

Multiple Imputation for Nonresponse in Surveys

Multiple Imputation for Nonresponse in Surveys

DONALD B. RUBIN

Department of Statistics
Harvard University

JOHN WILEY & SONS

New York • Chichester • Brisbane • Toronto • Singapore

A NOTE TO THE READER:

This book has been electronically reproduced from digital information stored at John Wiley & Sons, Inc. We are pleased that the use of this new technology will enable us to keep works of enduring scholarly value in print as long as there is a reasonable demand for them. The content of this book is identical to previous printings.

Copyright © 1987 by John Wiley & Sons, Inc.

All rights reserved. Published simultaneously in Canada.

Reproduction or translation of any part of this work beyond that permitted by Section 107 or 108 of the 1976 United States Copyright Act without the permission of the copyright owner is unlawful. Requests for permission or further information should be addressed to the Permissions Department, John Wiley & Sons, Inc.

Library of Congress Cataloging in Publication Data:

Rubin, Donald B.

Multiple imputation for nonresponse in surveys.

(Wiley series in probability and mathematical statistics. Applied probability and statistics, ISSN 0271-6232)

Bibliography: p.

Includes index.

1. Multiple imputation (Statistics) I. Title.

II. Series.

HH31.2.R83 1987 001.4'225 86-28935

ISBN 0-471-08705-X

Printed in the United States of America

10 9 8

To the family of my childhood
and the family of my parenthood

Preface

Multiple imputation is a statistical technique designed to take advantage of the flexibility in modern computing to handle missing data. With it, each missing value is replaced by two or more imputed values in order to represent the uncertainty about which value to impute. The ideas for multiple imputation first arose in the early 1970s when I was working on a problem of survey nonresponse at Educational Testing Service, here summarized as Example 1.1. This work was published several years later as Rubin (1977a).

The real impetus for multiple imputation, however, came from work encouraged and supported by Fritz Scheuren, then of the United States Social Security Administration and now head of the Statistics of Income Division at the United States Internal Revenue Service. His concern for problems of nonresponse in the Current Population Survey led to a working paper for the Social Security Administration (Rubin, 1977b), which explicitly proposed multiple imputation. Fritz's continued support and encouragement for the idea of multiple imputation resulted in (1) an American Statistical Association invited address on multiple imputation (Rubin, 1978a); (2) continued research, such as published in Rubin (1979a); (3) joint work with Fritz and Thomas N. Herzog in the late 1970s, summarized in several papers including Herzog and Rubin (1983); and (4) application of the ideas in 1980 to file matching, which eventually was published as Rubin (1986).

Another important contributor to the development of multiple imputation has been the United States Census Bureau, which several years ago supported the production of a monograph on multiple imputation (Rubin, 1980a). This monograph was the first of four nearly complete drafts that were supposed to become this book.

The second such draft was composed of the collection of chapters distributed to my class on survey nonresponse at the University of Chicago, Winter Quarter 1983. These stopped short of becoming the book primarily because of two Ph.D. students there, Kim Hung Li and Nathaniel Schenker, both of whom wrote theses on aspects of multiple imputation (Li, 1985; Schenker, 1985). Our efforts provided the foundation for the next level of sophistication, and I am extremely grateful for their involvement and for the outstandingly collegial atmosphere at the University of Chicago, which made this period so productive.

The third draft owed its demise to continued work involving Schenker and two Ph.D. students at Harvard University, T. E. Raghunathan and Leisa Weld, both of whom are completing theses on aspects of multiple imputation. This fourth and final version has benefitted from many suggestions from Raghunathan, Weld, Roderick J. A. Little and Alan Zaslavsky, and was facilitated by Raghunathan's computing help, and Bea Shube's and Rosalyn Farkas's editorial advice and patience. It too could have been postponed, waiting for improved results to come from ongoing research, but I believe the existing perspective is highly useful and that publication will stimulate new work. In fact, although many of the problems at the end of the chapters are rather standard exercises designed to check understanding of the material being presented, other problems involve issues that I consider research topics for term papers in a graduate-level course on survey methods or even points of departure for Ph.D. theses.

Since the summer of 1983, support for my work on multiple imputation and my graduate students' work at the University of Chicago and Harvard University has been primarily provided by a grant from NSF (SES-83-11428), and I am very grateful for this funding as well as additional support in 1986 from NSF (DMS-85-04332). The SES grant deals explicitly with the problem of the comparability of Census Bureau occupation and industry codes between 1970 and 1980, summarized here as Example 1.3. The creation of 1970 public-use files with multiply-imputed 1980 codes will be, I believe, an important milestone in the handling of missing values in public-use files.

This text is directed at applied survey statisticians with some theoretical background, but presents the necessary Bayesian and frequentist theory in the background Chapter 2. Chapter 3 derives, from the Bayesian perspective, general procedures for analyzing multiply-imputed data sets, and Chapter 4 evaluates the operating characteristics of these procedures from the randomization theory perspective. Particular procedures for creating multiple imputations are presented in Chapter 5 for cases with ignorable nonresponse and in Chapter 6 for cases with nonignorable nonresponse. Chapter 1 and the detailed table of contents are designed to allow the

reader to obtain a rapid overview of the theory and practice of multiple imputation.

