SURVEY METHODOLOGY

Catalogue No. 12-001-XPB

A JOURNAL PUBLISHED BY STATISTICS CANADA

JUNE 2004

•

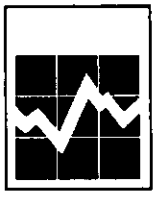
VOLUME 30

•

NUMBER 1

Statistics Canada Statistique Canada

Canada



SURVEY METHODOLOGY

A JOURNAL PUBLISHED BY STATISTICS CANADA

JUNE 2004 • VOLUME 30 • NUMBER 1

Published by authority of the Minister
responsible for Statistics Canada

July 2004

Catalogue no. 12-001-XPB

Frequency: Semi-annual

ISSN 0714-0045

Ottawa

Statistics Canada Statistique Canada

Canada

SURVEY METHODOLOGY

A Journal Published by Statistics Canada

Survey Methodology is abstracted in The Survey Statistician, Statistical Theory and Methods Abstracts and SRM Database of Social Research Methodology, Erasmus University and is referenced in the Current Index to Statistics, and Journal Contents in Qualitative Methods.

EDITORIAL POLICY

Survey Methodology publishes articles dealing with various aspects of statistical development relevant to a statistical agency, such as design issues in the context of practical constraints, use of different data sources and collection techniques, total survey error, survey evaluation, research in survey methodology, time series analysis, seasonal adjustment, demographic studies, data integration, estimation and data analysis methods, and general survey systems development. The emphasis is placed on the development and evaluation of specific methodologies as applied to data collection or the data themselves. All papers will be refereed. However, the authors retain full responsibility for the contents of their papers and opinions expressed are not necessarily those of the Editorial Board or of Statistics Canada.

Submission of Manuscripts

Survey Methodology is published twice a year. Authors are invited to submit their articles in English or French in electronic form, preferably in Word to the Editor, Dr. M.P. Singh, singhmp@statcan.ca (Household Survey Methods Division, Statistics Canada, Tunney's Pasture, Ottawa, Ontario, Canada, K1A 0T6). For formatting instructions, please see the guidelines provided in the Journal.

Subscription Rates

The price of *Survey Methodology* (Catalogue no. 12-001-XPB) is CDN $58 per year. The price excludes Canadian sales taxes. Additional shipping charges apply for delivery outside Canada: United States, CDN $12 ($6 × 2 issues); Other Countries, CDN $30 ($15 × 2 issues). Subscription order should be sent to Statistics Canada, Dissemination Division, Circulation Management, 120 Parkdale Avenue, Ottawa, Ontario, Canada, K1A 0T6 or by dialling 1 800 700-1033, by fax 1 800 889-9734 or by E-mail: order@statcan.ca. A reduced price is available to members of the American Statistical Association, the International Association of Survey Statisticians, the American Association for Public Opinion Research, the Statistical Society of Canada and l'Association des statisticiennes et statisticiens du Québec.

SURVEY METHODOLOGY

A journal Published by Statistics Canada

Volume 30, Number 1, June 2004

CONTENTS

In This Issue 1

Waksberg Invited Paper Series

NORMAN M. BRADBURN
Understanding the Question-Answer Process 5

Discussion Paper

ABDELLATIF DEMNATI and J.N.K. RAO
Linearization Variance Estimators for Survey Data 17

Comment:
PHILLIP S. KOTT 27
BABUBHAI V. SHAH 29
CHRIS SKINNER 30
Response from the authors 32

Regular Papers

CARY T. ISAKI, JULIE H. TSAY and WAYNE A. FULLER
Weighting Sample Data Subject to Independent Controls 35

D. NASCIMENTO DA SILVA and JEAN D. OPSOMER
Properties of the Weighting Cell Estimator Under a Nonparametric Response Mechanism 45

J. MICHAEL BRICK, GRAHAM KALTON and JAE KWANG KIM
Variance Estimation with Hot Deck Imputation Using a Model 57

MICHAEL A. HIDIROGLOU and ZDENEK PATAK
Domain Estimation Using Linear Regression 67

MICHAIL SVERCHKOV and DANNY PFEFFERMANN
Prediction of Finite Population Totals Based on the Sample Distribution 79

LEONARDO GRILLI and MONICA PRATESI
Weighted Estimation in Multilevel Ordinal and Binary Models in the Presence of Informative Sampling Designs 93

GEOFF ROWE and HUAN NGUYEN
Longitudinal Analysis of Labour Force Survey Data 105

MARC CALLENS and CHRISTOPHE CROUX
Contact and Cooperation in the Belgian Fertility and Family Survey 115

Survey Methodology, June 2004
Vol. 30, No. 1, pp. 1-2
Statistics Canada

In This Issue

This issue of Survey Methodology contains the fourth in the annual invited paper series in honour of Joseph Waksberg. A brief description of the series and a short biography of Joseph Waksberg were given in the June 2001 issue of the journal. I would like to thank the members of the awards selection committee for having selected Norman Bradburn as the author of this year's Waksberg invited paper.

In his paper entitled "Understanding the Question-Answer Process", Bradburn traces the history of conceptualization of the survey process over the past couple of decades, in which concepts from social and cognitive psychology and linguistics have been applied to improving our understanding of this process, and cognitive tools and approaches have been adapted for use in formulating survey instruments. He presents a conceptual model for the survey interview, and discusses various cognitive processes in survey response such as comprehension, retrieval, answer formulation and response. In his concluding summary he outlines challenges and priorities for further research in this area.

In Demnati and Rao, the authors present an approach for obtaining Taylor linearization variance estimators that is easier to apply than the usual Taylor linearization approach. The new method leads to a unique variance estimator and is applicable in many situations and estimators. The method is illustrated for calibration estimators, estimating equations and under two-phase sampling. For calibration estimators, the calibration weight is automatically captured in the variance formulae thus justifying what is commonly done in practice. Discussions of this paper are provided by Phil Kott, Babubhai Shah, and Chris Skinner.

Isaki, Tsay and Fuller propose a new method of household weighting for the 2000 U.S. Census long form, using quadratic programming to ensure that the weighted sums of household and individual characteristics match control totals derived either from the Census short form or from the Accuracy and Coverage Evaluation (A.C.E.) study. The weights are then rounded to integer values. They propose a jackknife procedure for estimation of the variance that incorporates the effects of both rounding and the random controls from A.C.E. Results of the proposed weighting procedures are compared to the 1990 weighting procedures using the 1990 Census data.

The theoretical properties of the estimator through reweighting within cells are studied in the article by da Silva and Opsomer. In contrast with numerous other studies on the subject, which involve a response model in which the population units are homogeneous within cells, it is not necessary to correctly specify the response model. It is necessary, however, to determine an auxiliary variable that is correlated with the response probability. The proposed approach can thus be seen as non-parametric. A simulation study explores the properties of the estimator being considered under various scenarios. The authors also provide some recommendations on the size and number of reweighting cells.

Brick, Kalton and Kim deal with the estimation of variance in the presence of hot-deck imputation within imputation cells for linear estimators. Särndal's decomposition (1992) and a model for the variable of interest are used to estimate variance. The originality of the proposed approach comes from the fact that, not only are the sampled and responding units conditioned, but also the units selected at the time of imputation. The article also deals with estimation for domains and a simulation study is carried out to evaluate the proposed method when certain model assumptions do not hold.

Hidiroglou and Patak study the properties of a number of small area estimators. They classify the estimators into two types, Horvitz-Thompson and Hájek, and by the detail of auxiliary information required. Conditional and unconditional properties of the estimators are investigated both analytically and in a simulation study. They conclude that the Hájek-type estimators have the best conditional properties, both in terms of bias and coverage, but these estimators do not have the additive property and their weights are domain dependent.

In their paper, Sverchkov and Pfeffermann develop prediction of finite population totals using a model for a variable of interest conditional on the unit not being in the sample (the sample-complement distribution) and possibly some covariates. They first describe the sample distribution and the sample-complement distribution, and then develop semi-parametric estimation of the sample complement model. A resampling procedure is proposed for mean-square error estimation. The method is illustrated by examples and it is compared to alternative approaches in a simulation study.

The article by Grilli and Pratesi considers the problem of parametric estimation for ordinal and binary models at a number of levels for informational sample plans. The authors extend the pseudo maximum likelihood method to deal with this problem. This method uses the inverse of the inclusion probabilities at each degree to weight the logarithm of the likelihood function. The estimator's properties thereby obtained are tested in a simulation study. The bootstrap method is also used to obtain a variance estimator.

Rowe and Nguyen explore longitudinal analysis using data from an overlapping panel survey, specifically, the Canadian Labour Force Survey. Successive six-month longitudinal panels can be used to provide estimates relating to cohorts of people over time, provided that cohort members can be identified in each panel. They develop a likelihood function for the longitudinal data observed in each six-month window, and show how this can be used to obtain estimates of parameters of interest. They then give an illustration of this approach for estimating transition probabilities between employment states and validate it by comparing simulated and observed data.

Finally, in a paper somewhat related to Bradburn's, Callens and Croux look at individual level and municipality level predictors of contact and cooperation in the Belgian Fertilily and Family Survey using multilevel logistic regression models. They discuss some social theory models for contact and cooperation that imply an important role for different indicators, and then fit models using data from the survey. Their qualitative findings, in particular with respect to socio-economic status (SES) indicators, seem to conflict with the results of similar studies in the literature. In this study, SES was found to be positively related to cooperation. Some possible explanations of the observed results are offered.

M.P. Singh

Survey Methodology, June 2004
Vol. 30, No. 1, pp. 3-4
Statistics Canada

Waksberg Invited Paper Series

Survey Methodology has established an annual invited paper series in honor of Joseph Waksberg, who has made many important contributions to survey methodology. Each year, a prominent survey researcher will be chosen to author a paper that will review the development and current state of a significant topic in the field of survey methodology. The author receives a cash award, made possible through a grant from Westat in recognition of Joe Waksberg's contributions during his many years of association with Westat. The grant is administered financially and managed by the *American Statistical Association.* The author of the paper is selected by a four-person committee appointed by *Survey Methodology* and the *American Statistical Association.*

The author of the Waksberg paper is announced at the annual Joint Statistical Meeting during the American Statistical Association Presidential Address and Awards session. In this session, recipients of awards such as Section, Chapter, Continuing Education-Excellence and other co-sponsored awards are congratulated. In particular, the Waksberg Award for outstanding contributions in the theory and practice of survey methodology is highlighted. Finally, the winner of the Waksberg award appears in the Awards program booklet.

Previous Waksberg Award Winners:

Gad Nathan (2001)
Wayne A. Fuller (2002)
Tim Holt (2003)

Nominations:

Nominations of individuals to be considered as authors or suggestions for topics should be sent by December 3, 2004 to the chair of the committee, David Bellhouse by e-mail at: bellhouse@stats.uwo.ca or by fax (519) 661-3813.

2004 WAKSBERG INVITED PAPER

Author: Norman M. Bradburn

Norman Bradburn is the Tiffany and Margaret Blake Distinguished Service Professor Emeritus in the University of Chicago. He has spent most of his career as a survey methodologist at the National Opinion Research Center (NORC) at the University of Chicago where he is currently a Senior Fellow. His research has concentrated on the study of non-sampling errors in surveys with particular emphasis on the cognitive aspects of the survey question/answer process.

MEMBERS OF THE WASKBERG PAPER SELECTION COMMITTEE (2004-2005)

David R. Bellhouse, (Chair), *University of Western, Ontario*
Gordon Brackstone, *Statistics Canada, Ontario*
Wayne Fuller, *Iowa State University*
Sharon Lohr, *Arizona State University*

Past Chairs:

Graham Kalton (1999 - 2001)
Chris Skinner (2001 - 2002)
David A. Binder (2002 - 2003)
J. Michael Brick (2003 - 2004)

Survey Methodology, June 2004
Vol. 30, No. 1, pp. 5-15
Statistics Canada

Understanding the Question-Answer Process

NORMAN M. BRADBURN [1]

ABSTRACT

Survey statisticians have long known that the question-answer process is a source of response effects that contribute to non-random measurement error. In the past two decades there has been substantial progress toward understanding these sources of error by applying concepts from social and cognitive psychology to the study of the question-answer process. This essay reviews the development of these approaches, discusses the present state of our knowledge, and suggests some research priorities for the future.

KEY WORDS: Measurement errors; Response effects; Cognitive psychology; Questionnaire design.

1. INTRODUCTION

When I was in graduate school, I was deeply impressed by Gordon Allport's comment to the effect that the best way to find out something was to ask a direct question. Later, as I began to study and do research on methodological problems in sample surveys of human populations, I became more convinced of the wisdom on this remark. I have even formulated it into Bradburn's Law for Questionnaires: "Ask what you want to know, not something else."

The trouble with this law is that it is extremely difficult to put into practice for several reasons. First, it presumes that we know what we want to know. Often when we start out to construct a questionnaire, we are not sure what we want to know and use the questionnaire construction process in an iterative fashion to refine our ideas about what we want to know. Until we have a clear understanding of what we are trying to ask about, there is little hope that we will be able to ask meaningful questions.

Second, even if we know what we want to know, we need to understand how people answer questions. The complexities of human communication make it difficult to construct of single, standardized instrument that will enable us to ask our questions so that respondents will understand them in the way that we intend and that we will understand their answers in the way they intend. Belson (1968), who has done extensive studies on the comprehension of questions by respondents, estimates that even with the best-constructed questionnaires, less than half of the sample will understand the questions the way the researcher intended. He does not present any data on how well the researchers understand the responses.

Even if this estimate is too pessimistic, we are faced with a difficult problem of measurement error that comes from the question-answer process itself, rather than from sample design or survey execution. The existence of this source of measurement error has been recognized since the beginning of scientific surveys, that is, since the development of sampling theory and its application to human populations. Unlike sampling theory, which rests on firm mathematical principles, the understanding of measurement error due to the question-answering process has not, until recently, been based on the theoretical understanding of human communication and cognition. This situation is beginning to change.

In the past two decades there has been substantial progress in the conceptualization of the survey interview applying concepts from social and cognitive psychology (Jabine, Straf, Tanur and Tourangeau 1984, Sudman and Bradburn 1974, Sudman, Bradburn and Schwarz 1996, Tourangeau, Rips and Rasinski 2000). In this essay I will review briefly the development of these approaches, discuss the present state of our knowledge regarding the question-answer process, and suggest some research priorities for the future.

Some History

The collaboration between cognitively oriented psychologists and survey researchers began about 25 years ago. Like many innovations it had many progenitors and seemed to spring up from several independent sources. One of the earliest, if not the earliest instance, was a seminar held in 1978 by the British Social Science Research Council and the Royal Statistical Society on problems in the collection and interpretation of recall data in social surveys. Particularly noteworthy was the participation of the Cambridge cognitive psychologist Alan Baddeley whose paper, "The Limitations of Human Memory: Implications for the Design of Retrospective Surveys," is perhaps the first paper by a psychologist interested in memory directly related to survey design (Baddeley 1979).

[1] Norman M. Bradburn, National Opinion Research Center, University of Chicago.

Two important events occurred in the United States in 1980. The first was a workshop convened by the Bureau of Social Science Research in connection with its work in the redesign of the National Crime Victimization Survey. This workshop brought together cognitive scientists and survey statisticians and methodologists to discuss what contributions cognitive scientists could make to under standing response errors in behavioral reports (Biderman 1980). One of the results of this conference was to stimulate some of the cognitive psychologists who participated to begin to study problems in survey questions in a laboratory setting. One of the earliest of such papers was "Since the eruption of Mt. St. Helens has anyone beaten you up? Improving the accuracy of retrospective reports with landmark events, " (Loftus and Marburger 1985) which demonstrated experimentally the value of using landmark events to improve the quality of dating events in survey reports.

The second event was the establishment of a panel on the measurement of subjective phenomena by the Committee on National Statistics. This panel produced two large volumes that reviewed a considerable amount of research on response effects involved in the measurement of subjective phenomena. It complemented the work that had been done by the earlier seminars on measuring behavior or more "objective" phenomena. (Turner and Martin 1982)

A big stimulus came in 1983 when the Committee on National Statistics with funding from NSF organized a 6-day seminar in St. Michaels, Maryland on Cognitive Aspects of Survey Methodology. Two papers, "Potential contributions of cognitive research to survey questionnaire design" (Bradburn and Danis 1984) and "Cognitive science and survey methods," (Tourangeau 1984) reviewed how new developments in cognitive psychology could contribute to survey methodology and how developments in survey methodology could contribute to the further development of cognitive psychology. The conference was extraordinarily fruitful and led to a whole new field of research in survey methodology both as applied to objective and subjective phenomena. The results of this conference were published in Jabine *et al.* (1984).

The final instance of independent work that may be thought of a progenitor of this field was a conference organized by Norbert Schwarz and his associates in Germany. Perhaps the most influential paper from this conference was the model proposed by Strack and Martin (1987) "Thinking, judging and communicating: A process account of context effects in attitude surveys." The results of the conference are published in Hippler, Schwarz and Sudman, *Social Information processing and survey methodology* (1987).

In the ensuing years, there has been a stream of research that has refined and elaborated the research agenda that came from these early seminars. Some of the work sponsored by the Social Science Research Council is published in "Questions about questions: Inquiries into the cognitive bases of surveys" (Tanur 1992). Subsequent research has been updated in a series of volumes edited by Schwarz and Sudman (1992, 1994, 1996).

A Conceptual Approach to the Survey Interview

A survey interview is a structured social interaction between two people who play distinctive roles-the interviewer and the respondent. It has been described as a "conversation with a purpose" (Bingham and Moore 1934). The purpose, to put it succinctly, is to get a series of questions answered. In scientific surveys, these questions are usually embodied in a structured questionnaire designed by a third party, the researcher. It is this type of survey activity that I will consider, although the analysis could be extended to other, less structured interviews.

Like all social interactions among people from the same culture, there are implicit rules that influence the way the participants behave. Some of these are general and apply to all social interactions between social equals; some are general to the peculiar type of interaction we call the survey interview; some are general to this survey; and some are idiosyncratic and apply to only this particular interview. Thus we think of these rules as hierarchically organized from the most general, which apply to all survey interviews, to the particular rules involved in a particular interview.

At the most general level the interaction is governed by the rules for voluntary interactions between strangers. The interaction is initiated by one party, the interviewer, who must establish the nature of the encounter. The important elements that must be established are: 1) that it is non-threatening, that is the interviewer is not going to do any harm to the respondents; 2) the purpose of the encounter, and 3) what are the costs and benefits to the respondents if they agree to participate in the interview. The interaction is thus viewed as neutral, purposive, and worthwhile. As with any structured social interaction, it is governed by the norms related to such interactions.

What are the norms that are important for the interview? The first is mutual respect for individuals, particularly the privacy of the respondents. This principle has become an important issue regarding the protection of research participants because of a number of instances in bio-medical research where the voluntary nature of participation was not made clear. For high-risk research written consent to participate is now required. In the survey interview, however, the context of the request for an interview makes it easy for respondents to refuse if they do not wish to participate and written consent is superfluous. Asking for written consent may actually raise suspicion that the

interviewer has not been truthful about the purpose of the interview because written consent is not normally part of a conversation between strangers who have established that the interaction is non-threatening.

A second important norm is truthfulness. It is part of the role obligation of both parties to be truthful. For the interviewer, this means telling the respondent pertinent facts about the purpose of the interview, what is required of the respondents, *e.g.*, how much time it will take, whether they will need to consult records, whether the questions may be sensitive, *etc.* and to answer any questions the respondents might ask. If providing some information at the beginning of the interview might bias responses, such as who the sponsor of the research is, the information can be given at the end of the interview.

The purpose of the interview is to obtain the information required by the research. The interviewer's role is to get the desired information and the questionnaire is the principal instrument for accomplishing this task. A well-designed questionnaire makes the interviewer's job easier and minimizes the need for the interviewer to have to answer questions about the meaning of questions in the questionnaire. While interviewers need to be trained about the purpose of questions and their meaning, interviewers may become a source of uncontrolled variance if they have to interpret questions for many respondents. Interviewers need to be alert to cues that respondents are misunderstanding questions and to act to correct them. The need for many interventions by interviewers indicates a bad questionnaire.

If respondents accept the role and agree to participate in the interview, they have the obligation, under the norm of truthfulness, to answer the questions as accurately and completely as possible. This norm, however, may conflict with the general desire of individuals to be well thought of and to present themselves in a favorable light. In many surveys, we ask questions about potentially embarrassing, sensitive or even illegal behavior or unpopular attitudes. The interviewer and the questionnaire both play an important role in minimizing this conflict and reinforce the norm of truthfulness. The empirical evidence, however, suggests that even with the best trained interviewers and the best techniques of questionnaire design, it is rarely possible to prevent some overreporting of socially desirable behavior and attitudes or underreporting of undesirable attitudes and behavior (See Bradburn, Sudman and Associates 1979; Wentland and Smith 1993).

Survey data are collected under a strong norm of confidentiality. The norm is so strong that even if it is not made explicit, respondents expect that information from interviews that have the form of scientific surveys, such as public opinion polls or employee attitude surveys, will not be identified with them. Violations of this norm such as occur with "sugging" (selling under the guise of a survey) or "frugging" (fund raising under the guise of a survey) threaten to erode public confidence in surveys and contribute to the increase in rates of refusal to participate. Unless the data are collected under "shield laws" or certificates of confidentiality that have the force of law, confidentiality promises, however, can be compromised by law enforcement activities.

Linguists have also noted that there are basic shared assumptions underlying conversations that facilitate the interactions. These have been systematically described by Grice and are referred to as Grician rules (Grice 1975, see also Sudman *et al.* 1996 for their application in surveys). According to Grice, conversations are based on a principle of "cooperativeness" which is embodied in four maxims. The maxim of quality enjoins speakers to be truthful and not to say things that they lack evidence for. The maxim of relation indicates that the utterances are relevant to the topic of the ongoing conversation. The maxim of quantity requires that speakers not repeat themselves and make the contributions to the conversation as informative as possible. The maxim of manner requires that the speakers be as clear as possible in their meaning. Thus, according to Grice, speakers are expected to be truthful, relevant, informative and clear.

These maxims apply equally to informal conversations and to interviews that have the form of a special type of conversation. Thus the questions asked by the interviewer are interpreted within the same framework, that is both questions and introductory material to questions are relevant to the topic, are supposed to be informative and clear. Violations of these maxims can lead to confusion on the part of respondents and produce response effects that are well documented. For example, violations of the maxim of relevance when questions are obscure (see for example, Schuman and Presser 1981) or deliberately about fictitious issues (Bishop, Oldendick and Tuchfarber 1986) lead to respondents trying to make sense of the question by looking to contextual cues about the meaning of the question. This produces what appears to be an erroneous response when viewed from the perspective of the researcher who does not understand the conversational assumptions of the respondents.

One of the most well documented order effects in surveys occurs when questions of differing levels of specificity occur together. When one question is general, *e.g.*, "Taking all things together, how happy are you these days? " and the other is specific, *e.g.*, "How happy is your marriage?", responses to the general question are affected by the order of the questions, while responses to the more specific question are not. The effect appears to be the result of the workings of the maxim of relevance. When the

general question comes first it is interpreted as intended, that is, respondents should include all aspects of their lives in making the judgment about their happiness. When the general question comes second after the specific question about marriage happiness, the maxim of relevance suggests that respondents should exclude from consideration their marriages because they have already reported on them. Thus, even though the question literally asks about "all things together", it is interpreted to mean "all things except those we have already asked about." It is only those things that have not been asked about that are still relevant.

What happens if the norms outlined above are not accepted in the interview either because the respondent rejects or redefines the role of respondent or does not observe the maxims of conversation? Of course the easiest form of rejection of the role of respondent is to refuse the interview altogether. Sometimes, however, a person sampled becomes a "reluctant respondent", that is, they are may feel pressured to participate in the study because of follow-up procedures, because they do not like to refuse a strong request from another person or for some other reason. In such cases they may care less about being a good respondent than just getting the interview finished. Thus they may take less time to think about questions, make less effort to recall information requested, or be less interested in a truthful answer than a "don't know" or even a false answer. Interviewers have told me that they often feel that the responses given by those that they have convinced to participate in an interview after many attempts at refusal conversion are less valid that those who participate more willingly. Extras efforts to obtain high completion rates may in fact produce less good data.

Respondents also may misunderstand the nature of the survey interview, simply want to convert it into a social conversation, or not be very skilled conversationalists, that is not abide by the Grician maxims and thus engage in an "inefficient" conversation. Such conversations are characterized by frequent asides or changes of topic, comments on topics of little or no relevance to the question at hand, relating personal anecdotes that may be triggered by some aspect of the question, or simple repetition of comments. In such cases the interviewer must politely but firmly teach the respondent the rules for the conversation and guide the respondent to keep focused on the questions in the interview. Skilled interviewers become experts in steering the conversation and, by selective reinforcement, shaping the respondents' behavior to follow the Grician maxims.

In summary, interviews take place in social contexts that have a structure governed by socially shared expectations and norms. These norms may differ from society to society and perhaps even within subcultures in the same society, but they have powerful effects on the way interviews are conducted and the way questions are interpreted. Violations of the expectations or norms may lead to "effects" that may be interpreted as error from the perspective of the researcher. If these norms and expectations are understood, they can be used to avoid problems or to mitigate the effects.

Data could also be obtained from interviewers about how much the interview deviated from the model outlined above. Although little research has been done assessing the quality of interviews from this point of view, a fruitful area for future research could be to investigate the decline in validity of data as the conditions of the interview increasingly deviate from the ideal model.

Cognitive Processes in Survey Response

Answering questions in a survey involves considerable cognitive work on the part of respondents. Much of what underlies recent advances in understanding survey response processes derives from the application of models of information processing to the question-answering process. While there is still much work to be done before we have complete and detailed understanding of how the brain processes information, there is sufficient agreement about the general approach to serve as the basis for a better understanding of the response process.

The mind is conceptualized as a large information processing system composed of a series of component systems. The physical sensations of sound and sight enter the system in the sensory register. The sensory register has capacity limitations so that only a portion of the information is transferred to short-term memory. Attention plays a large role in determining what is brought into short-term memory. Attention is a function of an executive monitor that enables and controls the information processing system much the way that programs enable what computers do. The executive system controls the entire system through goals and plans that are organized into priorities for action.

The storehouse of the system is the long-term memory system that has a very large capacity. Working memory refers to the system in which active thinking takes place. The activity here draws on short-term memory and retrievals from long-term memory. Short-term memory has limited capacity but rapid access, while long-term memory has large capacity but is relatively slow in access. Long-term memory appears to have two rather distinct subsystems, semantic memory and episodic memory, although this distinction is not universally agreed upon. Semantic memory refers to memory associated with vocabulary, language structure, rules and abstract knowledge, while episodic memory refers to memory for events that took place in time and space.

Information is represented as a list of features or concepts that are linked together in networks. Information is stored in memory in structures that are hierarchically organized with more general concepts being higher in the structure than more discrete instances of the concept or distinct features. The term "schema" is sometimes used to refer to larger, more complex shared and/or overlearned structures that organize our thoughts on familiar topics and may be retrieved as a whole rather than as individual parts.

Language is the medium through which information is primarily communicated and thus information, to be available for communication, must be associated with a linguistic code. The exact relationship between language and thought and whether or not all thoughts have verbal representation are still subjects of debate. It is clear, however, that meaning is encoded somehow in language and these codes play an important role in the acquisition, storage and retrieval of information. Emotion may also be part of the code, although its role is not well understood.

Knowledge structures facilitate and constrain patterns of activation in the mind. What comes to mind, that is, into consciousness, is limited and is the result of the activation of the networks. Activation is rapid but goes along pathways determined by the ways information is encoded. Encoding puts information into particular categories and structures the pathways by which the information will be retrieved. Cues are stimuli that are related to the codes and stimulate the activation of the networks. Activation is rapid but does take time. The amount of time it takes for someone to respond to a stimulus (reaction time) is often used in research as a clue to the way information is coded.

There are number of models of the question-answering process (Cannell, Miller and Oksenberg 1981; Strack and Martin 1987; Tourangeau and Rasinski 1988; Sudman *et al.* 1996;) that, while differing in details, generally agree on a series of processes respondents go through in answering questions. These processes are: 1) comprehending the meaning of the question; 2) retrieving relevant information; 3) formulating an answer; 4) formatting and editing the answer to meet the requirements of the interviewer and respondents self-presentation. While conceptually viewed as a linear sequence, it is recognized that in reality the processes occur in the flow of a conversation and that the different processes may go on in parallel or in rapid cycling back and forth. For purposes of considering the question-answer process, it is useful to consider them as if they were separate and proceeded in an orderly sequence.

Comprehension

In order to answer a question, respondents must first understand what they are being asked. The goal for the researcher is for respondents to understand the question in the same way that the researcher does. This goal is very difficult to reach because of the many subtleties and ambiguities of language. Indeed Belson (1981), who has studied extensively respondents' understanding of common terms such as "weekday", "children," "regularly" and "proportion," found widespread misunderstanding even in questions using such common terms.

Comprehension begins with a perceptual process of interpreting a string of sounds or written symbols as words in a language that respondents understand. The string of words is "parsed" into syntactical units that are understood, that is, the meaning that is encoded in the linguistic units is extracted by a process that is still poorly understood. Many comprehension problems occur because of ambiguities arising from words that have different meanings (lexical ambiguity) or are used in different ways (structural ambiguity). For example, the question "Where is the table?" is lexically ambiguous because the word "table" can refer to an object on which things can be placed or a set of numbers arranged in a sheet of paper. The sentence "Flying planes can be dangerous" is structurally ambiguous. The interpretation depends on whether "flying" is understood as a verb or as an adjective. Structural ambiguities can be resolved by careful wording of questions. Lexical ambiguities, on the other hand, are inherent in language and are usually resolved by the context within which the sentences appear.

Context plays an important role not only in resolving ambiguities but also aids in interpreting the meaning of words that are unfamiliar. For example, a study by Schuman and Presser (1981) found that a question about the Monetary Control Bill, an obscure piece of proposed legislation, was interpreted as referring to an anti-inflationary measure when it occurred after a series of questions about inflation, but was interpreted as referring to controls of the international transfer of money when it occurred after questions dealing with the balance of payments.

The underlying psychological mechanism for these types of context effects is priming. In order to interpret the stream of sounds or written symbols, we have to draw on our semantic memory that contains the store of linguistic information that enables us to understand the languages we know. Since this is a large store of knowledge, it takes time to retrieve information, and some things will be more easily accessible than others. Those bits of information that have been recently activated are more easily accessible and will be used first to interpret what is being said or read. Priming activates thoughts or "schemata", that is, organized thoughts about objects or concepts, so that they are more accessible to consciousness and thus more easily come into play in interpreting the questions. In the example above, previous questions have primed either thoughts about inflation or about international flows of money, so that when the

unfamiliar concept of the Monetary Control Bill is asked about, the thoughts that have been primed come more rapidly to the fore and affect the interpretation of the words.

Different meanings may be differentially accessible to different respondents because of the frequency with which they employ them in daily life. For example, Billiet (cited in Bradburn 1992, page 317) observed that, in response to the question "How many children do you have?" some respondents offered numbers between twenty and thirty. Further inspection of the data revealed that these respondents were teachers who interpreted the question to refer to the children in their classes, the meaning that was most accessible in their memories.

Information Retrieval

Once a question has been comprehended, respondents must retrieve from memory the information necessary to answer the question. In almost all cases this means retrieving the information from long-term memory. If the question is about behavior, the relevant information is likely to be stored in episodic memory. If the question is about attitudes, the relevant information is likely to be stored in semantic memory, but may require some retrieval from episodic memory.

Remembering is a process by which the memory storehouse is searched to retrieve a particular item that is being sought. If we think of memory as a big storehouse, it is clear that it must be organized in some way in order for us to be able to retrieve things from it. Just as we must label files when we put them in file drawers, so we must attach some kind of labels to information in the memory storehouse. The labeling process, often called "encoding," refers to various aspects of the information or the experience, including emotional tone, attached to the item when we stored is it memory so that we can retrieve it. (For a more complete discussion of memory models see Tourangeau *et al.* 2000, Chapter 3).

Barsalou (1988) has proposed a theory that provides a good framework for understanding how information about personal events is stored in memory. He notes that information about activities or event types in episodic memory includes not only specific events but also extensive idiosyncratic, generic knowledge about the events, that is, having a generic mental image of some types of activity, *e.g.*, visiting a pediatrician, rather than an image of a particular event, *e.g.*, going to Dr. Jones about your daughter's rash (Brewer 1986, 1994). For activities to be stored in memory, they must be comprehended. In other words they must be understood within some meaning system, usually linguistic, that brings to bear knowledge of past activities and generic knowledge about similar event types as well as specifics of the event itself and the context within which it occurred. This complex set of information that goes into the comprehension of the event becomes integrated into the memory of the event. The comprehension process determines how the memories are encoded.

Information, such as the wording of the question and any explanatory material available to respondents at the time they are asked to recall an event, acts as retrieval cues. Retrieval cues are any words, images, emotions, *etc.* that activate or direct the memory search process. If retrieval cues do not specify the event type, *e.g.*, pediatrician visits, then the event types must be inferred before the search can begin. This inference can come from the wording of the question or from the larger context in which the question is asked, including the preceding questions or the introductory material to the survey.

Retrieval is an active process that is facilitated by cues in the question that activate the pathways of association leading to the desired information. Because information, both in episodic and semantic memory, is encoded in many different ways, the cues in the question or in the context surrounding the question including previous questions, may facilitate or constrain the activation and produce better or less good retrieval.

Retrieval takes time. One clear empirical finding is that giving respondents more time to answer questions produces more accurate reports, particularly for behavioral questions. But time is not all there is to it. Memories for events in one's life appear to be organized in event sequences (Barsalou 1988), for example, a summer vacation or a hospitalization, which are hierarchically organized. Giving respondents cues to remind them about the sequence is more effective than trying to get them to retrieve information about a specific event. For example, in questions about alcohol consumption, giving examples of the kinds of situations in which one might drink increases consumption reports.

Examples are an important aid to recall, but they are not a panacea. Giving respondents of list of magazines that they might have read improves reports of reading; a list of organizational types helps respondents remember all the organizations they belong to. While examples may help reduce omissions, they have the effect also of being direct cues for memory and result in greater reports for the types of items on the list. If an important type of activity or event is omitted from a list, the lack of a cue for that type of activity may result in underreporting. The cuing effect of question wording can scarcely be overestimated.

When thinking about retrieval, we mostly think about forgetting or failure to retrieve relevant information. Some times, however, incorrect information may be retrieved that results in overreporting behavior. The best-known example is the phenomenon observed by Neter and Waksberg (1964)

called "telescoping", that is, recalling events that took place at a time other than the time period asked about. Telescoping occurs in response to questions about behavior in a defined time period such as: "How many times have you been to the doctor in the past 6 months?" Neter and Waksberg found in analyzing data from the Consumer Expenditure Survey that when respondents reported on purchases in different reference periods, there was a systematic overreporting of purchases that came from reporting purchases made in a pervious period as if they had been purchased in the period being asked about. While the phenomenon has been observed in a number of studies, there had been no cognitive explanation for it until recently.

Memory for the time of events becomes more uncertain the further back in time the event happened, even though there is no systematic bias in the reports. Telescoping results from the conjunction of two processes-rounding and bounding. Rounding refers to the fact that respondents round their estimates for when things took place in successively larger periods the further back in time an event occurred. For example, events are remembered as having occurred in "days ago" discretely up to about 7 days ago, then they are rounded to periods such as 10 days, two weeks, 4 weeks, 3 months, and 6 months ago. Bounding refers to the aspect of the question that limits the time of reports, *e.g.*, the last 6 months. The effect of this bounding is to truncate reports of events that are remembered as having occurred longer ago than 6 months. Since the variance in the memory for the dates of events becomes larger the further back the event occurred, a larger number of events will be incorrectly remembered as falling into the period the further back the events occurred. This overreporting of events from outside the period will not be offset by an underreporting of events in the near term because events cannot be reported that have not yet happened. Since there are no offsetting events remembered as occurring outside the period at the other end of the time boundary, *i.e.*, the future, the result is a net overreport. (For a full explication of the model see Huttenlocher, Hedges and Bradburn 1990).

Formulating an Answer

Taking into account the information activated by the cues provided by the questions and the context in which they are asked and retrieved from memory, respondents must formulate an answer to the question. Some information is easily accessible. For example, if the questions are about well-rehearsed topics, such as birthdates or marital status, or about topics for which the respondents have an already well-articulated position, respondents may retrieve the answers directly. They spring, as it were, fully formed from memory and can be reported directly. This kind of information we call chronically accessible.

On the other hand, if the questions are about behavior that has not been thought about recently and is not well-remembered or about attitudes that have not been well thought out or discussed, respondents must construct answers on the spot using all the information from whatever source available to them in working memory. This construction process utilizes not only chronically available information but also, importantly, information that is temporarily accessible because it has been activated by the question itself, contextual cues, previous questions, or any other aspects of the interview situation.

There are several general cognitive processes that are pervasive strategies used to process information efficiently. Assimilation and contrast are two such fundamental processes that affect communications. In the study of perception, assimilation refers to the tendency to perceive stimuli as more alike that they actually are. Contrast refers to the tendency to perceive stimuli as more different than they actually are. Applying these principles to survey answering leads to what has been called the inclusion/exclusion model (Schwarz and Bless 1992; Sudman *et al.* 1996). Information that is included in the temporary representation that respondents form of the target of the question will result in assimilation effects because the judgment required to answer the question is based on information included in the representation used. If the information is positive, the judgment will be more positive. If the information is negative, the judgment will be more negative. The size of the effect depends on the amount and extremity of the temporarily accessible information

Previous questions may activate thoughts that are then included in the representation of topics of later questions. The impact of a given question decreases as the number of other context questions increases. For example, answering a question about marital happiness had a pronounced effect on answers to subsequent questions about general life satisfaction when respondents' marriages were the only specific life domain asked about. When respondents were asked about their leisure time and their jobs in addition to questions about their marriages before reporting on life satisfaction, the effect was significantly reduced. (Schwarz, Strack and Mai 1991).

Information that is excluded rather than included in the temporary representation of the target will lead to a contrast effect. In this case, if the information excluded is positive, the judgment will become more negative; if the information is negative, the judgment will become more positive. Similarly the size of the effect depends on the amount and extremity of the temporarily accessible information. In effect, the excluded information is subtracted from the representation of the attitude object.

Excluded information, however, may play an additional role in formulating judgments. In addition to being excluded from the representation of the target, the information may be used in constructing a standard or scale anchor. In this case we speak of comparison-based contrast effects. The effect here is not caused so much by the subtraction of the excluded information from the evaluation of the attitude target, but by the comparison of the target with some standard or evaluated on some scale.

Which of these processes drives the emergence of a contrast effect determines whether the contrast effect is limited to the single object or generalizes across related objects. If the contrast effect is based on simple subtraction, the effect is limited to that particular target. If the contrast effect is based on a comparison, the effects are apt to appear in each judgment where that standard of comparison is relevant.

An example of a contrast effect based on using information from previous questions is provided in a study by Schwarz, Muenkel and Hippler (1990). Respondents were asked to rate a number of beverages according to how "typically German" they were. When this question was preceded by a question about the frequency with which Germans drink beer or vodka, contrast effects appear in the typicality ratings. Respondents who had estimated the consumption of beer first (a high frequency item), rated wine, milk and coffee as less typical German drinks than did respondents who had estimated the consumption of vodka first (a low frequency item), thus showing a contrast effect that extended across the three target drinks. This contrast effect, however, did not appear when the preceding question was about the caloric context of beer or vodka because the information activated by this question was not relevant to a judgment about typicality.

Formatting and Editing Responses

After respondents have formulated their responses, there remains the task of fitting these answers into the response formats that the interviewer offers. Rarely in surveys does the researcher allow respondents to answer questions in a free format. Open-ended questions have a multitude of problems not least of which is the cost and difficulty of transforming free-form answers in a format that can be treated quantitatively. Today almost all questionnaires depend on closed or pre-coded questions.

Research on response alternatives is less well developed theoretically than the study of question wording and context effects. In general, the empirically observed effects are thought to stem from two sources-memory limitations and cognitive elaboration stimulated by the response alternatives.

Memory limitations create some order effects among response alternatives. Primacy and recency are two well-known effects in the memory literature. When a series of stimuli are present visually, those that come early in the series are remembered better than those later in the series (primacy). When a series of stimuli are present in an auditory mode, those that come late in the series are remembered better (recency). Thus there is an interaction between the order in which stimuli are presented and the mode by which they are presented.

The research literature has shown that there are persistent, although in general samples fairly small, primacy and recency effects in the serial position of response alternatives depending on the mode presentation. Primacy effects appear when the response alternatives are presented visually, as in show cards in personal interviewing, and recency effects appear in telephone interviewing when the respondents have to depend entirely on auditory memory for the response alternatives. More recent research (Knaeuper 1999; Schwarz and Knaeuper 2000), however, reveals that the effect is very much a function of memory capacity and is sharply increased among older respondents whose memory is poorer and who depend more on the primacy or recency of the stimuli as supported by mode of presentation. Among older respondents, the primacy/recency effects can be quite large, on the order to 20 percentage points (Schwarz and Knaeuper 2000). Among younger respondents the effects are small.

