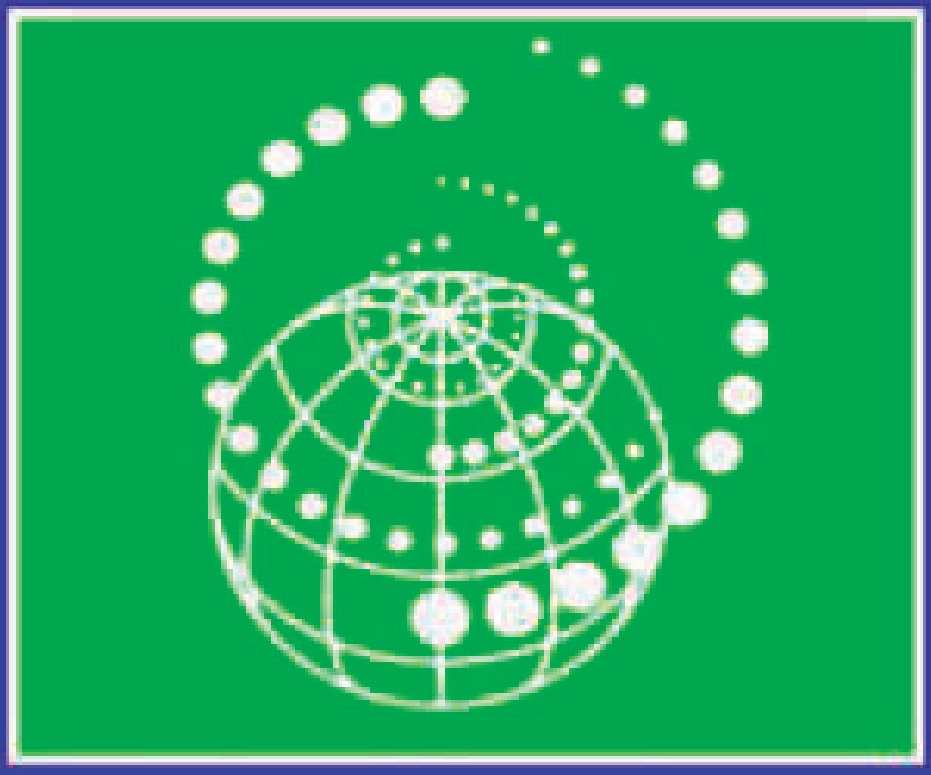
WILEY

Small Area Estimation



J. N. K. Rao

WILEY SERIES IN SURVEY METHODOLOGY

Small Area Estimation

WILEY SERIES IN SURVEY METHODOLOGY

Established in part by WALTER A. SHEWHART AND SAMUEL S. WILKS

A complete list of the titles in this series appears at the end of this volume.

Small Area Estimation

J. N. K. RAO
Carleton University

A JOHN WILEY & SONS, INC., PUBLICATION

Published by John Wiley & Sons, Inc., Hoboken, New Jersey.
Published simultaneously in Canada.

Library of Congress Cataloging-in-Publication Data:

Rao, J. N. K., 1937–
Small area estimation / J.N.K. Rao.
p. cm. — (Wiley series in survey methodology)
Includes bibliographical references and index.
ISBN 0-471-41374-7 (cloth)
1. Sampling (Statistics) 2. Estimation theory. I. Title. II. Series.

QA276.6 .R344 2003
519.5'2—dc21 2002033197

Printed in the United States of America.

10 9 8 7 6 5 4 3 2

To my mother, Sakuntalamma
and
to my wife, Neela

Contents

List of Figures

List of Tables

Foreword

The history of modern sample surveys dates back to the nineteenth century, but the field did not fully emerge until the 1930. It grew considerably during the World War II, and has been expanding at a tremendous rate ever since. Over time, the range of topics investigated using survey methods has broadened enormously as policy makers and researchers have learned to appreciate the value of quantitative data and as survey researchers – in response to policy makers' demands – have tackled topics previously considered unsuitable for study using survey methods. The range of analyses of survey data has also expanded, as users of survey data have become more sophisticated and as major developments in computing power and software have simplified the computations involved. In the early days, users were mostly satisfied with national estimates and estimates for major geographic regions and other large domains. The situation is very different today: more and more, policy makers are demanding estimates for small domains for use in making policy decisions. For example, population surveys are often required to provide estimates of adequate precision for domains defined in terms of some combination of such factors as age, sex, race/ethnicity, and poverty status. A particularly widespread demand from policy makers is for estimates at a finer level of geographic detail than the broad regions that were commonly used in the past. Thus estimates are frequently needed for such entities as states, provinces, counties, school districts, and health service areas.

The need to provide estimates for small domains has led to developments in two directions. One direction is toward the use of sample designs that can produce domain estimates of adequate precision within the standard design-based mode of inference used in survey analysis (i.e., "direct estimates"). Many sample surveys are now designed to yield sufficient sample sizes for key domains to satisfy the precision requirements for those domains. This approach is generally used for socio-economic domains and for some larger geographic domains. However, the increase in overall sample size that this approach entails may well exceed the survey's funding resources and capabilities, particularly so when estimates are required for many geographic areas. In the United States, for example, few surveys are large enough to be capable of providing reliable subpopulation estimates for all 50 states, even if the sample is optimally allocated across states for this purpose. For very small geographic areas like school districts, either a complete census or a sample of at least the size of the

census long-form sample (on average about 1 in 6 households nationwide) is required. Even censuses, however, although valuable, cannot be the complete solution for the production of small area estimates. In most countries censuses are conducted only once a decade. They cannot, therefore, provide satisfactory small area estimates for intermediate time points during a decade for population characteristics that change markedly over time. Furthermore, census content is inherently severely restricted, so a census cannot provide small area estimates for all the characteristics that are of interest. Hence another approach is needed.

The other direction for producing small area estimates is to turn away from conventional direct estimates toward the use of indirect model-dependent estimates. The model-dependent approach employs a statistical model that "borrows strength" in making an estimate for one small area from sample survey data collected in other small areas or at other time periods. This approach moves away from the design-based estimation of conventional direct estimates to indirect model-dependent estimates. Naturally, concerns are raised about the reliance on models for the production of such small area estimates. However, the demand for small area estimates is strong and increasing, and models are needed to satisfy that demand in many cases. As a result, many survey statisticians have come to accept the model-dependent approach in the right circumstances, and the approach is being used in a number of important cases. Examples of major small area estimation programs in the United States include: the Census Bureau's Small Area Income and Poverty Estimates program, which regularly produces estimates of income and poverty measures for various population subgroups for states, counties and school districts; the Bureau of Labor Statistics' Local Area Unemployment Statistics program, which produces monthly estimates of employment and unemployment for states, metropolitan areas, counties and certain subcounty areas; the National Agricultural Statistics Service's County Estimates Program, which produces county estimates of crop yield; and the estimates of substance abuse in states and metropolitan areas, which are produced by the Substance Abuse and Mental Health Services Administration (see Chapter 1).

The essence of all small area methods is the use of auxiliary data available at the small area level, such as administrative data or data from the last census. These data are used to construct predictor variables for use in a statistical model that can be used to predict the estimate of interest for all small areas. The effectiveness of small area estimation depends initially on the availability of good predictor variables that are uniformly measured over the total area. It next depends on the choice of a good prediction model. Effective use of small area estimation methods further depends on a careful, thorough evaluation of the quality of the model. Finally, when small area estimates are produced, they should be accompanied by valid measures of their precision.

Early applications of small area estimation methods employed only simple methods. At that time the choice of the method for use in particular case was relatively simple, being limited by the computable methods then in existence. However, the situation has changed enormously in recent years, and partic-

ularly in the last decade. There now exists a wide range of different, often complex models that can be used, depending on the nature of the measurement of the small area estimate (e.g., a binary or continuous variable) and on the auxiliary data available. One key distinction in model construction is between situations where the auxiliary data are available for the individual units in the population and those where they are available only at the aggregate level for each small area. In the former case, the data can be used in unit level models, whereas in the latter they can be used only in area level models. Another feature involved in the choice of model is whether the model borrows strength cross-sectionally, over time, or both. There are also now a number of different approaches, such as empirical best linear prediction (EBLUP), empirical Bayes (EB) and hierarchical Bayes (HB), that can be used to estimate the models and the variability of the model-dependent small area estimates. Moreover, complex procedures that would have been extremely difficult to apply a few years ago can now be implemented fairly straightforwardly, taking advantage of the continuing increases in computing power and the latest developments in software.

The wide range of possible models and approaches now available for use can be confusing to those working in this area. J.N.K. Rao's book is therefore a timely contribution, coming at a point in the subject's development when an integrated, systematic, treatment is needed. Rao has done a great service in producing this authoritative and comprehensive account of the subject. This book will help to advance the subject and be a valuable resource for practitioners and theorists alike.

GRAHAM KALTON

Preface

Sample surveys are widely used to provide estimates of totals, means and other parameters not only for the total population of interest but also for subpopulations (or domains) such as geographic areas and socio-demographic groups. Direct estimates of a domain parameter are based only on the domain-specific sample data. In particular, direct estimates are generally "design-based" in the sense that they make use of "survey weights" and the associated inferences (standard errors, confidence intervals, etc.) are based on the probability distribution induced by the sample design, with the population values held fixed. Standard sampling texts (e.g., the 1977 Wiley book *Sampling Techniques* by W.G. Cochran) provide extensive accounts of design-based direct estimation. Models that treat the population values as random may also be used to obtain model dependent direct estimates. Such estimates in general do not depend on survey weights and the associated inferences are based on the probability distribution induced by the assumed model (e.g., the 2001 Wiley book *Finite Population Sampling and Inference: A Prediction Approach* by R. Valliant, A.H. Dorfman and R.M. Royall).

We regard a domain as large if the domain sample size is large enough to yield direct estimates of adequate precision; otherwise, the domain is regarded as small. In this text, we generally use the term "small area" to denote any subpopulation for which direct estimates of adequate precision cannot be produced. Typically, domain sample sizes tend to increase with the population size of the domains, but this is not always the case. For example, due to oversampling of certain domains in the U.S. Third Health and Nutrition Examination Survey, sample sizes in many states were small (or even zero).

It is seldom possible to have a large enough overall sample size to support reliable direct estimates for all the domains of interest. Therefore, it is often necessary to use indirect estimates that "borrow strength" by using values of the variables of interest from related areas, thus increasing the "effective" sample size. These values are brought into the estimation process through a model (either implicit or explicit) that provides a link to related areas (domains) through the use of supplementary information related to the variables of interest, such as recent census counts and current administrative records. Availability of good auxiliary data and determination of suitable linking models are crucial to the formation of indirect estimates.

In recent years, the demand for reliable small area estimates has greatly increased worldwide. This is due, among other things, to their growing use in formulating policies and programs, the allocation of government funds and in regional planning. Demand from the private sector has also increased because business decisions, particularly those related to small businesses, rely heavily on the local conditions. Small area estimation is particularly important for studying the economies in transition in central and eastern European countries and the former Soviet Union countries because these countries are moving away from centralized decision making.

The main aim of this text is to provide a comprehensive account of the methods and theory of small area estimation, particularly indirect estimation based on explicit small area linking models. The model-based approach to small area estimation offers several advantages, most importantly, increased precision. Other advantages include the derivation of "optimal" estimates and associated measures of variability under an assumed model, and the validation of models from the sample data.