Multiple Imputation for Nonresponse in Surveys can serve as the basis for a course on survey methodology at the graduate level in a department of statistics, as I have done with earlier drafts at the University of Chicago and Harvard University. When utilized this way, I believe it should be supplemented with a more standard text, such as Cochran (1977), and readings from the National Academy of Sciences volumes on Incomplete Data (Madow et al., 1983).

I hope that the reader finds the material presented here to be a stimulating and useful contribution to the theory and practice of handling nonresponse in surveys.

DONALD B. RUBIN

Cambridge, Massachusetts
January 1987

Contents

| | |
|--|--------------|
| TABLES AND FIGURES | xxiv |
| GLOSSARY | xxvii |
| 1. INTRODUCTION | 1 |
| 1.1. Overview | 1 |
| Nonresponse in Surveys | 1 |
| Multiple Imputation | 2 |
| Can Multiple Imputation Be Used in Nonsurvey Problems? | 3 |
| Background | 4 |
| 1.2. Examples of Surveys with Nonresponse | 4 |
| Example 1.1. Educational Testing Service's Sample Survey of Schools | 4 |
| Example 1.2. Current Population Survey and Missing Incomes | 5 |
| Example 1.3. Census Public-Use Data Bases and Missing Occupation Codes | 6 |
| Example 1.4. Normative Aging Study of Drinking | 7 |
| 1.3. Properly Handling Nonresponse | 7 |
| Handling Nonresponse in Example 1.1 | 7 |
| Handling Nonresponse in Example 1.2 | 9 |
| Handling Nonresponse in Example 1.3 | 10 |
| | xi |

| | |
|--|-----------|
| Handling Nonresponse in Example 1.4 | 10 |
| The Variety of Objectives When Handling Nonresponse | 11 |
| 1.4. Single Imputation | 11 |
| Imputation Allows Standard Complete-Data Methods of Analysis to Be Used | 11 |
| Imputation Can Incorporate Data Collector's Knowledge | 12 |
| The Problem with One Imputation for Each Missing Value | 12 |
| Example 1.5. Best-Prediction Imputation in a Simple Random Sample | 13 |
| Example 1.6. Drawing Imputations from a Distribution (Example 1.5 continued) | 14 |
| 1.5. Multiple Imputation | 15 |
| Advantages of Multiple Imputation | 15 |
| The General Need to Display Sensitivity to Models of Nonresponse | 16 |
| Disadvantages of Multiple Imputation | 17 |
| 1.6. Numerical Example Using Multiple Imputation | 19 |
| Analyzing This Multiply-Imputed Data Set | 19 |
| Creating This Multiply-Imputed Data Set | 22 |
| 1.7. Guidance for the Reader | 22 |
| Problems | 23 |
| 2. STATISTICAL BACKGROUND | 27 |
| 2.1. Introduction | 27 |
| Random Indexing of Units | 27 |
| 2.2. Variables in the Finite Population | 28 |
| Covariates X | 28 |
| Outcome Variables Y | 29 |
| Indicator for Inclusion in the Survey I | 29 |
| Indicator for Response in the Survey R | 30 |
| Stable Response | 30 |
| Surveys with Stages of Sampling | 31 |

| | |
|---|-----------|
| CONTENTS | xiii |
| 2.3. Probability Distributions and Related Calculations | 31 |
| Conditional Probability Distributions | 32 |
| Probability Specifications Are Symmetric in Unit Indices | 32 |
| Bayes's Theorem | 33 |
| Finding Means and Variances from Conditional Means and Variances | 33 |
| 2.4. Probability Specifications for Indicator Variables | 35 |
| Sampling Mechanisms | 35 |
| Examples of Unconfounded Probability Sampling Mechanisms | 37 |
| Examples of Confounded and Nonprobability Sampling Mechanisms | 38 |
| Response Mechanisms | 38 |
| 2.5. Probability Specifications for (X, Y) | 39 |
| de Finetti's Theorem | 40 |
| Some Intuition | 40 |
| Example 2.1. A Simple Normal Model for Y_i | 40 |
| Lemma 2.1. Distributions Relevant to Example 2.1 | 41 |
| Example 2.2. A Generalization of Example 2.1 | 42 |
| Example 2.3. An Application of Example 2.2: The Bayesian Bootstrap | 44 |
| Example 2.4. Y_i Approximately Proportional to X_i | 46 |
| 2.6. Bayesian Inference for a Population Quantity | 48 |
| Notation | 48 |
| The Posterior Distribution for $Q(X, Y)$ | 48 |
| Relating the Posterior Distribution of Q to the Posterior Distribution of Y_{nob} | 49 |
| Ignorable Sampling Mechanisms | 50 |
| Result 2.1. An Equivalent Definition for Ignorable Sampling Mechanisms | 50 |
| Ignorable Response Mechanisms | 51 |
| Result 2.2. Ignorability of the Response Mechanism When the Sampling Mechanism Is Ignorable | 51 |
| Result 2.3. The Practical Importance of Ignorable Mechanisms | 52 |