An intriguing theory to account for some observed response order effects within a question is that of cognitive elaboration. This theory draws on early work by Krosnick and Alwin (1987) and cognitive research on persuasion (Eagly and Chaiken 1993; Petty and Cacioppo 1986). This theory hypothesizes that the order and mode in which response alternatives are presented affects respondents' opportunity to elaborate on their content. Such elaboration, in turn, activates thoughts in response to the question and provides retrieval cues in response to behavioral questions. The response alternatives provide supplementary cues that activate a range of thoughts that become temporarily accessible and may become part of the answer formulation process. In effect, the response alternatives are an essential part of the question but may be processed later in time after the question itself has been processed.

The cognitive elaboration hypothesis suggests a number of complex predictions, few of which have yet been tested. One example for which there is considerable evidence, is an interaction between serial position and mode of administration in long lists. The primacy effect evident in visually presented material gives respondents time and stimulus to think more about alternatives early in the list before giving an answer. The crowding out of early alternatives by the

reading of later alternatives and recency effects evident in lists presented in an auditory mode suggest that the later alternatives can be more deeply processed cognitively. These effects are more robust that the primacy and recency effects that appear to depend more on simple memory limitations.

Once a response alternative has been chosen in the respondents' mind, the respondent may still edit the response. As mentioned earlier, the interview is a social situation and respondents may be concerned with self-presentation. There is ample evidence that social desirability is an important aspect of the response process and responses to sensitive question may be seriously distorted by unwillingness to admit to behavior or attitudes that would put the respondent in a bad light in the interviewer's eyes or by the desire to over claim socially desirable behavior (Bradburn, Sudman and Associates 1979; Sudman and Bradburn 1974). There are several techniques for reducing social desirability bias, although there is no technique that totally and reliably eliminates it. The general strategy is to increase social distance between respondents and interviewers. This can be done by changing the mode of administration by eliminating or reducing the presence of the interviewer. Computer Assisted Personal Interviews (CAPI) which allow respondents to directly enter responses to sensitive questions into the computer as part of a face-to-face interview enable researchers to combine the benefits of a personal interview with a self-administered questionnaire. The use of audio enhanced CAPI (Audio-CAPI) which enables respondents to listen to a recorded voice reading the questions, although somewhat more expensive, overcomes literacy and language problems that might arise when respondents have to read questions from a computer screen.

Research on mode effects generally indicates that self-administration of a questionnaire, particularly in an anonymous, group setting, minimizes, but does not entirely eliminate desirability bias. Interviews done on the telephone generaly produce results that are intermediate between a face-to-face interview and a totally anonymous self-administration, although the results are not entirely consistent.

In addition to reducing the social distance between interviewer and respondent by altering the mode of administration there are techniques for increasing the real or perceived anonymity of respondents that also reduce social desirability bias. For example, respondents may put their responses in a sealed envelope and mail them back to a central office so that they know that the interviewer cannot see their responses.

Another technique is the so-called random response technique, although it is more properly a random question technique (Greenberg, Abul-Ela, Simmons and Horvitz 1969; Horvitz, Shah and Simmons 1967; Warner 1965). The interviewer asks two questions, one sensitive and the other non-sensitive. Both questions have the same possible answers, "yes" and "no". Which question the respondent answers is determined by a probability mechanism, such as flipping a coin or using a plastic box containing two colored beads, *e.g.*, red and blue beads, in differing proportions, *e.g.*, 70% red beads and 30% blue beads. The box is designed so that when it is shaken by respondents a red or a blue bead seen only by the respondent will appear in the window of the box. If the bead is red, the sensitive question is answered; if blue, the non-sensitive question is answered. The interviewer does not know which question is answered.

By using this procedure you can estimate the behavior of a group on the sensitive questions, but not that of any single individual. Thus with this method you cannot relate individual characteristics of respondents to individual behavior. If you have a very large sample, group characteristics can be related to the estimates obtained from randomized responses. For example, you could look at all the answers of young women and compare them to all the answers of men or young versus older age groups. On the whole, however much information is lost when randomized response is used.

While, compared with other methods, randomized response greatly reduces the under reporting of undesirable behavior, it does little to reduce the overreporting of desirable behavior. It also does not entirely eliminate underreporting of undesirable behavior (Bradburn *et al.* 1979).

CONCLUSION

In this essay, I have tried to present the outlines of a social psychological approach to the understanding of the question-answer process in the survey interview. This approach draws on theory from sociology, cognitive psychology and linguistics, to present a comprehensive framework for research on response effects. Much, however, remains uncertain or unknown.

While social role theory provides a good starting point for conceptualizing the social relations among researchers, interviewers and respondents, there is much we do not know about how these roles are played by their respective actors and how they may be changing. Contemporary concerns about privacy and confidentiality of data and protection of human participants in research are changing to an unknown degree the way respondents view surveys and social research. Technology is changing respondents' ability to protect their privacy and researchers' ability to protect confidentiality of data. Response rates have been declining and greater efforts are required to convince sampled persons to respond. Interviewing is increasingly mediated by

computer-assistance, which may change the way in which respondents and interviewers interact and the way respondents view the interview situation.

The cognitive processes involved in formulating an answer are complex and not yet fully understood. The application of our understanding of fundamental cognitive processes to the study of question formulation and order goes a long way toward improving our understanding of context effects. Cognitive science is making great strides in understanding how the brain works and how we organize and process information. New knowledge in these areas grows at a rapid pace. As we learn more, many of the conceptualizations outlined in this essay will change and either shown to be wrong or greatly elaborated.

Finally there is a great challenge to linguistics. Many of the effects we have discussed in this essay occur because of ambiguities in language. Understanding how meaning is encoded in language and how we extract that meaning from spoken and written language is a formidable challenge. Perhaps more than anything else, our ability to resolve some of the most fundamental problems in questionnaire construction depends on progress in these areas.

What are the high priority areas for research? In the short run, I would concentrate on better understanding of the biasing effects of declining respondent participation, particularly on possible distortions of responses from reluctant respondents. We must develop response effect models that not only account for missing data, whether at the item level or at the whole person level, but also for response effects introduced by reluctant respondents who give only partial answers or not well-considered answers. Multiple imputation models such as those developed by Little and Rubin (1987) and latent variable approaches such as developed by O'Muircheartaigh and Moustaki (1999) are promising. More empirical work is needed on the effects of pushing people into responding who inititially are unwilling to participate in a survey.

In the longer run, further research is needed on the mechanisms by which questions and answer categories stimulate cognitive elaboration and activate thoughts that are then used in answering questions. We need to know what it is about questions that cause respondents to exclude information in making a judgment as contrasted with those that stimulate them to include information when they make judgments. Progress in this area will require a close collaboration between cognitive psychologists and survey methodologists and involve both laboratory and field survey work.

In the end, however, fundamental understanding of the question-answer process will only come when we under stand how meaning is communicated between human beings. Questions have meaning that we expect respondents to comprehend. We can only go so far in improving the process of clear communication without a much deeper understanding of the basic mechanisms of communication. We need a concerted multidisciplinary effort by linguists, psychologists, statisticians, and cognitive scientists and others to crack the meaning code much as natural scientists cracked the genetic code. It is one of the grand scientific challenges of our time.

REFERENCES

BADDELEY, A. (1979). The limitations of human memory: Implications for the design of retrospective surveys. In *The Recall Method in Social Surveys*, (Eds. L. Moss and H. Goldstein). London: NFER Publishing Co., Ltd.

BARSALOU, L.W. (1988). The content and organization of autobiographical memories. In *Remembering Reconsidered: Ecological and Traditional Approaches to the Study of Memory*, (Eds. U. Neisser and E. Winograd). Cambridge, England: Cambridge University Press.

BELSON, W.A. (1968). Respondent understand of survey questions. *Polls*. 3(1), 1-13.

BELSON, W.A. (1981). *The Design and Understanding of Survey Questions*. Aldershot, England: Gower.

BIDERMAN, A. (1980). *Report of a Workshop on Applying Cognitive Psychology to Recall Problems of the National Crime Survey*. Washington, D.C.: Bureau of Social Science Research.

BINGHAM, W.V.D., and MOORE, B.V. (1934). *How to Interview* (Rev. ed.). New York: Harper Collins.

BISHOP, G.F., OLDENDICK, R.W. and TUCHFARBER, R.J. (1986). Opinions on fictitious issues: The pressure to answer survey questions. *Public Opinion Quarterly*. 50, 240-250.

BRADBURN, N.M. (1992). What have we learned? In *Context Effects in Social and Psychological Research*, (Eds. N. Schwarz and S. Sudman). New York: Springer-Verlag.

BRADBURN, N.M., and DANIS, C. (1984). Potential contributions of cognitive research to survey questionnaire design. In *Cognitive Aspects of Survey Methodology: Building a Bridge Between Disciplines*, (Eds. T.B. Jabine, M.L. Straf, J.M. Tanur, and R. Tourangeau). Washington, D.C.: National Academy Press.

BRADBURN, N.M., and SUDMAN, S. (1974). *Response Effects in Surveys: A Review and Synthesis*. Chicago: Aldine.

BRADBURN, N.M., SUDMAN, S. and ASSOCIATES (1979). *Improving Interview Method and Questionnaire Design*. San Francisco: Jossey-Bass.

BREWER, W.E. (1986). What is autobiographical memory? In *Autobiographical Memory*, (Ed. D.C. Rubin). Cambdrige, England: Cambridge University Press.

BREWER, W.E. (1994). Autobiographical memory and survey research. In *Autobiographical Memory and the Validity of Retrospective Reports*, (Eds. S. Schwarz and S. Sudman). New York: Springer-Verlag.

CANNELL, C.F., MILLER, P. and OKSENBERG, L. (1981). Research on interviewing techniques. In *Sociological Methodology 1981*, (Ed. S. Leinhardt). San Francisco: Jossey-Bass.

EAGLY, A.H., and CHAIKEN, S. (1993). *The Psychology of Attitudes*. Orlando, Fl.: Harcourt, Brace, Jovanovich.

GREENBERG, B.G., ABUL-ELA, A.L., SIMMONS W.R. and HORVITZ, D.G. (1969). The unrelated question randomized response model: Theoretical framework. *Journal of the American Statistical Association*. 64, 520-539.

GRICE, H.P. (1975). Logic and conversation. In *Syntax and Semantics 3: Speech Acts*, (Eds. P. Cole and J.L. Morgan). New York: Academic Press. 41-58.

HIPPLER, H.J., SCHWARZ, N. and SUDMAN, S. (Eds.) (1985). *Social Information Processing And Survey Methodology*. New York: Springer-Verlag.

HORVITZ, D.G., SHAH, B.V. and SIMMONS, W.R. (1967). The unrelated question randomized response model. In *Proceedings of the Social Statistics Section*, American Statistical Association. 65-72.

HUTTENLOCHER, J., HEDGES, L.V. and BRADBURN, N.M. (1990). Reports of elapsed time: Bounding and rounding processes in estimation. *Journal of Experimental Psychology, Learning, Memory and Cognition*. 16, 196-213.

JABINE, T., STRAF, M., TANUR, J. and TOURANGEAU, R. (Eds.) (1984). *Cognitive Aspects of Survey Methodology: Building a Bridge Between Disciplines*. Washington, D.C.: National Academy Press.

KNAEUPER, B. (1999). The impact of age and education on response order effects in attitude measurement. *Public Opinion Quarterly*. 63, 347-370.

KROSNICK, J.A., and ALWIN, D.F. (1987). An evaluation of a cognitive theory of response order effects in survey measurement. *Public Opinion Research Quarterly* 51,201-219.

LITTLE, R., and RUBIN, D.B. (1987). *Statistical Analysis with Missing Data*. New York: John Wiley & Sons, Inc.

LOFTUS, E.F., and MARBURGER, W. (1985). Since the eruption of Mt. St. Helens, has anyone beaten you up? Improving the accuracy of retrospective reports with landmark events. *Memory and Cognition*. 11, 114-120.

NETER, J., and WAKSBERG, J. (1964). A study of response errors in expenditure data from household interviews. *Journal of the American Statistical Association*. 59, 18-55.

O'MUIRCHEARTAIGH, C., and MOUSTAKI, I. (1999). Symmetric pattern models: a latent variable approach to item non-response in attitude scales. *Journal of the Royal Statistical Society, A*. 162, 2, 177-194.

PETTY, R.E., and CACIOPPO, J.T. (1986). *Communication and Persuasion: Central and Peripheral Routes to Attitude Change*. New York: Springer-Verlag.

SCHUMAN, H., and PRESSER, S. (1981). *Questions and Answers in Attitude Surveys*. New York: Academic Press.

SCHWARZ, N., and BLESS, H. (1992). Constructing reality and its alternatives: Assimilation and contrast effects in social judgment. In *The Construction of Social Judgments*, (Eds. L.L, Martin and A. Tesser). Hillsdale, N.J.: Erlbaum. 217-245.

SCHWARZ, N., and KNAEUPER, B. (2000). Cognition, aging, and self-reports. In *Cognitive Aging: A Primer*, (Eds. D.C. Park and N. Schwarz). Philadelphia: Psychology. 233-252.

SCHWARZ, N., MUENKEL, T. and HIPPLER, H.J. (1990). What determines a perspective? Contrast effects as a function of the dimension tapped by preceding questions. *European Journal of Social Psychology*. 20, 357-361.

SCHWARZ, N., STRACK, F. and MAI, H.F. (1991). Assimilation and contrast effects in part-whole question sequences: A conversational logic analysis. *Public Opinion Quarterly*. 55, 3-23.

SCHWARZ, N., and SUDMAN, S. (Eds.) (1992). *Context Effect in Social and Psychological Research*. New York: Springer-Verlag.

SCHWARZ, N., and Sudman, S. (Eds.) (1994). *Autobiographical Memory and the Validity of Retrospective Reports*. New York: Springer-Verlag.

SCHWARZ, N., and SUDMAN, S. (Eds.) (1996). *Answering Questions: Methodology of Determining Cognition and Communication Processes in Survey Research*. San Francisco: Jossey-Bass.

STRACK, F., and MARTIN, L.L. (1987). Thinking, judging, and communicating: A process account of context effects in attitude surveys. In *Social Information Processing and Survey Methodology*, (Eds. H.J. Hippler, N. Schwarz, and S. Sudman). New York: Springer-Verlag. 123-148.

SUDMAN, S., and BRADBURN, N.M. (1974). *Response Effects in Surveys: A Review and Synthesis*. Chicago: Aldine.

SUDMAN, S., BRADBURN, N.M. and SCHWARZ, N. (1996). *Thinking About Answers*. San Francisco: Jossey-Bass.

TANUR, J.M. (Ed.) (1992). *Questions About Questions: Inquiries Into the Cognitive Bases of Surveys*. New York: Russell Sage Foundation.

TOURANGEAU, R. (1984). Cognitive sciences and survey methods. In *Cognitive Aspects of Survey Methodology: Building a Bridge Between Disciplines*, (Eds. T.B. Jabine, M.L. Straf, J.M. Tanur, and R. Tourangeau). Washington, D.C.: National Academy Press

TOURANGEAU, R., and RASINSKI, K. (1988). Cognitive processes underlying context effects in attitude measurement. *Psychological Bulletin*. 103, 299-314.

TOURANGEAU, R., RIPS, L.J. and RASINSKI, K. (2000). *The Psychology of Survey Response*. Cambridge, England: Cambridge University Press.

TURNER, C.F., and MARTIN, E. (1982). *Surveys of Subjective Phenomena*. Cambridge, MA: Harvard University Press.

WARNER, S.L. (1965). Randomized response: A survey technique for eliminating error answer bias. *Journal of the American Statistical Association*. 60, 63-69.

WENTLAND, E.J., and SMITH, K.W. (1993). *Survey Responses: An Evaluation of Their Validity*. San Diego: Academic Press.

Survey Methodology, June 2004
Vol. 30, No. 1, pp. 17-26
Statistics Canada

Linearization Variance Estimators for Survey Data

ABDELLATIF DEMNATI and J.N.K. RAO[1]

ABSTRACT

In survey sampling, Taylor linearization is often used to obtain variance estimators for calibration estimators of totals and nonlinear finite population (or census) parameters, such as ratios, regression and correlation coefficients, which can be expressed as smooth functions of totals. Taylor linearization is generally applicable to any sampling design, but it can lead to multiple variance estimators that are asymptotically design unbiased under repeated sampling. The choice among the variance estimators requires other considerations such as (i) approximate unbiasedness for the model variance of the estimator under an assumed model, (ii) validity under a conditional repeated sampling framework. In this paper, a new approach to deriving Taylor linearization variance estimators is proposed. It leads directly to a variance estimator which satisfies the above considerations at least in a number of important cases. The method is applied to a variety of problems, covering estimators of a total as well as other estimators defined either explicitly or implicitly as solutions of estimating equations. In particular, estimators of logistic regression parameters with calibration weights are studied. It leads to a new variance estimator for a general class of calibration estimators that includes generalized raking ratio and generalized regression estimators. The proposed method is extended to two-phase sampling to obtain a variance estimator that makes fuller use of the first phase sample data compared to traditional linearization variance estimators.

KEY WORDS: Calibration; Design weights; Estimating equations; Raking ratio estimator; Regression estimators; Two-phase sampling.

1. INTRODUCTION

Taylor linearization is a popular method of variance estimation for complex statistics such as ratio and regression estimators and logistic regression coefficient estimators. It is generally applicable to any sampling design that permits unbiased variance estimation for linear estimators, and it is computationally simpler than a resampling method such as the jackknife. However, it can lead to multiple variance estimators that are asymptotically design unbiased under repeated sampling. The choice among the variance estimators, therefore, requires other considerations such as (i) approximate unbiasedness for the model variance of the estimator under an assumed model, (ii) validity under a conditional repeated sampling framework. For example, in the context of simple random sampling and the ratio estimator, $\hat{Y}_R = (\bar{y}/\bar{x})X$, of the population total Y, Royall and Cumberland (1981) showed that a commonly used linearization variance estimator, $v_L = N^2(n^{-1} - N^{-1})s_z^2$, does not track the conditional variance of $\hat{Y}_R$ given $\bar{x}$, unlike the jackknife variance estimator v_J. Here $\bar{y}$ and $\bar{x}$ are the sample means, X is the known population total of an auxiliary variable x, s_z^2 is the sample variance of the residuals $z_k = y_k - (\bar{y}/\bar{x})x_k$ and (n, N) denote the sample and population sizes. By linearizing the jackknife variance estimator, v_J, a different linearization variance estimator, $v_{JL} = (\bar{X}/\bar{x})^2 v_L$, is obtained. This variance estimator also tracks the conditional variance as well as the unconditional variance, where $\bar{X} = X/N$ is the mean of x. As a result, v_{JL} or v_J may be preferred over v_L. Yung and Rao (1996) considered generalized regression and ratio-adjusted post-stratified estimators under stratified multistage sampling and obtained a jackknife linearization variance estimator, v_{JL} by linearizing v_J. Valliant (1993) also obtained v_{JL} for the ratio-adjusted post-stratified estimator and conducted a simulation study to demonstrate that both v_J and v_{JL} possess good conditional properties given the estimated post-strata counts. Särndal, Swensson and Wretman (1989) showed that v_{JL} is both asymptotically design unbiased and approximately model unbiased in the sense of $E_m(v_{JL}) \approx V_m(\hat{Y}_R)$, where E_m denotes model expectation and $V_m(\hat{Y}_R)$ is the model variance of $\hat{Y}_R$ under a "ratio model": $E_m(y_k) = \beta x_k$; $k = 1, ..., N$ and the y_k's are independent with model variance $V_m(y_k) = \sigma^2 x_k$, $\sigma^2 > 0$. Thus, v_{JL} is a good choice from either the design-based or the model-based perspective.

Binder (1996) presented an elegant "cookbook" approach to Taylor linearization that leads directly to v_{JL}-type linearization variance estimators. He applied the method to smooth functions of estimated totals, $g(\hat{Y}_1, ..., \hat{Y}_m)$, generalized regression estimators and the Wilcoxon rank sum statistic. To illustrate Binder's method, consider a ratio estimator

$$\hat{Y}_R = (\hat{Y}/\hat{X})X = \hat{R}X,$$

[1] Abdellatif Demnati, Social Survey Methods Division, Statistics Canada, R.H. Coats Bldg, 15th Floor, Ottawa, Ontario, Canada, K1A 0T6; J.N.K. Rao, School of Mathematics and Statistics, Carleton University, Ottawa, Ontario, Canada, K1S 5B6.

where $\hat{Y} = \sum_{k=1}^{N} d_k(s)\, y_k = \hat{Y}(y)$, $\hat{X} = \sum_{k=1}^{N} d_k(s) x_k = \hat{Y}(x)$ and the $d_k(s)$ are the design weights with $d_k(s) = 0$ if the population element k is not in the sample s, *e.g.*, $d_k(s) = (1/\pi_k) a_k(s)$ where π_k is the probability of including the element k in the sample s, $a_k(s) = 1$ if $k \in s$, $a_k(s) = 0$ otherwise, and Σ denotes summation over the population elements. The weights are assumed to provide a design unbiased estimator $\hat{Y}$ of Y, *i.e.*, $E(d_k(s)) = 1$ for $k = 1, \ldots, N$. Now take the total differential of $\hat{Y}_R$ to get

$$(d\hat{Y}_R) = (d\hat{R})X = \frac{X}{\hat{X}}\left[(d\hat{Y}) - \hat{R}(d\hat{X})\right], \tag{1.1}$$

and replace all the total differentials in (1.1) by deviations of estimators from their respective population parameters, *e.g.*, $d\hat{Y}_R$ is changed to $\hat{Y}_R - Y$. Then (1.1) yields

$$\hat{Y}_R - Y \doteq \sum d_k(s)\, z_k - \frac{X}{\hat{X}}(Y - \hat{R}X), \tag{1.2}$$

where

$$z_k = \frac{X}{\hat{X}}(y_k - \hat{R}x_k). \tag{1.3}$$

The term $\Sigma d_k(s)\, z_k$ in (1.2) reduces to zero, but it is retained for variance estimation. On the other hand, the last term of (1.2) is ignored for variance estimation. Thus, $\hat{Y}_R - Y$ is represented as $\Sigma d_k(s)\, z_k = \hat{Y}(z)$ for the purpose of variance estimation. Denoting an unbiased variance estimator of $\hat{Y} = \hat{Y}(y)$ as $v(y)$, Binder's variance estimator of $\hat{Y}_R$ is given by $v(z)$. The linearization variance estimator $v(z)$, obtained from (1.3), agrees with v_{JL} for simple random sampling and stratified multistage sampling if the sample is treated as if the primary sampling units are sampled with replacement. Note that the jackknife method is not applicable generally for any sampling design.

For the estimator $\hat{\theta} = g(\hat{Y}_1, \ldots, \hat{Y}_m)$ of a smooth function of totals, $\theta = g(Y_1, \ldots, Y_m)$, Binder's (1996) method leads to

$$\hat{\theta} - \theta \doteq \sum d_k(s)\, z_k + \ldots$$

with

$$z_k = \sum_{i=1}^{m} \left(\partial g(\boldsymbol{a})/\partial a_i\big|_{\boldsymbol{a} = \hat{\boldsymbol{Y}}}\right) y_{ki}, \tag{1.4}$$

where $\hat{\boldsymbol{Y}} = (\hat{Y}_1, \ldots, \hat{Y}_m)^T$ and $\boldsymbol{a} = (a_1, \ldots, a_m)^T$. It follows from (1.4) that the partial derivatives, $\partial g(\boldsymbol{a})/\partial a_i$, are evaluated at $\hat{\boldsymbol{Y}}$ to obtain z_k's, whereas in the standard method (see *e.g.*, Andersson and Nordberg 1994) they are evaluated at $\boldsymbol{Y} = (Y_1, \ldots, Y_m)^T$ before getting z_k and then substituting estimates for the unknown components. For example, for the ratio estimator $\hat{Y}_R$ the term $X/\hat{X}$ disappears from z_k in the standard procedure because $X/\hat{X}$ becomes 1 when $\hat{X}$ is replaced by X.

Although Binder's (1996) approach is simple and attractive, a more rigorous and broadly applicable method is needed. In section 2, we propose an alternative approach that is theoretically justifiable and at the same time leads directly to a v_{JL}-type variance estimator for general designs. We apply the method, in section 3, to a variety of problems, covering regression calibration estimators of a total Y and other estimators defined either explicitly or implicitly as solutions of estimating equations, *e.g.*, estimators of logistic regression parameters with design weights calibrated to known auxiliary population totals. We also obtain a new variance estimator for a general class of calibration estimators that includes generalized raking ratio and generalized regression estimators. Section 4 extends the proposed method to two-phase sampling to obtain a variance estimator that makes fuller use of the first phase sample data compared to traditional linearization variance estimators.

For the case of independent and identically (iid) random variables $y_1, \ldots, y_n$ with distribution function $F(y)$, estimation of general parameters $\theta = T(F)$ has been studied extensively in the literature (see *e.g.*, Huber 1981). A natural estimator of $\theta = T(F)$ is $\hat{\theta} = T(\hat{F})$, where $\hat{F}(y)$ is the empirical distribution function given by $\hat{F}(y) = n^{-1} \sum_{k=1}^{n} I(y_k \le y)$ with $I(y_k \le y) = 1$ if $y_k \le y$ and $I(y_k \le y) = 0$ if $y_k > y$. For example, if $T(F)$ is the population mean $\int y\, dF(y)$, then $T(\hat{F}) = \int y\, d\hat{F}(y) = n^{-1} \sum_{k=1}^{n} y_k = \bar{y}$, the sample mean. Note that $\hat{F}$ assigns equal mass, $1/n$ to each of the sample values $y_1, \ldots, y_n$. If T is "sufficiently regular", then $T(\hat{F})$ may be linearized near F in terms of the influence curve (or function) of $T(\cdot)$ given by

$$\mathrm{IC}(y, F, T) = \lim_{a \to 0}\left[T((1-a)F + a\delta_y) - T(F)\right]/a, \tag{1.5}$$

where δ_y denotes the point mass 1 at y. We have

$$\begin{aligned}\sqrt{n}[T(\hat{F}) - T(F)] &= \sqrt{n}\int \mathrm{IC}(y, F, T)\, d\hat{F}(y) + \sqrt{n}\, R_n \\ &= \frac{1}{\sqrt{n}} \sum_{k=1}^{n} \tilde{z}_k + \sqrt{n}\, R_n \end{aligned} \tag{1.6}$$

where $\tilde{z}_k = \mathrm{IC}(y_k, F, T)$ and $\sqrt{n}\, R_n$ is a remainder term. If $\sqrt{n}\, R_n$ is asymptotically negligible in the sense that $\sqrt{n}\, R_n$ converges in probability to zero as $n \to \infty$ (denoted $\sqrt{n}\, R_n \to_p 0$) then it follows from (1.6) that $\sqrt{n}[T(\hat{F}) - T(F)]$ is asymptotically normal with mean 0 and variance

$$A(F, T) = \int [\mathrm{IC}(y, F, T)]^2\, dF(y), \tag{1.7}$$

noting that the terms $\tilde{z}_k$ in (1.6) are iid random variables. As noted by Huber (1981, page 13), $\sqrt{n}\, R_n$ is "often" asymptotically negligible, but the proof of this property may not be easy for general functionals $T(F)$. Serfling (1980, section 6.2) gave the following two conditions for

$\sqrt{n}\,R_n \to_p 0$, applicable for general random variables $y_1, ..., y_n$ (not necessarily iid): (i) $T(\cdot)$ is "stochastically differentiable" at F; (ii) $\sqrt{n}\,sup|\hat{F}(y) - F(y)|$ is bounded in probability, where *sup* is over y. Condition (ii) is satisfied in the iid case, but it may not be easy to prove (ii) for complex sampling designs. Condition (i) means that there exists a functional $T(F; F_n - F)$ such that $T(F_n) - T(F) = n^{-1}\sum_{k=1}^{n} T(F; \delta_{y_k} - F) + R_n$, where R_n is of lower order in probability than $sup|F_n(y) - F(y)|$ as the latter tends to zero. This condition may not be easy to verify for general $T(\cdot)$. Serfling (1980) suggested that in practice it is more effective to analyse R_n directly using "the method of differential inequalities".

A natural estimator of the asymptotic variance $A(F,T)$ is

$$A(\hat{F}, T) = \frac{1}{n}\sum_{k=1}^{n} [\mathrm{IC}(y_k, \hat{F}, T)]^2, \qquad (1.8)$$

where $\mathrm{IC}(y, \hat{F}, T)$ is the influence curve evaluated at $F = \hat{F}$. It follows that a linearization variance estimator of $T(\hat{F})$ is

$$v_L[T(\hat{F})] = A(\hat{F}, T)/n. \qquad (1.9)$$

Practical implementation of $v_L[T(\hat{F})]$ involves the computation of $\mathrm{IC}(y_k, \hat{F}, T)$ for each specified T. The latter can be avoided by using the jackknife method. Substituting $\hat{F}$ for F and $-1/(n-1)$ for a in (1.5), we obtain a jackknife estimator of $\mathrm{IC}(y_k, F, T)$ as $z_{kJ} = (n-1)[T(\hat{F}) - T(\hat{F}_{-k})]$, where $\hat{F}_{-k}(y)$ is the empirical distribution function obtained when y_k is omitted. The resulting jackknife variance estimator $T(\hat{F})$ is

$$v_J[T(\hat{F})] = \frac{1}{n(n-1)}\sum_{k=1}^{n} z_{kJ}^2$$

$$= \frac{n-1}{n}\sum_{k=1}^{n} [T(\hat{F}_{-k}) - T(\hat{F})]^2; \qquad (1.10)$$

see *e.g.*, Hampel, Ronchetti, Rousseeuw and Stahel (1986, page 95). If $\mathrm{IC}(y, F, T)$ does not depend smoothly on F, then the jackknife variance estimator may not be consistent for the variance of $T(\hat{F})$; for example, when $T(\hat{F})$ is the sample median.

Campbell (1980) attempted to extend the above results for the iid case to general sampling designs, using the design weights $d_k(s)$. The population (or census) parameter θ is now given by $\theta = T(F_N)$, where $F_N(y)$ is the population distribution function that assigns equal mass, $1/N$, to each of the N population values $y_1, ..., y_N$. An empirical distribution function is given by $\hat{F}(y) = \sum_{k\in s} \tilde{d}_k(s) I(y_k \le y)$, where $\tilde{d}_k(s) = d_k(s)/\sum_{l\in s} d_l(s)$ are the normalized design weights. Note that $\hat{F}(y)$ assigns the mass $\tilde{d}_k(s)$ to the element $k \in s$. An estimator of $\theta = T(F_N)$ is given by $\hat{\theta} = T(\hat{F})$. For example, if $T(F_N)$ is the population mean $\int y\,dF_N(y)$, then $T(\hat{F}) = \int y\,d\hat{F}(y) = \sum_{k\in s} d_k(s) y_k / \sum_{k\in s} d_k(s)$, the design-weighted sample mean. Campbell (1980) followed the linearization (1.6) for the iid case and concluded that $\sqrt{n}\,[T(\hat{F}) - T(F_N)]$ is asymptotically normal with mean 0 and variance

$$A(F_N, T) = n\,\mathrm{Var}\left[\sum_{k\in s} d_k(s)\tilde{z}_k \Big/ \sum_{k\in s} d_k(s)\right]$$

$$\approx n\,\mathrm{Var}\left[\sum_{k\in s} d_k(s)\{(\tilde{z}_k - R)/N\}\right], \qquad (1.11)$$

using the approximate variance of a ratio, where $R = \sum_{k\in s} \tilde{z}_k / N$ is the population mean of $\tilde{z}_k$'s and $\tilde{z}_k = \mathrm{IC}(y_k, F_N, T)$. Denoting the unbiased variance estimator of $\hat{Y} = \hat{Y}(y) = \sum_{k\in s} d_k(s) y_k$ as $v(y)$, it follows from (1.11) that a linearization variance estimator of $T(\hat{F})$ is given by

$$v_L[T(\hat{F})] = v[(z - \hat{R})/\hat{N}], \qquad (1.12)$$

where

$$z_k = \mathrm{IC}(y_k, \hat{F}, T), \qquad (1.13)$$

and

$$\hat{R} = \sum_{k\in s} d_k(s)\, z_k \Big/ \sum_{k\in s} d_k(s). \qquad (1.14)$$

To avoid the computation of z_k's, Campbell (1980) proposed a jackknife estimator of $\tilde{z}_k$ for each $k \in s$. It is given by

$$z_{kJ} = \frac{1 - \tilde{d}_k(s)}{\tilde{d}_k(s)}[T(\hat{F}) - T(\hat{F}_{-k})], \qquad (1.15)$$

where

$$d\hat{F}_{-k}(y) = \begin{cases} \dfrac{d\hat{F}(y) - \tilde{d}_k(s)}{1 - \tilde{d}_k(s)} & \text{if } y = y_k \\ \dfrac{d\hat{F}(y)}{1 - \tilde{d}_k(s)} & \text{if } y \ne y_k. \end{cases} \qquad (1.16)$$

The resulting linearization variance estimator is given by $v[(z_J - \hat{R}_J)/\hat{N}]$. Note that the proposed jackknife method is different from the customary jackknife for survey sampling. For example, for stratified multistage sampling, the customary jackknife deletes sample clusters in turn whereas the Campbell method deletes elements in turn. Also, the customary jackknife is not always applicable (*e.g.*, unequal probability sampling without replacement) unlike the Campbell method which uses the unbiased variance estimator $v(y)$ of the total $\hat{Y}$ for the given design and then replaces y by $(z_J - \hat{R}_J)/\hat{N}$. However, the computations involved in the Campbell method can be very heavy because it requires the computation of $T(\hat{F}_{-k})$ for each element $k \in s$; in large-scale surveys the number of sample

elements can be very large, as in the Canadian Labour Force Survey.

Deville (1999) and Berger (2002) obtained results very similar to those of Campbell (1980). Instead of using the natural probability measure $\hat{F}$, they considered functionals of the form $T(\hat{M})$, where $\hat{M}$ denotes a measure that allocates the design weight $d_k(s)$ to any point y_k for k in s and zero to units k not in s. For example, $T(\hat{M}) = \int x d\hat{M}(x) = \Sigma\, d_k(s)\, y_k$ if the population parameter is the total $T(M) = \int x dM(x) = Y$, where the measure M allocates a unit mass to each of the N points y_k in the finite population U. Suppose that $T(\cdot)$ is of degree α in the sense that $N^{-\alpha} T(\cdot)$ tends to a limit for some $\alpha \geq 0$. Typically, $\alpha = 0$ or 1; for example, $\alpha = 1$ if $T(M)$ is the total Y and $\alpha = 0$ if $T(M)$ is the ratio $R = Y/X$. Deville (1999) used the following asymptotic approximation:

$$\sqrt{n}\, N^{-\alpha}\left[T(\hat{M}) - T(M)\right] \approx \frac{\sqrt{n}}{N} \sum (d_k(s) - 1)\, \tilde{z}_k, \quad (1.17)$$

where $d_k(s) = 0$ if k is not in the sample s. Further $\tilde{z}_k = \mathrm{IT}(M; y_k)$ with IT denoting the influence function of $T(M)$ defined by

$$\mathrm{IT}(M; y) = \lim_{t \to 0} \frac{1}{t} \left[T(M + t\delta_y) - T(M) \right]. \quad (1.18)$$

As noted earlier, it is not easy to justify the approximation (1.17) for general functionals $T(\cdot)$. Deville (1999) developed rules for evaluating $\mathrm{IT}(M; y)$ for selected functionals $T(\hat{M})$. Berger (2002) used the jackknife method to estimate $\tilde{z}_k = \mathrm{IT}(M, y_k)$, similar to Campbell (1980).

Noting that $\Sigma d_k(s) \tilde{z}_k = \hat{Y}(\tilde{z})$ it follows from (1.17) that a linearization variance estimator of $N^{-\alpha} T(\hat{M})$ is given by $N^{-2} v(\tilde{z})$. But $\tilde{z}_k$ depends on unknown parameters and the corresponding estimator, z_k, may not be unique. For example, suppose $T(\hat{M}) = \hat{Y}_R = (\hat{Y}/\hat{X})X$, then $\alpha = 1$ and $\tilde{z}_k = y_k - R x_k$, where $R = Y/X$. In this case, two possible candidates for z_k are $z_k = y_k - \hat{R} x_k$ and $z_k = (X/\hat{X})\ (y_k - \hat{R} x_k)$. Thus, the choice of z_k in the presence of auxiliary information, such as a known total X, is not unique under Deville's approach. Unlike Deville's approach, our method leads to a unique choice z_k and it avoids the calculation of $\tilde{z}_k$ to determine z_k. Our z_k satisfies desirable properties mentioned in section 1, at least in a number of important cases.

2. THE METHOD

To motivate the method, we start with a simple general case where the estimator $\hat{\theta}$ of a parameter θ can be expressed as a smooth function $g(\hat{\boldsymbol{Y}})$ of estimated totals $\hat{\boldsymbol{Y}} = (\hat{Y}_1, ..., \hat{Y}_i, ..., \hat{Y}_m)^T$, where $\hat{Y}_i = \Sigma_{k \in U} d_k(s)\, y_{ik}$, $i = 1, ..., m$, is an estimator of the total $Y_i = \Sigma_{k \in U} y_{ik}$, and $\theta = g(\boldsymbol{Y})$ with $\boldsymbol{Y} = (Y_1, ..., Y_i, ..., Y_m)^T$. We may write $\hat{\theta}$ as $\hat{\theta} = f(\boldsymbol{d}(s), \boldsymbol{A}_y)$ and $\theta = f(\boldsymbol{1}, \boldsymbol{A}_y)$, where $\boldsymbol{A}_y$ is an $m \times N$ matrix with k^{th} column $\boldsymbol{y}_k = (y_{k1}, ..., y_{ki}, ..., y_{km})^T$, $k = 1, ..., N$, $\boldsymbol{d}(s) = (d_1(s), ..., d_N(s))^T$ and $\boldsymbol{1}$ is the N-vector of 1's. For example, if $\hat{\theta}$ denotes the ratio estimator $\hat{Y}_R = [(\Sigma d_k(s)\, y_k)/(\Sigma d_k(s)\, x_k)]\, X$, then $m = 2$, $y_{1k} = y_k$, $y_{2k} = x_k$ and $f(\boldsymbol{1}, \boldsymbol{A}_y)$ reduces to the total Y, noting that $(Y/X)X = Y$. Note that $\hat{Y}_R$ is a function of $\boldsymbol{d}(s), \boldsymbol{y}$ and $\boldsymbol{x}$ and the known total X, but we dropped X for simplicity and write $\hat{Y}_R = f(\boldsymbol{d}(s), \boldsymbol{y}, \boldsymbol{x})$.

Taylor linearization of $\hat{\theta}$ around $\boldsymbol{Y}$ gives the approximation

$$\sqrt{n}\, N^{-\alpha}(\hat{\theta} - \theta) \approx \frac{\sqrt{n}}{N} \left(\partial g(\boldsymbol{a})/\partial \boldsymbol{a}\right)^T \Big|_{\boldsymbol{a} = \boldsymbol{Y}} \left(\hat{\boldsymbol{Y}} - \boldsymbol{Y}\right) \quad (2.1)$$

where $\partial g(\boldsymbol{a})/\partial \boldsymbol{a} = (\partial g(\boldsymbol{a})/\partial a_1, ..., \partial g(\boldsymbol{a})/\partial a_m)^T$ and $N^{-\alpha} g(\cdot)$ tends to a limit for some $\alpha \geq 0$. Asymptotic normality of $\sqrt{n} N^{-\alpha}(\hat{\theta} - \theta)$ follows from (2.1), provided a central limit theorem for $\sqrt{n} N^{-1}(\hat{Y} - Y)$ holds and $g(\cdot)$ has continuous first derivatives in a neighbourhood of the mean $\overline{Y}$. Krewski and Rao (1981) justified (2.1) for stratified sampling.

Let $\check{\boldsymbol{Y}} = \Sigma b_k \boldsymbol{y}_k$ for arbitrary real numbers $\boldsymbol{b} = (b_1, ..., b_N)^T$, and $g(\check{\boldsymbol{Y}}) = f(\boldsymbol{b}, \boldsymbol{A}_y) = f(\boldsymbol{b})$. Noting that $\hat{\boldsymbol{Y}} = \boldsymbol{A}_y \boldsymbol{d}(s)$ and $\boldsymbol{Y} = \boldsymbol{A}_y \boldsymbol{1}$, we can express (2.1) as

$$\sqrt{n} N^{-\alpha}(\hat{\theta} - \theta) \approx \frac{\sqrt{n}}{N} \left(\partial g(\check{\boldsymbol{Y}})/\partial \check{\boldsymbol{Y}}\right)^T \Big|_{\check{\boldsymbol{Y}} = \boldsymbol{Y}} \boldsymbol{A}_y (\boldsymbol{d}(s) - \boldsymbol{1})$$

$$= \frac{\sqrt{n}}{N} \sum_{k=1}^{N} \left(\partial f(\boldsymbol{b})/\partial \check{\boldsymbol{Y}}\right)^T \Big|_{\boldsymbol{b} = \boldsymbol{1}} \boldsymbol{y}_k (d_k(s) - 1), \quad (2.2)$$

noting that $\check{\boldsymbol{Y}} = \boldsymbol{Y}$ is equivalent to $\boldsymbol{b} = \boldsymbol{1}$. Now we substitute $\boldsymbol{y}_k = \partial \check{\boldsymbol{Y}}/\partial b_k \big|_{\boldsymbol{b} = \boldsymbol{1}}$ in (2.2) to get

$$\sqrt{n} N^{-\alpha}(\hat{\theta} - \theta) \approx \frac{\sqrt{n}}{N} \sum_{k=1}^{N} \left(\partial f(\boldsymbol{b})/\partial b_k\right) \Big|_{\boldsymbol{b} = \boldsymbol{1}} (d_k(s) - 1)$$

$$= \frac{\sqrt{n}}{N} \tilde{\boldsymbol{z}}^T (\boldsymbol{d}(s) - \boldsymbol{1}), \quad (2.3)$$

where $\tilde{\boldsymbol{z}} = (\tilde{z}_1, ..., \tilde{z}_N)^T$ with $\tilde{z}_k = \partial f(\boldsymbol{b})/\partial b_k \big|_{\boldsymbol{b} = \boldsymbol{1}}$.

