27	42	53	11	144
Much	7	15	16	13	1	52
Totals	18	42	58	66	12	196

(b) *Coarsely Classified Data*

	Clinical Change			
Degree of Infiltration	Improvement	Stationary	Worse	Totals
Little	144	120	16	280
Much	92	24	4	120
Totals	236	144	20	400

(c) *ML Estimates of Cell Probabilities from (a) and (b)*

	Clinical Change				
	Improvement				
Degree of Infiltration	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).

asymptotically to order $1/r_{j+}^2$. Some algebra yields

$$\operatorname{Var}\hat{\pi}_{jk} \approx \frac{\pi_{jk}(1-\pi_{jk})}{r}\left\{1 - \frac{\pi_{k\cdot j}-\pi_{jk}}{1-\pi_{jk}}\frac{n-r}{n} + c_j\frac{1-\pi_{k\cdot j}}{1-\pi_{jk}}\right\},$$

where $c_j = r\pi_{j+}/r_{j+} - 1$. Substituting estimates of the parameters yields

$$\operatorname{Var}(\pi_{jk}-\hat{\pi}_{jk}) \approx \frac{\hat{\pi}_{jk}(1-\hat{\pi}_{jk})}{r}\left\{1 - \frac{\hat{\pi}_{k\cdot j}-\hat{\pi}_{jk}}{1-\hat{\pi}_{jk}}\frac{n-r}{n} + c_j\frac{1-\hat{\pi}_{k\cdot j}}{1-\hat{\pi}_{jk}}\right\}. \tag{13.7}$$

The left side of Eq. (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

$$\operatorname{Cov}(\pi_{jk}-\hat{\pi}_{jk}, \pi_{jl}-\hat{\pi}_{jl}) \approx \frac{-\hat{\pi}_{jk}\hat{\pi}_{jl}}{r}\left\{1 + \frac{(1-\hat{\pi}_{j+})}{\hat{\pi}_{j+}}\frac{n-r}{n} + \frac{c_j}{\hat{\pi}_{j+}}\right\}, \qquad k \neq l,$$

$$\operatorname{Cov}(\pi_{ik}-\hat{\pi}_{ik}, \pi_{jl}-\hat{\pi}_{jl}) \approx \frac{-\hat{\pi}_{ik}\hat{\pi}_{jl}}{n}, \qquad i \neq j.$$

For confidence intervals, a generally more satisfactory approach, especially when the samples are small, is to calculate asymptotic variances on a transformation of π_{jk} that better satisfies normality, for example $\operatorname{logit}(\pi_{jk})$. 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 precision 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 $\tilde{\pi}_{11} = 0.0909$ that ignores the supplementary margin has large sample variance $\pi_{11}(1-\pi_{11})/r$, which on substituting ML estimates yields

$$\operatorname{Var}(\tilde{\pi}_{11}) = \frac{(0.1030)(1-0.1030)}{220} = 0.000420. \tag{13.8}$$

Similarly, from Eq. (13.7), the ML estimate $\hat{\pi}_{11} = 0.1030$ has estimated large sample variance

$$\operatorname{Var}(\hat{\pi}_{11}) \approx 0.00042(0.9384 + 0.1151) = 0.000442. \tag{13.9}$$

Thus, the inclusion of the supplemental margin does not appear to have improved the precision. However, as noted in Example 13.2, the data do not appear to be MCAR, so $\tilde{\pi}_{11}$ is biased for π_{11}, and Eq. (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 unbiased for π_{11}

and 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

$$\widehat{\text{MSE}}(\hat{\pi}_{11}) \approx 0.0121^2 + \text{Var}(\tilde{\pi}_{11}) = 0.000566.$$

Comparing this with Eq. (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 $\text{Var}(\pi_{11}|\text{data}) = 0.000444$, 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 the posterior distribution.

13.3. ML AND BAYES ESTIMATION FOR MULTINOMIAL SAMPLES WITH GENERAL PATTERNS OF MISSING DATA

As with normal data, incomplete multinomial data that do not form a monotone data pattern require an iterative procedure for ML estimation. The EM algorithm is particularly simple, since 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 in the full table, using conditional probabilities calculated from the fully classified data. The E step of the EM algorithm for general patterns has the same function, 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 completed data. The algorithm first appeared in the statistical literature in Hartley (1958). We provide a quite general formulation of the algorithm, and then apply the algorithm to a special case.

Suppose that the hypothetical complete data are a multinomial sample of size n, with C cells, n_c observations classified in cell c, and parameters $\theta = (\pi_1, \ldots, \pi_C)$, where π_c is the classification probability for cell c. The observed data consist of r completely classified observations, with r_c belonging in cell c for $c = 1, \ldots, C$, and $r = \sum_{c=1}^{C} r_c$ and $n - r$ partially classified observations, 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, so that within each group, all units have the same set of possible cells. Suppose that partially classified observations fall in the kth group, and let S_k denote the set of cells to which these observations might belong. Furthermore, define the indicator functions $\delta(c \in S_k)$, $c = 1, \ldots, C$, $k = 1, \ldots, K$, 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, let $\{\pi_c^{(t)},\ c = 1, \ldots, C\}$ denote the current (tth iterate) estimate of the parameters. The hypothetical complete data belong to the regular exponential family with complete-data sufficient statistics

$$\{n_c, c = 1, \ldots, C\}.$$

Hence the E step consists in calculating

$$n_c^{(t)} = E\{n_c|\text{data}, \pi_1^{(t)}, \ldots, \pi_C^{(t)}\} = r_c + \sum_{k=1}^{K} m_k \psi_{c \cdot S_k}^{(t)},$$

where

$$\psi_{c \cdot S_k}^{(t)} = \pi_c^{(t)} \delta(c \in S_k) \Big/ \left[\sum_{j=1}^{C} \pi_j^{(t)} \delta(j \in S_k) \right]$$

is the current estimate of the conditional probability of falling in cell j given that an observation falls in the set of categories S_k. The E step effectively distributes the partially classified observations into the table according to these probabilities.

The M step calculates new parameter estimates as

$$\pi_c^{(t+1)} = n_c^{(t)}/n.$$

Here is a simple numerical example:

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 observations as indicated. 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

$$\hat{\pi}_{11} = 0.28, \qquad \hat{\pi}_{12} = 0.17, \qquad \hat{\pi}_{21} = 0.24, \qquad \hat{\pi}_{22} = 0.31.$$

Table 13.5 A 2×2 Table with Supplemental Margins for Both Variables

(1) Classified by Y_1 and Y_2

		Y_2 = 1	Y_2 = 2	Total
Y_1 = 1		100	50	150
Y_1 = 2		75	75	150
	Total	175	125	300

(2) Classified by Y_1

Y_1 = 1		30^a
Y_1 = 2		60^b
	Total	90

(3) Classified by Y_2

Y_2 = 1	Y_2 = 2	Total
28^c	60^d	88

Note: The superscripts a, b, c, and d refer to the partially classified cells and are used in Table 13.6.

A Bayesian analysis is analogous, except that it uses the DA algorithm to simulate values rather than EM to find ML estimates. 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 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) 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.

Table 13.6 The EM Algorithm for Data in Table 13.5, Ignoring the Missing-Data Mechanism

Estimated Probabilities	Fractional Allocation of Units
Step 1	
Y_2: 1 / 2	Y_2: 1 / 2
Y_1 = 1: 100/300 / 50/300	Y_1 = 1: $100 + 20^a + 16^c$ / $50 + 10^a + 24^d$ → 30^a
Y_1 = 2: 75/300 / 75/300	Y_1 = 2: $75 + 30^b + 12^c$ / $75 + 30^b + 36^d$ → 60^d
	28^c / 60^d
Step 2	
136/478 / 84/478	$100 + 18.6 + 15.1$ / $50 + 11.4 + 22.4$
117/478 / 141/478	$75 + 27.2 + 12.9$ / $75 + 32.8 + 37.6$
Step 3	
0.28 / 0.18	$100 + 18.4 + 15.1$ / $50 + 11.6 + 21.9$
0.24 / 0.30	$75 + 26.5 + 12.9$ / $75 + 33.5 + 38.1$
Step 4	
0.28 / 0.17	
0.24 / 0.31	

Note: The superscripts 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 $Y_1 = 2$.

The P 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 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, Shepp and Kaufman (1985) give an interesting application of 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 hypothesized to consist of B boxes or pixels, each characterized by a distinct intensity parameter $\lambda(b)$, $b = 1, \ldots, B$ governing the rate of emission. Physical considerations provide a $D \times B$ matrix of known conditional probabilities, Pr(detector = d|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 hypothetical complete data are n iid observations, $\delta_i = \{\delta_i(d, b)\}$ where $\delta_i(d, b) = 1$ if the ith emission emanated from pixel b and is recorded in detector d and zero otherwise. The observed (or incomplete) data consist of 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 $d = 1, \ldots, D$, 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 $\{\pi_{jk\ldots t}\}$, it is often important to consider 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

$$\pi_{jkl\cdots t} = \tau\tau_j^{(1)}\tau_k^{(2)}\cdots\tau_t^{(V)}, \tag{13.10}$$

for suitable multiplicative factors τ and $\{\tau_j^{(1)}\}$, $\{\tau_k^{(2)}\}, \ldots, \{\tau_t^{(V)}\}$. It is convenient to express Eq. (13.10) as a loglinear model:

$$\ln \pi_{jkl\cdots t} = \alpha + \alpha_j^{(1)} + \alpha_k^{(2)} + \cdots + \alpha_t^{(V)}, \tag{13.11}$$

where $\alpha_j^{(1)} = \ln \tau_j^{(1)}$, and so forth. Different sets of α's on the right side of Eq. (13.11) yield the same set of cell probabilities $\{\pi_{jk\cdots t}\}$, and V constraints are needed to define the α's uniquely. A common choice is to set

$$\sum_{j=1}^{J} \alpha_J^{(1)} = \cdots = \sum_{t=1}^{T} \alpha_t^{(V)} = 0.$$

Equations (13.10) or (13.11) define 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 a 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 $V =$ three-way table, we write

$$\ln \pi_{jkl} = \alpha + \alpha_j^{(1)} + \alpha_k^{(2)} + \alpha_l^{(3)} + \alpha_{jk}^{(12)} + \alpha_{jl}^{(13)} + \alpha_{kl}^{(23)} + \alpha_{jkl}^{(123)}, \tag{13.12}$$

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, respectively. The terms $\{\alpha_{ik}^{(12)}\}$, $\{\alpha_{jl}^{(13)}\}$, $\{\alpha_{kl}^{(23)}\}$ are called two-way associations between Y_1 and Y_2, Y_1 and Y_3, and Y_2 and Y_3, respectively. Finally, the terms $\{\alpha_{jkl}^{(123)}\}$ 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 Eq. (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 α_s^* between a set of factors S implies inclusion of all $(V-1)$-way and lower-order associations and main effects involving subsets of the factors in S. 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), (4), (5), (7), and (8).

ML estimation for hierarchical models varies in complexity according to 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 fitting procedure such as iterative proportional fitting (IPF) is necessary.

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}	$\{\alpha_{jkl}^{(123)}, \alpha_{jk}^{(12)}, \alpha_{jl}^{(13)}, \alpha_{kl}^{(23)}, \alpha_{j}^{(1)}, \alpha_{k}^{(2)}, \alpha_{l}^{(3)}\}$

Two asymptotically equivalent goodness-of-fit statistics are widely used to compare the fit of loglinear models. The likelihood ratio statistic is

$$G^2 = 2\sum_c n_c \ln \frac{n_c}{\hat{n}_c}, \tag{13.13}$$

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

$$X^2 = \sum_c \frac{(n_c - \hat{n}_c)^2}{\hat{n}_c}. \tag{13.14}$$

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, Fienberg, and Holland (1975), and Fienberg (1980).

EXAMPLE 13.8. *A Complete Three-Way Table.* Table 13.8a presents a 2^3 contingency table on survival of infants, previously analyzed in Bishop, Fienberg and Holland (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 13.9a 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 Tables 13.9b and c have very low values of G^2 and X^2, indicating good fits, namely, the model {SC, PC}, which indicates that survival is related to clinic, but survival and prenatal

Table 13.8 A 2^3 Contingency Table with Partially Classified Observations

Clinic (C)	Prenatal Care (P)	Survival (S): Died	Survival (S): Survived	
	(a) *Completely Classified Cases*			
A	Less	3	176	
	More	4	293	
B	Less	17	197	
	More	2	23	$r = 715$ cases
	(b) *Partially Classified Cases (Clinic Missing)*			
	Less	10	150	
	More	5	90	$m = 255$ cases

Source: (a) Bishop, Fienberg and Holland (1975), Table 1.4-2. (b) Artificial data.

Table 13.9 Estimated Cell Probabilities $\{\hat{\pi}_{jkl}\} \times 100$ from Saturated Model {SPC} and Three Loglinear Models, Fitted to Data in Table 13.8a (that is discarding the partially classified cases)

Clinic	Prenatal Care (P)	Survival (S): Died	Survival (S): Survived	Goodness-of-Fit
		(a) *Model*: {SPC}		
A	Less	0.42	24.62	
	More	0.56	40.98	
				$\text{df} = 0, G^2 = 0, X^2 = 0$
B	Less	2.38	27.55	
	More	0.28	3.22	
		(b) *Model:* {SP, SC, PC}		
A	Less	0.39	24.64	
	More	0.59	40.95	
				$\text{df} = 1, G^2 = 0.04, X^2 = 0.04$
B	Less	2.41	27.52	
	More	0.25	3.24	
		(c) *Model:* {SC, PC}		
A	Less	0.36	24.67	
	More	0.62	40.92	
				$\text{df} = 2, G^2 = 0.08, X^2 = 0.08$
B	Less	2.38	27.55	
	More	0.28	3.22	
		(d) *Model:* {SP, SC}		
A	Less	0.76	35.51	
	More	0.22	30.08	
				$\text{df} = 2, G^2 = 188.1, X^2 = 169.5$
B	Less	2.04	16.66	
	More	0.62	14.11	

care are not associated conditional on clinic, and the model {SC, PC, SP}, which adds the association SP to the previous model. Since 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 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. These constraints can increase the computation required to compute ML estimates for two reasons. First, factorizations of the likelihood for monotone patterns do not necessarily lead to explicit ML estimates, since the parameters in the factors are not necessarily distinct. Second, the M step of the EM algorithm for nonmonotone patterns may itself involve iteration.

The standard fitting algorithm in that case 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, Fienberg and Holland, 1975, Chapter 3). The method is described in 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. Since IPF increases the likelihood at each iteration, replacing the M step by a single iteration of IPF yields a generalized EM algorithm; Meng and Rubin (1991) show that it is also an ECM algorithm, and hence shares asymptotic properties similar to EM.

For Bayesian analyses, the imputation (I) step of DA is unaffected by the model restrictions, and allocates each partially classified counts into the set of possible cells as draws from a multinomial distribution, with probabilities from the previous posterior (P) step. The P step consists in generating a draw from the complete-data posterior distribution of the constrained parameters, with data filled in from the I step. In models with explicit complete-data ML estimates, the joint distribution factorizes into components with unrestricted multinomial parameters. If independent Dirichlet priors are assumed for these factors, the posterior distributions are also Dirichlet, and the P step consists of drawing from these distributions. In the case of three-way tables 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 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, a Bayesian analog of IPF. The CM steps of IPF are replaced by analogous CP (conditional posterior) steps, which are Dirichlet draws of sets of loglinear model parameters of 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 $t \to \infty$:

EXAMPLE 13.9. *Bayesian IPF for the No Three-Way Association Model for a 2×2×2 Table. (Example 8.7 continued).* Suppose we have complete data in a $2 \times 2 \times 2$ table, and $\{y_{ij+}\}$, $\{y_{i+k}\}$, and $\{y_{+jk}\}$ are 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:

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 Eq. (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:

CP2: $$\theta_{ijk}^{(t+2/3)} = \theta_{i(j)k}^{(t+1/3)}(g_{i+k}^{(t+2/3)}/g_{+++}^{(t+2/3)})$$

CP3: $$\theta_{ijk}^{(t+3/3)} = \theta_{ij(k)}^{(t+2/3)}(g_{+jk}^{(t+2/3)}/g_{+++}^{(t+3/3)}),$$

where $\{g_{i+k}^{(t+2/3)}\}$ are draws from the Dirichlet distribution

$$p(\theta_{i+k}|\theta_{ij+}^{(t)}, \theta_{+jk}^{(t)}, Y^{(t)}) \propto \prod_{i=1}^{2}\prod_{k=1}^{2} \theta_{i+k}^{(\alpha_{i+k}+y_{i+k}^{(t)}-1)},$$

and $\{g_{+jk}^{(t+3/3)}\}$ are draws from the Dirichlet distribution

$$p(\theta_{+jk}|\theta_{ij+}^{(t)}, \theta_{i+k}^{(t)}, Y^{(t)}) \propto \prod_{j=1}^{2}\prod_{k=1}^{2} \theta_{ij+}^{(\alpha_{+jk}+y_{+jk}^{(t)}-1)}.$$

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 13.8b are added to the data in Table 13.8a analyzed in Example 13.8. Survival (S) and Prenatal Care (P) are recorded in the supplemental data, but Clinic (C) is not recorded. The resulting incomplete data form a monotone pattern with P and S more observed than C.

The likelihood for the combined data in Tables 13.8a and b factors into a term for the distribution of SP, involving all $r + m = 970$ cases, and a term for the distribution of C given SP, involving the $m = 715$ completely classified cases. 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 13.10a shows ML estimates of $100\pi_{jkl}$ for the saturated model {SPC}, calculated by the methods of Section 13.2. Tables 13.10b and c show ML estimates for {SP, SC, PC} and {SP, SC}. Since 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 cases. 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. 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.

In the preceding example starting values for the EM algorithm were based on analysis of the completely classified table. With sparse tables containing zero cells,

Table 13.10 ML Estimates for Models {SPC}, {SP, SC, PC}, and {SP, SC} Fitted to Data in Tables 13.8a and b

		Survival	
Clinic	Prenatal Care	Died	Survived
(a) *Model:* {SPC}			
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 = 100.00
(b) *Model:* {SP, SC, PC}			
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 = 100.00
(c) *Model:* {SP, SC}			
A	Less	100(5.4/20)(30/970) = 0.84	100(253.9/373)(523/970) = 36.70
	More	100(l.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 = 100.00

Table 13.11 ML Estimates for Model {SC, PC} Fitted to Data in Tables 13.8a and b, via the EM Algorithm

M Step: Estimated Cell Probabilities×100					E Step: Filled-In Cell Counts	
			Survival		Survival	
Iteration	Clinic	Prenatal	Died	Survival	Died	Survived
I	A	Less	0.36	24.67	$3 + (10)(0.36)/2.74 = 4.33$	$176 + 150(24.62)/52.22 = 246.86$
		More	0.62	40.92	$4 + 5(0.62)/0.90 = 7.44$	$293 + 90(40.92)/44.14 = 376.44$
	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

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.

13.4.3. Goodness-of-Fit Tests for Partially Classified Data

Chi-squared goodness-of-fit statistics analogous to Eqs. (13.13) and (13.14) 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 X^2 and G^2 are obtained for the saturated model ({SPC} in Example 13.10); the values of X^2 and 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 (or X^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 chi-squared test for the restricted model with complete data.

This procedure appears to assume that data are MCAR, but in fact the tests remain valid provided the missing data are MAR. Under the latter assumption the component of the loglikelihood for the missing-data mechanism cancels out when the values of G^2 (or X^2) for the two models are subtracted.

EXAMPLE 13.11. *Goodness-of-Fit Statistics for Incomplete Three-Way Table. (Example 13.10 continued).* Goodness-of-fit statistics for the saturated model {SPC} in Example (13.10) are

$$X^2(\text{SPC}) = 7.96, \qquad G^2(\text{SPC}) = 7.80, \qquad \text{df} = 3.$$

To calculate degrees of freedom, 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 $\text{df} = 11 - 7 - 1 = 3$. Since 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 less than 0.05 using χ^2, and approximately 0.05 using G^2. The unsaturated models yield

$$\begin{aligned} X^2(\text{SP, SC, PC}) &= 7.99, \quad G^2(\text{SP, SC, PC}) = 7.84, \quad \text{df} = 11 - 6 - 1 = 4, \\ X^2(\text{SP, PC}) &= 8.29, \quad G^2(\text{SP, PC}) = 7.84, \quad \text{df} = 11 - 5 - 1 = 5, \\ X^2(\text{SP, SC}) &= 178.55, \quad G^2(\text{SP, SC}) = 195.92, \quad \text{df} = 11 - 5 - 1 = 5. \end{aligned}$$

Subtracting the chi-squared values for the saturated model yields

$$\Delta X^2(\text{SP, SC, PC}) = 0.03, \quad \Delta G^2(\text{SP, SC, PC}) = 0.04, \quad \Delta\text{df} = 8 - 6 - 1 = 1,$$
$$\Delta X^2(\text{SP, PC}) = 0.33, \quad \Delta G^2(\text{SP, PC}) = 0.20, \quad \Delta\text{df} = 8 - 5 - 1 = 2,$$
$$\Delta X^2(\text{SP, SC}) = 170.59, \quad \Delta G^2(\text{SP, SC}) = 188.12, \quad \Delta\text{df} = 8 - 5 - 1 = 2,$$

which can be compared with the goodness-of-fit statistics based on the completely classified cases in Table 13.9. We conclude as before that {SP, PC} is the preferred model.

PROBLEMS

13.1. Show that for complete data the Poisson and multinomial models for multiway counted 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 variances for the likelihood (13.1). (Hint: Remember the constraint that the cell probabilities sum to 1.)

13.3. Verify the results of the chi-squared 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 13.3b, and compare the answers with those obtained from complete cases.

13.6. Suppose that in Example 13.3 there are no cases with pattern d. Which parameters are inestimable, in that they do not appear in the likelihood? Estimate the cell probabilities, assuming specific values for the inestimable parameters. (See Section 7.5.)

13.7. State in words the assumption about the missing-data mechanism under which the estimates in Table 13.4c 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. Compute 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 cases. Are they identical?

13.12. Show that in Example 13.10, the factors 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 13.8b 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.

(a) Show that c_j in Eq. (13.7) is of smaller order than other terms in the expression.

(b) Show that Eq. (13.7) is asymptotically equal to

$$\mathrm{Var}(\pi_{jk} - \hat{\pi}_{jk}) \approx \frac{\hat{\pi}_{jk}(1-\hat{\pi}_{jk})}{r}\left[1 - \frac{\hat{\pi}_{k\cdot j} - \hat{\pi}_{jk}}{1-\hat{\pi}_{jk}}\frac{n-r}{n}\right].$$

Hence, state the proportionate reduction in variance of $\hat{\pi}_{jk}$ over the complete-case estimate, and describe situations where it is large and small. (The analogous situation for normal data is discussed in Section 7.2.2.)

CHAPTER 14

Mixed Normal and Non-normal Data with Missing Values, Ignoring the Missing-Data Mechanism

14.1. INTRODUCTION

In Chapters 11 and 12 we considered a variety of missing-data models for continuous variables, based on the multivariate normal distribution and longer-tailed distributions. The role of categorical variables was confined to that of fully observed covariates in regression models. In Chapter 13 we discussed models for categorical variables with missing values. In this chapter we consider missing-data methods for mixtures of normal and non-normal variables, assuming that the missing-data mechanism is ignorable.

Little and Schluchter (1985) discuss a model for missing data with mixed normal and categorical variables and provide relatively simple and computationally feasible EM algorithms with missing 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 hypothetical 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 $C = \Pi_{j=1}^{V} I_j$ cells. For subject 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 E_c if case i belongs to cell c of the contingency table, where E_c is a $(1 \times C)$ vector with 1 as the cth entry and 0s elsewhere.

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 iid multinomial random variables with cell probabilities

$$\Pr(w_i = E_c) = \pi_c, \qquad c = 1, \ldots, C; \Sigma\pi_c = 1. \tag{14.1}$$

2. Given that $w_i = E_c$,

$$(x_i | w_i = E_c) \sim_{\text{ind}} N_K(\mu_c, \Omega), \tag{14.2}$$

the K-variate normal distribution with mean $\mu_c = (\mu_{c1}, \ldots, \mu_{cK})$ and covariance matrix Ω. We write $\Pi = (\pi_1, \ldots, \pi_C)$ for the $(1 \times C)$ vector of cell probabilities and $\Gamma = \{\mu_{ck}\}$ for the matrix of cell means. There are $C - 1 + KC + \frac{1}{2}K(K+1)$ parameters in the model, $\theta = (\Pi, \Gamma, \Omega)$.

The following properties of this model are worth noting:

(i) 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.

(ii) 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 ML and Bayes estimation for partially classified contingency tables, as discussed in Chapter 13.

(iii) The within-cell covariance matrix Ω is assumed the same across all the cells of the contingency table. This is an important assumption of the basic model, although it can be relaxed, as noted in Section 14.5.

(iv) If a particular binary variable (say Y_1), with values 1 and 0, is chosen as a dependent variable, then the conditional distribution of Y_1, given the other variables, is Bernoulli with $\Pr(Y_1 = 1) = e^L/(1 + e^L)$, 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 model 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).