| | |
|--|-----------|
| Relating Ignorable Sampling and Response Mechanisms to Standard Terminology in the Literature on Parametric Inference from Incomplete Data | 53 |
| 2.7. Interval Estimation | 54 |
| General Interval Estimates | 55 |
| Bayesian Posterior Coverage | 55 |
| Example 2.5. Interval Estimation in the Context of Example 2.1 | 56 |
| Fixed-Response Randomization-Based Coverage | 56 |
| Random-Response Randomization-Based Coverage | 58 |
| Nominal versus Actual Coverage of Intervals | 58 |
| 2.8. Bayesian Procedures for Constructing Interval Estimates, Including Significance Levels and Point Estimates | 59 |
| Highest Posterior Density Regions | 59 |
| Significance Levels— p -Values | 60 |
| Point Estimates | 62 |
| 2.9. Evaluating the Performance of Procedures | 62 |
| A Protocol for Evaluating Procedures | 63 |
| Result 2.4. The Average Coverages Are All Equal to the Probability That C Includes Q | 64 |
| Further Comments on Calibration | 64 |
| 2.10. Similarity of Bayesian and Randomization-Based Inferences in Many Practical Cases | 65 |
| Standard Asymptotic Results Concerning Bayesian Procedures | 66 |
| Extensions of These Standard Results | 66 |
| Practical Conclusions of Asymptotic Results | 67 |
| Relevance to the Multiple-Imputation Approach to Nonresponse | 67 |
| Problems | 68 |
| 3. UNDERLYING BAYESIAN THEORY | 75 |
| 3.1. Introduction and Summary of Repeated-Imputation Inferences | 75 |
| Notation | 75 |

| | |
|--|-----------|
| Combining the Repeated Complete-Data Estimates and Variances | 76 |
| Scalar Q | 77 |
| Significance Levels Based on the Combined Estimates and Variances | 77 |
| Significance Levels Based on Repeated Complete-Data Significance Levels | 78 |
| Example 3.1. Inference for Regression Coefficients | 79 |
| 3.2. Key Results for Analysis When the Multiple Imputations Are Repeated Draws from the Posterior Distribution of the Missing Values | 81 |
| Result 3.1. Averaging the Completed-Data Posterior Distribution of Q over the Posterior Distribution of Y_{mis} to Obtain the Actual Posterior Distribution of Q | 82 |
| Example 3.2. The Normal Model Continued | 82 |
| The Posterior Cumulative Distribution Function of Q | 83 |
| Result 3.2. Posterior Mean and Variance of Q | 84 |
| Simulating the Posterior Mean and Variance of Q | 85 |
| Missing and Observed Information with Infinite m | 85 |
| Inference for Q from Repeated Completed-Data Means and Variances | 86 |
| Example 3.3. Example 3.2 Continued | 87 |
| 3.3. Inference for Scalar Estimands from a Modest Number of Repeated Completed-Data Means and Variances | 87 |
| The Plan of Attack | 88 |
| The Sampling Distribution of S_m Given (X, Y_{obs}, R_{inc}) | 88 |
| The Conditional Distribution of $(\bar{Q}_\infty, \bar{U}_\infty)$ Given S_m and B_∞ | 89 |
| The Conditional Distribution of Q Given S_m and B_∞ | 89 |
| The Conditional Distribution of B_m Given S_m | 90 |
| The Conditional Distribution of $\bar{U}_m + (1 + m^{-1})B_\infty$ Given S_m | 90 |
| Approximation 3.1 Relevant to the Behrens-Fisher Distribution | 91 |
| Applying Approximation 3.1 to Obtain (3.3.9) | 92 |
| The Approximating t Reference Distribution for Scalar Q | 92 |

| | |
|--|------------|
| Example 3.4. Example 3.3 Continued | 92 |
| Fraction of Information Missing Due to Nonresponse | 93 |
| 3.4. Significance Levels for Multicomponent Estimands from a Modest Number of Repeated Completed-Data Means and Variance–Covariance Matrices | 94 |
| The Conditional Distribution of Q Given S_m and B_∞ | 94 |
| The Bayesian p -Value for a Null Value Q_0 Given S_m : General Expression | 95 |
| The Bayesian p -Value Given S_m with Scalar Q | 95 |
| The Bayesian p -Value Given S_m with Scalar Q —Closed-Form Approximation | 96 |
| <p>p-Values with B_∞ <i>a Priori</i> Proportional to T_∞</p> | 96 |
| <p>p-Values with B_∞ <i>a Priori</i> Proportional to T_∞—Closed-Form Approximation</p> | 97 |
| <p>p-Values When B_∞ Is Not <i>a Priori</i> Proportional to \bar{U}_∞</p> | 98 |
| 3.5. Significance Levels from Repeated Completed-Data Significance Levels | 99 |
| A New Test Statistic | 99 |
| The Asymptotic Equivalence of \tilde{D}_m and \hat{D}_m —Proof | 100 |
| Integrating over r_m to Obtain a Significance Level from Repeated Completed-Data Significance Levels | 100 |
| 3.6. Relating the Completed-Data and Complete-Data Posterior Distributions When the Sampling Mechanism Is Ignorable | 102 |
| Result 3.3. The Completed-Data and Complete-Data Posterior Distributions Are Equal When Sampling and Response Mechanisms Are Ignorable | 103 |
| Using i.i.d. Modeling | 104 |
| Result 3.4. The Equality of Completed-Data and Complete-Data Posterior Distributions When Using i.i.d. Models | 104 |
| Example 3.5. A Situation in Which Conditional on θ_{XY} , the Completed-Data and Complete-Data Posterior Distributions of Q Are Equal—Condition (3.6.7) | 105 |
| Example 3.6. Cases in Which Condition (3.6.7) Nearly Holds | 105 |