A variance estimator of the right hand side of (2.3) is given by $(n/N^2) v(\tilde{z})$, where $v(\tilde{z})$ is the variance estimator of the estimated total $\Sigma d_k(s) \tilde{z}_k = \hat{Y}(\tilde{z})$. Since $\tilde{z}_k$'s are unknown, we replace $\tilde{z}_k$ by $z_k = \partial f(\boldsymbol{b})/\partial b_k \big|_{\boldsymbol{b} = \boldsymbol{d}(s)}$, to get $(n/N^2) v(z)$. Thus, a linearization variance estimator of $\hat{\theta}$ is given by

$$v_L(\hat{\theta}) = (N^{2\alpha}/N^2) v(z), \quad (2.4)$$

which reduces to $v(z)$ if $\alpha = 1$. Note that $v_L(\hat{\theta})$ given by (2.4) is simply obtained from the formula $v(y)$ for $\hat{Y}$ by replacing y_k by z_k for $k \in s$. Note that we do not first

evaluate the partial derivatives $\partial f(\boldsymbol{b})/\partial b_k$ at $\boldsymbol{b} = \mathbf{1}$ to get $\tilde{z}$ and then substitute estimates for the unknown components of $\tilde{z}$. Our method, therefore, is similar in spirit to Binder's approach. The variance estimator $v_L(\hat{\theta})$ is valid because z_k is a consistent estimator of $\tilde{z}_k$.

Example 2.1 Suppose $\hat{\theta}$ is the ratio estimator $\hat{Y}_R = X[(\Sigma d_k(s) y_k)/(\Sigma d_k(s) x_k)]$ of the total Y. Then $f(\boldsymbol{b}) = X[(\Sigma b_k y_k)/(\Sigma b_k x_k)]$ and

$$\partial f(\boldsymbol{b})/\partial b_k = X\frac{y_k\sum b_k x_k - x_k\sum b_k y_k}{\left(\sum b_k x_k\right)^2}.$$

Therefore,

$$z_k = \partial f(\boldsymbol{b})/\partial b_k\big|_{b=d(s)} = \frac{X}{\hat{X}}\left(y_k - \hat{R}x_k\right)$$

which agrees with (1.3). Thus, our variance estimator $v_L(\hat{Y}_R)$ is identical to Binder's (1996) variance estimator, $v(z)$, noting that $\alpha = 1$.

Our derivation is simple and natural. On the other hand, in the standard linearization method, $\hat{\theta}$ is first expressed in terms of elementary components $\hat{Y}_1, ..., \hat{Y}_m$ as $g(\hat{\boldsymbol{Y}})$ and the partial derivatives $\partial g(\boldsymbol{a})/\partial a_j$ are then evaluated at $\boldsymbol{a} = \boldsymbol{Y}$. It is interesting to note that all the components of $\hat{\boldsymbol{Y}}$ use the same weights $d_k(s)$ and our approach always takes first derivatives of $f(\boldsymbol{b})$ with respect to b_k at $\boldsymbol{b} = \boldsymbol{d}(s)$. It is not necessary to first express $\hat{\theta}$ in terms of elementary components.

3. CALIBRATION ESTIMATORS

The ratio estimator can be viewed as a calibration estimator, $\hat{Y}_R = \Sigma w_k(s) y_k$, with explicit weights $w_k(s) = (X/\hat{X}) d_k(s)$ and satisfying the calibration constraint $\Sigma w_k(s) x_k = X$. Calibration estimators of a total Y of the form $\hat{Y}_w = \Sigma w_k(s) y_k$ with explicit weights $w_k(s)$ and satisfying the calibration constraints $\Sigma w_k(s)\boldsymbol{x}_k = \boldsymbol{X}$ are widely used, where $\boldsymbol{x}_k = (x_{1k}, ..., x_{qk})^T$ and $\boldsymbol{X} = (X_1, ..., X_q)^T$ is the vector of known totals of auxiliary variables x_j, $j = 1, ..., q$. In subsection 3.1 we consider the generalized regression (GREG) estimator and then study a general class of regression calibration estimators in subsection 3.2. Extension to estimators, $\hat{\theta}$, obtained as solutions of estimating equations is presented in subsection 3.3. The case of general calibration estimators is investigated in subsection 3.4.

3.1 Generalized Regression Estimator

The GREG estimator of total Y is given by $\hat{Y}_w$ with calibration weights $w_k(s) = d_k(s) g_k(\boldsymbol{d}(s))$, where

$$g_k(\boldsymbol{d}(s)) = 1 + (\boldsymbol{X} - \hat{\boldsymbol{X}})^T\left(\Sigma d_k(s) c_k \boldsymbol{x}_k \boldsymbol{x}_k^T\right)^{-1} c_k \boldsymbol{x}_k \quad (3.1)$$

with specified constants c_k and $\hat{\boldsymbol{X}} = \Sigma d_k(s)\boldsymbol{x}_k$ (*cf.*, Särndal *et al.* 1989). The ratio estimator, $\hat{Y}_R$, is a special case with $q=1$ (*i.e.*, scalar x_k) and $c_k = x_k^{-1}$, and $g_k(\boldsymbol{d}(s))$, given by (3.1), reduces to $X/\hat{X}$.

The GREG estimator may be expressed as a differentiable function of estimated totals. Hence, the general theory of section 2 is applicable and it remains to evaluate $z_k = \partial f(\boldsymbol{b})/\partial b_k|_{b=d(s)}$, where $f(\boldsymbol{b}) = \Sigma(b_k g_k(\boldsymbol{b})) y_k$ is obtained by replacing $\boldsymbol{d}(s)$ by $\boldsymbol{b}$ in the formula for $\hat{Y}_w$. Noting that $\partial \boldsymbol{A}(\boldsymbol{b})^{-1}/\partial b_k = -\boldsymbol{A}(\boldsymbol{b})^{-1}(\partial \boldsymbol{A}(\boldsymbol{b})/\partial b_k)\boldsymbol{A}(\boldsymbol{b})^{-1}$, where $\boldsymbol{A}(\boldsymbol{b}) = \Sigma b_k c_k \boldsymbol{x}_k \boldsymbol{x}_k^T$, we get

$$\begin{aligned}\partial(b_k g_k(\boldsymbol{b}))/\partial b_k &= g_k(\boldsymbol{b}) - \boldsymbol{x}_k^T \boldsymbol{A}(\boldsymbol{b})^{-1} b_k c_k \boldsymbol{x}_k \\ &\quad - (\boldsymbol{X} - \hat{\boldsymbol{X}}(\boldsymbol{b}))^T \boldsymbol{A}(\boldsymbol{b})^{-1}(c_k \boldsymbol{x}_k \boldsymbol{x}_k^T)\boldsymbol{A}(\boldsymbol{b})^{-1}(b_k c_k \boldsymbol{x}_k) \quad (3.2)\end{aligned}$$

and for $l \neq k$

$$\begin{aligned}\partial(b_l g_l(\boldsymbol{b}))/\partial b_k &= -\boldsymbol{x}_k^T \boldsymbol{A}(\boldsymbol{b})^{-1}(b_l c_l \boldsymbol{x}_l) \\ &\quad - (\boldsymbol{X} - \hat{\boldsymbol{X}}(\boldsymbol{b}))^T \boldsymbol{A}(\boldsymbol{b})^{-1}(c_k \boldsymbol{x}_k \boldsymbol{x}_k^T)\boldsymbol{A}(\boldsymbol{b})^{-1}(b_l c_l \boldsymbol{x}_l). \quad (3.3)\end{aligned}$$

It now follows from (3.2) and (3.3), that

$$\partial f(\boldsymbol{b})/\partial b_k = g_k(\boldsymbol{b})\, e_k(\boldsymbol{b}), \quad (3.4)$$

where

$$e_k(\boldsymbol{b}) = y_k - \boldsymbol{x}_k^T \boldsymbol{B}(\boldsymbol{b}) \quad (3.5)$$

with $\boldsymbol{B}(\boldsymbol{b}) = \boldsymbol{A}^{-1}(\boldsymbol{b})(\Sigma_k b_k c_k \boldsymbol{x}_k y_k)$. Therefore, $z_k = \partial f(\boldsymbol{b})/\partial b_k|_{b=d(s)}$ reduces to

$$z_k = g_k(\boldsymbol{d}(s))\, e_k, \quad (3.6)$$

where $e_k = y_k - \boldsymbol{x}_k^T \hat{\boldsymbol{B}}$ with $\hat{\boldsymbol{B}} = \boldsymbol{B}(\boldsymbol{d}(s))$.

The variance estimator of $\hat{Y}_w$, resulting from (3.6), namely $v(z)$, takes account of the g-weights, $g_k(\boldsymbol{d}(s))$, unlike the standard linearization variance estimator (see *e.g.*, Särndal *et al.* 1991, page 237). It agrees with the model-assisted variance estimator of Särndal *et al.* (1989). It also agrees with the jackknife linearization variance estimator when the latter is applicable (Yung and Rao 1996).

3.2 A General Class of Regression Calibration Weights

We now turn to a general class of regression calibration weights of the form $w_k(s) = d_k(s)\, h_k(\boldsymbol{d}(s))$ with

$$h_k(\boldsymbol{d}(s)) = 1 + (\boldsymbol{X} - \hat{\boldsymbol{X}})^T \hat{\boldsymbol{Q}}^{-1}\left(c_k \boldsymbol{x}_k + \textstyle\sum_{l\neq k} d_l(s) c_{kl} \boldsymbol{x}_l\right), \quad (3.7)$$

where the ab-th element of $\hat{\boldsymbol{Q}}$ is given by

$$\hat{q}_{ab} = \textstyle\sum_{k=1}^N d_k(s) c_k x_{ak} x_{bk} + \sum_{k=1}^N \sum_{l\neq k}^N d_k(s) d_l(s) c_{kl} x_{ak} x_{bl}$$

for specified constants c_k and $c_{kl}(=c_{lk})$. The class (3.7) covers the GREG estimator as well as the "optimal" linear regression estimator with $d_k(s) = (1/\pi_k)a_k(s)$. In the former case $c_{kl} = 0$ while the optimal linear regression estimator uses $c_k = (1-\pi_k)/\pi_k$ and $c_{kl} = (\pi_{kl} - \pi_k\pi_l)/\pi_{kl}$, $k \neq l$, where π_{kl} is the probability of including both elements k and l in the sample s (Montanari 1998).

The calibration weights $w_k(s)$ may be rewritten as

$$w_k(s) = d_k(s) + (\boldsymbol{X} - \hat{\boldsymbol{X}})^T \hat{\boldsymbol{Q}}^{-1} \left(d_k(s) c_k \boldsymbol{x}_k + \sum_{l \neq k} d_{kl}(s) c_{kl}^* \boldsymbol{x}_l \right), \quad (3.8)$$

where $d_{kl}(s) = d_k(s) d_l(s) / E[d_k(s) d_l(s)]$, $c_{kl}^* = c_{kl} E[d_k(s) d_l(s)]$ and

$$\hat{q}_{ab} = \sum_{k=1}^N d_k(s) c_k x_{ak} x_{bk} + \sum_{k=1}^N \sum_{l \neq k}^N d_{kl}(s) c_{kl}^* x_{ak} x_{bl}.$$

Note that $E\,d_k(s) = 1$ and $E\,d_{kl}(s) = 1$. If $d_k(s) = (1/\pi_k)a_k(s)$ then $d_{kl}(s)$ reduces to $d_{kl}(s) = a_k(s)a_l(s)/\pi_{kl}$ and $c_{kl}^* = (\pi_{kl} - \pi_k\pi_l)/(\pi_k\pi_l)$. We can regard the calibration estimator $\hat{Y}_w$ resulting from (3.8) as a function of totals, by expressing a quadratic form as a total of synthetic variables (Sitter and Wu 2002). Therefore, we can use the method of section 2 and write $\hat{Y}_w = f(\boldsymbol{d}^{(1)}(s), \boldsymbol{d}^{(2)}(s), \boldsymbol{y}) = \sum d_k(s)\, h(\boldsymbol{d}^{(1)}(s), \boldsymbol{d}^{(2)}(s))\, y_k$ where $\boldsymbol{d}^{(1)}(s) = \boldsymbol{d}(s)$ and $\boldsymbol{d}^{(2)}(s)$ is the vector of elements $d_{kl}(s)$, $k < l$, arranged in a sequence. Now, following the derivation of (2.3), we get

$$\hat{Y}_w - Y \approx \sum_k \tilde{z}_k (d_k(s) - 1) + 2 \sum \sum_{k<l} \tilde{z}_{kl} (d_{kl}(s) - 1) \quad (3.9)$$

where

$$\tilde{z}_k = \partial f(\boldsymbol{b}^{(1)}, \boldsymbol{b}^{(2)}, \boldsymbol{y}) / \partial b_k \big|_{\boldsymbol{b}^{(1)} = 1, \boldsymbol{b}^{(2)} = 1},$$

$$\tilde{z}_{kl} = \partial f(\boldsymbol{b}^{(1)}, \boldsymbol{b}^{(2)}, \boldsymbol{y}) / \partial b_{kl} \big|_{\boldsymbol{b}^{(1)} = 1, \boldsymbol{b}^{(2)} = 1},$$

$\boldsymbol{b}^{(1)} = \boldsymbol{b} = (b_1, ..., b_N)^T$ and $\boldsymbol{b}^{(2)}$ is the vector of arbitrary real numbers b_{kl}, $k < l$, arranged in the same order as the elements $d_{kl}(s)$ in $\boldsymbol{d}^{(2)}(s)$. Using (3.9), a variance estimator of $\hat{Y}_w$ is approximately given by the variance estimator of $\sum_k \tilde{z}_k d_k(s) + 2\sum\sum_{k<l} \tilde{z}_{kl} d_{kl}(s)$, denoted by $v(\tilde{\boldsymbol{z}}^{(1)}, \tilde{\boldsymbol{z}}^{(2)})$.

Since $v(\tilde{\boldsymbol{z}}^{(1)}, \tilde{\boldsymbol{z}}^{(2)})$ involves the unknown values $\tilde{z}_k$ and $\tilde{z}_{kl}$, we replace $\tilde{z}_k$ by $z_k = \partial f(\boldsymbol{b}^{(1)}, \boldsymbol{b}^{(2)}, \boldsymbol{y}) / \partial b_k \big|_{\boldsymbol{b}^{(1)} = \boldsymbol{d}^{(1)}(s), \boldsymbol{b}^{(2)} = \boldsymbol{d}^{(2)}(s)}$ and $\tilde{z}_{kl}$ by $z_{kl} = \partial f(\boldsymbol{b}^{(1)}, \boldsymbol{b}^{(2)}, \boldsymbol{y}) / \partial b_{kl} \big|_{\boldsymbol{b}^{(1)} = \boldsymbol{d}^{(1)}(s), \boldsymbol{b}^{(2)} = \boldsymbol{d}^{(2)}(s)}$ to get $v(\boldsymbol{z}^{(1)}, \boldsymbol{z}^{(2)})$. Unfortunately, the variance estimator $v(\boldsymbol{z}^{(1)}, \boldsymbol{z}^{(2)})$ involves third order and fourth order moments $E[d_k(s) d_l(s) d_q(s)]$ and $E[d_k(s) d_l(s) d_q(s) d_r(s)]$ in addition to the second moments $E[d_k(s) d_l(s)]$, whereas the variance estimator for the generalized regression estimator requires only the second moments. In particular, if $d_k(s) = (1/\pi_k)a_k(s)$ we required third and fourth order inclusion probabilities π_{klq} and π_{klqr} as well as the second order inclusion probabilities π_{kl}.

The calculation of z_k and z_{kl} involves the derivatives $\partial[b_l h(\boldsymbol{b}^{(1)}, \boldsymbol{b}^{(2)})] / \partial b_k$ for $l = k$ and $l \neq k$ and the derivatives $\partial[b_l h(\boldsymbol{b}^{(1)}, \boldsymbol{b}^{(2)})] / \partial b_{kl}$ for $l = k$ and $l \neq k$. After simplification, we get

$$z_k = \left[1 + (\boldsymbol{X} - \hat{\boldsymbol{X}})^T \hat{\boldsymbol{Q}}^{-1} c_k \boldsymbol{x}_k \right] e_k^*$$

and

$$z_{kl} = (\boldsymbol{X} - \hat{\boldsymbol{X}})^T \hat{\boldsymbol{Q}}^{-1} c_{kl}^* \boldsymbol{x}_l e_k^*,$$

where

$$e_k^* = y_k - \boldsymbol{x}_k^T \hat{\boldsymbol{B}}^*$$

with $\hat{\boldsymbol{B}}^* = \hat{\boldsymbol{Q}}^{-1}(\sum_k d_k(s) c_k \boldsymbol{x}_k y_k + \sum_k \sum_{l \neq k} d_k(s) d_l(s) c_{kl} \boldsymbol{x}_l y_k)$. Note that the customary Taylor linearization variance estimation uses $v(e^*)$, while $v(\boldsymbol{z}^{(1)}, \boldsymbol{z}^{(2)})$ would involve the residuals e_k^* as well as the g-weights $1 + (\boldsymbol{X} - \hat{\boldsymbol{X}})^T \hat{\boldsymbol{Q}}^{-1} c_k \boldsymbol{x}_k$ and $(\boldsymbol{X} - \hat{\boldsymbol{X}})^T \hat{\boldsymbol{Q}}^{-1} c_{kl}^* \boldsymbol{x}_l$. If $c_{kl} = 0$ for all $k \neq l$, then $z_{kl} = 0$ and $v(\boldsymbol{z}^{(1)}, \boldsymbol{z}^{(2)})$ reduces to $v(\boldsymbol{z})$ with z_k given by (3.6). Thus the GREG result of subsection 3.1 is a special case.

3.3 Estimating Equations

We now turn to a vector parameter $\boldsymbol{\theta} = (\theta_1, ..., \theta_p)^T$ defined either explicitly or implicitly as the solution to "census" estimating equations $\boldsymbol{S}(\boldsymbol{\theta}) = \sum_{k=1}^N \boldsymbol{u}_k(\boldsymbol{\theta}) = \boldsymbol{0}$. A calibration estimator $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, ..., \hat{\theta}_p)^T$ with GREG calibration weights $w_k(s) = d_k(s)\, g_k(\boldsymbol{d}(s))$ is obtained as the solution to sample estimating equations:

$$\hat{\boldsymbol{S}}(\hat{\boldsymbol{\theta}}) = \sum w_k(s) \boldsymbol{u}_k(\hat{\boldsymbol{\theta}}) = \boldsymbol{0}, \quad (3.10)$$

where $\boldsymbol{u}_k(\hat{\boldsymbol{\theta}})$ and $\hat{\boldsymbol{S}}(\hat{\boldsymbol{\theta}})$ are $(p \times 1)$ vectors (Binder 1983). For example for logistic regression with scalar θ, we have $u_k(\theta) = (y_k - p_k(\theta))a_k$, where $p_k(\theta) = P(y_k = 1 | a_k) = \exp(\theta a_k)/(1 + \exp(\theta a_k))$ and a_k is the predictor variable. Note that $\hat{\theta}$, in this case, is the implicit solution to (3.10) and obtained iteratively using Newton-Raphson or Fisher scoring method.

The estimator of a ratio of totals Y and $A = \sum a_k$ is obtained as the explicit solution of (3.10) with $u_k(\theta) = y_k - \theta a_k$: $\hat{\theta} = \sum w_k(s) y_k / \sum w_k(s) a_k = \hat{Y}/\hat{A}$. In this case, $\hat{\theta}$ is a function of estimated totals and hence our method for functions of totals is applicable. It remains to evaluate $\partial f(\boldsymbol{b}) / \partial b_k$, where $f(\boldsymbol{b}) = \sum b_k g_k(\boldsymbol{b}) y_k / \sum b_k g_k(\boldsymbol{b}) a_k$. We have

$$\partial f(\boldsymbol{b}) / \partial b_k = \sum_{l=1}^N [\partial(b_l g_l(\boldsymbol{b})) / \partial b_k]\, \hat{A}(\boldsymbol{b})^{-1} (y_l - f(\boldsymbol{b}) a_l),$$

where $\hat{A}(\boldsymbol{b}) = \sum b_l g_l(\boldsymbol{b})\, a_l$. Now using (3.4) and (3.5), it is easy to verify that z_k reduces to

$$z_k = g_k(\boldsymbol{d}(s))\, \hat{A}^{-1} e_k^*$$

where

$$e_k^* = u_k(\hat{\theta}) - \boldsymbol{x}_k^T \hat{\boldsymbol{B}}_u$$

with $\hat{\boldsymbol{B}}_u$ obtained from $\hat{\boldsymbol{B}}$ by changing y_k to $u_k(\hat{\boldsymbol{\theta}})$. Note that the residuals e_k^* has the same form as the GREG residuals e_k with y_k changed with $u_k(\hat{\boldsymbol{\theta}})$.

In general, the solution $\hat{\boldsymbol{\theta}}$ to the estimating equations (3.10) may not be expressable as a function of estimated totals. We therefore follow Binder's (1983) approach and write the linearization estimator of the covariance matrix of $\hat{\boldsymbol{\theta}}$ as

$$v_L(\hat{\boldsymbol{\theta}}) = \left[\hat{\boldsymbol{J}}(\hat{\boldsymbol{\theta}})\right]^{-1} \hat{\boldsymbol{\Sigma}}_S(\hat{\boldsymbol{\theta}}) \left[\hat{\boldsymbol{J}}(\hat{\boldsymbol{\theta}})\right]^{-1}, \tag{3.11}$$

where $\hat{\boldsymbol{J}}(\boldsymbol{\theta}) = -\partial \hat{\boldsymbol{S}}(\boldsymbol{\theta})/\partial \boldsymbol{\theta}$ and $\hat{\boldsymbol{\Sigma}}_S(\hat{\boldsymbol{\theta}})$ is the estimated covariance matrix $v_L(\hat{\boldsymbol{S}}(\boldsymbol{\theta})) = \hat{\boldsymbol{\Sigma}}_S(\boldsymbol{\theta})$ evaluated at $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$. Binder (1983) gave regularity conditions for the validity of (3.11). Noting that $\hat{\boldsymbol{S}}(\boldsymbol{\theta})$ is a vector of estimated totals with GREG weights $d_k(s)\, g_k(\boldsymbol{d}(s))$, it follows from (3.6) and (3.11) that

$$v_L(\hat{\boldsymbol{\theta}}) = v(z) \tag{3.12}$$

where

$$z_k = \left[\hat{\boldsymbol{J}}(\hat{\boldsymbol{\theta}})\right]^{-1} g_k(\boldsymbol{d}(s)) \boldsymbol{e}_k^* \tag{3.13}$$

with $\boldsymbol{e}_k^* = (e_{k1}^*, \ldots, e_{kp}^*)^T$ and

$$e_{kj}^* = u_{jk}(\hat{\boldsymbol{\theta}}) - \boldsymbol{x}_k \hat{\boldsymbol{B}}_{ju};\ j = 1, \ldots, p.$$

Further, $\hat{\boldsymbol{B}}_{ju}$ is obtained from $\hat{\boldsymbol{B}}_j$ by changing y_k to $u_{jk}(\hat{\boldsymbol{\theta}})$ and $v(z)$ is the estimated covariance matrix of the vector of estimated totals $\hat{\boldsymbol{Z}} = \Sigma d_k(s) z_k$, where $u_{jk}(\hat{\boldsymbol{\theta}})$ is the j^{th} element of $\boldsymbol{u}_k(\hat{\boldsymbol{\theta}})$. The result (3.12) agrees with the jackknife linearization variance estimator, v_{JL}, for stratified multistage sampling obtained by Rao, Yung and Hidiroglou (2002).

The result (3.12)-(3.13) may also be obtained directly by writing $\hat{\boldsymbol{\theta}}$ as $f(\boldsymbol{d}(s))$ and evaluating $z_k = \partial f(\boldsymbol{b})/\partial b_k|_{\boldsymbol{b}=\boldsymbol{d}(s)}$. We denote $\hat{\boldsymbol{\theta}}(\boldsymbol{b}) = f(\boldsymbol{b})$ as the solution of $\Sigma (b_k g_k(\boldsymbol{b}))\, \boldsymbol{u}_k(\boldsymbol{\theta}) = \boldsymbol{0}$, *i.e.*,

$$\sum (b_k g_k(\boldsymbol{b}))\, \boldsymbol{u}_k(\hat{\boldsymbol{\theta}}(\boldsymbol{b})) = \boldsymbol{0}, \tag{3.14}$$

We now take the derivative of (3.14) with respect to b_k to get

$$\sum_{l=1}^{N} \left[\partial (b_l g_l(\boldsymbol{b}))/\partial b_k\right] \boldsymbol{u}_l(\hat{\boldsymbol{\theta}}(\boldsymbol{b})) + \sum_{l=1}^{N} (b_l g_l(\boldsymbol{b})) \left[\partial \boldsymbol{u}_l(\hat{\boldsymbol{\theta}}(\boldsymbol{b}))/\partial(\hat{\boldsymbol{\theta}}(\boldsymbol{b}))\right] \partial(\hat{\boldsymbol{\theta}}(\boldsymbol{b}))/\partial b_k. \tag{3.15}$$

Substituting (3.2) and (3.3) for $\partial (b_l g_l(\boldsymbol{b}))/\partial b_k$ in (3.15), we obtain (3.13) after simplification. This result shows that our method is also directly applicable to general estimators $\hat{\boldsymbol{\theta}}$ under Binder's (1983) regularity conditions.

3.4 A General Class of Calibration Estimators

The calibration weights, $w_k(s)$, associated with the GREG estimator $\hat{Y}_w$ may not be always nonnegative. To get around this difficulty, generalized raking ratio weights are often used. These weights are always nonnegative, but the method can lead to some extreme weights (Deville and Särndal 1992).

The generalized raking weights belong to the class

$$w_k(s) = d_k(s)\, F(\boldsymbol{x}_k^T \hat{\boldsymbol{\lambda}}) \tag{3.16}$$

with $F(a) = e^a$, where the LaGrange multiplier $\hat{\boldsymbol{\lambda}}$ is determined by solving the calibration equations

$$\sum w_k(s) \boldsymbol{x}_k = \sum d_k(s) F(\boldsymbol{x}_k^T \hat{\boldsymbol{\lambda}}) \boldsymbol{x}_k = \boldsymbol{X}. \tag{3.17}$$

The GREG weights correspond to $F(a) = 1 + a$ in which case $\hat{\boldsymbol{\lambda}} = (\Sigma d_k(s) \boldsymbol{x}_k \boldsymbol{x}_k^T)^{-1}(\boldsymbol{X} - \hat{\boldsymbol{X}})$.

In general, the calibration estimator $\hat{Y}_w = \Sigma w_k(s)\, y_k$ with weights $w_k(s)$ given by (3.16) may not be expressable as a function of estimated totals. We therefore follow Binder's (1983) approach and expand $F(\boldsymbol{x}_k^T \hat{\boldsymbol{\lambda}})$ around $\boldsymbol{\lambda}$, where $\boldsymbol{\lambda}$ denotes the probability limit of $\hat{\boldsymbol{\lambda}}$. We get

$$F(\boldsymbol{x}_k^T \hat{\boldsymbol{\lambda}}) \approx F(\boldsymbol{x}_k^T \boldsymbol{\lambda}) + f(\boldsymbol{x}_k^T \boldsymbol{\lambda}) \boldsymbol{x}_k^T (\hat{\boldsymbol{\lambda}} - \boldsymbol{\lambda}), \tag{3.18}$$

where $f(a) = \partial F(a)/\partial a$. Further, by expanding the calibration equations (3.17) around $\boldsymbol{\lambda}$, we obtain after simplification,

$$\hat{\boldsymbol{\lambda}} - \boldsymbol{\lambda} \approx -\hat{\boldsymbol{Q}}_\lambda^{-1} \left(\hat{\boldsymbol{S}}_\lambda - \boldsymbol{X}\right) \tag{3.19}$$

where $\hat{\boldsymbol{Q}}_\lambda = \Sigma d_k(s) f(\boldsymbol{x}_k^T \boldsymbol{\lambda}) \boldsymbol{x}_k \boldsymbol{x}_k^T$ and $\hat{\boldsymbol{S}}_\lambda = \Sigma_k d_k(s)$ $F(\boldsymbol{x}_k^T \boldsymbol{\lambda}) \boldsymbol{x}_k$. Note that both $\hat{\boldsymbol{Q}}_\lambda$ and $\hat{\boldsymbol{S}}_\lambda$ are of the form of estimated totals. Substituting (3.19) into (3.18) gives

$$F(\boldsymbol{x}_k^T \hat{\boldsymbol{\lambda}}) \approx F(\boldsymbol{x}_k^T \boldsymbol{\lambda}) - f(\boldsymbol{x}_k^T \boldsymbol{\lambda}) \boldsymbol{x}_k^T \hat{\boldsymbol{Q}}_\lambda^{-1} \left(\hat{\boldsymbol{S}}_\lambda - \boldsymbol{X}\right). \tag{3.20}$$

Using the approximation (3.20) in (3.16), it follows that $\hat{Y}_w$ is approximated by a differentiable function of estimated totals. Hence, the general theory of section 2 is applicable and it remains to evaluate $z_k = \partial h(\boldsymbol{b})/\partial b_k|_{\boldsymbol{b}=\boldsymbol{d}(s)}$, where $h(\boldsymbol{b}) = \Sigma b_k g_k^*(\boldsymbol{b}) y_k$ with

$$g_k^*(\boldsymbol{b}) = F(\boldsymbol{x}_k^T \boldsymbol{\lambda}) - f(\boldsymbol{x}_k^T \boldsymbol{\lambda}) \boldsymbol{x}_k^T \boldsymbol{Q}_\lambda(\boldsymbol{b})^{-1} \left(\boldsymbol{S}_\lambda(\boldsymbol{b}) - \boldsymbol{X}\right)$$

where $\boldsymbol{Q}_\lambda(\boldsymbol{b}) = \Sigma b_k f(\boldsymbol{x}_k^T \boldsymbol{\lambda}) \boldsymbol{x}_k \boldsymbol{x}_k^T$ and $\boldsymbol{S}_\lambda(\boldsymbol{b}) = \Sigma b_k F(\boldsymbol{x}_k^T \boldsymbol{\lambda}) \boldsymbol{x}_k$. After simplification, we get

$$z_k = F(\boldsymbol{x}_k^T \hat{\boldsymbol{\lambda}})(y_k - \boldsymbol{x}_k^T \hat{\boldsymbol{B}}_\lambda) = F(\boldsymbol{x}_k^T \hat{\boldsymbol{\lambda}}) e_{k\lambda}, \tag{3.21}$$

where

$$\hat{\boldsymbol{B}}_\lambda = \left(\sum d_k(s) f(\boldsymbol{x}_k^T \hat{\boldsymbol{\lambda}}) \boldsymbol{x}_k \boldsymbol{x}_k^T\right)^{-1} \sum d_k(s) f(\boldsymbol{x}_k^T \hat{\boldsymbol{\lambda}}) \boldsymbol{x}_k y_k.$$

Singh and Folsom (2000) obtained a similar result, using a somewhat different approach.

The result (3.21) may also be obtained directly along the lines of (3.2) and (3.3) by writing $\hat{Y}_w$ as $f(\boldsymbol{d}(s))$ and evaluating $z_k = \partial f(\boldsymbol{b})/\partial b_k|_{\boldsymbol{b}=\boldsymbol{d}(s)}$, where $f(\boldsymbol{b}) = \Sigma b_k g_k(\boldsymbol{b}) y_k$ with $g_k(\boldsymbol{b}) = F(\boldsymbol{x}_k^T \hat{\boldsymbol{\lambda}}(\boldsymbol{b}))$. We have

$$\partial(b_k g_k(\boldsymbol{b}))/\partial b_k = g_k(\boldsymbol{b}) + b_k f(\boldsymbol{x}_k^T\hat{\boldsymbol{\lambda}}(\boldsymbol{b}))\boldsymbol{x}_k^T(\partial\hat{\boldsymbol{\lambda}}(\boldsymbol{b})/\partial b_k), \quad (3.22)$$

and for $l \neq k$

$$\partial(b_l g_l(\boldsymbol{b}))/\partial b_k = b_l f(\boldsymbol{x}_l^T\hat{\boldsymbol{\lambda}}(\boldsymbol{b}))\boldsymbol{x}_l^T(\partial\hat{\boldsymbol{\lambda}}(\boldsymbol{b})/\partial b_k). \quad (3.23)$$

To evaluate $\partial\hat{\boldsymbol{\lambda}}(\boldsymbol{b})/\partial b_k$, we take the derivatives of the calibration equations (3.17) with $\boldsymbol{d}(s)$ replaced by $\boldsymbol{b}$: $\sum b_k F(\boldsymbol{x}_k^T\hat{\boldsymbol{\lambda}}(\boldsymbol{b}))\boldsymbol{x}_k - \boldsymbol{X} = \boldsymbol{0}$. This gives

$$\boldsymbol{0} = F(\boldsymbol{x}_k^T\hat{\boldsymbol{\lambda}}(\boldsymbol{b}))\boldsymbol{x}_k + \sum_l b_l f(\boldsymbol{x}_l^T\hat{\boldsymbol{\lambda}}(\boldsymbol{b}))\boldsymbol{x}_l\boldsymbol{x}_l^T(\partial\hat{\boldsymbol{\lambda}}(\boldsymbol{b})/\partial b_k)$$

or

$$\partial\hat{\boldsymbol{\lambda}}(\boldsymbol{b})/\partial b_k = -\Big(\sum b_k f(\boldsymbol{x}_k^T\hat{\boldsymbol{\lambda}}(\boldsymbol{b}))\boldsymbol{x}_k\boldsymbol{x}_k^T\Big)^{-1}F(\boldsymbol{x}_k^T\hat{\boldsymbol{\lambda}}(\boldsymbol{b}))\boldsymbol{x}_k. \quad (3.24)$$

Substituting (3.24) into (3.22) and (3.23), we get (3.21) after simplification.

Deville and Särndal (1992) showed that the asymptotic variance of $\hat{Y}_w$ for general $F(\cdot)$ is equivalent to the asymptotic variance of the GREG estimator which involves the "census" regression coefficient $\boldsymbol{B}$. Using this result they obtained a variance estimator of $\hat{Y}_w$ for general $F(\cdot)$, by replacing $\boldsymbol{B}$ by $\hat{\boldsymbol{B}} = (\sum w_k(s)\boldsymbol{x}_k\boldsymbol{x}_k^T)^{-1}\sum w_k(s)\boldsymbol{x}_k y_k$, where $w_k(s) = d_k(s)F(\boldsymbol{x}_k^T\hat{\boldsymbol{\lambda}})$. The resulting z_k agrees with our z_k given by (3.21) if $f(a) = F(a)$, *i.e.*, in the case of generalized raking weights. In the case of GREG estimator, we have $F(x) = 1 + x$, $f(x) = 1$ and $\hat{\boldsymbol{\lambda}} = (\sum d_k(s)\boldsymbol{x}_k\boldsymbol{x}_k^T)^{-1}(\boldsymbol{X} - \hat{\boldsymbol{X}})$. It readily follows that $F(\boldsymbol{x}_k^T\hat{\boldsymbol{\lambda}})$ reduces to the customary g-weight $g_k(\boldsymbol{d}(s)) = 1 + (\boldsymbol{X} - \hat{\boldsymbol{X}})^T(\sum d_k(s)\boldsymbol{x}_k\boldsymbol{x}_k^T)^{-1}\boldsymbol{x}_k$, and $e_{k\lambda} = y_k - \boldsymbol{x}_k^T\hat{\boldsymbol{B}}_\lambda$ reduces to $e_k = y_k - \boldsymbol{x}_k^T\hat{\boldsymbol{B}}$ with $\hat{\boldsymbol{B}} = (\sum d_k(s)\boldsymbol{x}_k\boldsymbol{x}_k^T)^{-1}\sum d_k(s)\boldsymbol{x}_k y_k$. Note that our z_k in this case is different from the z_k of Deville and Särndal (1992), but agrees with a commonly used z_k (Särndal, Swensson and Wretman 1989).

Our method, along the lines of section 3.3, can be extended to implicitly defined estimators, $\hat{\theta}_w$, obtained as solutions to estimating equations (3.10) based on the general calibration weights (3.16). Details are omitted for simplicity.

4. TWO-PHASE SAMPLING

We extend our method to two-phase sampling, assuming the estimator $\hat{\theta}$ of a parameter θ can be expressed as a differentiable function, $g(\hat{\boldsymbol{Y}}, \hat{\boldsymbol{X}}^{(1)})$, of estimated totals, $\hat{\boldsymbol{Y}} = (\hat{Y}_1, ..., \hat{Y}_m)^T$, from the second-phase sample and estimated totals, $\hat{\boldsymbol{X}}^{(1)} = (\hat{X}_1^{(1)}, ..., \hat{X}_p^{(1)})^T$, from the first-phase sample only. Here $\hat{Y}_i = \sum_{k=1}^N d_k(s)y_{ik}$, $i = 1, ..., m$, $\hat{X}_j^{(1)} = \sum_{k=1}^N d_k^{(1)}(s_1)x_{jk}$, $j = 1, ..., p$, $d_k^{(1)}(s_1)$ denotes the first-phase design weight attached to the k^{th} element with $d_k(s_1) = 0$ if k is not in the first-phase sample s_1, and $d_k(s)$ is the final design weight attached to the k^{th} element with $d_k(s) = 0$ if k is not in the second-phase sample s. Further, the parameter $\theta = g(\boldsymbol{Y}, \boldsymbol{X})$ with $\boldsymbol{Y} = (Y_1, ..., Y_m)^T$ and $\boldsymbol{X} = (X_1, ..., X_p)^T$ denoting the vectors of Y- and X- totals. For example, the two-phase ratio estimator, $\hat{Y}_{R2}$, is of the form $\hat{\theta} = g(\hat{Y}, \hat{X}, \hat{X}^{(1)})$:

$$\hat{Y}_{R2} = \frac{\hat{Y}}{\hat{X}}\hat{X}^{(1)} = \hat{R}\hat{X}^{(1)}$$

$$= \frac{\sum d_k(s)y_k}{\sum d_k(s)x_k}\Big(\sum d_k^{(1)}(s_1)x_k\Big). \quad (4.1)$$

Note that $\hat{\boldsymbol{Y}} = (\hat{Y}_1, \hat{Y}_2)^T$ with $\hat{Y}_1 = \hat{Y}$, $\hat{Y}_2 = \hat{X}$, and $\hat{\boldsymbol{X}}^{(1)} = \hat{X}^{(1)}$. Also, $\theta = g(Y, X, X^{(1)}) = Y$.

