OXFORD STATISTICAL SCIENCE SERIES • 37

An Introduction to Model-Based Survey Sampling with Applications

Raymond L. Chambers and Robert G. Clark

An Introduction to Model-Based Survey Sampling with Applications

Raymond L. Chambers
Centre for Statistical and Survey Methodology, University of Wollongong, Australia

Robert G. Clark
Centre for Statistical and Survey Methodology, University of Wollongong, Australia

OXFORD
UNIVERSITY PRESS

OXFORD
UNIVERSITY PRESS

Great Clarendon Street, Oxford OX2 6DP

Oxford University Press is a department of the University of Oxford.
It furthers the University's objective of excellence in research, scholarship,
and education by publishing worldwide in

Oxford New York

Auckland Cape Town Dar es Salaam Hong Kong Karachi
Kuala Lumpur Madrid Melbourne Mexico City Nairobi
New Delhi Shanghai Taipei Toronto

With offices in

Argentina Austria Brazil Chile Czech Republic France Greece
Guatemala Hungary Italy Japan Poland Portugal Singapore
South Korea Switzerland Thailand Turkey Ukraine Vietnam

Oxford is a registered trade mark of Oxford University Press
in the UK and in certain other countries

Published in the United States
by Oxford University Press Inc., New York

First published 2012

British Library Cataloguing in Publication Data
Data available

Library of Congress Cataloguing in Publication Data
Data available

Typeset by SPI Publisher Services, Pondicherry, India
Printed and bound by
CPI Group (UK) Ltd, Croydon, CR0 4YY

ISBN 978–0–19–856662–5

1 3 5 7 9 10 8 6 4 2

OXFORD STATISTICAL SCIENCE SERIES

Series Editors

A. C. ATKINSON R. J. CARROLL D. J. HAND
D. M. TITTERINGTON J.-L. WANG

OXFORD STATISTICAL SCIENCE SERIES

For a full list of titles please visit
http://www.oup.co.uk/academic/science/maths/series/osss/

10. J.K. Lindsey: *Models for Repeated Measurements*
11. N.T. Longford: *Random Coefficient Models*
12. P.J. Brown: *Measurement, Regression, and Calibration*
13. Peter J. Diggle, Kung-Yee Liang, and Scott L. Zeger: *Analysis of Longitudinal Data*
14. J.I. Ansell and M.J. Phillips: *Practical Methods for Reliability Data Analysis*
15. J.K. Lindsey: *Modelling Frequency and Count Data*
16. J.L. Jensen: *Saddlepoint Approximations*
17. Steffen L. Lauritzen: *Graphical Models*
18. A.W. Bowman and A. Azzalini: *Applied Smoothing Techniques for Data Analysis*
19. J.K. Lindsey: *Models for Repeated Measurements, Second Edition*
20. Michael Evans and Tim Swartz: *Approximating Integrals via Monte Carlo and Deterministic Methods*
21. D.F. Andrews and J.E. Stafford: *Symbolic Computation for Statistical Inference*
22. T.A. Severini: *Likelihood Methods in Statistics*
23. W.J. Krzanowski: *Principles of Multivariate Analysis: A User's Perspective, Revised Edition*
24. J. Durbin and S.J. Koopman: *Time Series Analysis by State Space Methods*
25. Peter J. Diggle, Patrick Heagerty, Kung-Yee Liang, and Scott L. Zeger: *Analysis of Longitudinal Data, Second Edition*
26. J.K. Lindsey: *Nonlinear Models in Medical Statistics*
27. Peter J. Green, Nils L. Hjort, and Sylvia Richardson: *Highly Structured Stochastic Systems*
28. Margaret Sullivan Pepe: *The Statistical Evaluation of Medical Tests for Classification and Prediction*
29. Christopher G. Small and Jinfang Wang: *Numerical Methods for Nonlinear Estimating Equations*
30. John C. Gower and Garmt B. Dijksterhuis: *Procrustes Problems*
31. Margaret Sullivan Pepe: *The Statistical Evaluation of Medical Tests for Classification and Prediction, Paperback*
32. Murray Aitkin, Brian Francis, and John Hinde: *Statistical Modelling in GLIM4, Second Edition*
33. Anthony C. Davison, Yadolah Dodge, N. Wermuth: *Celebrating Statistics: Papers in Honour of Sir David Cox on his 80th Birthday*
34. Anthony Atkinson, Alexander Donev, and Randall Tobias: *Optimum Experimental Designs, with SAS*
35. M. Aitkin, B. Francis, J. Hinde, and R. Darnell: *Statistical Modelling in R*
36. Ludwig Fahrmeir and Thomas Kneib: *Bayesian Smoothing and Regression for Longitudinal, Spatial and Event History Data*
37. Raymond L. Chambers and Robert G. Clark: *An Introduction to Model-Based Survey Sampling with Applications*