Chapter 1 introduces some basic terminology related to small area estimation and presents some important applications as motivating examples. Chapter 2 contains a brief account of direct estimation, which provides a background for later chapters. It also addresses survey design issues that have a bearing on small area estimation. Traditional demographic methods that employ indirect estimates based on implicit linking models are studied in Chapter 3. Typically, demographic methods only use administrative and census data and sampling is not involved, whereas indirect estimation methods studied in later chapters are largely based on sample survey data in conjunction with auxiliary population information. Chapter 4 gives a detailed account of traditional indirect estimation based on implicit linking models. The well-known James-Stein method of composite estimation is also studied in the context of sample surveys.

Explicit small area models that account for between area variation are presented in Chapter 5, including linear mixed models and generalized linear mixed models, such as logistic models with random area effects. The models are classified into two broad groups: (i) Area level models that relate the small area means to area-specific auxiliary variables, (ii) Unit level models that relate the unit values of study variables to unit-specific auxiliary variables. Several extensions to handle complex data structures, such as spatial dependence and time series structures, are also presented. Chapters 6–8 study in more detail linear mixed models involving fixed and random effects. General results on empirical best linear unbiased prediction (EBLUP) under the frequentist approach are presented in Chapter 6. The more difficult problem of estimating the mean squared error (MSE) of EBLUP estimators is also considered. A basic area level model and a basic unit level model are studied thoroughly in Chapter 7 by applying the EBLUP results developed in Chapter 6. Several important applications are also presented in this chapter. Various extensions of the basic models are considered in Chapter 8.

Chapter 9 presents empirical Bayes (EB) estimation. This method is more generally applicable than the EBLUP method. Various approaches to measuring the variability of EB estimators are presented. Finally, Chapter 10 presents a self-contained account of hierarchical Bayes (HB) estimation, by assuming prior distributions on model parameters. Both chapters include actual applications with real data sets.

Throughout the text, we discuss the advantages and limitations of the different methods for small area estimation. We also emphasize the need for both internal and external evaluations for model selection. To this end, we provide various methods of model validation, including comparisons of estimates derived from a model with reliable values obtained from external sources, such as previous census values.

Proofs of basic results are given in Sections 2.7, 3.5, 4.4, 6.4, 9.9 and 10.14, but proofs of results that are technically involved or lengthy are omitted. The reader is referred to relevant papers for details of omitted proofs. We provide self-contained accounts of direct estimation (Chapter 2), linear mixed models (Chapter 6), EB estimation (Chapter 9) and HB estimation (Chapter 10). But prior exposure to a standard course in mathematical statistics, such as the 1990 Wadsworth & Brooks/Cole book *Statistical Inference* by G. Casella and R.L. Berger, is essential. Also, a course in linear mixed models, such as the 1992 Wiley book *Variance Components* by S.R. Searle, G. Casella and C.E. McCulloch, would be helpful in understanding model based small area estimation. A basic course in survey sampling methods, such as the 1977 Wiley book *Sampling Techniques* by W.G. Cochran, is also useful but not essential.

This book is intended primarily as a research monograph, but it is also suitable for a graduate level course on small area estimation. Practitioners interested in learning small area estimation methods may also find portions of this text useful; in particular, Chapters 4, 7, 9, and Sections 10.1–10.3 and 10.5 as well as the applications presented throughout the book.

Special thanks are due to Gauri Datta, Sharon Lohr, Danny Pfeffermann, Graham Kalton, M.P. Singh, Jack Gambino and Fred Smith for providing many helpful comments and constructive suggestions. I am also thankful to Yong You, Ming Yu and Wesley Yung for typing portions of this text, to Gill Murray for the final typesetting and preparation of the text, and to Roberto Guido of Statistics Canada for designing the logo on the cover page. Finally, I am grateful to my wife Neela for her long enduring patience and encouragement and to my son, Sunil, and daughter, Supriya, for their understanding and support.

J.N.K. RAO

Ottawa, Canada
January, 2003

Small Area Estimation. J. N. K. Rao

ISBN: 0-471-41374-7

Chapter 1

Introduction

1.1 What is a Small Area?

Sample surveys have long been recognized as cost-effective means of obtaining information on wide-ranging topics of interest at frequent intervals over time. They are widely used in practice to provide estimates not only for the total population of interest but also for a variety of subpopulations (domains). Domains may be defined by geographic areas or socio-demographic groups or other subpopulations. Examples of a geographic domain (area) include a state/province, county, municipality, school district, unemployment insurance (UI) region, metropolitan area and health service area. On the other hand, a socio-demographic domain may refer to a specific age-sex-race group within a large geographic area. An example of "other domains" is the set of business firms belonging to a census division by industry group.

In the context of sample surveys, we refer to a domain estimator as "direct" if it is based only on the domain-specific sample data. A direct estimator may also use known auxiliary information, such as the total of an auxiliary variable, x, related to the variable of interest, y. A direct estimator is typically "design based" but it can also be motivated by and justified under models (see Section 2.1 of Chapter 2). Design based estimators make use of survey weights, and the associated inferences are based on the probability distribution induced by the sampling design with the population values held fixed (see Chapter 2). "Model assisted" direct estimators that make use of "working" models are also design based, aiming at making the inferences "robust" to possible model misspecification (see Chapter 2).

A domain (area) is regarded as large (or major) if the domain-specific sample is large enough to yield "direct estimates" of adequate precision. A domain is regarded as "small" if the domain-specific sample is not large enough to support direct estimates of adequate precision. Some other terms used to denote a domain with small sample size include "local area", "subdomain", "small subgroup", "subprovince" and "minor domain". In some applications,

many domains of interest (such as counties) may have zero sample size.

In this text, we generally use the term "small area" to denote any domain for which direct estimates of adequate precision cannot be produced. Typically domain sample size tends to increase with the population size of the domain, but this is not always the case. Sometimes the sampling fraction is made larger than the average fraction in small domains in order to increase the domain sample sizes and thereby increase the precision of domain estimates. Such oversampling was, for example, used in the U.S. Third Health and Nutrition Examination Survey (NHANES III) for certain domains in the cross-classification of sex, race/ethnicity, and age, in order that direct estimates of acceptable precision could be produced for those domains. This oversampling led to a greater concentration of the sample in certain states (e.g., California and Texas) than normal, and thereby exacerbated the common problem in national surveys that sample sizes in many states are small (or even zero). Thus, while direct estimates may be used to estimate characteristics of demographic domains with NHANES III, they cannot be used to estimate characteristics of many states. States may therefore be regarded as small areas in this survey. Even when a survey has large enough state sample sizes to support the production of direct estimates for the total state populations, these sample sizes may well not be large enough to support direct estimates for subgroups of the state populations, such as school-age children or persons in poverty. Clearly, it is seldom possible to have a large enough overall sample size to support reliable direct estimates for all domains. Furthermore, in practice it is not possible to anticipate all uses of the survey data, and "the client will always require more than is specified at the design stage" (Fuller (1999), p. 344).

In making estimates for small areas with adequate level of precision, it is often necessary to use "indirect" estimators that "borrow strength" by using values of the variable of interest, y, from related areas and/or time periods and thus increase the "effective" sample size. These values are brought into the estimation process through a model (either implicit or explicit) that provides a link to related areas and/or time periods through the use of supplementary information related to y, such as recent census counts and current administrative records. Three types of indirect estimators can be identified (Schaible (1996), Chapter 1): "domain indirect', "time indirect" and "domain and time indirect". A domain indirect estimator makes use of y-values from another domain but not from another time period. A time indirect estimator uses y-values from another time period for the domain of interest but not from another domain. On the other hand, a domain and time indirect estimator uses y-values from another domain as well as another time period. Some other terms used to denote an indirect estimator include "non-traditional", "small area", "model dependent" and "synthetic".

Availability of good auxiliary data and determination of suitable linking models are crucial to the formation of indirect estimators. As noted by Schaible (1996, Chapter 10), expanded access to auxiliary information through coordination and cooperation among different agencies is needed.

1.2 Demand for Small Area Statistics

Historically, small area statistics have long been used. For example, such statistics existed in eleventh century England and seventeenth century Canada, based on either census or on administrative records (Brackstone (1987)). Demographers have long been using a variety of indirect methods for small area estimation of population and other characteristics of interest in postcensal years. Typically, sampling is not involved in the traditional demographic methods (Chapter 3).

In recent years, the demand for small area statistics has greatly increased worldwide. This is due, among other things, to their growing use in formulating policies and programs, in the allocation of government funds and in regional planning. Legislative acts by national governments have increasingly created a need for small area statistics, and this trend will likely continue. Demand from the private sector has also increased because business decisions, particularly those related to small businesses, rely heavily on the local socio-economic, environmental and other conditions. Schaible (1996) provides an excellent account of the use of traditional and model-based indirect estimators in U.S. Federal Statistical Programs.

Small area estimation is of particular interest for the economies in transition in central and eastern European countries and the former Soviet Union countries. In the 1990s, these countries have moved away from centralized decision making. As a result, sample surveys are now used to produce estimates for large areas as well as small areas. Prompted by the demand for small area statistics, an International Scientific Conference on Small Area Statistics and Survey Designs was held in Warsaw, Poland, in 1992 and an International Satellite Conference on Small Area Estimation was held in Riga, Latvia, in 1999 to disseminate knowledge on small area estimation. See Kalton, Kordos and Platek (1993) and IASS Satellite Conference (1999) for the published conference proceedings.

Some other proceedings of conferences on small area estimation include National Institute on Drug Abuse (1979), Platek and Singh (1986) and Platek, Rao, Särndal and Singh (1987). Review papers on small area estimation include Rao (1986), Chaudhuri (1992), Ghosh and Rao (1994), Rao (1999), Marker (1999), Rao (2001b) and Pfeffermann (2002). A text on the theory of small area estimation has also appeared (Mukhopadhyay (1998)).

1.3 Traditional Indirect Estimators

Traditional indirect estimators, based on implicit linking models, include synthetic and composite estimators (Chapter 4). These estimators are generally design based and their design variances (i.e., variances with respect to the probability distribution induced by the sampling design) are usually small relative to the design variances of direct estimators. However, the indirect estimators will be generally design biased and the design bias will not decrease

as the overall sample size increases. If the implicit linking model is approximately true, then the design bias will be small, leading to significantly smaller design mean squared error (MSE) compared to the MSE of a direct estimator. Reduction in MSE is the main reason for using indirect estimators.

1.4 Small Area Models

Explicit linking models based on random area-specific effects that account for between area variation beyond that is explained by auxiliary variables included in the model will be called "small area models" (Chapter 5). Indirect estimators based on small area models will be called "model-based estimators". We classify small area models into two broad types: (i) Aggregate (or area) level models that relate small area direct estimators to area-specific covariates. Such models are necessary if unit (or element) level data are not available. (ii) Unit level models that relate the unit values of a study variable to unit-specific covariates. A basic area level and a basic unit level are introduced in Sections 5.2 and 5.3, respectively. Various extensions of the basic small area models are outlined in Section 5.4. Models given in Sections 5.2–5.4 are relevant for continuous responses y, and may be regarded as special cases of a general linear mixed model (Section 6.3). However, for binary or count variables y, generalized linear mixed models are used (Section 5.6); in particular, logistic linear mixed models for the binary case and loglinear mixed models for the count case.

A critical assumption for the unit level models is that the sample values obey the assumed population model, that is, sample selection bias is absent (see Section 5.3). For area level models, we assume the absence of informative sampling of the areas in situations where only some of the areas are selected to the sample, that is, the sample area values (the direct estimates) obey the assumed population model.

Inferences from model-based estimators refer to the distribution implied by the assumed model. Model selection and validation, therefore, play a vital role in model-based estimation. If the assumed models do not provide a good fit to the data, the model-based estimators will be model biased which, in turn, can lead to erroneous inferences. Several methods of model selection and validation are presented in Chapters 6, 7 and 10. It is also useful to conduct external evaluations by comparing indirect estimates (both traditional and model-based) to more reliable estimates or census values based on past data; see Examples 7.1.1 and 7.1.2 for both internal and external evaluations.