For simplicity, consider a $g(\cdot)$ such that $N^{-1}g(\cdot)$ tends to a limit. Taylor linearization of $\hat{\theta} = g(\hat{\boldsymbol{Y}}, \hat{\boldsymbol{X}}^{(1)})$ around $(\boldsymbol{Y}, \boldsymbol{X})$ gives

$$\hat{\theta} - \theta = g(\hat{\boldsymbol{Y}}, \hat{\boldsymbol{X}}^{(1)}) - g(\boldsymbol{Y}, \boldsymbol{X})$$

$$\approx (\partial g(\boldsymbol{a}, \boldsymbol{a}^{(1)})/\partial\boldsymbol{a})^T\big|_{\boldsymbol{a}=\boldsymbol{Y}, \boldsymbol{a}^{(1)}=\boldsymbol{X}}(\hat{\boldsymbol{Y}} - \boldsymbol{Y})$$

$$+\big(\partial g(\boldsymbol{a}, \boldsymbol{a}^{(1)})/\partial\boldsymbol{a}^{(1)}\big)^T\big|_{\boldsymbol{a}=\boldsymbol{Y}, \boldsymbol{a}^{(1)}=\boldsymbol{X}}(\hat{\boldsymbol{X}}^{(1)} - \boldsymbol{X}). \quad (4.2)$$

Let $\check{\boldsymbol{Y}} = \sum b_k\boldsymbol{y}_k$ and $\check{\boldsymbol{X}}^{(1)} = \sum b_k^{(1)}\boldsymbol{x}_k$ for arbitrary real numbers $\boldsymbol{b} = (b_1, ..., b_N)^T$ and $\boldsymbol{b}^{(1)} = (b_1^{(1)}, ..., b_N^{(1)})^T$. Also, let $g(\check{\boldsymbol{Y}}, \check{\boldsymbol{X}}^{(1)}) = f(\boldsymbol{b}, \boldsymbol{b}^{(1)}, \boldsymbol{A}_y, \boldsymbol{A}_x) = f(\boldsymbol{b}, \boldsymbol{b}^{(1)})$, where $\boldsymbol{A}_y$ is an $m \times N$ matrix with k^{th} column $\boldsymbol{x}_k = (x_{k1}, ..., x_{kp})^T$, $k = 1, ..., N$, and $\boldsymbol{A}_x$ is an $p \times N$ matrix with k^{th} column $\boldsymbol{y}_k = (y_{k1}, ..., y_{km})^T$, $k = 1, ..., N$. Now following the derivation of (2.3) and noting that $\hat{\boldsymbol{Y}} = \boldsymbol{A}_y\boldsymbol{d}(s)$, $\boldsymbol{Y} = \boldsymbol{A}_y\boldsymbol{1}$, $\hat{\boldsymbol{X}}^{(1)} = \boldsymbol{A}_x\boldsymbol{d}^{(1)}(s_1)$, $\boldsymbol{X} = \boldsymbol{A}_x\boldsymbol{1}$, it can be shown that (4.2) reduces to

$$\hat{\theta} - \theta \approx \tilde{\boldsymbol{z}}^T(\boldsymbol{d}(s) - \boldsymbol{1}) + \tilde{\boldsymbol{z}}^{(1)T}(\boldsymbol{d}^{(1)}(s_1) - \boldsymbol{1}), \quad (4.3)$$

where $\boldsymbol{d}(s) = (d_1(s), ..., d_N(s))^T$ and $\boldsymbol{d}^{(1)}(s_1) = (d_1^{(1)}(s_1), ..., d_N^{(1)}(s_1))^T$. Further, $\tilde{\boldsymbol{z}} = (\tilde{z}_1, ..., \tilde{z}_N)^T$ with $\tilde{z}_k = \partial f(\boldsymbol{b}, \boldsymbol{b}^{(1)})/\partial b_k\big|_{\boldsymbol{b}=\boldsymbol{1}, \boldsymbol{b}^{(1)}=\boldsymbol{1}}$, and $\tilde{\boldsymbol{z}}^{(1)} = (\tilde{z}_1^{(1)}, ..., \tilde{z}_N^{(1)})^T$ with $\tilde{z}_k^{(1)} = \partial f(\boldsymbol{b}, \boldsymbol{b}^{(1)})/\partial b_k^{(1)}\big|_{\boldsymbol{b}=\boldsymbol{1}, \boldsymbol{b}^{(1)}=\boldsymbol{1}}$. It follows from (4.3) that a variance estimator of $\hat{\theta}$ is approximately given by the variance estimator of the estimated total $\sum d_k(s)\tilde{z}_k + \sum d_k^{(1)}(s_1)\tilde{z}_k^{(1)} = \hat{Y}(\tilde{z}) + \hat{X}^{(1)}(\tilde{z}^{(1)})$. We denote the latter variance estimator as $v(\tilde{z}, \tilde{z}^{(1)})$. Now we replace $\tilde{z}_k$ and $\tilde{z}_k^{(1)}$ by $z_k = \partial f(\boldsymbol{b}, \boldsymbol{b}^{(1)})/\partial b_k\big|_{\boldsymbol{b}=\boldsymbol{d}(s), \boldsymbol{b}^{(1)}=\boldsymbol{d}^{(1)}(s_1)}$ and $z_k^{(1)} = \partial f(\boldsymbol{b}, \boldsymbol{b}^{(1)})/\partial b_k^{(1)}\big|_{\boldsymbol{b}=\boldsymbol{d}(s), \boldsymbol{b}^{(1)}=\boldsymbol{d}^{(1)}(s_1)}$ respectively, since $\tilde{z}_k$ and $\tilde{z}_k^{(1)}$ are unknown. This leads to a linearization variance estimator

$$v_L(\hat{\theta}) = v(z, z^{(1)}). \quad (4.4)$$

We now consider the special case of a "double expansion" estimator $\hat{Y}(y) = \sum d_k(s)y_k$ with $d_k(s) = \pi_{1k}^{-1}\pi_{2k/1}^{-1}$ for $k \in s$ and the Horvitz-Thompson (H-T) estimator $\hat{X}^{(1)}(x) = \sum d_k^{(1)}(s_1)x_k$ with $d_k^{(1)}(s_1) = \pi_{1k}^{-1}$ for $k \in s_1$, where π_{1k} is the probability of including element k in s_1, and $\pi_{2k/l}$ is the conditional probability of including element k in s

given s_1. In this case, an unbiased H-T type estimator of $\hat{Y}(y)+\hat{X}^{(1)}(x)$ is given by

$$v(y,x)=\sum\sum_{k,l\in s_1}\frac{\pi_{1kl}-\pi_{1k}\pi_{1l}}{\pi_{1kl}}\frac{x_k}{\pi_{1k}}\frac{x_l}{\pi_{1l}}$$

$$+\sum\sum_{k,l\in s}\frac{\pi_{1kl}-\pi_{1k}\pi_{1l}}{\pi^*_{kl}}\left(\frac{y_k}{\pi_{1k}}\frac{y_l}{\pi_{1l}}+2\frac{y_k}{\pi_{1k}}\frac{x_l}{\pi_{1l}}\right)$$

$$+\sum\sum_{k,l\in s}\frac{\pi_{2kl/1}-\pi_{2k/1}\pi_{2l/1}}{\pi_{2kl/1}}\frac{y_k}{\pi^*_k}\frac{y_l}{\pi^*_l} \qquad (4.5)$$

where $\pi^*_k=\pi_{1k}\pi_{2k/1}$, $\pi^*_{kl}=\pi_{1kl}\pi_{2kl/1}$, π_{1kl} is the probability of including both elements k and l in s_1 and $\pi_{2kl/1}$ is the conditional probability of including both elements k and l in s given s_1. A proof of (4.5) is given in the Appendix. The variance estimator (4.4) is obtained from (4.5) by changing y_k and x_k to z_k and $z_k^{(1)}$ respectively.

Example 4.1 We illustrate the calculation of $v(z,z^{(1)})$ for the two-phase ratio estimator $\hat{Y}_{R2}$, given by (4.1), for the special case of simple random sampling at both phases: s_1 is a simple random sample of size n and s is a simple random subsample of size m from s_1. In this case, $\pi_{1k}=n/N$ and $\pi_{2k/1}=m/n$. Further, it follows from (4.1) that for general two-phase design,

$$z_k=\frac{\hat{X}^{(1)}}{\hat{X}}(y_k-\hat{R}x_k)=\frac{\hat{X}^{(1)}}{\hat{X}}e_k \qquad (4.6)$$

and

$$z_k^{(1)}=\hat{R}x_k. \qquad (4.7)$$

Under simple random sampling at both stages, (4.6) and (4.7) reduce to $z_k=(\bar{x}^{(1)}/\bar{x})e_k$ and $z_k^{(1)}=(\bar{y}/\bar{x})x_k$, where $e_k=y_k-(\bar{y}/\bar{x})x_k$, $\bar{y}$ and $\bar{x}$ are the second-phase sample means of y and x respectively, and $\bar{x}^{(1)}$ is the first-phase sample mean of x. Now substituting z_k and $z_k^{(1)}$ for y and x in (4.5) and nothing that $\pi_{1kl}=n(n-1)/[N(N-1)]$, $\pi_{2k/1}=m(m-1)/[n(n-1)]$, $\pi_{1kk}=\pi_{1k}$ and $\pi_{2kk/1}=\pi_{2k/1}$, we get

$$v_L(\hat{Y}_{R2})=N^2\left(\frac{1}{n}-\frac{1}{N}\right)\hat{R}^2s_{1x}^2+N^2\left(\frac{1}{m}-\frac{1}{n}\right)\left(\frac{\bar{x}^{(1)}}{\bar{x}}\right)^2s_{2e}^2$$
$$+2N^2\left(\frac{1}{n}-\frac{1}{N}\right)\hat{R}\frac{\bar{x}^{(1)}}{\bar{x}}s_{ex}, \qquad (4.8)$$

where

$$\hat{R}=\bar{y}/\bar{x},\ s_{1x}^2=(n-1)^{-1}\sum_{k\in s_1}(x_k-\bar{x}^{(1)})^2,$$

$$s_{2e}^2=(m-1)^{-1}\sum_{k\in s}(e_k-\bar{e})^2,$$

$$s_{2ex}=(m-1)^{-1}\sum_{k\in s}(e_k-\bar{e})(x_k-\bar{x})$$

and $\bar{e}$ is the second-phase sample mean of e. The formula (4.8) agrees with the formula derived by Rao and Sitter (1995). It is different from the customary formula (Sukhatme and Sukhatme 1970, page 176) which fails to make use of the full x-data $\{x_k, k\in s_1\}$. Rao and Sitter (1995) demonstrated through simulation that $v_L(\hat{Y}_{R2})$ is more efficient than the customary variance estimator. Also, $v_L(\hat{Y}_{R2})$ performed better in tracking the conditional mean squared error of $\hat{Y}_{R2}$; see Rao and Sitter (1995, section 3) for details of the simulation study.

CONCLUDING REMARKS

We have presented a unified approach to deriving Taylor linearization variance estimators and applied it to a variety of problems. It leads directly to a variance estimator that has some desirable properties at least in a number of important special cases; in particular, approximate unbiasedness for the model variance of the estimator under an assumed model and validity under a conditional repeated sampling framework. It would be useful to investigate whether such desirable properties also hold for more complex cases such as the general class of calibration estimators (section 3.2), the estimators based on estimating equations (section 3.3) and two-phase sampling (section 4). We are currently investigating various extensions of our method, including variance estimation under imputation for item nonresponse and variance estimation from longitudinal survey data.

ACKNOWLEDGMENTS

We thank the Associate Editor and a referee for constructive comments and suggestions. We also thank several colleagues in Statistics Canada for useful suggestions and encouragement, especially Linda Standish, David Binder, Geoff Hole, Richard Burgess and Larry Swain. Demnati's work was made possible by the Small Area and Administrative Data Division of Statistics Canada. J.N.K. Rao's work was supported by a grant from the Natural Sciences and Engineering Research Council of Canada.

APPENDIX

Unbiased Variance Estimator of $\hat{Y}(y)+\hat{X}^{(1)}(x)$

The variance of $\hat{Y}(y)+\hat{X}^{(1)}(x)$ is the sum of the variance of $\hat{Y}(y)$, the variance of $\hat{X}^{(1)}(x)$ and twice the covariance of $\hat{Y}(y)$ and $\hat{X}^{(1)}(x)$. An unbiased H-T type estimator of

$V[\hat{Y}(y)]$ is given by Särndal, Swensson and Wretman (1991, chapter 9, page 348):

$$v\left[\hat{Y}(y)\right] = \sum\sum_{k,l\in s} \frac{\pi_{1kl}-\pi_{1k}\pi_{1l}}{\pi^{*}_{kl}} \frac{y_k}{\pi_{1k}} \frac{y_l}{\pi_{1l}} + \sum\sum_{k,l\in s} \frac{\pi_{2kl/1}-\pi_{2k/1}\pi_{2l/1}}{\pi_{2kl/1}} \frac{y_k}{\pi^{*}_{k}} \frac{y_l}{\pi^{*}_{l}}. \quad (A.1)$$

An unbiased H-T type estimator of $V[\hat{X}^{(1)}(x)]$ is given by

$$v\left[\hat{X}^{(1)}(x)\right] = \sum\sum_{k,l\in s_1} \frac{\pi_{1kl}-\pi_{1k}\pi_{1l}}{\pi_{1kl}} \frac{x_k}{\pi_{1k}} \frac{x_l}{\pi_{1l}}. \quad (A.2)$$

Further,

$$\mathrm{Cov}\left[\hat{Y}(y), \hat{X}^{(1)}(x)\right] = E\mathrm{Cov}_2\left[\hat{Y}(y), \hat{X}^{(1)}(x)\right] + \mathrm{Cov}\left[E_2(\hat{Y}(y)), E_2(\hat{X}^{(1)}(x))\right],$$

where E_2 and Cov_2 denote conditional expectation and conditional covariance given s_1. Noting that

$$E_2\,\hat{Y}(y) = \hat{X}^{(1)}(y), E_2\,\hat{X}^{(1)}(x) = \hat{X}^{(1)}(x)$$

and $\mathrm{Cov}_2[\hat{Y}(y), \hat{X}^{(1)}(x)] = 0$, we get

$$\mathrm{Cov}\left[\hat{Y}(y), \hat{X}^{(1)}(x)\right] = \mathrm{Cov}\left[\hat{X}^{(1)}(y), \hat{X}^{(1)}(x)\right].$$

An unbiased H-T type estimator of $2\,\mathrm{Cov}\,[\hat{X}^{(1)}(y), \hat{X}^{(1)}(x)]$ is given by

$$2\,\mathrm{cov}\left[\hat{X}^{(1)}(y), \hat{X}^{(1)}(x)\right] = 2\sum\sum_{k,l\in s} \frac{\pi_{1kl}-\pi_{1k}\pi_{1l}}{\pi^{*}_{kl}} \frac{y_k}{\pi_{1k}} \frac{x_l}{\pi_{1l}}. \quad (A.3)$$

The sum of (A.1), (A.2) and (A.3) equals (4.5).

REFERENCES

ANDERSON, C., and NORDBERG, L. (1994). A method for variance estimation of non-linear functions of totals in surveys - theory and software implementation. *Journal of Official Statistics*. 10, 395-405

BERGER, Y.G. (2002). A generalized jackknife variance estimator for nonlinear statistics in probability sampling. Technical Report, Department of Social Statistics, University of Southampton.

BINDER, D. (1983). On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review*. 51, 279-292.

BINDER, D. (1996). Linearization methods for single phase and two-phase samples: a cookbook approach. *Survey Methodology*. 22, 17-22.

CAMPBELL, C. (1980). A different view of finite population estimation. *Proceeding of the Section on Survey Research Methods*, American Statistical Association. 319-324.

DEVILLE, J.C. (1999). Variance estimation for complex statistics and estimators: linearization and residual techniques. *Survey Methodology*. 25, 193-203.

DEVILLE, J.C., and SÄRNDAL, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*. 87, 376-382.

HAMPEL, F.R., RONCHETTI, E.M., ROUSSEEUW, P.J. and STAHEL, W.A. (1986). *Robust Statistics: The Approach Based on Influence Functions*. New York: John Wiley & Sons, Inc.

HUBER, P.J. (1981). *Robust Statistics*. New York: John Wiley & Sons, Inc.

KREWSKI, D., and RAO, J.N.K. (1981). Inference from stratified samples: properties of the linearization, jackknife and balanced repeated replication methods. *Annals of Statistics*. 9, 1010-1019.

MONTANARI, G.E. (1998). On regression estimation of finite population means. *Survey Methodology*. 24, 69-77.

RAO, J.N.K., and SITTER, R.R. (1995). Variance estimation under two-phase sampling with application to imputation for missing data. *Biometrika*. 82, 453-460.

RAO, J.N.K., YUNG, W. and HIDIROGLOU, M. (2002). Estimating equations for the analysis of survey data using poststratification information. *Sankhyā*.

ROYALL, R.M., and CUMBERLAND, W.G. (1981). An empirical study of the ratio estimator and estimators of its variance. *Journal of the American Statistical Association*. 76, 66-77.

SÄRNDAL, C.-E., SWENSSON, B. and WRETMAN, J.H. (1989). The weighted residual technique for estimating the variance of the general regression estimator of the finite population total. *Biometrika*. 76, 527-537.

SÄRNDAL, C.-E., SWENSSON, B. and WRETMAN, J.H. (1991). *Model Assisted Survey Sampling*. New York: Springer-Verlag.

SERFLING, R.J. (1980). *Approximation Theorems of Mathematical Statistics*. New York: John Wiley & Sons, Inc.

SINGH, A.C., and FOLSOM, R.E. (2000). Bias correcting estimating function approach for variance estimation adjusted for poststratification. *Proceeding of the Section on Survey Research Methods*, American Statistical Association. 610-615.

SITTER, R.R., and WU, C. (2002). Efficient estimation of quadratic finite population functions in the presence of auxiliary information. *Journal of the American Statistical Association*. 97, 535-544.

SUKHATME, P.V., and SUKHATME, B.V. (1970). *Sampling Theory of Surveys with Applications*. 2nd ed. London: Asia Publishing House.

VALLIANT, R. (1993). Postsratification and conditional variance estimation. *Journal of the American Statistical Association*. 88, 89-96.

YUNG, W., and RAO, J.N.K. (1996). Jackknife linearization variance estimators under stratified multi-stage sampling. *Survey Methodology*. 22, 23-31.

Comment

PHILLIP S. KOTT[1]

The article addresses an impressive number of contexts, many of which have only recently been investigated in the literature, often by Professor Rao himself. I will have little to say here about estimating functions with calibration weights or two-phase sampling, except (mostly) to agree with the solutions advocated in the text. Instead, I will focus on three applications: the ratio estimator under simple random sampling discussed in the Introduction, the general class of regression calibration weights from section 3.2, and the general class of calibration estimators from section 3.4. I will end with a question about the linearization variance estimator in full Horvitz-Thompson form, which has bothered me for some time.

The Ratio Under Simple Random Sampling

Before beginning, let me confess to a certain skepticism about the general method proposed in section 2. I find that techniques of this sort work best when you already know what the answer is. Godambe and Thompson (1986) tried to use estimating functions to settle a controversy then surrounding the best variance estimator for the ratio under simple random sampling. Using the notation in the text, they demonstrated that $(\bar{X}/\bar{x})\,v_L$ was the proper way to estimate the variance of a ratio estimator, $\hat{Y}_R = (\bar{X}/\bar{x})\,\bar{y}$. Later, Binder (1996) corrected them. He showed that when done properly, $v_{JL} = (\bar{X}/\bar{x})^2\,v_L$ is produced from estimating-function technology. It helped that he already knew that was the better answer.

As Demanti and Rao state, v_{JL} has both good randomization (design) and model-based properties (here and hereafter I omit the qualifier, "under mild conditions which I assume to hold"). In fact, *when n/N is ignorably small*, v_{JL} has a relative bias of $O(1/n)$ as an estimator for the model variance of $\hat{Y}_R$. If the y_k are uncorrelated, then this is not only true when $V_m(y_k) = \sigma^2 x_k$ as stated in the text, but, more generally, when $V_m(y_k) = \sigma_k^2$. Unfortunately, the result is less general when *n/N* is *not* ignorably small. In that context, when the y_k are uncorrelated and $V_m(y_k) = \sigma^2 x_k$, a more appropriate estimator for the model variance of $\hat{Y}_R$ is $v_m = [(\bar{X}/\bar{x})^2 - (n/N)(\bar{X}/\bar{x})][1 - (n/N)]^{-1} v_L$ (Kott and Brewer 2001). As an estimator for the randomization mean squared error of $\hat{Y}_R$, v_m has a relative bias of $O(1/\sqrt{n})$, just like v_{JL} and v_L.

When simple random sampling is used in practice the sampling fraction is almost always small. Thus, v_{JL} is an attractive variance/mean-squared-error estimator, and my criticism of Demnati and Rao for advocating it is mild.

A General Class of Regression Calibration Weights

I would generalize the results of section 3.1 in a different manner than the authors do in section 3.2. Following Estavao and Särndal (2002), replace $c_k \boldsymbol{x}_k$ in equation (3.1) with a vector $\boldsymbol{q}_k$ having the same dimension as $\boldsymbol{x}_k$. The rest of that section follows easily.

One choice for $\boldsymbol{q}_k$ is

$$\boldsymbol{q}_{(1)k} = \sum_{j \in U} (\pi_{kj} - \pi_k \pi_j)\boldsymbol{x}_j / (\pi_k \pi_j),$$

the use of which results in a variant of the randomization-optimal regression estimator proposed by Tillé (1999). Observe that $(\Sigma_U \boldsymbol{q}_{(1)k} \boldsymbol{x}_k^T)^{-1}(\Sigma_U \boldsymbol{q}_{(1)k} y_k^T) = [\mathrm{Var}(\hat{\boldsymbol{X}})]^{-1} \mathrm{Cov}(\hat{\boldsymbol{X}}, \hat{Y})$, where Var and Cov denote randomization-based properties.

Another choice, investigated indirectly by Demnati and Rao and likewise resulting into a variant of the randomization-optimal estimator, is

$$\boldsymbol{q}_{(2)k} = \sum_{j \in s} (\pi_{kj} - \pi_k \pi_j)\boldsymbol{x}_j / (\pi_{kj} \pi_j).$$

Since $\boldsymbol{q}_{(2)k}$ is a function of the sample, the authors take us through the complications of section 3.2. This was only necessary for randomization-based inference. I would have gone a different way. Observe that $d_k(s)\boldsymbol{q}_{(2)k} - d_k(s)\boldsymbol{q}_{(1)k} = O_P(1/\sqrt{n})$. Replacing one for the other has an asymptotically ignorable effect on $w_k(s)$ (*i.e.*, the relative difference is $O_P(1/n)$).

A General Class of Calibration Estimators

A mild generalization of equation (3.16) allows calibration weights of the form,

$$w_k(s) = d_k(s) F(\boldsymbol{q}_k^T \hat{\boldsymbol{\lambda}}),$$

where $\boldsymbol{q}_k$ again has the same dimension as $\boldsymbol{x}_k$. For convenience F is assumed positive and twice differentiable around $\boldsymbol{q}_k^T \boldsymbol{\lambda}$. Without loss of generality, one can assume $\boldsymbol{\lambda}$ (the limit of $\hat{\boldsymbol{\lambda}}$) is $\mathbf{0}$, and $f(0) > 1$. When $\hat{Y}_{GC} = \Sigma_U w_k(s) y_k$ is a randomization consistent estimator, as I assume it is, $F(0)$ is equal to 1.

Paralleling the development in the text leads ultimately to

$$z_k = F(\boldsymbol{q}_k^T \hat{\boldsymbol{\lambda}})(y_k - \boldsymbol{x}_k^T \hat{\boldsymbol{B}}_\lambda) = F(\boldsymbol{q}_k^T \hat{\boldsymbol{\lambda}}) e_{k\lambda},$$

[1] Phillip S.Kott, USDA / NASS, 3251 Old Lee Hwy, Fairfax, VA 22030, U.S.A.

where $\hat{B}_\lambda = [\sum d_k(s) f(q_k^T\hat{\lambda}) q_k x_k^T]^{-1} \sum d_k(s) f(q_k^T\hat{\lambda}) q_k y_k$. The presence of the $f(\cdot)$ in the expession of $\hat{B}_\lambda$ may be a bit of a surprise, but, it turns out, not a meaningful one in this context. For inference under the prediction model, $E_m(y_k|x_k) = x_k^T\beta$, the derivative can be replaced by any constant without asymptotic consequence; $\hat{B}_\lambda$ remains a model unbiased estimator for β. For randomization-based inference, since $q_k^T\hat{\lambda} = O_P(1/\sqrt{n})$ and $F(0), f(0)>0, z_k$ would be unaffected asymptotically if $f(q_k^T\hat{\lambda})$ were replaced by 1 or by $F(q_k^T\hat{\lambda})$.

Things change, however, if we push the envelop a bit. Fuller, Loughin and Baker (1994) use calibration to adjust for unit nonresponse by treating sample response as a second phase of sampling. They assume that every element k in the population has a Poisson probability of sample response, π_{2k}, which is independent of whether it is actually chosen for the sample. They further assume $\pi_{2k} = 1/(1+x_k^T\lambda)$, where λ is unknown and implicitly estimated by calibration. Here we generalize that and assume $\pi_{2k} = 1/F(q_k^T\lambda)$, where F is known, positive, and twice differentiable. In practice, q_k will likely be identical to x_k, but it may be reasonable to replace one of more components of x_k with variables conjectured to be more strongly correlated with response/ nonresponse.

Redefining s as the respondent sample and $d_k(s)$ as $(1/\pi_{1k})$ when $k\in s, 0$ otherwise, everything proceeds as before. The difference is that $f(q_k^T\hat{\lambda})$ in $\hat{B}_\lambda$ need no longer need be asymptotically identical across the k. Thus, the term can matter even with a large sample.

Now $V(\hat{Y}_{GC}) \approx V(\sum_U d_k(s) z_k)$, where $\sum_U d_k(s) z_k = \sum_U d_k(s) F(q_k^T\hat{\lambda}) e_{k\lambda}$ is the double expansion estimation. Substituting $1/F(q_k^T\hat{\lambda})$ for π_{2k}, the variance estimator for $\hat{Y}_{GC}$ becomes (from equation (A.1) with $\pi_{2kjl} = \pi_{2kj}\pi_{2k}\pi_{2j}$)

$$v(\hat{Y}_{GC}) = \sum_{k,j\in s} [(\pi_{1kj} - \pi_{1k}\pi_{1j})/\pi_{1kj}]$$
$$d_k(s) F(q_k^T\hat{\lambda}) e_{k\lambda} d_j(s) F(q_j^T\hat{\lambda}) e_{j\lambda}$$
$$+\sum_{k\in s} \pi_{1k}\{[F(q_k^T\hat{\lambda})]^2 - [F(q_k^T\hat{\lambda})]\}[d_k(s) e_{k\lambda}]^2.$$

This differs from the variance estimator in Folsom and Singh (2000) mainly because those authors assume the original sample is chosen using a stratified multistage design employing with-replacement sampling in the first. That, among other things, annihilates the second summation on the right hand side.

Not only does $v(\hat{Y}_{GC})$ estimate the quasi-randomization mean squared error of $\hat{Y}_{GC}$– "quasi" because a response model is assumed, it also estimates the model variance of $\hat{Y}_{GC}$. In fact, the relative bias of $v(\hat{Y}_{GC})$ under the prediction model, $E_m(y_k|x_k, q_k) = x_k\beta$, is $O(1/n)$ when the y_k are uncorrelated and $V_m(y_k|x_k, q_k) = x_k^T\gamma$, where γ (like β) need not be specified. Surprisingly, the second term in $v(\hat{Y}_{GC})$ provides the model-based correction I recommended for the ratio estimator under simple random sampling *in the absence of nonresponse.*

Does the "Plug-in" Variance Estimator Really Work for the Full Horvitz-Thompson Form?

As I warned parenthetically early on, I have omitted the key phrase, "under mild conditions which I assume to hold," repeatedly in these comments. Now, I want to turn my attention to what may be one of those conditions. It is standard in variance estimation to replace population (or model) values with sample analogues since their difference is asymptotically ignorable. That is done, for example, by Demnati and Rao in equation (2.4) when they plug in z_k for $\tilde{z}_k$. The question I want to raise, and for which I do not know the answer, is this. Suppose one is estimating a total with a calibration estimator. The total is $O(N)$, and $O(n) = O(N)$. The estimator's model variance and randomization mean squared error are also $O(n)$. Is it legitimate to plug in z_k for $\tilde{z}_k$, where $z_k - \tilde{z}_k = O_P(1/\sqrt{n})$, when there are $n(n-1)/2$ terms in the Horvitz-Thompson – or Yates-Grundy – variance/mean-squared-error estimator? In most practical applications, this is a non-issue, because the variance estimator can be re-expressed with $O(n)$ terms. What if that is not the case?

Let me conclude these remarks by thanking Drs. Demnati and Rao for their stimulating article and *Survey Methodology* for both publishing it and allowing me to provide some comments.

ADDITIONAL REFERENCES

ESTEVAO, V.M., and SÄRNDAL, C-E. (2002). The ten cases of auxiliary information for calibration in two-phase sampling. *Journal of Official Statistics*. 18, 233-255.

FULLER, W.A., LOUGHIN, M.M. and BAKER, H.D. (1994). Regression weighting for the 1987-88 National Food Consumption Survey. *Survey Methodology*. 20, 75-85.

GODAMBE, V.P., and THOMPSON, M.E. (1986). Parameters of superpopulation and survey population: their relationship and estimation. *International Statistical Review*. 54, 2, 127-138.

KOTT, P.S., and BREWER K.R.W. (2001). Estimating the model variance of a randomization-consistent regression estimator. *Proceedings of the Section on Survey Research Methods*, American Statistical Association.

RAO, J.N.K., and SHAO, J. (1992). Jackknife variance estimation with survey data under hot deck imputation. *Biometrika*. 811-822.

TILLÉ, Y. (1999). Estimation in surveys using conditional inclusion probabilities: complex designs. *Survey Methodology*. 25, 57-66.

Comment

BABUBHAI V. SHAH[1]

This is an excellent paper that removes the mystery underlying Taylor linearization. Most data analysis applications use Horvitz-Thompson weights that are reciprocals of the probabilities of selection. The simplest prescription for deriving the linearization for an estimator $\hat{\theta}$ is as follows:

1. For each observation, create a new variable $z_i = \partial\hat{\theta} / \partial w_i$, where w_i is the reciprocal of the selection probability for the i-th observation selected in the sample. In cases where the estimator $\hat{\theta}$ is defined implicitly through estimating equations, the derivative can be computed by differentiating the implicit equations.
2. Define weighted $\hat{T} = \Sigma w_i z_i$ total.
3. Compute the variance $\hat{V}$ of the total $\hat{T}$ based on the sample design.
4. The variance $\hat{V}$ is the approximate variance of the estimator $\hat{\theta}$.

If the parameter θ is a vector then the variable z_i and the total T are also vectors and $\hat{V}$ is an approximate estimate of the variance covariance matrix of the estimator $\hat{\theta}$.

The steps (1) and (2) specified above produce the correct linearization in the following cases:

a. Means, proportions, and ratio estimates.

b. Generalized linear regression models.

c. Predicted marginal for generalized linear model.

d. Estimate of the mean from regression imputed data.

e. Generalized linear regression models with calibrated weights.

f. Wilcoxon two sample rank sum test.

g. Estimates of coefficients and the hazard rate in Cox's proportional hazard model.

h. Estimates of predicted marginal survival in Cox's proportional hazard model.

i. Two-phase sample survey.

The derivation in the step (1) is uniquely defined and does not contain the true value of the parameter θ, and does not require substitution by the estimator $\hat{\theta}$.

The independence of step (3) for variance computation from the linearization in steps (1) and (2) is aptly demonstrated by the discussion on two-phase sampling in section 4. In most cases, one assumes with replacement sample design to estimate the variance of the total in the step (3). Of course, a better estimate of the variance of the total may be obtained by using all the available information about the sample design. For the case of a two-phase design, step (1) can be performed by using Horvitz Thompson weights for the phase one sampling, and treating the multipliers m_i as data. The multiplier m_i is equal to zero if the observation i is not selected in phase two and is equal to the inverse of the conditional probability $\pi^{-1}_{2k/1}$. The resulting step (2) produces the same total as presented in the paragraph between equations (4.3) and (4.4). The subsequent discussion in section 4, describes the appropriate way to estimate the variance of this total for a two-stage sample design without replacement at each stage, and that calculation is independent of the linearization.

The steps (1) and (2) generate appropriate linearization in all known cases except where the estimator is not a continuous function of the weights w_i, *e.g.*, quantile.

[1] Babubhai V. Shah, SAFAL Institute, Inc. E-mail: babushah@earthlink.net

Comment

CHRIS SKINNER[1]

Linearization and replication approaches provide two broad classes of methods for variance estimation in surveys. Both have their relative advantages and it seems important to keep a place for both in the survey statistician's 'toolkit'. This paper deepens our understanding of linearization methods, proposes a general procedure to generate such variance estimators uniquely and provides valuable illustrations of this procedure in some important areas of application.

A linearization method approximates the variance of a statistic of interest by the variance of a linear statistic, for which it is assumed a suitable variance estimator is available. The main issue here is the method used to determine the linear statistic. The standard approach assumes the statistic of interest may be expressed as a differentiable function of a vector of linear statistics (of fixed dimension) and uses Taylor series expansion to determine the approximation. The approach proposed in this paper applies to a more general class of sample-weighted statistics, illustrated by the complex examples in sections 3.2. and 4. The variance estimator is constructed by differentiating the statistic with respect to the sample weights. The approach to linear approximation is closely related to methods based upon the influence function (*e.g.*, equations 1.6 and 1.13) and the paper provides a helpful review of such methods in section 1. The authors note that it is not easy to verify the validity of such methods for statistics which are not smooth functions of (or a fixed number of) linear statistics and it would be interesting to know how far the proposed approach does indeed provide valid variance estimators for statistics, such as quantiles, which are not of this form.

A key feature of the proposed approach, which ensures the unique construction of the variance estimator, is that derivatives are evaluated at values based on the achieved sample, without any initial evaluation of the approximating linear statistic at theoretical population values. Such initial evaluation may lead to non-uniqueness when auxiliary information is available, for example on a population mean, $\bar{X}$, and it is assumed that this value is equal to the limiting value of a corresponding sample statistic, $\bar{x}$. For statistics which are smooth functions of linear statistics, it appears that the variance estimator generated by the proposed method may also be constructed by conventional Taylor series methods, provided no initial simplification of the variance estimator takes place based on such assumptions about auxiliary information. Such construction may, however, be less clear-cut than for the proposed approach.

Assumptions employed by linearization methods differing from the proposed approach, such as that an auxiliary value $\bar{X}$ is the theoretical limiting value of a sample value $\bar{x}$, are based upon unconditional distributions and so it might be anticipated that the incorporation of such assumptions into a variance estimator might damage the method's conditional properties, especially with respect to statistics such as $\bar{x}$. The proposed procedure avoids dependence upon such assumptions and, by evaluating derivatives at achieved sample values, may be expected to track conditional properties more closely. (There appear to be parallels with Efron and Hinkley's (1978) arguments in favour of the observed versus the expected information, although the context is rather different.)

The avoidance of dependence upon such assumptions may not only benefit the conditional properties of the proposed approach, but also protect the variance estimator against possible biasing effects of non-sampling errors. The auxiliary population information may differ from the limiting values of the corresponding sample statistics either because of non-response or non-coverage or because of discrepancies in the way the auxiliary variables are measured. In such circumstances, linearization methods differing from the proposed approach might lead to inconsistent variance estimation. For this reason, Fuller (2002, page 10) recommends the use of the g-weights in (3.6), as proposed, especially in the presence of nonresponse (page 15). With regards to the latter case, it seems worth noting that the validity of the proposed procedure does not appear to depend on the requirement that $E(\boldsymbol{d}(s)) = \mathbf{1}$, provided $\mathbf{1}$ is replaced by $E(\boldsymbol{d}(s))$ in the development in section 2. In particular, if s denotes unit respondents and non-response may be represented by Poisson sampling with unknown response probabilities then the proposed approach to variance estimation may still be consistent (when based on many standard variance estimators for linear statistics), even if $d(s)$ is based only on sampling inclusion probabilities.

Julia d'Arrigo and I have recently studied the properties of linearization variance estimators under nonresponse in simulation studies as part of the DACSEIS research project (www.dacseis.de) using data from the UK Labour Force

[1] Chris Skinner, Southampton Statistical Sciences Research Institute, University of Southampton, Southampton S017 1BJ, United Kingdom. E-mail: cjs@socsci.soton.ac.uk.

Survey and the German Income and Expenditure Survey. We considered various calibration estimators under Poisson models for unit non-response which were ignorable given the calibrating variables, using standard variance estimators for linear statistics under stratified multi-stage sampling. We indeed found that nonresponse could lead to serious biases in the linearization variance estimators if they failed to take account of the g-weights for GREG estimation (section 3.1.) or ignored the $F(\boldsymbol{x}_k^T\hat{\boldsymbol{\lambda}})$ term in (3.21). Such biases were absent in the proposed approach.

We also investigated the alternative calibration estimators discussed in section 3.4. Deville and Särndal's (1992) theoretical finding that the asymptotic variance of $\hat{Y}_w$ does not depend on the form of the function $F(\cdot)$ is based on the assumption that $\Sigma d_k(s)\boldsymbol{x}_k$ is consistent for $\boldsymbol{X}$. This assumption may not hold under various sources of non-sampling error, and is not required for the proposed approach. Hence, the appropriate approximate linear statistic (under departures from this assumption) is defined by (3.21) and the resulting variance estimator may depend on the form of $F(\cdot)$, even asymptotically. The standard linearization variance estimators in which $d_k(s)f(\boldsymbol{x}_k^T\hat{\boldsymbol{\lambda}})$ in $\hat{\boldsymbol{B}}_\lambda$ is replaced by $d_k(s)$ or $w_k(s)$ may be inconsistent if these weights differ from $d_k(s)f(\boldsymbol{x}_k^T\hat{\boldsymbol{\lambda}})$. Despite this theoretical fact, we observed little difference in our simulation study (for each of the functions, $1+u$, $\exp(u)$, and $(1-u)^{-1}$, used for $F(u)$) between the statistical properties of variance estimators based upon these three different choices of weight, $d_k(s)f(\boldsymbol{x}_k^T\hat{\boldsymbol{\lambda}})$, $d_k(s)$ or $w_k(s)$, in the $\hat{\boldsymbol{B}}_\lambda$ vector in (3.21). Others studies might produce different findings.

A disadvantage of the linearization methods considered here compared to replication methods is the need for analytic differentiation. It would appear from the examples presented in this paper that the analytic differentiation involved in the proposed method is at least as straightforward as that in standard methods of Taylor series expansion of smooth functions of linear statistics. Nevertheless, in some applications, it may be advantageous to replace the human labour and possible human error arising with analytic differentiation by the use of 'numerical differentiation'. The proposed approach might be described as an *infinitestimal jackknife* method since it perturbs the weight given to each sample observation by an infinitesimal amount to determine the approximating linear statistic. The derivative with respect to a weight in the proposed approach may be approximated numerically by a finite difference approach in which the statistic is recalculated with the weight perturbed by a finite amount for each observation in turn. This approach may be described as a *jackknife* method of linearization. A conventional approach would be to change each weight to zero in turn, perhaps standardizing for unequal weights as in (1.15). It does not seem essential to replace the original weight by zero and, in principle, each weight might be perturbed in some other way, for example by reducing it by a fixed amount δ, smaller than the minimum value of $d_k(s)$. It seems likely that in many applications the variance estimator arising from such jackknife linearization will have very similar statistical properties to that constructed by the proposed approach. The choice between the estimators is likely to depend more on practical and computational considerations.

My final comments are on terminology. There are practical reasons why it may be helpful to give the z_k variable a name. In particular, this may be helpful for the practitioner who, for some complex statistics, has to employ two separate computational steps: (a) construction of the z_k variable, for example using least squares routines when calibration weighting is used, and (b) use of standard variance estimation software for linear statistics. Different names are used for z_k in the literature. Woodruff (1971) is usually acknowledged as the first paper in the survey sampling literature to draw attention to the role of z_k and Andersson and Nordberg (1994) refer to z_k as the *Woodruff transformation*. Woodruff and Causey (1976) refer to the approximating linear statistic as the *linear substitute* and z_k as the *substitute variable*. In the more mainstream statistical literature, Davison and Hinkley (1997, page 46) refer to the z_k as the *empirical influence values*. The term *linearized variable*, as used by Deville (1999), seems to me a simple and natural one. It is consistent with the use of the term *linearized statistic* to denote the approximating linear statistic and the term *linearization* for the method (which is a more suitable general term than Taylor series method for the broad class of approaches considered here).

ADDITIONAL REFERENCES

DAVISON, A.C., and HINKLEY, D.V. (1997). *Bootstrap Methods and their Applications*. Cambridge: Cambridge University Press.

EFRON, B., and HINKLEY, D.V. (1978). Assessing the accuracy of the maximum likelihood estimator: observed versus expected Fisher information (with discussion). *Biometrika*. 65, 457-487.

FULLER, W.A. (2002). Regression estimation for survey samples. *Survey Methodology*. 28, 5-23.

WOODRUFF, R.S. (1971). A simple method for approximating the variance of a complicated estimate. *Journal of the American Statistical Association*. 66, 411-414.

WOODRUFF, R.S., and CAUSEY, B.D. (1976). Computerized method for approximating the variance of a complicated estimate. *Journal of the American Statistical Association*. 71, 315-321.

Response from the Authors

1. INTRODUCTION

We thank the three discussants, Phillip Kott, Babubhai Shah and Chris Skinner, for their insightful comments. Our rejoinder will attempt to address some of the issues raised by the discussants. The main aim of our paper was to study variance estimation for calibration estimators of population totals and nonlinear parameters, θ, defined as solutions to "census" estimating equations. We proposed a new Taylor linearization approach that provides a unique variance estimator, by avoiding initial evaluation of the linearized statistic at the population values. We have also shown that the variance estimator satisfies some desirable considerations, such as approximate model unbiasedness and validity under a conditional repeated sampling frame work, at least in a number of important cases. We have also shown that in two-phase sampling the variance estimator makes fuller use of the first phase sample data compared to traditional linearization variance estimators.

Kott

Kott's discussion focused on three applications in our paper: (i) the jackknife linearization variance estimator, v_{JL}, of the ratio estimator $\hat{Y}_R = (\bar{y}/\bar{x})X$ in simple random sampling mentioned in section 1; (ii) the general class of regression calibration weights considered in section 3.2; (iii) the general class of calibration weights studied in section 3.4. Regarding (i), we noted the result that v_{JL} is both asymptotically design unbiased and approximately model unbiased under the ratio model $E_m(y_k) = \beta x_k$ and $V_m(y_k) = \sigma^2 x_k$. Kott is correct in saying that the model bias may not be negligible if the sampling fraction, n/N, is not small. If n/N is "ignorably small", then model unbiasedness is, in fact, valid under a general variance function $V_m(y_k) = \sigma_k^2$, as noted by Kott and previously by Särndal *et al.* (1989). Under the ratio model, Kott proposes a more appropriate variance estimator, v_m, that is model unbiased even if n/N is not small and also valid under repeated sampling. The leading terms of v_m and v_{JL} are identical, and our new approach captures only the leading term. It should be noted that model-unbiasedness of v_m depends on the validity of the assumption $\sigma_k^2 = \sigma^2 x_k$.