Preface

The theory and methods of survey sampling are often glossed over in statistics education, with undergraduate programmes in statistics mainly concerned with introducing students to designs and procedures for choosing statistical models, checking model fit to available data, and estimating and making inferences about model parameters. Students may learn about models of considerable complexity, for example generalised linear models can be used for modelling the relationship of a range of explanatory variables to a response variable that can be continuous, binary or categorical. Increasingly, students are introduced to mixed models, time series models and models for spatial data, all of which are suitable for complex, correlated data sets. Non-parametric and semi-parametric methods based on kernel smoothing and spline smoothing are also increasingly important topics. In contrast, survey sampling is often only covered relatively briefly, and in contrast to these other topics, models either do not appear or are simple and are de-emphasised.

This is surprising because survey sampling is one of the most satisfying and useful fields of statistics:

- **The target of inference is satisfyingly solid and observable**. In classical modelling theory, the focus is on estimating model parameters that are intrinsically unobservable. In contrast, a primary aim in surveys is to estimate quantities defined on a finite population – quantities that can in principle be directly observed by carrying out a census of this population. For example, an aim in classical statistical modelling might be to estimate the expected value of income, assuming that the distribution of income can be characterised by a specified distributional family; in contrast, the aim in a survey could be to estimate the mean income for the population of working age citizens of a country at a certain point in time. This mean income actually exists, and so it is possible to check the performance of statistical procedures for estimating its value in specific populations, in a way that is not possible when estimating model parameters. The practicalities of running a survey are also considered by the statistician and the statistical researcher, perhaps more so than in other fields of statistics.
- **Survey sampling is a major field of application of statistics, and is one of the great success stories of mathematical statistics**. Before the mid-twentieth century, national statistics were based by and large on complete censuses of populations. This was enormously expensive and meant that only a limited range of data could be collected. Since then, the use of samples

has become widely accepted, due mainly to the leadership of mathematical statisticians in government statistical agencies, and to the rapid development of a body of theory and methods for probability sampling. Surveys remain a major area of application of statistics – probably **the** major area in terms of dollars spent. A high proportion of graduates from statistics programmes spend some of their career at organisations that conduct surveys.

- **The rich range of models and associated methods used in 'mainstream' statistics can also be used in survey sampling**. The key inferential objectives in survey sampling are fundamentally about prediction, and it is not difficult to transfer theoretical insights from mainstream statistics to survey sampling. Unfortunately, however, this remains a rather under-developed area because the use of models is often de-emphasised in undergraduate courses on survey sampling.

One of the reasons why modelling does not play much part when students are first taught sampling theory is that this theory has essentially evolved within the so-called design-based paradigm, with little or no reliance on models for inference. Instead, inference is based on the repeated sampling properties of estimators, where the repeated sampling is from a fixed finite population consisting of arbitrary data values. This is an attractive and logically consistent approach to inference, but is limiting because methods are required to work for virtually **any** population, and so cannot really exploit the properties of the **particular** population at hand. The model-assisted framework, which has existed in some form since the early 1970s, makes use of models for the population at hand, but in a limited way, so that the potential risks and benefits from modelling are likewise limited. This book is an introduction to the *model-based approach* to survey sampling, where estimators and inference are based on a model that is assumed to summarise the population of interest for a survey.

