Interdisciplinary Statistics

# MEASUREMENT ERROR and MISCLASSIFICATION in STATISTICS and EPIDEMIOLOGY

## Impacts and Bayesian Adjustments

## CHAPMAN & HALL/CRC

*Interdisciplinary Statistics Series*

Series editors: N. Keiding, B. Morgan, T. Speed, P. van der Heijden

| | |
|---|---|
| **AN INVARIANT APPROACH TO STATISTICAL ANALYSIS OF SHAPES** | S. Lele and J. Richtsmeier |
| **ASTROSTATISTICS** | G. Babu and E. Feigelson |
| **CLINICAL TRIALS IN ONCOLOGY SECOND EDITION** | J. Crowley, S. Green, and J. Benedetti |
| **DESIGN AND ANALYSIS OF QUALITY OF LIFE OF STUDIES IN CLINICAL TRIALS** | Diane L. Fairclough |
| **DYNAMICAL SEARCH** | L. Pronzato, H. Wynn, and A. Zhigljavsky |
| **GRAPHICAL ANALYSIS OF MULTI-RESPONSE DATA** | K. Basford and J. Tukey |
| **INTRODUCTION TO COMPUTATIONAL BIOLOGY: MAPS, SEQUENCES, AND GENOMES** | M. Waterman |
| **MARKOV CHAIN MONTE CARLO IN PRACTICE** | W. Gilks, S. Richardson, and D. Spiegelhalter |
| **STATISTICAL ANALYSIS OF GENE EXPRESSION MICROARRAY DATA** | Terry Speed |
| **STATISTICS FOR ENVIRONMENTAL BIOLOGY AND TOXICOLOGY** | A. Bailer and W. Piegorsch |
| **STATISTICS IN MUSICOLOGY** | J. Beran |
| **MEASUREMENT ERROR AND MISCLASSIFICATION IN STATISTICS AND EPIDEMIOLOGY: IMPACTS AND BAYESIAN ADJUSTMENTS** | Paul Gustafson |

*Interdisciplinary Statistics*

# MEASUREMENT ERROR and MISCLASSIFICATION in STATISTICS and EPIDEMIOLOGY

## Impacts and Bayesian Adjustments

Paul Gustafson

CHAPMAN & HALL/CRC

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

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

Gustafson, Paul.
Measurement error and misclassification in statistics and epidemiology : impacts and Bayesian adjustments / Paul Gustafson
p. cm. (Chapman & Hall/CRC Interdisciplinary Statistics Series; 13)
Includes bibliographical references and index.
ISBN 1-58488-335-9
1. Error analysis (Mathematics). 2. Sequential analysis. 3. Bayesian statistical decision theory. I. Title.

QA275.G93 2003
511′.43—dc22 2003047290

This book contains information obtained from authentic and highly regarded sources. Reprinted material is quoted with permission, and sources are indicated. A wide variety of references are listed. Reasonable efforts have been made to publish reliable data and information, but the author and the publisher cannot assume responsibility for the validity of all materials or for the consequences of their use.

Direct all inquiries to CRC Press LLC, 2000 N.W. Corporate Blvd., Boca Raton, Florida 33431.

**Trademark Notice:** Product or corporate names may be trademarks or registered trademarks, and are used only for identification and explanation, without intent to infringe.

No claim to original U.S. Government works
International Standard Book Number 1-58488-335-9
Library of Congress Card Number 2003047290
Printed in the United States of America 1 2 3 4 5 6 7 8 9 0
Printed on acid-free paper

To Reka, Joseph, Lucas, and Anna

# Contents

# Preface

One of the most fundamental tasks in statistical science is to understand the relationship between some *explanatory* variables and a *response* or *outcome* variable. This book discusses problems and solutions associated with statistical analysis when an explanatory variable is *mismeasured*. That is, the variable in question cannot be observed, though a rough surrogate for it can be measured. While it is tempting to simply 'plug-in' this surrogate and proceed as usual, one must be mindful that scientific interest will be focussed on how the actual but unobservable variable impacts the response variable. When the variable in question is continuous in nature, this sort of scenario is often referred to as involving *errors-in-variables*, *errors-in-covariables*, or simply *measurement error*. When the explanatory variable in question is categorical, often *misclassification* is the term used. We deliberately use *mismeasurement* to encompass both scenarios.

Mismeasurement of explanatory variables is quite common in many scientific areas where statistical modelling is brought to bear. While the statistical discussion and development herein is general, our examples are drawn from biostatistics and epidemiology. In these fields, statistical models are often applied to explanatory variables which reflect human *exposure* of one kind or another. Many exposures of medical interest cannot be measured very precisely at the level of the individual, with prime examples being exposures to indoor and outdoor pollutants, and exposures based on food and drug intakes. In recognition of this problem, much of the literature on mismeasured explanatory variables has a biostatistical or epidemiological focus.