(v) If a particular continuous variable (say X_1) is chosen as a dependent variable, then a normal linear regression model results. That is, the conditional distribution of X_1 given the other variables is normal, with mean given by 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 $\theta = (\Pi, \Gamma, \Omega)$ 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

$$\begin{aligned}\ell(\Gamma, \Omega, \Pi) &= \sum_{i=1}^{n} \ln f(x_i|w_i, \Gamma, \Omega) + \sum_{i=1}^{n} \ln f(w_i|\Pi) \\ &= h(\Omega) - \tfrac{1}{2}\mathrm{tr}\left(\Omega^{-1}\sum_{i=1}^{n} x_i^{\mathrm{T}} x_i\right) + \mathrm{tr}\,\Omega^{-1}\Gamma\left(\sum_{i=1}^{n} w_i^{\mathrm{T}} x_i\right) \\ &\quad + \sum_{c=1}^{C}\left[\left(\sum_{i=1}^{n} w_{ic}\right)(\ln \pi_c - \tfrac{1}{2}\mu_c \Omega^{-1}\mu_c^{\mathrm{T}})\right],\end{aligned} \tag{14.3}$$

where w_{ic} is the cth component of w_i, tr means "trace of the matrix," and $h(\Omega) = -\frac{1}{2}n[K\ln(2\pi) + \ln|\Omega|]$. Maximizing Eq. (14.3) yields complete-data ML estimates

$$\begin{aligned}\hat{\Pi} &= n^{-1}\sum_{i=1}^{n} w_i, \\ \hat{\Gamma} &= (\sum_{i=1}^{n} x_i^{\mathrm{T}} w_i)(\sum_{i=1}^{n} w_i^{\mathrm{T}} w_i)^{-1}, \\ \hat{\Omega} &= n^{-1}\sum_{i=1}^{n}(x_i - w_i\hat{\Gamma})^{\mathrm{T}}(x_i - w_i\hat{\Gamma}),\end{aligned} \tag{14.4}$$

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 Xs and Ws are missing. For subject i, let $x_{\mathrm{obs},i}$ denote the vector of observed continuous variables, $x_{\mathrm{mis},i}$ denote the vector of missing continuous variables, and S_i denote the set of cells in the contingency table where subject i could lie, given the observed categorical variables. We now consider the EM algorithm for ML estimation of θ given data $\{x_{\mathrm{obs},i}, S_i: i = 1, \ldots, n\}$.

The density (14.3) belongs to the regular exponential family with complete-data sufficient statistics $\sum x_i^{\mathrm{T}} x_i$, $\sum w_i^{\mathrm{T}} x_i$, and $\sum w_i$, which are, respectively, the raw sum of squares and cross-products of the Xs, the cell totals of the Xs, 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 $\{x_{\mathrm{obs},i}, S_i: i = 1, \ldots, n\}$ and current parameter estimates $\theta^{(t)} = (\Pi^{(t)}, \Gamma^{(t)}, \Omega^{(t)})$. The contributions from case i are

E Step:

$$T_{1i}^{(t)} = E(x_i^{\mathrm{T}} x_i | x_{\mathrm{obs},i}, S_i, \theta^{(t)}), \tag{14.5}$$

$$T_{2i}^{(t)} = E(w_i^{\mathrm{T}} x_i | x_{\mathrm{obs},i}, S_i, \theta^{(t)}), \tag{14.6}$$

$$T_{3i}^{(t)} = E(w_i | x_{\mathrm{obs},i}, S_i, \theta^{(t)}). \tag{14.7}$$

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:

$$\begin{aligned}
\Pi^{(t+1)} &= n^{-1} \sum_{i=1}^{n} T_{3i}^{(t)}, \\
\Gamma^{(t+1)} &= D^{-1} \left(\sum_{i=1}^{n} T_{2i}^{(t)} \right), \\
\Omega^{(t+1)} &= n^{-1} \left[\sum_{i=1}^{n} T_{1i}^{(t)} - \left(\sum_{i=1}^{n} T_{2i}^{(t)} \right)^{\mathrm{T}} D^{-1} \left(\sum_{i=1}^{n} T_{2i}^{(t)} \right) \right],
\end{aligned} \tag{14.8}$$

where D 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 Eqs. (14.5)–(14.7) with the new parameter estimates, and cycles back and forth 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_k and V_k are standardized reading and verbal comprehension scores for the kth child in a family, $k = 1, 2$. The variable G is always observed, but the other variables are missing in a variety of different combinations.

An analysis of the missing-data pattern suggests that all the parameters of the general location model are estimable, despite the sparseness of the data matrix. For example, even R_2 is not observed in the fully classified table when $G = 1$, $D_1 = 2$, $D_2 = 1$, there are five other families with R_2 measured that could possibly be in that cell, and these observations provide the information to estimate the mean of R_2 in that cell.

ML estimates computed using the EM algorithm under the unrestricted model are displayed in Table 14.1 (Model A). The maximized loglikelihood under the

Table 14.1 Maximum Likelihood Estimates for Data in Table 11.1[a]

(a) *Expected Frequencies and Cell Means*

Cell			Expected Frequencies		Cell Means							
					R_1		R_2		V_1		V_2	
G	D_1	D_2	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)
A	13.2	11.9	20.7	24.1	0.701	0.832	0.825	0.663	0.835	0.885
B	13.1	11.9	20.1	23.3	0.685	0.832	0.822	0.654	0.836	0.881

[a] A = model with no restrictions on means or cell probabilities. B = model with no restrictions on means, cell probabilities restricted so that D_1 and D_2 are independent of G.

unrestricted model is -872.73. Perhaps because of the relatively high 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 the algorithm. 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, since 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 $\{T_{1i}^{(t)}, T_{2i}^{(t)}, T_{3i}^{(t)}, \quad i = 1, \ldots, n\}$ 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 $E(w_i|x_{\text{obs},i}, S_i, \theta^{(t)})$ for each subject $i = 1, \ldots, n$. The cth component of this vector will be denoted $\omega_{ic} = \Pr(w_i = E_c|x_{\text{obs},i}, S_i, \theta^{(t)})$. That is, for $c = 1, \ldots, C$, ω_{ic} is the conditional posterior probability that subject i belongs in cell c, given the observed continuous variables $x_{\text{obs},i}$, the knowledge that subject i is restricted to be in one of the cells in S_i, and $\theta = \theta^{(t)}$. This is positive when $c \in S_i$, where it takes the form

$$\omega_{ic} = \exp(\delta_{ic}) / \sum_{c' \in S_i} \exp(\delta_{ic'}), \tag{14.9}$$

where

$$\delta_{ic} = x_{\text{obs},i} \Omega_{\text{obs},i}^{-1} \mu_{\text{obs},i,c}^{\mathrm{T}} - \tfrac{1}{2} \mu_{\text{obs},i,c} \Omega_{\text{obs},i}^{-1} \mu_{\text{obs},i,c}^{\mathrm{T}} + \ln(\pi_c) \tag{14.10}$$

and $\mu_{\text{obs},i,c}$ and $\Omega_{\text{obs},i}$ are the mean and covariance matrix in cell c of the continuous variables $x_{\text{obs},i}$ present for subject i.

To calculate $T_{1i}^{(t)}$ and $T_{2i}^{(t)}$, write the continuous variables for subject i as $\{x_{ij}, j = 1, \ldots, K\}$. If x_{ij} is missing, define $\hat{x}_{ij}^{(c)} = E(x_{ij}|x_{\text{obs},i}, w_i = E_c, \theta^{(t)})$, the predicted value of x_{ij} from the regression in cell c of X_j on $x_{\text{obs},i}$, evaluated at $\theta = \theta^{(t)}$. The element in the cth row and jth column of $T_{2i}^{(t)}$, for $c = 1, \ldots, C$ and $j = 1, \ldots, K$, is obtained by multiplying x_{ij} or its estimate by the conditional posterior probability that subject i falls in cell c:

$$E(w_{ic}x_{ij}|x_{\text{obs},i}, S_i, \theta^{(t)}) = \begin{cases} \omega_{ic}\hat{x}_{ij}^{(c)} & \text{if } x_{ij} \text{ is missing,} \\ \omega_{ic}x_{ij} & \text{if } x_{ij} \text{ is present.} \end{cases}$$

When both x_{ij} and x_{ik} are missing, let $\sigma_{jk\cdot\text{obs},i}$ denote the conditional covariance of x_{ij} and x_{ik} given $x_{\text{obs},i}$ and given that $w_i = E_c$. Then the jk element of $T_{1i}^{(t)}$, for $j, k = 1, \ldots, K$ is

$$\begin{aligned} E(x_{ij}x_{ik}|x_{\text{obs},i}, S_i, \theta^{(t)}) &= \sum_{c \in S_i} \omega_{ic} E(x_{ij}x_{ik}|x_{\text{obs},i}, w_i = E_c, \theta^{(t)}) \\ &= \begin{cases} x_{ij}x_{ik}, & x_{ij}, x_{ik} \text{ both present;} \\ x_{ik} \sum_{c \in S_i} \omega_{ic}\hat{x}_{ij}^{(c)}, & x_{ij} \text{ missing, } x_{ik} \text{ present;} \\ x_{ij} \sum_{c \in S_i} \omega_{ic}\hat{x}_{ik}^{(c)}, & x_{ik} \text{ missing, } x_{ij} \text{ present;} \\ \sigma_{jk\cdot\text{obs},i} + \sum_{c \in S_i} \omega_{ic}\hat{x}_{ij}^{(c)}\hat{x}_{ik}^{(c)}, & x_{ij}, x_{ik} \text{ both missing.} \end{cases} \end{aligned}$$

The computations are easily performed by sweep operations, discussed in Section 7.4.3. Consider the matrix

$$H = \begin{bmatrix} \hat{\Omega}_{\text{obs},i} & \hat{\Omega}_{\text{cov},i} & \hat{\Gamma}_{\text{obs},i} \\ \hat{\Omega}^{\text{T}}_{\text{cov},i} & \hat{\Omega}_{\text{mis},i} & \hat{\Gamma}_{\text{mis},i} \\ \hat{\Gamma}^{\text{T}}_{\text{obs},i} & \hat{\Gamma}^{\text{T}}_{\text{mis},i} & P \end{bmatrix},$$

where P is a $C \times C$ diagonal matrix, having cth diagonal element equal to $2 \ln \pi_c$, for $c = 1, \ldots, C$,

and

$$\hat{\Omega} = \begin{bmatrix} \hat{\Omega}_{\text{obs},i} & \hat{\Omega}_{\text{cov},i} \\ \hat{\Omega}^{\text{T}}_{\text{cov},i} & \hat{\Omega}_{\text{mis},i} \end{bmatrix}$$

and

$$\hat{\Gamma} = [\hat{\Gamma}_{\text{obs},i}\ \hat{\Gamma}_{\text{mis},i}]$$

are current estimates of Ω and Γ, partitioned according to the observed and missing X variables in case i. Sweeping on the elements of H corresponding to present Xs yields

$$\text{SWP}[x_{\text{obs},i}]H = \begin{bmatrix} G_{11} & G_{12} & G_{13} \\ G^{\text{T}}_{12} & G_{22} & G_{23} \\ G^{\text{T}}_{13} & G^{\text{T}}_{23} & G_{33} \end{bmatrix},$$

where $G_{11} = -\hat{\Omega}^{-1}_{\text{obs},i}$; $G_{12} = \hat{\Omega}^{-1}_{\text{obs},i}\hat{\Omega}_{\text{cov},i}$ are regression coefficients of the missing Xs on $x_{\text{obs},i}$; $G_{22} = \hat{\Omega}_{\text{mis},i} - \hat{\Omega}^{\text{T}}_{\text{cov},i}\hat{\Omega}^{-1}_{\text{obs},i}\hat{\Omega}_{\text{cov},i}$ contains the residual variances and covariances $\sigma_{jk\cdot\text{obs},i}$ for x_{ij}, $x_{ik} \in x_{\text{obs},i}$, $G_{13} = \hat{\Omega}^{-1}_{\text{obs},i}\hat{\Gamma}_{\text{obs},i}$ yields the coefficients of $x_{\text{obs},i}$ in the linear discriminant function (14.10); and the cth diagonal element of $\frac{1}{2}G_{33} = \frac{1}{2}P - \frac{1}{2}\hat{\Gamma}^{\text{T}}_{\text{obs},i}\hat{\Omega}^{-1}_{\text{obs},i}\hat{\Gamma}_{\text{obs},i}$ equals the sum of the second and third terms on the right side of Eq. (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 Eq. (14.9). Considerable savings in computation may be obtained if subjects with the same pattern of missing Xs are grouped together to avoid unnecessary sweep operations.

14.2.4. Bayes Computations 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 (Schafer, 1997). The I and P steps of data augmentation parallel the E and M steps of EM. 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 observation i involves two steps, say I1 and I2: I1 imputes the missing categorical variables, which corresponds to assigning the case to a particular cell of the contingency table formed by the categorical variables. Specifically, observation 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 $x_{\text{mis},i}$ from the conditional multivariate normal distribution of $x_{\text{mis},i}$ given $x_{\text{obs},i}$, the cell c determined by the I1 step, and $\theta^{(t)}$. These steps create a filled-in data set $Y^{(t)}$.

The P step computes new draws $\theta^{(i+1)}$ 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:

$$p(\Pi|\{Y^{(t)}\}) = \prod_{c=1}^{C} \pi_c^{n_c^{(t)}-1/2}, \tag{14.11}$$

where $n_c^{(t)}$ is the number of observed or imputed observations 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:

$$(\Omega|\Pi^{(t+1)}, Y^{(t)}) \sim \text{Inv-Wishart}(S^{(t)}, n - C), \tag{14.12}$$

where $S^{(t)}$ is the pooled within-cell covariance matrix of the continuous variables from the filled-in data. The new $\mu_c^{(t+1)}$ $(c = 1, \ldots, C)$ is drawn from the posterior distribution of μ_c given $\Omega^{(t+1)}$ and $Y^{(t)}$, which is multivariate normal:

$$(\mu_c|\Pi^{(t+1)}, \Omega^{(t+1)}, Y^{(t)}) \sim N_K(\bar{y}_c^{(t)}, \Omega^{(t+1)}, n_c^{(t)}), \tag{14.13}$$

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, and can be compared with the ML estimates in Table 14.1. Figure 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 similar results for four transformed correlation 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 sequences and histograms of draws look reasonably stable for the means, but the sequences for the covariances display some jumpiness, reflecting severe lack of information to estimate some of these parameters. We prefer the

Table 14.2 Posterior Means and Standard Deviations of Parameters from Data Augmentation Applied to the Data in Table 11.1, Unrestricted General Location Model

(a) *Expected Frequencies and Cell Means*

Cell			Expected Frequencies		Cell Means							
					R_1		R_2		V_1		V_2	
G	D_1	D_2	*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

Bayesian results to ML since they tend to average over plausible regions of the likelihood, and display 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 data set.

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 more general model that allows ANOVA-like restrictions on the $\{\mu_c\}$, and models the $\{\pi_c\}$ by a restricted loglinear model. This more general model is considered for complete-data discriminant analysis by Krzanowski (1980, 1982).

14.3.2. Restricted Models for the Cell Means

For $u \leqslant C$, let z_i be a $1 \times u$ vector of design variables for case i, which can be obtained from the cell indicator vector w_i as $z_i = w_i A$ where A is a known $C \times u$

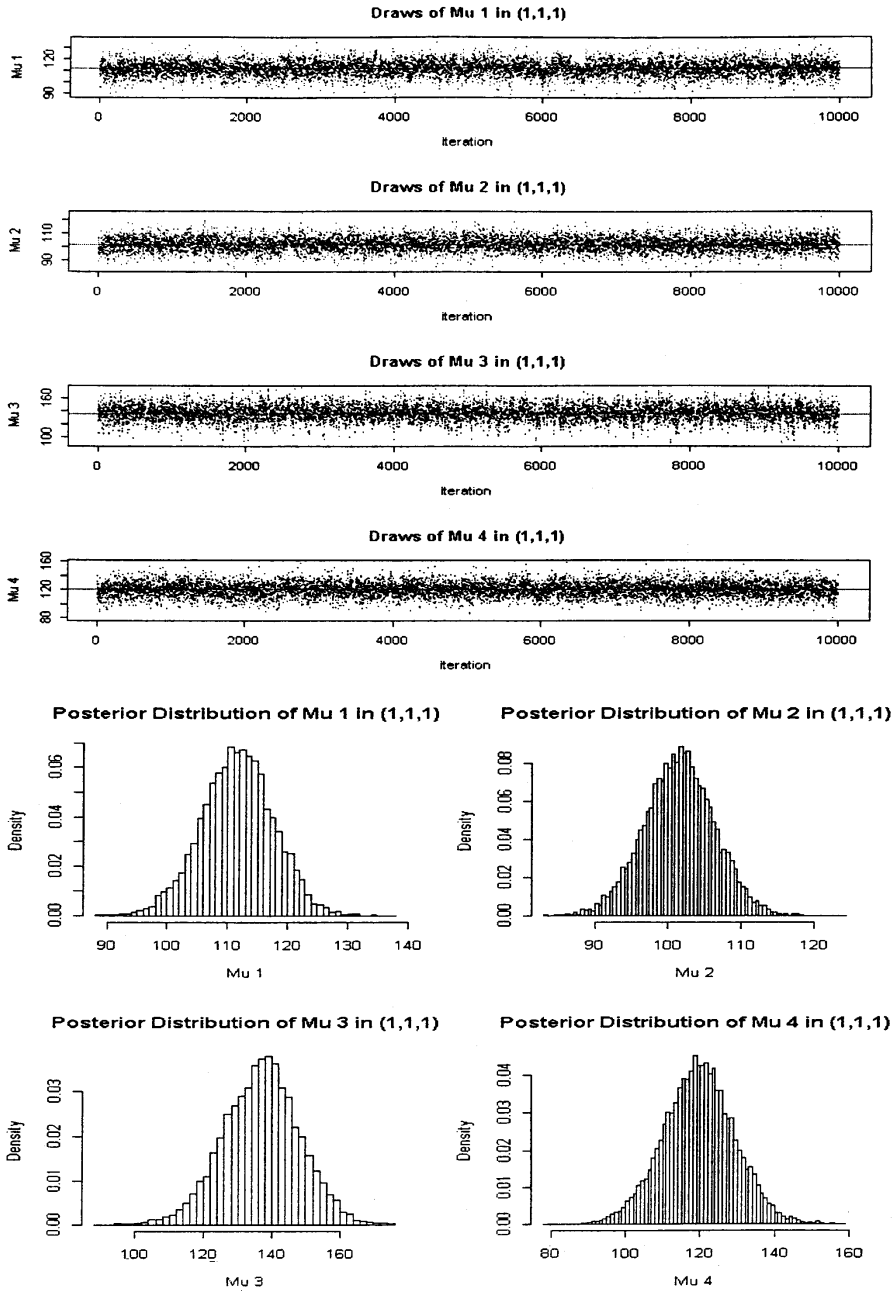


Figure 14.1A. Gibbs' sequences and posterior distributions of the means ($1 = R_1$, $2 = R_2$, $3 = V_1$, $4 = V_2$) in cell (1,1,1). Lines in sequences are the corresponding Gibbs' means.

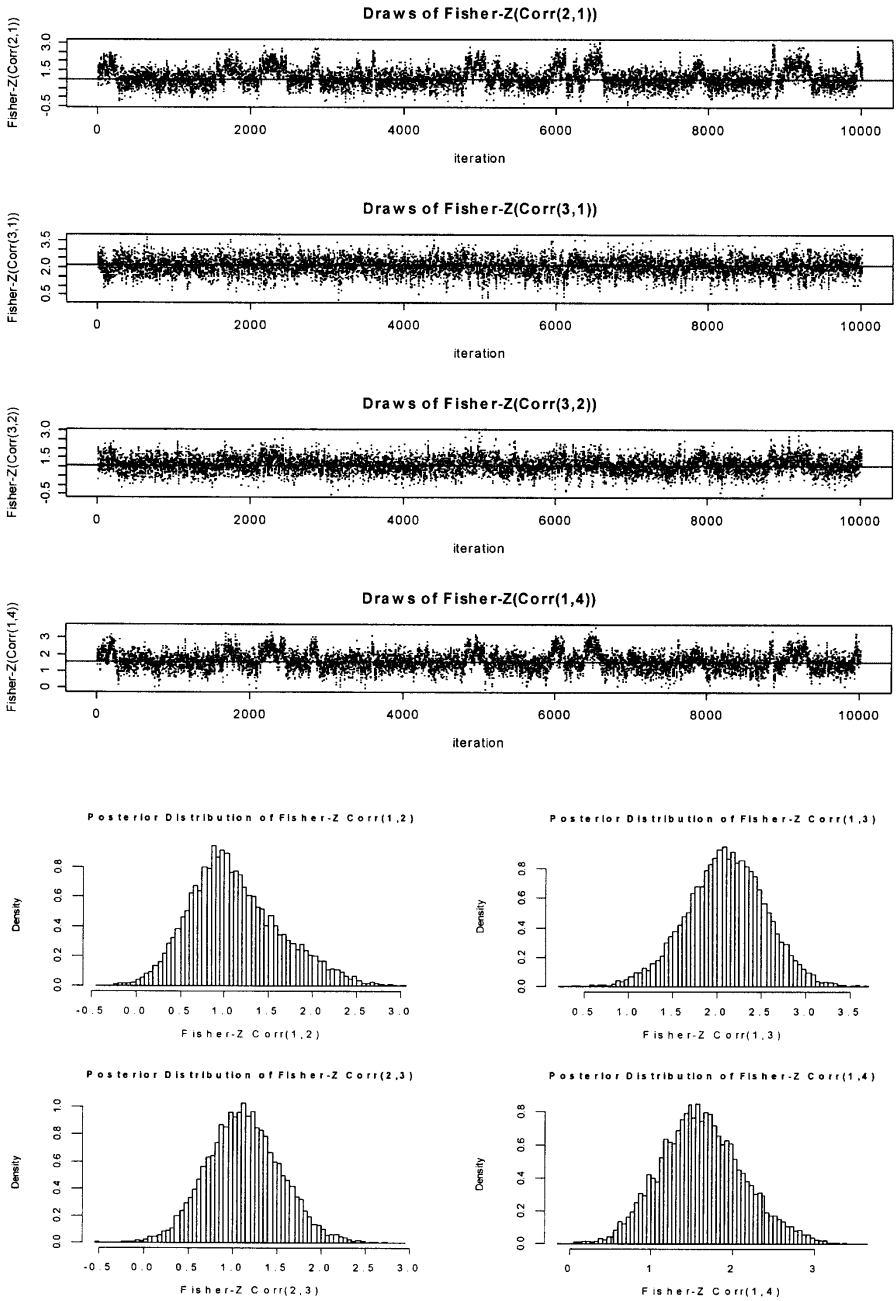


Figure 14.1B. Gibbs' sequences posterior distributions of Fisher Z transforms of selected correlations. Lines in series of draws are corresponding Gibbs' mean.

matrix that represents the chosen design. 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 $f(x_i|w_i) \sim N_k(z_iB, \Omega)$, where B is a $(u \times k)$ matrix of unknown parameters. Note that $E(x_i|w_i) = w_iAB$, 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 dimensionality of the model 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 $j = 1, \ldots, I_1$, $k = 1, \ldots, I_2$, and $l = 1, \ldots, I_3$. 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 zero. See Section 13.4 for more details.

14.3.4. Modifications to the Algorithms of Sections 14.2.2 and 14.2.3 for Parameter Restrictions

For a general V-way table with $C = \Pi_{j=1}^{V} I_j$ cells, let α denote the nonzero α terms in the loglinear model, and write $\pi_c(\alpha)$ for the constrained probability of falling in cell c, $c = 1, \ldots, C$. 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 and 14.3.3, let $\alpha^{(0)}$, $\Omega^{(0)}$, and $B^{(0)}$ be initial estimates of the parameters, perhaps calculated from complete cases. Also let $\Gamma^{(0)} = AB^{(0)}$, where A is a known design matrix, and $\pi_c^{(0)} = \pi_c(\alpha^{(0)})$, $c = 1, \ldots, C$. The restricted models of Sections 14.3.2 and 14.3.3 lie in the regular exponential family, with complete-data minimal sufficient statistics $\sum x_i^{\mathrm{T}} x_i$, $\sum w_i^{\mathrm{T}} w_i A$ and linear combinations of the counts $\sum w_i$ determined by the margins fitted in the loglinear model. Since 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 forms 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)]. This table contains 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 the 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., Anderson, 1965, Chapter 8) are $\hat{B} = (\sum z_i^{\mathrm{T}} z_i)^{-1} \sum z_i^{\mathrm{T}} x_i$ and $\hat{\Omega} = n^{-1} \sum (x_i - z_i \hat{B})^{\mathrm{T}} (x_i - z_i \hat{B})$. The M step estimates of B and Ω are obtained by writing $z_i = w_i A$, in the preceding equations for $\hat{B}$ and $\hat{\Omega}$, and then replacing $\sum x_i^{\mathrm{T}} x_i$, $\sum w_i^{\mathrm{T}} x_i$, and $\sum w_i^{\mathrm{T}} w_i$ 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

$$B^{t+1} = (A^T D^{(t)} A)^{-1} A^T \left(\sum T_{2i}^{(t)} \right), \tag{14.14}$$

$$\Gamma^{(t+1)} = A B^{(t+1)}, \tag{14.15}$$

and

$$\Omega^{(t+1)} = n^{-1} \left[\sum_{i=1}^{n} T_{1i}^{(t)} - \left(\sum_{i=1}^{n} T_{2i}^{(t)} \right)^{\mathrm{T}} A (A^{\mathrm{T}} D^{(t)} A)^{-1} A^{\mathrm{T}} \left(\sum_{i=1}^{n} T_{2i}^{(t)} \right) \right]. \tag{14.16}$$

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 Eqs. (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 inv-Wishart distribution, as in the unrestricted case given by Eq. (14.12), but with $S^{(t)}$ replaced by the right side of Eq. (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 Eq. (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 general location model was fitted to the data in Table 11.1. The model is somewhat overparameterized, with 69 parameters for 69 incomplete cases. In this section we fit and test models with fewer parameters that correspond to hypotheses of substantive interest. In particular, suppose we wish to test the hypothesis that the occurrence of adverse psychiatric symptoms in children is unrelated to the risk group of the parent. This hypothesis implies that

$$\pi_{jkl} = \pi_{j++} \pi_{+kl}, \qquad j = 1, 2, 3; \; k, l = 1, 2,$$

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 chi-square statistic for testing whether D_1 and D_2 are independent of G was thus $2(-872.73 + 877.64) = 9.82$ with six degrees of freedom, suggesting no evidence of lack of fit. Another local maximum (loglikelihood -877.72) was found for this model.