1.5 Model-Based Estimation

It is now generally accepted that when indirect estimators are to be used they should be based on explicit small area models. Such models define the way that the related data are incorporated in the estimation process. The

model-based approach to small area estimation offers several advantages: (1) "Optimal" estimators can be derived under the assumed model. (2) Area-specific measures of variability can be associated with each estimator unlike global measures (averaged over small areas) often used with traditional indirect estimators. (3) Models can be validated from the sample data. (4) A variety of models can be entertained depending on the nature of the response variables and the complexity of data structures (such as spatial dependence and time series structures).

In this text, we focus on empirical best linear unbiased prediction (EBLUP) estimators (Chapters 6–8), parametric empirical Bayes (EB) estimators (Chapter 9) and parametric hierarchical Bayes (HB) estimators (Chapter 10) derived from small area models. For the HB method, a further assumption on the prior distribution of model parameters is also needed. EBLUP is applicable for linear mixed models, whereas EB and HB are more generally valid.

The EBLUP method for general linear mixed models has been extensively used in animal breeding and other applications to estimate realized values of linear combinations of fixed and random effects. An EBLUP estimator is obtained in two steps: (i) The best linear unbiased predictor (BLUP), which minimizes the model MSE in the class of linear model unbiased estimators of the quantity of interest is first obtained. It depends on the variances (and covariances) of random effects in the model. (ii) An EBLUP estimator is obtained from the BLUP by substituting suitable estimators of variance parameters. Chapter 6 presents some unified theory of the EBLUP method for the general linear mixed model, which covers many small area models considered in the literature (Chapters 7 and 8). Estimation of model MSE of EBLUP estimators is studied in detail in Chapters 6–8. PROC MIXED in SAS software can be used to implement the EBLUP method.

Under squared error loss, the best predictor (BP) of a (random) small area quantity of interest (such as mean or proportion) is the conditional expectation of the quantity given the data and the model parameters. Distributional assumptions are needed for calculating the BP. The empirical BP (or EB) estimator is obtained from BP by substituting suitable estimators of model parameters. On the other hand, the HB estimator under squared error loss is obtained by integrating the BP with respect to the (Bayes) posterior distribution of model parameters derived from an assumed prior distribution of model parameters. The HB estimator under squared error loss is the posterior mean of the estimand, where the expectation is with respect to the posterior distribution of the quantity of interest given the data. The HB method uses the posterior variance as a measure of uncertainty associated with the HB estimator. Posterior (or credible) intervals for the quantity of interest can also be constructed from the posterior distribution of the quantity of interest. Currently, the HB method is being extensively used for small area estimation because it is straightforward, inferences are "exact" and complex problems using recently developed Markov chain Monte Carlo (MCMC) methods can be handled. Software for implementing the HB method is also available (sub-

section 10.2.5). Chapter 10 gives a self-contained account of the HB method and its applications to small area estimation.

"Optimal" model-based estimates of small area totals or means may not be suitable if the objective is to produce an ensemble of estimates whose distribution is in some sense close enough to the distribution of the corresponding estimands. We are also often interested in the ranks (e.g., ranks of schools, hospitals or geographical areas) or in identifying domains (areas) with extreme values. Ideally, it is desirable to construct a set of "triple-goal" estimates that can produce good ranks, a good histogram and good area-specific estimates. However, simultaneous optimization is not feasible, and it is necessary to seek a compromise set that can strike an effective balance between the three goals. Triple-goal EB estimation and constrained EB estimation that preserves the ensemble variance are studied in Section 9.6. A HB version of constrained estimation is studied in Section 10.13.

1.6 Some Examples

We conclude the introduction by presenting some important applications of small area estimation as motivating examples. Details of some of these applications, including auxiliary information used, are given in Chapters 7–10.

Health

Small area estimation of health related characteristics has attracted a lot of attention in the U.S. because of a continuing need to assess health status, health practices and health resources at both the national and subnational levels. Reliable estimates of health-related characteristics help in evaluating the demand for health care and the access individuals have to it. Health care planning often takes place at the state and sub-state levels because health characteristics are known to vary geographically. Health System Agencies in the United States, mandated by the National Health Planning Resource Development Act of 1994, are required to collect and analyze data related to the health status of the residents and to the health delivery systems in their health service areas (Nandram (1999)).

(i) The U.S. National Center for Health Statistics pioneered the use of synthetic estimation, based on implicit linking models, developing state estimates of disability and other health characteristics for different groups from the National Health Interview Survey(NHIS). Examples 4.2.2 and 10.11.3 give health applications from national surveys. Malec, Davis, and Cao (1999) studied HB estimation of overweight prevalence for adults by states, using data from NHANES III. Folsom, Shah and Vaish (1999) produced survey-weighted HB estimates of small area prevalence rates for states and age groups, for up to 20 binary variables related to drug use, using data from pooled National Household Surveys on Drug Abuse. Chattopadhyay, Lahiri, Larsen and Reimnitz (1999) studied EB estimates of state-wide prevalences of the use of alcohol and drugs (e.g., marijuana) among civilian non-institutionalized adults and

adolescents in the United States. These estimates are used for planning and resource allocation, and to project the treatment needs of dependent users.

(ii) *Mapping* of small area mortality (or incidence) rates of diseases, such as cancer, is a widely used tool in public health research. Such maps permit the analysis of geographical variation which may be useful for formulating and assessing aetiological hypotheses, resource allocation, and the identification of areas of unusually high risk warranting intervention; see Section 9.5. Direct (or crude) estimates of rates, called standardized mortality ratios (SMRs) can be very unreliable, and a map of crude rates can badly distort the geographical distribution of disease incidence or mortality because the map tends to be dominated by areas of low population. Disease mapping, using model-based estimators, has received increased attention in recent years. The October 2000 issue of *Statistics in Medicine* is devoted to the new developments in disease mapping. We give several examples of disease mapping in this text: see Examples 9.5.1, 9.7.1, 10.10.11 and 10.10.3. Typically, sampling is not involved in disease mapping applications.

Agriculture

The U.S. National Agricultural Service (NASS) publishes model-based county estimates of crop acreage using remote sensing satellite data as auxiliary information; see Example 7.2.1 for an application. County estimates assist the agricultural authorities in local agricultural decision making. Also, county crop yield estimates are used to administer federal programs involving payments to farmers if crop yields fall below certain levels. Example 4.2.2 gives an application of synthetic estimation to produce county estimates of wheat production in the state of Kansas based on a non-probability sample of farms. Chapters 6 and 7 of Schaible (1996) provide details of traditional and model-based indirect estimation methods used by NASS for county crop acreage and production.

Remote sensing satellite data and crop surveys are currently being used in India to produce crop yield estimates at the district level (Singh and Goel (2000)). Small area estimation techniques to obtain estimates of crop production at lower administrative units like "tehsil" or block level are also under study.

Income for small places

Example 7.1.1 gives details of an application of the EB (EBLUP) method of estimation of small area incomes, based on a basic area level linking model (see Section 7.1). This method, proposed originally by Fay and Herriot (1979), was adopted by the U.S. Bureau of the Census to form updated per capita income (PCI) for small places. This was the largest application (prior to 1990) of model-based estimators in a U.S. Federal Statistical Program. The PCI estimates are used to determine fund allocations to local government units (places) under the General Revenue Sharing Program.

Poverty counts

The Fay-Herriot method has been recently used to produce model-based current county estimates of poor school-age children in U.S.A. (National Research Council (2000)). Using these estimates, the U.S. Department of Education allocates annually over $7 billion of funds, (called Title I funds), to counties, and then states distribute the funds among school districts. The allocated funds support compensatory education programs to meet the needs of educationally disadvantaged children. In the past, funds were allocated on the basis of estimated counts from the previous census, but this allocation system had to be changed since the poverty counts vary significantly over time. EBLUP estimates in this application obtained from the Current Population Survey (CPS) using administrative data as auxiliary information. Example 7.1.2 presents details of this application. The reader is referred to National Research Council (2000) for further details.

Median income of four-person families

Estimates of the current median income of four-person families in each of the U.S. states are used to determine the eligibility for a program of energy assistance to low-income families administered by the U.S. Department of Health and Human Services. CPS and administrative information are used to produce model-based estimates, using extensions of the basic area level linking model; see Examples 8.1.1 and 8.3.3.

Small Area Estimation. J. N. K. Rao

ISBN: 0-471-41374-7

Chapter 2

Direct Domain Estimation

2.1 Introduction

Sample survey data are extensively used to provide reliable direct estimates of totals and means for the whole population and large areas or domains. As noted in Chapter 1, a direct estimator for a domain uses values of the variable of interest, y, only from the sample units in the domain. Sections 2.2–2.5 provide a brief account of direct estimation under the design-based or repeated sampling framework. We refer the reader to standard text books on sampling theory (e.g., Cochran (1977), Hedayat and Sinha (1991), Särndal, Swensson and Wretman (1992), Thompson (1997), Lohr (1999)) for a more extensive treatment of direct estimation.

Model-based methods have also been used to develop direct estimators and associated inferences. Such methods provide valid conditional inferences referring to the particular sample that has been drawn, regardless of the sampling design (see Brewer (1963), Royall (1970) and Valliant, Dorfman and Royall (2001)). But unfortunately, model-based strategies can perform poorly under model misspecification as the sample size in the domain increases. For instance, Hansen, Madow and Tepping (1983) introduced a model misspecification that is not detectable through tests of significance from sample sizes as large as 400, and then showed that the repeated sampling coverage probabilities of model-based confidence intervals on the population mean, $\bar{Y}$, are substantially less than the desired level and that the understatement becomes worse as the sample size decreases. This poor performance is largely due to asymptotic design-inconsistency of the model-based estimator with respect to the stratified random sampling design employed by Hansen et al. (1983). We do not consider model-based direct estimators in this monograph, but model-based methods will be extensively used in the context of indirect estimators and small sample sizes in the domains of interest. As noted in Chapter 1, an indirect estimator for a domain "borrows strength" by using the values of the study variable, y, from sample units outside the domain of interest.

The main intention of Chapter 2 is to provide some background material for later chapters and to indicate that direct estimation methods may sometimes suffice, particularly after addressing survey design issues that have a bearing on small area estimation (see Section 2.6). Effective use of auxiliary information through ratio and regression estimation is also useful in reducing the need for indirect small area estimators (Sections 2.3–2.5).

2.2 Design-based Approach

We assume the following somewhat idealized set-up and focus on the estimation of a population total or mean in Section 2.3. Direct estimators for domains are obtained in Section 2.4, using the results for population totals. A survey population U consists of N distinct elements (or ultimate units) identified through the labels $j = 1, \ldots, N$. We assume that a characteristic of interest, y, associated with element j can be measured exactly by observing element j. Thus measurement errors are assumed to be absent. The parameter of interest is the population total $Y = \sum_U y_j$ or the population mean $\bar{Y} = Y/N$, where $\sum_U$ denotes summation over the population elements j.

A sampling design is used to select a sample s from U with probability $p(s)$. The sample selection probability $p(s)$ can depend on known design variables such as stratum indicator variables and size measures of clusters. In practice, a sampling scheme is used to implement a sampling design. For example, a simple random sample of size n can be obtained by drawing n random numbers from 1 to N without replacement. Commonly used sampling designs include stratified simple random sampling (e.g., establishment surveys) and stratified multistage sampling (e.g., large-scale socio-economic surveys such as the Canadian Labour Force Survey and the Current Population Survey of the United States).