One way of presenting the model-based approach is to start with a very general linear model, or even generalised linear model, allowing also for any correlation structure in the population. Most of the methods in general use would then be special cases of this general model. We have instead chosen to start with simple models and build up from there, discussing the models suitable to different practical situations. With this aim in mind, this book is divided into three parts, with Part 1 focusing on estimating population totals under a range of models. Chapters 1 and 2 introduce survey sampling, and the model-based approach, respectively. Chapter 3 considers the simplest possible model, the homogenous population model. Chapter 4 extends this model to stratified populations. The stratified model is also quite simple, but nevertheless is very widely used in practice and is a good approximation to many populations of interest. Chapter 5 discusses linear regression models for populations with a single auxiliary variable, and Chapter 6 considers two level hierarchical populations made up of units grouped into clusters, with sampling carried out in two stages. Chapter 7 then integrates these results via the general linear population model. The approach in

Chapters 3 through 7 is to present a model and discuss its applicability, to derive efficient predictors of a population total, and then to explore sample design issues for these predictors.

Robustness to incorrectly specified models is of crucial importance in model-based survey sampling, particularly since much of the sample surveys canon has been model-free. Part 2 of this book therefore considers the properties of estimators based on incorrectly specified models. In practice, all statistical models are incorrect to a greater or lesser extent. To quote from Box and Draper (1987, page 74), 'all models are wrong; the practical question is how wrong do they have to be to not be useful'. Chapter 8 shows that robust sample designs exist, and that, under these designs, predictors of population totals will still be approximately unbiased (although perhaps less efficient), even if the assumed model is incorrect. Chapter 9 extends this exploration of robustness to the important problem of robustifying prediction variance estimators to model misspecification. Chapter 10 completes Part 2 of the book with an exploration of how survey sampling methods can be made robust to outliers (extreme observations not consistent with the assumed model), and also how flexible modelling methods like non-parametric regression can be used in survey sampling.

Parts 1 and 2 of this book are concerned with the estimation of population totals, and more generally with linear combinations of population values. This has historically been the primary objective of sample surveys, and still remains very important, but other quantities are becoming increasingly important. Part 3 therefore explores how model-based methods can be used in a variety of new problem areas of modern survey sampling. Chapter 11 discusses prediction of non-linear population quantities, including non-linear combinations of population totals, and population medians and quantiles. Prediction variance estimation for such complex statistics is the focus of Chapter 12, which discusses how subsampling methods can be used for this purpose. In practice, most surveys are designed to estimate a range of quantities, not just a single population total, and Chapter 13 considers issues in design and estimation for multipurpose surveys. Chapter 14 discusses prediction for domains, and Chapter 15 explores small area estimation methods, which are rapidly becoming important for many survey outputs. Finally, in Chapters 16 and 17 we consider efficient prediction of population distribution functions and the use of transformations in survey inference.

The book is designed to be accessible to undergraduate and graduate level students with a good grounding in statistics, including a course in the theory of linear regression. Matrix notation is not introduced until Chapter 7, and is avoided where possible to support readers less familiar with this notation. The book should also be a useful introduction to applied survey statisticians with some familiarity with surveys and statistics and who are looking for an introduction to the use of models in survey design and estimation.

Using models for survey sampling is a challenge, but a rewarding one. It can go wrong – if the model is not checked carefully against sample data, or

if samples are chosen poorly, then estimates and inferences will be misleading. But if the model is chosen well and sampling is robust, then the rich body of knowledge that exists on modelling can be used to understand the population of interest, and to exploit this understanding through tailored sample designs and estimators. We hope that this book will help in this process.