In a 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 are set to zero, with the same restricted model for the cell probabilities. The restrictions on the means of continuous variables can be written as $E(x_i|z_i) = z_i B$, where B is a 6×4 matrix of parameters, and $z_i = w_i A$, where

$$A^{\mathrm{T}} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & -1 & -1 & -1 & -1 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 \\ 1 & 0 & 0 & -1 & 1 & 0 & 0 & -1 & 1 & 0 & 0 & -1 \\ 0 & 1 & 0 & -1 & 0 & 1 & 0 & -1 & 0 & 1 & 0 & -1 \\ 0 & 0 & 1 & -1 & 0 & 0 & 1 & -1 & 0 & 0 & 1 & -1 \end{bmatrix},$$

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 the 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. The global maximized loglikelihood for this model was -910.46, so that the likelihood ratio chi-square testing the fit of this model versus the full model was $\chi^2 = 75.46$ ($\mathrm{df} = 30$), 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 zero, with the same restriction on cell probabilities. This model showed evidence of lack of fit when compared to the full model ($\chi^2 = 59.39$, $\mathrm{df} = 14$). P values were not reported because they are inappropriate given the multiple maxima of the likelihood. Nonetheless, 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 the Categorical Variables are More Observed than the 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 present in cases where one or more of the continuous variables are present. The incomplete-data likelihood then

Variables						
Cases	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 simpler ML estimates. 0 = Observed; 1 = Missing; × = Observed or missing. *Source:* Little and Schluchter (1985).

factorizes into the likelihood for the marginal distribution of $(Y_1, \ldots, Y_V)$ and the likelihood for the conditional distribution of $(X_1, \ldots, X_K)$ given $(Y_1, \ldots, Y_V)$. 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. Since 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 from the algorithm, yielding estimates $\hat{B}$, $\hat{\boldsymbol{\Omega}}$ 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.

Similar 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 readily 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

$$(X_1|X_2, \ldots, X_K, Y_1, \ldots, Y_V) \sim N\left(\beta_{c0}(\theta) + \sum_{j=2}^{K} \beta_j(\theta) X_j, \sigma^2(\theta)\right), \qquad (14.17)$$

where c is the cell of the contingency table formed by $(Y_1, \ldots, Y_V)$, 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 observation 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, Chen and Lipsitz (1999). In a similar vein, Schluchter and Jackson (1989) discuss EM for survival analysis with missing categorical covariates.

When the continuous variables are completely observed and Y consists of a k-category covariate that is completely 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. Since 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

Table 14.3 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*

Starting Specifications (Observations in Component 1)	− 2 loglikelihood	$\hat{\mu}_1$	$\hat{\mu}_2$	$\hat{\sigma}^2$	$15\hat{p}$
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)			

likelihood are a definite possibility (Aitkin and Rubin, 1985), so it is advisable to run 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 14.3a. The standard ML estimates, assuming a single normal sample with mean μ and variance σ^2, are displayed in Table 14.3b along with -2 loglikelihood (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 observations 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 14.3c and demonstrate the sensitivity of the final estimate to starting values. The likelihood appears to be bimodal, with a high peaked mode at the estimates obtained starting from the first or third set of starting values and a low broad mode at the estimates obtained from the second set of starting values.

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 $(Y_2, \ldots, Y_V)$ and $(X_1, \ldots, X_K)$ is Bernoulli, with

$$\text{logit}[\Pr(Y_1 = 1 | Y_2, \ldots, Y_V, X_1, \ldots, X_K)] = \gamma_{c0}(\theta) + \sum_{j=1}^{K} \gamma_{cj}(\theta) X_j, \qquad (14.18)$$

where c indexes the cell defined by the values of $(Y_2, \ldots, Y_V)$, and θ again represents the location model parameters. ML or Bayes inference for the model (14.18) with missing data is obtained by fitting the general location model, and computing the regression parameters $\{\gamma_{c0}(\theta), \gamma_{cj}(\theta)\}$ in Eq. (14.18) with θ replaced by ML estimates $\hat{\theta}$ or draws $\theta^{(d)}$ from the posterior distribution. Restrictions on the general location model induce corresponding restrictions on the parameters of Eq. (14.18), yielding other logistic models. For more discussion of various methods for analyzing missing data in logistic regression, see Vach (1994).

As discussed in Chapter 10, an alternative way of implementing 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 MI methods in Section 10.2 applied to each data set. 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 para-

meters, estimate the parameters of (14.8) directly by a standard logistic regression applied to each filled-in data set, that is, apply the model

$$(Y_1|Y_2,\ldots,Y_V,X_1,\ldots,X_K) \sim \text{Bern}\left(\gamma_{c0}+\sum_{j=1}^{K}\gamma_{cj}X_j\right).$$

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, 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 (a) replacing the common covariance matrix with different but proportional covariance matrices, where the proportionality constants are to be estimated; and (b) 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 very useful for multiple imputation of missing values, assuming an ignorable missing-data mechanism. Liu and Rubin (1998) discuss ML estimation for these extensions using the AECM algorithm of Section 8.5.2. They also present a monotone-data Gibbs' sampling scheme for drawing parameters and missing values from their posterior distributions.

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 missing values.

14.3. Suppose that in Problem 14.2, X is fully observed and Y has 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 factorization is distinct and display ML estimates for this model.

14.4. Compare the properties of discriminant analysis and logistic regression for classifying observations 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 $x_{\text{obs},i}$, 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):

Case	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	2	1	1	2	2	1	1	1	2	2
Race	1	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 cases 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 chi-squared statistic for testing independence of race and sex. Note that the sample size is too small for this statistic to be considered 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 on 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 $\{\gamma_{c0}, \gamma_{cj}\}$ as functions of the general location model parameters $\theta = (\Pi, \Gamma, \Omega)$. Consider the impact of imposing constraints on on the parameters of the logistic model (14.18).

CHAPTER 15

Nonignorable Missing-Data Models

15.1. INTRODUCTION

In Section 6.2 we introduced the partition $Y = (Y_{obs}, Y_{mis})$ of the complete data Y into observed values, Y_{obs}, and missing values, Y_{mis}, and the missing-data indicator matrix M that identifies the pattern of missing data. We formulated models in terms of a probability distribution for Y with density $f(Y|\theta)$ indexed by unknown vector parameter θ, and a probability distribution $f(M|Y, \psi)$ for M given Y indexed by a vector parameter ψ. The likelihood ignoring the missing-data mechanism was defined to be any function of θ proportional to $f(Y_{obs}|\theta)$:

$$L_{ign}(\theta|Y_{obs}) \propto f(Y_{obs}|\theta), \tag{15.1}$$

where $f(Y_{obs}|\theta)$ is obtained by integrating Y_{mis} out of the density $f(Y|\theta) = f(Y_{obs}, Y_{mis}|\theta)$. The full likelihood was defined to be any function of θ and ψ proportional to $f(Y_{obs}, M|\theta, \psi)$:

$$L_{full}(\theta, \psi|Y_{obs}, M) \propto f(Y_{obs}, M|\theta, \psi), \tag{15.2}$$

where $f(Y_{obs}, M|\theta, \psi)$ is obtained by integrating Y_{mis} out of the density $f(Y, M|\theta, 4) = f(Y_{obs}, Y_{mis}|\theta)f(M|Y_{obs}, Y_{mis}, 4)$. Inference about θ based on Eq. (15.1) was shown to be equivalent to ML estimation based on Eq. (15.2) when (1) the missing data are MAR, that is, $f(M|Y_{obs}, Y_{mis}, \psi) = f(M|Y_{obs}, \psi)$ for all ψ and Y_{mis} evaluated at the observed values of M and Y_{obs}, and (2) the parameters θ and ψ are distinct, as defined in Section 6.2. Examples in Chapters 7 to 14 all concern models with likelihoods of the form (15.1), and hence were based on the assumption that conditions (1) and (2) apply. In this chapter we discuss models where the data are *not* MAR, so ML estimation requires a model for the missing-data mechanism and maximization of the full likelihood (15.2).

An important distinction arises here between models where the missing-data mechanism is nonignorable but *known*, in the sense that the distribution of M

given $Y = (Y_{obs}, Y_{mis})$ depends on Y_{mis} but does not involve unknown parameters ψ, and models where the missing-data mechanism is nonignorable and unknown, with lack of knowledge reflected in the unknown parameters ψ. A simple example with a known nonignorable mechanism is the censored exponential sample leading to the likelihood (6.51); since values are missing when they are greater than a *known* censoring point c, the distribution of M given Y is fully determined. Other examples of known nonignorable mechanisms are presented in Section 15.3. The EM algorithm or its extensions, discussed in Section 15.2 for the general case of known or unknown nonignorable mechanisms, can typically achieve ML estimation in these examples.

Sections 15.4 to 15.6 concern nonignorable models where the missing-data mechanism is nonignorable and ψ is unknown. That is, nonresponse is considered to be related in some partially unknown way to the values of Y_{mis}, even after adjusting for covariates known for respondents and nonrespondents.

Three approaches to formulating nonignorable missing-data models can be distinguished. These approaches are most easily described in the case where the observations are modeled as independent, that is, $f(M, Y|\theta, \psi) = \prod_{i=1}^{n} f(M_i, y_i|\theta, \psi)$. *Selection* models write the joint distribution of M_i and y_i in the form

$$f(M_i, y_i|\theta, \psi) = f(y_i|\theta) f(M_i|y_i, \psi), \tag{15.3}$$

where the first factor characterizes the distribution of y_i in the population, the second factor models the incidence of missing data as a function of y_i, and θ and ψ are distinct. (Any conditioning on fully observed covariates X is suppressed in the notation here.) This factorization motivates the theory of Section 6.2. Alternatively, *pattern-mixture* models write

$$f(M_i, y_i|\xi, \omega) = f(y_i|M_i, \xi) f(M_i|\omega), \tag{15.4}$$

where the first distribution characterizes the distribution of y_i in the strata defined by different patterns of missing data, M_i, the second distribution models the incidence of the different patterns (Glynn, Laird and Rubin, 1986, 1993), and ξ and ω are distinct. More generally, write $M_i = (M_i^{(1)}, M_i^{(2)})$ where $M_i^{(1)}$ indexes sets of missing-data patterns, and $M_i^{(2)}$ indexes individual patterns within each set. *Pattern-set mixture models* (Little, 1993b) write the joint distribution of M and Y in the form

$$f(M_i^{(1)}, M_i^{(2)}, y_i|\xi, \psi, \omega) = f(y_i|M_i^{(1)}, \xi) f(M_i^{(2)}|y_i, M_i^{(1)}, \psi) f(M_i^{(1)}|\omega), \tag{15.5}$$

where ξ, ψ, and ω are distinct; Eq. (15.5) includes Eqs. (15.3) and (15.4) as special cases where $M_i^{(1)}$ or $M_i^{(2)}$ contain just single patterns.

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 $i = 1, \ldots, r$ and missing for $i = r + 1, \ldots, n$. Let $M_i = M_{i2} = 1$ if y_{i2} is missing

and $M_{i2} = 0$ if y_{i2} is observed. Then the density of Y_{obs} and M is

$$f(Y_{\text{obs}}, M|\xi, \omega) = \prod_{i=1}^{r} f(y_{i1}, y_{i2}|M_{i2} = 0, \xi)\Pr(M_{i2} = 0|\omega) \times \prod_{i=r+1}^{n} f(y_{i1}|M_{i2} = 1, \xi)\Pr(M_{i2} = 1|\omega).$$

This factorization reveals a basic difficulty, namely we have no data with which to estimate directly the distribution $f(y_{i2}|y_{i1}, M_{i2} = 1, \xi)$, since all observations with $M_{i2} = 1$ have y_{i2} missing. To make progress, the distribution $f(y_{i2}|y_{i1}, M_{i2} = 1, \xi)$ for nonrespondents must be related to the corresponding distribution $f(y_{i2}|y_{i1}, M_{i2} = 0, \xi)$ for respondents. In Section 15.5 we discuss approaches to this problem.

The selection model formulation (15.3) for this problem is:

$$f(y_i, M_{i2}|\theta, \psi) = f(y_{i1}, \theta) f(y_{i2}|y_{i,s}\theta) f(M_{i2}|y_{i1}, y_{i2}, \psi),$$

and underlies the models discussed in Sections 15.3 and 15.4. We shall see that in some cases the parameters of the models can be estimated without the explicit inclusion of prior information relating respondents and nonrespondents, unlike models based on the pattern-mixture factorization. However, this property is deceptive, since there still is an implicit specification of prior information relating respondents and nonrespondents. Consequently, sensitivity to model specification is an equally serious scientific problem for *both* pattern-mixture and selection models, and in many applications it is prudent to consider estimation for a variety of missing-data models, rather than to rely exclusively on one model.

EXAMPLE 15.2. *Pattern-Set Mixture Models for Survey Nonresponse*. Unit and item nonresponse in a sample survey can be 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 nonresponse for the set of survey items. Then $M_i^{(2)} = (1, 1, \ldots, 1)$ for unit nonrespondents, and $M_i^{(2)}$ contains at least some components equal to zero for unit respondents. The pattern-set mixture model for unit i is:

$$f(M_i^{(1)}, M_i^{(2)}, y_i|\xi, \psi, \omega) = f(y_i|M_i^{(1)}, \xi) f(M_i^{(2)}|y_i, M_i^{(1)}, \psi) f(M_i^{(1)}|\omega),$$

where y_i are the survey variables, and conditioning on fully observed survey design variables is implicit. This models unit nonresponse via a pattern-mixture model, with distributions $f(y_i|M_i^{(1)} = 1, \xi)$ for unit nonrespondents and $f(y_i|M_i^{(1)} = 0, \xi)$ for unit respondents and mixing distribution $f(M_i^{(1)}|\omega)$, and item nonresponse for unit respondents as a selection model with components $f(y_i|M_i^{(1)} = 0, \xi)$ and

$f(M_i^{(2)}|y_i, M_i^{(1)} = 0, \psi)$; the remaining distribution $f(M_i^{(2)}|y_i, M_i^{(1)} = 1, \psi)$ equals 1 when $M_i^{(2)} = (1, 1, \ldots, 1)$ and zero otherwise.

A special case of this formulation that may make substantive sense is to allow unit nonresponse to be nonignorable, but to assume item nonresponse for unit respondents to be ignorable, given sufficient observed characteristics to characterize differences between item respondents and item nonrespondents. In that case, the factor $f(M_i^{(2)}|y_i, M_i^{(1)} = 0, \psi) = f(M_i^{(2)}|y_{\text{obs},i}, M_i^{(1)} = 0, \psi)$ can be ignored for likelihood inference about ψ and ω, and thus for inference about the distribution of Y.

15.2. LIKELIHOOD THEORY FOR NONIGNORABLE MODELS

Likelihood theory based on the full likelihood (15.2) parallels that for θ when the nonresponse mechanism is ignorable, as discussed in Chapters 6 to 8. In particular, ML estimates are obtained by maximizing (15.2), and a large-sample covariance matrix for the parameters can be estimated using the bootstrap, or the inverse of the information matrix obtained by differentiating the loglikelihood twice with respect to (θ, ψ). Explicit ML estimates can be derived in special situations, such as the pattern-mixture model in Section 15.5.2 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 nonignorable selection models: (1) find initial estimates $(\theta^{(0)}, \psi^{(0)})$ of $(\hat{\theta}, \hat{\psi})$; (2) at iteration t, given current estimates $(\theta^{(t)}, \psi^{(t)})$ of (θ, ψ), the E step calculates

$$Q(\theta, \psi|\theta^{(t)}, \psi^{(t)}) = \int \ell(\theta, \psi|Y_{\text{obs}}, Y_{\text{mis}}, M) f(Y_{\text{mis}}|Y_{\text{obs}}, M, \theta = \theta^{(t)}, \psi = \psi^{(t)}) dY_{\text{mis}},$$

where $\ell(\theta, \psi|Y_{\text{obs}}, Y_{\text{mis}}, M)$ is the complete-data loglikelihood and $f(Y_{\text{mis}}|Y_{\text{obs}}, M, \theta, \psi)$ is the density of the conditional distribution of the missing data given the observed values, θ and ψ. The M step finds $\theta^{(t+1)}, \psi^{(t+1)}$ to maximize Q:

$$Q(\theta^{(t+1)}, \psi^{(t+1)}|\theta^{(t)}, \psi^{(t)}) \geqslant Q(\theta, \psi|\theta^{(t)}, \psi^{(t)}) \qquad \text{for all } \theta, \psi.$$

Then $\theta^{(t+1)}, \psi^{(t+1)}$ replace $\theta^{(t)}, \psi^{(t)}$ in the next iteration of the algorithm. By theory analogous to that in Section 8.4, each iteration of this algorithm increases $\text{L}(\theta, \psi|Y_{\text{obs}}, M)$, and under rather general conditions the algorithm converges to a stationary value of the likelihood. Extensions of EM, such as ECM or PX-EM, can be helpful here, as with the ignorable case. However, for nonignorable models that are weakly identified or for which there is a large fraction of missing information, convergence to a maximum may be very slow. Also, particular attention needs to be paid to the possibility of multiple maxima of the likelihood function.

15.3. MODELS WITH KNOWN NONIGNORABLE MISSING-DATA MECHANISMS: GROUPED AND ROUNDED DATA

Kulldorff (1961) discusses scoring algorithms for ML estimation from data where some observations 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 hypothetical complete data are an independent random sample $(y_1, \ldots, y_n)$ from the exponential distribution with mean θ, but y_i is observed for $i = 1, \ldots, r < n$. The remaining $n - r$ cases are grouped into J categories, such that the jth category contains values of y_i known to lie between a_j and b_j, and the observed data for these $n - r$ cases are counts m_j of observations in the jth category, for $j = 1, \ldots, J$, $\sum_{j=1}^{J} m_j = n - r$. 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 assumed coarsened at random, using the terminology in Section 6.4.

We expand the binary missing-data indicator M_i of Section 15.1 to a variable with $J + 1$ outcomes for this example. Specifically, let $M_i = 0$ if y_i is observed, and $M_i = j$ if y_i falls in the jth nonresponse category, that is, lies between a_j and b_j $(j = 1, \ldots, J)$.

The hypothetical complete data belong to the regular exponential family with complete-data sufficient statistics $\sum_{i=1}^{n} y_i$. Hence the E step of the EM algorithm consists in calculating at iteration t

$$E\left(\sum_{i=1}^{n} y_i | Y_{\text{obs}}, M, \theta = \theta^{(t)}\right) = \sum_{i=1}^{r} y_i + \sum_{j=1}^{J} m_j \hat{y}_j^{(t)},$$

where the predicted values are given by

$$\begin{aligned}\hat{y}_j^{(t)} &= E(y | a_j \leqslant y < b_j; \theta^{(t)}) \\ &= \int_{a_j}^{b_j} y \exp\left(-\frac{y}{\theta^{(t)}}\right) dy \Bigg/ \int_{a_j}^{b_j} \exp\left(-\frac{y}{\theta^{(t)}}\right) dy,\end{aligned}$$

from the definition of the exponential distribution. Integrating by parts gives

$$\hat{y}_j^{(t)} = \theta^{(t)} + \frac{b_j e^{-b_j/\theta^{(t)}} - a_j e^{-a_j/\theta^{(t)}}}{e^{-b_j/\theta^{(t)}} - e^{-a_j/\theta^{(t)}}}. \tag{15.6}$$

The M step of EM calculates

$$\theta^{(t+1)} = n^{-1}\left(\sum_{i=1}^{r} y_i + \sum_{j=1}^{J} m_j \hat{y}_j^{(t)}\right). \tag{15.7}$$

The predicted value for an observation censored at a_j is obtained by setting $b_j = \infty$, yielding

$$\hat{y}_j^{(t)} = \theta^{(t)} + a_j.$$

If all the $n - r$ grouped observations are censored, then an explicit ML estimate can be derived. Substituting Eq. (15.6) in Eq. (15.7) 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 observations 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 6.22.

EXAMPLE 15.4. *Grouped Normal Data with Covariates*. Suppose that data on an outcome variable Y are grouped as in Example 15.3, where a case $i > r$ is classified in group j if it is known to lie between a_j and b_j, but now the hypothetical complete Y values are independent normal with a linear regression on fully observed covariates $X_1, X_2, \ldots, X_p$. 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 $\sum y_i$, $\sum y_i x_{ik}$ $(k = 1, \ldots, p)$, and $\sum y_i^2$. Hence the E step of the EM algorithm computes

$$E\left(\sum_{i=1}^{n} y_i | Y_{\text{obs}}, M, \theta = \theta^{(t)}\right) = \sum_{i=1}^{r} y_i + \sum_{i=r+1}^{n} \hat{y}_i^{(t)},$$

$$E\left(\sum_{i=1}^{n} y_i x_{ik} | Y_{\text{obs}}, M, \theta = \theta^{(t)}\right) = \sum_{i=1}^{r} y_i x_{ik} + \sum_{i=r+1}^{n} \hat{y}_i^{(t)} x_{ik}, \qquad k = 1, 2, \ldots, p,$$

$$E\left(\sum_{i=1}^{n} y_i^2 | Y_{\text{obs}}, M, \theta = \theta^{(t)}\right) = \sum_{i=1}^{r} y_i^2 + \sum_{i=r+1}^{n} (\hat{y}_i^{(t)2} + \hat{s}_i^{(t)2}),$$

where $\theta = (\beta_0, \beta_1, \ldots, \beta_p, \sigma^2)$, $\theta^{(t)} = (\beta_0^{(t)}, \beta_1^{(t)}, \ldots, \beta_p^{(t)}, \sigma^{(t)2})$ 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 nonignorable nonresponse, which 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, and for units i in the jth category ($M_i = j$ or equivalently, $a_j < Y \leqslant b_j$),

$$c_i^{(t)} = (a_j - \mu_i^{(t)})/\sigma^{(t)} \qquad \text{and} \qquad d_i^{(t)} = (b_j - \mu_i^{(t)})/\sigma^{(t)}.$$

The M step calculates the regression of Y on $X_1, \ldots, X_p$ 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, Stead, and Galke (1980).

EXAMPLE 15.5. *Censored Normal Data with Covariates (Tobit Model)*. An important special case of the previous example occurs when positive values of Y are fully recorded but negative values are censored, that is, can lie anywhere in the interval $(-\infty, 0)$. In the notation of Example 15.4, all observed y_i are positive, $J = 1$, $a_1 = -\infty$, and $b_1 = 0$. We have for censored cases $c_i^{(t)} = -\infty$, $d_i^{(t)} = -\mu_i^{(t)}/\sigma^{(t)}$, $\delta_i^{(t)} = -\phi(d_i^{(t)})/\Phi(d_i^{(t)})$, $\gamma_i^{(t)} = \delta_i^{(t)}(\delta_i^{(t)} + \mu_i^{(t)}/\sigma_i^{(t)})$. Hence

$$\hat{y}_i^{(t)} = E(y_i|\theta^{(t)}, x_i, y_i \leqslant 0) = \mu_i^{(t)} - \sigma^{(t)}\lambda(-\mu_i^{(t)}/\sigma^{(t)}),$$

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

$$\hat{y}_i^{(t)} = E(y_i|\hat{\theta}, x_i, y_i \leqslant 0) = \hat{\mu}_i - \hat{\sigma}\lambda(-\hat{\mu}_i/\hat{\sigma}) \tag{15.8}$$

for censored cases, where $\hat{\mu}_i = \hat{\beta}_0 + \sum_{k=1}^{p} \hat{\beta}_k x_{ik}$. This model is sometimes called the Tobit model in the econometric literature (see Amemiya, 1984), after an earlier econometric application (Tobin, 1958).