To make inferences on the total Y, we observe the y-values associated with the selected sample s. For simplicity, we assume that all the elements $j \in s$ can be observed, that is, complete response. In the design-based approach, an estimator $\hat{Y}$ of Y is said to be design-unbiased (or p–unbiased) if the design expectation of $\hat{Y}$ equals Y; that is,

$$E_p(\hat{Y}) = \sum p(s)\hat{Y}_s = Y, \tag{2.2.1}$$

where the summation is over all possible samples s under the specified design and $\hat{Y}_s$ is the value of $\hat{Y}$ for the sample s. The design variance of $\hat{Y}$ is denoted as $V_p(\hat{Y}) = E_p\left[\hat{Y} - E_p(\hat{Y})\right]^2$. An estimator of $V_p(\hat{Y})$ is denoted as $v(\hat{Y}) = s^2(\hat{Y})$, and the variance estimator $v(\hat{Y})$ is p-unbiased for $V(\hat{Y})$ if $E_p[v(\hat{Y})] \equiv V_p(\hat{Y})$. An estimator of $\hat{Y}$ is design-consistent (or p-consistent) if $\hat{Y}$ is p-unbiased (or its p-bias tends to zero as the sample size increases) and $V_p(\hat{Y})$ tends to zero as the sample size increases. Strictly speaking, we need to consider p-consistency in the context of a sequence of populations U_ν such that

both the sample size n_ν and the population size N_ν tend to ∞ as $\nu \to \infty$. p-consistency of the variance estimator $v(\hat{Y})$ is similarly defined. If the estimator $\hat{Y}$ and variance estimator $v(\hat{Y})$ are both p-consistent, then the design-based approach provides valid inferences on Y regardless of the population values in the sense that the pivotal $t = (\hat{Y} - Y)/s(\hat{Y})$ converges in distribution ($\to_d$) to a $N(0,1)$ variable as the sample size increases. Thus in repeated sampling about $100(1-\alpha)\%$ of the confidence intervals $\left[\hat{Y} - z_{\alpha/2}s(\hat{Y}), \hat{Y} + z_{\alpha/2}s(\hat{Y})\right]$ contain the true value Y as the sample size increases, where $z_{\alpha/2}$ is the upper $\alpha/2$- point of a $N(0,1)$ variable. In practice, one often reports only the estimate (realized value of $\hat{Y}$) and associated standard error (realized value of $s(\hat{Y})$) or coefficient of variation (realized value of $c(\hat{Y}) = s(\hat{Y})/\hat{Y}$). Coefficient of variation or standard error is used as a measure of variability associated with the estimate.

The design-based (or probability sampling) approach has been criticized on the grounds that the associated inferences, although assumption-free, refer to repeated sampling instead of just the particular sample s that has been drawn. A conditional design-based approach that allows us to restrict the set of samples used for inference to a "relevant" subset has also been proposed. This approach leads to conditionally valid inferences. For example, in the context of poststratification (i.e., stratification after selection of the sample) it makes sense to make design-based inferences conditional on the realized poststrata sample sizes (Holt and Smith (1979); Rao (1985)). Similarly, when the population total X of an auxiliary variable x is known, conditioning on the estimator $\hat{X}$ of X is justified because the distance $|\hat{X} - X|/X$ provides a measure of imbalance in the realized sample (Robinson (1987); Rao (1992); Casady and Valliant (1993)).

2.3 Estimation of Totals

2.3.1 Design-unbiased Estimator

Design weights $w_j(s)$ play an important role in constructing design-based estimators $\hat{Y}$ of Y. These basic weights may depend both on s and the element j ($j \in s$). An important choice is $w_j(s) = 1/\pi_j$, where $\pi_j = \sum_{\{s:j\in s\}} p(s)$, $j = 1, 2, \ldots, N$ are the inclusion probabilities and $\{s : j \in s\}$ denotes summation over all samples s containing the element j. To simplify the notation, we write $w_j(s) = w_j$ except when the full notation $w_j(s)$ is needed. The weight w_j may be interpreted as the number of elements in the population represented by the sample element j.

In the absence of auxiliary population information, we use the expansion estimator

$$\hat{Y} = \Sigma_s w_j y_j, \tag{2.3.1}$$

where Σ_s denotes summation over $j \in s$. In this case, the design-unbiasedness

condition (2.2.1) reduces to

$$\sum_{\{s:j\in s\}} p(s)w_j(s) = 1;\ j = 1,\ldots,N. \tag{2.3.2}$$

The choice $w_j(s) = 1/\pi_j$ satisfies the unbiasedness condition (2.3.2) and leads to the well-known Horvitz-Thompson (H-T) estimator (see Cochran (1977), p. 259).

It is convenient to denote $\hat{Y} = \Sigma_s w_j y_j$ in an operator notation as $\hat{Y} = \hat{Y}(y)$. Using this notation, we have $\hat{Y}(x) = \Sigma_s w_j x_j$ for another variable x, whereas the traditional notation is to denote $\Sigma_s w_j x_j$ as $\hat{X}$. Similarly, we denote a variance estimator of $\hat{Y}$ as $v(\hat{Y}) = v(y)$. Using this notation, we have $v(\hat{X}) = v[\hat{Y}(x)] = v(x)$. Note that $\hat{Y}(x)$ and $v(x)$ are obtained by attaching the subscript j to the character, x, in the brackets and then replacing y_j by x_j in the formulas for $\hat{Y}(y)$ and $v(y)$. Hartley (1959) introduced the operator notation. We refer the reader to Cochran (1977), Särndal et al. (1992) and Wolter (1985) for details on variance estimation. Rao (1979) has shown that a nonnegative unbiased quadratic estimator of variance of $\hat{Y}$ is necessarily of the form

$$v(\hat{Y}) = v(y) = -\sum_{j<k,\ j,k\in s}\sum w_{jk}(s)b_j b_k\left(\frac{y_j}{b_j} - \frac{y_k}{b_k}\right)^2, \tag{2.3.3}$$

where the weights $w_{jk}(s)$ satisfy the unbiasedness condition and the nonzero constants b_j are such that the variance of $\hat{Y}$ becomes zero when $y_j \propto b_j$ for all j. For example, in the special case of $w_j = 1/\pi_j$ and a fixed sample size design, we have $b_j = \pi_j$ and $w_{jk}(s) = (\pi_{jk} - \pi_j\pi_k)/(\pi_{jk}\pi_j\pi_k)$ where $\pi_{jk} = \sum_{\{s:(j,k)\in s\}} p(s)$ $j < k = 1,\ldots,N$ are the joint inclusion probabilities assumed to be positive. The variance estimator (2.3.3) in this case reduces to the well-known Sen-Yates-Grundy (S-Y-G) variance estimator (see Cochran (1977), p. 261).

Under stratified multistage sampling, the expansion estimator $\hat{Y}$ may be written as

$$\hat{Y} = \hat{Y}(y) = \Sigma_s w_{hlk} y_{hlk}, \tag{2.3.4}$$

where w_{hlk} is the design weight associated with the kth element in the l-th primary sampling unit (cluster) belonging to the hth stratum, y_{hlk} is the associated y-value and Σ_s is the summation over all elements $j = (hlk) \in s$ $(h = 1,\ldots,L; l = 1,\ldots,n_h)$. It is common practice to treat the sample as if the clusters are sampled with replacement and subsampling is done independently each time a cluster is selected. This leads to overestimation of the variance but the variance estimator is greatly simplified. We have

$$v(\hat{Y}) = v(y) = \sum_h \frac{1}{n_h(n_h-1)}\sum_l (y_{hl} - \bar{y}_h)^2, \tag{2.3.5}$$

where $y_{hl} = \sum_k (n_h w_{hlk}) y_{hlk}$ are weighted sample cluster totals and $\bar{y}_h = \sum_l y_{hl}/n_h$. Note that $v(\hat{Y})$ depends on the y_{hlk}'s only through the totals y_{hl}. The relative bias of $v(\hat{Y})$ is small if the first-stage sampling fraction is small in each stratum. Note that $\hat{Y}(x)$ and $v(x)$ are obtained by attaching the subscripts hlk to the character, x, in the brackets and then replacing y_{hlk} by x_{hlk} in the formulas (2.3.4) and (2.3.5) for $\hat{Y}(x)$ and $v(y)$, respectively.

2.3.2 Generalized Regression Estimator

Suppose now that auxiliary information in the form of known population totals $\mathbf{X} = (X_1, \ldots, X_p)^T$ is available and that the auxiliary vector $\mathbf{x}_j$ for $j \in s$ is also observed, that is, the data $(y_j, \mathbf{x}_j)$ for each element $j \in s$ are observed. An estimator that makes efficient use of this auxiliary information is the generalized regression (GREG) estimator which may be written as

$$\hat{Y}_{\mathrm{GR}} = \hat{Y} + (\mathbf{X} - \hat{\mathbf{X}})^T \hat{\mathbf{B}}, \tag{2.3.6}$$

where $\hat{\mathbf{X}} = \Sigma_s w_j \mathbf{x}_j = \hat{Y}(\mathbf{x})$ and $\hat{\mathbf{B}} = (\hat{B}_1, \ldots, \hat{B}_p)^T = \hat{\mathbf{B}}(y)$ is the solution of the sample weighted least squares equations:

$$\left(\Sigma_s w_j \mathbf{x}_j \mathbf{x}_j^T / c_j\right) \hat{\mathbf{B}} = \Sigma_s w_j \mathbf{x}_j y_j / c_j \tag{2.3.7}$$

with specified constants $c_j (> 0)$. It is also useful to write $\hat{Y}_{\mathrm{GR}}$ in the expansion form with design weights w_j changed to "revised" weights w_j^*. We have

$$\hat{Y}_{\mathrm{GR}} = \Sigma_s w_j^* y_j = \hat{Y}_{\mathrm{GR}}(y) \tag{2.3.8}$$

in operator notation, where $w_j^* = w_j^*(s) = w_j(s) g_j(s)$ with

$$g_j(s) = 1 + (\mathbf{X} - \hat{\mathbf{X}})^T \left(\Sigma_s w_j \mathbf{x}_j \mathbf{x}_j^T / c_j\right)^{-1} \mathbf{x}_j / c_j. \tag{2.3.9}$$

The revised weight $w_j^*(s)$ is the product of the design weight $w_j(s)$ and the estimation weight $g_j(s)$. The form (2.3.8) shows that the same weight w_j^* is applied to all variables of interest as in the case of the expansion estimator. This ensures consistency of results when aggregated over different variables, that is,

$$\hat{Y}_{\mathrm{GR}}(y_1) + \ldots + \hat{Y}_{\mathrm{GR}}(y_r) = \hat{Y}_{\mathrm{GR}}(y_1 + \ldots + y_r)$$

for different variables $y_1, \ldots, y_r$ attached to each element.

An important property of the GREG estimator is that it ensures consistency with the known auxiliary totals $\mathbf{X}$ in the sense

$$\hat{Y}_{\mathrm{GR}}(\mathbf{x}) = \Sigma_s w_j^* \mathbf{x}_j = \mathbf{X}. \tag{2.3.10}$$

A proof of (2.3.10) is given in subsection 2.7.1. This property does not hold for the basic expansion estimator $\hat{Y}$. Many agencies regard this property as desirable from the user's viewpoint. Because of the property (2.3.10), the GREG

estimator is also called a calibration estimator (Deville and Särndal (1992)). In fact, among all calibration estimators of the form $\Sigma_s b_j y_j$ with weights b_j satisfying the calibration constraints $\Sigma_s b_j \mathbf{x}_j = \mathbf{X}$, the GREG weights w_j^* minimize a chi-squared distance, $\Sigma_s c_j (w_j - b_j)^2 / w_j$, between the basic weights w_j and the calibration weights b_j (see subsection 2.7.2). Thus the GREG weights w_j^* modify the design weights as little as possible subject to the calibration constraints.