| | |
|---|----------------|
| Example 3.7. Situations in Which the Completed-Data and Complete-Data Posterior Distributions of θ_{XY} Are Equal—Condition (3.6.8) | 106 |
| Example 3.8. A Simple Case Illustrating the Large-Sample Equivalence of Completed-Data and Complete-Data Posterior Distributions of θ_{XY} | 106 |
| The General Use of Complete-Data Statistics | 106 |
| Problems | 107 |
| 4. RANDOMIZATION-BASED EVALUATIONS | 113 |
| 4.1. Introduction | 113 |
| Major Conclusions | 113 |
| Large-Sample Relative Efficiency of Point Estimates | 114 |
| Large-Sample Coverage of t -Based Interval Estimates | 114 |
| Outline of Chapter | 115 |
| 4.2. General Conditions for the Randomization-Validity of Infinite-m Repeated-Imputation Inferences | 116 |
| Complications in Practice | 117 |
| More General Conditions for Randomization-Validity | 117 |
| Definition: Proper Multiple-Imputation Methods | 118 |
| Result 4.1. If the Complete-Data Inference Is Randomization-Valid and the Multiple-Imputation Procedure Is Proper, Then the Infinite- m Repeated-Imputation Inference Is Randomization-Valid under the Posited Response Mechanism | 119 |
| 4.3. Examples of Proper and Improper Imputation Methods in a Simple Case with Ignorable Nonresponse | 120 |
| Example 4.1. Simple Random Multiple Imputation | 120 |
| Why Variability Is Underestimated Using the Multiple-Imputation Hot-Deck | 122 |
| Example 4.2. Fully Normal Bayesian Repeated Imputation | 123 |
| Example 4.3. A Nonnormal Bayesian Imputation Procedure That Is Proper for the Standard Inference—The Bayesian Bootstrap | 123 |
| Example 4.4. An Approximately Bayesian yet Proper Imputation Method—The Approximate Bayesian Bootstrap | 124 |

| | |
|--|------------|
| Example 4.5. The Mean and Variance Adjusted Hot-Deck | 124 |
| 4.4. Further Discussion of Proper Imputation Methods | 125 |
| Conclusion 4.1. Approximate Repetitions from a Bayesian Model Tend to Be Proper | 125 |
| The Heuristic Argument | 126 |
| Messages of Conclusion 4.1 | 126 |
| The Importance of Drawing Repeated Imputations Appropriate for the Posited Response Mechanism | 127 |
| The Role of the Complete-Data Statistics in Determining Whether a Repeated Imputation Method Is Proper | 127 |
| 4.5. The Asymptotic Distribution of $(\bar{Q}_m, \bar{U}_m, B_m)$ for Proper Imputation Methods | 128 |
| Validity of the Asymptotic Sampling Distribution of S_m | 128 |
| The Distribution of $(\bar{Q}_m, \bar{U}_m, B_m)$ Given (X, Y) for Scalar Q | 129 |
| Random-Response Randomization-Based Justification for the t Reference Distribution | 130 |
| Extension of Results to Multicomponent Q | 131 |
| Asymptotic Efficiency of \bar{Q}_m Relative to \bar{Q}_∞ | 131 |
| 4.6. Evaluations of Finite-m Inferences with Scalar Estimands | 132 |
| Small-Sample Efficiencies of Asymptotically Proper Imputation Methods from Examples 4.2–4.5 | 132 |
| Large-Sample Coverages of Interval Estimates Using a t Reference Distribution and Proper Imputation Methods | 134 |
| Small-Sample Monte Carlo Coverages of Asymptotically Proper Imputation Methods from Examples 4.2–4.5 | 135 |
| Evaluation of Significance Levels | 135 |
| 4.7. Evaluation of Significance Levels from the Moment-Based Statistics D_m and \tilde{D}_m with Multicomponent Estimands | 137 |
| The Level of a Significance Testing Procedure | 138 |
| The Level of D_m —Analysis for Proper Imputation Methods and Large Samples | 138 |
| The Level of D_m —Numerical Results | 139 |