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 valued responses, bracketed (or interval-censored) replies, and completely missing data. Heeringa, Little and Raghunathan (2002) develop multiple imputations of coarsened and missing data for 12 asset and liability variables in the Health and Retirement survey, 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 non-negative asset or liability measures for subject 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

$$y_i = \begin{cases} (\exp(z_{i1}), \exp(z_{i2})) & \text{if } t_i = (1, 1) \\ (\exp(z_{i1}), 0) & \text{if } t_i = (1, 0) \\ (0, \exp(z_{i2})) & \text{if } t_i = (0, 1) \\ (0, 0) & \text{if } t_i = (0, 0), \end{cases} \tag{15.9}$$

and

$$(z_i | t_i = (j, k)) \sim N_2(\mu_{jk}, \Sigma). \tag{15.10}$$

The exponential transformations in Eq. (15.9) imply that the non-zero 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 13.2. Note that it is unrealistic to apply Eq. (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 were constrained to zero. However, an alternative approach that may speed the convergence of the algorithm is to treat them as parameters to be estimated, 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, and the observed or drawn values of all the other variables. Since 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.

Gibbs' sequences are created with 20 different random starts to yield 20 multiply imputed data sets. Multiple imputation inference methods described in Section 10.2

Table 15.1 HRS Wave 1 Estimates of the Distribution of Total Household Net Worth, in Thousands of Dollars

	1. MI Bayes		2. Complete Cases		3. Mean Imputation		4. MI Hot Deck	
Estimand	Estimate	s.e.	Estimate	s.e.	Estimate	s.e.	Estimate	s.e.
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	2642.3	264.1	1995.0	107.1	2142.1	62.035	2317.8	216.0
Max	15,663	9458	6202	322.0	9096	469.4	9645	3070

then yield estimates, confidence intervals, and test statistics for model parameters. Imputations in the highest open-ended dollar categories are highly sensitive to model specification, and the method described here truncates them so that they did not exceed the highest recorded value in the data set. Extensions of the log-normal model presented here might be developed to fit the tails of the distributions more closely.

Table 15.1 shows the results of applying the 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 4566 (60.0%) of the 7607 cooperating HRS Wave 1 households who provided complete information on holding and amounts for each of the 23 asset and liability components.

Mean Imputation. The mean of observed values that fell within the known bounds was imputed for bracketed missing values. For example, if a missing observation was reported to fall in the bracket \$5000–\$49,999, the mean of observed values for this interval (e.g., \$21,000) was imputed for each missing case. If an observation for a variable was completely missing (no bracket information), the overall mean or median of the observed values for that variable was imputed for the case.

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. Each asset and liability variable was imputed independently—better methods would attempt to exploit associations between the items, but the hot deck is severely limited in this regard by the need to find matches with

observed cases. All observed and missing cases were assigned to hot deck cells based on the bracket boundaries for the variable and discrete categories defined by covariate information including the age, race, sex, and marital status of the household head. Each missing value was then imputed using the observed value of a randomly selected observed case within the same hot-deck cell. If a missing observation had no bracketing information, the random donor was drawn from observed values in the collapsed hot-deck cell formed by the discrete cross-classifications of household head's age, race, sex, and marital status. Repeating this hot-deck procedure with different random donors chosen within each adjustment cell created 20 multiply imputed data sets.

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 (JRR) method (Wolter, 1985), and reflect the influences of weighting, stratification, and clustering of the complex multistage HRS sample design. Standard errors for the mean and median 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 formulas 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 (a) complete-case analysis appears to markedly underestimate the distribution of household net worth for HRS households; (b) 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 standard deviation in net worth amounts imputed by the stochastic hot deck and Bayes alternatives; (c) The hot-deck method produces lower estimated values for the mean and upper quantiles of the distribution than the Bayes algorithm. 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.

15.4. NORMAL SELECTION MODELS

An interesting extension of the model of Example 15.5 concerns an incompletely observed variable (Y_1) that again has a linear regression on covariates and is observed if and only if the value of another completely unobserved variable (Y_2) exceeds a threshold (say zero). A common specification is given in the following example.

EXAMPLE 15.7. *Bivariate Normal Stochastic Censoring Model*. Suppose Y_i is incompletely observed, Y_2 is never observed, p covariates X are fully observed, and for case i, $f(Y|X, \theta)$ is specified by

$$\begin{pmatrix} y_{i1} \\ y_{i2} \end{pmatrix} \sim_{\text{ind}} N_2\left[\begin{pmatrix} x_i\beta_1 \\ x_i\beta_2 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \rho\sigma_1 \\ \rho\sigma_1 & 1 \end{pmatrix}\right], \tag{15.11}$$

where $x_i = (x_{i0}, x_{i1}, \ldots, x_{ip})$ are the constant term ($x_{i0} \equiv 1$) and predictors $(x_{i1}, \ldots, x_{ip})$ for case i, β_1 and β_2 are $(p+1) \times 1$ vectors of regression coefficients, some of which may be set to zero *a priori*. Further, let $M_i = (M_{i1}, M_{i2})$, where M_{ij} is the missing-data indicator for y_{ij}; $f(M_i|x_i, y_i, \psi)$ is specified by the degenerate distribution

$$\begin{aligned} M_{i1} &= \begin{cases} 1, & \text{if } y_{i2} \leqslant 0, \\ 0, & \text{if } y_{i2} > 0, \end{cases} \\ M_{i2} &\equiv 1. \end{aligned} \tag{15.12}$$

Since y_{i2} is always missing, we can integrate it out of the model and omit M_{i2}. From Eqs. (15.11) and (15.12), the distribution of M_{i1} given y_{i1} and x_i is Bernoulli with probability of nonresponse

$$\begin{aligned} \Pr(M_{i1} = 1|y_{i1}, x_i) &= \Pr(y_{i2} \leqslant 0|y_{i1}, x_i) \\ &= 1 - \Phi\left[\frac{\mu_{i2} + \rho\sigma_1^{-1}(y_{i1} - \mu_{i1})}{\sqrt{1-\rho^2}}\right], \end{aligned} \tag{15.13}$$

where $\mu_{i1} = x_i\beta_1$, $\mu_{i2} = x_i\beta_2$. When $\rho \neq 0$ (that is Y_1 and Y_2 are correlated), this probability is a monotonic function of the values y_{i1}, which are sometimes missing, so the missing-data mechanism is nonignorable.[1] If, on the other hand, $\rho = 0$ and (β_1, σ_1^2) and β_2 are distinct, then the missing-data mechanism is ignorable, and ML estimates of (β_1, σ_{11}) are obtained by least squares linear regression based on the complete cases. This model was introduced by Heckman (1976) to describe selection of women into the labor force. Amemiya (1984) calls it a Type II Tobit model; note that the Tobit model of Example 15.5 is obtained when $Y_1 = \sigma_2 Y_2$.

Two estimation procedures have been proposed for this model, ML and the two-step method of Heckman (1976). ML estimation was originally achieved using the algorithm of Berndt et al. (1974). We describe the EM algorithm for the case where no constraints are placed on the coefficients β_1, β_2, with hypothetical complete data defined as cases with both Y_1 and Y_2 completely observed. The complete-data sufficient statistics are then $\{\sum_i y_{i1}x_{ij}, \sum_i y_{i2}x_{ij}, \sum_i y_{i1}y_{i2}, \sum_i y_{i1}^2, \sum_i y_{i2}^2\}$ for

[1] When $\rho = 0$ the parameters in Eqs. (15.11) and (15.13) are not distinct, so these equations do not define a selection model according to the strict definition in Section 15.1; they are based on the selection model factorization, however, and distinctness can be achieved by a reparametrization.

$j = 0, 1, \ldots, p$. Since $\{x_{ij}\}$ are fully observed, the E step consists of replacing missing values of y_{i1}, y_{i2}, $y_{i1}y_{i2}$, y_{i1}^2, and y_{i2}^2 by their expectations given the parameters and the observed data. Properties of the bivariate normal distribution yield, for cases with y_{i1} missing:

$$\begin{aligned}
E(y_{i2}|y_{i2} \leqslant 0) &= \mu_{i2} - \lambda(-\mu_{i2}),\\
E(y_{i1}|y_{i2} \leqslant 0) &= \mu_{i1} - \rho\sigma_1\lambda(-\mu_{i2}),\\
E(y_{i2}^2|y_{i2} \leqslant 0) &= 1 + \mu_{i2}^2 - \mu_{i2}\lambda(-\mu_{i2}),\\
E(y_{i1}^2|y_{i2} \leqslant 0) &= \mu_{i1}^2 + \sigma_1^2 - \rho\sigma_1\lambda(-\mu_{i2})(2\mu_{i1} - \rho\sigma_1\mu_{i2}),\\
E(y_{i1}y_{i2}|y_{i2} \leqslant 0) &= \mu_{i1}[\mu_{i2} - \lambda(-\mu_{i2})] + \rho\sigma_1,
\end{aligned}$$

and for cases with y_{i1} observed:

$$\begin{aligned}
E(y_{i2}|y_{i1}, y_{i2} > 0) &= \mu_{i2\cdot 1} + \sqrt{1-\rho^2}\,\lambda(\mu_{i2\cdot 1}/\sqrt{1-\rho^2}),\\
E(y_{i2}^2|y_{i1}, y_{i2} > 0) &= 1 - \rho^2 + \mu_{i2\cdot 1}^2 + \mu_{i2\cdot 1}\sqrt{1-\rho^2}\,\lambda(\mu_{i2\cdot 1}/\sqrt{1-\rho^2}),
\end{aligned}$$

where conditioning on x_i and the parameters is implicit in these expressions, $\lambda(\cdot)$ is the inverse Mills ratio, as defined in Example 15.5, and $\mu_{i2\cdot 1} = \mu_{i2} + \rho\sigma_1^{-1}(y_{i1} - \mu_{i1})$. Current values of the parameters are substituted to yield estimates for the E step.

The M step consists of the following computations, performed with complete-data sufficient statistics replaced by estimates from the E step:

1. Regress Y_2 on X, yielding coefficients $\hat{\beta}_2$ of the response equation.
2. Regress Y_1 on Y_2 and X, yielding coefficients $\hat{\delta}$ for Y_2 and $\hat{\beta}_1^*$ for X, and residual variance $\hat{\sigma}_{1\cdot 2}^2$.
3. Set $\hat{\beta}_1 = \hat{\beta}_1^* + \hat{\delta}\hat{\beta}_2$, $\hat{\sigma}_1^2 = \hat{\sigma}_{1\cdot 2}^2 + \hat{\delta}^2$, and $\hat{\rho} = \hat{\delta}/\hat{\sigma}_1$.

When *a priori* constraints are imposed on the coefficients β_1 and β_2, the M step involves iterative calculations, so the simplicity of the algorithm is lost, although ECM may be applicable.

The model of Example 15.7 implies the following predicted values for missing cases:

$$E(y_{i1}|y_{i2} \leqslant 0, x_i, \hat{\theta}) = \hat{\mu}_{i1} - \hat{\rho}\hat{\sigma}_1\lambda(-\hat{\mu}_{i2}), \tag{15.14}$$

where $\hat{\mu}_{i1} = x_i\hat{\beta}_1$, $\hat{\mu}_{i2} = x_i\hat{\beta}_2$. Note that the censoring correction $-\hat{\rho}\hat{\sigma}_1\lambda(-\hat{\mu}_{i2})$ depends on the estimated correlation $\hat{\rho}$, a quantity that does not arise in the predictions (15.8) for the pure censoring model of Example 15.5. Thus, even though Y_1 and Y_2 are never jointly observed, their correlation ρ must be estimated.

Two assumptions of the model yield information on ρ that is exploited in ML estimation: (1) any *a priori* restriction placed on the coefficients β_1 and β_2 and (2)

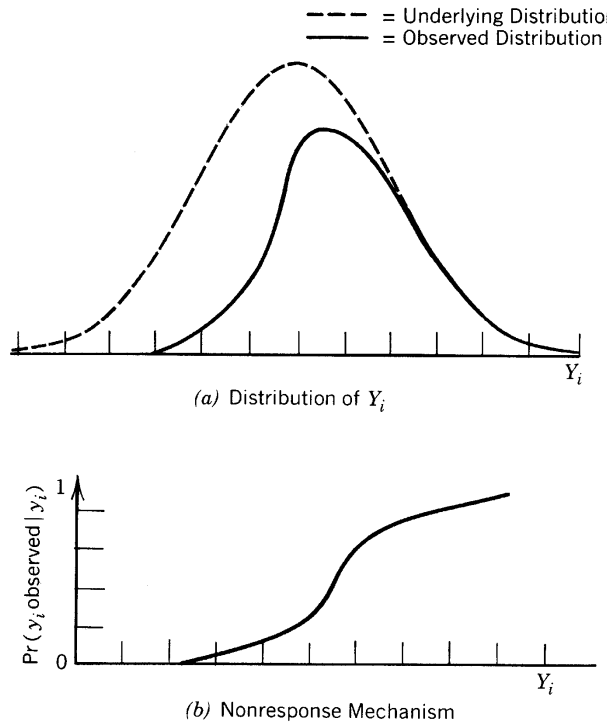


Figure 15.1. The normal stochastic censoring model for a single sample.

the assumption of normality of y_{i1} given x_i in the unrestricted population. To see the role of the latter assumption, consider the model in the absence of covariates, where $x_i \equiv 1$, the constant term. Figure 15.1*a* displays the model distributions of Y_1 in the unrestricted and the respondent populations, with the latter scaled so that the area underneath it is the fraction of respondents. By assumption, the unrestricted distribution of Y_1 is normal. The respondent distribution is skewed by the stochastic censoring, unless $\rho = 0$. Given a sample of respondents, ρ is estimated to account for the lack of normality in the sample. In other words, it fills in the nonrespondents to make the unrestricted sample look as normal as possible. Clearly, this procedure relies totally on the untestable assumption that the values of Y_1 in the unrestricted population are normal. In the absence of knowledge about this distribution, an equally plausible assumption might be that nonrespondents have the same skewed distribution as that observed for respondents in Figure 15.1. If this hypothesis were in fact true, the correction for selectivity in the normal model would add bias rather than eliminate it. The following example illustrates this possibility.

EXAMPLE 15.8. *Income Nonresponse in the Current Population Survey.* Lillard, Smith, and Welch (1982, 1986) apply the model of Example 15.7 to income nonresponse in four rounds of the Current Population Survey Income Supplement, conducted in 1970, 1975, 1976, and 1980. In 1980 their sample consisted of 32,879

employed white civilian males aged 16–65 years 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 is a power transformation of the kind proposed in Box and Cox (1964). The predictors X were chosen as

Constant term
Education (five dummy variables; Sch 8, Sch 9–11, Sch 12, Sch 13–15, Sch 16+)
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 (yes, no)
Personal interview (1 = yes, 0 = no)
Year in survey (1 or 2)

The last four variables were omitted from the earnings equation; that is, their coefficients in the vector β_1 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 β_2 were set to zero.

Most empirical studies model the logarithm of earnings, the transformation obtained by letting $\gamma \to 0$. Table 15.2 shows estimates of the regression coefficients

Table 15.2 Estimates for the Regression of ln (Earnings) on Covariates: 1980 CPS. (Example 15.8.)

Variable	Coefficient (β_1) OLS on Respondents	ML for Selection Model	Two-Step Method
Constant	9.5013 (0.0039)	9.6816 (0.0051)	10.0373 (0.0173)
Sch 8	0.2954 (0.0245)	0.2661 (0.0202)	0.2615 (0.0241)
Sch 9–11	0.3870 (0.0206)	0.3692 (0.0169)	0.3718 (0.0203)
Sch 12	0.6881 (0.0188)	0.6516 (0.0158)	0.6713 (0.0185)
Sch 13–15	0.7986 (0.0201)	0.7694 (0.0176)	0.8096 (0.0198)
Sch 16+	1.0519 (0.0199)	1.0445 (0.0178)	1.0418 (0.0195)
Exp 0–5	−0.0225 (0.0119)	−0.0294 (0.0111)	−0.0425 (0.0117)
Exp 5–10	0.0534 (0.0038)	0.0557 (0.0039)	0.0561 (0.0037)
Exp 10–20	0.0024 (0.0016)	0.0240 (0.0016)	0.0448 (0.0017)
Exp 20+	−0.0052 (0.0008)	−0.0036 (0.0008)	−0.0033 (0.0008)
Prob 1	−1.8136 (0.1075)	−1.7301 (0.0945)	−1.5311 (0.1059)
South	−0.0654 (0.0087)	−0.0649 (0.0085)	−0.0893 (0.0086)
$\rho = \text{Corr}(y_1, y_2\|x)$	0 (by assumption)	−0.6842	(not estimated)
N	27,909	32,879	32,879

Table 15.3 The Maximized Loglikelihood as a Function of γ with Associated Values of $\hat{\rho}$

γ	Maximized Loglikelihood	$\hat{\rho}$
0	−300,613.4	−0.6812
0.45	−298,169.7	−0.6524
1.0	−300,563.1	0.8569

Source: Lillard, Smith, and Welch (1982).

β_1 for this transformation of earnings, calculated (1) by ordinary least squares (OLS) on the respondents, a procedure that effectively assumes ignorable nonresponse ($\rho = 0$); and (2) by ML for the model of Example 15.7. For the ML procedure, $\hat{\rho} = -0.6812$, implying positive corrections $-\hat{\rho}\hat{\sigma}_1\lambda(-\hat{\mu}_{i2})$ for nonignorability to the income amounts of nonrespondents (Eq. 15.9). The regression coefficients from OLS and ML in Table 15.2 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 the predicted income amounts for nonignorable nonresponse, a substantial correction.

Lillard, Smith, and Welch (1982) fit the stochastic censoring model for a variety of other choices of γ. Table 15.3 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 for $\hat{\gamma} = 0.45$ is much larger than that for $\gamma = 0$ or $\gamma = 1$, indicating that the normal selection models for raw and log-transformed earnings are not supported by the data.

Table 15.3 also shows values of $\hat{\rho}$ as a function of γ. Note that for $\gamma = 0$ and $\hat{\gamma} = 0.45$, $\hat{\rho}$ 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, Smith, and Welch's best-fitting model, $\hat{\gamma} = 0.45$, predicts large income amounts for nonrespondents, in fact, 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 ignorable and the unrestricted residuals follow the same (skewed) distribution as that in the respondent sample. Indeed, comparisons of Census Bureau imputations with IRS income amounts from matched CPS/IRS files do not indicate substantial underestimation (David et al., 1986).

The last column of Table 15.2 shows estimates from the original least squares fitting procedure proposed by Heckman (1976), which is not ML but is easier to compute. It relies heavily on untestable assumptions about *a priori* zeros in the

regression coefficient vectors, β_1 and β_2, and may yield very misleading results if these assumptions are incorrect. See Problem 15.6, Little (1985a), and Little and Rubin (1987). Hence we do not recommend this method.

Variants of the normal stochastic censoring model replace the probit model for response by the uniform (Olsen, 1980) and the logistic (Olsen, 1980; Greenlees, Reece, and Zieschang, 1982), the latter reference in the context of CPS income nonresponse. Estimates from these models are also highly sensitive to the choice of *a priori* zeros in β_1 and β_2.

15.5. NORMAL PATTERN-MIXTURE MODELS

15.5.1. Univariate Normal Pattern-Mixture Models

A normal pattern-mixture model for a single variable Y subject to missing values is:

$$(y_i|M_i = m) \sim_{\text{ind}} N(\mu_{(m)}, \sigma^2_{(m)}), \qquad m = 0, 1; \; M_i \sim_{\text{ind}} \text{Bern}(\pi). \tag{15.15}$$

This model implies that marginally Y is a mixture of two normal distributions, with mean $\mu = \pi\mu_{(1)} + (1 - \pi)\mu_{(0)}$ and variance $\pi\sigma^2_{(1)} + (1 - \pi)\sigma^2_{(0)} + \pi(1 - \pi)(\mu_{(1)} - \mu_{(0)})^2$. When the data are MCAR, $\mu_{(m)} = \mu$, $\sigma^2_{(m)} = \sigma$ for $m = 0, 1$. When the data are not MCAR, more assumptions are needed to identify $\mu_{(1)}, \sigma_{(1)}$.

EXAMPLE 15.9. *Sensitivity of a Mean to Nonignorable Nonresponse*. Rubin (1977) considers a version of the following model. Assume that $\sigma_{(1)} = \sigma_{(0)} = \sigma$, and that $\mu^{(1)}$ is identified by the following prior distribution relating nonrespondents to respondents:

$$p(\mu_{(0)}, \log\sigma^2) \propto \text{const.}, \qquad p(\mu_{(1)}|\mu_{(0)}, \sigma^2) \sim N(\mu_{(0)}, \psi_2^2\mu^2_{(0)}), \tag{15.16}$$

for a selected value of ψ_2. The mean of the prior distribution of $\mu_{(1)}$ is $\mu_{(0)}$, which implies that it is as likely for the nonrespondent mean to be above or below the respondent mean. The quantity ψ_2 measures the subjective coefficient of variation for the nonrespondent mean about the respondent mean. In particular, it implies that the investigator is 95% sure that the nonrespondent mean will fall in the interval

$$\mu_{(0)}(1 \pm 1.96\psi_2).$$

If $\psi_2 = 0$ then the distributions of Y for respondents and nonrespondents are equal, and the missing-data mechanism is MCAR. Let $\bar{y}_{(0)}$ be the mean for the r respondents and $p = (n - r)/n$, the fraction of nonrespondents. Rubin (1977) considers predictive Bayesian inference for the average of Y in the whole sample, namely $\bar{y}$.

Applying Bayes Theorem, the posterior distribution of $\bar{y}$ given $\bar{y}_{(0)}, \sigma^2$ is normal with mean $\bar{y}_{(0)}$ and variance

$$\begin{aligned}\mathrm{Var}(\bar{y}|\bar{y}_{(0)}, \sigma^2) &= p^2\mathrm{Var}(\bar{y}_{(1)}|\bar{y}_{(0)}, \sigma^2) \\ &= p^2\mathrm{Var}(\mu_{(0)}|\bar{y}_{(0)}, \sigma^2) + p^2E\{\mathrm{Var}(\bar{y}_{(1)}|\mu_{(0)})|\bar{y}_{(0)}, \sigma^2\} \\ &\quad + p^2E\{\mathrm{Var}(\mu_{(1)}|\mu_{(0)})|\bar{y}_{(1)}, \sigma^2\} \\ &= p\sigma^2/r + p^2\psi_2^2(\bar{y}_{(0)}^2 + \sigma^2/r).\end{aligned}$$

from Eq. (15.16). Hence the posterior distribution of μ is normal with mean $\bar{y}_{(0)}$ and variance

$$\mathrm{Var}(\mu|\bar{y}_{(0)}, \sigma^2) = \sigma^2/r + p^2\psi_2^2(\bar{y}_{(0)}^2 + \sigma^2/r).$$

The second term on the right side is zero when $\psi_2^2 = 0$, and represents the added uncertainty from the nonignorable component of the model.

EXAMPLE 15.10. *Sensitivity of the Sample Mean to Nonignorable Nonresponse, in the Presence of Covariates*. An extension of the previous example includes q covariates $X = (X_1, \ldots, X_q)$, available for all units in the sample. Suppose:

$$\begin{aligned}(y_i|x_i, M_i = m) &\sim_{\text{ind}} N(\phi_{(m)} + \beta_{(m)}(x_i - \bar{x}_{(0)}), \sigma^2); \\ p(\phi_{(0)}, \beta_{(0)}, \log\sigma^2) &\propto \text{const.}, \\ p(\beta_{(1)}|\phi_{(0)}, \beta_{(0)}, \sigma^2) &\sim N_q(\beta_{(0)}, \psi_1^2\beta_{(0)}\beta_{(0)}^{\mathrm{T}}), \\ p(\phi_{(1)}|\phi_{(0)}, \beta_{(0)}, \beta_{(1)}, \sigma^2) &\sim N(\phi_{(0)}, \psi_2^2\phi_{(0)}^2),\end{aligned} \tag{15.17}$$

where $\phi_{(0)}$ and $\phi_{(1)}$ 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 coefficients. The parameter ψ_2 measures uncertainty in the adjusted mean and corresponds to ψ_2 in Example 15.9 for the case of no covariates. The nonresponse mechanism is ignorable for likelihood-based inferences if $\psi_1 = \psi_2 = 0$.

The posterior distribution of $\bar{y}$ is normal with mean, $\bar{y}_{(0)} + \hat{\beta}_{(0)}^{\mathrm{T}}(\bar{x} - \bar{x}_{(0)})$, namely the regression estimator, and variance $\bar{y}_{(0)}^2(\psi_1^2h_1^2 + \psi_2^2h_2^2 + h_3^2)$, where

$$\begin{aligned}h_1^2 &= (\sigma^2/\bar{y}_{(0)}^2)\{[\hat{\beta}_{(0)}^{\mathrm{T}}(\bar{x} - \bar{x}_{(0)})]^2/\sigma^2 + (\bar{x} - \bar{x}_{(0)})^{\mathrm{T}}S_{xx}^{-1}(\bar{x} - \bar{x}_{(0)})\} \\ h_2^2 &= p^2\{1 + \sigma^2/(r\bar{y}_{(0)})\}\end{aligned}$$

and

$$h_3^2 = (\sigma^2/\bar{y}_{(0)}^2)[(p/r) + (\bar{x} - \bar{x}_{(0)})^{\mathrm{T}}S_{xx}^{-1}(\bar{x} - \bar{x}_{(0)})];$$

here S_{xx} is the sum-of-squares and cross-products matrix of X for respondents. The width of the associated 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 nonresponse 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 (Ys) and 35 background variables (Xs) 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 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 95% interval for $\bar{y}$ for seven of these outcome variables, expressed as a percentage of the 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, which ranged from 0.003 to 0.077. The values of h_2 which ranged from 0.285 to 0.290, were are only marginally greater than the proportion of missing values, $p = 0.2848$. Thus, the contribution to the interval width of

Table 15.4 Widths of Subjective 95% Intervals of $\bar{y}$, as Percentages of $\bar{y}_{(0)}$

	$\psi_1 = 0$				$\psi_1 = 0.4$			
Variable	$\psi_2 = 0$	$\psi_2 = 0.1$	$\psi_2 = 0.2$	$\psi_2 = 0.4$	$\psi_2 = 0$	$\psi_2 = 0.1$	$\psi_2 = 0.2$	$\psi_2 = 0.4$
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.