The GREG estimator takes a simpler form when $c_j = \boldsymbol{\nu}^T \mathbf{x}_j$ for all $j \in U$ and some constant column vector $\boldsymbol{\nu}$. In this case we have

$$\hat{Y}_{\mathrm{GR}} = \mathbf{X}^T \hat{\mathbf{B}} = \Sigma_s \tilde{w}_j y_j \tag{2.3.11}$$

because $\Sigma_s w_j e_j(s) = \hat{Y} - \hat{\mathbf{X}}^T \hat{\mathbf{B}} = 0$ (see subsection 2.7.3), where $e_j(s) = e_j = y_j - \mathbf{x}_j^T \hat{\mathbf{B}}$ are the sample residuals and $\tilde{w}_j = \tilde{w}_j(s) = w_j(s) \tilde{g}_j(s)$ with

$$\tilde{g}_j(s) = \mathbf{X}^T \left(\Sigma_s w_j \mathbf{x}_j \mathbf{x}_j^T / c_j \right)^{-1} \mathbf{x}_j / c_j. \tag{2.3.12}$$

The GREG estimator covers many practically useful estimators as special cases. For example, in the case of a single auxiliary variable x we get the well-known ratio estimator

$$\hat{Y}_R = \frac{\hat{Y}}{\hat{X}} X, \tag{2.3.13}$$

by setting $c_j = x_j$ in (2.3.12) and noting that $\tilde{g}_j(s) = X/\hat{X}$. The ratio estimator $\hat{Y}_R$ uses the weights $\tilde{w}_j = w_j(X/\hat{X})$. If only the population size N is known, we set $x_j = 1$ in (2.3.13) so that $X = N$ and $\hat{X} = \hat{N} = \Sigma_s w_j$. If we set $\mathbf{x}_j = (1, x_j)^T$ and $c_j = 1$, then $\boldsymbol{\nu} = (1, 0)^T$ and (2.3.11) reduces to the familiar linear regression estimator

$$\hat{Y}_{\mathrm{LR}} = \hat{Y} + \hat{B}_{\mathrm{LR}}(X - \hat{X}), \tag{2.3.14}$$

where

$$\hat{B}_{\mathrm{LR}} = \hat{B}_2 = \Sigma_s w_j (x_j - \hat{\bar{X}})(y_j - \hat{\bar{Y}}) / \Sigma_s w_j (x_j - \hat{\bar{X}})^2$$

with $\hat{\bar{Y}} = \hat{Y}/\hat{N}$ and $\hat{\bar{X}} = \hat{X}/\hat{N}$. The GREG estimator also covers the familiar poststratified estimator as a special case. Suppose we partition U into G poststrata U_g (e.g., age/sex groups) with known population counts N_g ($g = 1, \ldots, G$). Then we set $\mathbf{x}_j = (x_{1j}, \ldots, x_{Gj})^T$ with $x_{gj} = 1$ if $j \in U_g$ and $x_{gj} = 0$ otherwise so that $\mathbf{X} = (N_1, \ldots, N_G)^T$. Noting that $\mathbf{1}^T \mathbf{x}_j = 1$ for all j we take $c_j = 1$ and $\boldsymbol{\nu} = \mathbf{1}$ and the GREG estimator (2.3.11) reduces to the poststratified estimator

$$\hat{Y}_{\mathrm{PS}} = \sum_g \frac{N_{.g}}{\hat{N}_{.g}} \hat{Y}_{.g}, \tag{2.3.15}$$

where $\hat{N}_{.g} = \sum_{s_{.g}} w_j$ and $\hat{Y}_{.g} = \sum_{s_{.g}} w_j y_j$, with $s_{.g}$ denoting the sample of elements belonging to poststratum g.

Turning to variance estimation, the traditional Taylor linearization method gives

$$v_L(\hat{Y}_{\mathrm{GR}}) = v(e), \tag{2.3.16}$$

which is obtained by substituting the residuals e_j for y_j in $v(y)$. Simulation studies have indicated that $v_L(\hat{Y}_{\mathrm{GR}})$ may lead to slight underestimation, whereas an alternative estimator

$$v(\hat{Y}_{\mathrm{GR}}) = v(ge), \tag{2.3.17}$$

obtained by substituting $g_j e_j$ for y_j in $v(y)$, reduces this underestimation, where $g_j = g_j(s)$ (Estevao, Hidiroglou and Särndal (1995)). The alternative variance estimator $v(\hat{Y}_{\mathrm{GR}})$ also performs better for conditional inference in the sense that it is approximately unbiased for the model variance of $\hat{Y}_{\mathrm{GR}}$ conditionally on s for several designs, under the following linear regression model (or GREG model):

$$y_j = \mathbf{x}_j^T \boldsymbol{\beta} + \epsilon_j, \ j \in U \tag{2.3.18}$$

where $E_m(\epsilon_j) = 0, V_m(\epsilon_j) = c_j \sigma^2, \mathrm{Cov}_m(e_j, e_k) = 0$ for $j \neq k$, and E_m, V_m and Cov_m, respectively, denote the model expectation, variance and covariance (Särndal, Swensson and Wretman (1989); Rao (1994)). In the model-based framework, y_j is a random variable and s is fixed. The GREG estimator is also model-unbiased under (2.3.18) in the sense $E_m(\hat{Y}_{\mathrm{GR}}) = E_m(Y)$ for every s. In the design-based framework, $\hat{Y}_{\mathrm{GR}}, v_L(\hat{Y}_{\mathrm{GR}})$ and $v(\hat{Y}_{\mathrm{GR}})$ are p-consistent.

We refer the reader to Särndal et al. (1992) for a detailed account of GREG estimation and to Estevao et al. (1995) for the highlights of a Generalized Estimation System at Statistics Canada based on GREG estimation theory.

2.4 Domain Estimation

2.4.1 Case of no Auxiliary Information

Suppose U_i denotes a domain (or subpopulation) of interest and that we are required to estimate the domain total $Y_i = \sum_{U_i} y_j$ or the domain mean $\bar{Y}_i = Y_i / N_i$, where N_i, the number of elements in U_i, may or may not be known. If y_i is binary (1 or 0), then $\bar{Y}_i$ reduces to the domain proportion P_i; for example, the proportion in poverty in the ith domain. Much of the theory in Section 2.3 for a total can be adapted to domain estimation by using the following relationships. Writing Y in the operator notation as $Y(y)$ and defining

$$y_{ij} = \begin{cases} y_j & \text{if } j \in U_i \\ 0 & \text{otherwise,} \end{cases}$$

$$a_{ij} = \begin{cases} 1 & \text{if } j \in U_i \\ 0 & \text{otherwise,} \end{cases}$$

we have

$$Y(y_i) = \sum_{j \in U} y_{ij} = \sum_{j \in U_i} y_j = Y_i \tag{2.4.1}$$

and

$$Y(a_i) = \sum_{j \in U} a_{ij} = \sum_{j \in U_i} 1 = N_i. \tag{2.4.2}$$

Note that y_{ij} may also be written as $a_{ij}y_j$. If the domains of interest, say, $U_1, \ldots, U_m$ form a partition of U (or of a larger domain), it is desirable from a user's viewpoint to ensure that the estimates of domain totals add up to the estimate of population total.

In the absence of auxiliary population information, we use the expansion estimator

$$\hat{Y}_i = \hat{Y}(y_i) = \sum_{j \in s} w_j y_{ij} = \sum_{j \in s_i} w_j y_j, \tag{2.4.3}$$

where s_i denotes the sample of elements belonging to domain U_i. It readily follows from (2.4.1) that $\hat{Y}_i$ is p-unbiased for Y_i if $\hat{Y}$ is p-unbiased for Y. It is also p-consistent if the expected domain sample size is large. Similarly, $\hat{N}_i = \hat{Y}(a_i)$ is p-unbiased for N_i, using (2.4.2). We note from (2.4.3) that the additive property is satisfied: $\hat{Y}_1 + \ldots + \hat{Y}_m = \hat{Y}$.

Noting that $v(\hat{Y}) = v(y)$, an estimator of the variance of $\hat{Y}_i$ is simply obtained from $v(y)$ by changing y_j to y_{ij}:

$$v(\hat{Y}_i) = v(y_i). \tag{2.4.4}$$

It follows from (2.4.3) and (2.4.4) that no new theory is required for domain estimation.

The domain mean $\bar{Y}_i = Y(y_i)/Y(a_i)$ is estimated by

$$\hat{\bar{Y}}_i = \frac{\hat{Y}(y_i)}{\hat{Y}(a_i)} = \frac{\hat{Y}_i}{\hat{N}_i}. \tag{2.4.5}$$

If $y_j = 1$ or 0, then $\hat{\bar{Y}}_i$ reduces to $\hat{P}_i$, an estimator of the domain proportion P_i. If the expected domain sample size is large, the ratio estimator (2.4.5) is p-consistent, and a Taylor linearization variance estimator is given by

$$v_L(\hat{\bar{Y}}_i) = v(\tilde{e}_i)/\hat{N}_i^2, \tag{2.4.6}$$

where $\tilde{e}_{ij} = y_{ij} - \hat{\bar{Y}}_i a_{ij}$. Note that $v(\tilde{e}_i)$ is obtained from $v(y)$ by changing y_j to $\tilde{e}_{ij}$. It follows from (2.4.5) and (2.4.6) that no new theory is required for domain means as well. Note that $\tilde{e}_{ij} = 0$ if $j \in s$ and $j \notin U_i$. We refer the reader to Hartley (1959) for domain estimation.

2.4.2 GREG Estimation

GREG estimation of Y is also easily adapted to estimation of a domain total Y_i. It follows from (2.3.8) that the GREG estimator of Y_i is

$$\hat{Y}_{i\mathrm{GR}} = \hat{Y}_{\mathrm{GR}}(y_i) = \sum_{j \in s_i} w_j^* y_j, \tag{2.4.7}$$

when the population total of the auxiliary vector $\mathbf{x}$ is known. It follows from (2.4.7) that the GREG estimator also satisfies the additive property: $\hat{Y}_{1\mathrm{GR}} + \ldots + \hat{Y}_{m\mathrm{GR}} = \hat{Y}_{\mathrm{GR}}$. The estimator $\hat{Y}_{i\mathrm{GR}}$ is approximately p-unbiased if the overall sample size is large, but p-consistency requires a large expected domain sample size as well. The special case of a ratio estimator (2.3.13) gives

$$\hat{Y}_{iR} = \frac{\hat{Y}_i}{\hat{X}} X$$

by changing y_j to $a_{ij}y_j$. Similarly, a poststratified estimator is obtained from (2.3.15) by changing y_{gj} to $a_{ij}y_{gj}$:

$$\hat{Y}_{i\mathrm{PS}} = \sum_g \frac{N_{.g}}{\hat{N}_{.g}} \sum_{s_{ig}} w_j y_j,$$

where s_{ig} is the sample falling in the (ig)th cell of the cross-classification of domains and poststrata.