Turning to (ii), we have shown in section 3.2 that if the general class of regression calibration weights, (3.7), are used, our approach leads to a variance estimator that is quite complex, involving third and fourth order moments of the design weights $d_k(s)$ with $d_k(s) = 0$ if the k^{th} population element is not in the sample s. Kott proposes an attractive choice of weights obtained by replacing $c_k \boldsymbol{x}_k$ in the GREG weight (3.1) with $\boldsymbol{q}_{(1)k} = \sum_{l=1}^{N} (\pi_{kl} - \pi_k \pi_l) \boldsymbol{x}_l / (\pi_k \pi_l)$. This choice gives a variant of the "optimal" linear regression estimator and also avoids the complexities associated with the variance estimator based on the weights (3.7). This is an interesting and useful proposal, but $\boldsymbol{q}_{(1)k}$ requires the knowledge of the $\boldsymbol{x}$-vector for all the population elements, unlike (3.7) which depends only on the population total $\boldsymbol{X}$; in practice, only $\boldsymbol{X}$ may be available. Moreover, $\boldsymbol{q}_{(1)k}$ depends on all the $N(N-1)/2$ joint inclusion probabilities π_{kl} and hence computation of $\boldsymbol{q}_{(1)k}$ may become cumbersome when the sampling design is based on unequal probability sampling without replacement.

Turning to (iii), Kott proposes a generalization of the calibration weights $w_k(s) = d_k(s) F(\boldsymbol{x}_k^T \hat{\boldsymbol{\lambda}})$ in section 3.4 by replacing $\boldsymbol{x}_k$ with "instrumental" variables $\boldsymbol{q}_k$ having the same dimension as $\boldsymbol{x}_k$. The corresponding z-variable in the variance estimator $v(z)$ is similar to our (3.21) with $\boldsymbol{x}_k \boldsymbol{x}_k^T$ and $\boldsymbol{x}_k y_k$ in $\hat{\boldsymbol{B}}_\lambda$ changed to $\boldsymbol{q}_k \boldsymbol{x}_k^T$ and $\boldsymbol{q}_k y_k$ respectively and $F(\boldsymbol{x}_k^T \hat{\boldsymbol{\lambda}})$ changed to $F(\boldsymbol{q}_k^T \hat{\boldsymbol{\lambda}})$. This is an useful extension. Kott notes that $\hat{\boldsymbol{B}}_\lambda$ remains a model unbiased estimation of $\boldsymbol{B}_\lambda$ if $f(\boldsymbol{q}_k^T \boldsymbol{\lambda})$ in $\hat{\boldsymbol{B}}_\lambda$ is replaced by any constant and the resulting z_k is unaffected asymptotically under repeated sampling. However, Kott also notes that the term $f(\boldsymbol{q}_k^T \boldsymbol{\lambda})$ can matter even asymptotically if the calibration is used to adjust for unit nonresponse by treating sample response as a second phase of sampling. Using the result for two-phase sampling given in the Appendix, Kott then obtains a corresponding variance estimator, $v(\hat{Y}_{GC})$. This extension for nonresponse setting is also useful. It is indeed surprising that the second term in $v(\hat{Y}_{GC})$ provides the model based correction he recommended for the ratio estimator $\hat{Y}_R$ under simple random sampling in the absence of nonresponse.

Finally, Kott raises a question on the customary "plug-in" or "substitution" method used for variance estimation, as done in (2.4), where we plug in z_k for $\tilde{z}_k$. He asks if it is legitimate to plug in z_k for $\tilde{z}_k$, where $z_k - \tilde{z}_k = O_p(1/\sqrt{n})$, when they are $n(n-1)/2$ terms in the variance estimator $v(\tilde{z}_k)$, as in the case of Sen-Yates-Grundy variance estimator. We are not sure if we have understood his point correctly, but as long as $O_p(1/\sqrt{n})$ is uniform in k, say $a/\sqrt{n}$, then $v(z) = (\tilde{z}) +$ lower order terms.

Shah

Shah's prescription (steps 1-4) clearly summarizes our method. Shah also notes that his steps 1 and 2, leading to our z-variable, produces the "correct" linearization in many other important applications not studied in our paper,

including Wilcoxon two sample rank sum test and estimation of regression coefficients and hazard rate in the Cox proportional hazard model. Shah's unpublished paper (seen by courtesy of the author) spells out the z-variable for those applications, but using design weights. Extension to calibration weights should follow along the lines of section 3.

Shah makes an important point that step 3 for the computation of the variance estimate is independent of the linearization in step 1 and 2 and that it is "aptly demonstrated by the discussion on two-phase sampling in section 4". He also notes that for two-phase sampling, linearization (step 1) can be performed using only the first-phase H-T weights π_{1k}^{-1}, by treating the second phase weights, $\pi_{2k/1}^{-1}$, if $k \in s$ and 0 if k is not in the second-phase sample s as data, and that the resulting step 2 produces the same approximation as given in our paper. We have verified this equivalence result for the two-phase ratio estimator in Example 4.1, and it is likely to hold generally. Shah's proposal might simplify the implementation of step 1 to some extent.

Skinner

Skinner gives a clear appraisal of our linearization method and raises a number of important points: (i) terminology, (ii) possible extensions to non-smooth statistics such as quantiles, (iii) modifications of the method to handle unit nonresponse, (iv) possible use of numerical differentiation to calculate the z_k-variables.

With regard to point (i), Skinner notes that it would be useful to give the z_k variable a name since different names have been used in the literature. He suggests that the term *linearized variable*, as used by Deville (1999), is a simple and natural one since it is consistent with the usage of *linearized statistic* to denote the approximating linear statistic and linearization for the method. We are in agreement with Skinner's suggestion.

Turning to point (ii), a difficulty in extending our proposal to nonsmooth statistics $\hat{\theta} = f(\hat{\boldsymbol{d}}(s))$, such as quantiles, is that $f(\cdot)$ is not a differentiable function. A way to get around this difficulty is to approximate $\hat{\theta} - \theta$ by a differentiable function and then apply our method to the approximation. For example, in the case of the p^{th} quantile θ, Francisco and Fuller (1991) and Shao (1991) established the following asymptotic approximation valid for stratified multistage designs:

$$\hat{\theta} - \theta \approx -\frac{1}{h(\theta)}\left\{\hat{F}_w(\theta) - p\right\},$$

where $\hat{F}_w(\theta) = \Sigma w_k(s) I(y_k \le \theta) / \Sigma w_k(s)$ is the calibration estimator of the distribution function $F(\cdot)$ at θ, $F(\theta) = N^{-1}\Sigma I(y_k \le \theta) = p$, and $h(\theta)$ is the value of the density function $h(\cdot)$ at θ. The definition of $h(\cdot)$ requires reference to a sequence of populations (Shao and Rao 1993) or to a superpopulation (Francisco and Fuller 1991). We used $h(\cdot)$ to denote the density rather than the customary $f(\cdot)$ because we used $f(\boldsymbol{d}(s))$ to denote the estimator $\hat{\theta}$. Now, suppose $w_k(s) = d_k(s) g_k(\boldsymbol{d}(s))$, where $g_k(\boldsymbol{d}(s))$ is the GREG weight given by (3.1). We can then use (3.2) and (3.3) to get the linearized variable z_k from the above approximation to $\hat{\theta} - \theta$, by replacing $h(\theta)$ with a suitable estimator $\hat{h}(\hat{\theta})$; for example the kernel-based estimator of $h(\cdot)$ used by Berger and Skinner (2003). Similarly, one can apply the method to general calibration weights, $w_k(s)$, using the results of section 4. Variance estimators of a low income proportion, say $\theta = F(\tau/2)$ where τ is the median income, can also be obtained using the asymptotic approximation for $\hat{\theta} - \theta$ developed by Shao and Rao (1993). Berger and Skinner (2003) studied variance estimation for a low income proportion when generalized raking ratio weights, $w_k(s)$, are used. We can apply the results in section 3.2 to this case, and the resulting linearized variable z_k will account for the calibration. Also, it will be different from the Deville z-variable (10) in Berger and Skinner (2003).

The modification suggested in point (iii) to handle unit nonresponse is very important, and it broadens the applicability of our method. As noted by Skinner, Kott and Fuller (2002), it is important to retain the g-weights in variance estimation whenever the limiting values of the estimators $\hat{\boldsymbol{X}}$ differ from the corresponding control totals $\boldsymbol{X}$, as in the case of non-response or non-coverage. Our method automatically accounts for the g-weights and may lead to consistent variance estimators in such cases. Empirical results of Skinner with d'Arrigo in this context are very interesting. The case of variance estimators for alternative calibration estimators, studied in section 3.4, relative to customary variance estimators that replace $d_k(s) f(\boldsymbol{x}_k^T \hat{\boldsymbol{\lambda}})$ in the expression for $\hat{\boldsymbol{B}}_\lambda$ by $d_k(s)$ or $w_k(s)$ need further study, as noted by Skinner.

It may be noted that unit nonresponse is typically treated as second phase sampling (*e.g.*, Poisson sampling with unknown response probabilities) and Skinner notes that our method may lead to consistent variance estimators even when the estimators are based only on the sampling inclusion probabilities. However, control totals $\boldsymbol{X}$ are needed to get valid estimators of the total Y, under some assumptions on the response probabilities (Fuller 2002, equation (8.4)). We have extended our method to handle weight adjustment for unit nonresponse and imputation for item nonresponse when control totals are not available, assuming uniform response within classes (Demnati and Rao 2002). The resulting variance estimators are naturally more complex compared to Skinner's modification for unit nonresponse in the presence of control totals.

Turning to point (iv) on the possible use of numerical differentiation to calculate the linearized variables z_k, Woodroff and Causey (1976) used such a method to calculate the derivatives $\partial g(\boldsymbol{a})/\partial a_i|_{\boldsymbol{a}=\hat{\boldsymbol{Y}}}$ given in (1.4) when $\hat{\theta} = g(\hat{\boldsymbol{Y}})$. Skinner proposes perturbing each weight $d_k(s)$ in turn and then recalculating $\hat{\theta}$; for example, by replacing it by a fixed amount δ smaller than the minimum value of $d_k(s), k \in s$. He conjectures that the proposed approach should lead to variance estimators very similar to those obtained through analytical differentiation. It would be useful to study the statistical properties of the proposed approach to analytic differentiation of $f(\boldsymbol{d}(s))$ with respect to weights $d_k(s)$.

We hope the discussions by Kott, Shah and Skinner will stimulate further work on the approach to variance estimation presented in our paper.

REFERENCES

BERGER, Y.G., and SKINNER, C.J. (2003). Variance estimation for a low income proportion. *Applied Statistics*. 52, 457-468.

DEMNATI, A., and RAO, J.N.K. (2002). Linearization variance estimators for survey data with missing responses. *Proceeding of the Section Survey Research Methods*, American Statistical Association. 736-740.

FRANCISCO, C.A., and FULLER, W.A. (1991). Quantile estimation wih a complex survey design. *Annals of Statistics*. 19, 454-469.

SHAO, J. (1991). L-statistics in complex problems. Technical Report, University of Ottawa, Ottawa.

SHAO, J., and RAO, J.N.K. (1993). Standard errors for low income proportions estimated from stratified multistage samples. *Sankhyā*, Series B. 55, 393-414.

WOODRUFF, R.S., and CAUSEY, B.D. (1976). Computerized method for approximating the variance of a complicated estimate. *Journal of the American Statistical Association*. 71, 315-321.

Survey Methodology, June 2004
Vol. 30, No. 1, pp. 35-44
Statistics Canada

Weighting Sample Data Subject to Independent Controls

CARY T. ISAKI, JULIE H. TSAY and WAYNE A. FULLER[1]

ABSTRACT

In the U.S. Census of Population and Housing, a sample of about one-in-six of the households receives a longer version of the census questionnaire called the long form. All others receive a version called the short form. Raking, using selected control totals from the short form, has been used to create two sets of weights for long form estimation; one for individuals and one for households. We describe a weight construction method based on quadratic programming that produces household weights such that the weighted sum for individual characteristics and for household characteristics agree closely with selected short form totals. The method is broadly applicable to situations where weights are to be constructed to meet both size bounds and sum-to-control restrictions. Application to the situation where the controls are estimates with an estimated covariance matrix is described.

KEY WORDS: Raking; Regression; Quadratic programming; Coverage adjustment; Integer weights; Weighting area.

1. INTRODUCTION

Given the availability of known characteristic totals, it is common among survey practitioners to use such information in estimators of the post stratified, ratio and regression type. The known characteristic totals are sometimes called independent controls because they are derived outside of the survey situation. Use of independent controls tends to reduce the variance of most estimates. Independent controls also often compensate for coverage problems in surveys. See Deville and Särndal (1992) and Fuller (2002).

The U.S. decennial census utilizes a sample for the measurement of selected characteristics. The questionnaire for these characteristics is called the long form and the sample for the long form consists of a random sample of addresses. The long form questionnaire requests information that is asked of all individuals (called short form information) plus information on a set of additional characteristics. In previous Censuses, raking to controls based on short form information was used to construct weights for the long form sample. Two sets of sample weights were created, one for person characteristics and one for housing unit characteristics.

The set of categories used for person weighting was a classification of individuals by race, Hispanic origin, age and sex, family type, and household size. For households, the categories were the cross classification of race by Hispanic-origin-of-householder by tenure by household type and size. In the 1990 Census long form weighting process, persons and housing units were each classified by four sets of classifications for raking in four dimensions. When raking was completed, the long form sample weights were converted to integers. Integer weights are desirable because, unlike real weights, integer weights provide arithmetically consistent totals of integral characteristics. For details, see Schindler, Griffin and Swan (1992).

Long form weighting using short form census information is a part of the Canadian Census of population and housing. Unlike the procedure used by the U.S. Census Bureau (USCB), the procedure used at Statistics Canada constructs a single set of household weights using regression estimation. See Bankier, Houle and Luc (1997). Should the initial weights generated by the regression procedure exceed prescribed bounds, collapsing of cells defining explanatory variables is carried out. Linear dependencies and near linear dependencies among the explanatory variables are also removed by eliminating variables. See Bankier, Rathwell and Majkowski (1992).

Lemaître and Dufour (1987) used a generalized least squares estimator (GLS) to construct weights meeting person and household constraints. Alexander (1987) considers a procedure for constructing household weights in the census setting. One of his distance functions is similar to the one used in this paper.

The use of quadratic programming to compute regression weights in the survey context was suggested by Husain (1969). An application of quadratic programming (QP) in a Census environment is that in Isaki, Ikeda, Tsay and Fuller (2000) where household weights for Census households were obtained using person totals as controls. Motivation for the use of various distance functions can be found in these two papers and in Deville and Särndal (1992) who discuss a general class of estimators called calibration estimators. Fuller, Laughin and Baker (1994) consider a regression weight generation procedure that is modified so that all weights are positive and very large weights are made

[1] Cary T. Isaki and Julie H. Tsay, U.S. Bureau of the Census, Statistical Research Division, Washington, D.C. 20233, U.S.A. E-mail: Julie.Hsu.Ling.L.Tsay@census.gov; Wayne A. Fuller, Iowa State University, Department of Statistics, 221 Snedecor Hall, Ames, Iowa 50011, U.S.A.

smaller than the corresponding least squares weight. Jayasuriya and Valliant (1996) also consider a restricted regression. Fuller (2002) is a review of regression estimation.

Our proposed long form weighting method is a type of regression estimation and, like the Statistics Canada approach, provides a single set of household weights that maintain given independent controls. We generate household weights using quadratic programming with the restrictions that the weights fall within a specified range and that the weights maintain control totals. In the following, we refer to the suggested method as the quadratic programming method or QP.

2. THE QUADRATIC PROGRAMMING METHOD

The purpose of quadratic programming is to produce sample weights that i) are close to initial weights, ii) are within reasonable bounds, iii) maintain specified control totals and iv) produce a design consistent estimator. Apart from the bounds on the weight, the weights from quadratic programming are those of a simple regression estimator. We first describe the mathematical form of the QP and then discuss the implementation. Let

i) $\{W_i ; i = 1, 2, ..., n\}$ denote the set of final housing unit weights, where i denotes the i^{th} long form sample household and n is the size of the long form sample,

ii) $\{W_i^{(2)}; i = 1, 2, \ldots, n\}$ denote the set of initial housing unit weights,

iii) X_{ji}, $j = 1, 2, \ldots, m_p$, $i = 1, 2, \ldots, n$; denote the observation on the j^{th} person control variable for the i^{th} sample household,

iv) Z_{ji}, $j = 1, 2, \ldots, m_h$, $i = 1, 2, \ldots, n$; denote the observation on the j^{th} household control variable for the i^{th} sample household,

v) X_j, $j = 1, 2, ..., m_p$, denote the j^{th} person control,

vi) Z_j, $j = 1, 2, ..., m_h$, denote the j^{th} household control.

The quadratic programming method seeks W_i, $i = 1, 2, ..., n$, that minimize a quadratic objective function subject to linear constraints. In our application we minimize

$$g(W) = \sum_{i=1}^{n} \left(W_i - W_i^{(2)}\right)^2 \left[W_i^{(2)}\right]^{-1}, \qquad (1)$$

subject to

$$\sum_{i=1}^{n} W_i X_{ji} = X_j, \quad \text{for } j = 1, 2, ..., m_p, \qquad (2)$$

$$\sum_{i=1}^{n} W_i Z_{ji} = Z_j, \quad \text{for } j = 1, 2, ..., m_h, \qquad (3)$$

$$1 \le W_i \le K \qquad (4)$$

where the summations are over housing units in the long form sample. Observe that the long form household weights are bounded below by one. This is on the basis that an element in the sample should at least "represent" itself. In our program, K was set equal to 48 but the bound was never attained. The lower bound of one was attained. The FORTRAN subroutine from IMSL was used to solve the QP. Other programs, such as LCP of SAS®/IML, are available.

The USCB's current long form weighting procedure rakes the initially weighted long form sample counts to the census counts for the control categories. The weighting is done by subdivisions of the country called weighting areas and is done separately for person and household characteristics. The nominal sample rates for the long form are one-in-two, one-in-six, and one-in-eight. The nominal sampling weights are the inverses of the nominal sampling rates and are denoted by $W_i^{(1)}$. A second set of weights, denoted by $W_i^{(2)}$, are the realized sampling rates calculated for cells, where the cells are required to contain at least five sample households. For details on the USCB's procedures see Schindler *et al.* (1992).

Since we intend to compare the raking and QP methods, we use most of the USCB's person and household categories as the X_j and Z_j control totals in the quadratic program, but some changes were instituted. For example, while we maintained all of the age-race-sex person categories, we did not use a category based on the nominal sampling rates.

We used the USCB's specifications for determining whether a cell category would be retained as a separate control or would be combined with another cell and we used the USCB's procedure for determining the cells to be combined. This capitalized on the USCB's experience and minimized differences between the USCB's set of long form control totals and the set used by the QP method. The procedure used to define $W_i^{(2)}$ is given in the appendix.

Two possibilities exist for the control totals to be used in the construction of weights for the long form of the U.S. 2000 Census. One possibility is to use controls from the 2000 Census short form. That is, the independent controls to be maintained in long form weighting are those that are tabulated from the Census short form. When the Census is used as the control, the person control (X_j) categories include a cross classification of age and sex-race/ethnicity. Other characteristics, such as tenure, were used as additional

controls. The majority of the household control categories (Z_j) are defined by a cross classification of household type (*e.g.*, family with children under 18) and household size (*e.g.*, number of persons in the family). The Z_j also include race/ethnicity of the householder cross-classified by tenure.

The other possible set of controls for the 2000 Census is the set of estimates from the post enumeration survey, called the Accuracy and Coverage Evaluation (A.C.E.) survey. The A.C.E. survey is designed to estimate person characteristics only. The X_j for the A.C.E. include age-sex-race/ethnicity-tenure controls.

The last step in long form weighting is to round the W_i to integers. Integer weights prevent discrepancies between sets of estimates caused by rounding of real valued estimates. Sample housing units were grouped by race/ethnicity of the householder and by tenure. Then within each group, the sample was sorted by family type by household size. The weights were then rounded to integers using the cumulate-and-round procedure. Table 1 illustrates the method. The partial sums of the weights are formed (cumulated) as shown in the column CW. The partial sums are then rounded as shown in the column RCW. The integer weight for element i is the difference between successive entries $i-1$ and i in the RCW column.

Table 1
Illustration of Cumulate and Round

Sample Unit	Initial Weight	CW	RCW	Integer Weight
1	3.333	3.333	3	3
2	2.500	5.833	6	3
3	1.428	7.261	7	1
4	1.250	8.511	9	2
5	1.111	9.622	10	1
6	5.021	14.643	15	5

3. VARIANCE ESTIMATION

Variances of long form estimates were estimated using the jackknife method. In the numerical results using census controls, sixteen replicates were formed. Sixteen was chosen for convenience and a larger number could have been used. The long form sample was ordered by the census identification number within blocks and sixteen replicates were formed as the sixteen one-in-sixteen systematic samples. Sixty seven replicates were formed for the estimates using ACE controls.

3.1 Replicates for Census Controls

The jackknife replicate is created by deleting the i^{th} group of elements, computing the quadratic programming weights and rounding the weights to integers. Because of the rounding, the usual jackknife variance estimation procedure required modification. To isolate the effect of rounding, we consider the replicate estimate constructed with real-valued weights. Let

$\hat{\theta}_w$ = the sample estimator with weights rounded to integers,

$\hat{\theta}_R$ = the sample estimator with real-valued weights,

$\hat{\theta}_{R(i)}$ = jackknife replicate estimate with i^{th} group deleted and real-valued weights,

$\hat{\theta}_{w(i)}$ = jackknife replicate estimate with i^{th} group deleted weights rounded to integers,

and let

$$\bar{\theta}_w = r^{-1}\sum_{i=1}^{r}\hat{\theta}_{w(i)}, \qquad (5)$$

where r is the total number of replicates. Then the jackknife deviation for the estimator with integer weights can be decomposed as

$$\hat{\theta}_{w(i)} - \bar{\theta}_w = \hat{\theta}_{R(i)} - \hat{\theta}_R + \left[\hat{\theta}_{w(i)} - \bar{\theta}_w - (\hat{\theta}_{R(i)} - \hat{\theta}_R)\right]. \qquad (6)$$

We assume that the error in the rounding operation is independent of the group chosen for deletion, a reasonable assumption, given that the deletion produces an entire new set of weights to be rounded. Then

$$E\left\{(\hat{\theta}_{w(i)} - \bar{\theta}_w)^2\right\} \doteq E\left\{(\hat{\theta}_{R(i)} - \hat{\theta}_R)^2\right\} + E\left\{\left[(\hat{\theta}_{w(i)} - \hat{\theta}_{R(i)}) - (\bar{\theta}_w - \hat{\theta}_R)\right]^2\right\}. \qquad (7)$$

Assume that the average of the $\hat{\theta}_{R(i)}$ is equal to $\hat{\theta}_R$. Then the last term of (7) is a replicate deviation for the difference between the real and rounded estimates. Then

$$E\left\{\left[\left(\hat{\theta}_{w(i)} - \hat{\theta}_{R(i)}\right) - \left(\bar{\theta}_w - \hat{\theta}_R\right)\right]^2\right\} = r^{-1}(r-1)V\left\{\hat{\theta}_{w(i)} - \hat{\theta}_{R(i)}\right\} = V\left\{\hat{\theta}_w - \hat{\theta}_R\right\} \qquad (8)$$

where $V\{\hat{\theta}_{w(i)} - \hat{\theta}_{R(i)}\}$ is the variance due to rounding for a sample of $r-1$ groups and $V\{\hat{\theta}_w - \hat{\theta}_R\}$ is the variance due to rounding for a sample of r groups. In obtaining (8) we assumed the variance due to rounding for a sample of r groups is the variance for $r-1$ groups multiplied by $r^{-1}(r-1)$. Thus

$$E\left\{(r-1)^{-1}\sum_{i=1}^{r}(\hat{\theta}_{w(i)}-\bar{\theta}_w)^2\right\}\doteq E\left\{(r-1)^{-2}r\hat{V}_R\{\hat{\theta}_R\}\right\} + V\{\hat{\theta}_w-\hat{\theta}_R\}, \quad (9)$$

where

$$\hat{V}_R\{\hat{\theta}_R\} = r^{-1}(r-1)\sum_{i=1}^{r}(\hat{\theta}_{R(i)}-\hat{\theta}_R)^2$$

is the jackknife variance estimator for the estimator with real weights. Then an estimator of the variance due to rounding is

$$\hat{V}\{\hat{\theta}_w-\hat{\theta}_R\} = r^{-1}(r-1)\left[\begin{array}{c}(r-1)^{-1}\sum_{i=1}^{r}(\hat{\theta}_{w(i)}-\bar{\theta}_w)^2 \\ -r(r-1)^{-2}\hat{V}_R\{\hat{\theta}_R\}\end{array}\right] = r^{-1}\left[\begin{array}{c}\sum_{i=1}^{r}(\hat{\theta}_{w(i)}-\bar{\theta}_w)^2 \\ -r(r-1)^{-1}\hat{V}_R\{\hat{\theta}_R\}\end{array}\right]. \quad (10)$$

Based on these results, the estimated variance for the rounded estimator is

$$\hat{V}\{\hat{\theta}_w\} = (r-1)^{-1}(r-2)\hat{V}_R\{\hat{\theta}_R\} + r^{-1}\sum_{i=1}^{r}(\hat{\theta}_{w(i)}-\bar{\theta}_w)^2. \quad (11)$$

3.2 Replicates for A.C.E. Controls

The replicates for estimates constructed with A.C.E. controls were modified so that the estimated variances contained a component for the error in the A.C.E. estimates. The data in a weighting area were assigned to 67 replicates where 67 is the number of controls. The procedure requires the number of replicates to equal or exceed the number of controls if the covariance matrix of the estimated control totals is to be reproduced. More replicates than controls can be used. See Fuller (1998).

The estimator of the total of a characteristic for the long form is a type of regression estimator using the A.C.E. numbers as controls. We write the estimator for the total based on real valued weights as

$$\hat{\theta}_R = \hat{\mathbf{X}}_A\hat{\boldsymbol{\beta}}, \quad (12)$$

where $\hat{\mathbf{X}}_A$ is the vector of A.C.E. estimates and $\hat{\boldsymbol{\beta}}$ is the regression coefficient computed with the long form data.

Let $\hat{\mathbf{V}}_{AA}$ be the $r \times r$ covariance matrix of the vector of A.C.E. controls, where $\hat{\mathbf{V}}_{AA}$ is estimated as part of the A.C.E. process, and $r = 67$. Let $\lambda_1, \lambda_2, ..., \lambda_r$ be the roots of $\hat{\mathbf{V}}_{AA}$ and let

$$\mathbf{Q}'\mathbf{V}_{AA}\mathbf{Q} = \boldsymbol{\Lambda}, \quad (13)$$

where $\boldsymbol{\Lambda} = \text{diag}(\lambda_1, \lambda_2, ..., \lambda_r)$, $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_r$, and $\mathbf{Q}$ is the matrix composed of the characteristic vectors of $\hat{\mathbf{V}}_{AA}$. Recall that

$$\hat{\mathbf{V}}_{AA} = \mathbf{Q}\boldsymbol{\Lambda}\mathbf{Q}'$$

and

$$\hat{\mathbf{V}}_{AA} = \sum_{j=1}^{r}\mathbf{q}_{\bullet j}\lambda_j\mathbf{q}'_{\bullet j} =: \sum_{j=1}^{r}\mathbf{z}_{\bullet j}\mathbf{z}'_{\bullet j} \quad (14)$$

where $\mathbf{q}_{\bullet j}$ is the j^{th} column of $\mathbf{Q}$ and $\mathbf{z}_{\bullet j} = \lambda_j^{1/2}\mathbf{q}_{\bullet j}$.

Using result (14), controls for the r replicates were constructed as

$$\ddot{\mathbf{X}}_{A(i)} = \hat{\mathbf{X}}_A + c\mathbf{z}'_{\bullet i}, \quad i = 1, 2, ..., r, \quad (15)$$

where $\hat{\mathbf{X}}_A$ is the row vector of the original controls and c is a constant. The constant c is determined so that the expectation of the sum of the jackknife squared deviations for the elements of the vector $\mathbf{X}$ are the diagonal elements of $\hat{\mathbf{V}}_{AA}$. In our application, the constant c is $(r-1)^{-1/2}r^{1/2}$ and

$$(r-1)r^{-1}\sum_{j=1}^{r}c^2\mathbf{z}_{\bullet j}\mathbf{z}'_{\bullet j} = \sum_{j=1}^{r}\mathbf{z}_{\bullet j}\mathbf{z}'_{\bullet j} = \hat{\mathbf{V}}_{AA}. \quad (16)$$

Thus, if the characteristic being "estimated" is one of the controls used in the QP, the jackknife procedure returns the A.C.E. estimated variance for that characteristic. The $\mathbf{z}_{\bullet j}$ are assigned at random to the r replicates.

Using the regression representation, we write the estimator for the i^{th} replicate as

$$\ddot{\theta}_{R(i)} = \ddot{\mathbf{X}}_{A(i)}\hat{\boldsymbol{\beta}}_{(i)} = \hat{\mathbf{X}}_A\hat{\boldsymbol{\beta}}_{(i)} + \left(\ddot{\mathbf{X}}_{A(i)}-\hat{\mathbf{X}}_A\right)\hat{\boldsymbol{\beta}}_{(i)} =: \hat{\theta}_{R(i)} + c\mathbf{z}'_{\bullet i}\hat{\boldsymbol{\beta}}_{(i)}, \quad (17)$$

where $\ddot{\theta}_{R(i)}$ is the real-valued estimator computed with the i^{th} group deleted using $\ddot{\mathbf{X}}_{A(i)}$ as the control vector, $\hat{\boldsymbol{\beta}}_{(i)}$ is the regression coefficient computed with the i^{th} group deleted, and $\hat{\theta}_{R(i)}$ is the real-valued estimator computed with the i^{th} group deleted using $\hat{\mathbf{X}}_A$ as the control vector. Then

$$\ddot{\theta}_{R(i)} - \hat{\theta}_R = \hat{\theta}_{R(i)} - \hat{\theta}_R + c\mathbf{z}'_{\bullet i}\hat{\boldsymbol{\beta}}_{(i)}.$$

Because $\mathbf{q}_{\bullet j}$ are assigned to replicates at random, the expectation of the replicate variance estimator for the real-valued estimator based on A.C.E. controls is

$$E\left\{\hat{V}_R\left(\hat{\theta}_R\right)\right\} = E\left\{ r^{-1}(r-1)\sum_{i=1}^{r} (\ddot{\theta}_{R(i)} - \hat{\theta}_R)^2 \right\}$$

$$= E\left\{ r^{-1}(r-1)\sum_{i=1}^{r} (\hat{\theta}_{R(i)} - \hat{\theta}_R)^2 \right\} + E\left\{\hat{\boldsymbol{\beta}}'_{(i)}\hat{\mathbf{V}}_{AA}\hat{\boldsymbol{\beta}}_{(i)}\right\}. \quad (18)$$

Now, assuming $E\{\hat{\mathbf{V}}_{AA}\} = \mathbf{V}_{AA}$, $E\{\hat{\boldsymbol{\beta}}_{(i)}\} = \boldsymbol{\beta}$, and that $\hat{\mathbf{V}}_{AA}$ is independent of $\hat{\boldsymbol{\beta}}_{(i)}$,

$$E\left\{\hat{\boldsymbol{\beta}}'_{(i)}\hat{\mathbf{V}}_{AA}\hat{\boldsymbol{\beta}}_{(i)}\right\} = \boldsymbol{\beta}'\mathbf{V}_{AA}\boldsymbol{\beta} + tr\left\{V\{\hat{\boldsymbol{\beta}}_{(i)}\}\mathbf{V}_{AA}\right\},$$

where $\text{tr}\{\mathbf{V}_{AA}\}$ is the trace of the matrix. It follows that

$$E\left\{ r^{-1}(r-1)\sum_{i=1}^{r} (\ddot{\theta}_{R(i)} - \hat{\theta}_R)^2 \right\}$$

$$= E\left\{ r^{-1}(r-1)\sum_{i=1}^{r} (\hat{\theta}_{R(i)} - \hat{\theta}_R)^2 \right\} + \boldsymbol{\beta}'\mathbf{V}_{AA}\boldsymbol{\beta} + O(n^{-2}), \quad (19)$$

where we assume $\text{tr}\{\mathbf{V}_{AA}\} = O(n^{-1})$ and $\text{tr}[V\{\hat{\boldsymbol{\beta}}_{(i)}\}] = O(n^{-1})$, where n is the sample size. The first term on the right of the equality in (19) is the expectation of the variance estimator for the variance due to the sampling of long forms from the census. The second term is the contribution of the variance of the error in the A.C.E. estimates to the total variance. Thus, the variance estimator based on $\ddot{\theta}_{R(i)}$ estimates both components of variation. Observe that the estimated covariance matrix for the controls is $\hat{\mathbf{V}}_{AA}$, as it should be.

4. NUMERICAL RESULTS

We used the USCB's 1990 Census data file to illustrate the application of the QP method to actual data. The file provides data for households and for persons in households, together with long form weights as developed for the 1990 U.S. Census. Hence, the file provides data appropriate for comparing the performance of the USCB's 1990 long form weighting method with the QP method.

The USCB long form sample weighting is done by weighting area, where the weighting areas usually contain two to three thousand housing units. There were about 56,000 weighting areas in the U.S. in 1990. For our numerical work we chose weighting area (WA) 1788 that contains 8,034 occupied housing units and 25,145 persons.

In Table 2 we provide estimates of some person and housing unit characteristics for weighting area 1788. The characteristics in the table, except the number of rented units, were suggested by subject matter personnel at the USCB. In Table 2, Est.(H) is the long form sample weighted estimate computed with housing unit weights, Est.(P) is the long form sample weighted estimate computed with person weights. The quadratic programming estimator constructed with Census controls is called QP in the table, while QPG is used to denote the generalization of the quadratic programming estimator with objective function (20). The QPG estimator is discussed subsequently. The USCB housing unit estimates in Table 2 that are based on person weights were created by using the householder weight as the housing unit weight. Every occupied unit contains a single householder. The householder procedure is called the *principal person method* by Alexander (1987). All estimates in the table are given as a percent of the census count.

Estimates constructed by the two USCB methods can differ by several percentage points with the differences between Est.(P) and Est.(H) for rented units, persons aged 0 to 4 years, persons aged 65 and over, Hispanic, Asian, and persons in rented units being noticeable. The Est.(H) estimate for persons in rented units is closer to 100 than the Est.(P) estimate.

The cell collapsing rules produced 45 person and 22 housing unit controls for WA 1788. An example of a person control is the total number of Non-Hispanic Black males aged 65 and over, while an example of a housing unit control is the total number of Non-Hispanic White owned housing units. Total Black persons is an implicit control in WA 1788. Controls for total persons 18-44, total persons 45-64, total males, total renters and total number of rented housing units were added to the QP. Apart from the controls mentioned above, none of the remaining characteristics in Table 2 is also used as a control in the QP procedure.

The QP estimates and standard errors of the QP estimates are given, as a percent of the census counts, in the fourth and fifth columns of Table 2. The agreement between count and QP estimates for household characteristics are comparable to the USCB household based estimates and superior to USCB person based estimates. For person counts, the QP estimates are generally closer to the census counts than either of the USCB raking estimates.

The largest difference between a QP estimate and the census count relative to the standard error is for the estimate of the number of households with own children present, where the difference is about 1.6 standard errors. The majority of the QP estimates differ from the census count by less than one standard error. A number of the USCB person estimates deviate from the census count by more than one QP standard error.

Table 2
Estimated Occupied Housing Unit and Person Characteristics for WA 1788

	Census Count	Est.(H)* Count	(%)	Est.(P)** Count	(%)	QP† Count	(%)	se (QP) Count	(%)	QPG†† Count	(%)	se (QPG) Count	(%)
Housing unit characteristics													
With Own Children	4,349	100.18		100.45		100.21		0.13		100.18		0.14	
Not With Own Children	3,685	99.78		99.67		99.76		0.15		99.78		0.16	
With 1 to 4 Persons	6,785	100.00		100.57		100.04		0.05		100.07		0.05	
With 5+ Persons	1,249	100.00		97.51		99.76		0.30		99.60		0.30	
Rented Unit	2,559	100.00		95.97		100.00		0.19		99.92		0.16	
Owned Unit	5,475	100.00		102.02		100.00		0.09		100.04		0.08	
Person characteristics													
Age 0–4 years	2,493	101.92		97.95		98.84		1.68		99.96		0.29	
Age 5–17 years	6,339	103.91		101.07		100.63		0.71		99.98		0.18	
Age 18–44 years	12,711	99.50		99.69		100.01		0.05		100.00		0.06	
Age 45–64 years	3,028	101.65		101.95		99.90		0.09		99.97		0.09	
Age 65+ years	574	81.18		93.73		100.17		0.85		100.00		027	
Males	12,473	99.95		99.64		100.06		0.08		99.98		0.09	
Females	12,672	101.43		100.36		99.95		0.10		100.01		0.09	
Hispanic	2,385	95.38		103.40		99.96		0.38		99.87		0.38	
Not Hispanic	22,760	101.25		99.64		100.03		0.07		100.00		0.10	
Black	1,285	101.08		101.79		100.86		1.22		99.77		0.54	
White	22,372	100.69		99.91		100.03		0.07		100.00		0.10	
Asian	257	92.60		80.05		96.83		2.32		99.76		0.50	
Remainder	1,231	101.94		103.89		105.84		9.54		100.78		1.75	
In Rented Unit	7,978	102.04		95.41		100.01		0.24		99.92		0.19	
In Owned Unit	17,167	100.06		102.13		100.00		0.09		100.02		0.13	

* USCB weights for households
** USCB weights for persons
† QP weights with 82 constraints
†† Generalized QP with 13 constraints and objective function (20)

Because the number of rented units, persons aged 18-44, persons aged 45-64, males, and persons in rented units were used as controls in the QP procedure, differences between QP estimates and census totals for those categories are due to rounding. The standard errors demonstrate that the rounding can lead to sizeable deviations from the controls.

The 45 person and 22 housing unit control totals obtained by the collapsing rules are such that a margin estimate, such as total males, may not be constrained to agree with the count. In addition, for different weighting areas, USCB's collapsing procedure gives different person and housing unit constraints. Thus we considered adding some margin totals

to the set of control totals. To reduce the impact of the added controls on the weights, we replaced the original constraints with additional terms in the objective function. The terms are deviations between the final estimates and the control totals. The objective function becomes

$$G(W)=g(W)+\sum_{j=1}^{67}\alpha_j\left(\sum_i W_i X_{ji}-X_j\right)^2, \quad (20)$$

where $g(W)$ is defined in expression (1), the $\{X_{ji}, j=1,2,\ldots,67\}$ is the set of auxiliary variables defining the 45 person and 22 housing unit controls, and α_j are constants to be specified. The X_{ji} for category j of household i for a person characteristic is the number of individuals in category j in the housing unit. The X_{ji} for a housing unit characteristic is one if the housing unit has the characteristic and zero otherwise. In our application, the function is minimized subject to two household controls and eleven person controls. The housing unit controls are rented housing units and owned housing units. The person controls are persons 0 to 4 years, persons 5 to 17 years, persons 18 to 44 years, persons 45-64 years, persons 65 years and over, males, black, white, Asian, Hispanic, and renters. The α_j are $10[\overline{W}^{(2)}]^{-1}[\sigma_j^2]^{-1}$, where $\overline{W}^{(2)}=8.95$ is the mean of the $W_i^{(2)}$, $\sigma_j^2=P_j(1-P_j)$, and P_j is the proportion of the population in cell j. The α_j would minimize the mean square error of an estimated total if there was a single control variable and the squared correlation between the control variable and the dependent variable was about 0.9. Thus, the function exerts considerable pressure for the final estimate to be close to the control total.

The QP solution to (20) gives a type of regression estimator. See Fuller (2002) and Fuller and Isaki (2001). Rao and Singh (1997) and Bardsley and Chambers (1984) consider related estimators.

Using $G(W)$ of (20) and the 13 linear constraints, the results in the final two columns of Table 2, under the heading "QPG", were obtained. As expected, the estimates are close to Census totals because the Census marginals were used as constraints. The relative percent differences between the QP estimate and the census count for the 67 characteristics in $G(W)$ of (20) ranged from –3.50% to 3.75% with about 50 of the differences being less than one percent.

The sample weights obtained by the two programming approaches are compared to those of the USCB's household raking method in Table 3. The number and type of controls used under the USCB raking was not determined exactly because the number depends on the execution of the USCB collapsing procedure and on some preliminary files that are not readily available. However, we believe the number to be about 67 because the collapsing procedure used to form the 67 cells is basically that used by the USCB. The QP procedure used 82 controls and the QPG procedure used 90 controls. The range of weights for the two QP methods are similar with a smaller range for raking. There are modest differences among the three sums of squares of the weights. The $g(W)$ values are also similar, with the value for (20) being the largest. The $g(W)$ value is the quantity being minimized by the weights of the first line of the table. The sum of squares of the weights for the QP of (20) could be reduced by reducing the α_j in the objective function.

We also used data from the 1990 Census to simulate the situation in which the controls come from adjusted census counts. For 1990, person estimates from the 1990 Post Enumeration Survey are available, but there are no housing unit estimates based on that survey. We call these estimates A.C.E. estimates. See Hogan (1993) and Isaki, Tsay and Fuller (2000). Estimates for WA 1788 were created by the QP method, using the A.C.E. estimates as controls. We used $G(W)$ of (20) as the objective function with 63 age-race-sex-tenure person characteristics in the second term of the objective function and 11 person constraints. The person constraints are persons 0 to 4 years, 5 to 17 years, 18 to 44 years, 45 to 64 years, 65 and over, total males, total Hispanic, total Black, total White, total Asian and total persons in rented units.