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 five. 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 non-negligible for $\psi_2 \geqslant 0.1$. 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.

One way to reduce sensitivity of inference to nonignorable 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 15.11. *Decreased Sensitivity of Inference with Follow-ups*. Glynn, Laird, and Rubin (1986) performed a series of simulations using normal and log-normal data, which can be 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 nonrespondents. The resultant data consisted of (y_i, M_i) for respondents and followed-up nonrespondents, but only M_i for nonrespondents who were not followed up.

Two models were used to analyze the data. First, the pattern-mixture model (15.15) 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:

$$\begin{aligned}(y_i|\mu, \sigma^2) &\sim N(\mu, \sigma^2),\\ \operatorname{logit}(\Pr(M_i = 1|y_i, \alpha_0, \alpha_1)) &= \alpha_0 + \alpha_1 y_i,\end{aligned}$$

where the 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 log-normal values (exponentiated standard normal deviates) and 88 nonrespondents created using the nonignorable logistic response 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 log-normal data. Note that, whereas for the normal

Table 15.5 Sample Moments of Generated Data[a] for Example 15.11

	Normal Data			Lognormal Data		
	N	Sample Mean	Sample Standard Deviation	N	Sample Mean	Sample Standard Deviation
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

[a] Normal data are sampled from the normal $(0, 1)$ distribution; log-normal data are the exponentiated normal values. Response is determined by a logistic response function: $\Pr(M_i = 1|y_i) = [1 + \exp(1 + y_i)]^{-1}$, where $(\alpha_0, \alpha_1) = (1, 1)$ for normal data and $(0, 1)$ for lognormal data.

data the selection model was correct and the pattern-mixture model incorrect, for the log-normal data both models are incorrect.

Table 15.5 summarizes the generated data, both normal and log-normal. Table 15.6 gives estimates of the population means for both models with both the normal and log-normal 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, even a few follow-ups yield reasonable estimates. Glynn, Laird, and Rubin (1986) use multiple imputation to draw inferences from survey data of retired men with follow-ups, using an extension of the mixture model that includes covariates.

Table 15.6 Estimates of Population Mean Using Respondent Data of Table 15.5 and Follow-up Data from Some Nonrespondents

Normal Data			Lognormal Data		
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

15.5.2. Bivariate Normal Pattern-Mixture Models Identified via Parameter Restrictions

Suppose now that the data consist of a random sample of n observations (y_{i1}, y_{i2}), $i = 1, \ldots, n$ on two variables (Y_1, Y_2), with $\{y_{i1}\}$ fully observed and $\{y_{i2}\}$ observed for $i = 1, \ldots, r$ and missing for $i = r+1, \ldots, n$. The following defines a normal pattern-mixture model for this bivariate data set:

$$(y_{i1}, y_{i2}|M_i = m) \sim_{\text{ind}} N(\mu^{(m)}, \Sigma^{(m)}); \qquad M_i \sim_{\text{ind}} \text{Bern}(\pi), \quad m = 0, 1. \tag{15.18}$$

When the data are MCAR, missingness of y_{i2} is independent of y_{i1} and y_{i2}, and $\mu^{(m)} = \mu$, $\Sigma^{(m)} = \Sigma$ for $m = 0, 1$. We consider inferences about parameters averaged over patterns when the data are not MCAR. The pattern-mixture model (15.18) has 11 parameters $\phi^{(m)} = (\mu_1^{(m)}, \mu_2^{(m)}, \sigma_{11}^{(m)}, \sigma_{22}^{(m)}, \sigma_{12}^{(m)})$, $m = 0, 1$, and π, but only eight can be identified from the data, in the sense of appearing in the likelihood and having unique ML estimates, namely

$$\phi_{\text{id}} = (\pi, \mu_1^{(0)}, \mu_2^{(0)}, \sigma_{11}^{(0)}, \sigma_{12}^{(0)}, \sigma_{22}^{(0)}, \mu_1^{(1)}, \sigma_{11}^{(1)}). \tag{15.19}$$

The likelihood has the form

$$\text{L}(\phi_{\text{id}}) = (1-\pi)^r \pi^{n-r} \prod_{i=1}^{r} f(y_{i1}, y_{i2}|m_i = 0, \phi^{(0)}) \prod_{i=r+1}^{n} f(y_{i1}|m_i = 1, \mu_1^{(1)}, \sigma_{11}^{(1)}), \tag{15.20}$$

yielding ML estimates $\hat{\pi} = (n-r)/n$, $\hat{\mu}_j^{(0)} = \bar{y}_j$, $\hat{\sigma}_{jk}^{(0)} = s_{jk}$, $\hat{\mu}_1^{(1)} = \bar{y}_1^{(1)}$, and $\hat{\sigma}_{11}^{(1)} = s_{11}^{(1)}$, where the latter two statistics are the mean and variance of Y_1 from the $n-r$ incomplete cases. The three parameters of the conditional distribution of Y_2 given Y_1 for incomplete cases, say $\phi_{2\cdot1}^{(1)} = (\beta_{20\cdot1}^{(1)}, \beta_{21\cdot1}^{(1)}, \sigma_{22\cdot1}^{(1)})$, do not appear in the likelihood, and are identified by restrictions on the parameters that are based on the assumed nature of the missing-data mechanism.

First, suppose that missingness of Y_2 is assumed to depend only on Y_1; that is, the mechanism is MAR. Then the distribution of Y_2 given Y_1 is the same for both patterns, implying the restriction:

$$\phi_{2\cdot1}^{(1)} = \phi_{2\cdot1}^{(0)}. \tag{15.21}$$

Little (1993b) calls Eq. (15.21) complete-case missing-variable restrictions, since they tie the unidentified parameters $\phi_{2\cdot1}^{(1)}$ to their complete-case analogs $\phi_{2\cdot1}^{(0)}$. Eq. (15.21) implies that $\mu_2^{(1)} = \beta_{20\cdot1}^{(1)} + \beta_{21\cdot1}^{(1)}\mu_1^{(1)} = \beta_{20\cdot1}^{(0)} + \beta_{21\cdot1}^{(0)}\mu_1^{(1)}$, and hence

$$\hat{\mu}_2 = (1-\hat{\pi})\hat{\mu}_2^{(0)} + \hat{\pi}\hat{\mu}_2^{(1)} = (r/n)\bar{y}_2 + [(n-r)/n](b_{20\cdot1} + b_{21\cdot1}\bar{y}_1),$$

which is the ML estimate of μ_2 for the ignorable selection model, Eq. (7.9). ML estimates of the other parameters of the marginal distribution of Y_1 and Y_2 are easily shown to be the same as those for the ignorable selection model. Hence the pattern-mixture model (15.18) with restrictions (15.21), and the ignorable normal selection models of Section 7.2.1 yield the same ML estimates, despite differing distributional assumptions.

Now suppose that missingness of Y_2 is assumed to depend on Y_2 rather than Y_1. This mechanism implies that the distribution of Y_1 given Y_2 is independent of pattern, so that

$$\phi_{1\cdot 2}^{(1)} = \phi_{1\cdot 2}^{(0)}, \tag{15.22}$$

where $\phi_{1\cdot 2}^{(m)}$ are the parameters of the distribution of Y_1 given Y_2 for pattern m. Hence

$$\mu_2^{(1)} = (\mu_1^{(1)} - \beta_{10\cdot 2}^{(1)})/\beta_{12\cdot 2}^{(1)} = (\mu_1^{(1)} - \beta_{10\cdot 2}^{(0)})/\beta_{12\cdot 2}^{(0)},$$
$$\sigma_{22}^{(1)} = (\sigma_{11}^{(1)} - \sigma_{11\cdot 2}^{(1)})/(\beta_{12\cdot 2}^{(1)})^2 = (\sigma_{11}^{(1)} - \sigma_{11\cdot 2}^{(0)})/(\beta_{12\cdot 2}^{(0)})^2,$$

and $\sigma_{12}^{(1)} = \beta_{12\cdot 2}^{(1)}\sigma_{22}^{(1)} = \beta_{12\cdot 2}^{(0)}\sigma_{22}^{(1)}$. Substituting ML estimates of the identified parameters in these expressions yields

$$\hat{\mu}_2 = \bar{y}_2 + (\hat{\mu}_1 - \bar{y}_1)/b_{12\cdot 2} \tag{15.23}$$
$$\hat{\sigma}_{22} = s_{22} + (\hat{\sigma}_{11} - s_{11})/b_{12\cdot 2}^2 \tag{15.24}$$
$$\hat{\sigma}_{12} = s_{12} + (\hat{\sigma}_{11} - s_{11})/b_{12\cdot 2}. \tag{15.25}$$

Equations (15.23)–(15.25) define the protective estimators proposed by Brown (1990); Eq. (15.23) effectively imputes for missing values of Y_2 by the inverse regression of Y_1 on Y_2, as in calibration problems.

The restrictions (15.22), unlike Eq. (15.21), involve parameters that are not distinct from the identified parameters ϕ_{id}. As a result, Eqs. (15.23)–(15.25) may need modifications to ensure that parameter estimates 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 cases with the variance of Y_1 for incomplete cases.

More generally, suppose that missingness of Y_2 given Y_1 and Y_2 depends only on $Y_2^* = Y_1 + \lambda Y_2$, for $\lambda \neq 0$, and assume for the present that λ is known. Then the conditional distribution of Y_1 given Y_2^* is independent of pattern, that is, for observation i:

$$f(y_{i1}|y_{i2}^*, \phi^{(1)}, M_i = 1) = f(y_{i1}|y_{i2}^*, \phi^{(0)}, M_i = 0). \tag{15.26}$$

ML estimates of the mean and variance of Y_2^* and the covariance of Y_1 and Y_2^* then have the form of Eqs. (15.23)–(15.25), subject to the parameter-space restrictions

noted above. They can be transformed to obtain ML estimates of μ and Σ. This process yields (after some algebra),

$$\hat{\mu}_2 = \bar{y}_2 + b_{21\cdot1}^{(\lambda)}(\hat{\mu}_1 - \bar{y}_1) \tag{15.27}$$

$$\hat{\sigma}_{22} = s_{22} + (b_{21\cdot1}^{(\lambda)})^2(\hat{\sigma}_{11} - s_{11}) \tag{15.28}$$

$$\hat{\sigma}_{12} = s_{12} + b_{21\cdot1}^{(\lambda)}(\hat{\sigma}_{11} - s_{11}), \tag{15.29}$$

where

$$b_{21\cdot1}^{(\lambda)} = \frac{\lambda s_{22} + s_{12}}{\lambda s_{12} + s_{11}}. \tag{15.30}$$

These expressions yield the ignorable ML estimates of Section 7.2.1 when $\lambda = 0$, and equal Eqs. (15.23)–(15.25) in the limit as $\lambda \to \infty$. In assessing choices of λ, let us suppose that Y_1 and Y_2 are positively correlated, as when they are repeated measurements of the same quantity; if not, Y_2 can be replaced by $-Y_2$. It may then be reasonable to assume that λ is non-negative. In that case λ serves as an index of nonignorability, taking the value zero when missingness depends solely on Y_1, the value one when missingness depends on the sum or mean of Y_1 and Y_2, and the value infinity when missingness depends entirely on Y_2.

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 15.11 and 15.12) that when $\lambda = -\beta_{12\cdot2}^{(0)}$, 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.

As with nonignorable 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 is inherent in the fact that no data are available for estimating the distribution of Y_2 given Y_1 for the incomplete cases. 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 missing-data mechanism, as shown in the next example.

The large-sample variances of Eqs. (15.27)–(15.29) 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 the posterior distribution. With Jeffreys' priors, draws of ϕ_{id} from their posterior distribution can be obtained via the following eight steps:

(i) $\pi \sim \text{Beta}(n - r + 0.5, r + 0.5)$
(ii) $1/\sigma_{22}^{(0)} \sim \chi_{r-1}^2/(rs_{22})$
(iii) $1/\sigma_{11}^{(1)} \sim \chi_{n-r-1}^2/[(n-r)s_{11}^{(1)}]$
(iv) $1/\sigma_{11\cdot2}^{(0)} \sim \chi_{r-2}^2/(rs_{11\cdot2})$

(v) $\beta_{12\cdot2}^{(0)} \sim N[b_{12\cdot2}, \sigma_{11\cdot2}^{(0)}/(rs_{22})]$
(vi) $\beta_{10\cdot2}^{(0)} \sim N(\bar{x}_1 - \beta_{12\cdot2}^{(0)}\bar{x}_2, \sigma_{11\cdot2}^{(0)}/r)$
(vii) $\mu_2^{(0)} \sim N(\bar{x}_2, \sigma_{22}^{(0)}/r)$
(viii) $\mu_1^{(1)} \sim N[\bar{x}_1^{(1)}, \sigma_{11}^{(1)}/(n-r)]$.

To satisfy the parameter constraints, the drawn value of $\sigma_{11}^{(1)}$ from (iii) must be greater than the drawn value of $\sigma_{11\cdot2}^{(0)}$ from (iv); if this is not the case then these draws are discarded and steps (iii) and (iv) 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}. For the more general model where missingness depends on $Y_2^* = Y_1 + \lambda Y_2$, $\lambda \neq 0$, the above algorithm can be applied to obtain draws of parameters of the joint distribution of Y_1 and Y_2^*, and then the draws transformed to draws of the parameters of the distribution of Y_1 and Y_2.

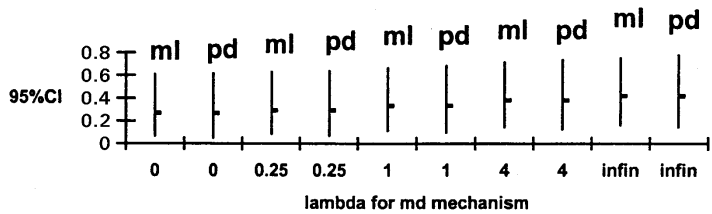


Figure 15.2. 95% confidence intervals for generated data with $\rho = 0.8$. (Example 15.12.)

EXAMPLE 15.12. *Numerical Illustration of ML and Bayes Estimates for the Bivariate Pattern-Mixture Model with Parameter Restrictions*. We illustrate these methods on two artificial data sets, with statistics $(\bar{x}_1, \bar{x}_2, s_{11}, s_{22}, \bar{x}_1^{(1)}, s_{11}^{(1)})$ 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 15.2 displays inferences for μ_2 for the pattern-mixture model for a range of values of λ, ranging from 0 to infinity, for data with $\rho = 0.8$; Figure 15.3 shows the results for data with $\rho = 0.4$. For each value of λ, two 95% intervals are displayed:

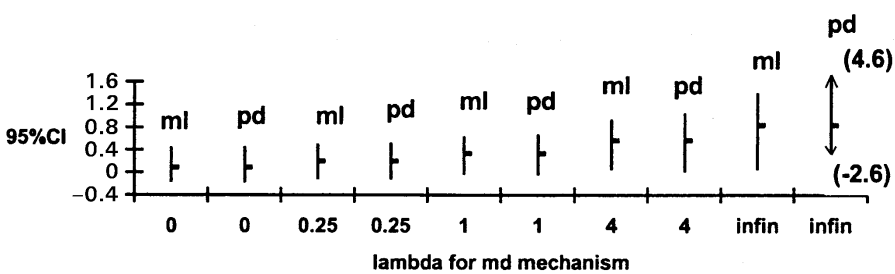


Figure 15.3. 95% confidence intervals for generated data with $\rho = 0.4$. (Example 15.1.)

(i) the ML estimate ± 2 asymptotic standard errors (labeled ml); and
(ii) the posterior mean ± 2 posterior standard errors, computed from 5000 draws from the posterior distribution of μ_2 with Jeffreys' priors (labeled pd).

The mark in each interval is the true value of μ_2 assuming the chosen value of λ is in fact correct. This quantity is computed via a population analog of Eq. (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 upwards 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 standard error. The location and size 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 predictor Y_1.

The Bayes intervals are always wider than the ML intervals. We conjecture that when the intervals deviate significantly, the Bayesian intervals provide a better reflection of uncertainty than the ML intervals. The differences between Bayes and ML are positively associated with λ and ρ. For the most extreme case with $\lambda = \infty$ and $\rho = 0.4$, the Bayesian interval $(-2.6, 4.6)$ is extreme and off the chart. This reflects problems when the drawn values of $\beta_{12\cdot 2}$ approach zero, since 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 $\beta_{12\cdot 2}$ away from zero and hence provide reliable estimates of μ_2. Histograms of draws from the posterior distributions, not shown here, look reasonably normal when λ is small, but for $\lambda = 9$ or $\lambda = \infty$ display right-skewness and outliers reflecting the occurrence of draws of $\beta_{12\cdot 2}$ near zero, particularly when $\rho = 0.4$. The distribution for $\lambda = 9$ is considerably more normal than the distribution for $\lambda = \infty$, suggesting that even modest dependence of selection on X_1 can stabilize the inference. This experiment was repeated with the sample sizes multiplied by five. Asymptotic and Bayes intervals were very close, and outliers no longer occurred with 1000 draws.

For any one of these intervals to provide a correctly calibrated inference for μ_2, the right choice of λ is needed. Presenting intervals for a range of values of λ is one way of reflecting sensitivity. Another way is to create draws from the posterior distribution, but with a different value of λ for each draw sampled from a prior distribution of plausible values. The result would be a single wider interval for μ_2. Of course this interval is sensitive to the choice of prior, but a choice that provides for nonzero λs may be more plausible than the prior implied by the standard ignorable-model analysis, which puts all the probability at the value $\lambda = 0$.

Little and Wang (1996) extend the bivariate normal pattern-mixture model (15.18) with parameter restrictions to multivariate Y_1 and Y_2 and covariates X, with Y_1 and X fully observed and Y_2 observed for r cases and missing for $n - r$

cases. Unfortunately extensions of the model to more than two patterns of missing data appear problematic (Tang, Little, and Raghunathan, 2002).

15.6. NONIGNORABLE MODELS FOR NORMAL REPEATED-MEASURES DATA

We now consider nonignorable models involving random effects for the analysis of repeated-measures data that are subject to missing data because subjects drop out prior to the end of the study.[2] For subject i, let $y_i = (y_{i1}, \ldots, y_{iK})$ represent the set of repeated measures for the ith subject, some of which may be missing. Let $y_{\text{obs},i}$ denote the observed values and $y_{\text{mis},i}$ the missing values of y_i, and let X_i denote fixed covariates. Let M_i denote a dropout indicator, taking the value 0 for the complete cases, and the value m if the subject drops out at the mth time point, $m = 1, \ldots, K$.

As in Section 11.5, correlations of the repeated-measures are modeled by subject-specific random coefficients β_i that vary across the subjects. The complete-data likelihood for the ith case is based on a statistical model for joint distribution of, Y_i, M_i, β_i, conditional on covariates X_i and fixed parameters. Selection models factor the joint distribution as:

$$f(y_i, M_i, \beta_i|X_i, \theta, \phi, \psi) = f(y_i|X_i, \beta_i, \theta) f(\beta_i|X_i, \phi) f(M_i|X_i, y_i, \beta_i, \psi). \quad (15.31)$$

Pattern-mixture models factor the joint distribution as:

$$f(y_i, M_i, \beta_i|X_i, \xi, \delta, \omega) = f(y_i|X_i, \beta_i, M_i, \xi) f(\beta_i|X_i, M_i, \delta) f(M_i|X_i, \omega). \quad (15.32)$$

The first two components of the selection model (15.31) define the joint distribution of y_i, β_i, and represent the complete-data model in the absence of missing values. For normal data, the model discussed in Section 11.5 is widely used:

$$\begin{aligned} (y_i|X_i, \beta_i) &\sim_{\text{ind}} N_K(X_{1i}\alpha + X_{2i}\beta_i, \Sigma) \\ (\beta_i|X_i) &\sim_{\text{ind}} N_q(0, \Gamma), \end{aligned} \quad (15.33)$$

where $N_p(a, B)$ denotes the p-variate normal distribution with mean α and covariance matrix B, X_{1i} is a known $(K \times p)$ design matrix containing fixed within-subject and between-subject covariates, with associated unknown $(p \times 1)$ parameter vector α, X_{2i} is a known $(K \times q)$ matrix for modeling random effects, and β_i is an unknown $(q \times 1)$ random-coefficient vector. ML estimation assuming an ignorable missing-data mechanism is widely available in programs such as Proc Mixed in SAS (SAS, 1992) or function lme in S-Plus (Pinheiro and Bates, 2000).

[2] In clinical trial settings, Meinert (1980) distinguishes *analysis* dropouts, leading to missing data, and *treatment* dropouts, where participants fail to comply with their assigned treatment or otherwise deviate from their treatment protocol. We consider here analyis dropouts, and ignore issues of compliance; when an analysis dropout is associated with treatment noncompliance, missing values can be interpreted as the values that would have occurred under continued treatment. For further discussion of the subtle issues when faced with both noncompliance and missing data, see Frangakis and Rubin (1999).

The third component of the selection model (15.31) models the probability of dropping out at a particular time as a function of other variables and random effects. There are many possible choices, depending on the assumed reasons for dropout. The simplest model is MCAR which assumes that dropping out does not depend on X_i, Y_i, or β_i . Of course, typically this assumption is too strong, because dropout does depend on treatment arm or other covariates, or on repeated measures. A weaker assumption is that dropout depends on the covariates X_i – for example dropout may depend on the treatment arm, other observed characteristics of subjects or on the time of the repeated measures – but does not depend on the repeated measures y_i or on the slopes β_i. That is,

$$f(M_i|X_i, y_i, \beta_i, \psi) = f(M_i|X_i, \psi). \tag{15.34}$$

This assumption is called covariate-dependent dropout in Little (1995). Complete-case analysis is valid under covariate-dependent dropout, but invalid if dropout depends on y_i or β_i.

A less restrictive assumption is MAR, which implies that the probability of dropping out depends only on values of variables that are observed, not on values of variables that are missing. That is:

$$f(M_i|X_i, y_{\text{obs},i}, y_{\text{mis},i}, \beta_i, \psi) = f(M_i|X_i, y_{\text{obs},i}, \psi). \tag{15.35}$$

As noted in Chapter 1, Murray and Findlay (1988) provide an instructive example of missing at random for data from a study of hypertensive drugs where the outcome was diastolic blood pressure. The subject dropped out by protocol when the diastolic blood pressure got too large—the subject was forced out of the study and changed treatments. The mechanism is not covariate-dependent here, since it depends on the values of blood pressure. But blood pressure at the time of dropout was observed before the subject dropped out. Hence the mechanism is MAR, because dropout only depends on the observed part of Y. Murray and Findlay show that results from an ML analysis differ somewhat from results from complete-case analysis, which is biased for this mechanism.

Nonignorable models are needed when the missing-data mechanism depends on the missing values or unobserved random coefficients. Little (1995) distinguishes these two kinds of nonignorable dropout. The first is outcome-dependent dropout, where dropout depends on Y. In symbols:

$$f(M_i|X_i, y_{\text{obs},i}, y_{\text{mis},i}, \beta_i, \psi) = f(M_i|X_i, y_{\text{obs},i}, y_{\text{mis},i}, \psi). \tag{15.36}$$

For example, if the repeated measures are of pain and dropout depends on the value of the pain variable at the time of dropout, that would be outcome-dependent, because missingness then would depend on the (unobserved) value of Y at the time of dropout. Diggle and Kenward (1994) consider models where the dropout depends on the current and previous values of Y. Another class of models assumes

random coefficient-dependent dropout, where the probability of dropping out depends on the underlying random coefficients β_i. In symbols:

$$f(M_i|X_i, y_i, \beta_i, \psi) = f(M_i|X_i, \beta_i, \psi). \tag{15.37}$$

For example, one of the random coefficients may represent a slope. If people who have more rapid decline tend to drop out more frequently than those with less rapid decline, dropping out depends on this underlying, unobserved slope. Wu and Carroll (1988) and Wu and Bailey (1989) consider models for slope-dependent dropout, using the term *informative censoring* to describe the dropout mechanism. We use terminology that is more explicit about the assumed form of the mechanism. In summary, Eqs. (15.36) and (15.37) represent two kinds of nonignorable missing-data models, one where missingness depends on the missing components of Y_i, and one where missingness depends on underlying random effects. The key in all these models is to decide the plausible mechanism in the particular applied setting.

Nonignorable models for repeated-measures data can also be formulated based on the pattern-mixture factorization (15.32), as in the following example.