| | |
|---|------------|
| CONTENTS | xix |
| The Level of \tilde{D}_m —Analysis | 139 |
| The Effect of Unequal Fractions of Missing Information on \tilde{D}_m | 141 |
| Some Numerical Results for \tilde{D}_m with $k' = (k + 1)\nu/2$ | 141 |
| 4.8. Evaluation of Significance Levels Based on Repeated Significance Levels | 144 |
| The Statistic $\hat{\hat{D}}_m$ | 144 |
| The Asymptotic Sampling Distribution of \bar{d}_m and s_d^2 | 144 |
| Some Numerical Results for $\hat{\hat{D}}_m$ | 145 |
| The Superiority of Multiple Imputation Significance Levels | 145 |
| Problems | 148 |
| 5. PROCEDURES WITH IGNORABLE NONRESPONSE | 154 |
| 5.1. Introduction | 154 |
| No Direct Evidence to Contradict Ignorable Nonresponse | 155 |
| Adjust for All Observed Differences and Assume Unobserved Residual Differences Are Random | 155 |
| Univariate Y_i and Many Respondents at Each Distinct Value of X_i That Occurs Among Nonrespondents | 156 |
| The More Common Situation, Even with Univariate Y_i | 156 |
| A Popular Implicit Model—The Census Bureau's Hot-Deck | 157 |
| Metric-Matching Hot-Deck Methods | 158 |
| Least-Squares Regression | 159 |
| Outline of Chapter | 159 |
| 5.2. Creating Imputed Values under an Explicit Model | 160 |
| The Modeling Task | 160 |
| The Imputation Task | 161 |
| Result 5.1. The Imputation Task with Ignorable Nonresponse | 162 |
| The Estimation Task | 163 |
| Result 5.2. The Estimation Task with Ignorable Nonresponse When $\theta_{Y X}$ and θ_X Are <i>a Priori</i> Independent | 164 |

| | |
|---|------------|
| Result 5.3. The Estimation Task with Ignorable Nonresponse, $\theta_{Y X}$ and θ_X <i>a Priori</i> Independent, and Univariate Y_i | 165 |
| A Simplified Notation | 165 |
| 5.3. Some Explicit Imputation Models with Univariate Y_i and Covariates | 166 |
| Example 5.1. Normal Linear Regression Model with Univariate Y_i | 166 |
| Example 5.2. Adding a Hot-Deck Component to the Normal Linear Regression Imputation Model | 168 |
| Extending the Normal Linear Regression Model | 168 |
| Example 5.3. A Logistic Regression Imputation Model for Dichotomous Y_i | 169 |
| 5.4. Monotone Patterns of Missingness in Multivariate Y_i | 170 |
| Monotone Missingness in Y —Definition | 171 |
| The General Monotone Pattern—Description of General Techniques | 171 |
| Example 5.4. Bivariate Y_i and an Implicit Imputation Model | 172 |
| Example 5.5. Bivariate Y_i with an Explicit Normal Linear Regression Model | 173 |
| Monotone-Distinct Structure | 174 |
| Result 5.4. The Estimation Task with a Monotone-Distinct Structure | 175 |
| Result 5.5. The Imputation Task with a Monotone-Distinct Structure | 177 |
| 5.5. Missing Social Security Benefits in the Current Population Survey | 178 |
| The CPS–IRS–SSA Exact Match File | 178 |
| The Reduced Data Base | 179 |
| The Modeling Task | 179 |
| The Estimation Task | 180 |
| The Imputation Task | 181 |
| Results Concerning Absolute Accuracies of Prediction | 181 |
| Inferences for the Average OASDI Benefits for the Nonrespondents in the Sample | 184 |
| Results on Inferences for Population Quantities | 185 |

| | |
|---|------------|
| CONTENTS | xxi |
| 5.6. Beyond Monotone Missingness | 186 |
| Two Outcomes Never Jointly Observed—Statistical Matching of Files | 186 |
| Example 5.6. Two Normal Outcomes Never Jointly Observed | 187 |
| Problems Arising with Nonmonotone Patterns | 188 |
| Discarding Data to Obtain a Monotone Pattern | 189 |
| Assuming Conditional Independence Among Blocks of Variables to Create Independent Monotone Patterns | 190 |
| Using Computationally Convenient Explicit Models | 191 |
| Iteratively Using Methods for Monotone Patterns | 192 |
| The Sampling/Importance Resampling Algorithm | 192 |
| Some Details of SIR | 193 |
| Example 5.7. An Illustrative Application of SIR Problems | 194 |
| | 195 |
| 6. PROCEDURES WITH NONIGNORABLE NONRESPONSE | 202 |
| 6.1. Introduction | 202 |
| Displaying Sensitivity to Models for Nonresponse | 202 |
| The Need to Use Easily Communicated Models | 203 |
| Transformations to Create Nonignorable Imputed Values from Ignorable Imputed Values | 203 |
| Other Simple Methods for Creating Nonignorable Imputed Values Using Ignorable Imputation Models | 203 |
| Essential Statistical Issues and Outline of Chapter | 204 |
| 6.2. Nonignorable Nonresponse with Univariate Y_i and No X_i | 205 |
| The Modeling Task | 205 |
| The Imputation Task | 206 |
| The Estimation Task | 206 |
| Two Basic Approaches to the Modeling Task | 207 |
| Example 6.1. The Simple Normal Mixture Model | 207 |
| Example 6.2. The Simple Normal Selection Model | 209 |
| 6.3. Formal Tasks with Nonignorable Nonresponse | 210 |
| The Modeling Task—Notation | 210 |