Perhaps it is easiest to characterize this book in terms of what it is not. First, it is not a textbook, nor is it strictly a research monograph. Rather, it presents a mix of expository material and research-oriented material. However, much of the expository material is presented in a new light, with nuances and emphases that differ from those of existing references. Second, this book is not comprehensive. The statistical literature on adjusting for mismeasurement is surprisingly large, and also somewhat unwieldy. I do not claim to cover all facets of it. Rather, the book focusses on some of the main ideas, and then delves selectively into more advanced topics. The choice of topics reflects my own views on what are interesting, relevant, and topical issues in understanding and adjusting for mismeasurement.

If asked to identify novel aspects of this book, I would start by noting its dual treatment of mismeasurement in both continuous and categorical variables. Historically there seems to be a somewhat artificial divide, with the continuous case dealt with in the statistics and biostatistics literature, and the categorical case dealt with in the epidemiology literature. In other statistical research areas the distinction between continuous and categorical explanatory variables is much less important. This book aims to draw the two topics together as much as possible, and to be of interest to both the (bio)statistics community, and at least the more quantitative subset of the epidemiology community.

A second novel feature of the book is that all the statistical modelling to adjust for mismeasurement is done within the *Bayesian* paradigm for statistical inference. This paradigm provides a conceptually simple and general approach to statistical analysis. The Bayesian approach has received much attention over the last decade or so, now that powerful *Markov chain Monte Carlo* computational algorithms are available to implement Bayesian methods in complex scenarios. While a number of research articles have demonstrated the utility of the Bayesian approach in adjusting for mismeasurement, to date there has not been an overarching reference on this topic.

A third noteworthy feature of the book is that a slightly sceptical 'wrong-model' view of the world is adopted in many places. In contrast to much of the literature, when considering the impact of unchecked mismeasurement it is usually not assumed that the postulated statistical models are correct. And there is some emphasis on how adjustments for mismeasurement will fare when the entertained models are not correct. The famous adage of G.E.P. Box that "all models are wrong, but some are useful" is quite central to much of the book's developments.

The book should be accessible to graduate students, practitioners, and researchers in Statistics and Biostatistics, and to Epidemiologists with particular interest in quantitative methods. Much of the mathematical detail in each chapter is collected in a final section, with the goal of making the main body of the chapter more accessible. As well, there is a largely self-contained Appendix introducing the main ideas of Bayesian analysis and Markov chain Monte Carlo algorithms, for the sake of readers without previous exposure to these topics.

I am grateful to a number of organizations and people for help in bringing this book to fruition. I thank the Natural Sciences and Engineering Research Council of Canada and the Canadian Institutes of Health Research for research grant support. I thank Kirsty Stroud, Jasmin Naim, and Jamie Sigal at Chapman and Hall/CRC for their help, and particularly for their patience when I adopted the typical professorial interpretation of deadlines. I thank all my colleagues in the Department of Statistics at U.B.C. for collectively providing such a convivial environment for academic pursuits, with particular thanks to Nhu Le for venturing into the mismeasurement world with me, and to Jim Zidek for suggesting I write this book in the first place. Above all, I thank my wife, and my children—of whom there were two when I started writing and three when I finished. They were all much better sports about this endeavour than they ought to have been!

# Guide to Notation

Some commonly used notation is as follows.

| | |
|---|---|
| $AF$ | attenuation factor |
| $Y$ | outcome or response variable (continuous or binary) |
| $X$ | unobservable continuous explanatory variable of interest |
| $X^*$ | observable surrogate for $X$ |
| $V$ | unobservable binary explanatory variable of interest |
| $V^*$ | observable surrogate for $V$ |
| $Z$ | other observable explanatory variables (of any type) |
| $\theta_M$ | all the parameters in the measurement model |
| $\tau^2$ | the variance of $X^*$ given $X$ in the measurement model |
| $\theta_R$ | all the parameters in the response model |
| $\beta$ | regression coefficients in the response model |
| $\theta_{E1}$ | all the parameters in the conditional exposure model |
| $\alpha$ | regression coefficients in the conditional exposure model |
| $(SN, SP)$ | sensitivity and specificity for $V^*$ as a surrogate for $V$ |
| $(p, q)$ | more compact notation for sensitivity and specificity |
| $(\tilde{p}, \tilde{q})$ | (incorrectly) assumed values of $(p, q)$ |
| $r_i$ | prevalence of binary exposure in $i$-th population |
| $\Psi$ | odds-ratio |
| $\psi$ | log odds-ratio |