EXAMPLE 15.13. *A Pattern-Mixture Model Where Dropout Depends on Random Effects*. If the dropout mechanism depends on X_i and the regression parameters β_i, then the pattern-mixture factorization yields

$$f(y_i|X_i, \beta_i, M_i, \xi) = f(y_i|X_i, \beta_i, \xi).$$

A pattern-mixture extension of the normal model (15.33) is then:

$$\begin{aligned}(y_i|X_i, \beta_i, M_i) = (y_i|X_i, \beta_i) &\sim_{\text{ind}} N_K(X_{1i}\alpha + X_{2i}\beta_i, \Sigma)\\ (\beta_i|X_i, M_i = m) &\sim_{\text{ind}} N_q(X_{3i}\beta^{(m)}, \Gamma^{(m)}),\\ (M_i|X_i) &\sim \text{Multinomial}(\pi_{i0}, \ldots, \pi_{iK})\end{aligned} \tag{15.38}$$

where α, Σ, and $\{\beta^{(m)}, \Gamma^{(m)}, m = 1, \ldots, K\}$ are the fixed parameters; X_{1i}, X_{2i}, X_{3i} are known design matrices, and $\{\pi_{im}, m = 1, \ldots, K\}$ may be functions of unknown fixed parameters ρ. The introduction of X_{3i} in Eq. (15.38) leads to some redundancy in the notation, but it allows random coefficients to depend directly on between-subject covariates.

A special case of this model for repeated measures data on subjects in J treatment groups is:

$$\begin{aligned}(y_i|\beta_i) &\sim_{\text{ind}} N_K\left(\begin{bmatrix}1 & t_{i1}\\ \cdots & \cdots\\ 1 & t_{iK}\end{bmatrix}\begin{bmatrix}\beta_{i0}\\ \beta_{i1}\end{bmatrix}, \sigma_e^2 I\right),\\ (\beta_i|x_i = j, M_i = m) &\sim_{\text{ind}} N_2(\beta_j^{(m)}, \Gamma), \qquad j = 1, \ldots, J,\\ (M_i|x_i = j) &\sim_{\text{ind}} \text{Multinomial}(\pi_{j0}, \ldots, \pi_{jK}),\end{aligned} \tag{15.39}$$

where $\beta_i^{\mathrm{T}} = (\beta_{i0}, \beta_{i1})$ represents a random intercept and slope for subject i, x_i is an indicator for treatment group, π_{jk} represents the proportion of the population given treatment j for which $M_i = m$, $\beta_j^{(m)} = (\beta_{j0}^{(m)}, \beta_{j1}^{(m)})^{\mathrm{T}}$ is the expected intercept and slope for subjects in treatment group j with dropout pattern $M_i = m$, $\Sigma = \sigma_e^2 I_K$, and $\Gamma^{(m)} = \Gamma$.

This model can be estimated using existing software such as SAS Proc Mixed by including as covariates dummy variables for dropout time and treatment; $\{\pi_{jm}, m = 1, \ldots, K\}$ are estimated as the sample proportion in treatment group j with each missing-data pattern m. The quantities of interest are the expected intercept and slope for subjects in treatment group j, averaged over the missing-data pattern, that is

$$E(\beta_i | x_{3i} = j) = \sum_{m=1}^{K} \pi_{jm} \beta_j^{(m)}. \tag{15.40}$$

ML estimates of these parameters are obtained as a weighted sum of the ML estimates $\hat{\beta}_j^{(m)}$ of the expected intercept and slope for treatment j, pattern m, with weights given as the proportion $\hat{\pi}_{jm}$ of cases with treatment j that have pattern m. This contrasts with a MAR model, where estimates for each pattern are effectively weighted by their precisions. A problem here is that $\beta_j^{(1)}$ and $\beta_j^{(2)}$ are not identified, since ML estimation of straight lines requires at least two observed time points, and these patterns contain at most one observation. These cases might simply be omitted from the analysis.

Simulations (Wang-Clow et al., 1995) suggest that estimates from this model can have poor precision. One way to address this problem is to put additional structure on $\beta_j^{(m)}$. One possible assumption is that the expected intercept is independent of pattern, and the expected slope is linearly related to the dropout time $t^{(m)}$ for pattern m:

$$\beta_{j0}^{(m)} = \beta_{j0}; \qquad \beta_{j1}^{(m)} = \alpha_{j0} + \alpha_{j1} t^{(m)}. \tag{15.41}$$

Equations (15.39) and (15.41) formalize the model underlying the "conditional linear model" estimates of Wu and Bailey (1989), also considered by Mori, Woolson and Woodsworth (1994) in the context of empirical Bayes estimation of the random slopes. Clearly, many other models might be postulated within the class given by Eq. (15.38).

15.7. NONIGNORABLE MODELS FOR CATEGORICAL DATA

We conclude by discussing some nonignorable models for categorical data. Two types of nonignorable models for incomplete categorical data have been considered. Pregibon (1977), Little (1982), and Nordheim (1984) introduce prior odds of response for categories of the 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 (1985), Fay (1986), and Little (1985b). We consider the latter approach here, as it is closer in spirit to the contingency tables models discussed in Chapter 13. Unlike those models, the nonignorable 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 15.14. *Two-Way Contingency Table with One Supplemental Margin.* Suppose data are as in Example 13.1, with n observations on two categorical variables, Y_1 with levels $j = 1, \ldots, J$ and Y_2 with levels $k = 1, \ldots, K$, 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 shall fit models to the data set in Table 15.7, with $J = K = 2$.

Now define M to take the value 1 if Y_2 is missing, 0 if Y_2 is observed. Suppose that for fixed n, complete observations have a multinomial distribution over the $J \times K \times 2$ 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 $\Pr(Y_1 = j, Y_2 = k, M = 0) = \pi_{jk}(1 - \phi_{jk})$ where $\{\pi_{jk}\}$ and $\{\phi_{jk}\}$ are assumed distinct. 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 $2JK - 1 - (JK + J - 1) = J(K - 1)$ too many parameters in the unrestricted (saturated) model for unique ML estimates. 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 Ys; here we are modeling the joint distribution of both the Ys and the response indicator, M.)

All the hierarchical models that include the main effects of Y_1, Y_2, and M are displayed in Table 15.8. 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, in the sense that they do not

Table 15.7 A 2×2 Contingency Table with One Partially Classified Margin

Fully Classified ($M = 0$)

Y_1	$Y_2 = 1$	$Y_2 = 2$	
1	$r_{11} = 100$	$r_{12} = 20$	$r_{1+} = 120$
2	$r_{21} = 30$	$r_{22} = 50$	$r_{2+} = 80$
	$r_{+1} = 130$	$r_{+2} = 70$	$r = 200$

Partially Classified ($M = 1$)

Y_1	$Y_2 = 1$	$Y_2 = 2$	
1	$m_{11} = ?$	$m_{12} = ?$	$m_1 = 40$
2	$m_{21} = ?$	$m_{22} = ?$	$m_2 = 60$
			$m = 100$

appear in the likelihood and hence are not identified. 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 15.7—the likelihood ratio chi-squared statistic for lack of fit, its associated degrees of freedom, and estimates of the cell probabilities ($\times 100$).

The following properties of the models in Table 15.8 merit some discussion:

1. *Inestimability.* The models $\{Y_1Y_2M\}$, $\{Y_1Y_2, Y_1M, Y_2M\}$, $\{Y_1M, Y_2M\}$, $\{Y_1, Y_2M\}$, and, if $K > J$, $\{Y_1Y_2, Y_2M\}$ have inestimable parameters. Additional information is needed to estimate the cell probabilities for these models, so estimated probabilities are not given in the table.

   Note that two of these models, $\{Y_1M, Y_2M\}$ and $\{Y_1, Y_2M\}$ 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 $\{Y_1M, Y_2M\}$. The model has $2J + 2K - 3$ parameters—one for the marginal probability of response, $J + K - 2$ 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, less $(J - 1)(K - 1)$ degrees of freedom since Y_1 and Y_2 are independent, given M. The incomplete-data likelihood factorizes into three 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 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 conditional independence assumption of Y_1 and Y_2 given $M = 0$.
2. *Ignorability.* The models $\{Y_1Y_2, Y_1M\}$ and $\{Y_2, Y_1M\}$ are ignorable since missingness depends only on Y_1, which is fully observed. These models can be fitted using the methods of Chapter 13. The models $\{Y_1Y_2, M\}$ and $\{Y_1, Y_2, M\}$ are also ignorable, since they assume that nonresponse is independent of Y_1 and Y_2 that is, that the data are MCAR. They yield the same estimates of $\{\pi_{jk}\}$ as their MAR counterparts, $\{Y_1Y_2, Y_1M\}$ and $\{Y_2, Y_1M\}$ respectively.
3. *Lack of Fit.* The lack-of-fit chi-squared for $\{Y_1Y_2, M\}$ is based on a test of independence of Y_1 and M, using the $Y_1 \times M$ two-way margin. The lack-of-fit chi-squared for $\{Y_1M, Y_2\}$ is based on a test of independence of Y_1 and Y_2, using the fully classified data. The lack-of-fit chi-squared for $\{Y_1, Y_2, M\}$ is found by summing the chi-squared statistics for $\{Y_1Y_2, M\}$ and $\{Y_1M, Y_2\}$.

Table 15.8 Models for a Two-Way Table with One Supplemental Margin

					Example from Table 15.7					
		Degrees of Freedom			Lack of Fit		Estimated Cell Prob × 100			
	Model	Model	Lack of Fit	Inestimable	χ^2	df	π_{11}	π_{12}	π_{21}	π_{22}
(1)	$\{Y_1Y_2M\}$	$2JK-1$	0	$J(K-1)$	—	—	—	—	—	—
(2)	$\{Y_1Y_2, Y_1M, Y_2M\}$	$JK+J+K-2$	0	$K-1$	—	—	—	—	—	—
(3)	$\{Y_1Y_2, Y_1M\}$	$JK+J-1$	0	0	0	0	44.4	8.9	17.5	29.2
(4)	$\{Y_1Y_2, Y_2M\}$	$JK+K-1$	$\max(J-K,0)$	$\max(K-J,0)$	0	0	39.4	14.0	11.8	34.9
(5)	$\{Y_1Y_2, M\}$	JK	$J-1$	0	10.75	1	44.4	8.9	17.5	29.2
(6)	$\{Y_1M, Y_2M\}$	$2(J+K)-3$	$(J-1)(K-1)$	$K-1$	44.99	1	—	—	—	—
(7)	$\{Y_1M, Y_2\}$	$2J+K-2$	$(J-1)(K-1)$	0	44.99	1	34.7	18.7	30.3	16.3
(8)	$\{Y_1, Y_2M\}$	$2K+J-2$	$(J-1)K$	$K-1$	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

4. *Estimation*. The ML estimate of π_{jk} for $\{Y_1Y_2, Y_1M\}$ or $\{Y_1Y_2, M\}$ is $\hat{\pi}_{jk} = (r_{jk} + \hat{m}_{jk})/(r+m)$ 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 $\{m_j\}$ into the table to match the *row* distributions $\{r_{jk}/r_{j+}\}$ of the fully observed data, as in Examples 13.1 and 13.2.

Only one of the five nonignorable models in Table 15.8 can be fitted to data without additional prior information, namely, $\{Y_1Y_2, Y_2M\}$ which can be estimated if $K \leqslant J$. The model supposes that response to 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 $\Sigma_k \hat{m}^*_{jk} = m_j$ for all j, yield $JK - K + J$ linear equations for the JK unknowns $\hat{m}^*_{jk}$. When $K > J$ there are fewer equations than parameters, and hence *a priori* constraints are required to define uniquely $\{\hat{m}^*_{jk}\}$ (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 (1985).] When $K = J$, the JK linear equations can be solved directly, yielding estimates without resorting to the EM algorithm iterations. In particular, for $J = K = 2$ we obtain the following equations for $\hat{m}^*_{11}$, $\hat{m}^*_{12}$, $\hat{m}^*_{21}$, and $\hat{m}^*_{22}$:

$$\hat{m}^*_{21} = \hat{m}^*_{11} r_{21}/r_{11}; \qquad \hat{m}^*_{22} = \hat{m}^*_{12} r_{22}/r_{12}; \qquad \hat{m}^*_{11} + \hat{m}^*_{12} = m_1;$$
$$\hat{m}^*_{21} + \hat{m}^*_{22} = m_2.$$

Solving 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

$$\hat{m}^*_{11} = 200/11, \qquad \hat{m}^*_{12} = 240.11, \qquad \hat{m}^*_{21} = 60/11, \qquad \hat{m}^*_{22} = 600/11,$$

which yield the estimates of $\{\pi_{jk}\}$ in row (4) of Table 15.8.

The estimates obtained from solving these linear equations can be negative, and hence not ML. Baker and Laird (1985) show that to yield non-negative estimates $\{\hat{m}^*_{jk}\}$, the marginal column odds $\{m_j/m_l\}$ must lie between the smallest and largest values of the column odds $\{r_{jk}/r_{lk}\}$ $k = 1, \ldots, K$. 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 $\hat{m}^*_{jk} \geqslant 0$ for all j, k. Details are given in Baker and Laird (1985).

5. *Choice Between Models*. It is important to note that in our example both the models $\{Y_1Y_2, Y_1M\}$ and $\{Y_1Y_2, Y_2M\}$ yield perfect fits to the data with no degrees of freedom for fit. Thus it is not possible to choose between the estimates of $\{\pi_{jk}\}$ they supply, except by *a priori* reasoning about which nonresponse 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 response to Y_1 and Y_2, and models for the four-way table of Y_1, Y_2, M_1 and M_2 are considered. Higher-order tables can also be considered, at least in principle.

EXAMPLE 15.15. *Predicting Results of the Slovenian Plebiscite with Polling Data*. Prior to the Slovenian Plebiscite in 1991, in which 88.5% of eligible Slovenians voted to create an independent state, the Slovenian Public Opinion Survey (SPOS) collected information on the likely outcome of that vote. 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 ignorable and nonignorable nonresponse models.

Table 15.9, taken from Rubin, Stern, and Vehovar (1995), summarizes results from the survey for three categorical variables: "Attendance" concerns whether the respondents will 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 "Don't 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. The reason is 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, since a "non- vote" from an eligible voter counts the same as a "No" vote against independence. The critical questions in Table 15.9 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 useful covariate information for the other two questions.

Table 15.9 Results of Slovenian Public Opinion Survey. (Example 15.15.)

		Independence		
Secession	Attendance	Yes	No	Don't Know
	Yes	1191	8	21
Yes	No	8	0	4
	Don't Know	107	3	9
	Yes	158	68	29
No	No	7	14	3
	Don't Know	18	43	31
	Yes	90	2	109
Don't Know	No	1	2	25
	Don't Know	19	8	96

Table 15.10 Slovenian Public Opinion Poll Survey: Comparison of Estimates on Independence Question for Different Methods of Handling Missing Data. (Example 15.15.)

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
Nonignorable	0.782	0.218	0.122
Plebiscite	0.885	0.115	0.065

The results of various approaches to address nonresponse in the SPOS are displayed in Table 15.10. The conservative approach assumed every "Don't 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 ignorable ML estimates are based on the EM algorithm applied to a saturated ignorable multinomial model for the $2 \times 2 \times 2$ data of Table 15.9, as in Section 13.3. A DA algorithm for the same model and data with a Jeffreys' prior distribution as in Example 6.17, yielded a posterior median that was the same as the ML estimate to within a tenth of a percent.

The nonignorable model in Table 15.10 assumed that nonresponse on a question was a function of the answer to that question. More specifically, including the missing-data indicators for the data of Table 15.9 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 15.10 shows the results of the plebiscite that the opinion poll was attempting to predict. The only estimates that are close to this outcome are ones based on the ignorable model, despite the fact that the nonignorable model might be regarded as reasonably sensible.

In our limited experience, this is not an uncommon outcome. When good information on nonrespondents is available in carefully conducted surveys, ignorable missing-data models are often seen to outperform nonignorable models. This is not to argue that the missingness mechanisms operating in these surveys are really ignorable, but rather that the formulation of nonignorable models that are superior to ignorable models is very context-specific and not easy.

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 expressions for the E step in Example 15.7.

15.4. Justify the M step computations given at the end of Example 15.7. In particular, why is the estimate of σ_1^2 not given simply from the regression of Y_1 on X?

15.5. 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 Example 15.7 (see, e.g., Little 1985a).

15.6. Consider the selection model of Example 15.7 when $(x_i = x_{i1}, z_i)$, where z_i is a single binary variable predictive of selection but with coefficient zero in the regression on Y_1 on X. The following table gives the means of y_{i1} given x_i, classified by z_i and by whether y_{i1} is observed $(M_{i1} = 0)$ or missing $(M_{i1} = 1)$.

		z_i	
		0	1
M_{i1}	0	$x_i\beta_1 + \rho\sigma_1\lambda(\gamma_{i0})$	$x_i\beta_1 + \rho\sigma_1\lambda(\gamma_{i1})$
	1	$x_i\beta_1 + \rho\sigma_1\lambda(-\gamma_{i0})$	$x_i\beta_1 + \rho\sigma_1\lambda(-\gamma_{i1})$
	ALL	$x_i\beta_1$	$x_i\beta_1$

In the table, $\lambda(\cdot)$ is defined as in Example 15.5, and γ_{ij} is the mean of Y_2 for units with values $(x_i, z_i = j)$ of the covariates.

(a) Derive the expressions in the table.

(b) By considering the difference in means of Y_1 between responding and nonresponding cases for several values of γ_{ij}, show that the model implies that the means in the table have an approximately additive structure. (See Little, 1985a, for more details.)

15.7. Suppose that for the model of Example 15.7 a random subsample of nonrespondents to Y_1 is followed up, and values of Y_1 obtained. Write down the loglikelihood for the resulting data and describe the E and M steps of the EM algorithm.

15.8. Derive the expressions for the posterior mean and variance of $\bar{y}$ in Example 15.10. What is the posterior mean and variance of variable 32D when $\psi_1 = \psi_2 = 0.5$?

15.9. Section 15.5.2 shows that for the pattern-mixture model (15.18) with MAR restrictions (15.21), 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).

15.10. Fill in the details leading to the ML estimates (15.23)–(15.25) for the pattern-mixture model (15.18) under restrictions (15.22), and Eqs. (15.27)–(15.29) for the pattern-mixture model (15.18) under restrictions (15.26).

15.11. 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.

15.12. For the pattern-mixture model (15.18) 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 ignorable ML 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).

15.13. For suitable parameterizations of the models, write down factored likelihoods for the models $\{Y_1Y_2, Y_1M\}$, $\{Y_1Y_2, Y_2M\}$, $\{Y_1M, Y_2M\}$, $\{Y_1, Y_2M\}$ in Example 15.14. State for each model which parameters (if any) are inestimable.

15.14. Verify the five sets of estimated cell probabilities in Table 15.8.

15.15. Redo Table 15.8 for the data in Table 15.7 with m_1 and m_2 multiplied by a factor of ten.

15.16. Reproduce the EM estimates for the ignorable model in Table 15.10. Use the bootstrap to compute the standard error of the estimated proportion voting yes.

15.17. Reproduce the Bayes estimates for the ignorable model in Table 15.10 by data augmentation, and provide a histogram of draws. Use five 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.

15.18. Repeat the calculations in Problem 15.17 for the nonignorable model in Table 15.10.

References

Afifi, A.A., and Elashoff, R.M. (1966). Missing observations in multivariate statistics: Review of the literature, *J. Am. Statist. 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. Roy. Statist. Soc. B* **47**, 67–75.

Aitkin, M., and Wilson, G.T. (1980). Mixture models, outliers, and the EM algorithm, *Technometrics* **22**, 325–331.

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. Econometrics* **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. Statist. Assoc.* **52**, 200–203.

Anderson, T.W. (1965). *An Introduction to Multivariate Statistical Analysis*. New York: Wiley.

Angrist, J.D., Imbens, G.W., and Rubin, D.B. (1996). Identification of causal effects using instrumental variables (with discussion), *J. Am. Statist. Assoc.* **91**, 444–472.

Azen, S., and Van Guilder, M. (1981). Conclusions regarding algorithms for handling incomplete data, *Proc. Stat. Computing Sec.*, *Am. Statist. Assoc. 1981*, 53–56.

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*, *Vol. III: Symposium on Incomplete Data*, *Proceedings* (W.G. Madow and I. Olkin, eds.), New York: Academic Press.

Bailar, B.A., Bailey, L., and Corby, C. (1978). A comparison of some adjustment and weighting procedures for survey data, *Proc. Survey Res. Meth. Sec.*, *Am. Statist. Assoc. 1978*, 175–200.

Baker, S., and Laird, N. (1985). Categorical response subject to nonresponse, Technical Report, Department of Biostatistics, Harvard School of Public Health, Boston, MA.

Bard, Y. (1974). *Nonlinear Parameter Estimation*. New York: Academic Press.

Barnard, J., Du, J., Hill, J., and Rubin, D.B. (1998). A broader template for analyzing broken randomized experiments, *Sociol. Meth. Res.* **27**, 285–318.

Barnard, J., and Rubin, D.B. (1999), Small-sample degrees of freedom with multiple imputation, *Biometrika* **86**, 949–955.

Bartlett, M.S. (1937). Some examples of statistical methods of research in agriculture and applied botany, *J. Roy. Statist. 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. Statist.* **41**, 164–171.

Beale, E.M.L., and Little, R.J.A. (1975). Missing values in multivariate analysis, *J. Roy. Statist. 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, *Statist. Meth. Med. Res.* **6**, 38–54.

Bentler, P.M., and Tanaka, J.S. (1983). Problems with EM for ML factor analysis, *Psychometrika* **48**, 247–253.

Berndt, E.B., Hall, B., Hall, R., and Hausman, J.A. (1974). Estimation and inference in nonlinear structural models, *Ann. Econ. Soc. Meas.* **3**, 653–665.

Besag, J. (1986). On the statistical analysis of dirty pictures, *J. Roy. Statist. Soc. B* **48**, 259–279.

Bethlehem, J.G. (2002). Weighting adjustments for ignorable nonresponse, Chapter 18, in *Survey Nonresponse* (R.M. Groves, D.A. Dillman, J.L. Eltinge, and R.J. Little, eds.), 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.

Box, G.E.P., and Cox, D.R. (1964). An analysis of transformations, *J. Roy. Statist. Soc. B* **26**, 211–252.

Box, G.E., Hunter, J.S., and Hunter, W.G. (1985). *Statistics for Experimenters: an Introduction to Design*, *Data Analysis & Model Building*. New York: Wiley.

Box, G.E.P., and Jenkins, G.M. (1976). *Time Series Analysis: Forecasting and Control*. San Francisco: 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.

Breslow, N.E., and Clayton, D.G. (1993). Approximate inference in generalized linear mixed models, *J. Am. Statist. 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. Roy. Statist. Soc. B* **22**, 302–306.

Cassel, C.M., Sarndal, C.E., and Wretman, J.H. (1983). Some uses of statistical models in connection with the nonresponse problem, in *Incomplete Data in Sample Surveys*, *Vol. III: Symposium on Incomplete Data*, *Proceedings* (W.G. Madow and I. Olkin, eds.), New York: Academic Press.

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*, 3rd ed. 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.

Colledge, M.J., Johnson, J.H., Paré, R., and Sande, I.G. (1978). Large scale imputation of survey data, *Proc. Survey Res. Meth. Sec.*, *Am. Statist. Assoc. 1978*, 431–436.

Cox, D.R., and Hinkley, D.V. (1974). *Theoretical Statistics*. New York: Wiley.

David, M.H., Little, R.J.A., Samuhel, M.E., and Triest, R.K. (1983). Imputation methods based on the propensity to respond, *Proc. Bus. Econ. Sec.*, *Am. Statist. Assoc. 1983*, 168–173.

David, M.H., Little, R.J.A., Samuhel, M.E., and Triest, R.K. (1986). Alternative methods for CPS income imputation, *J. Am. Statist. 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. Statist.* **8**, 264–278.

Dempster, A.P. (1969). *Elements of Continuous Multivariate Analysis*. Reading, MA: Addison-Wesley.

Dempster, A.P., Laird, N.M., and Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion), *J. Roy. Statist. Soc. B* **39**, 1–38.

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 Analysis* **V**, 35–37.

Dempster, A.P., and Rubin, D.B. (1983). Introduction, pp. 3–10, in *Incomplete Data in Sample Surveys (Volume 2): Theory and Bibliography* (W.G Madow, I. Olkin, and D.B. Rubin, eds.), New York: Academic Press.

Dempster, A.P., Rubin, D.B., and Tsutakawa, R.K. (1981). Estimation in covariance component models, *J. Am. Statist. Assoc.* **76**, 341–353.

Diggle, P., and Kenward, M.G. (1994). Informative drop-out in longitudinal data analysis, *J. Roy. Statist. Soc. C* **43**, 49–73.

Dixon, W.J. (1988). *BMDP Statistical Software*. Berkeley CA: University of California Press.

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.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife, *Ann. Stat.* **7**, 1–26.

Efron, B. (1987). Better bootstrap confidence intervals, *J. Am. Statist. Assoc.* **82**, 171–200 (with discussion).

Efron, B. (1994). Missing data, imputation, and the bootstrap, *J. Am. Statist. 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.

Ekholm, A., and Skinner, C. (1998). The Muscatine children's obesity data reanalysed using pattern mixture models, *Appl. Statist.* **47**, 251–264.

Ernst, L.R. (1980). Variance of the estimated mean for several imputation procedures, *Proc. Survey Res. Meth. Sec.*, *Am. Statist. Assoc. 1980*, 716–721.

