# Multiple Imputation for Nonresponse in Surveys

DONALD B. RUBIN

Department of Statistics
Harvard University

JOHN WILEY & SONS

New York • Chichester • Brisbane • Toronto • Singapore

This Page Intentionally Left Blank

# Multiple Imputation for Nonresponse in Surveys

This Page Intentionally Left Blank

# Multiple Imputation for Nonresponse in Surveys

DONALD B. RUBIN

Department of Statistics
Harvard University

JOHN WILEY & SONS

New York • Chichester • Brisbane • Toronto • Singapore

*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

This Page Intentionally Left Blank

# 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 colleageal 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. Raghunathen 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 Raghunathen, Weld, Roderick J. A. Little and Alan Zaslavsky, and was facilitated by Raghunathen'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*

This Page Intentionally Left Blank

# Contents

# Tables and Figures

# 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