Table 3
Properties of Long Form Housing Unit Sample Weights in WA 1788

Method	Minimum Weight	Maximum Weight	$\sum_i W_i^2$	$g(W)$
QP with $g(W)$ of (1) 72 constraints	1	26.5	78,028	326
QP with $G(W)$ of (20) 13 exact constraints	1	29.9	78,672	383
Raking	4	22	77,000	369

Table 4 contains the estimates for WA 1788 identified as QPG and given as a percent of the census counts. The QPG estimates for these eleven person characteristics agree with the A.C.E estimates, except for rounding error. The standard errors reflect the error in the A.C.E estimates and, hence, are much larger than the standard deviation of rounding error. For example, the rounding error standard deviation for persons 18 - 44 is 0.06 in Table 2, while the standard error for the ACE estimate of persons 18 - 44 is 0.63. The QP estimates for household characteristics seem very reasonable. The estimated total number of households is 1.8% larger than the census count while the A.C.E. estimated number of persons is 2.0% larger than the census count. The quadratic programming total number of persons differs slightly from the A.C.E. estimate because of rounding of the weights. The difference is about 7% of the standard error.

Table 4
The Census Count, A.C.E. Estimates and QP Estimates with A.C.E. Controls – WA 1788

	Census Count	$\frac{\text{A.C.E}}{\text{Count}}$	$\frac{\text{QPG}}{\text{Count}}(\%)$	$\frac{\text{s.e.}(\text{QPG})}{\text{Count}}(\%)$
Housing unit characteristics				
With Own Children	4,349	–	101.89	2.09
Not With Own Children	3,685	–	101.66	3.07
With 1 to 4 Persons	6,785	–	102.03	2.03
With 5^+ Persons	1,249	–	100.40	5.92
Rented Unit	2,559	–	104.57	2.62
Owned Unit	5,475	–	100.47	1.50
Total	8,034	–	101.78	1.22
Person characteristics				
Age 0–4 years	2,493	103.17	102.81	1.00
Age 5–17 years	6,339	103.09	103.08	0.96
Age 18–44 years	12,711	101.67	101.67	0.63
Age 45–64 years	3,028	100.26	100.33	0.59
Age 65^+years	574	99.48	98.95	0.70
Males	12,473	102.18	102.01	0.68
Females	12,672	101.74	101.82	0.62
Hispanic	2,385	104.95	104.91	1.09
Not Hispanic	22,760	101.64	101.60	0.60
Black	1,285	104.59	104.82	1.01
White	22,372	101.69	101.69	0.61
Asian	257	100.00	101.95	1.95
Remainder	1,231	104.47	102.92	1.14
In Rented Unit	7,978	104.25	104.21	0.89
In Owned Unit	17,167	100.89	100.84	0.68
Total	25,145	101.96	101.91	0.57

5. CONCLUSIONS

The QP method is shown to work well on actual USCB long form data. The QP single household weight method possesses several advantages over the USCB separate weights method. With one set of weights, there will be no confusion as to which weights to use for estimating a given characteristic. Also, estimates of relationships such as ratios of person characteristics to household characteristics are expected to be less variable when a single set of weights is used for both characteristics.

Given that a single set of weights is easier to compute and easier for analysts to use, one would only construct two sets of weights if the weights designed for one type of characteristic give estimates with smaller variance for that type of characteristic. This did not seem to be the case in our example. The single set of QP weights gave favorable

results for both household and person characteristics when compared with the USCB weights for the specific category.

The QP estimation module is computationally feasible and can replace the raking estimation module in the USCB operational setting. The QP method can produce long form sample weights for households in an adjustment situation in which only person controls are available.

ACKNOWLEDGEMENTS

This article reports the results of research and analysis undertaken by the authors. It has undergone a more limited review than official U.S. Census Bureau publications. Research results and conclusions expressed are those of the authors and have not been endorsed by the U.S. Census Bureau. The report is released to inform interested parties of research and to encourage discussion.

This research was partly supported by Cooperative Agreement 43-3AEU-3-80088 between Iowa State University, the National Agricultural Statistics Service and the U.S. Bureau of the Census.

We gratefully acknowledge the comments of the Associate Editor and two referees which led to a much improved paper.

APPENDIX

Procedure used to define cells and initial weights $W_i^{(2)}$

We used the USCB's procedure to determine the order in which cells are combined (collapsed). The cell collapsing rules specify that each cell contain at least 5 sample households. The procedure below is our extension of the USCB rules for defining $W_i^{(2)}$.

Let two cells under consideration be identified as Cell 1 and Cell 2.

i) Cell 1 is not to be collapsed and $n_1^{-1}N_1 \le B$, where N_1 is the Census count of households in Cell 1 and n_1 is the long form sample count in Cell 1. The constant B is provided by the sponsor and in our work, 27 is used. For household i in Cell 1, let

$$W_i^{(2)} = \max\{1.2, \ddot{W}_i\}, \qquad \text{(A.1)}$$

where $\ddot{W}_i = \min\{Q_1 W_i^{(1)}, B\}$,

$$Q_1 = \left[\sum_{i \in A_1} W_i^{(1)}\right]^{-1} N_1,$$

and A_1 is the set of indices in Cell 1. The number 1.2 is an arbitrary lower bound chosen greater than one and less than the minimum of $W_i^{(1)}$ which is two. Note that the $W_i^{(2)}$ provides reasonable estimated totals for Cell 1. If $n_1^{-1}N_1 > B$, collapse cell 1 with cell 2 as in ii) below.

ii) Cells 1 and 2 are designated for collapse, $(n_1 + n_2)^{-1}(N_1 + N_2) \le B$, $n_1 + n_2 \ge 5$, and $n_1^{-1}N_1 > n_2^{-1}N_2$. Then for i in Cell 1, $W_i^{(2)}$ is defined by (A.1). For i in Cell 2,

$$W_i^{(2)} = \max\{1.2, \ddot{W}_i\},$$

where

$$\ddot{W}_i = \min\{Q_2 W_i^{(1)}, B\},$$

$$Q_2 = \left[\sum_{i \in A_2} W_i^{(1)}\right]^{-1} (N_1 + N_2 - \hat{N}_1),$$

and

$$\hat{N}_1 = \sum_{i \in A_1} W_i^{(2)}.$$

The $W_i^{(2)}$ in $A_1 \cup A_2$, the union of cells 1 and 2, maintains the total households in $A_1 \cup A_2$ and also provide an estimated total for Cell 1 that is reasonably close to the true total.

iii) Cells 1 and 2 are designated for collapse, $n_1 + n_2 \ge 5$, and $(n_1 + n_2)^{-1}(N_1 + N_2) > B$. Then it is necessary to initiate further collapsing. The combined cell becomes the Cell 1 of case (ii). Continue cell collapsing until $(n_1 + n_2 + ..)^{-1}(N_1 + N_2 + ..) \le B$. Case (iii) was not observed in the study data set.

One could repeat the weight construction procedure in an iterative manner by using the $W_i^{(2)}$ as $W_i^{(1)}$ in a second cycle. We tried a second cycle on the data described in the text. There was no discernable improvement in the estimates from using a second cycle.

REFERENCES

ALEXANDER, C.H. (1987). A class of methods for using person controls in household weighting. *Survey Methodology*. 13, 183-198.

BANKIER, M.D., RATHWELL, S. and MAJKOWSKI, M. (1992). Two step generalized least squares estimation in the 1991 canadian census. Working Paper-Methodology Branch, Census operations section, Social Survey Methods Division, Statistics Canada. SSMD92-007E.

BANKIER, M., HOULE, A.M. and LUC, M. (1997). Calibration estimation in the 1991 and 1996 canadian censuses. *Proceedings of the Section on Survey Research Methods,* American Statistical Association. 66-75.

BARDSLEY, P., and CHAMBERS, R.L. (1984). Multipurpose estimation from unbalanced samples. *Applied Statistics.* 33, 290-299.

DEVILLE, J., and SÄRNDAL, C.-E. (1992). Calibration estimators in survey sampling, *Journal of the American Statistical Association.* 87, 376–382.

FULLER, W.A. (1998). Replication variance estimation for two phase samples, *Statistica Sinica.* 8, 1153–1164.

FULLER, W.A. (2002). Regression estimation for survey samples. *Survey Methodology.* 28, 5-23.

FULLER, W.A., and ISAKI, C.T. (2001). Estimation using estimated coverage in a census. Presented at the CAESAR conference, June, Rome, Italy.

FULLER, W.A., LAUGHIN, M.M. and BAKER, H.D. (1994). Regression weighting in the presence of nonresponse with application to the 1987-1988 Nationwide Food Consumption Survey. *Survey Methodology.* 20, 75-85.

HOGAN, H. (1993). The 1990 post-enumeration survey: operations and results. *Journal of the American Statistical Association.* 88, 1047-1060.

HUSAIN, M. (1969). Construction of regression weights for estimation in sample surveys. Unpublished M.S. Thesis, Iowa State University, Ames, Iowa.

ISAKI, C.T., TSAY, J.H. and FULLER, W.A. (2000). Estimation of census adjustment factors. *Survey Methodology.* 26, 31–42.

ISAKI, C.T., IKEDA, M.M., TSAY, J.H. and FULLER, W.A. (2000). An estimation file that incorporates auxiliary information. *Journal of Official Statistics.* 16, 155–172.

JAYASURIYA, B.R., and VALLIANT, R. (1996). An application of restricted regression estimation in a household survey. *Survey Methodology.* 22, 127-137.

LEMAÎTRE, G., and DUFOUR, J. (1987). An integrated method for weighting persons and families. *Survey Methodology.* 13, 199-207.

RAO, J.N.K., and SINGH, A.C. (1997). A ridge-shrinkage method for range-restricted weight calibration in survey sampling. *Proceedings of the Section on Survey Research Methods,* American Statistical Association. 57-65.

SCHINDLER, E., GRIFFIN, R. and SWAN, C. (1992). Weighting the 1990 census sample. *Proceedings of the Section on Survey Research Methods,* American Statistical Association. 664–669.

Survey Methodology, June 2004
Vol. 30, No. 1, pp. 45-55
Statistics Canada

Properties of the Weighting Cell Estimator Under a Nonparametric Response Mechanism

D. NASCIMENTO DA SILVA and JEAN D. OPSOMER[1]

ABSTRACT

The weighting cell estimator corrects for unit nonresponse by dividing the sample into homogeneous groups (cells) and applying a ratio correction to the respondents within each cell. Previous studies of the statistical properties of weighting cell estimators have assumed that these cells correspond to known population cells with homogeneous characteristics. In this article, we study the properties of the weighting cell estimator under a response probability model that does not require correct specification of homogeneous population cells. Instead, we assume that the response probabilities are a smooth but otherwise unspecified function of a known auxiliary variable. Under this more general model, we study the robustness of the weighting cell estimator against model misspecification. We show that, even when the population cells are unknown, the estimator is consistent with respect to the sampling design and the response model. We describe the effect of the number of weighting cells on the asymptotic properties of the estimator. Simulation experiments explore the finite sample properties of the estimator. We conclude with some guidance on how to select the size and number of cells for practical implementation of weighting cell estimation when those cells cannot be specified a priori.

KEY WORDS: Finite population asymptotics; Quasi-randomization inference; Weighting cell selection.

1. INTRODUCTION

Item and unit nonresponse occur in almost all large-scale surveys, and proper estimation techniques need to account for it. While item nonresponse is often dealt with through imputation, unit nonresponse is most often accounted for through weighting adjustments. Cell weighting adjustments for nonresponse have been applied since at least the 1950s in survey estimation, *e.g.* U.S. Bureau of the Census (1963, page 53), and continue to be widely used in practice today, because they have intuitive appeal and are relatively easy to implement in practice. Reviews of common weighting procedures are given in Kalton (1983) and Kalton and Kasprzyk (1986). A number of authors have studied the properties of the weighting cell estimator under a variety of theoretical frameworks. Oh and Scheuren (1983) derive the mean and variance of the weighting cell estimator under simple random sampling, conditional on the sample size and the number of respondents in each cell. See also Kalton and Maligalig (1991). Särndal, Swensson and Wretman (1992, page 578) use the term "response homogeneity group" for cells in which the nonresponse is assumed to be constant, and derive the properties of the resulting weighting cell estimator for general designs. The recently introduced *fully efficient fractional imputation* (FEFI) of Kim and Fuller (1999) can also be expressed as a weighting cell estimator, and these authors derive its model properties under the assumption that the variables are independent and identically distributed (iid) within each cell.

While the specific assumptions vary, a common thread among all these results is that the weighting cells are correctly specified, in the sense that units within each cell are indeed fully "exchangeable" (the precise definition of this term depends on the framework selected: equal response probabilities for randomization-based inference, or iid observations for model-based inference). In the terminology of Little and Rubin (2002, Chapter 1), this is the case of observations *missing at random* (MAR), where auxiliary information (*i.e.*, cell membership in this case) can be used to correct the inference for the nonresponse.

In this article, we depart from this framework. We will assume that the response mechanism depends on a known continuous auxiliary variable, but the exact functional form of this relationship is left almost completely unspecified (details on this *nonparametric response mechanism* are provided in the next section). Knowledge of such a variable could be used to construct more sophisticated nonresponse adjustments such as *propensity weighting* (Cassel, Särndal and Wretman (1983), Little (1986), and Da Silva and Opsomer (2003)) or post-stratification, but we will instead limit our use of this auxiliary variable to the division of the population into weighting cells. Our primary goal with this approach is to study the robustness of the popular weighting cell estimator to model misspecification, and in particular, the effect of the number of cells. Hence, in contrast to the approach of the authors discussed above, the weighting cells are used as a practical way to construct an survey estimator, but they will not be assumed as part of the

[1] D. Nascimento Da Silva, Departamento de Estatística, Universidade Federal do Rio Grande do Norte, Natal, RN 59072-970, Brazil. E-mail: damiao@ccet.ufrn.br; Jean D. Opsomer, Department of Statistics, Iowa State University, Ames IA 50011, U.S.A. E-mail: jopsomer@iastate.edu.

statistical framework. This is similar to the "adjustment by subclassification" idea proposed by Cochran (1968) for removing the bias due to a continuous covariate in observational studies.

We will study the properties of the estimator under *quasi-randomization*, a term used by Oh and Scheuren (1983) to denote joint inference under the sampling design and the response mechanism. The asymptotic properties of the estimator will be established by embedding the finite population and the corresponding sampling design and response mechanism in a sequence of such populations and random mechanisms, as will be explained in later sections. This asymptotic framework is very similar to that advocated by Hansen, Madow and Tepping (1983) and used in Isaki and Fuller (1982), among others.

The remainder of this paper is as follows. In section 2, we introduce the notation and framework for the sampling design and the nonresponse model, and discuss the weighting cell estimator. In the following section, we derive the asymptotic design properties of the estimator. In section 4, we report on a simulation study to examine the practical behavior of the estimator, compare its practical behavior with that predicted by the asymptotic theory, and provide some guidance on the choice of the weighting cells.

2. THE WEIGHTING CELL ESTIMATOR

Before describing the weighting cell estimator, we introduce our survey design framework and the response generating mechanism. We consider a population $U = \{1, 2, ..., N\}$, where N is finite and known. For every element i in U, let $\boldsymbol{Y}_i = (Y_{1,i}, Y_{2,i}, ..., Y_{p,i})$ be the associated vector of values of p characteristics of interest, $Y_1, Y_2, ..., Y_p$. Likewise, let $\boldsymbol{X}_i = (X_{1,i}, X_{2,i}, ..., X_{q,i})$ be the vector of values of q auxiliary variables, $X_1, X_2, ..., X_q$, corresponding to the i^{th} unit, $i \in U$. We assume that $\boldsymbol{X}_i$ is known $\forall i \in U$. If $p = 1$, we denote $\boldsymbol{Y}_i$ by Y_i and, for $q = 1$, X_i is used to denote $\boldsymbol{X}_i$. Let s represent a sample drawn from U according to some sampling design $p(\cdot)$. This sampling design $p(\cdot)$ is chosen by the survey sampler and may be based on information available in the $\boldsymbol{X}_i, i \in U$.

The goal of the sample survey is to estimate unknown population quantities such as the population mean or total, or a function of these quantities. To simplify the presentation, we will focus on the estimation of the population total of the $\boldsymbol{Y}_i$,

$$t_y = \sum_U \boldsymbol{Y}_i.$$

When there is no nonresponse, this quantity will be estimated by a sample-based estimator of the form

$$\hat{t}_y = \sum_s w_i \boldsymbol{Y}_i = \sum_U w_i \boldsymbol{Y}_i I_i \qquad (1)$$

where the w_i, $i \in s$, are the sampling weights and I_i is an indicator for whether the i^{th} unit is in the sample or not. In this article, we will assume that the sampling weights are the inverse of the inclusion probabilities, or $w_i = \pi_i^{-1}$, with $\pi_i = \Pr(i \in s)$, so that the estimator (1) is the classical Horvitz-Thompson estimator (Horvitz and Thompson 1952). Also, let $\boldsymbol{I} = (I_1, I_2, ..., I_N)^T$ represent the vector of inclusion indicators for the population.

In the context of nonresponse, it is convenient to assume that each unit in the population is either a *respondent* or a *nonrespondent* for the variable of interest $\boldsymbol{Y}$. Consider the vector $\boldsymbol{R} = (R_1, R_2, ..., R_N)^T$, where R_i indicates if the i^{th} unit is a respondent or not. The distribution of $\boldsymbol{R}$ is called the *response mechanism*. In analogy to the definition of the sample s, we use $r \subseteq U$ to denote the (realized) set of respondents in the population, *i.e.*, those elements for which $R_i = 1$. Since the distribution of r and $\boldsymbol{R}$ is typically unknown and can in principle depend on the realized value of $\boldsymbol{I}$ as well as on the $\boldsymbol{Y}$, we need to assume a model for the response mechanism. When this assumed model is used to develop an estimator for a population quantity, the properties of this estimator become dependent on the response model. Hence, a misspecified model for $\boldsymbol{R}$ has the potential to cause significant and difficult to measure bias in both the estimator and its associated measures of precision. To avoid this problem, we will keep the response mechanism quite general in this article. Specifically, we will assume that the R_i are independent Bernoulli variables with

$$\Pr\{R_i = 1 \mid \boldsymbol{I}, \boldsymbol{Y}\} = \varphi_i, \; 0 < \varphi_i \leq 1, \forall i \in U,$$

and that the φ_i can be written as $\varphi_i = \varphi(\boldsymbol{X}_i)$, with $\varphi(\cdot)$ a continuous and differentiable but otherwise unspecified function of the $\boldsymbol{X}_i$. Note that this includes the uniform response mechanism, where $\varphi_i \equiv \varphi$ for all $i \in U$, as a special case.

When some of the selected elements do not respond, the estimator (1) can no longer be computed, and an estimator that includes a nonresponse adjustment is required. In this article, we are using the weighting cell estimator for this purpose. For simplicity, we will describe the situation in which both the $\boldsymbol{Y}_i$ and $\boldsymbol{X}_i$ are univariate variables, but the approach can be generalized to the multi-dimensional case. Let $s_r = s \cap r$ represent the subset of the selected elements that actually respond to the survey.

Let $U_g, g = 1, ..., G$, represent G groups obtained by dividing the population into groups based on the values of the known auxiliary variable X. Specific implementations might generate groups of equal size, or divide the range of

X into equal-length intervals. We shall leave the implementation unspecified for now, and state some general assumptions about G and the size of the groups in the next section. Note that we are considering the groups as fixed with respect to the sampling design and the response mechanism, which excludes the situation in which groups are formed based on the *observed* sample values $\{X_i : i \in s\}$. This was done primarily to simplify the theoretical derivations, and is similar to the approach of Särndal *et al.* (1992) and Kim and Fuller (1999), among others.

Let $s_g = s \cap U_g$ be the portion of the sample that falls in group g, and define similarly $s_{r,g} = s_r \cap U_g$. The weighting cell estimator is defined as

$$\hat{t}_{\text{WC}} = \sum_{g=1}^{G} \left(\frac{\sum_{s_g} w_i}{\sum_{s_{r,g}} w_i} \right) \sum_{i \in s_{r,g}} w_i Y_i. \qquad (2)$$

From this expression, is it easy to see that in each group, the estimator of the group total is ratio-adjusted by the inverse of the weighted proportion of respondents in the cell. This estimator is also the FEFI estimator of Kim and Fuller (1999). The properties of this estimator will be studied in next section.

3. PROPERTIES UNDER QUASI-RANDOMIZATION

3.1 Asymptotic Framework and Assumptions

The quasi-randomization properties of the weighting cell estimator will be studied in the usual finite population asymptotic context, in which the population U is treated as an element in an increasing sequence $U_1, U_2, ..., U_\nu$ with $\nu \to \infty$, with a corresponding sequence of sampling designs $p_\nu(\cdot)$ (see Isaki and Fuller (1982) for an early example of this framework). Let N_ν be the size of the ν^{th} population with $N_\nu > N_{\nu-1}$, let $\boldsymbol{Y}_\nu = (Y_1, Y_2, ..., Y_{N_\nu})^T$ denote the set of values of the characteristic of interest, Y, associated with U_ν, and similarly, $\boldsymbol{X}_\nu = (X_1, X_2, ..., X_{N_\nu})^T$. We assume that $\boldsymbol{X}_\nu$ is known. For each ν, a sample of size $n_\nu (n_\nu \geq n_{\nu-1})$ is selected from U_ν, according to a sampling design $p_\nu(\cdot)$. As before, let $\boldsymbol{I}_\nu = (I_1, I_2, ..., I_{N_\nu})^T$ be the corresponding sample inclusion vector. We will denote the K^{th} order central moment of the sample membership indicators $I_{i_1}, ..., I_{i_K}$ by

$$\Delta_{i_1,...,i_K} = E\left(\prod_{k=1}^{K} (I_{i_k} - \pi_{i_k}) \right). \qquad (3)$$

It is assumed that U_ν can be divided into $G_\nu (G_\nu \geq G_{\nu-1})$ mutually exclusive and exhaustive groups, U_g, $g = 1, ..., G_\nu$. These groups are constructed by sorting the population according to their X values and dividing the population into G_ν groups. We will assume that there are at least G_ν distinct values among the elements of $\boldsymbol{X}_\nu$. Let N_g represent the number of elements in U_g.

As mentioned in the previous section, we are treating the groups as fixed with respect to the population. The problem created by this approach is that in general, there is a non-zero chance of obtaining a group without any respondents. We solve this problem by adding a small constant in the denominators in each of the groups, or

$$\hat{t}^{*}_{\text{WC}} = \sum_{g=1}^{G_\nu} \left(\frac{\sum_{s_g} w_i}{\max\left(\sum_{s_{r,g}} w_i, N_g G_\nu n_\nu^{-1} \right)} \right) \sum_{i \in s_{r,g}} w_i Y_i. \qquad (4)$$

Hence, the difference between $\hat{t}^{*}_{\text{WC}}$ and $\hat{t}_{\text{WC}}$ in (2) is asymptotically negligible. This is similar to what is often done in practice to avoid overly large weights in ratio estimation.

Fuller and Kim (2003) give the limiting distribution of the FEFI estimator under the assumption that the response probabilities are constant within these cells. We will study the case where the response probabilities are a smooth function of an auxiliary variable and the number of cells are allowed to vary. Let $\mathbf{R}_\nu = (R_1, R_2, ..., R_{N_\nu})^T$ be the response indicator vector for the ν^{th} population. We assume that the distribution of $\mathbf{R}_\nu$ satisfies the *nonparametric response mechanism* assumptions, specified as follows:

(R1) $R_1, R_2, ..., R_{N_\nu}$ are independent random variables,

(R2) $\Pr\{R_i = 1 \mid \mathbf{I}_\nu, \boldsymbol{Y}_\nu\} = \varphi_i, \forall i \in U_\nu$,

(R3) $\varphi_i = \varphi(X_i) \forall i \in U_\nu$, where $\varphi(\cdot)$ is differentiable with bounded first derivative, and the $X_i \in [x_m, x_M]$, with x_m, x_M fixed constants and $x_m < x_M$.

The remaining assumptions are technical conditions that will be used extensively in the proofs. We assume that there are positive constants $\lambda_1, \lambda_2, ..., \lambda_9$ such that:

(A1) $\lambda_1 < N_\nu n_\nu^{-1} \pi_i < \lambda_2 < \infty$, $\forall i \in U_\nu$, and $n_\nu N_\nu^{-1} \to \pi \in (0, 1)$, as $\nu \to \infty$;

(A2) For distinct $i_1, ..., i_K \in U_\nu$, $K = 2, 3, ..., 8$,

$$|\Delta_{i_1,...,i_K}| \leq \begin{cases} \left(\prod_{k=1}^{K} (N-k+1) \right)^{-1} n_\nu^{K/2} \lambda_3, & \text{if K is even} \\ \left(\prod_{k=1}^{K} (N-k+1) \right)^{-1} n_\nu^{(K-1)/2} \lambda_4, & \text{if K is odd} \end{cases}$$

(A3) $\lim_{\nu \to \infty} \frac{1}{N_g} \sum_{i \in U_g} \varphi_i = \varphi_g^*, \forall g = 1, 2, ..., G_\nu$ and $\nu \geq 1$;

(A4) $\max_{i \in U_\nu} |Y_i| \leq \lambda_5$;

(A5) $\lambda_6 < \min_{i \in U_\nu} \varphi_i \leq 1$;

(A6) $\lambda_7 G_\nu^{-1} \le N_g N_\nu^{-1} \le \lambda_8 G_\nu^{-1}, \forall\, g = 1, 2, ..., G_\nu$;

(A7) $1 \le G_\nu \le n_\nu^\gamma \lambda_9$, with $0 \le \gamma \le 1/2$.

Assumptions (A1) – (A2) imply that, asymptotically, the sampling design is "well behaved," in the sense that the moments of the sample membership indicators are of the same order of magnitude as those in simple random sampling without replacement. This is a common assumption in finite population asymptotic theory. (A1) also requires that the sampling fraction converges to a constant in the interval (0, 1). The boundedness assumption (A4) on the observations will significantly simplify the proofs for some of the theorems in the article, and could be relaxed to the existence of bounded moments if desired. Similarly, some technical regularity conditions are required to avoid degenerate response mechanisms: (A3) provides that the limit for the average response probability in a cell exists, and (A5) excludes the situation in which some units might have $\varphi_i = 0$. Finally, assumptions (A6) and (A7) on the weighting cells require that all the cells grow at a similar rate, and that the total number of cells does not increase "too fast" relative to the sample size.

3.2 Main Results

The approach we will use in the study of the properties of the weighting cell estimator follows that commonly used in the study of finite population estimators. First, we show the asymptotic equivalence between the non-linear weighting cell estimator and a "linearized" approximation. Next, we derive the mean squared error properties of the linearized estimator and consider those as the asymptotic properties of the weighting cell estimator or, more precisely, the properties of the asymptotic distribution of the weighting cell estimator. See, for instance, Särndal *et al.* (1992, Chapter 5) for a description of this approach.

The following theorem formally states our first results. The proof is in the appendix.

Theorem 3.1. *Consider the sequence of populations* $\{U_\nu : \nu \ge 1\}$. *Assume that for each* ν, *a probabilistic sample of fixed size* $n_\nu (n_\nu \ge n_{\nu-1})$ *is selected from* U_ν *according to sampling design* $p_\nu(\cdot)$, *and that the response mechanism satisfies the conditions (R1) – (R2). Finally, assume that (A1) – (A7) hold. Then, the estimator* $\hat{t}_{WC}^*$ *is asymptotically equivalent to a linearized random variable* $\tilde{t}_{WC}$, *in the sense that*

$$\frac{1}{N_\nu}(\hat{t}_{WC}^* - \tilde{t}_{WC}) = O_p(G_\nu n_\nu^{-1}). \qquad (5)$$

The bias and variance of $\tilde{t}_{WC}/N_\nu$ *are given by*

$$E\left(\frac{\tilde{t}_{WC}}{N_\nu}\right) - \bar{Y}_\nu = \frac{1}{N_\nu}\sum_{g=1}^{G_\nu}\sum_{U_g}\left(\frac{\varphi_i - \bar{\varphi}_g}{\bar{\varphi}_g}\right)(Y_i - \tilde{Y}_g) \qquad (6)$$

and

$$\mathrm{Var}\left(\frac{\tilde{t}_{WC}}{N_\nu}\right) = \frac{1}{N_\nu^2}\sum_{g=1}^{G_\nu}\sum_{g'=1}^{G_\nu}\left[\sum_{U_g}\sum_{U_{g'}}\Delta_{ij}\tilde{Y}_{ig}\tilde{Y}_{jg'}\right] + \frac{1}{N_b^2}\sum_{g=1}^{G_\nu}\sum_{U_g}\pi_i^{-2}\frac{\varphi_i(1-\varphi_i)}{\bar{\varphi}_g^2}(Y_i - \tilde{Y}_g)^2, \qquad (7)$$

where

$$\bar{\varphi}_g = \frac{1}{N_g}\sum_{U_g}\varphi_i,\ \bar{Y}_g = \frac{1}{N_g}\sum_{U_g}Y_i,\ \tilde{Y}_g = \frac{\sum_{U_g}\varphi_i Y_i}{\sum_{U_g}\varphi_i}$$

and

$$\tilde{Y}_{ig} = \frac{\varphi_i(Y_i - \tilde{Y}_g) + \bar{\varphi}_g\tilde{Y}_g}{\pi_i\bar{\varphi}_g},\ \forall\, i \in U_g \text{ and } \forall\, g = 1, 2, ..., G_\nu.$$

Remark 1. The asymptotic equivalence between $\hat{t}_{WC}^*$ and $\tilde{t}_{WC}$ depends on the number of groups G_ν, with a faster convergence rate achieved when G_ν grows more slowly. The intuition behind this result is that the goodness of the linear approximation depends on how well the true cell ratio response adjustments φ_g^* are estimated by the sample-based estimators $\sum_{s_{r,g}} w_i / \sum_{s_g} w_i$. Since the cell ratios will be better estimators as the sample size grows larger, this would argue that G_ν should be chosen to be small, which corresponds to the current practice in applications of weighting cell estimation. However, as will be shown below, the MSE properties of $\tilde{t}_{WC}$ under the nonparametric response mechanism improve as G_ν gets larger. A more detailed discussion of the selection of the number of groups will be provided after Theorem 3.2 below and in section 4.

Remark 2. The results in Theorem 3.1 depend on the population groups U_g, $g = 1, ..., G_\nu$ and on the φ_i, $i \in U_\nu$, but do not rely on the fact that the response probabilities are a smooth function of the auxiliary variable X. Hence, the explicit expressions for the asymptotic bias and variance can be used to derive results for other response mechanisms that follow (R1) – (R2). In particular, results for the *response homogeneity group* model (see Särndal *et al.* 1992, page 577) follow directly from Theorem 3.1. This is also the model studied by Fuller and Kim (2003). Under that model, one assumes that $\varphi_i \equiv \varphi_g$ for all

$i \in U_g$, $g = 1, ..., G$, and it can easily be shown that the bias of $\tilde{t}_{WC}$ is 0 and its variance is

$$\mathrm{Var}\left(\frac{\tilde{t}_{WC}}{N_\nu}\right) = \mathrm{Var}\left(\frac{\hat{t}_y}{N_\nu}\right) + \frac{1}{N_\nu^2}\sum_{g=1}^{G_\nu}\frac{1-\varphi_g}{\varphi_g}\sum_{U_g}\pi_i^{-2}(Y_i - \bar{Y}_g)^2.$$

The first term in the variance is the variance of the estimator without nonresponse, and the second term represents the variance inflation caused by the nonresponse under a homogeneous within-cell response mechanism.

The following corollary follows directly from Theorem 3.1 and Fuller (1996, Theorem 5.2.1). A proof is given in the appendix.

Corollary 3.1. *Under the conditions of Theorem 3.1 with* $\gamma < 1/2$ *in (A7), for any sampling design* $p_\nu(\cdot)$ *such that*

$$n_\nu^{1/2}\left(\frac{\tilde{t}_{WC}}{N_\nu} - \bar{Y}_\nu - B_\nu\right) \xrightarrow{L} N(0, V),$$

where B_ν *corresponding to the bias of* $\tilde{t}_{WC}/N_\nu$ *given in Theorem 3.1 and*

$$V \equiv \lim_{\nu\to\infty} n_\nu \mathrm{Var}(\tilde{t}_{WC}/N_\nu) \in (0, \infty),$$

then

$$\left[\mathrm{Var}\left(\frac{\tilde{t}_{WC}}{N_\nu}\right)\right]^{-1/2}\left(\frac{\hat{t}^*_{WC}}{N_\nu} - \bar{Y}_\nu - B_\nu\right) \xrightarrow{L} N(0, 1).$$

Corollary 3.1 states that, whenever the linearized estimator $\tilde{t}_{WC}$ achieves asymptotic normality, then so does $\hat{t}^*_{WC}$. Since $\tilde{t}_{WC}$ can be written as a classical expansion estimator of the form (1), this result is quite general.

Under the nonparametric response mechanism described in (R1) – (R3), it is possible to describe the effect of the number of groups G_ν on the asymptotic bias and variance of $\hat{t}^*_{WC}$. The next theorem gives the asymptotic rates for the bias and variance, and is proven in the appendix.

Theorem 3.2. *Assume that (R3) and the conditions of Theorem 3.1. Then,*

$$\mathrm{E}\left(\frac{\tilde{t}_{WC}}{N_\nu}\right) - \bar{Y}_\nu = O\left(\frac{1}{G_\nu}\right)$$

and

$$\mathrm{Var}\left(\frac{\tilde{t}_{WC}}{N_\nu}\right) = O\left(\frac{1}{n_\nu}\right) + O\left(\frac{1}{n_\nu G_\nu}\right).$$

Remark 3. Theorem 3.2 shows that both the asymptotic bias and variance of the weighting cell estimator $\hat{t}^*_{WC}$ become smaller as the number of groups G_ν increases. An intuitive explanation of that fact is that the approximation of the function $\varphi_i = \varphi(X_i)$ by the step function $\varphi_i = \varphi^*_g$ improves as the number of cells increases. The asymptotic variance has a term that is independent of G_ν. This "residual variance" is due to the inherent variability of the sampling design and the response mechanism, and cannot be reduced by changing G_ν.

Remark 4. As noted in Remark 1, constructing a good linear approximation $\tilde{t}_{WC}$ requires G_ν to be small, while Theorem 3.2 states that the MSE of $\tilde{t}_{WC}$ is minimized by taking G_ν as large as possible. Taken together, this can be interpreted to mean that, once the sample size in every cell is sufficiently large to obtain a "valid" ratio estimator for the average cell response probability φ^*_g, it is preferable to increase the number of cells than to increase the sample size per cell. The simulation experiments discussed in section 4 will further explore this recommendation.

The following corollary follows directly from Corollary 3.1, Theorem 3.2, and Chebyshev's inequality, and establishes the consistency of the weighting cell estimator under the nonparametric response mechanism.

Corollary 3.2. *Under the conditions of Theorem 3.2,* $\hat{t}^*_{WC}$ *is a consistent estimator for* t_y, *in the sense that for any* $\epsilon > 0$,

$$\Pr\left(\left|\frac{\tilde{t}^*_{WC} - t_y}{N_\nu}\right| > \epsilon\right) \to 0, \quad \nu \to \infty.$$

Remark 5. As Corollary 3.2 shows, as long as a variable X can be found that is sufficiently related to the nonresponse, in the sense of assumptions (R1) – (R3), construction of weighting cells does not require knowledge of homogeneous response probability cells in order to construct a consistent estimator. However, as discussed in Remarks 1 and 4, the choice of the number of cells still has an effect on the properties of the estimator.

Remark 6. Assumption (R3) can easily be relaxed to allow for a small number of points of discontinuity in both $\varphi(\cdot)$ and its first derivative. A "small" number can mean that the number is either fixed as $\nu \to \infty$ or increases at a rate slower than G_ν. This would make it possible to account for situations such as stratified designs or the presence of domains within U_ν. The present theory can be extended

directly to these situations, if the values for the variable X fall in non-overlapping segments for the different strata or domains.

4. SIMULATION EXPERIMENTS

4.1 Description of the Experiment

In order to investigate the practical implications of the results of section 3, we carried out a Monte Carlo experiment on a fixed population of $N = 3{,}000$ units. We consider the case of one covariate, X, whose population values are generated as:

$$X_1, X_2, ..., X_N \sim \text{i.i.d. } U(0, 1),$$

and two different variables of interest, Y_1 and Y_2. We are interested in evaluating the effects of (1) the (model) relationship between Y and X, (2) the response mechanism $\varphi(X)$, (3) the sample size n and (4) the number of cells G, on the bias and on the mean square error of the $\hat{t}_{WC}$ estimator. Since our theoretical results rely on the approximation of $\hat{t}_{WC}$ (or $\hat{t}^*_{WC}$) by a linearized estimator $\tilde{t}_{WC}$, we will also compare the behavior of $\hat{t}_{WC}/N_v$ and $\tilde{t}_{WC}/N_v$ as estimators of the population mean, $\bar{Y}_v = N_v^{-1}\sum_U Y_i$. Finally, we compare $\hat{t}_{WC}/N_v$ to the "naive" estimator of the mean, which is defined for the variable Y as:

$$\bar{y}_r = \frac{\sum_{i\in s_r} w_i Y_i}{\sum_{i\in s_r} w_i},$$

corresponding to a ratio adjustment of the respondent sample to the original sample. This estimator is appropriate under the assumption of uniform response mechanism or, to use the terminology of Little and Rubin (2002, chapter 1), when observations are *missing completely at random* (MCAR). Note that $\bar{y}_r$ is equivalent to the weighting cell estimator with a single cell.

The levels of the four factors used in the experiment are given in Table 1. The "levels" of the variable Y correspond to two populations of independent values. The variable Y_1 was generated as $N(40, 58)$, truncated to -3 to +3 standard deviations, corresponding to the "white noise" case. The variable Y_2 is related to X and was generated through the linear model $Y_2 = 27.12 + 26.06X + \varepsilon$, where $\varepsilon \sim N(0, 9)$. The population mean and variance for the two variables were, respectively, (39.9, 55.3) for Y_1, and (40.0, 63.9) for Y_2.

The four levels of the response mechanisms contain two different scenarios regarding the response probabilities: constant ($C1$, $C2$), and linearly related to X ($L1$, $L2$). The response probabilities are:

- $\varphi_{C1}(X) = 0.5$
- $\varphi_{C2}(X) = 0.8$
- $\varphi_{L1}(X) = 0.20 + 0.60X$
- $\varphi_{L2}(X) = 0.65 + 0.30X$

The levels of the linear response mechanisms were chosen so that the average probabilities (over X) were approximately equal to 0.5 and 0.8, respectively.

Table 1
Overview of Factors in the Simulation Experiment

Factor	Levels
Y variable	Y_1, Y_2
Response mechanism $\varphi(\cdot)$	$C1, C2, L1, L2$
Sample size n	200, 500
Number of cells G	2, 3, 5, 8

For a given G, the groups were created by dividing the range of X into G equal segments and assigning the element i to the group g if the value X_i was in the g^{th} segment, $i = 1, 2, ..., N$ and $g = 1, 2, ..., G$. The simulations were carried out through a completely randomized factorial experiment $2\times4\times2\times4$. For each combination of the levels of the factors in Table 1, $B = 5{,}000$ independent realizations of the vector indicator of responses, $\boldsymbol{R} = (R_1, R_2, ..., R_N)^T$, were generated according to the corresponding response mechanism. For each one of such realizations, a simple random sample (without replacement and of size n), s, was selected from the overall population. Within each selected sample, the respondents were the values of $i \in s$ such that $R_i = 1$.

This procedure could in principle lead to a group not containing any sampled and responding element, in which case the weighting cell estimator (ignoring the adjustment in (4)) cannot be computed. If that happened, the realization was discarded and a new sample drawn from the population. Out of the 5,000 repetitions for each combination of factors, this happened 13 times in the factor combination $(Y_1, \varphi_{L1}, 200, 8)$ and 15 times with $(Y_2, \varphi_{L1}, 200, 8)$. It did not occur with any of the other factor combinations. Hence, the number of samples discarded was very small and this has a negligible effect on the simulation results.

With $n = 200$ and $G = 8$, we expect approximately 25 sampled elements in each cell, to be further reduced by the nonresponse. Since the estimator relies on ratio estimation in each cell, we judged this to be a reasonable lower bound on the number of observations per cell to consider in the simulations. In practice, a number of procedures could be used when groups have too few elements, such as picking

a smaller value for G or collapsing neighboring groups. We also implemented an estimator that collapses the empty cell with a neighboring cell as well as a version with a lower bound on the value of the denominator in the weighting adjustment (*i.e.*, $\hat{t}^{*}_{WC}$), and the results are virtually indistinguishable from those reported below, so they will not be further discussed here.

4.2 Results

Tables 2 and 3 show the simulated bias of the weighting cell estimator for the variables Y_1 and Y_2 as a fraction of the standard deviation. As a comparison, the last column of Tables 2 and 3 displays the bias of the naive estimator, $\bar{y}_r$. The bias as a fraction of the standard deviation, referred to here as the *relative bias*,

$$RB(\hat{t}_{WC}, \hat{t}_y) = \frac{E(\hat{t}_{WC} - \bar{Y})}{(Var(\hat{t}_{WC}))^{1/2}}$$

was also used in Cochran (1977, page 14), where it is shown that as the relative bias increases, inferential results rapidly become unreliable. In a simple simulation example, Cochran (1977) shows that a relative bias of ±0.50 or more leads to highly inaccurate 95% confidence intervals.