Ezzati-Rice, T., Johnson, W., Khare, M., Little, R.J.A., Rubin, D., and Schafer, J. (1995). A simulation study to evaluate the performance of model-based multiple imputations in NCHS health examination surveys. *Proc. 1995 Annual Res. Conf.*, *U.S. Bureau of the Census*, 257–266.

Fay, R.E. (1986). Causal models for patterns of nonresponse, *J. Am. Statist. Assoc.* **81**, 354–365.

Fay, R.E. (1992). When are inferences from multiple imputation valid? *Proc. Survey Res. Meth. Sec.*, *Am. Statist. Assoc.*, 227–232.

Fay, R.E. (1996). Alternative paradigms for the analysis of imputed survey data, *J. Am. Statist. Assoc.* **91**, 490–498.

Fienberg, S.E. (1980). *The Analysis of Crossclassified Data*, 2nd ed. Cambridge, MA: MIT Press.

Firth, D. (1991). Generalized linear models, pp. 55–82, in *Statistical Theory and Modelling in Honour of Sir David Cox*. New York: Chapman & Hall.

Ford, B.N. (1983). An overview of hot deck procedures, in *Incomplete Data in Sample Surveys*, *Vol. II: Theory and Annotated Bibliography* (W.G. Madow, I. Olkin, and D.B. Rubin, eds.), 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 idiosyncracy in estimating survival curves using double sampling in the presence of self-selected right censoring, *Biometrics* **57**, 333–342. With discussion and rejoinder, pp. 343–353.

Frangakis, C., and Rubin, D.B. (2002). Principal stratification in causal inference, *Biometrics* **58**, 21–29.

Fuchs, C. (1982). Maximum likelihood estimation and model selection in contingency tables with missing data, *J. Am. Statist. Assoc.* **77**, 270–278.

Gelfand, A.E., and Smith, A.F.M. (1990). Sampling-based approaches to calculating marginal densities, *J. Am. Statist. 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. Statist. Assoc.* **85**, 972–985.

Gelman, A.E., and Carlin, J.B. (2002). Poststratification and weighting adjustments, Chapter 19, in *Survey Nonresponse* (R.M. Groves, D.A. Dillman, J.L. Eltinge, and R.J.A. Little, eds.), New York: Wiley.

Gelman, A.E., Carlin, J.B, Stern, H.S., and Rubin, D.B. (1995). *Bayesian Data Analysis*. London: Chapman & Hall.

Gelman, A.E., and Meng, X.L. (1998). Computing normalizing constants: from importance sampling to bridge sampling to path sampling, *Statist. Sci.* **13**, 163–185.

Gelman, A.E., and Rubin, D.B. (1992). Inference from iterative simulation using multiple sequences (with discussion), *Statist. Sci.* **7**, 457–472.

Geyer, C.J. (1992). Practical Markov chain Monte Carlo (with discussion), *Statist. Sci.* **7**, 473–503.

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, pp. 115–142, in *Drawing Inferences from Self-Selected Samples* (H. Wainer, ed.), New York: Springer-Verlag.

Glynn, R.J., Laird, N.M., and Rubin, D.B. (1993). Multiple imputation in mixture models for nonignorable nonresponse with follow-ups, *J. Am. Statist. Assoc.* **88**, 984–993.

Goodman, L.A. (1970). The multivariate analysis of qualitative data: Interaction among multiple classifications, *J. Am. Statist. Assoc.* **65**, 225–256.

Goodman, L.A. (1979). Simple models for the analysis of association in crossclassifications having ordered categories, *J. Am. Statist. Assoc.* **74**, 537–552.

Goodnight, J.H. (1979). A tutorial on the SWEEP operator, *Am. Statist.* **33**, 149–158.

Goodrich, R.L., and Caines, P.E. (1979). Linear system identification from nonstationary cross-sectional data, *IEEE Trans. Aut. Control* **AC-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. Statist. Assoc.* **77**, 251–261.

Groves, R., Dillman, D., Eltinge, J., and Little, R.J.A. (2002) *Survey Nonresponse*. New York: Wiley.

Gupta, N.K., and Mehra, R.K. (1974). Computational aspects of maximum likelihood estimation and reduction in sensitivity function calculations, *IEEE Trans. Aut. Control* **AC-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. Roy. Statist. Soc. B* **30**, 67–81.

Hajek, J. (1960). Limiting distributions in simple random sampling from a finite population, *Pub. Math. Inst. Hung. Acad. Sci.* **4**, 49–57.

Hansen, M.H., Hurwitz, W.N., and Madow, W.G. (1953). *Sample Survey Methods and Theory*, two volumes. New York: Wiley.

Hanson, R.H. (1978). The Current Population Survey: design and methodology, Technical Paper No. 40, U.S. Bureau of the Census.

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.

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 (with discussion), *J. Am. Statist. Assoc.* **72**, 320–340.

Hasselblad, V., Stead, A.G., and Galke, W. (1980). Analysis of coarsely grouped data from the lognormal distribution, *J. Am. Statist. 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. Statist.* **5**, 203–206.

Heckman, J.I. (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.A., and Raghunathan, T. (2002). Multivariate imputation of coarsened survey data on household wealth, Chapter 24, in *Survey Nonresponse* (R.M. Groves, D.A. Dillman, J.L. Eltinge, and R.J.A. Little, eds.), New York: Wiley.

Heitjan, D. F. (1994). Ignorability in general incomplete-data models, *Biometrika*. **81**, 701–708.

Heitjan, D., and Rubin, D.B. (1991). Ignorability and coarse data, *Ann. Statist.* **19**, 2244–2253.

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*, *Volume 2: Theory and Bibliography* (W.G. Madow, I. Olkin, and D.B. Rubin, eds.), New York: Academic Press, pp. 209–245.

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., and Wightman, L.E. (1982). Section pre-equating: A preliminary investigation, in *Test Equating* (P.W. Holland and D.B. Rubin, eds.), New York: Academic Press.

Holt, D., and Smith, T.M.F. (1979). Post stratification, *J. Roy. Statist. Soc. A* **142**, 33–46.

Horton, N.J., and Laird, N.M. (1998). Maximum likelihood analysis of generalized linear models with missing covariates, *Statist. Meth. 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. Statist. Assoc.* **47**, 663–685.

Huber, P.J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions, *Proc 5th. Berkeley Symp. in Math. Stat. and Prob.*

Ibrahim, J.G. (1990). Incomplete data in generalized linear models, *J. Am. Statist. 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. Roy. Statist. Soc. B* **61**, 173–190.

Ireland, C.T., and Kullback, S. (1968). Contingency tables with given marginals, *Biometrika* **55**, 179–188.

Jamshidian, M., and Jennrich, R.I. (1993). Conjugate gradient acceleration of the EM algorithm, *J. Am. Statist. Assoc.* **88**, 221–228.

Jarrett, R.G. (1978). The analysis of designed experiments with missing observations, *Appl. Statist.* **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.

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, *Proc. Survey Res. Meth. Sec.*, *Am. Statist. Assoc. 1981*, 146–151.

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, *Proc. Sec. Surv. Res. Meth.*, *Am. Statist. Assoc.*, 1–10.

Kenward, M.G., and Molenberghs, G. (1998). Likelihood based frequentist inference when data are missing at random, *Statist. Sci.* **13**, 236–247.

Kent, J.T., Tyler, D.E., and Vardi, Y. (1994). A curious likelihood identity for the multivariate T-distribution, *Commun. Statist. B—Simulation and Computation* **23**, 441–453.

Kim, J.O., and Curry, J. (1977). The treatment of missing data in multivariate analysis, *Sociol. Meth. Res.* **6**, 215–240.

Kleinbaum, D.G., Morgenstern, H., and Kupper, L.L. (1981). Selection bias in epidemiological studies, *Am. J. Epidem.* **113**, 452–463.

Kong, A., Liu, J.S., and Wong, W.H. (1994). Sequential imputations and Bayesian missing data problems, *J. Am. Statist. 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: Almquist and Wiksell, and New York: 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. Roy. Statist. Soc. B* **57**, 425–437.

Lange, K. (1995b). A quasi-Newtonian acceleration of the EM algorithm, *Statistica Sinica* **5**, 1–18.

Lange, K., Little, R.J.A., and Taylor, J.M.G. (1989). Robust statistical inference using the *t* distribution, *J. Am. Statist. Assoc.* **84**, 881–896.

LaVange, L.M., and Helms, R.W. (1983). The analysis of incomplete longitudinal data with modeled covariance matrices, Mimeo 1449, Inst. of Statistics, University of N. Carolina.

Lavori, P.W., Dawson, R., and Shera, D. (1995). A multiple imputation strategy for clinical trials with truncation of patient data, *Stat. Med.* **14**, 1913–1925.

Lazzeroni, L.C., and Little, R.J. (1998). Random-effects models for smoothing post-stratification weights, *J. Official Statist.* **14**(1), 61–78.

Ledolter, J. (1979). A recursive approach to parameter estimation in regression and time series problems, *Commun. Statist. Theory Meth.* **A8**, 1227–1245.

Lee, H., Rancourt, E., and Sarndal, C.E. (2002). Variance estimation from survey data under single imputation, Chapter 21, in *Survey Nonresponse* (R.M. Groves, D.A. Dillman, J.L.A. Eltinge, and R.J.A. Little, eds.), New York: Wiley.

Lee, Y., and Nelder, J.A. (1996). Hierarchical generalized linear models, *J. Roy. Statist. Soc. Ser. B* **58**, 619–678 (with discussion).

Li, K.H., Meng, X.-L., Raghunathan, T.E., and Rubin, D.B. (1991). Significance levels from repeated *p*-values with multiply-imputed data, *Statist. Sinica* **1**, 65–92.

Li, K.H., Raghunathan, T.E., and Rubin, D.B. (1991). Large sample significance levels from multiply-imputed data using moment-based statistics and an F reference distribution, *J. Am. Statist. Assoc.* **86**, 1065–1073.

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, The Rand Corporation, Santa Monica, CA.

Lillard, L., Smith, J.P., and Welch, F. (1986). What do we really know about wages? The importance of nonreporting and census imputation, *J. Political Econ.* **94**, 489–506.

Lindley, D.V. (1965). *Introduction to Probability and Statistics from a Bayesian Viewpoint*, *Part 2*, *Inference*. Cambridge, UK: 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. Statist. Assoc.* **94**, 1147–1160.

Little, R.J.A. (1976). Inference about means from incomplete multivariate data, *Biometrika* **63**, 593–604.

Little, R.J.A. (1979). Maximum likelihood inference for multiple regression with missing values: A simulation study, *J. Roy. Statist. Soc. B* **41**, 76–87.

Little, R.J.A. (1982). Models for nonresponse in sample surveys, *J. Am. Statist. Assoc.* **77**, 237–250.

Little, R.J.A. (1985a). A note about models for selectivity bias, *Econometrica* **53**, 1469–1474.

Little, R.J.A. (1985b). Nonresponse adjustments in longitudinal surveys: models for categorical data, *Bulletin Int. Statist. Inst.* **15**(1), 1–15.

Little, R.J.A. (1986). Survey nonresponse adjustments, *Int. Statist. Rev.* **54**, 139–157.

Little, R.J.A. (1988a). Small sample inference about means from bivariate normal data with missing values, *Comput. Statist. Data Analysis* **7**, 161–178.

Little, R.J.A. (1988b). Robust estimation of the mean and covariance matrix from data with missing values, *Appl. Statist.* **37**, 23–38.

Little, R.J.A. (1992). Regression with missing X's: a review, *J. Am. Statist. Assoc.* **87**, 1227–1237.

Little, R.J.A. (1993a). Post-stratification: a modeler's perspective, *J. Am. Statist. Assoc.* **88**, 1001–1012.

Little, R.J.A. (1993b). Pattern-mixture models for multivariate incomplete data, *J. Am. Statist. Assoc.* **88**, 125–134.

Little, R.J.A. (1994). A class of pattern-mixture models for normal missing data, *Biometrika* **81**(3), 471–483.

Little, R.J.A. (1995). Modeling the drop-out mechanism in longitudinal studies, *J. Am. Statist. Assoc.* **90**, 1112–1121.

Little, R.J.A. (1997). Biostatistical analysis with missing data, in *Encyclopedia of Biostatistics* (P. Armitage and T. Colton, eds.), London: Wiley.

Little, R.J.A., and Rubin, D.B. (1983a). Incomplete data, *Encyclopedia of the Statistical Sciences* **4**, 46–53.

Little, R.J.A., and Rubin, D.B. (1983b). On jointly estimating parameters and missing data by maximizing the complete-data likelihood, *Am. Statist.* **37**, 218–220.

Little, R.J.A., and Rubin, D.B. (1987). *Statistical Analysis with Missing Data*, 1st edition. New York: Wiley.

Little, R.J.A., and Schenker, N. (1994). Missing Data, in *Handbook for Statistical Modeling in the Social and Behavioral Sciences* (G. Arminger, C.C. Clogg, and M.E. Sobel, eds.), pp. 39–75. New York: Plenum.

Little, R.J.A., 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.A., and Su, H.L. (1987). Missing-data adjustments for partially-scaled variables, *Proc. Survey Res. Methods Sec.*, *Am. Statist. Assoc. 1987*, 644–649.

Little, R.J.A., and Su, H.L. (1989). Item nonresponse in panel surveys, pp. 400–425, in *Panel Surveys* (D. Kasprzyk, G. Duncan, and M.P. Singh, eds.), New York: Wiley.

Little, R.J.A., and Vartivarian, S. (2002). On weighting the rates in nonresponse weights. To appear in *Stat. Med.* **21**.

Little, R.J.A., and Wang, Y.-X. (1996). Pattern-mixture models for multivariate incomplete data with covariates, *Biometrics* **52**, 98–111.

Little, R.J.A., and Yau, L. (1996). Intent-to-treat analysis in longitudinal studies with drop-outs, *Biometrics* **52**, 1324–1333.

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. Statist. Assoc.* **91**, 1219–1227.

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). Markov-normal analysis of iterative simulations before their convergence: redesign for better performance, to appear in *Statist. Sinica*.

Liu, C.H., Rubin, D.B., and Wu, Y. (1998). Parameter expansion to accelerate EM: the PX-EM algorithm, *Biometrika* **85**, 755–770.

Liu, J.S. (2001). *Monte Carlo Strategies in Scientific Computing*. New York: Springer.

Liu, J.S., and Chen, R. (1998). Sequential Monte Carlo methods for dynamic systems, *J. Am. Statist. Assoc.* **93**, 1032–1044.

Lord, F.M. (1955). Estimation of parameters from incomplete data, *J. Am. Statist. Assoc.* **50**, 870–876.

Louis, T.A. (1982). Finding the observed information when using the EM algorithm, *J. Roy. Statist. Soc. B* **44**, 226–233.

Madow, W.G., and Olkin, I. (eds.) (1983). *Incomplete Data in Sample Surveys*, *Vol. 3: Proceedings of the Symposium*. New York: Academic Press.

Madow, W.G., Nisselson, H., and Olkin, I. (eds.) (1983). *Incomplete Data in Sample Surveys*, *Vol. 1: Report and Case Studies*. New York: Academic Press.

Madow, W.G., Olkin, I., and Rubin, D.B. (eds.) (1983). *Incomplete Data in Sample Surveys*, *Vol. 2: Theory and Bibliographies*. 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, *Sociological Methodology 1980*, San Francisco: Jossey Bass.

Marker, D.A., Judkins, D.R., and Winglee, M. (2002). Large-scale imputation for complex surveys, Chapter 22, in *Survey Nonresponse* (R.M. Groves, D.A. Dillman, J.L. Eltinge, and R.J.A. Little, eds.), 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. Roy. Statist. Soc. B* **42**, 109–142.

McCullagh, P., and Nelder, J. (1989). *Generalized Linear Models*, 2nd Edition. New York: CRC Press.

McCulloch, C.E. (1997). Maximum likelihood algorithms for generalized linear mixed models, *J. Am. Statist. 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.

Meilijson, I. (1989). A fast improvement to the EM algorithm on its own terms, *J. Roy. Statist. 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., Cosler, J., Baghelai, C., and Bobula, J. (1980). Develop physician and physician extender data bases, G-155, Final Report, Applied Management Sciences, Inc., Silver Springs, MD.

Meng, X.-L. (1995). Multiple imputation with uncongenial sources of input (with discussion), *Statist. Sci.* **10**, 538–573.

Meng, X.L. (2002). A congenial overview and investigation of multiple imputation inferences under uncongeniality, Chapter 23, in *Survey Nonresponse* (R. Groves, D. Dillman, J. Eltinge, and R. Little, eds.), New York: Wiley.

Meng, X.-L., and Pedlow, S. (1992). EM: a bibliographic review with missing articles, *Proc. Statist. Computing Sec.*, *Am. Statist. Assoc.*, 24–27.

Meng, X.-L., and Rubin, D.B. (1991), Using EM to obtain asymptotic variance-covariance matrices: the SEM algorithm, *J. Am. Statist. 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 and its Applications* **199**, 413–425.

Meng, X.-L., and Wong, W.H. (1996). Simulating ratios of normalizing constants via a simple identity: a theoretical exploration, *Statist. Sinica* **6**, 831–860.

Meng, X.L., and van Dyk, D.A. (1997). The EM algorithm—an old folk song sung to a fast new tune (with discussion), *J. Roy. Statist. Soc. B* **59**, 511–567.

Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H., and Teller, E. (1953). Equations of state calculations by fast computing machines, *J. Chem. Physics* **21**, 1087–1091.

Miller, R.G. (1974). The jackknife—a review, *Biometrika* **61**, 1–15.

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. Statist. 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, *Statist. Med.* **7**, 941–946.

Neyman, J. (1934). On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection, *J. Roy. Statist. Soc. A* **97**, 558–606.

Nie, N.H., Hull, C.H., Jenkins, J.G., Steinbrenner, K., and Bent, D.H. (1975). *SPSS*, 2nd ed. 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. Statist. Assoc.* **79**, 772–780.

Oh, H.L., and Scheuren, F.S. (1983). Weighting adjustments for unit nonresponse, in *Incomplete Data in Sample Surveys*, *Vol. II: Theory and Annotated Bibliography* (W.G. Madow, I. Olkin, and D.B. Rubin, eds.), New York: Academic Press.

Olkin, I., and Tate, R.F. (1961). Multivariate correlation models with mixed discrete and continuous variables, *Ann. Math. Statist.* **32**, 448–465.

Olsen, R.J. (1980). A least squares correction for selectivity bias, *Econometrica* **48**, 1815–1820.

Orchard, T., and Woodbury, M.A. (1972). A missing information principle: theory and applications, *Proc. 6th Berkeley Symposium on Math. Statist. and Prob.* **1**, 697–715.

Park, T. (1993). A comparison of the generalized estimating equation approach with the maximum likelihood approach for repeated measurements, *Statist. Med.* **12**, 1723–1732.

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. Roy. Statist. Soc. B* **47**, 253–261.

Pinheiro, J.C., and Bates, D.M. (2000). *Mixed-Effects Models in S and S-PLUS*. New York: Springer-Verlag.

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, *Survey Methodol.* **2**, 70–102.

Press, S.J., and Scott, A.J. (1976). Missing variables in Bayesian regression, II, *J. Am. Statist. Assoc.* **71**, 366–369.

Press, S.J., and Wilson, S. (1978). Choosing between logistic regression and discriminant analysis, *J. Am. Statist. Assoc.* **73**, 699–705.

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, *Survey Methodol*. (in press). For associated IVEWARE software see http://www.isr.umich.edu/src/smp/ive/

Raghunathan, T.E., and Rubin, D.B. (1998). Roles for Bayesian techniques in survey sampling, *Proc. Silver Jubilee Meeting Statist. Soc. Canada*, 51–55.

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. Statist. 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.

Robins, J.M., and Rotnitzky, A. (1995). Semiparametric efficiency in multivariate regression models with missing data, *J. Am. Statist. Assoc.* **90**, 122–129.

Robins, J.M., Rotnitzky, A., and Zhao, L.P. (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data, *J. Am. Statist. Assoc.* **90**, 106–121.

Robins, J.M., and Wang, N. (2002). Inference for imputation estimators, *Biometrika* **89**, in press.

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. Statist.* **39**, 33–38.

Rubin, D.B. (1972). A non-iterative algorithm for least squares estimation of missing values in any analysis of variance design, *Appl. Statist.* **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. Statist. Assoc.* **69**, 467–474.

Rubin, D.B. (1976a). Inference and missing data (with discussion), *Biometrika* **63**, 581–592.

Rubin, D.B. (1976b). Non-iterative least squares estimates, standard errors and F-tests for any analysis of variance with missing data, *J. Roy. Statist. 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. Statist. Assoc.* **72**, 538–543.

Rubin, D.B. (1978a). Bayesian inference for causal effects: the role of randomization, *Ann. Statist.* **7**, 34–58.

Rubin, D.B. (1978b). Multiple imputations in sample surveys, *Proc. Survey Res. Meth. Sec.*, *Am. Statist. Assoc. 1978*, 20–34.

Rubin, D.B. (1980). Illustrating the use of multiple imputation to handle nonresponse in sample surveys, *Proc. 42nd Session Int. Stat. Inst.*, *1979*, *Book 2*, 517–532.

Rubin, D.B. (1983a). Iteratively reweighted least squares, *Encyclopedia of the Statistical Sciences* **4**, 272–275.

Rubin, D.B. (1983b). Imputing income in the CPS, in *The Measurement of Labor Cost* (Jack Triplett, ed.). Chicago: University of Chicago Press.

Rubin, D.B. (1985a). The use of propensity scores in applied Bayesian inference, in *Bayesian Statistics 2* (J.M. Bernardo, M.H. De Groot, D.V. Lindley, and A.F.M. Smith, eds.), Amsterdam: North Holland, pp. 463–472.

Rubin, D.B. (1985b). Comment on "A statistical model for positron emission tomography," *J. Am. Statist. Assoc.* **80**, 31–32.

Rubin, D.B. (1986). Statistical matching using file concatenation with adjusted weights and multiple imputations, *J. Bus. Econ. Statist.* **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. Statist. Assoc.* **82**, 543–546.

Rubin, D.B. (1994). Comment on "Missing Data, Imputation, and the Bootstrap" by Bradley Efron, *J. Am. Statist. Assoc.* **89**, 475–478.

Rubin, D.B. (1996). Multiple imputation after 18+ years (with discussion), *J. Am. Statist. Assoc.* **91**, 473–489.

Rubin, D.B. (2000). The utility of counterfactuals for causal inference. Comment on A.P. Dawid, "Causal Inference Without Counterfactuals," *J. Am. Statist. Assoc.* **95**, 435–438.

Rubin, D. B. (2002). Multiple imputation of NMES, to appear in *Statistica Nierlandica*.

Rubin, D.B., and Schenker, N. (1986). Multiple imputation for interval estimation from simple random samples with ignorable nonresponse, *J. Am. Statist. 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. Official Statist.* **3**, 375–387.

Rubin, D.B., Stern, H., and Vehovar, V. (1995). Handling "don't know" survey responses: the case of the Slovenian plebiscite, *J. Am. Statist. Assoc.* **90**, 822–828.

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. Statist.* **20**, 1079–1093.

Rubin, D.B., and Thomas, N. (2000). Combining propensity score matching with additional adjustments for prognostic covariates, *J. Am. Statist. Assoc.* **95**, 573–585.

Sande, I.G. (1983). Hot deck imputation procedures, in *Incomplete Data in Sample Surveys*, *Vol. III: Symposium on Incomplete Data*, *Proceedings* (W.G. Madow and I. Olkin, eds.), 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, SAS Institute, Inc., Cary: NC.

Schafer, J.L. (1997). *Analysis of Incomplete Multivariate Data*. New York: CRC Press.

Schafer, J.L. (1998a). Multiple imputation: a primer, *Statist. Meth. Med. Res.* **8**, 3–15.

Schafer, J.L. (1998b). Some improved procedures for linear mixed models. Technical Report, Department of Statistics, Pennsylvania State University.

Scharfstein, D., Rotnitsky, A., and Robins, J. (1999). Adjusting for nonignorable dropout using semiparametric models, *J. Am. Statist. Assoc.* **94**, 1096–1146 (with discussion).

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, *Proc. Survey Res. Meth. Sec.*, *Am. Statist. Assoc. 1978*, 212–218.

Schluchter, M.D., and Jackson, K.L. (1989). Log-linear analysis of censored survival data with partially observed covariates, *J. Am. Statist. Assoc.* **84**, 42–52.

Shao, J. (2002). Replication methods for variance estimation in complex surveys with imputed data, Chapter 20, in *Survey Nonresponse* (R.M. Groves, D.A. Dillman, J.L. Eltinge, and R.J.A. Little, eds.), New York: Wiley.

Shao, J., Chen, Y., and Chen, Y. (1998). Balanced repeated replications for stratified multistage survey data under imputation, *J. Am. Statist. Assoc.* **93**, 819–831.

Shumway, R.H., and Stoffer, D.S. (1982). An approach to time series smoothing and forecasting using the EM algorithm, *J. Time Series Anal.* **3**, 253–264.

Skinner, C.J., Smith, T.M.F., and Holt, D. (eds.) (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. Statistician* **46**, 84–88.

Snedecor, G.W., and Cochran, W.G. (1967). *Statistical Methods*. Ames: Iowa State University Press.

Stuart, A., and Ord, J.K. (1994). *Kendall's Advanced Theory of Statistics*, 6th ed., Vol 1: Distribution Theory. New York: Arnold.