| | |
|---|------------|
| Two General Approaches to the Modeling Task | 211 |
| Similarities with Ignorable Case | 211 |
| The Imputation Task | 212 |
| Result 6.1. The Imputation Task with Nonignorable Nonresponse | 212 |
| Result 6.2. The Imputation Task with Nonignorable Nonresponse When Each Unit Is Either Included in or Excluded from the Survey | 212 |
| The Estimation Task | 213 |
| Result 6.3. The Estimation Task with Nonignorable Nonresponse When $\theta_{YR X}$ Is <i>a Priori</i> Independent of θ_X | 213 |
| Result 6.4. The Estimation Task with Nonignorable Nonresponse When $\theta_{Y XR}$ Is <i>a Priori</i> Independent of $(\theta_{R X}, \theta_X)$ and Each Unit Is Either Included in or Excluded from the Survey | 213 |
| Result 6.5. The Imputation and Estimation Tasks with Nonignorable Nonresponse and Univariate Y_i | 214 |
| Monotone Missingness | 214 |
| Result 6.6. The Estimation and Imputation Tasks with a Monotone-Distinct Structure and a Mixture Model for Nonignorable Nonresponse | 214 |
| Selection Modeling and Monotone Missingness | 215 |
| 6.4. Illustrating Mixture Modeling Using Educational Testing Service Data | 215 |
| The Data Base | 216 |
| The Modeling Task | 216 |
| Clarification of Prior Distribution Relating Nonrespondent and Respondent Parameters | 217 |
| Comments on Assumptions | 218 |
| The Estimation Task | 219 |
| The Imputation Task | 219 |
| Analysis of Multiply-Imputed Data | 221 |
| 6.5. Illustrating Selection Modeling Using CPS Data | 222 |
| The Data Base | 223 |
| The Modeling Task | 224 |
| The Estimation Task | 225 |
| The Imputation Task | 225 |

| | |
|---|------------|
| CONTENTS | xxiii |
| Accuracy of Results for Single Imputation Methods | 226 |
| Estimates and Standard Errors for Average $\log(\text{wage})$ for Nonrespondents in the Sample | 227 |
| Inferences for Population Mean $\log(\text{wage})$ | 229 |
| 6.6. Extensions to Surveys with Follow-Ups | 229 |
| Ignorable Nonresponse | 231 |
| Nonignorable Nonresponse with 100% Follow-Up Response | 231 |
| Example 6.3. 100% Follow-Up Response in a Simple Random Sample of Y_i | 232 |
| Ignorable Hard-Core Nonresponse Among Follow-Ups | 233 |
| Nonignorable Hard-Core Nonresponse Among Follow- Ups | 233 |
| Waves of Follow-Ups | 234 |
| 6.7. Follow-Up Response in a Survey of Drinking Behavior Among Men of Retirement Age | 234 |
| The Data Base | 235 |
| The Modeling Task | 235 |
| The Estimation Task | 235 |
| The Imputation Task | 235 |
| Inference for the Effect of Retirement Status on Drinking Behavior | 239 |
| Problems | 240 |
| REFERENCES | 244 |
| AUTHOR INDEX | 251 |
| SUBJECT INDEX | 253 |

Tables and Figures

| | | |
|--------------------|--|-----|
| Figure 1.1. | Data set with m imputations for each missing datum. | 3 |
| Table 1.1. | Artificial example of survey data and multiple imputation. | 20 |
| Table 1.2. | Analysis of multiply-imputed data set of Table 1.1. | 21 |
| Figure 2.1. | Matrix of variables in a finite population of N units. | 29 |
| Figure 2.2. | Contours of the posterior distribution of Q with the null value Q_0 indicated. The significance level of Q_0 is the posterior probability that Q is in the shaded area and beyond. | 62 |
| Table 4.1. | Large-sample relative efficiency (in %) when using a finite number of proper imputations, m , rather than an infinite number, as a function of the fraction of missing information, γ_0 : $RE = (1 + \gamma_0/m)^{-1/2}$. | 114 |
| Table 4.2. | Large-sample coverage probability (in %) of interval estimates based on the t reference distribution, (3.1.8), as a function of the number of proper imputations, $m \geq 2$; the fraction of missing information, γ_0 ; and the nominal level, $1 - \alpha$. Also included for contrast are results based on single imputation, $m = 1$, using the complete-data normal reference distribution (3.1.1) with \hat{Q} replaced by $\bar{Q}_1 = \hat{Q}_{*1}$ and U replaced by $\bar{U}_1 = U_{*1}$. | 115 |
| Table 4.3. | Simulated coverages (in %) of asymptotically proper multiple ($m = 2$) imputation procedures with nominal levels 90% and 95%, using t -based inferences, response rates n_1/n , and normal and nonnormal data (Laplace, lognormal = $\exp N(0, 1)$); maximum standard error $< 1\%$. | 136 |