For Y_1 (Table 2), the relative bias of the weighting cell estimator is small and is similar to the relative bias of the naive estimator, for all sample sizes, response mechanisms and cells sizes considered. For the variable Y_2 (Table 3), similar results hold when the response mechanism is uniform (C1, C2). However, when the response probabilities are a linear function of X (L1, L2), the naive estimator becomes severely biased. This relative bias decreases as the number of cells increases, and three to five cells appear sufficient to remove most of the bias. This finding agrees with that of Cochran (1968) in the context of bias reduction for observational studies.

Table 2
Relative Bias of the Weighting Cell and Naive Estimators for the Mean Y_1

Sample size	Response mechanism	Number of Cells 2	3	5	8	Naive estimator
200	C1	-0.00	-0.01	0.01	0.01	-0.00
	C2	0.01	-0.00	-0.01	0.00	0.00
	L1	-0.02	0.03	-0.04	-0.01	-0.00
	L2	-0.00	-0.02	0.00	-0.02	-0.00
500	C1	-0.00	-0.01	0.04	-0.01	0.00
	C2	0.01	0.02	-0.01	-0.01	0.00
	L1	0.05	0.02	-0.01	-0.02	0.01
	L2	0.01	0.01	-0.00	-0.01	0.01

Table 3
Relative Bias of the Weighting Cell and Naive Estimators for the Mean of Y_2

Sample size	Response mechanism	Number of Cells 2	3	5	8	Naive estimator
200	C1	0.01	-0.01	-0.02	0.02	-0.01
	C2	-0.03	-0.00	0.02	0.01	-0.00
	L1	1.16	0.59	0.22	0.07	3.57
	L2	0.36	0.18	0.06	0.03	1.36
500	C1	0.01	0.01	-0.02	-0.00	0.00
	C2	0.02	-0.00	-0.00	-0.01	-0.01
	L1	1.98	0.96	0.32	0.15	5.84
	L2	0.61	0.29	0.09	0.02	2.26

Hence, when the variable of interest is totally unrelated to the response mechanism, as in the cases of Y_1 under all mechanisms considered and of Y_2 under the uniform response mechanism, the bias does not depend on the number of cells. When the variable of interest and the response mechanism are related, multiple cells are required to remove the bias.

The relative mean squared error (RMSE) for the two variables of interest, defined as the MSE of the weighting cell estimator divided by the MSE of the estimator with no non-response,

$$RMSE(\hat{t}_{WC}, \hat{t}_y) = \frac{E(\hat{t}_{WC} - t_y)^2}{E(\hat{t}_y - t_y)^2},$$

are in Tables 4 and 5. In these tables, the last column again corresponds to the relative MSE of the naive estimator. Note that with the exception of the two L1 cases for variable Y_2, the Tables 4 and 5 are really variance tables, since the bias is so small.

For Y_1 (Table 4), the variable uncorrelated with X, the number of cells has relatively little effect on the relative mean square error, with results around 2.3 for a 50% response rate, and around 1.3 for the 80% rate. However, a relatively modest increase in MSE is observed, especially for the high nonresponse cases (C1, L1). For Y_2 (Table 5), the variable correlated with X, increasing the number of cells improves the results for all response mechanisms, but the effect is much more pronounced when the response mechanism is also correlated with the variable of interest. As for the relative bias, three to five cells achieve most of the efficiency gain, while the naive estimator is extremely inefficient.

Table 4
Relative Mean Squared Error of the Weighting Cell Estimator Compared to the Estimator Without Nonresponse for Y_1

Sample size	Response mechanism	Number of Cells 2	3	5	8	Naive estimator
200	C1	2.02	2.13	2.11	2.21	2.08
	C2	1.25	1.31	1.29	1.28	1.28
	L1	2.34	2.32	2.61	2.70	2.08
	L2	1.30	1.29	1.29	1.31	1.28
500	C1	2.25	2.21	2.19	2.31	2.23
	C2	1.30	1.32	1.34	1.29	1.30
	L1	2.55	2.57	2.62	2.70	2.22
	L2	1.32	1.35	1.33	1.34	1.31

Table 5
Relative Mean Squared Error of the Weighting Cell Estimator Relative to the Estimator Without Nonresponse for Y_2

Sample size	Response mechanism	Number of Cells 2	3	5	8	Naive estimator
200	C1	1.33	1.17	1.10	1.07	2.07
	C2	1.09	1.05	1.02	1.02	1.26
	L1	3.14	1.57	1.16	1.12	26.32
	L2	1.23	1.07	1.03	1.01	3.57
500	C1	1.35	1.19	1.10	1.09	2.22
	C2	1.09	1.05	1.03	1.03	1.30
	L1	6.60	2.30	1.23	1.13	69.75
	L2	1.50	1.14	1.04	1.02	7.83

The difference between the results for both variables is surprising at first, but it can be explained using the results from section 3. Clearly, the results for Y_2 follow the asymptotic theory, in that the MSE improves as the number of cells improves (as long as sufficient observations are available in each cell). In the case of Y_1, note first that the bias is negligible relative to the standard deviation for all values of G (see Table 2), so that the change in MSE is due almost exclusively to differences in variance. It turns out that when a variable is iid in the population and sampling is equal-probability, the asymptotic variance in Theorem 3.1 is relatively insensitive to the number of cells. In that case, the increase in MSE is influenced by the variability implied in the linear approximation in Theorem 3.1, which increases with the number of cells.

The theory described in this article applies to response functions that can have arbitrary smooth shape. In order to evaluate results for more complicated functions, we also created a variable $Y_3 = 25 + 95X - 95X^2 + \varepsilon$, where $\varepsilon \sim N(0, 3)$, so that the Y_3 has mean 40.9 and variance 51.8, and two additional quadratic response mechanisms

- $\varphi_{Q1}(X) = 0.17 + 1.96X - 1.96X^2$
- $\varphi_{Q2}(X) = 0.50 + 1.80X - 1.80X^2$.

The results (not shown) broadly reflect the findings for the previous variables. When the response mechanism and the variables are correlated (the linear variable is correlated with the linear response mechanism, and the quadratic variable is correlated with the linear and quadratic response mechanisms), significant bias occurs but can be removed by increasing the number of cells. In the case of the quadratic response mechanism and the quadratic variable, eight or more cells appear to be required to remove the bias. Similarly, the relative efficiency improves for all response mechanisms for both the linear (Y_2) and quadratic variable, with the most dramatic results found for the linear variable/linear response and quadratic variable/quadratic response cases.

In the previous sections of this article, we approximated the weighting cell estimator by a "linearized" estimator $\tilde{t}_{WC}$, and then derived the asymptotic properties of that estimator. It is therefore of interest to compare the statistical properties of both estimators in simulated settings. For all the scenarios in Table 1, we calculated the relative efficiencies of the weighting cell estimator compared to the linearized estimator. These relative efficiencies were all close to 1.00, with the largest deviation being a value of 1.08. Hence, the statistical properties of weighting cell estimator appear to be well approximated by those of the linearized estimator.

5. CONCLUSIONS

We have shown that the weighting cell estimator, corresponding also to the FEFI estimator proposed by Kim and Fuller (1999), is consistent with respect to the sampling design and a nonparametric response model. That model does not require the correct specification of homogeneous response probability cells, as long as a variable related to the response probability can be identified.

The statistical properties of the estimator depend on the number of cells used in the estimation, but the relationship is rather complex. Asymptotically, there appears to be a trade-off between the goodness of the approximation of the weighting cell estimator by a linearized estimator, which requires a small number of cells, and the mean squared error of that linearized estimator, which is reduced when a large number of cells are used. While useful in understanding the asymptotic behavior of the estimator, these

findings only provide limited guidance for choosing the number of cells for a particular survey. However, these findings show that reliable inference for weighting cell estimators will require cells with reasonable sample sizes, because variance estimates typically rely on the variance of the linearized estimator as an approximation of the variance of the weighting cell estimator.

The simulation experiments show that when the variable of interest and the response mechanism are uncorrelated, the number of cells has virtually no effect on the design bias of the estimator. When the variable of interest and the response mechanism are uncorrelated, even the estimator with a single weighting cell (corresponding to a simple ratio adjustment) is essentially unbiased, while models with multiple cells perform equally well. When the response mechanism and the variable of interest are related, however, the bias properties of the weighting cell estimator depend critically on the number of cells. In particular, estimators with a single cell are severely biased, but even a relatively small number of cells is sufficient to reduce both the bias and variance of the estimator. This result holds for both linear and nonlinear relationships between the response mechanism and the variable of interest.

The design efficiency of estimators depends on the relationship between the variable of interest and the variable(s) used to form weighting cells. When those two variables are uncorrelated, the number of cells has no effect on the efficiency of the estimator. Conversely, when those two variables are correlated, increasing the number of cells improves the design efficiency of the estimator. Even a small number of cells dramatically improves the performance of the estimator.

Overall, it appears that in the presence of nonresponse, forming at least a small number of weighting cells based on a variable related to the non-response provides a good "insurance policy" against design bias and design inefficiency. This article has shown that this adjustment does not require the assumption that the cells be based on a priori knowledge of constant nonresponse groups. The resulting weighting cell estimator will never perform worse than the naive estimator with a single ratio adjustment for the whole sample, and it might perform significantly better.

6. ACKNOWLEDGEMENTS

The authors thank Wayne Fuller for many helpful comments made during the development of this manuscript. We also are grateful for the comments of the associate editor and the two referees. This research was supported by a subcontract between Westat and Iowa State University under Contract No. ED–99–CO–0109 between Westat and the U.S. Department of Education. The first author gratefully acknowledges the support of CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico), Brazil, during his Ph. D. studies at Iowa State University.

APPENDIX

Derivations of Theoretical Results

Lemma 1. *Assume that the conditions (A1) – (A3) and (R1) – (R2) hold. For* $i_1, i_2, ..., i_k \in U_v$, *define*

$$\Gamma_{i_1,...,i_k} = \mathrm{E}\left(\prod_{l=1}^{k} (I_{i_l} R_{i_l} - \pi_{i_l} \varphi_{i_l}) \right),$$

where $\varphi_i = \varphi(X_i)$. *Consider the* $\Delta_{i_1,...,i_k}$ *of (3). Let* A^r *denotes the r-fold Cartesian product of the set A, where r is a fixed positive integer,* $A_{1,r,v} = \{(i_1, i_2, ..., i_r) \in U_v^r : i_1 = i_2 = ... = i_r\}$ *and* $A_{k,r,v} = \{(i_1, i_2, ..., i_r) \in U_v^r$: *exactly k components are distinct*$\}$, $k = 2, 3, ..., r$. *Then, for* $r = 8$,

$$N_v^8 n_v^{-8} \max_{i_1,...,i_8 \in A_{k,8,v}} (|\Gamma_{i_1,...,i_8}|, |\Delta_{i_1,...,i_8}|) = \begin{cases} O(N_v^3 n_v^{-4}), & \text{if } k=5 \\ O(N_v^3 n_v^{-5}), & \text{if } k=6 \\ O(N_v n_v^{-4}), & \text{if } k=7 \\ O(n_v^{-4}), & \text{if } k=8. \end{cases}$$

Proof of Lemma 1. See Da Silva (2003).

Lemma 2. *Suppose the conditions of Theorem 3.1 hold. Consider the vectors* $\hat{t}_{gv} = (\hat{t}_{1,g}, \hat{t}_{2,g}, \hat{t}_{3,g})' = \sum_{U_g} \pi_i^{-1} (1, Y_i R_i, R_i)' I_i$ *and* $\hat{t}_{gv}^* \equiv (\hat{t}_{1,g}, \hat{t}_{2,g}, \hat{t}_{3,g}^*)'$, *with* $\hat{t}_{3,g}^* = \max\{\hat{t}_{3,g}, N_g G_v / n_v\}$. *Let* $t_{gv} = E(\hat{t}_{gv})$. *Then for all* $g = 1, 2, ..., G_v$,

$$\frac{1}{N_g^8} (\mathrm{E} \|\hat{t}_{gv}^* - t_{gv}\|^8, \mathrm{E} \|\hat{t}_{gv} - t_{gv}\|^8) = O\left((G_v / n_v)^4\right).$$

Proof of Lemma 2: See Da Silva (2003).

Proof of Theorem 3.1: Consider the proof of (5). Let $\boldsymbol{a} = (a_1, a_2, a_3)' \in \mathbb{R}^3$ and $h: \mathbb{R}^3 \to \mathbb{R}$, where $h(\boldsymbol{a}) = a_1 a_2 / a_3, a_3 \neq 0$. Define

$$\eta_{gv}(\boldsymbol{a}) = h(N_g^{-1} t_{gv}) + \sum_{k=1}^{3} h^{(k)}(N_g^{-1} t_{gv})(a_k - N_g^{-1} t_{gv}),$$

where $h^{(k)}(\boldsymbol{a}) = \partial h(\boldsymbol{a}) / \partial a_k$, and let $e_{gv} = h(\boldsymbol{a}) - \eta_{gv}(\boldsymbol{a})$. Note that $\hat{t}_{\mathrm{WC}}^* = \sum_{g=1}^{G_v} N_g h(N_g^{-1} \hat{t}_{gv})$, and hence, defining the "linearized" estimator $\tilde{t}_{\mathrm{WC}} = \sum_{g=1}^{G_v} N_g \eta_{gv}(N_g^{-1} \hat{t}_{gv})$, we can write

$$\frac{1}{N_v} (\hat{t}_{\mathrm{WC}}^* - \tilde{t}_{\mathrm{WC}}) = \bar{e}_v + \bar{\eta}_v,$$

where

$$\bar{e}_v = \frac{1}{N_v}\sum_{g=1}^{G_v} N_g e_{gv}(N_g^{-1}\hat{t}^*_{gv})$$

and

$$\bar{\eta}_v = \frac{1}{N_v}\sum_{g=1}^{G_v} N_g\left(\eta_{gv}(N_g^{-1}\hat{t}^*_{gv}) - \eta_{gv}(N_g^{-1}\hat{t}_{gv})\right).$$

Consider first the term $\bar{\eta}_v$. Observe that

$$|\eta_{gv}(N_g^{-1}\hat{t}^*_{gv}) - \eta_{gv}(N_g^{-1}\hat{t}_{gv})| =$$

$$|h^{(3)}(N_g^{-1}t_{gv})|\frac{1}{N_g}|\hat{t}^*_{3,g} - \hat{t}_{3,g}|.$$

By (A4) and (A5), it is straightforward to check that $h(\cdot)$ and $h^{(k)}(\cdot)$, $k = 1,2,3$, are O(1) when evaluated at $N_g^{-1}t_{gv}$, for all $g = 1,2,...,G_v$. Since by construction, we have $1/N_g|\hat{t}^*_{3,g} - \hat{t}_{3,g}| \le G_v/n_v$, we conclude that $|\bar{\eta}_v| = O(G_v/n_v)$. Thus, to complete the proof of (5), it remains to show that $\bar{e}_v = O_p(G_v n_v^{-1})$. Let $f_{gv}(a) \equiv (e_{gv}(a))^2$. By the C_r the inequality (Sen and Singer 1993, page 21),

$$|f_{gv}(a)|^2 \le 5^3\left(|h(a)|^4 + |h(N_g^{-1}t_{gv})|^4 + \sum_{k=1}^{3}|h^{(k)}(N_g^{-1}t_{gv})|^4|a_k - N_g^{-1}t_{gv}|^4\right).$$

Using (A1) and (A4), straightforward bounding arguments show that $|h(N_g^{-1}\hat{t}^*_{gv})|^4 = O((n_v/G_v)^4)$ and that $N_g^{-4}|\hat{t}_{k,g} - t_{k,g}|^4 = O(1)$ for $k = 1, 2, 3$. Therefore,

$$|f_{gv}(N_g^{-1}\hat{t}^*_{gv})|^2 = O\left(\frac{n_v^4}{G_v^4}\right).$$

Since by Lemma 2, $N_g^{-8}E\|\hat{t}^*_{gv} - t_{gv}\|^8 = O((n_v/G_v)^{-4})$, and $v|f_{gv}(N_g^{-1}\hat{t}^*_{gv})|^2$ is continuous at any realization of $N_g^{-1}\hat{t}^*_{gv}$, then the sequence $\{|f_{gv}(N_g^{-1}\hat{t}^*_{gv})|^2\}$ satisfies the conditions of Theorem 5.4.4 (with $\eta = 1$, $p = 4$) of Fuller (1996, page 247). Therefore,

$$E\left[|f_{gv}(N_g^{-1}\hat{t}^*_{gv})|^2\right] = O(1), \quad \forall\, g = 1, 2, ..., G_v.$$

Now, from the continuity of $f_{gv}(\cdot)$ and its derivatives up to order three, $\{f_{gv}(N_g^{-1}\hat{t}^*_{gv})\}$ satisfies the conditions of Theorem 5.4.3 (with $\delta = 1$, $s = 4$ and $a_v = O(\sqrt{G_v/n_v})$ of Fuller (1996, pages 244-245). Hence,

$$Ef_{gv}(N_g^{-1}\hat{t}^*_{gv}) = O(a_v^4) = O\left(\frac{G_v^2}{n_v^2}\right), \quad \forall\, g = 1, 2, ..., G_v,$$

because $f_{gv}(\cdot)$ and all of its derivatives up to order three are zero at $N_g^{-1}t_{gv}$. Therefore, we conclude that

$$E|\bar{e}_v| \le \frac{1}{N_v}\sum_{U_g} N_g E|e_{gv}(N_g^{-1}\hat{t}^*_{gv})| \le \frac{1}{N_v}\sum_{U_g} N_g\left(Ef_{gv}(N_g^{-1}\hat{t}^*_{gv})\right)^{1/2} = O\left(\frac{G_v}{n_v}\right),$$

which leads to $\bar{e}_v = O_p(G_v n_v^{-1})$ by an application of Markov's inequality.

Expressions (6) and (7) are obtained by direct computation of the moments of the linear estimator $\tilde{t}_{WC}$ under the sampling design and the response mechanism.

Proof of Corollary 3.1: Let

$$Z_v = \frac{1}{V_v^{1/2}}\left(\frac{\tilde{t}_{WC}}{N_v} - \bar{Y}_v - B_v\right)$$

and

$$W_v = \frac{1}{V_v^{1/2}}\left(\frac{\hat{t}^*_{WC}}{N_v} - \frac{\tilde{t}_{WC}}{N_v}\right),$$

where $V_v = \mathrm{Var}(\tilde{t}_{WC}/N_v)$. Hence,

$$\left[\mathrm{Var}\left(\frac{\tilde{t}_{WC}}{N_v}\right)\right]^{1/2}\left(\frac{\hat{t}^*_{WC}}{N_v} - \bar{Y}_v - B\right) = Z_v + W_v.$$

Since $V/n_v V_v \to 1$, as $v \to \infty$, then,

$$Z_v = \frac{1}{V^{1/2}}\left(\frac{V}{n_v V_v}\right)^{1/2} n_v^{1/2}\left(\frac{\tilde{t}_{WC}}{N_v} - \bar{Y}_v - B_v\right) \xrightarrow{L} \frac{1}{V^{1/2}}Z,$$

where $Z \sim N(0, V)$. Also, (A7) with $\gamma < 1/2$ implies that $n_v^{1/2}O_p(G_v n_v^{-1}) = o_p(1)$. Hence, by Theorem 3.1,

$$W_v = \frac{1}{V^{1/2}}\left(\frac{V}{n_v V_v}\right)^{1/2} n_v^{1/2}\left(\frac{\hat{t}^*_{WC}}{N_v} - \frac{\tilde{t}_{WC}}{N_v}\right) = o_p(1).$$

The result of the corollary follows, therefore, from Fuller (1996, Theorem 5.2.1).

Proof of Theorem 3.2: Fix a $g \in \{1, 2, ..., G_v\}$. The conditions of the theorem imply, by the Intermediate Value Theorem, that there exists X_{0g} inside the interval defined by the lowest and the highest values of $X_i \in U_g$ such that $\bar{\varphi}_g = N_g^{-1}\Sigma_{U_g}\varphi_i = \varphi(X_{0g})$. Also, by the mean Value Theorem, $\forall\, i \in U_g$,

$$\varphi_i = \varphi(X_i) = \varphi(X_{0g}) + \varphi'(c^*)(X_i - X_{0g}),$$

where c^* is between X_i and X_{0g}. So,

$$|\varphi_i - \bar{\varphi}_g| = |\varphi'(c^*)|\,|X_i - X_{0g}| \le C\frac{X_{(N)} - X_{(1)}}{G_v}, \qquad (8)$$

for some constant $C \in (0, \infty)$ and, by (A5) and (A6),

$$\left| \text{Bias}\left(\frac{\tilde{t}_{\text{WC}}}{N_v} \right) \right| \leq C \lambda_6^{-1} \lambda_5 \frac{X_{(N)} - X_{(1)}}{G_v}.$$

Observe now that since

$$|\tilde{Y}_{ig}| \leq \frac{1}{\pi_i} \frac{\varphi_i}{\bar{\varphi}_g} |Y_i| + \frac{1}{\pi_i} \frac{|\varphi_i - \bar{\varphi}_g|}{\bar{\varphi}_g} |\tilde{Y}_g|,$$

then, by (A1, (A6) and (8),

$$\tilde{Y}_{ig} = O\left(\frac{N_v}{n_v} \right) + O\left(\frac{N_v}{n_v G_v} \right), \ \forall U_g, \forall g = 1, 2, ..., G_v,$$

which implies that

$$\tilde{Y}_{ig} \tilde{Y}_{jg'} = O\left(\frac{N_v^2}{n_v^2} \right) + O\left(\frac{N_v^2}{n_v^2 G_v} \right), \ \forall U_g, \forall g = 1, 2, ..., G_v.$$

Using the facts that, by (A7), $N_g / N_v = O(1/G_v)$, by (A2) and (A3), $\Sigma_{U_g} \Sigma_{U_g} \Delta_{ij} = O(n_v / G_v)$ and, for $g \neq g'$, $\Sigma_{U_g} \Sigma_{U_{g'}} \Delta_{ij} = O(n_v / G_v^2)$, then, the first term of $\text{Var}(\tilde{t}_{\text{WC}} / N_v)$ is bounded by

$$O\left(\frac{1}{n_v} \right) + O\left(\frac{1}{n G_v} \right).$$

Since the second terms of $\text{Var}(\tilde{t}_{\text{WC}} / N_v)$ is bounded by $O(1/n_v)$, the conclusion follows.

REFERENCES

CASSEL, C.-M., SÄRNDAL, C.-E. and WRETMAN J.H. (1983). Some uses of statistical models in connection with the nonresponse problem. In *Incomplete data in sample surveys: Theory and bibliographies*, (Eds. W.G. Madow, I. Olkin and D. B. Rubin). Academic Press, New York: London. 3, 143–160.

COCHRAN, W.G. (1968). The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics*. 24, 295–313.

COCHRAN, W.G. (1977). *Sampling Techniques* (3rd Ed.). New York: John Wiley & Sons, Inc.

DA SILVA, D.N. (2003). Adjustments for Survey Unit Nonresponse Under Nonparametric Response Mechanisms. Ph. D. Thesis, Iowa State University, Ames, IA.

DA SILVA, D.N., and OPSOMER, J.D. (2003). A kernel smoothing method to adjust for unit nonresponse in sample surveys. In *Proceedings of the Section on Survey Research Methods*, American Statistical Association [CD-ROM]. Alexandria, VA. Article #00605.

FULLER, W.A. (1996). *Introduction to Statistical Time Series* (Second Edition). Wiley.

FULLER, W.A., and KIM, J.-K. (2003). Hot deck imputation for the response model. Submitted for publication.

HANSEN, M.H., MADOW, W.G. and TEPPING, B.J. (1983). An evaluation of model-dependent and probability-sampling inferences in sample surveys. *Journal of the American Statistical Association*. 78, 776–793.

HORVITZ, D.G., and THOMPSON, D.J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*. 47, 663–685.

ISAKI, C.T., and FULLER, W.A. (1982). Survey design under the regression superpopulation model. *Journal of the American Statistical Association*. 77, 89–96.

KALTON, G. (1983). *Compensating for Missing Survey Data*. Institute of Social Research.

KALTON, G., and KASPRZYK, D. (1986). The treatment of missing survey data. *Survey Methodology*. 12, 1–16.

KALTON, G., and MALIGALIG, D.S. (1991). A comparison of methods of weighting adjustment for nonresponse. In *Proceedings of the Bureau of the Census Annual Research Conference*. U.S. Bureau of the Census (Suitland, MD). 409–428.

KIM, J.-K., and FULLER, W.A. (1999). Jackknife variance estimation after hot deck imputation. In *Proceedings of the Section on Survey Research Methods*, American Statistical Association. Alexandria, VA. 825–830.

LITTLE, R.J.A. (1986). Survey nonresponse adjustments for estimates of means. *International Statistical Review*. 54, 139–157.

LITTLE, R.J.A., and RUBIN, D.B. (2002). *Statistical Analysis With Missing Data*. Wiley. 20

OH, H. L., and SCHEUREN, F.J. (1983). Weighting adjustments for unit non-response. In *Incomplete data in sample surveys: Theory and bibliographies*, (Eds. W.G. Madow, I. Olkin, and D.B. Rubin). Academic Press New York: London. 2, 143–184.

SÄRNDAL, C.-E., SWENSSON, B. and WRETMAN, J. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.

SEN, P.K., and SINGER, J.D.M. (1993). *Large Sample Methods in Statistics: An Introduction with Applications*. Chapman and Hall Ltd.

U.S. BUREAU OF THE CENSUS (1963). The Current Population Survey: A report on methodology. Technical Paper No. 7, Washington, DC.

Survey Methodology, June 2004
Vol. 30, No. 1, pp. 57-66
Statistics Canada

Variance Estimation with Hot Deck Imputation Using a Model

J. MICHAEL BRICK, GRAHAM KALTON and JAE KWANG KIM [1]

ABSTRACT

When imputation is used to assign values for missing items in sample surveys, naïve methods of estimating the variances of survey estimates that treat the imputed values as if they were observed give biased variance estimates. This article addresses the problem of variance estimation for a linear estimator in which missing values are assigned by a single hot deck imputation (a form of imputation that is widely used in practice). We propose estimators of the variance of a linear hot deck imputed estimator using a decomposition of the total variance suggested by Särndal (1992). A conditional approach to variance estimation is developed that is applicable to both weighted and unweighted hot deck imputation. Estimation of the variance of a domain estimator is also examined.

KEY WORDS: Missing data; Model-assisted approach; Conditional variance estimation.

1. INTRODUCTION

The important practical problem of estimating the variance of an estimate computed from a data set in which some of the items are missing and values are assigned by imputation has been addressed in a number of different ways (*e.g.*, see Rubin 1987 and Rao and Shao 1992). The approach used in this article is based on the model-assisted approach introduced by Särndal (1992). In the initial application, Särndal used the model-assisted approach with a simple random sample in which the missing data were imputed using deterministic ratio imputation. Subsequently, the approach has been extended to other imputation methods and sample designs (*e.g.*, Deville and Särndal 1994; Rancourt, Särndal and Lee 1994; and Gagnon, Lee, Rancourt and Särndal 1996). This article extends the model-assisted approach to general forms of linear estimators in which missing values have been assigned by hot deck imputation within imputation cells. This form of hot deck imputation, which replaces a missing item by the value observed for a responding unit in the same cell, is one of the most frequently used methods of imputing for missing items in household sample surveys (Brick and Kalton 1996). This paper employs a conditional approach to develop a variance estimator for hot deck imputed estimators that is valid for general sample designs and a variety of estimation strategies.

In the model-assisted approach, the difference between an imputed estimator (the term used here to denote an estimator based in part on imputed values), $\hat{\theta}_I$, and the corresponding finite population parameter, θ_N, is written as

$$\hat{\theta}_I - \theta_N = \left(\hat{\theta}_n - \theta_N\right) + \left(\hat{\theta}_I - \hat{\theta}_n\right), \qquad (1)$$

where $\hat{\theta}_n$ is the usual, approximately design unbiased, estimator of θ_N with complete response. The first term on the right hand side of (1) is called the sampling error and depends only on the sampling distribution of the estimator based on the sample design used to select the full sample, denoted by p. The second term is the imputation error; it depends on the sampling distribution, the response mechanism (R) that generates the respondents from the full sample, and the imputation mechanism (I) for filling in the missing values. This paper is restricted to estimators $\hat{\theta}_I$ that involve only one variable subject to missing data.

We use a model-assisted approach that makes assumptions about the distribution of the variable of interest in the population. We refer to these assumptions as a superpopulation model, denoted by ξ. In general, the aim of imputation is to create a multi-purpose data set that can be validly analyzed in many different ways, potentially involving the associations of a variable subject to imputation with any of the other variables in the data set. Since a superpopulation model is needed to impute for item nonresponses in a way that preserves such associations, it is natural to use that approach also in variance estimation.

Under the superpopulation model, the total variance for an imputed estimator is given by

$$V_{\text{TOT}} = E_\xi E_p E_R E_I \left(\hat{\theta}_I - \theta_N\right)^2, \qquad (2)$$

where E_ξ, E_p, E_R, and E_I refer to expectations with respect to the superpopulation model, the sampling mechanism, the response mechanism, and the imputation mechanism, respectively. We assume that the sample design, response mechanism, and the imputation mechanism are unconfounded as described by Rubin (1987) and used by Särndal (1992) and all of the other literature cited above

[1] J. Michael Brick and Graham Kalton, Westat, 1650 Research Blvd., Rockville, MD 20850, U.S.A. E-mail: mikebrick@westat.com; Jae Kwang Kim, Department of Applied Statistics Yonsei University, Seoul 120-749, Korea.

on the model-assisted approach. Essentially, unconfounded mechanisms allow the order of the expectations to be changed so that the expectation with respect to the model can be taken first. Thus, the total variance can be re-written as $V_{TOT} = E_p E_R E_I E_\xi (\hat{\theta}_I - \theta_N)^2$. Roughly speaking, unconfounded sampling, response, and imputation mechanisms imply that the mechanisms are independent of the distribution of the y-value being analyzed after conditioning on auxiliary variables (*e.g.*, stratification variables for sampling or imputation cells for imputing). Thus, for example, we assume the value of the variable being imputed is independent of the probability of response within each hot-deck cell. Rubin (1987, pages 36-39) has a more detailed discussion of unconfounded mechanisms.

Using the decomposition given in equation (1), Särndal (1992) expressed the total variance for the imputed estimator as

$$V_{TOT} = E_\xi E_p E_R E_I \left(\hat{\theta}_I - \theta_N\right)^2 = V_{SAM} + V_{IMP} + 2V_{MIX}, \quad (3)$$

where $V_{SAM} = E_\xi E_p (\hat{\theta}_n - \theta_N)^2$ is the sampling variance, $V_{IMP} = E_\xi E_p E_R E_I (\hat{\theta}_I - \hat{\theta}_n)^2$ is the imputation variance, and $V_{MIX} = E_\xi E_p E_R E_I [(\hat{\theta}_I - \hat{\theta}_n)(\hat{\theta}_n - \theta_N)]$ is a mixed component. In this formulation, the total variance and its components are more aptly described as anticipated variances because they incorporate the added expectation with respect to the superpopulation model.

The model-assisted approach to variance estimation with imputed data used in this paper should be distinguished from model-assisted sampling (Särndal, Swensson and Wretman 1992). With model-assisted sampling, models are used to guide the choice of efficient sample designs and estimators, but the validity of statistical inferences is not dependent on the validity of the models. In contrast, when some data are missing, reliance on models for inferences is essential, both for point estimators and for variance estimators for them. In this paper, the general approach to inference employs the imputation model assumptions (*i.e.*, superpopulation model and unconfoundedness assumptions) only to the extent necessary to account for imputed data. Both the point estimators and the variance estimators are the standard design-based estimators when no data are missing. Whether the variance estimators are approximately unbiased for V_{SAM} depends on the validity of the imputation model. Also, the estimators for V_{IMP} and V_{MIX} rely completely on the imputation model. Thus the validity of the model is much more critical with model-assisted variance estimation with imputed data than it is with model-assisted sampling. Särndal (1992) argues that if we are willing to accept the validity of the model in point estimation with imputed data, we should also be willing to accept its validity for variance estimation.

Variance estimators are obtained by conditioning on the realized set of sampled units, responding units, and imputations. We develop estimators of $V'_{SAM} = E_\xi[(\hat{\theta}_n - \theta_N)^2 \mid \mathbf{A}, \mathbf{A_R}, \mathbf{d}]$, $V'_{IMP} = E_\xi[(\hat{\theta}_I - \hat{\theta}_n)^2 \mid \mathbf{A}, \mathbf{A_R}, \mathbf{d}]$, and $V'_{MIX} = E_\xi[(\hat{\theta}_n - \theta_N)(\hat{\theta}_I - \hat{\theta}_n) \mid \mathbf{A}, \mathbf{A_R}, \mathbf{d}]$, where $\mathbf{A}$ and $\mathbf{A_R}$ denote matrices of indices for the sampled and responding units, respectively, and $\mathbf{d}$ is the set of indices for the imputations. The conditioning is on the set of indices, not on the values of the units. The matrix $\mathbf{d}$ is an $r \times (n - r)$ matrix in which the rows refer to respondents and the columns to nonrespondents. In this paper, we consider only single imputation methods, in which case all but one of the $d_{ij} = 0$ in every column. The exception occurs in the row of the donor respondent when $d_{ij} = 1$.

By considering the conditional expectations of V'_{IMP} and V'_{MIX}, the estimators reflect the number of times responding units are used as donors in the given application rather than taking the expectation over all possible imputation outcomes. We argue below that these are the appropriate variances to estimate in a given application. If the variance estimators are conditionally unbiased, they are also, of course, unconditionally unbiased.

A conditional approach is useful for two reasons. First, when an estimator is conditionally unbiased and consistent (as $\hat{\theta}_I$ is assumed to be for $\hat{\theta}_n$), the conditional variance is generally a more appropriate estimator for making inferences from a realized sample than an unconditional variance (Holt and Smith 1979, Rao 1999, Kalton 2002). Thus, a variance estimator that conditions on the actual number of times each donor is used is to be preferred to a variance estimator that averages over all possible donor selections. Second, the results apply to any unconfounded sampling, response, and imputation mechanisms that produce the same set of sampled units, respondents, and imputations. Therefore, the results given below for hot deck imputation apply to any unconfounded imputation scheme that substitutes observed values for missing ones and for which $E_\xi(\hat{\theta}_I) = E_\xi(\hat{\theta}_n)$.

2. HOT DECK IMPUTATION

We consider a simple model for which hot deck imputation is appropriate. Assume that the finite population (U) is composed of G classes or cells. Within cell g ($g = 1, ..., G$), the elements in U are realizations of independently and identically distributed random variables with mean μ_g and variance σ_g^2. This cell mean model can be written as

$$Y_i \overset{i.i.}{\sim} \left(\mu_g, \sigma_g^2\right), i \in U_g. \quad (4)$$

where $\overset{i.i.}{\sim}$ is an abbreviation for independently and identically distributed.

A linear estimator of θ_N with complete item response from a complex sample survey can be written as

$$\hat{\theta}_n = \sum_{i \in A} w_i y_i, \qquad (5)$$

where w_i is the weight that accounts for unequal selection probabilities and the estimation strategy. When the cell mean model holds, a more efficient estimator of θ_N uses the unweighted group means, *i.e.*, $\hat{\theta}_n' = \sum\sum w_{gi} \bar{y}_g$ where $\bar{y}_g = \sum_i y_{gi} / n_g$. However, the model-assisted approach does not place complete reliance on the model; rather, it uses the standard design-based approach to the extent possible and the model is used only for the missing data. The weights in (5) can be the inverse of the probability of selection weights or calibration adjusted weights, as described below.

The hot deck imputed value for y_j is $y_j^* = \sum_{i \in A_R} d_{ij} y_i$ and the imputed estimator is

$$\hat{\theta}_I = \sum_{i \in A} w_i \tilde{y}_i = \sum_{i \in A_R} w_i y_i + \sum_{j \in A_M} w_j \sum_{i \in A_R} d_{ij} y_i, \qquad (6)$$

where $\tilde{y}_i = y_i$ for $i \in A_R$ and $\tilde{y}_i = y_i^*$ for $i \in A_M$. We assume throughout that imputed values are selected from respondents in the same imputation cells, and that each cell contains at least one respondent.

This imputation formulation does not specify the way in which donors are selected. It thus covers both unweighted hot deck imputation in which donors are selected with equal probabilities within each cell and weighted hot deck imputation. Weighted hot decks are typically used when assumptions are made only about the response distribution. The form (6) also covers with and without replacement imputation methods. For example, it covers the common hot deck procedure in which a respondent is randomly selected to be a donor within a cell, and then that respondent is not used as a donor again until every other respondent in the cell has been used.

While not explicitly considered here, nearest neighbor imputation procedures that use continuous variables to identify a small set of the most similar respondents and then randomly select one as the donor, satisfy the above requirements. Furthermore, researchers often use hot deck methods even when continuous variables are available. Little (1986) discusses strategies for forming imputation cells using variables that are predictive of the y-variable and notes that imputation within cells and regression imputation should produce similar results in many circumstances. Cochran (1968) and Aigner, Goldberger and Kalton (1975) show that a relatively small number of well-constructed cells formed from a continuous variable can capture a large proportion of the predictive power of the variable.

The conditional bias of the imputed estimator under the cell mean model is

$$E_\xi\left(\hat{\theta}_I - \hat{\theta}_n \mid \mathbf{A}, \mathbf{A_R}, \mathbf{d}\right) =$$

$$E_\xi\left[\sum_{j \in A_M} w_j\left(y_j^* - y_j\right) \mid \mathbf{A}, \mathbf{A_R}, \mathbf{d}\right] = 0,$$

since $E_\xi(y_j^*) = E_\xi(\sum_{i \in A_R} d_{ij} y_i) = \sum_{i \in A_R} d_{ij} E_\xi(y_i) = \sum_{i \in A_R} d_{ij} \mu_g = \mu_g$ for j in cell g. This expectation is conditioned on the indices of the sampled units, the responding units, and the donors. However, since the estimator is conditionally unbiased for any sample, it is also unconditionally unbiased. Kim and Fuller (1999) also use this conditioning argument. Estimators for each component of the variance of the hot deck imputed estimator are given in the next section.

3. ESTIMATION OF THE COMPONENTS OF THE TOTAL VARIANCE

This section contains the main results about estimators of the three components of the total variance of a linear hot deck imputed estimator. Throughout, we assume unconfounded sampling, response, and imputation mechanisms and a linear complete sample estimator of the form (5). The results require that the cell mean model holds and that there is at least one respondent in each imputation cell. We begin with the variance due to sampling, V_{SAM}.

We assume that there exists a complete sample variance estimator, $\hat{V}_n$, that is design unbiased for the sampling variance of $\hat{\theta}_n$, is a quadratic in the y-variable, and is of the form

$$\hat{V}_n = \sum_{i \in A} \sum_{j \in A} \Omega_{ij} y_i y_j = \sum_{i \in A} \sum \Omega_{ii} y_i^2 + 2 \sum_{\substack{i<j \\ i,j \in A}} \Omega_{ij} y_i y_j, \qquad (7)$$

for known coefficients Ω_{ij}. This formulation covers the Horvitz-Thompson estimator, where the Ω_{ij} are determined by the single and joint probabilities of selection. It also covers the linearized variance estimator for the generalized regression (GREG) estimator. Rao, Yung and Hidiroglou (2002) show that the linearized variance estimator for the GREG estimator can be written by substituting $g_{is} e_i$ for y_i in the variance estimator for the Horvitz-Thompson estimator of a total. Here, g_{is} is the sample-dependent g-weight and $e_i = y_i - \mathbf{x}_i' \hat{\mathbf{B}}$, where $\mathbf{x}_i$ is the vector of auxiliary variables and $\hat{\mathbf{B}}$ is the vector of estimated regression coefficients. Since g_{is} is not a function of y and $\hat{\mathbf{B}}$ is linear in the y-variable, $g_{is} e_i$ is linear in y. Therefore, the linearized variance estimator for the GREG estimator is quadratic in y and can be expressed in the form given by

equation (7). Note that in this case the Ω_{ij} may be dependent on the specific sample as well as on the selection probabilities. Deville and Särndal (1992) show that any calibration estimator has the same asymptotic variance as the GREG. Thus, asymptotic variance estimators for calibration estimators in general have the required quadratic form.

The naïve variance estimator treats imputed values as if they were observed values and can be written as

$$\hat{V}_0 = \sum_{i \in A} \sum_{j \in A} \Omega_{ij} \tilde{y}_i \tilde{y}_j. \tag{8}$$

Lemma 1 gives the bias of the naïve variance estimator as an estimator for V'_{SAM}. As noted earlier, the naïve variance estimator is proposed as the estimator of V'_{SAM} to be as consistent as possible with design-based inference. An additional practical reason for using the naïve variance estimator is to take advantage of existing software programs that estimate the sampling variance under complex samples.