Sundberg, R. (1974). Maximum likelihood theory for incomplete data from an exponential family, *Scand. J. Statist.* **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. Statist. Math.* **30**, 81–88.

Tang, G., Little, R.J.A., and Raghunathan, T. (2002). A pseudo-likelihood method for multivariate monotone missing data with nonignorable nonresponse. Submitted to *Biometrika*.

Tanner, M.A. (1996). *Tools for Statistical Inference*, 3rd edition. New York: Springer-Verlag.

Tanner, M.A., and Wong, W.H. (1987). The calculation of posterior distributions by data augmentation (with discussion), *J. Am. Statist. Assoc.* **82**, 528–550.

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. Roy. Statist. Soc. B* **14**, 45–100.

Trawinski, I.M., and Bargmann, R.W. (1964). Maximum likelihood with incomplete multivariate data, *Ann. Math. Statist.* **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. Statist. Assoc.* **88**, 26–36.

Vach, W. (1994). *Logistic Regression with Missing Values in the Covariates*. New York: Springer-Verlag.

Van Buuren, S., and Oudshoorn, C.G.M (1999). Flexible multivariate imputation by MICE. Leiden: TNO Preventie en Gezondheid, TNO/VGZ/PG 99.054. For associated software see http://www.multiple-imputation.com.

Van Dyk, D.A., Meng, X.L., and Rubin, D.B. (1995). Maximum likelihood estimation via the ECM algorithm: computing the asymptotic variance, *Statist. Sinica* **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. Statist. Assoc.* **80**, 8–37.

Von Neumann, J. (1951). Various techniques used in connection with random digits, *National Bureau of Standards Applied Mathematics Series* **12**, 36–38.

Wachter, K.W., and Trussell, J. (1982). Estimating historical heights, *J. Am. Statist. Assoc.* **77**, 279–301.

Wang-Clow, F., Lange, M., Laird, N.M., and Ware, J.H. (1995). A simulation study of estimators for rates of change in longitudinal studies with attrition, *Statist. Med.* **14**, 283–297.

Ware, J.H. (1985). Linear models for the analysis of longitudinal studies, *Am. Statist.* **39**, 95–101.

Weisberg, S. (1980). *Applied Linear Regression*. New York: Wiley.

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. Roy. Statist. Soc. A* **147**, 87–99.

Wu, C.F.J. (1983). On the convergence properties of the EM algorithm, *Ann. Statist.* **11**, 95–103.

Wu, C.F.J., and Hamada, M. (2000). *Experiments: Planning*, *Analysis*, *and Parameter Design Optimization*. New York: John Wiley.

Wu, M.C., and Bailey, K.R. (1989). Estimation and comparison of changes in the presence of informative right censoring: conditional linear model, *Biometrics* **45**, 939–955.

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.

Yates, F. (1933). The analysis of replicated experiments when the field results are incomplete, *Emp. J. Exp. Agric.* **1**, 129–142.

Zellner, A. (1962). An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias, *J. Am. Statist. Assoc.* **57**, 348–368.

Zhao, L.P., and Lipsitz, S. (1992). Designs and analysis of two-stage studies, *Statist. Med.* **11**, 769–782.

Author Index

Subject Index

WILEY SERIES IN PROBABILITY AND STATISTICS

ESTABLISHED BY WALTER A. SHEWHART AND SAMUEL S. WILKS

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.

ABRAHAM and LEDOLTER · Statistical Methods for Forecasting
AGRESTI · Analysis of Ordinal Categorical Data
AGRESTI · An Introduction to Categorical Data Analysis
AGRESTI · Categorical Data Analysis, *Second Edition*
ANDĚL · Mathematics of Chance
ANDERSON · An Introduction to Multivariate Statistical Analysis, *Second Edition*
*ANDERSON · The Statistical Analysis of Time Series
ANDERSON, AUQUIER, HAUCK, OAKES, VANDAELE, and WEISBERG · Statistical Methods for Comparative Studies
ANDERSON and LOYNES · The Teaching of Practical Statistics
ARMITAGE and DAVID (editors) · Advances in Biometry
ARNOLD, BALAKRISHNAN, and NAGARAJA · Records
*ARTHANARI and DODGE · Mathematical Programming in Statistics
*BAILEY · The Elements of Stochastic Processes with Applications to the Natural Sciences
BALAKRISHNAN and KOUTRAS · Runs and Scans with Applications
BARNETT · Comparative Statistical Inference, *Third Edition*
BARNETT and LEWIS · Outliers in Statistical Data, *Third Edition*
BARTOSZYNSKI and NIEWIADOMSKA-BUGAJ · Probability and Statistical Inference
BASILEVSKY · Statistical Factor Analysis and Related Methods: Theory and Applications
BASU and RIGDON · Statistical Methods for the Reliability of Repairable Systems
BATES and WATTS · Nonlinear Regression Analysis and Its Applications
BECHHOFER, SANTNER, and GOLDSMAN · Design and Analysis of Experiments for Statistical Selection, Screening, and Multiple Comparisons
BELSLEY · Conditioning Diagnostics: Collinearity and Weak Data in Regression
BELSLEY, KUH, and WELSCH · Regression Diagnostics: Identifying Influential Data and Sources of Collinearity
BENDAT and PIERSOL · Random Data: Analysis and Measurement Procedures, *Third Edition*

*Now available in a lower priced paperback edition in the Wiley Classics Library.

BERRY, CHALONER, and GEWEKE · Bayesian Analysis in Statistics and Econometrics: Essays in Honor of Arnold Zellner
BERNARDO and SMITH · Bayesian Theory
BHAT and MILLER · Elements of Applied Stochastic Processes, *Third Edition*
BHATTACHARYA and JOHNSON · Statistical Concepts and Methods
BHATTACHARYA and WAYMIRE · Stochastic Processes with Applications
BILLINGSLEY · Convergence of Probability Measures, *Second Edition*
BILLINGSLEY · Probability and Measure, *Third Edition*
BIRKES and DODGE · Alternative Methods of Regression
BLISCHKE AND MURTHY · Reliability: Modeling, Prediction, and Optimization
BLOOMFIELD · Fourier Analysis of Time Series: An Introduction, *Second Edition*
BOLLEN · Structural Equations with Latent Variables
BOROVKOV · Ergodicity and Stability of Stochastic Processes
BOULEAU · Numerical Methods for Stochastic Processes
BOX · Bayesian Inference in Statistical Analysis
BOX · R. A. Fisher, the Life of a Scientist
BOX and DRAPER · Empirical Model-Building and Response Surfaces
*BOX and DRAPER · Evolutionary Operation: A Statistical Method for Process Improvement
BOX, HUNTER, and HUNTER · Statistics for Experimenters: An Introduction to Design, Data Analysis, and Model Building
BOX and LUCEÑO · Statistical Control by Monitoring and Feedback Adjustment
BRANDIMARTE · Numerical Methods in Finance: A MATLAB-Based Introduction
BROWN and HOLLANDER · Statistics: A Biomedical Introduction
BRUNNER, DOMHOF, and LANGER · Nonparametric Analysis of Longitudinal Data in Factorial Experiments
BUCKLEW · Large Deviation Techniques in Decision, Simulation, and Estimation
CAIROLI and DALANG · Sequential Stochastic Optimization
CHAN · Time Series: Applications to Finance
CHATTERJEE and HADI · Sensitivity Analysis in Linear Regression
CHATTERJEE and PRICE · Regression Analysis by Example, *Third Edition*
CHERNICK · Bootstrap Methods: A Practitioner's Guide
CHILÈS and DELFINER · Geostatistics: Modeling Spatial Uncertainty
CHOW and LIU · Design and Analysis of Clinical Trials: Concepts and Methodologies
CLARKE and DISNEY · Probability and Random Processes: A First Course with Applications, *Second Edition*
*COCHRAN and COX · Experimental Designs, *Second Edition*
CONGDON · Bayesian Statistical Modelling
CONOVER · Practical Nonparametric Statistics, *Second Edition*
COOK · Regression Graphics
COOK and WEISBERG · Applied Regression Including Computing and Graphics
COOK and WEISBERG · An Introduction to Regression Graphics
CORNELL · Experiments with Mixtures, Designs, Models, and the Analysis of Mixture Data, *Third Edition*
COVER and THOMAS · Elements of Information Theory
COX · A Handbook of Introductory Statistical Methods
*COX · Planning of Experiments
CRESSIE · Statistics for Spatial Data, *Revised Edition*
CSÖRGŐ and HORVÁTH · Limit Theorems in Change Point Analysis
DANIEL · Applications of Statistics to Industrial Experimentation
DANIEL · Biostatistics: A Foundation for Analysis in the Health Sciences, *Sixth Edition*
*DANIEL · Fitting Equations to Data: Computer Analysis of Multifactor Data, *Second Edition*

*Now available in a lower priced paperback edition in the Wiley Classics Library.

DAVID · Order Statistics, *Second Edition*
*DEGROOT, FIENBERG, and KADANE · Statistics and the Law
DEL CASTILLO · Statistical Process Adjustment for Quality Control
DETTE and STUDDEN · The Theory of Canonical Moments with Applications in Statistics, Probability, and Analysis
DEY and MUKERJEE · Fractional Factorial Plans
DILLON and GOLDSTEIN · Multivariate Analysis: Methods and Applications
DODGE · Alternative Methods of Regression
*DODGE and ROMIG · Sampling Inspection Tables, *Second Edition*
*DOOB · Stochastic Processes
DOWDY and WEARDEN · Statistics for Research, *Second Edition*
DRAPER and SMITH · Applied Regression Analysis, *Third Edition*
DRYDEN and MARDIA · Statistical Shape Analysis
DUDEWICZ and MISHRA · Modern Mathematical Statistics
DUNN and CLARK · Applied Statistics: Analysis of Variance and Regression, *Second Edition*
DUNN and CLARK · Basic Statistics: A Primer for the Biomedical Sciences, *Third Edition*
DUPUIS and ELLIS · A Weak Convergence Approach to the Theory of Large Deviations
*ELANDT-JOHNSON and JOHNSON · Survival Models and Data Analysis
ETHIER and KURTZ · Markov Processes: Characterization and Convergence
EVANS, HASTINGS, and PEACOCK · Statistical Distributions, *Third Edition*
FELLER · An Introduction to Probability Theory and Its Applications, Volume I, *Third Edition,* Revised; Volume II, *Second Edition*
FISHER and VAN BELLE · Biostatistics: A Methodology for the Health Sciences
*FLEISS · The Design and Analysis of Clinical Experiments
FLEISS · Statistical Methods for Rates and Proportions, *Second Edition*
FLEMING and HARRINGTON · Counting Processes and Survival Analysis
FULLER · Introduction to Statistical Time Series, *Second Edition*
FULLER · Measurement Error Models
GALLANT · Nonlinear Statistical Models
GHOSH, MUKHOPADHYAY, and SEN · Sequential Estimation
GIFI · Nonlinear Multivariate Analysis
GLASSERMAN and YAO · Monotone Structure in Discrete-Event Systems
GNANADESIKAN · Methods for Statistical Data Analysis of Multivariate Observations, *Second Edition*
GOLDSTEIN and LEWIS · Assessment: Problems, Development, and Statistical Issues
GREENWOOD and NIKULIN · A Guide to Chi-Squared Testing
GROSS and HARRIS · Fundamentals of Queueing Theory, *Third Edition*
*HAHN · Statistical Models in Engineering
HAHN and MEEKER · Statistical Intervals: A Guide for Practitioners
HALD · A History of Probability and Statistics and their Applications Before 1750
HALD · A History of Mathematical Statistics from 1750 to 1930
HAMPEL · Robust Statistics: The Approach Based on Influence Functions
HANNAN and DEISTLER · The Statistical Theory of Linear Systems
HEIBERGER · Computation for the Analysis of Designed Experiments
HEDAYAT and SINHA · Design and Inference in Finite Population Sampling
HELLER · MACSYMA for Statisticians
HINKELMAN and KEMPTHORNE: · Design and Analysis of Experiments, Volume 1: Introduction to Experimental Design
HOAGLIN, MOSTELLER, and TUKEY · Exploratory Approach to Analysis of Variance
HOAGLIN, MOSTELLER, and TUKEY · Exploring Data Tables, Trends and Shapes

*Now available in a lower priced paperback edition in the Wiley Classics Library.

*HOAGLIN, MOSTELLER, and TUKEY · Understanding Robust and Exploratory Data Analysis
HOCHBERG and TAMHANE · Multiple Comparison Procedures
HOCKING · Methods and Applications of Linear Models: Regression and the Analysis of Variables
HOEL · Introduction to Mathematical Statistics, *Fifth Edition*
HOGG and KLUGMAN · Loss Distributions
HOLLANDER and WOLFE · Nonparametric Statistical Methods, *Second Edition*
HOSMER and LEMESHOW · Applied Logistic Regression, *Second Edition*
HOSMER and LEMESHOW · Applied Survival Analysis: Regression Modeling of Time to Event Data
HØYLAND and RAUSAND · System Reliability Theory: Models and Statistical Methods
HUBER · Robust Statistics
HUBERTY · Applied Discriminant Analysis
HUNT and KENNEDY · Financial Derivatives in Theory and Practice
HUSKOVA, BERAN, and DUPAC · Collected Works of Jaroslav Hajek—with Commentary
IMAN and CONOVER · A Modern Approach to Statistics
JACKSON · A User's Guide to Principle Components
JOHN · Statistical Methods in Engineering and Quality Assurance
JOHNSON · Multivariate Statistical Simulation
JOHNSON and BALAKRISHNAN · Advances in the Theory and Practice of Statistics: A Volume in Honor of Samuel Kotz
JUDGE, GRIFFITHS, HILL, LÜTKEPOHL, and LEE · The Theory and Practice of Econometrics, *Second Edition*
JOHNSON and KOTZ · Distributions in Statistics
JOHNSON and KOTZ (editors) · Leading Personalities in Statistical Sciences: From the Seventeenth Century to the Present
JOHNSON, KOTZ, and BALAKRISHNAN · Continuous Univariate Distributions, Volume 1, *Second Edition*
JOHNSON, KOTZ, and BALAKRISHNAN · Continuous Univariate Distributions, Volume 2, *Second Edition*
JOHNSON, KOTZ, and BALAKRISHNAN · Discrete Multivariate Distributions
JOHNSON, KOTZ, and KEMP · Univariate Discrete Distributions, *Second Edition*
JUREČKOVÁ and SEN · Robust Statistical Procedures: Aymptotics and Interrelations
JUREK and MASON · Operator-Limit Distributions in Probability Theory
KADANE · Bayesian Methods and Ethics in a Clinical Trial Design
KADANE AND SCHUM · A Probabilistic Analysis of the Sacco and Vanzetti Evidence
KALBFLEISCH and PRENTICE · The Statistical Analysis of Failure Time Data, *Second Edition*
KASS and VOS · Geometrical Foundations of Asymptotic Inference
KAUFMAN and ROUSSEEUW · Finding Groups in Data: An Introduction to Cluster Analysis
KEDEM and FOKIANOS · Regression Models for Time Series Analysis
KENDALL, BARDEN, CARNE, and LE · Shape and Shape Theory
KHURI · Advanced Calculus with Applications in Statistics
KHURI, MATHEW, and SINHA · Statistical Tests for Mixed Linear Models
KLUGMAN, PANJER, and WILLMOT · Loss Models: From Data to Decisions
KLUGMAN, PANJER, and WILLMOT · Solutions Manual to Accompany Loss Models: From Data to Decisions
KOTZ, BALAKRISHNAN, and JOHNSON · Continuous Multivariate Distributions, Volume 1, *Second Edition*
KOTZ and JOHNSON (editors) · Encyclopedia of Statistical Sciences: Volumes 1 to 9 with Index

*Now available in a lower priced paperback edition in the Wiley Classics Library.

KOTZ and JOHNSON (editors) · Encyclopedia of Statistical Sciences: Supplement Volume
KOTZ, READ, and BANKS (editors) · Encyclopedia of Statistical Sciences: Update Volume 1
KOTZ, READ, and BANKS (editors) · Encyclopedia of Statistical Sciences: Update Volume 2
KOVALENKO, KUZNETZOV, and PEGG · Mathematical Theory of Reliability of Time-Dependent Systems with Practical Applications
LACHIN · Biostatistical Methods: The Assessment of Relative Risks
LAD · Operational Subjective Statistical Methods: A Mathematical, Philosophical, and Historical Introduction
LAMPERTI · Probability: A Survey of the Mathematical Theory, *Second Edition*
LANGE, RYAN, BILLARD, BRILLINGER, CONQUEST, and GREENHOUSE · Case Studies in Biometry
LARSON · Introduction to Probability Theory and Statistical Inference, *Third Edition*
LAWLESS · Statistical Models and Methods for Lifetime Data
LAWSON · Statistical Methods in Spatial Epidemiology
LE · Applied Categorical Data Analysis
LE · Applied Survival Analysis
LEE and WANG · Statistical Methods for Survival Data Analysis, *Third Edition*
LEPAGE and BILLARD · Exploring the Limits of Bootstrap
LEYLAND and GOLDSTEIN (editors) · Multilevel Modelling of Health Statistics
LIAO · Statistical Group Comparison
LINDVALL · Lectures on the Coupling Method
LINHART and ZUCCHINI · Model Selection
LITTLE and RUBIN · Statistical Analysis with Missing Data, *Second Edition*
LLOYD · The Statistical Analysis of Categorical Data
MAGNUS and NEUDECKER · Matrix Differential Calculus with Applications in Statistics and Econometrics, *Revised Edition*
MALLER and ZHOU · Survival Analysis with Long Term Survivors
MALLOWS · Design, Data, and Analysis by Some Friends of Cuthbert Daniel
MANN, SCHAFER, and SINGPURWALLA · Methods for Statistical Analysis of Reliability and Life Data
MANTON, WOODBURY, and TOLLEY · Statistical Applications Using Fuzzy Sets
MARDIA and JUPP · Directional Statistics
MASON, GUNST, and HESS · Statistical Design and Analysis of Experiments with Applications to Engineering and Science
McCULLOCH and SEARLE · Generalized, Linear, and Mixed Models
McFADDEN · Management of Data in Clinical Trials
McLACHLAN · Discriminant Analysis and Statistical Pattern Recognition
McLACHLAN and KRISHNAN · The EM Algorithm and Extensions
McLACHLAN and PEEL · Finite Mixture Models
McNEIL · Epidemiological Research Methods
MEEKER and ESCOBAR · Statistical Methods for Reliability Data
MEERSCHAERT and SCHEFFLER · Limit Distributions for Sums of Independent Random Vectors: Heavy Tails in Theory and Practice
*MILLER · Survival Analysis, *Second Edition*
MONTGOMERY, PECK, and VINING · Introduction to Linear Regression Analysis, *Third Edition*
MORGENTHALER and TUKEY · Configural Polysampling: A Route to Practical Robustness
MUIRHEAD · Aspects of Multivariate Statistical Theory
MURRAY · X-STAT 2.0 Statistical Experimentation, Design Data Analysis, and Nonlinear Optimization

*Now available in a lower priced paperback edition in the Wiley Classics Library.

MYERS and MONTGOMERY · Response Surface Methodology: Process and Product Optimization Using Designed Experiments, *Second Edition*
MYERS, MONTGOMERY, and VINING · Generalized Linear Models. With Applications in Engineering and the Sciences
NELSON · Accelerated Testing, Statistical Models, Test Plans, and Data Analyses
NELSON · Applied Life Data Analysis
NEWMAN · Biostatistical Methods in Epidemiology
OCHI · Applied Probability and Stochastic Processes in Engineering and Physical Sciences
OKABE, BOOTS, SUGIHARA, and CHIU · Spatial Tesselations: Concepts and Applications of Voronoi Diagrams, *Second Edition*
OLIVER and SMITH · Influence Diagrams, Belief Nets and Decision Analysis
PANKRATZ · Forecasting with Dynamic Regression Models
PANKRATZ · Forecasting with Univariate Box-Jenkins Models: Concepts and Cases
*PARZEN · Modern Probability Theory and Its Applications
PEÑA, TIAO, and TSAY · A Course in Time Series Analysis
PIANTADOSI · Clinical Trials: A Methodologic Perspective
PORT · Theoretical Probability for Applications
POURAHMADI · Foundations of Time Series Analysis and Prediction Theory
PRESS · Bayesian Statistics: Principles, Models, and Applications
PRESS and TANUR · The Subjectivity of Scientists and the Bayesian Approach
PUKELSHEIM · Optimal Experimental Design
PURI, VILAPLANA, and WERTZ · New Perspectives in Theoretical and Applied Statistics
PUTERMAN · Markov Decision Processes: Discrete Stochastic Dynamic Programming
*RAO · Linear Statistical Inference and Its Applications, *Second Edition*
RENCHER · Linear Models in Statistics
RENCHER · Methods of Multivariate Analysis, *Second Edition*
RENCHER · Multivariate Statistical Inference with Applications
RIPLEY · Spatial Statistics
RIPLEY · Stochastic Simulation
ROBINSON · Practical Strategies for Experimenting
ROHATGI and SALEH · An Introduction to Probability and Statistics, *Second Edition*
ROLSKI, SCHMIDLI, SCHMIDT, and TEUGELS · Stochastic Processes for Insurance and Finance
ROSENBERGER and LACHIN · Randomization in Clinical Trials: Theory and Practice
ROSS · Introduction to Probability and Statistics for Engineers and Scientists
ROUSSEEUW and LEROY · Robust Regression and Outlier Detection
RUBIN · Multiple Imputation for Nonresponse in Surveys
RUBINSTEIN · Simulation and the Monte Carlo Method
RUBINSTEIN and MELAMED · Modern Simulation and Modeling
RYAN · Modern Regression Methods
RYAN · Statistical Methods for Quality Improvement, *Second Edition*
SALTELLI, CHAN, and SCOTT (editors) · Sensitivity Analysis
*SCHEFFE · The Analysis of Variance
SCHIMEK · Smoothing and Regression: Approaches, Computation, and Application
SCHOTT · Matrix Analysis for Statistics
SCHUSS · Theory and Applications of Stochastic Differential Equations
SCOTT · Multivariate Density Estimation: Theory, Practice, and Visualization
*SEARLE · Linear Models
SEARLE · Linear Models for Unbalanced Data
SEARLE · Matrix Algebra Useful for Statistics
SEARLE, CASELLA, and McCULLOCH · Variance Components
SEARLE and WILLETT · Matrix Algebra for Applied Economics
SEBER · Linear Regression Analysis

*Now available in a lower priced paperback edition in the Wiley Classics Library.

SEBER · Multivariate Observations
SEBER and WILD · Nonlinear Regression
SENNOTT · Stochastic Dynamic Programming and the Control of Queueing Systems
*SERFLING · Approximation Theorems of Mathematical Statistics
SHAFER and VOVK · Probability and Finance: It's Only a Game!
SMALL and McLEISH · Hilbert Space Methods in Probability and Statistical Inference
SRIVASTAVA · Methods of Multivariate Statistics
STAPLETON · Linear Statistical Models
STAUDTE and SHEATHER · Robust Estimation and Testing
STOYAN, KENDALL, and MECKE · Stochastic Geometry and Its Applications, *Second Edition*
STOYAN and STOYAN · Fractals, Random Shapes and Point Fields: Methods of Geometrical Statistics
STYAN · The Collected Papers of T. W. Anderson: 1943–1985
SUTTON, ABRAMS, JONES, SHELDON, and SONG · Methods for Meta-Analysis in Medical Research
TANAKA · Time Series Analysis: Nonstationary and Noninvertible Distribution Theory
THOMPSON · Empirical Model Building
THOMPSON · Sampling, *Second Edition*
THOMPSON · Simulation: A Modeler's Approach
THOMPSON and SEBER · Adaptive Sampling
THOMPSON, WILLIAMS, and FINDLAY · Models for Investors in Real World Markets
TIAO, BISGAARD, HILL, PEÑA, and STIGLER (editors) · Box on Quality and Discovery: with Design, Control, and Robustness
TIERNEY · LISP-STAT: An Object-Oriented Environment for Statistical Computing and Dynamic Graphics
TSAY · Analysis of Financial Time Series
UPTON and FINGLETON · Spatial Data Analysis by Example, Volume II: Categorical and Directional Data
VAN BELLE · Statistical Rules of Thumb
VIDAKOVIC · Statistical Modeling by Wavelets
WEISBERG · Applied Linear Regression, *Second Edition*
WELSH · Aspects of Statistical Inference
WESTFALL and YOUNG · Resampling-Based Multiple Testing: Examples and Methods for *p*-Value Adjustment
WHITTAKER · Graphical Models in Applied Multivariate Statistics
WINKER · Optimization Heuristics in Economics: Applications of Threshold Accepting
WONNACOTT and WONNACOTT · Econometrics, *Second Edition*
WOODING · Planning Pharmaceutical Clinical Trials: Basic Statistical Principles
WOOLSON and CLARKE · Statistical Methods for the Analysis of Biomedical Data, *Second Edition*
WU and HAMADA · Experiments: Planning, Analysis, and Parameter Design Optimization
YANG · The Construction Theory of Denumerable Markov Processes
*ZELLNER · An Introduction to Bayesian Inference in Econometrics
ZHOU, OBUCHOWSKI, and McCLISH · Statistical Methods in Diagnostic Medicine

*Now available in a lower priced paperback edition in the Wiley Classics Library.