| | | |
|--------------------|---|-----|
| Table 4.4. | Large-sample level (in %) of D_m with $F_{k,\nu}$ reference distribution as a function of nominal level, α ; number of components being tested, k ; number of proper imputations, m ; and fraction of missing information, γ_0 . Accuracy of results = 5000 simulations of (4.7.8) with ρ_0 set to I. | 140 |
| Table 4.5. | Large-sample level (in %) of \tilde{D}_m with $F_{k,(k+1)\nu/2}$ reference distribution as a function of number of components being tested, k ; number of proper imputations, m ; fraction of missing information, γ_0 ; and variance of fractions of missing information, 0 (zero), S (small), L (large). Accuracy of results = 5000 simulations of (4.7.9). | 142 |
| Table 4.6. | Large-sample level (in %) of \hat{D}_m with $F_{k,(1+k^{-1})\hat{\nu}/2}$ reference distribution as a function of number of components being tested, k ; number of proper imputations, m ; fraction of missing information, γ_0 ; and variance of fractions of missing information, 0 (zero), S (small), L (large). Accuracy of results = 5000 simulations of (4.7.7). | 146 |
| Table 4.7. | Large-sample level (in %) of d_{*1} with χ_k^2 reference distribution as a function of nominal level α ; number of components being tested, k ; and fraction of missing information, γ_0 . | 147 |
| Figure 5.1. | A monotone pattern of missingness, 1 = observed, 0 = missing. | 171 |
| Figure 5.2. | Artificial example illustrating hot-deck multiple imputation with a monotone pattern of missing data; parentheses enclose $m = 2$ imputations. | 172 |
| Table 5.1. | Multiple imputations of OASDI benefits for nonrespondents 62–71 years of age. | 182 |
| Table 5.2. | Multiple imputations of OASDI benefits for nonrespondents over 72 years of age. | 183 |
| Table 5.3. | Accuracies of imputation methods with respect to mean absolute deviation (MAD) and root mean squared deviation (RMS). | 183 |
| Table 5.4. | Comparison of estimates (standard errors) for mean OASDI benefits implied by imputation methods for nonrespondent groups in the sample. | 184 |

| | | |
|--------------------|---|-----|
| Table 5.5. | Comparison of estimates (standard errors) for mean OASDI benefits implied by imputation methods for groups in the population. | 185 |
| Table 5.6. | Example from Marini, Olsen and Rubin (1980) illustrating how to obtain a monotone pattern of missing data by discarding data; 1 = observed, 0 = missing. | 190 |
| Table 6.1. | Summary of repeated-imputation intervals for variable 17B in educational example. | 221 |
| Table 6.2. | Background variables X for GRZ example on imputation of missing incomes. | 223 |
| Table 6.3. | Root-mean-squared error of imputations of log-wage: Impute posterior mean given θ fixed at MLE, $\hat{\theta}$. | 226 |
| Table 6.4. | Repeated-imputation estimates (standard errors) for average log(wage) for nonrespondents in the sample under five imputation procedures. | 228 |
| Figure 6.1. | Schematic data structure with follow-up surveys of nonrespondents: boldface produces Y data. | 230 |
| Table 6.5. | Mean alcohol consumption level and retirement status for respondents and nonrespondents within birth cohort: Data from 1982 Normative Aging Study drinking questionnaire. | 236 |
| Table 6.6. | Summary of least-squares estimates of the regression of $\log(1 + \text{drinks/day})$ on retirement status (0 = working, 1 = retired), birth year, and retirement status \times birth year interaction. | 237 |
| Table 6.7. | Five values of regression parameters for nonrespondents drawn from their posterior distribution. | 237 |
| Table 6.8. | Five imputed values of $\log(1 + \text{drinks/day})$ for each of the 74 non-followed-up nonrespondents. | 238 |
| Table 6.9. | Sets of least-squares estimates from the five data sets completed by imputation. | 239 |
| Table 6.10. | Repeated-imputation estimates, standard errors, and percentages of missing information for the regression of $\log(1 + \text{drinks/day})$ on retirement status, birth year, and retirement status \times birth year interaction. | 239 |

Glossary

Basic Random Variables

| | |
|--|-----|
| $X = N \times q$ matrix of fully observed covariates = (X_{ij}) | 28 |
| $X_i = i$ th row of X = values of X for i th unit | 28 |
| $Y = N \times p$ matrix of partially observed outcome variables = (Y_{ij}) | 29 |
| $Y_i = i$ th row of Y = values of Y for i th unit | 29 |
| $Y_{[j]} = j$ th column of Y = j th outcome variable | 171 |
| $I = N \times p$ 0-1 indicator for inclusion of Y in survey = (I_{ij}) | 29 |
| $I_i = i$ th row of I = indicator for outcomes included for unit i | 29 |
| $R = N \times p$ 0-1 indicator for response on Y = (R_{ij}) | 30 |
| $R_i = i$ th row of R = indicator for response for i th unit | 30 |