A Taylor linearization variance estimator of $\hat{Y}_{i\mathrm{GR}}$ is simply obtained from $v(y)$ by changing y_j to $e_{ij} = y_{ij} - \mathbf{x}_{ij}^T \hat{\mathbf{B}}(y_i)$, where $\hat{\mathbf{B}}(y_i)$ is obtained from $\hat{\mathbf{B}}(y)$ by changing y_j to y_{ij}. Note that $e_{ij} = -\mathbf{x}_{ij}^T \hat{\mathbf{B}}(y_i)$ if $j \in s$ and $j \notin U_i$. The large negative residuals for all sampled elements not in U_i lead to inefficiency, unlike in the case of $\hat{Y}_{\mathrm{GR}}$ where the variability of the e_j's will be small relative to the variability of the y_j's. This inefficiency of the GREG estimator $\hat{Y}_{i\mathrm{GR}}$ is due to the fact that the auxiliary population information used here is not domain-specific. But $\hat{Y}_{i\mathrm{GR}}$ has the advantage that it is approximately p-unbiased even if the expected domain sample size is small, whereas the GREG estimator based on domain specific auxiliary population information is p-biased unless the expected domain sample size is also large.

2.4.3 Domain-specific Auxiliary Information

We now turn to GREG estimation of a domain total Y_i under domain-specific auxiliary information. We assume that the domain totals $\mathbf{X}_i = (X_{i1}, \ldots, X_{ip})^T = Y(\mathbf{x}_i)$ are known, where $\mathbf{x}_{ij} = \mathbf{x}_j$ if $j \in U_i$ and $\mathbf{x}_{ij} = 0$ otherwise. In this case, a GREG estimator of Y_i is given by

$$Y_{i\mathrm{GR}}^* = \hat{Y}_i + (\mathbf{X}_i - \hat{\mathbf{X}}_i)^T \hat{\mathbf{B}}_i, \tag{2.4.8}$$

where $\hat{\mathbf{X}}_i = \hat{Y}(\mathbf{x}_i)$ and

$$\left(\sum_{j \in s} w_j \mathbf{x}_{ij} \mathbf{x}_{ij}^T / c_j \right) \hat{\mathbf{B}}_i = \sum_{j \in s} w_j \mathbf{x}_{ij} y_{ij} / c_j.$$

We may also write (2.4.8) as

$$Y^*_{i\text{GR}} = \sum_{j \in s} w^*_{ij} y_{ij}, \tag{2.4.9}$$

where $w^*_{ij} = w_j g^*_{ij}$ with

$$g^*_{ij} = 1 + (\mathbf{X}_i - \hat{\mathbf{X}}_i)^T \left(\sum_{j \in s} w_j \mathbf{x}_{ij} \mathbf{x}^T_{ij} / c_j \right)^{-1} \mathbf{x}_{ij} / c_j.$$

Note that the weights w^*_{ij} now depend on i unlike the weights w^*_j. Therefore, the estimators $Y^*_{i\text{GR}}$ do not add up to $\hat{Y}_{\text{GR}}$. Also, $Y^*_{i\text{GR}}$ is not approximately p-unbiased unless the domain sample size is large.

In the special case of a single auxiliary variable x with known domain total X_i, we set $c_j = x_j$ in (2.4.9) to get the ratio estimator

$$Y^*_{iR} = \frac{\hat{Y}_i}{\hat{X}_i} X_i. \tag{2.4.10}$$

If domain-specific poststrata counts N_{ig} are known, then a poststratified count (PS/C) estimator is obtained from the GREG estimator (2.4.9) as

$$Y^*_{i\text{PS/C}} = \sum_g \frac{N_{ig}}{\hat{N}_{ig}} \hat{Y}_{ig}, \tag{2.4.11}$$

where $\hat{N}_{ig} = \Sigma_{s_{ig}} w_j$ and $\hat{Y}_{ig} = \Sigma_{s_{ig}} w_j y_j$. If the cell totals X_{ig} of an auxiliary variable x are known, then we can use a poststratified-ratio (PS/R) estimator

$$Y^*_{i\text{PS/R}} = \sum_g \frac{X_{ig}}{\hat{X}_{ig}} \hat{Y}_{ig}, \tag{2.4.12}$$

where $\hat{X}_{ig} = \Sigma_{s_{ig}} w_j x_j$.

If the expected domain size is large, then a Taylor linearization variance estimator of $Y^*_{i\text{GR}}$ is obtained from $v(y)$ by changing y_j to $e^*_{ij} = y_{ij} - \mathbf{x}^T_{ij} \hat{\mathbf{B}}_i$. Note that $e^*_{ij} = 0$ if $j \in s$ and $j \notin U_i$ unlike the large negative residuals e_{ij} in the case of $\hat{Y}_{i\text{GR}}$. Thus the domain-specific GREG estimator $Y^*_{i\text{GR}}$ will be more efficient than $\hat{Y}_{i\text{GR}}$, provided the expected domain-specific sample size is large.

Example 2.4.1 Wages and Salaries. Särndal and Hidiroglou (1989) and Rao and Choudhry (1995) considered a population U of $N = 1,678$ unincorporated tax filers (units) from the province of Nova Scotia, Canada, divided into 18 census divisions. This population is actually a simple random sample but it was treated as a population for simulation purposes. The population is also classified into four mutually exclusive industry groups: retail (515 units), construction (496 units) accommodation (114 units) and others (553 units).

Domains (small areas) are formed by a cross-classification of the four industry types with 18 census divisions. This leads to 70 nonempty domains out of 72 possible domains. The objective is to estimate the domain totals Y_i of the y-variable (wages and salaries), utilizing the only auxiliary variable x (gross business income) assumed to be known for all the N units in the population. A simple random sample, s, of size n is drawn from U and y-values observed. The sample s_i consist of n_i (≥ 0) units. The data consist of (y_j, x_j) for $j \in s$ and the auxiliary population information.

Under the above set-up, we have $\pi_j = n/N$, $w_j = N/n$ and the expansion estimator (2.4.3) reduces to

$$\hat{Y}_i = \begin{cases} (N/n)\sum_{s_i} y_j & \text{if } n_i \geq 1 \\ 0 & \text{if } n_i = 0. \end{cases} \tag{2.4.13}$$

The estimator $\hat{Y}_i$ is p-unbiased for Y_i unconditionally, but it is p-biased conditional on the realized domain sample size n_i. In fact, conditional on n_i, the sample s_i is a simple random sample of size n_i from U_i, and the conditional p-bias of $\hat{Y}_i$ is

$$B_2(\hat{Y}_i) = E_2(\hat{Y}_i) - Y_i = N\left(\frac{n_i}{n} - \frac{N_i}{N}\right)\bar{Y}_i, \; n_i \geq 1$$

where E_2 denotes conditional expectation. Thus the conditional bias is zero only if the sample proportion n_i/n equals the population proportion N_i/N.

Suppose we form G poststrata based on the x-variable known for all the N units. Then (2.4.11) and (2.4.12) reduce to

$$Y^*_{i\text{PS/C}} = \sum_g N_{ig}\bar{y}_{ig} \tag{2.4.14}$$

and

$$Y^*_{i\text{PS/R}} = \sum_g X_{ig}\frac{\bar{y}_{ig}}{\bar{x}_{ig}}, \tag{2.4.15}$$

where $\bar{y}_{ig}$ and $\bar{x}_{ig}$ are the sample means for the n_{ig} units falling in the cell (ig), and the cell counts N_{ig} and the cell totals X_{ig} are known. If $n_{ig} = 0$, we set $\bar{y}_{ig} = 0$ and $\bar{y}_{ig}/\bar{x}_{ig} = 0$. The estimator $Y^*_{i\text{PS/C}}$ is p-unbiased conditional on the realized sample sizes n_{ig} (≥ 1 for all g), whereas $Y^*_{i\text{PS/R}}$ is only approximately p-unbiased conditional on the n_{ig}'s, provided all the expected sample sizes $E(n_{ig})$ are large.

If poststratification is not used, we can use the ratio estimator (2.4.10):

$$Y^*_{iR} = X_i\frac{\bar{y}_i}{\bar{x}_i}, \tag{2.4.16}$$

where $\bar{y}_i$ and $\bar{x}_i$ are the sample means for the n_i units falling in domain i. If an x-variable is not observed but N_i is known, we can use an alternative estimator

$$Y^*_{iC} = N_i\bar{y}_i. \tag{2.4.17}$$

This estimator is p-unbiased conditional on the realized sample size n_i (≥ 1).

It is desirable to make inferences conditional on the realized sample sizes, but this may not be possible under designs more complex than simple random sampling. Even under simple random sampling, we require $n_{ig} \geq 1$ for all g which limits the use of poststratification when n_i is small.

2.5 Modified Direct Estimators

We now consider modified direct estimators that use y-values from outside the domain but remain p-unbiased or approximately p-unbiased as the overall sample size increases. In particular, we replace $\hat{\mathbf{B}}_i$ in (2.4.8) by the overall regression coefficient $\hat{\mathbf{B}}$, given by (2.3.7), to get

$$\tilde{Y}_{i\mathrm{GR}} = \hat{Y}_i + \left(\mathbf{X}_i - \hat{\mathbf{X}}_i\right)^T \hat{\mathbf{B}} = \sum_{j \in s} \tilde{w}_{ij} y_j \tag{2.5.1}$$

with

$$\tilde{w}_{ij} = w_j a_{ij} + \left(\mathbf{X}_i - \hat{\mathbf{X}}_i\right) \left(\Sigma_s w_j \mathbf{x}_j \mathbf{x}_j^T / c_j\right)^{-1} (w_j \mathbf{x}_j / c_j),$$

where a_{ij} is the domain indicator variable. The estimator $\tilde{Y}_{i\mathrm{GR}}$ is approximately p-unbiased as the overall sample size increases, even if the domain sample size is small. This estimator is also called the "survey regression" estimator (Battese, Harter and Fuller (1988), Woodruff (1966)). The modified estimator (2.5.1) may also be viewed as a calibration estimator $\Sigma_s \tilde{w}_{ij} y_j$ with weights $b_{ij} = \tilde{w}_{ij}$ minimizing a chi-squared distance $\Sigma_s c_j (w_j a_{ij} - b_{ij})^2 / w_j$ subject to the constraints $\Sigma_s b_{ij} \mathbf{x}_j = \mathbf{X}_i$ (Singh and Mian (1995)).

A ratio form of (2.5.1) in the case of a single auxiliary variable x with known domain total X_i is given by

$$\tilde{Y}_{iR} = \hat{Y}_i + \frac{\hat{Y}}{\hat{X}} (X_i - \hat{X}_i). \tag{2.5.2}$$

Although the modified direct estimator borrows strength for estimating the regression coefficient, it does not increase the effective sample size, unlike indirect estimators. To illustrate this, consider the simple random sample of Example 2.4.1. In this case, $\tilde{Y}_{iR}$ reduces to

$$\tilde{Y}_{iR} = N_i \left[\bar{y}_i + \frac{\bar{y}}{\bar{x}} (\bar{X}_i - \bar{x}_i)\right], \tag{2.5.3}$$

where $\bar{y}$ and $\bar{x}$ are the overall sample means and $\bar{X}_i = X_i / N_i$. For large n, we can replace $\bar{y}/\bar{x}$ by $R = \bar{Y}/\bar{X}$ and the conditional variance is

$$V_2(\tilde{Y}_{iR}) \approx N_i^2 \left(\frac{1}{n_i} - \frac{1}{N_i}\right) S_{Ei}^2, \tag{2.5.4}$$

where $S_{Ei}^2 = \sum_{j \in U_i}(E_j - \bar{E}_i)^2/(N_i - 1)$ with $E_j = y_j - Rx_j$ and $\bar{E}_i$ is the domain mean of the E_j's. It follows from (2.5.4) than $V_2(\tilde{Y}_{iR})/N_i^2$ is of order n_i^{-1} so that the effective sample is not increased, although the variability of the E_j's may be smaller than the variability of the y_j's for $j \in U_i$. Note that the variability of the E_j's will be larger than the variability of the domain specific residuals $y_j - R_i x_j$ for $j \in U_i$ unless $R_i = \bar{Y}_i/\bar{X}_i \approx R$.