Ray Chambers and Robert Clark

April 2011

Acknowledgements

This book owes its existence to the many people who have influenced our careers in statistics, and particularly our work in survey sampling. In this context, Ken Foreman stands out as the person whose inspiration and support in Ray's early years in the field set him on the path that eventually led to this book and to the model-based ideas that it promotes, while Ken Brewer and our many colleagues at the Australian Bureau of Statistics provided us with the theoretical and practical challenges necessary to ensure that these ideas were always grounded in reality.

Early in Ray's career he was enormously privileged to study under Richard Royall, who opened his eyes to the power of model-based ideas in survey sampling, and Alan Ross, who convinced him that it was just as necessary to ensure that these ideas were translated into practical advice for survey practitioners. To a large extent, the first part of this book is our attempt to achieve this aim. The book itself has its origin in a set of lectures that Ray presented to Eustat in Bilbao in 2003. Subsequently David Holmes was invaluable in providing advice on how these lectures should be organised into a book and with preparation of the exercises.

Robert also had the privilege to work with some great colleagues and mentors. Frank Yu of the Australian Bureau of Statistics encouraged Robert to undertake study and research into the use of models in survey sampling. David Steel's supervision of Robert's PhD developed his knowledge and interest in this area, as did a year in the stimulating environment of the University of Southampton, enriched by interaction with too many friends and colleagues to mention. Robert would also like to express his appreciation of his parents for their lifelong love and support, and for passing on their belief in education.

Many research colleagues have contributed over the years to the different applications that are described in this book, and we have tried to make sure that their inputs have been acknowledged in the text. However, special thanks are due to Hukum Chandra who helped considerably with the material presented in Chapter 15 on prediction for small areas and to Alan Dorfman whose long-standing collaboration on the use of transformations in sample survey inference eventually led to Chapter 17, and whose insightful and supportive comments on the first draft of the book resulted in it being significantly improved.

The book itself has been a long time in preparation, and we would like to thank the editorial team at Oxford University Press, and in particular Keith

Mansfield, Helen Eaton, Alison Jones and Elizabeth Hannon, for their patience and support in bringing it to a conclusion. Finally, we would express our sincere thanks to our wives, Pat and Linda, for freely giving us the time that we needed to develop the ideas set out in this book. Without their support this book would never have been written.

Contents

PART I

Basics of Model-Based Survey Inference

Statistical models for study populations were used in survey design and inference almost from the very first scientific applications of the sampling method in the late nineteenth century. Following publication of Neyman's influential paper (Neyman, 1934), however, randomisation or design-based methods became the dominant paradigm in scientific and official surveys in the mid-twentieth century, and models were effectively relegated to the secondary role of 'assisting' in the identification of efficient estimators for unknown population quantities. See Lohr (1999) for a development of sampling theory based on this approach. This situation has changed considerably over the last 30 years, with a resurgence of interest in the explicit use of models in finite population inference. Valliant *et al.* (2000) provide a comprehensive overview of the use of models in sample survey inference. To a large extent, this interest in the use of models is due to two mutually reinforcing trends in modern sample surveys. The first is the need to provide sample survey solutions for inferential problems that lie outside the domain of design-based theory, particularly situations where standard probability-based sampling methods are not possible. The second is the need for methods of survey inference that can efficiently integrate the increasing volume and complexity of data sources provided by modern information technology. In particular, it has been the capacity of the model-based paradigm to allow inference under a wider and more realistic set of sampling scenarios, as well as its capacity to efficiently integrate multiple sources of information about the population of interest, that has driven this resurgence.

This book aims to provide the reader with an introduction to the basic concepts of model-based sample survey inference as well as to illustrate how it is being used in practice. In particular, in Part 1 of this book we introduce the reader to model-based survey ideas via a focus on four basic model 'types' that are in wide use. These are models for homogeneous populations, stratified populations, populations with regression structure and clustered populations, as well as combinations of these basic structures.