Index Sets Describing Portions of Y

| | |
|---|-----|
| $inc = \{(i, j) I_{ij} = 1\}$ = included in survey | 48 |
| $exc = \{(i, j) I_{ij} = 0\}$ = excluded from survey | 48 |
| $obs = \{(i, j) I_{ij} R_{ij} = 1\}$ = observed | 48 |
| $nob = \{(i, j) I_{ij} R_{ij} = 0\}$ = not observed | 48 |
| $mis = \{(i, j) I_{ij}(1 - R_{ij}) = 1\}$ = missing (i.e., included but not observed) | 48 |
| $inc(i) = \{j I_{ij} = 1\}$ = included in survey for unit i | 162 |
| $exc(i) = \{j I_{ij} = 0\}$ = excluded from survey for unit i | 162 |
| $obs(i) = \{j I_{ij} R_{ij} = 1\}$ = observed for unit i | 162 |
| $nob(i) = \{j I_{ij} R_{ij} = 0\}$ = not observed for unit i | 162 |
| $mis(i) = \{j I_{ij}(1 - R_{ij}) = 1\}$ = missing for unit i | 162 |

| | |
|--|-----|
| $obs[j] = \{i I_{ij} R_{ij} = 1\}$ = units with $Y_{[j]}$ observed | 175 |
| $nob[j] = \{i I_{ij} R_{ij} = 0\}$ = units with $Y_{[j]}$ not observed | 176 |
| $mis[j] = \{i I_{ij}(1 - R_{ij}) = 1\}$ = units with $Y_{[j]}$ missing | 177 |
| $ob = \cup_j obs[j]$ = units with observed values | 165 |
| $ms = \cup_j mis[j]$ = units with missing values | 162 |

Complete-Data Statistics and Their Values in Multiply-Imputed Data Set

| | |
|--|----|
| $\hat{Q}(X, Y_{inc}, I)$ = complete-data estimate of k -component estimand | |
| $Q = Q(X, Y)$ | 75 |
| $U(X, Y_{inc}, I)$ = variance of $(Q - \hat{Q})$, $k \times k$ matrix | 75 |
| \hat{Q}_{*l} , $l = 1, \dots, m$ = values of \hat{Q} | 76 |
| U_{*l} , $l = 1, \dots, m$ = values of U | 76 |
| d_{*l} , $l = 1, \dots, m$ = values of χ^2 statistics testing $Q = Q_0$ | 78 |
| $S_m = \{\hat{Q}_{*l}, U_{*l}, l = 1, \dots, m\}$ | 88 |

Repeated-Imputation Summary Statistics Used to Create Repeated-Imputation Inference for Estimand Q

| | |
|--|----|
| $\bar{Q}_m = \sum_{l=1}^m \hat{Q}_{*l} / m$ = repeated-imputation estimate of Q , k -component row vector | 76 |
| $\bar{U}_m = \sum_{l=1}^m U_{*l} / m$, $k \times k$ matrix | 76 |
| $B_m = \sum_{l=1}^m (\hat{Q}_{*l} - \bar{Q}_m)'(\hat{Q}_{*l} - \bar{Q}_m) / (m - 1)$, $k \times k$ matrix | 76 |
| $T_m = \bar{U}_m + (1 + m)^{-1} B_m$ = total variance of $(Q - \bar{Q}_m)$, $k \times k$ matrix | 76 |
| $r_m = (1 + m^{-1}) \text{Tr}(B_m \bar{U}_m^{-1}) / k$ = fractional increase in variance due to nonresponse | 78 |
| $\nu = (m - 1)(1 + r_m^{-1})^2$ = degrees of freedom in reference distribution | 77 |
| $\gamma_m = \frac{r_m + 2/(\nu + 3)}{r_m + 1}$ = fraction of information missing due to nonresponse | 77 |
| $D_m = (Q_0 - \bar{Q}_m) T_m^{-1} (Q_0 - \bar{Q}_m)' / k$ = moment-based statistic testing $Q = Q_0$ | 77 |
| $\tilde{D}_m = (1 - r_m^{-1})(Q_0 - \bar{Q}_m) \bar{U}_m^{-1} (Q_0 - \bar{Q}_m)' / k$ = moment-based statistic testing $Q = Q_0$ | 78 |

| | |
|---|----|
| \hat{D}_m = statistic testing $Q = Q_0$; asymptotically equivalent to \tilde{D}_m | 78 |
| $\hat{\hat{D}}_m$ = test statistic approximating \hat{D}_m using only d_{*1}, \dots, d_{*m} | 79 |

**Variances Relevant for Randomization-Based Evaluations,
 $k \times k$ matrices**

| | |
|--|-----|
| $U_0 = U_0(X, Y) = E(U X, Y) = V(\hat{Q} X, Y)$ | 118 |
| $B_0 = B_0(X, Y) = E(B X, Y)$ where $B = V(\bar{Q}_\infty X, Y, I)$ | 119 |
| $T_0 = T_0(X, Y) = V(\bar{Q}_\infty X, Y)$ | 116 |
| γ_0 = eigenvalues of B_0 with respect to T_0 = population fractions of missing information | 131 |