# Analysis of Incomplete Multivariate Data

**J. L. Schafer**
Department of Statistics
The Pennsylvania State University
USA

CHAPMAN & HALL/CRC

A CRC Press Company
Boca Raton London New York Washington, D.C.

**Library of Congress Cataloging-in-Publication Data**

Catalog record is available from the Library of Congress.



First edition 1997
First CRC Press reprint 1999
Originally published by Chapman & Hall
No claim to original U.S. Government works
International Standard Book Number 0-412-04061-1
Printed in the United States of America 3 4 5 6 7 8 9 0
Printed on acid-free paper

# Contents

# Preface

The last quarter of a century has seen enormous developments in general statistical methods for incomplete data. The EM algorithm and its extensions, multiple imputation and Markov chain Monte Carlo provide a set of flexible and reliable tools for inference in large classes of missing-data problems. Yet, in practical terms, these developments have had surprisingly little impact on the way most data analysts handle missing values on a routine basis. My hope is that this book will help to bridge the gap between theory and practice, making a multipurpose kit of missing-data tools accessible to anyone who may need them.

This book is intended for applied statisticians, graduate students and methodologically-oriented researchers in search of practical tools to handle missing data. The focus is applied rather than theoretical, but technical details have been included where necessary to help readers thoroughly understand the statistical properties of these methods and the behavior of the accompanying algorithms.

The methods presented here rely on three fully parametric models for multivariate data: the unrestricted multivariate normal distribution, loglinear models for cross-classified categorical data and the general location model for mixed continuous and categorical variables. In addition, the missing data are assumed to be missing at random, in the sense defined by Rubin (1976). My reviewers have correctly pointed out that many other vitally important topics could (and perhaps should) have been addressed: non-normal models such as the contaminated normal and multivariate-t; repeated measures and restricted covariance structures; censored and coarsened data; models for nonignorable nonresponse; latent variables; and hierarchical or random-effects models. Imputation for

complex surveys and censuses, a topic in which I am deeply interested, deserves much more attention than it received. For better or worse, I decided to limit the material to a few important subjects, but to treat these subjects thoroughly and illustrate them with non-trivial data examples.

This book would not have been possible without the generous support and encouragement of many friends, colleagues and agencies. Don Rubin, whose countless contributions to the area of missing data provided a springboard for this work, was the first to suggest publishing it as a book. The initial round of software development was sponsored by Frank Sulloway, whose wonderfully incomplete dataset provided the first and most colorful application of these methods. Additional support was provided by the Bureau of the Census, the United States Department of Agriculture and the National Center for Health Statistics. Many helpful comments and suggestions were given by John Barnard, Rose Brunner, Andrew Gelman, Bonnie Ghosh, Xiao-Li Meng, Susan Murphy, Maren Olsen, Ritz Scheuren, Stef van Buuren, Recai Yucel and Alan Zaslavsky, and the editorial and production staff at Chapman & Hall. Data on the Foreign Language Attitude Scale were contributed by Mark Raymond. My parents, Chester and Dolores Schafer, created a loving and stable childhood environment, and my wife Sharon did not fail to encourage and inspire. Prayer support was provided by Dr. Samuel C. Lee and members of University Bible Fellowship.

Finally, I must acknowledge my debt to the late Clifford C. Clogg, to whom this book is dedicated. Cliff's steady encouragement and careful review greatly improved the quality of the book, especially the first five chapters. His warmth, love for learning, hard work and faith continue to inspire the many who were close to him. Personally and professionally, it is most gratifying to know that Cliff regarded this book as 'good stuff'.

Joseph L. Schafer
*University Park, Pennsylvania*
*October 1996*