Statistics for Social Science and Public Policy

Advisors:
S.E. Fienberg W. van der Linden

Nicholas T. Longford

Missing Data and Small-Area Estimation

Modern Analytical Equipment for the Survey Statistician

With 45 Figures

Springer

Nicholas T. Longford
SNTL
Oadby, Leicester LE2 5RL
England
NTL@SNTL.co.uk

Library of Congress Cataloging-in-Publication Data
Longford, Nicholas T., 1955–
Modern analytical equipment for the survey statistician : incomplete data and small-area estimation / Nicholas T. Longford.
p. cm. — (Statistics for social science and public policy)
Includes bibliographical references and index.
ISBN 1-85233-760-5 (alk. paper)
1. Surveys—Methodology. 2. Social surveys—Methodology. 3. Missing observations (Statistics) 4. Estimation theory. 5. Social sciences—Research—Statistical methods. I. Title. II. Series.
HA31.2.L66 2005
001.4′33—dc22 2005043229

ISBN-10: 1-85233-760-5 Printed on acid-free paper.
ISBN-13: 978-185233-760-5

Printed in the United States of America. (MVY)

9 8 7 6 5 4 3 2 1

springeronline.com

To all those who put me to the Task

Preface

This book evolved from lectures, courses and workshops on missing data and small-area estimation that I presented during my tenure as the first Campion Fellow (2000–2002). For the Fellowship I proposed these two topics as areas in which the academic statistics could contribute to the development of government statistics, in exchange for access to the operational details and background that would inform the direction and sharpen the focus of academic research. After a few years of involvement, I have come to realise that the separation of 'academic' and 'industrial' statistics is not well suited to either party, and their integration is the key to progress in both branches.

Most of the work on this monograph was done while I was a visiting lecturer at Massey University, Palmerston North, New Zealand. The hospitality and stimulating academic environment of their Institute of Information Science and Technology is gratefully acknowledged. I could not name all those who commented on my lecture notes and on the presentations themselves; apart from them, I want to thank the organisers and silent attendees of all the events, and, with a modicum of reluctance, the 'grey figures' who kept inquiring whether I was any nearer the completion of whatever stage I had been foolish enough to attach a date.

The first part of the book deals with analysis of incomplete data. The subject is a must for every survey analyst because large scale surveys without any missing data exist only in textbooks and superficial plans. Although [146] and [233] have exposed the deficiencies of trivial methods for handling incomplete data, they have influenced the practice in official statistics and epidemiology outside U.S.A. only at the margins. I have aimed the presentation at practicing and future survey analysts, setting aside much of the theory, and focussing on the general principles of letting all the substantial sources of uncertainty permeate through the entire estimation process, exploiting all the relevant information about the missing data and challenging the adopted untestable assumptions by sensitivity analysis that plays, within reason, the role of the devil's advocate. The solution, the method of multiple imputation,

is respectful of the analyst's work and is built around the methods, tools and software that are well suited, and may have been prepared at some cost, for the ideal complete-data setting.

The subject of the second part is small-area estimation. Although not a concern in every survey, it is becoming a prominent problem in government statistics as clients of established national surveys demand more and more detail about geographical and other divisions of the country, while the survey management is reluctant to conduct more extensive surveys because of escalating costs and increasing rates of nonresponse. Empirical Bayes models are the principal methodological tool at present. I review these methods and develop an approach that relies on a 'good' model much less than model-based methods do, pursuing the creed of making the best of the available information, irrespective of its format or source.

The third part, a single chapter, is a diversion from the focus on survey analysis. It addresses the problem of model uncertainty by drawing on the solution from small-area estimation. In brief, selection (of models, estimators, and the like) is replaced by synthesis, linearly combining estimators or predictors based on alternative models. In the process, I question some of the established wisdoms, such as the finite-sample efficiency of the maximum likelihood estimator and the imperative of basing all inferences on a model judged to be valid by error-prone criteria.

Chapters 5, 10 and 11 directly, and the other chapters indirectly, draw on several publications, some of them written with coauthors. The numerous anonymous referees and journal editors not only helped us to improve the manuscripts but also pointed to aspects and areas in which more rigour and further research was (and in some cases still is) required. I have been tested hardest of all by reactions to the material in Chapter 11. The encouraging comments prevailed, although I may have been a bit too harsh on some of the existing conventions. I want to thank the Editors of the *Journal of the Royal Statistical Society*, *Journal of Official Statistics* and *Statistics in Transition* for their permissions to use material from my publications in their journals.

All the computing described in this book was carried out in `Splus` and `R`. The data analysed in Sections 5 and 10 can be obtained from their original sources; I am not allowed to distribute them, but the code for their analysis as well as for the various illustrations is available from me on request (`NTL@sntl.co.uk`). I hope that the reader will realise early on that a computational and graphical environment in which all the statistical computations can be done and high-quality illustrations drawn as a matter of routine is an essential part of an effective statistician's toolkit. I want to convey my apologies to Grazia Pittau for treating her like a guinea pig in this regard.

Research, on small-area estimation in particular, involved some travel overseas, to NSD, Bergen, Norway; ZA, Cologne, Germany; ZUMA, Mannheim, Germany; and CEPS/INSTEAD, Differdange, Luxembourg. I want to acknowledge the support of the Travel and Mobility of Researchers, a EUROSTAT programme that funded some of this travel and the assistance of the

hosts with my research and general well-being. My former employer, De Montfort University, was generous in releasing me, up to a point, on these and other occasions.

I have greatly benefited from consulting engagements with Communities Scotland (formerly Scottish Homes), Edinburgh, and a three-year secondment at the Office for National Statistics, London. Ludi Simpson, by introducing me to a particular problem, and the U.S. National Center for Educational Statistics, by generous research grants, provided impetus that turned my attention to small-area estimation about a decade ago. Don Rubin has been an inexhaustible source of experience and wise advice on anything to do with missing data, and I have caught his incurable virus of multiple imputation.

Jim Ramsay gave me invaluable advice on manuscript preparation; I will not elaborate on the details of which elements I adhered to and which I have failed. Interactions with Albert Satorra have stimulated my interest in some 'missionary' aspects of the statistics profession, which probably come through in the text. Rolling the time back by a decade or two, Murray Aitkin helped to shape my ideas of how and why I want to work as a research statistician, and how this can be enjoyed, by myself and others.

I owe Nathan Jeffery for his competent IT support, on occasions beyond the call of duty. The best testament to the Springer-Verlag team is that this is not my first project with them ([152]).

Leicester, England *Nick Longford*
February 2005

Contents

Part I

Missing data