The modified GREG estimator (2.5.1) may be expressed as

$$\tilde{Y}_{i\mathrm{GR}} = \mathbf{X}_i^T \hat{\mathbf{B}} + \sum_{j \in s_i} w_j e_j. \tag{2.5.5}$$

The first term $\mathbf{X}_i^T \hat{\mathbf{B}}$ is the synthetic-regression estimator (see Chapter 4) and the second term $\sum_{s_i} w_j e_j$ approximately corrects the p-bias of the synthetic estimator. We can improve on $\tilde{Y}_{i\mathrm{GR}}$ by replacing the expansion estimator $\sum_{s_i} w_j e_j$ in (2.5.5) with a ratio estimator (Särndal and Hidiroglou (1989)):

$$\hat{E}_{iR} = N_i(\sum_{s_i} w_j e_j)/(\sum_{s_i} w_j); \tag{2.5.6}$$

note that $\hat{N}_i = \sum_{s_i} w_j$. The resulting estimator

$$\tilde{Y}_{i\mathrm{GR}}(m) = \mathbf{X}_i^T \hat{\mathbf{B}} + \hat{E}_{iR}, \tag{2.5.7}$$

however, suffers from the ratio bias when the domain sample size is small, unlike $\tilde{Y}_{i\mathrm{GR}}$.

A Taylor linearization variance estimator of $\tilde{Y}_{i\mathrm{GR}}$ is obtained from $v(y)$ by changing y_j to $a_{ij}e_j$:

$$v_L(\tilde{Y}_{i\mathrm{GR}}) = v(a_i e). \tag{2.5.8}$$

This variance estimator is valid even when the small area sample size is small, provided the overall sample size is large.

2.6 Design Issues

"Optimal" design of samples for use with direct estimators of large area totals or means has received a lot of attention over the past 60 years or so. In particular, design issues, such as number of strata, construction of strata, sample allocation and selection probabilities, have been addressed (see, e.g., Cochran (1977)). The ideal goal here is to find an "optimal" design that minimizes the MSE of a direct estimator subject to a given cost. This goal is seldom achieved in practice due to operational constraints and other factors. As a result, a "compromise" design that is "close" to the optimal design is adopted.

In practice, it is not possible to anticipate and plan for all possible areas (or domains) and uses of survey data as "the client will always require more than is specified at the design stage" (Fuller (1999), p. 344). As a result, indirect estimators will always be needed in practice, given the growing demand for reliable small area statistics. However, it is important to consider design issues that have an impact on small area estimation, particularly in the context of planning and designing large-scale surveys. In this section, we present a brief discussion on some of the design issues. A proper resolution of these issues could lead to enhancement in the reliability of direct (and also indirect) estimates for both planned and unplanned domains. For a more detailed discussion, we refer the reader to Singh, Gambino and Mantel (1994) and Marker (2001).

(i) *Minimization of clustering*

Most large-scale surveys use clustering to a varying degree in order to reduce the survey costs. Clustering, however, results in a decrease in the "effective" sample size. It can also adversely affect the estimation for unplanned domains because it can lead to situations where some domains become sample rich while others may have no sample at all. It is therefore useful to minimize the clustering in the sample. The choice of sampling frame plays an important role in this respect; for example, the use of a list frame, replacing clusters wherever possible, such as Business Registers for business surveys and Address Registers for household surveys. Also, the choice of sampling units, their sizes and the number of sampling stages have significant impact on the effective sample size.

(ii) *Stratification*

One method of providing better sample size distribution at the small area level is to replace large strata by many small strata from which samples are drawn. By this approach, it may be possible to make an unplanned small domain contain mostly complete strata. For example, each Canadian province is partitioned into both Economic Regions (ERs) and Unemployment Insurance Regions (UIRs), and there are 71 ERs and 61 UIRs in Canada. In this case, the number of strata may be increased by treating all the areas created by the intersections of the two partitions as strata. This strategy will lead to 133 intersections (strata). As another example of stratification, the United States National Health Interview Survey (NHIS) used stratification by region, metropolitan area status, labor force data, income and racial composition until 1994. The resulting sample sizes for individual states did not support state-level direct estimates for several states; in fact, two of the states did not have NHIS sampled units. The NHIS stratification scheme was replaced by state and metropolitan area status in 1995, thus enabling state-level direct estimation for all states; see Marker (2001) for details.

(iii) *Sample allocation*

By adopting compromise sample allocations, it may be possible to satisfy reliability requirements at a small area level as well as large area level, using

only direct estimates. Singh et al. (1994) presented an excellent illustration of compromise sample allocation in the Canadian LFS to satisfy reliability requirements at the provincial level as well as sub-provincial level. For the LFS with a monthly sample of 59,000 households, "optimizing" at the provincial level yields a coefficient of variation (CV) of the direct estimate for "unemployed" as high as 17% for some UIR's (small areas). On the other hand, a two-step compromise allocation with 42,000 households allocated at the first step to get reliable provincial estimates and the remaining 17,000 households allocated at the second step to produce best possible UIR estimates reduced the worst case of 17% CV for UIR to 9.4% at the expense of a small increase at the provincial and national levels: CV for Ontario increased from 2.8% to 3.4% and for Canada from 1.36% to 1.51%. Thus, by oversampling small areas it is possible to decrease the CV of direct estimates for these areas significantly at the expense of a small increase in CV at the national level. The U.S. National Household Survey on Drug Abuse used stratification and oversampling to produce direct estimates for every state. The 2000 Danish Health and Morbidity Survey used two national samples, each of 6,000 respondents, and distributed an additional 8,000 respondents to guarantee at least 1,000 respondents in each county (small area). Similarly, the Canadian Community Health Survey (CCHS) conducted by Statistics Canada accomplishes its sample allocation in two steps. First, it allocates 500 households to each of its 133 Health Regions and then the remaining sample (about one half of 130,000 households) is allocated to maximize the efficiency of provincial estimates; see Béland, Bailie, Catlin and Singh (2000) for details.

(iv) *Integration of surveys*

Harmonizing questions across surveys of the same population leads to increased effective sample sizes for the harmonized items. The increased sample sizes, in turn, lead to improved direct estimates for small areas. However, caution should be exercised because the data may not be comparable across surveys even if the questionnaire wording is consistent. As noted by Groves (1989), different modes of data collection and the placement of questions can cause differences.

A number of current surveys in Europe are harmonized both within countries and between countries. For example, the European Community Household Panel Survey (ECHP) collects consistent data across member countries. Statistics Netherlands uses a common procedure to collect basic information across social surveys.

(v) *Dual frame surveys*

Dual frame surveys can be used to increase the effective sample size in a small area. In a dual frame survey, samples are drawn independently from two overlapping frames that together cover the population of interest. For example, suppose frame A is a complete area frame and data collected by personal interviewing, while frame B is an incomplete list frame and data collected by telephone interviewing. In this case, the dual frame design augments the expensive frame A information with inexpensive additional information

from B. There are many surveys using dual frame designs. For example, the Dutch Housing Demand Survey collects data by personal interviewing, but uses telephone supplementation in over 100 municipalities to produce dual frame estimates for those municipalities (small areas). Statistics Canada's CCHS is another recent example, where an area sample is augmented with a telephone list sample in selected Health Regions.

Hartley (1974) discussed dual frame designs, and developed a unified theory for dual frame estimation of totals by combining information from the two samples. Skinner and Rao (1996) developed dual frame estimators that use the same survey weights for all the variables. Lohr and Rao (2000) applied the jackknife method to obtain variance estimators for dual frame estimators of totals.

(vi) *Repeated surveys*

Many surveys are repeated over time and effective sample size can be increased by combining data from two or more consecutive surveys. For example, the United States National Health Interview Survey (NHIS) is an annual survey that uses nonoverlapping samples across years. Combining consecutive annual NHIS samples leads to improved estimates, although the correlation between years, due to the use of the same psu's, reduces the effective sample size. Such estimates, however, can lead to significant bias if the characteristic of interest is not stable over the time period.

Marker (2001) studied the level of accuracy for state estimates by combining the 1995 NHIS sample with the previous year sample or the previous two years samples. He showed that aggregation helps achieve CV's of 30% and 20% for four selected variables, but 10% CV cannot be achieved for many states even after aggregation across 3 years.

Kish (1999) recommended "rolling samples" as a method of cumulating data over time. Unlike the customary periodic surveys, such as the NHIS with the same psu's over time or the Canadian LFS and the United States Current Population Survey (CPS) with the same psu's and partial overlap of sample elements, rolling samples (RS) aim at a much greater spread to facilitate maximal spatial range for cumulation over time. This, in turn, will lead to improved small area estimates when the periodic samples are cumulated. The American Community Survey (ACS), scheduled to begin in 2003, is an excellent example of RS design. It aims to provide monthly samples of 250,000 households and detailed annual statistics based on 3 million households spread across all counties in the United States. It will also provide quinquennial and decennial census samples later.

2.7 Proofs

2.7.1 Proof of $\hat{Y}_{\mathrm{GR}}(\mathbf{x}) = \mathbf{X}$

We have

$$\begin{aligned}\hat{Y}_{\mathrm{GR}}(\mathbf{x}^T) &= \Sigma_s w_j g_j \mathbf{x}_j^T \\ &= \Sigma_s w_j \left[\mathbf{x}_j^T + (\mathbf{X} - \hat{\mathbf{X}})^T \left(\Sigma_s w_j \mathbf{x}_j \mathbf{x}_j^T / c_j\right)^{-1} \mathbf{x}_j \mathbf{x}_j^T / c_j\right] \\ &= \hat{\mathbf{X}}^T + (\mathbf{X} - \hat{\mathbf{X}})^T = \mathbf{X}^T.\end{aligned}$$

2.7.2 Derivation of Calibration Weights w_j^*

We minimize the chi-squared distance $\Sigma_s c_j (w_j - b_j)^2 / w_j$ with respect to the b_j's subject to the calibration constraints $\Sigma_s b_j \mathbf{x}_j = \mathbf{X}$, that is, minimize $\phi = \Sigma_s c_j (w_j - b_j)^2 / w_j - 2\boldsymbol{\lambda}^T (\Sigma_s b_j \mathbf{x}_j - \mathbf{X})$, where $\boldsymbol{\lambda}$ is the vector of Lagrange multipliers. We get

$$b_j = w_j (1 + \mathbf{x}_j^T \boldsymbol{\lambda} / c_j),$$

where

$$\boldsymbol{\lambda} = \Sigma_s (w_j \mathbf{x}_j \mathbf{x}_j^T / c_j)^{-1} (\mathbf{X} - \hat{\mathbf{X}}).$$

Thus $b_j = w_j^*$, where $w_j^* = w_j g_j(s)$ and $g_j(s)$ is given by (2.3.9).

2.7.3 Proof of $\hat{Y} = \hat{\mathbf{X}}^T \hat{\mathbf{B}}$ when $c_j = \boldsymbol{\nu}^T \mathbf{x}_j$

If $c_j = \boldsymbol{\nu}^T \mathbf{x}_j$, then

$$\begin{aligned}\hat{\mathbf{X}}^T \hat{\mathbf{B}} &= \left(\Sigma_s w_j \mathbf{x}_j^T\right) \hat{\mathbf{B}} \\ &= \boldsymbol{\nu}^T \left(\Sigma_s w_j \mathbf{x}_j \mathbf{x}_j^T / c_j\right) \hat{\mathbf{B}} \\ &= \boldsymbol{\nu}^T \left(\Sigma_s w_j \mathbf{x}_j y_j / c_j\right) \\ &= \Sigma_s w_j y_j = \hat{Y}.\end{aligned}$$

This establishes the result $\Sigma_s w_j e_j(s) = 0$ noted below (2.3.11).

Small Area Estimation. J. N. K. Rao

ISBN: 0-471-41374-7

Chapter 3

Traditional Demographic Methods

3.1 Introduction

Population censuses are usually conducted at 10-year or 5-year intervals to provide population counts for detailed geographical areas of a country as well as for domains (or subpopulations) defined by age, sex, marital status and other demographic variables. Such counts serve a variety of purposes including the calculation of revenue transfers and grants from federal governments to state and local governments. For example, the U.S. census counts are used in the allocation of federal funds to the 50 states and the 39,000 general purpose governmental units.

Information from a census becomes outdated due to changes in the size and composition of the resident population over time. In the absence of population registers maintained over time (as in some Scandinavian countries), it becomes necessary to develop suitable methods of population estimation in the noncensal years, exploiting administrative files that contain valuable demographic information related to population changes. Such postcensal estimates of population are used for a variety of purposes both in the public and private sectors. Some specific uses include the determination of fund allocations, the calculation of social and economic indicators such as vital rates and unemployment rates in which the population count serves as the denominator, and the calculation of survey weights for use in ongoing large-scale sample surveys. The current population trends derived from postcensal population estimates are also used for planning purposes, market research, and making decisions about site locations, advertising and so on. For example, more than 200 health system agencies in the United States use the postcensal estimates to develop health plans and review proposed health programs (National Research Council (1980)).

Traditional demographic methods employ indirect estimators based on im-

plicit linking models. Typically, sampling is not involved in these methods, excepting the sample regression method (subsection 3.3.2).

Sections 3.2–3.4 give a brief account of traditional demographic methods used in developing postcensal estimates for local areas. These methods may be categorized as (a) symptomatic accounting techniques (SAT) and (b) regression symptomatic procedures. The SAT methods include the vital rates (VR) method (Bogue (1950)), the composite method (Bogue and Duncan (1959)), the composite method-II (CM-II) (U.S. Bureau of the Census, 1966), the administrative records (AR) method (Starsinic (1974)), and the housing unit (HU) method (Smith and Lewis (1980)). The regression symptomatic procedures include the ratio correlation method (Schmitt and Crosetti (1954)), the difference correlation method (O'Hare (1976)), and the sample regression method utilizing current survey estimates (Ericksen (1974)). We refer the reader to the following reports/monographs for detailed accounts of the traditional demographic methods: Purcell and Kish (1979), National Research Council (1980), Rives, Serow, Lee and Goldsmith (1989), Statistics Canada (1987) and Zidek (1982). In recent years, demographers have also been using sophisticated model-based methods, as noted in Section 1.3.

Demographic methods make use of census counts in conjunction with demographic information derived from administrative files, but censuses are often subject to omissions, duplications and misclassification. In fact, the issue of adjusting a census for undercount, utilizing estimates of net undercount from a postcensal survey, has received considerable attention in recent years because of the impact of undercount on the allocation of funds. Section 3.4 gives a brief account of the dual-system method of estimating the total population, using the census counts in conjunction with undercount data from a post-enumeration survey.

3.2 Symptomatic Accounting Techniques

Administrative registers contain current data on various demographic variables, changes in which are strongly related to changes in local population. Such variables are called "symptomatic" indicators. For example, the number of births and deaths and net migration during the period since the last census are obvious components of population change. The diverse registration data used in the U.S. include births and deaths as well as school enrollments and number of existing and new housing units. Data on the number of children receiving family allowance and the number of health care recipients are also used in some countries (e.g., Canada).

3.2.1 Vital Rates Method

The vital rates (VR) method uses only birth and death data for the current year t but assumes that an independent estimate of the current population

total, P_t, for a larger area (say a state) containing the local area of interest is available from official sources.

Let (b_t, d_t) and (b_0, d_0) respectively denote the annual number of births and deaths for the local area for the current year t and the last census year $(t = 0)$; b_t and d_t are determined from administrative registers. Further, let (r_{1t}, r_{2t}) and (r_{10}, r_{20}) respectively denote the crude birth and death rates for the local area for the current year and the last census. Note that $r_{1t} = b_t/p_t$ and $r_{2t} = d_t/p_t$, where p_t is the current population for the local area; r_{10} and r_{20} are similarly defined.

If r_{1t} and r_{2t} were known, then both b_t/r_{1t} and d_t/r_{2t} would yield the current population of the local area, p_t. This suggests the estimation of r_{1t} and r_{2t}, utilizing the census values r_{10}, r_{20} and the independent estimate, $\hat{P}_t$, of P_t for the larger area. The VR method essentially finds updating factors ϕ_1 and ϕ_2 such that $r_{1t} = \phi_1 r_{10}$ and $r_{2t} = \phi_2 r_{20}$. Estimates of ϕ_1 and ϕ_2 are obtained by assuming that the same factors also apply for the larger area, that is $R_{1t} = \phi_1 R_{10}$ and $R_{2t} = \phi_2 R_{20}$, where $R_{1a} = B_a/P_a, R_{2a} = D_a/P_a$ are the rates for the larger area and (B_a, D_a) denote the number of births and deaths in the larger area for time $a(= 0, t)$. Under this assumption, using the known census rates R_{10} and R_{20} we obtain estimates of ϕ_1 and ϕ_2 as

$$\hat{\phi_1} = \hat{R_{1t}}/R_{10} \qquad \text{and} \qquad \hat{\phi_2} = \hat{R_{2t}}/R_{20},$$

where $\hat{R_{1t}} = B_t/\hat{P}_t$ and $R_{2t} = D_t/\hat{P}_t$. The estimate of current population, p_t, is then obtained as

$$\hat{p}_t = \frac{1}{2}\left(\frac{b_t}{\hat{r_{1t}}} + \frac{d_t}{\hat{r_{2t}}}\right), \tag{3.2.1}$$

where $\hat{r_{1t}} = \hat{\phi_1} r_{10}$ and $\hat{r_{2t}} = \hat{\phi_2} r_{20}$.

The success of the VR method depends heavily on the validity of the implicit model that the updating factors, ϕ_1 and ϕ_2, for the local area remain valid for the larger area containing the local area. Such a strong assumption is often questionable in practice.

Example 3.2.1. County Population. Govindarajulu (1999, Chapter 17) considered the estimation of the population of a small county in Kentucky. Here $b_t = 400, d_t = 350$ for the current year t and from the 1990 census $R_{10} = 2\%$ and $R_{20} = 1.8\%$. It was assumed that $R_{10} = r_{10}$ and $R_{20} = r_{20}$. The current state rates are $\hat{R}_{1t} = 2.1\%$ and $\hat{R}_{2t} = 1.9\%$ so that $\hat{\phi}_1 = 2.1/2$ and $\hat{\phi}_2 = 1.9/1.8$. Also,

$$\hat{r}_{1t} = (2.1/2)2 = 2.1\%, \hat{r}_{2t} = (1.9/1.8)(1.8) = 1.9\%.$$

It now follows from (3.2.1) that

$$\hat{p}_t = \frac{1}{2}\left(\frac{400}{0.021} + \frac{350}{0.019}\right) \approx 18{,}735.$$

3.2.2 Composite Method

The composite method is a refinement on the VR method. It uses the VR method to compute group-specific population estimates separately and then sums these estimates across the groups to obtain a "composite" estimate of the current population, p_t. This method requires group-specific birth and death counts for the local area as well as current population in each group for the larger area. Alternative data sources may be used for specific groups. For example, the school age population, 5–14, in the United States is estimated by using school enrollment and school enrollment rates (see Zidek (1982)).

3.2.3 Component Methods

Component methods derive current population estimates by taking census values, adding births, subtracting deaths, and adding an estimate of net migration (which can be negative). Let $b_{0,t}$, $d_{0,t}$ and $m_{0,t}$ respectively denote the numbers of births and deaths and net migration in the local area during the period $[0, t]$. Net migration $m_{0,t}$ is the sum of immigration, $i_{0,t}$, and net interarea migration, $n_{0,t}$, minus emigration $e_{0,t}$. The current population, p_t, may be expressed as

$$p_t = p_0 + b_{0,t} - d_{0,t} + m_{0,t},$$

where p_0 is the baseline census population. Registration of births and deaths are usually complete in the United States and Canada, but net migration figures are not directly available. In the United States, net migration is divided into military and civilian migration since the former is obtainable from administrative records. For estimating civilian migration, the component method-II (CM-II) uses school enrollments, whereas the Administrative Records (AR) method employs income tax returns (see Zidek (1982)). An earlier version of CM-II is called CM-I. In Canada, emigration and interarea migration are inferred from personal income tax files for census divisions (or areas). Immigration statistics at the province level are obtained from official sources, but such official statistics are not available for all census divisions. As a result, immigration figures at the census division level are also estimated from income tax returns but benchmarked to the official figures for the provinces.

Post-censal population estimates by age, sex and marital status are produced in Canada using the cohort component method which is similar to the component method but uses certain modifications because of the nature of disaggregation (Statistics Canada (1987), Chapter II).

3.2.4 Housing Unit Method

Housing units and group quarters (e.g., college dormitories, prisons, nursing homes) may be distinguished as places of residence. Accordingly, the current population, p_t, may be expressed as

$$p_t = (h_t)(pph_t) + gq_t,$$

where h_t is the number of occupied housing units at time t, pph_t is the average number of persons per housing unit at time t and gq_t is the number of persons in group quarters at time t. The quantities h_t, pph_t, and gq_t all need to be estimated. In particular, h_t is estimated from the change in the number of housing units, $hu_t - hu_0$, which, in turn, is given by

$$hu_t - hu_0 = bp_t - dl_t.$$

Here bp_t is the number of building permits issued during $[0, t]$ adjusted for completion times and dl_t is the number of demolitions during $[0, t]$. It now follows that h_t may be obtained as

$$h_t = h_0 + (bp_t - dl_t)(ocr),$$

where ocr denotes the occupancy rate which may be estimated by, ocr_0, the value at the time of the last census. The quantity pph_t is estimated by the value pph_0 at the time of the last census or estimated by extrapolating pph values from the previous two censuses. Finally, gq_t is estimated by the value gq_0 at the time of the last census or determined by extrapolation.

Various refinements and modifications of the housing unit (HU) method have been proposed. For example, h_t may be obtained as $h_0 + rec_t - rec_0$, where rec_α is the number of active residential electrical units at time α, or as $(h_0/rec_0)(rec_t)$. A more accurate estimate of h_t may be obtained by stratifying the housing units according to type (single family, mobile homes, etc.) and then applying the HU method separately in each stratum. The resulting strata estimators are summed to get an estimate of h_t.

The HU method has been used in the United States to estimate county and sub-county level populations. It is an attractive method but getting the relevant data is not easy.

3.3 Regression Symptomatic Procedures

Regression symptomatic procedures use multiple linear regression to estimate local area populations, utilizing symptomatic variables as independent variables in the regression equation. Two such procedures are the ratio correlation and the difference correlation methods.

3.3.1 Ratio Correlation and Difference Correlation Methods

Let $0, 1$ and $t(> 1)$ denote two consecutive census years and the current year, respectively. Also, let p_{ia} and s_{ija} be the population size and the value of the jth symptomatic variable $(j = 1, \ldots, p)$ for the ith local area $(i = 1, \ldots, m)$ in the year a $(= 0, 1, t)$. Also, let p_{ia}/P_a and s_{ija}/S_{ja} be the corresponding proportions, where $P_a = \sum_i p_{ia}$ and $S_{ja} = \sum_i s_{ija}$ are the values for the larger area (state or province).