Lemma 1. *Under the cell mean model with unconfounded sampling, response, and imputation mechanisms and the assumptions that $\hat{\theta}_I$ is an unbiased hot deck imputed linear estimator given by (6) and $\hat{V}_n$ is an unbiased complete sample variance estimator given by (7), then the bias of the naïve variance estimator, $\hat{V}_0$, as an estimator of V_n is*

$$2 \sum_{g=1}^{G} \sum_{i \in A_{R_g}} \sum_{j \in A_{M_g}} \Omega_{ij} d_{ij} \sigma_g^2 + 2 \sum_{g=1}^{G} \sum_{\substack{i<j \\ i,j \in A_{M_g}}} \sum \Omega_{ij} \gamma_{ij} \sigma_g^2, \tag{9}$$

where $A_{R_g} = A_R \cap U_g$, $A_{M_g} = A_M \cap U_g$, *and*

$$\gamma_{ij} = \sum_{k \in A_R} d_{ki} d_{kj}. \tag{10}$$

For any two nonrespondents, i and j, that have the same donor, $\gamma_{ij} = 1$; $\gamma_{ij} = 0$ otherwise. By definition, $\gamma_{ii} = 1$.

Proof. We begin by noting that the difference between $\hat{V}_0$ and $\hat{V}_n$ can be written as:

$$\begin{aligned}
\hat{V}_0 - \hat{V}_n &= \sum_{i \in A} \Omega_{ii} \left(\tilde{y}_i^2 - y_i^2 \right) \\
&\quad + 2 \sum \sum_{\substack{i<j \\ i,j \in A}} \Omega_{ij} \left(\tilde{y}_i \tilde{y}_j - y_i y_j \right) \\
&= \sum_{i \in A_M} \Omega_{ii} \left(y_i^{*2} - y_i^2 \right) \\
&\quad + 2 \sum \sum_{\substack{i<j \\ i \in A_R, j \in A_M}} \Omega_{ij} \left(y_i y_j^* - y_i y_j \right) \\
&\quad + 2 \sum_{\substack{i<j \\ i,j \in A_M}} \sum \Omega_{ij} \left(y_i^* y_j^* - y_i y_j \right).
\end{aligned} \tag{11}$$

Under the cell mean model, the conditional expectation of the first term of (11) is zero. The conditional expectation $E_\xi(y_i y_j^* - y_i y_j) = E_\xi[y_i(y_j^* - y_j)] = 0$ unless respondent i is the donor for nonrespondent j; it is thus zero when units i and j are in different cells and is only nonzero for one i and j in the same cell g. It may be represented by $E_\xi[y_i(y_j^* - y_j)] = d_{ij}\sigma_g^2$. The conditional expectation $E_\xi(y_i^* y_j^* - y_i y_j)$ is zero unless nonresponding units i and j have the same donor, which can occur only if these units are in the same cell. It can be represented by $E_\xi(y_i^* y_j^* - y_i y_j) = \gamma_{ij}\sigma_g^2$ for $i \neq j$. Applying these results in equation (11) gives

$$\begin{aligned}
E_\xi\left(\hat{V}_0 - \hat{V}_n \mid \mathbf{A}, \mathbf{A_R}, \mathbf{d} \right) &= 2 \sum_{g=1}^{G} \sum_{i \in A_{R_g}} \sum_{j \in A_{M_g}} \Omega_{ij} d_{ij} \sigma_g^2 \\
&\quad + 2 \sum_{g=1}^{G} \sum_{\substack{i<j \\ i,j \in A_{M_g}}} \sum \Omega_{ij} \gamma_{ij} \sigma_g^2.
\end{aligned} \tag{12}$$

The proof is completed by noting that since $\hat{V}_n$ is unbiased under the design, it is also unbiased for V'_{SAM}. Substituting a model unbiased estimator for σ_g^2, say $\hat{\sigma}_g^2$, gives an unbiased estimator of the bias of the naïve variance estimator. Note that whenever respondents donate their values to more than one nonrespondent, the last term in equation (12) is positive; otherwise, it is zero.

Two simple examples illustrate applications of these results. Consider first the estimation of a population mean from a simple random sample selected with replacement. In this case, $\Omega_{ii} = n^{-2}$ and $\Omega_{ij} = -n^{-2}(n-1)^{-1}$ for $i \neq j$. Assume that the cell mean model holds with hot deck imputation and that no donor is used more than once. By Lemma 1, the bias of $\hat{V}_0$ is $-2n^{-2}(n-1)^{-1}\sum_g m_g \sigma_g^2$, where $m_g = \sum_{i \in A_{R_g}} \sum_{j \in A_{M_g}} d_{ij}$ is the number of imputed values in cell g. In this case, the bias of the naïve variance estimator is $O_p(n^{-2})$ and hence is negligible for large n. Now suppose that every missing value in each cell is imputed from the same donor. In this case, with $\sum\sum_{i<j \in A_M} \gamma_{ij} = m_g(m_g - 1)/2$, the bias of $\hat{V}_0$ is $-n^{-2}(n-1)^{-1}\sum_g (m_g^2 + m_g)\sigma_g^2$, which is $O_p(n^{-1})$ and is the same order as the sampling variance.

As the second example, consider a simple two-stage sample of size $n = ab$, in which a clusters are selected from a population of A equal-sized clusters by simple random sampling and b of B elements are selected by simple random sampling within each sampled cluster. Let $y_{\alpha i}$ be the value for y for sampled unit i in cluster α. Assume that the first stage sampling fraction is small enough to ignore. The estimate of the variance of the sample mean is of the form given by equation (7) where $\Omega_{\alpha i, \beta j} = a^{-2} b^{-2} = n^{-2}$ for $\alpha = \beta$, and $\Omega_{\alpha i, \beta j} = -n^{-2}(a-1)^{-1}$ for $\alpha \neq \beta$. These

values can now be inserted into equation (9) to compute an estimate of the bias. For example, suppose that all missing values are imputed using donors from the same cluster (the cells are the clusters) and no donor is used more than once. In this case, the bias of the naïve variance estimator is $2n^{-2}\sum_\alpha m_\alpha \sigma_\alpha^2$, where m_α is the number of nonrespondents in cluster α. Now, suppose an overall cell mean model hot deck is used and no donor can donate more than once, but that donors are always chosen from different clusters than their missing values. In this case, the bias of the naïve variance estimator is $-2n^{-2}(a-1)^{-1}\sigma^2\sum_\alpha m_\alpha$. This two-stage example shows the naïve variance estimator can be biased in either direction. In both of the cases considered, the bias is of lower order than the variance, and if a is large the bias will be negligible.

The second component of the total variance is the variance due to imputation, V_{IMP}. Lemma 2 gives an unbiased estimator for this component with hot deck imputation.

Lemma 2. *Under the assumptions used in Lemma 1, an unbiased estimator of* V'_{IMP} *is*

$$\hat{V}'_{\text{IMP}} = 2\sum_{g=1}^{G}\left\{\sum_{i\in A_{M_g}} w_i^2\,\hat{\sigma}_g^2 + \sum\sum_{\substack{i<j\\ i,j\in A_{M_g}}} w_i w_j \gamma_{ij}\,\hat{\sigma}_g^2\right\}. \quad (13)$$

where $\hat{\sigma}_g^2$ is an unbiased estimator for σ_g^2.

Proof. Since the variance due to imputation involves the squared difference between the imputed and complete response estimates, we begin by writing

$$\begin{aligned}\left(\hat{\theta}_I - \hat{\theta}_n\right)^2 &= \left[\sum_{i\in A} w_i\left(\tilde{y}_i - y_i\right)\right]^2 \\ &= \sum_{i\in A_M} w_i^2\left(y_i^* - y_i\right)^2 \\ &\quad + 2\sum\sum_{\substack{i<j\\ i,j\in A_M}} w_i w_j\left(y_i^* - y_i\right)\left(y_j^* - y_j\right).\end{aligned}$$

Noting that $E_\xi(y_i^* - y_i)^2 = 2\sigma_g^2$ for i in cell g and, from above, $E_\xi[(y_i^* - y_i)(y_j^* - y_j)] = E_\xi(y_i^* y_j^* - y_i y_j)] = \gamma_{ij}\sigma_g^2$, it follows that

$$V'_{\text{IMP}} = 2\sum_{g=1}^{G}\left\{\sum_{i\in A_{M_g}} w_i^2\,\sigma_g^2 + \sum\sum_{\substack{i<j\\ i,j\in A_{M_g}}} w_i w_j \gamma_{ij}\,\sigma_g^2\right\}. \quad (14)$$

Substituting $\hat{\sigma}_g^2$, a model unbiased estimator for σ_g^2, establishes the lemma.

Equation (14) shows that the imputation variance has positive contributions from each imputed value and also from using donors more than once. For example, suppose that the weights for all sampled cases are equal. The contribution to the imputation variance from cell g is then proportional to the sum of the number of missing cases in the cell and the number of pairs of nonrespondents that receive values from the same donors. Limiting the number of times donors are re-used can reduce the imputation variance.

The third term in the total variance is V_{MIX}, which previous research often considered small or negligible (*e.g.*, Särndal 1992; Deville and Särndal 1994). Lemma 3 gives an unbiased estimator for V'_{MIX}.

Lemma 3. *Under the assumptions used in Lemma 1, an unbiased estimator for* V'_{MIX} *is*

$$\sum_{g=1}^{G}\left[\sum_{i\in A_{R_g}}\sum_{j\in A_{M_g}} w_i w_j d_{ij} - \sum_{j\in A_{M_g}} w_j^2\right]\hat{\sigma}_g^2. \quad (15)$$

Proof. Begin by writing $(\hat{\theta}_I - \hat{\theta}_n)(\hat{\theta}_n - \theta_N) = \hat{\theta}_n(\hat{\theta}_I - \hat{\theta}_n) - \theta_N(\hat{\theta}_I - \hat{\theta}_n)$. Let θ_N be the finite population total, which can be written as $\sum_{i\in U-A} y_i + \sum_{i\in A_R} y_i + \sum_{i\in A_M} y_i$. Using this expression, the second component can be expanded as

$$\theta_N\left(\hat{\theta}_I - \hat{\theta}_n\right) = \left(\sum_{i\in U-A} y_i + \sum_{i\in A_R} y_i + \sum_{i\in A_M} y_i\right)\left[\sum_{j\in A_M} w_j\left(y_j^* - y_j\right)\right].$$

In taking the conditional expectation of this product, the only nonzero contributions occur either when unit i in A_R is the donor for y_j^*, or when unit i in A_M in the first set of parentheses is unit j in the second set. In the first case, $E_\xi[y_i(y_j^* - y_j)] = d_{ij}\sigma_g^2$ for $i\in A_{R_g}$, $j\in A_{M_g}$. In the second case, if nonrespondent unit i in A_{M_g} is the same as unit j in the second term, $i=j$, $E_\xi[y_i(y_j^* - y_j)] = -\sigma_g^2$, and this expectation is 0 otherwise. Thus,

$$E_\xi\left(\theta_N\left(\hat{\theta}_I - \hat{\theta}_n\right)\middle|\,\mathbf{A}, \mathbf{A_R}, \mathbf{d}\right) = \sum_g\sum_{j\in A_{M_g}} w_j\sigma_g^2 - \sum_g\sum_{j\in A_{M_g}} w_j\sigma_g^2 = 0.$$

The first term can be expressed as

$$\hat{\theta}_n\left(\hat{\theta}_I - \hat{\theta}_n\right) = \left(\sum_{i\in A_R} w_i y_i + \sum_{i\in A_M} w_i y_i\right)\left[\sum_{j\in A_M} w_j\left(y_j^* - y_j\right)\right].$$

Using the results for $E_\xi(y_i(y_j^* - y_j))$ given above,

$$V'_{\text{MIX}} = \sum_{g=1}^{G}\left(\sum_{i\in A_{R_g}}\sum_{j\in A_{M_g}} w_i w_j d_{ij} - \sum_{g=1}^{G}\sum_{i\in A_{M_g}} w_i^2\right)\sigma_g^2. \quad (16)$$

Substituting an unbiased estimator of σ_g^2 proves the lemma.

The estimator of V'_{MIX} is zero when the weights are constant, or more generally when the weights of the donors are equal to the weights of the missing cases to which they are assigned. Most of the simulations in the literature (*e.g.*, Särndal 1992; Lee, Rancourt and Särndal 1995) have used simple random samples so that the estimates of the mixed term from the simulations are approximately equal to zero.

To illustrate the effect of unequal weights, consider a stratified simple random sample selected from two equal size strata with replacement, and suppose that the sampling rate in stratum 2 is k times the rate in stratum 1. Let the imputation model be the overall cell mean model and let the hot deck procedure select donors with simple random sampling without replacement. For this simple situation, V'_{MIX} can be derived algebraically. Table 1 shows the percentage contribution of the mixed term to the total variance ($100^* 2V'_{MIX}/V'_{TOT}$) for various combinations of strata response rates. The table illustrates the fact that when the sampling weights are unequal, the contribution of the mixed term may be important and can be either positive or negative. The mixed term may also be important in domain estimation, as discussed in the next section.

Table 1
Percentage Contribution of the Mixed Term to V_{TOT}

Response rate		Oversampling rate in stratum 2		
Stratum 1	Stratum 2	$k = 2$	$k = 4$	$k = 6$
100%	80%	4.3	5	13.7
100	60	8.7	10.8	18.3
100	40	13.7	18.3	17.7
100	20	19.9	28.8	29.7
60	100	-15.4	-34.1	-44.5
60	80	-10.4	-27.1	-37.6
60	60	-5.2	-19	-29.3
60	40	1	-8.8	-18.2
60	20	9.4	6.5	0

Now consider estimating the total variance using the three lemmas for the hot deck estimator under the cell mean model. To estimate V'_{SAM} we can either use the naïve variance estimator, with its bias as given in Lemma 1, or correct for the bias with a procedure similar to that recommended by Särndal (1992). For a single stage sample, the bias correction given by Lemma 1 is easy to apply. However, with multi-stage sampling the correction involving Ω may be complicated and difficult to implement in practice. In this case, the naïve variance estimator should produce an adequate approximation provided that the number of sampled clusters is large, that no donor is used too often, and that the percentage of missing data in each cell is not extremely large.

For the other two components, the only unknown quantities that must be estimated from the sample are the cell variances, σ_g^2. These parameters could be estimated using either unweighted observations or weighted observations, where the weights are the selection weights. Fuller (2002) recommends the use of weighted observations to provide more robust estimates. Unbiased estimators of the conditional variance due to imputation and the mixed component are computed by substituting unbiased estimates of the cell variances, $\hat{\sigma}_g^2$. Then, adding $\hat{V}_0$, $\hat{V}'_{IMP}$, and $2\hat{V}'_{MIX}$ gives an estimator of the total variance

$$\hat{V}'_{TOT} = \hat{V}_0 + 2\sum_{g=1}^{G}\sum_{\substack{i<j \\ i,j\in A_{M_g}}}\sum w_i w_j \gamma_{ij} \hat{\sigma}_g^2 + 2\sum_{g=1}^{G}\sum_{i\in A_{R_g}}\sum_{j\in A_{M_g}} w_i w_j d_{ij} \hat{\sigma}_g^2 . \quad (17)$$

To examine this estimator, we give a few simple examples with known solutions. All of these examples involve samples with equal weights so the mixed component is zero. First, assume simple random sampling with replacement, hot deck imputation under the overall cell mean model, and no donor used more than once. Using the naïve variance estimator for V'_{SAM}, the estimated total variance is $n^{-1}s_{\tilde{y}}^2 + 2n^{-1}\hat{\sigma}^2(1 - rn^{-1})$, where $s_{\tilde{y}}^2 = (n-1)^{-1}\Sigma_{i\in A}(\tilde{y}_i - \bar{y}_I)^2$, r is the number of respondents, and m is the number of missing cases. If we use $\hat{\sigma}^2$ instead of $s_{\tilde{y}}^2$ (where $\hat{\sigma}^2$ is model unbiased while $s_{\tilde{y}}^2$ has a small sample bias), then this simplifies to $r^{-1}\hat{\sigma}^2[1 + m(r-m)n^{-2}]$. Taking the expectation of this estimator gives the unconditional variance of the without-replacement hot deck estimator given by Kalton (1983, page 25, 2.3.1.7).

If a multiple cell mean model rather than an overall cell mean model is used, then the estimated total variance is $n^{-1}s_{\tilde{y}}^2 + 2n^{-2}\Sigma_{g=1}^{G}\hat{\sigma}_g^2(n_g - r_g)$, which is similar to the result given by Tollefson and Fuller (1992).

Continuing with the simple random sampling example, now allow donors to be used more than once with the overall cell mean model. Again using $\hat{\sigma}^2$ instead of $s_{\tilde{y}}^2$, the estimated total variance is approximately

$$n^{-2}\hat{\sigma}^2\left(n + m + \sum\sum_{\substack{i<j \\ i,j\in A_M}}\gamma_{ij}\right). \quad (18)$$

For fixed m, the variance in equation (18) is minimized when no donor is used more often than any other donor, to the extent possible (thereby minimizing $\Sigma_{i\in A_M}\Sigma_{j\in A_M}\gamma_{ij}$). Therefore, an imputation scheme that uses any donor at most once more than any other donor minimizes the total variance.

If donors are selected by simple random sampling with replacement, then $E_I[\gamma_{ij}] = r^{-1}$ and the expected value of (18) is $r^{-1}\hat{\sigma}^2[1 + n^{-2}m(r-1)]$. This is the expected variance of the with-replacement hot deck estimator given by Kalton (1983, page 26, 2.3.1.9).

These examples show that the approach produces reasonable estimates for the total variance in simple cases and highlights the conditional nature of the variance estimates. For example, (18) is conditional on the actual number of times donors are used rather than on the expected number of times they are used (the unconditional result). The approach is flexible enough to allow a variety of imputation methods, including with- and without-replacement and weighted and unweighted versions of the hot deck.

4. DOMAIN ESTIMATION

This section considers the important problem of domain estimation under the cell mean model with hot deck imputed data. Previous research on this topic is limited (Lee *et al.* 1995). The standard estimator with complete response for a population total for domain v is $\hat{\theta}_{n_v} = \Sigma_{i\in A_v} w_i y_i$, which may be alternatively expressed as $\hat{\theta}_{n_v} = \Sigma_{i\in A} w_i' y_i$ where $w_i' = \delta_{vi} w_i$ with $\delta_{vi} = 1$ if $i \in A_v$ and $\delta_{vi} = 0$ otherwise. The hot deck imputed estimator is $\hat{\theta}_{I_v} = \Sigma_{i\in A} w_i \delta_{vi} \tilde{y}_i = \Sigma_{i\in A} w_i' \tilde{y}_i$. Throughout we assume that δ_{vi} is known for all $i \in A$.

The cell mean model assumes that all the elements in a cell have the same distribution. In general, some elements in a cell may be in the domain and others not. One version of the model assumes a separate cell mean model for the domain alone and then applies an appropriate imputation scheme. The theory given in the previous section covers this case, and it will, therefore, not be discussed further here. While it is feasible to account for key domains in the imputation stage, it is impossible to consider all possible domains analysts may wish to study. Thus, the focus in this section on domains that cut across imputation cells has important practical implications, especially for analysis of public use data files.

We now discuss the estimation of the three components of V'_{TOT_v}, the variance of an imputed domain total. Consider first the estimation of V'_{SAM_v}. In the case of complete response, by setting $y_i = 0$ for elements outside the domain, the estimated sampling variance can be expressed in the form of equation (7) as $\hat{V}_{n_v} = \Sigma_{i\in A_v} \Omega_{ii} y_i^2 + 2\Sigma\Sigma_{i<j\in A_v} \Omega_{ij} y_i y_j$. With domain membership known for all sample elements, the conditional bias of the imputed variance estimator $\hat{V}_{0_v}$, following the developments in section 3 is:

$$E_\xi\left(\hat{V}_{0_v} - \hat{V}_{n_v} \mid \mathbf{A}, \mathbf{A_R}, \mathbf{d}\right) = 2\sum_{g=1}^{G} \sum_{i\in A_{R_{gv}}} \sum_{j\in A_{M_{gv}}} \Omega_{ij} d_{ij} \sigma_g^2 + 2\sum_{g=1}^{G} \sum\sum_{\substack{i<j \\ i,\, j\in A_{M_{gv}}}} \Omega_{ij} \gamma_{ij} \sigma_g^2. \quad (19)$$

As discussed in section 3, with large samples $\hat{V}_{0_v}$ may be conveniently employed to estimate V'_{SAM_v} using standard survey sampling variance estimation software. It is interesting to note that the naïve variance estimator would be unbiased if all the donors were from outside the domain (thus, $d_{ij} = 0$) and no donor was used more than once ($\gamma_{ij} = 0$).

The derivation of $\hat{V}'_{\mathrm{IMP}_v}$ follows directly from Lemma 2, where the weights are treated as constants in the conditional expectation. Replacing w_i' for w_i in equation (13) gives

$$\hat{V}'_{\mathrm{IMP}_v} = 2\sum_{g=1}^{G} \left\{ \sum_{i\in A_{M_g}} w_i'^2 \hat{\sigma}_g^2 + \sum\sum_{\substack{i<j \\ i,\, j\in A_{M_g}}} w_i' w_j' \gamma_{ij} \hat{\sigma}_g^2 \right\}$$

$$= 2\sum_{g=1}^{G} \left\{ \sum_{i\in A_{M_{gv}}} w_i^2 \hat{\sigma}_g^2 + \sum\sum_{\substack{i<j \\ i,\, j\in A_{M_{gv}}}} w_i w_j \gamma_{ij} \hat{\sigma}_g^2 \right\}.$$

$\hat{V}'_{\mathrm{IMP}}$ does not depend on whether donors come from within or from outside the domain.

The derivation of $\hat{V}'_{\mathrm{MIX}_v}$ also follows from section 3. Substituting w_i' for w_i in equation (15) gives

$$\hat{V}'_{\mathrm{MIX}_v} = \sum_{g=1}^{G} \left(\sum_{i\in A_{R_g}} \sum_{j\in A_{M_g}} w_i' w_j' d_{ij} - \sum_{j\in A_{M_g}} w_j'^2 \right) \hat{\sigma}_g^2$$

$$= \sum_{g=1}^{G} \left(\sum_{i\in A_{R_{gv}}} \sum_{j\in A_{M_{gv}}} w_i w_j d_{ij} - \sum_{j\in A_{M_{gv}}} w_j^2 \right) \hat{\sigma}_g^2. \quad (20)$$

Note that the mixed component is not zero for a domain total, even if all the original weights are equal. With equal weights w (but not equal w'), the contribution to $\hat{V}'_{\mathrm{MIX}}$ is zero when the donor is from inside the domain whereas it is negative when the donor is from outside the domain. As a result, $\hat{V}'_{\mathrm{MIX}_v} = -w^2 \Sigma_g l_{gv} \hat{\sigma}_g^2$, where l_{gv} is the number of donors from outside the domain in cell g. In this case, ignoring the mixed component with domain estimation results in an overestimate of the total variance. With unequal weights, the bias due to ignoring the mixed component can be either positive or negative.

The total variance of a (linear) imputed domain estimator under the cell mean model is then estimated by

$$\hat{V}'_{\mathrm{TOT}_v} = \hat{V}_{0_v} + 2\sum_{g=1}^{G} \sum_{\substack{i<j \\ i,j \in A_{M_g}}} \sum w_i' w_j' \gamma_{ij} \hat{\sigma}_g^2 + 2\sum_{g}^{G} \sum_{i \in A_{R_g}} \sum_{j \in A_{M_g}} w_i' w_j' d_{ij} \hat{\sigma}_g^2 . \quad (21)$$

As an illustration, consider the case of equal weights within the domain ($w_{iv} = w_v$) and no donor used more than once. In this case, the second term on the right in (21) is zero and the third term reflects the variance increase from imputation. If all the missing values are imputed using donors from the domain, then the third term is $2w_v^2 \Sigma m_{gv} \hat{\sigma}_g^2$ where m_{gv} is the number of missing items in cell g and domain v. On the other hand, if no units are imputed from within the domain, then this term is zero. Thus, the total variance is minimized when the donors are selected from outside the domain rather than from within the domain. This result occurs because imputing from outside the domain in effect substitutes a new value for a missing value for domain estimation, thus maintaining the original domain sample size. On the other hand, imputing from within the domain does not increase domain sample size and there is also a penalty to the variance from reusing a domain respondent's value for the nonrespondent.

If the distribution of y varies by domain (*i.e.*, the imputation model is misspecified), then choosing donors from outside the domain results in biased estimates. Since all models are misspecified to some degree, it is therefore generally unwise to intentionally select donors from outside the domain in order to minimize the variance.

5. SIMULATION STUDY

A small simulation study was performed to examine the model-assisted variance estimates for estimating an overall total and a domain total. A sample of 40 clusters with exactly 5 units in each cluster was selected from an infinite superpopulation, where $y_{\alpha i}$ is the study variable for unit i in cluster α. The y-values were generated from $y_{\alpha i} = \tau a_\alpha + e_{\alpha i}$, where a_α and $e_{\alpha i}$ are independent random draws from the standard normal distribution. Thus, the y-values have mean zero, variance $(\tau^2 + 1)$, and correlation $\rho = \tau^2/(1 + \tau^2)$ if the units are from the same cluster and $\rho = 0$ otherwise. Values of $\tau = 0$ and $\tau = 0.5$ were chosen, giving correlations of 0 and 0.2, respectively. The value, $\rho = 0.2$, was chosen to illustrate the effect of a high intraclass correlation. In addition to the y-variable, an indicator variable for domain v was generated by independent sampling with the probability of being in the domain of 0.25. Respondents were selected from the full sample using a uniform response probability of 0.6 and missing values were imputed using a single-cell with-replacement hot deck. A total of 5,000 Monte Carlo samples was selected.

The simulated point estimators for the overall total and the domain total are unbiased. The means and biases of the model-assisted variance estimators ($\hat{V}'_{\mathrm{TOT}}$) are given in Table 2 (the tabulated values are divided by $N^2 10^{-4}$). When $\rho = 0$, the relative biases of the variance estimators for the overall and domain totals are very small. On the other hand, when $\rho = 0.2$, the variance estimators have negative relative biases that are not negligible (a relative bias of -13% for the overall total and -5% for the domain total). To identify the source of the bias, Table 2 also gives the means and biases of the three variance components. The tabled values show that V'_{IMP} and V'_{MIX} are approximately unbiased, and it is only $\hat{V}_0$ that has a non-negligible bias.

When $\rho = 0$ the cell mean model holds and $\hat{V}_0$ is unbiased as expected under the theory. When $\rho = 0.2$, the correlation of the y-values within clusters implies that the cell mean model assumption does not hold. The imputation procedure replaces some missing values using donors from outside the cluster, causing $\hat{V}_0$ to underestimate the sampling variance due to the underestimation of the intraclass correlation. In this particular situation, the model failures do not result in biased estimates for the other two components. However, these components could be biased under other types of model failure. The simulation illustrates the dependence of the model-assisted estimators on the model assumptions and this is discussed further in the next section.

Table 2
Mean and Bias of Simulated Variance Estimators, with Cluster Sampling of 40 Clusters with 5 Elements and Response Rate of 60 Percent#

		$\hat{V}'_{\mathrm{TOT}}$		$\hat{V}_0$		$\hat{V}'_{\mathrm{IMP}}$		$\hat{V}'_{\mathrm{MIX}}$	
Estimate	ρ	Mean	Bias	Mean	Bias	Mean	Bias	Mean	Bias
$\hat{y}$	0	104	-0.5	50	-1.9	54	-1	0	1.2
	0.2	126	-19.6	86	-21	61	0	0	0.9
$\hat{y}_v$	0	16	-0.1	12	0.3	11	0	-4	-0.2
	0.2	18	-1	16	-1	12	0.1	-4	-0.1

The values in the table are actual values divided by $N^2\ 10^{-4}$

6. DISCUSSION

This paper describes a method for estimating the variance of a survey estimate when some of the values are imputed using hot deck imputation. The method uses a model-assisted approach and conditions on indices for sample members, respondents, and hot deck donors. The approach extends the work of Deville and Särndal (1994) to variance estimation for hot deck imputation, probably the most widely used method of imputation in household surveys. The proposed variance estimator is valid for a general

sample design and for a variety of estimation procedures under the superpopulation model and unconfounded assumptions. The paper also extends the previous work by handling stochastic rather than deterministic imputation and giving conditions for the bias of the naïve variance estimator as an estimator of V'_{SAM} to be small.

The results focus attention on the need to take the mixed component into account when the sample elements have unequal weights. In particular, since domain estimates can be treated by assigning adjusted weights of zero for sample elements not in the domain, the mixed term needs to be taken into account in estimating the variance of domain imputed estimates even if the original weights were equal. Other statistics can also be covered by the approach used for domain estimates. For example, for the simple regression of y on x, with y including hot deck imputed values and x complete, the regression coefficient can be expressed as a weighted linear combination of the y's: $b = \Sigma w_i(x_i - \bar{x})y_i/\Sigma w_i(x_i - \bar{x})^2 = \Sigma w_i' y_i$, where $w_i' = w_i(x_i - \bar{x})/\Sigma w_i(x_i - \bar{x})^2$. Also the difference between two domain estimates, $\hat{\theta}_{v1}$ and $\hat{\theta}_{v2}$, can be expressed as $\hat{\theta}_{v1} - \hat{\theta}_{v2} = \Sigma_{i\in v1} w_i y_i - \Sigma_{i\in v2} w_i y_i = \Sigma w_i' y_i$, where $w_i' = w_i$ for $i \in v1$, $w_i' = -w_i$ for $i \in v2$, and $w_i' = 0$ for $i \notin v1 \cup v2$.

The last example, involving the difference between domain estimates where imputation cells cut across domains, highlights the importance of the model in the imputation process. In this example, the analytic interest in the difference between the domain statistics is incompatible with an imputation model that assumes no difference in y-distribution across domains within imputation cells. By imputing across domains with a hot deck cell imputation scheme, the sample domain means for y will be brought closer together, thus decreasing the estimate of the difference. Thus, a good imputation model is crucial for producing valid point estimates.

The model-assisted approach to variance estimation with imputed data described here assumes a linear estimator, but smooth nonlinear functions can also be included using a Taylor series approximation. Like the Rao and Shao (1992) adjusted jackknife method, the model-assisted method is applicable with general sample designs and estimation schemes. However, the adjusted jackknife method is applicable only with a weighted hot-deck whereas, as a result of its model assumptions, the model-assisted method can be employed with a variety of hot deck methods, including choosing donors with equal probability and with probabilities proportional to their weights. The model-assisted method of variance estimation could also be extended to other imputation schemes such as nearest neighbor imputation and fractional hot deck imputation (Kalton and Kish 1984; Fay 1996; Kim 2000), a technique which reduces the variance due to imputation.

Implementation of the model-assisted method with hot deck imputation requires the availability of the information needed to compute the three components of the total variance. Standard survey sampling variance estimation software can be used to compute an estimate of $\hat{V}_0$ that is approximately unbiased with large samples, but as the simulation study illustrates the estimate may be biased if the cell mean model does not hold. The computations of the other components require information on the identity of the donor for each imputed value and of the imputation cell membership of all sample members. From this information, d_{ij} and γ_{ij} can be determined. In addition, an estimate of σ_g^2 is required.

While the theory given above applies to variance estimation with many sample designs, including multi-stage samples, there are serious concerns about the validity of the imputation model in many cases. In the case of multi-stage sampling, the means of many survey variables differ across PSUs, yet hot deck cells are seldom formed within PSUs. Rather they are constructed in terms of other variables that cut across PSUs. Even within these cells there may be differences in means between PSUs. These differences may be offsetting to some extent and not introduce substantial biases for point estimation. However, their effect on variance estimation may be more significant. As indicated in the simulation, failure of the assumptions may have a greater impact on second order statistics than first order statistics. This issue merits more detailed investigation.

Imputation is more difficult when the goal is estimating a function of more than one variable with missing values. To produce an unbiased estimate of a parameter that involves several variables subject to imputation requires the development of an appropriate multivariate model and an imputation procedure consistent with that model. Given an appropriate model and a hot deck imputation that is consistent with it, the model-assisted approach to variance estimation can then be implemented. However, estimating the variance becomes considerably more complex with multivariate estimates. The development of practical methods of imputation and variance estimation for this situation is much needed.

ACKNOWLEDGEMENTS

The authors would like to acknowledge the National Center for Education Statistics, Institute for Education Sciences for funding this research, and in particular the support of the Project Officer, Marilyn Seastrom. We also would like to thank the Associate Editor and referees for their constructive comments.

REFERENCES

AIGNER, D.J., GOLDBERGER, A.S. and KALTON, G. (1975). On the explanatory power of dummy variable regressions. *International Economic Review*. 16, 503-509.

BRICK, J.M., and KALTON, G. (1996). Handling missing data in survey research. *Statistical Methods in Medical Research*. 5, 215-238.

COCHRAN, W.G. (1968). The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics*. 24, 295-313.

DEVILLE, J.C., and SÄRNDAL, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*. 87, 376-382.

DEVILLE, J.C., and SÄRNDAL, C.-E. (1994). Variance estimation for the regression imputed Horvitz-Thompson estimator. *Journal of Official Statistics*. 10, 381-394.

FAY, R.E. (1996). Alternative paradigms for the analysis of imputed survey data. *Journal of the American Statistical Association*. 91, 490-498.

FULLER, W.A. (2002). Regression estimation for sample surveys. *Survey Methodology*. 28, 5-23.

GAGNON, F., LEE, H., RANCOURT, E. and SÄRNDAL, C.-E. (1996). Estimating the variance of the generalized regression estimator in the presence of imputation for the Generalized Estimation System. *Proceedings of the Survey Methods Section*, Statistical Society of Canada. 151-156.

HOLT, D., and SMITH, T.M.F. (1979). Post stratification. *Journal of the Royal Statistical Society. Series A*, 142, Part 1. 33-46.

KALTON, G. (1983). *Compensating for Missing Survey Data*. Ann Arbor: Institute for Social Research, University of Michigan.

KALTON, G. (2002). Models in the practice of survey sampling (revisited). *Journal of Official Statistics*. 18, 129-154.

KALTON, G., and KISH, L. (1984). Some efficient random imputation methods. *Communications in Statistics*. 13(16), 1919-1939.

KIM, J.K. (2000). Variance estimation after imputation. Unpublished Ph.D. thesis, Iowa State University.

KIM, J.K., and FULLER, W.A. (1999). Jackknife variance estimation after hot deck imputation. *Proceedings of the Section on Survey Research Methods*, American Statistical Association. 825-830.

LEE, H., RANCOURT, E. and SÄRNDAL, C.-E. (1995). Variance estimation in the presence of imputed data for the Generalized Estimation System. *Proceedings of the Survey Methods Section*, Statistical Society of Canada. 384-389.

LITTLE, R. (1986). Survey nonresponse adjustments for estimates of means. *International Statistical Review*. 54, 139-157.

RANCOURT, E., SÄRNDAL, C.-E. and LEE, H. (1994). Estimation of the variance in the presence of nearest neighbor imputation. *Proceedings of the Section on Survey Research Methods*, American Statistical Association. 888-893.

RAO, J.N.K. (1999). Some current trends in sample survey theory and methods. *Sankhyā (B)*. 61, 1-57.

RAO, J.N.K., and SHAO, J. (1992). Jackknife variance estimation with survey data under hot deck imputation. *Biometrika*. 79, 811-822.

RAO, J.N.K., YUNG, W. and HIDIROGLOU, M.A. (2002). Estimating equations for the analysis of survey data using poststratification information. *Sankhyā (A)*. 64, 364-378.

RUBIN, D.B. (1987). *Multiple imputation for nonresponse in surveys*. New York: John Wiley & Sons, Inc.

SÄRNDAL, C.-E. (1992). Methods for estimating the precision of survey estimates when imputation has been used. *Survey Methodology*. 18, 241-252.

SÄRNDAL, C.-E., SWENSSON, B. and WRETMAN, J. (1992). *Model assisted survey sampling*. New York: Springer-Verlag.

TOLLEFSON, M., and FULLER, W.A. (1992). Variance estimation for samples with random imputation. *Proceedings of the Section on Survey Research Methods*, American Statistical Association. 140-145.

Survey Methodology, June 2004
Vol. 30, No. 1, pp. 67-78
Statistics Canada

Domain Estimation Using Linear Regression

MICHAEL A. HIDIROGLOU and ZDENEK PATAK [1]

ABSTRACT

One of the main objectives of a sample survey is the computation of estimates of means and totals for specific domains of interest. Domains are determined either before the survey is carried out (primary domains) or after it has been carried out (secondary domains). The reliability of the associated estimates depends on the variability of the sample size as well as on the y-variables of interest. This variability cannot be controlled in the absence of auxiliary information for subgroups of the population. However, if auxiliary information is available, the estimated reliability of the resulting estimates can be controlled to some extent. In this paper, we study the potential improvements in terms of the reliability of domain estimates that use auxiliary information. The properties (bias, coverage, efficiency) of various estimators that use auxiliary information are compared using a conditional approach.

KEY WORDS : Domain estimation; Auxiliary data; Conditional properties.

1. INTRODUCTION

One of the main objectives of a sample survey is to compute estimates of means and totals of a number of characteristics associated with the units of a finite population U. The data are often used for analytic studies such as the comparison of means and totals for subgroups of the population. Such subgroups are referred to as *domains of study*. Hartley's (1959) paper is one of the first attempts to unify the theory of domain estimation. Hartley provided the theory for a number of sample designs where domain estimation was of interest. His paper mostly discussed estimators that did not make use of auxiliary information. He did, however, consider the case of the ratio estimator where population totals were known for the domains. The use of auxiliary data in the context of domain estimation has been discussed in a number of articles. Särndal, Swensson and Wretman (1992) provided a unified treatment of domain estimation with auxiliary data. Estevao, Hidiroglou and Särndal (1995) were the first to recognize that the weights accounting for auxiliary data could be domain dependent or not domain dependent. Estevao and Särndal (1999) discussed desirable properties of regression estimators of domain totals using auxiliary data.

The existence of multivariate auxiliary data raises a number of questions in the context of domain estimation. Some of those questions are as follows. What is the effect of having auxiliary information that is not known on a population basis for the given domain of interest? How do we compute valid variance estimates in the context of domain estimators that use auxiliary data? If more than one estimator is possible for point estimation and/or variance estimation, what criteria should be used to choose the best estimator? Durbin (1969) supported the use of conditional inference to do such comparisons. He stated, "If the sample size is determined by a random mechanism and one happens to get a large sample, one knows perfectly well that the quantities of interest are measured more accurately than they would have been if the sample size had happened to be small. It seems self evident that one should use the information available on sample size in the interpretation of the result. To average over variations in sample size which might have occurred but did not occur, when in fact the sample size is exactly known, seems quite wrong from the standpoint of the analysis of the data actually observed". Holt and Smith (1979) favored conditional inference, and applied it to study the properties of the post-stratified estimator, given simple random sampling. Rao (1985) introduced the idea of "recognizable subsets" of the population to formalize the conditioning process. Recognizable subsets are defined *after* the sample has been drawn. In the case of domain estimation the number of units belonging to a particular domain is a random variable. Recognizable subsets in that context are those where the sample size is fixed within each domain. Comparison of the conditional statistical properties (*i.e.*, bias, mean squared error) of the different estimators can then be based on these subsets. The conditioning process assumes that population totals are known for each domain. In the case of simple random sampling, the number of units in the population domain is assumed known.

The main purpose of this paper is to study the unconditional and conditional properties of a number of domain estimators of totals in the presence of auxiliary data in the context of simple random sampling without replacement (SRSWOR). These conditional properties will be established by conditioning on fixed sample sizes within each domain.

[1] Michael Hidiroglou, Business Survey Methods Division, Statistics Canada, Ottawa, Ontario, Canada, K1A 0T6; Zdenek Patak, Business Survey Methods Division, Statistics Canada, Ottawa, Ontario, Canada, K1A 0T6.

The paper is organized as follows. In section 2 we will introduce several estimators of domain totals. Their unconditional and conditional properties are provided in section 3. In section 4, we will present the results of a simulation study for the case of the ratio estimator of domain totals, and provide some concluding remarks in section 5.

2. ESTIMATORS OF DOMAIN TOTALS

We first introduce some notation to set up the framework, under which we will be assessing the performance of various estimators of domain totals. Let $U = \{1, ..., k, ..., N\}$ denote the finite population. A sample "s" is selected from this population using a sampling plan $P(s)$. Let the first and second order inclusion probabilities be given by π_k and π_{kl}. The domain total $Y_d = \sum_{U_d} y_k$ is the parameter of interest for a variable "y". A domain U_d ($d = 1, ..., D$) is any subpopulation of U, for which a separate estimate may be required, before or after the planning stage. The number of population units in domain U_d is denoted N_d and $N = \sum_{d=1}^{D} N_d$ for D mutually exclusive and exhaustive domains spanning the entire population. The sample s is correspondingly divided into D domains $s_1, ..., s_d, ..., s_D$ where $s_d = U_d \cap s$. The realized sample size within s_d is a random variable that we denote n_d. Note that the sum of the n_d's over non-overlapping and exhaustive domains of the sample equals n. An estimator of the domain total $Y_d = \sum_{U_d} y_k$ that does not use auxiliary data is given by $\hat{Y}_{d,\mathrm{HT}} = \sum_{s_d} w_k\, y_k = \sum_s w_k\, y_{dk}$ where $w_k = \pi_k^{-1}$, and y_{dk} is equal to y_k if $k \in U_d$ and 0 otherwise.

Auxiliary information in the form of a p-dimensional vector $\mathbf{x}$ may be available at different levels of aggregation. It may be known for each unit in the population, or for subsets $U_g \subseteq U$ $(g = 1, ..., G)$ of the population U that may coincide with the domains U_d. We denote such known totals $\mathbf{X}_g = \sum_{U_g} \mathbf{x}_k$; they are estimated by $\hat{\mathbf{X}}_{g,\mathrm{HT}} = \sum_{s_g} w_k\, \mathbf{x}_k$. A modified set of weights $\tilde{w}_k$ incorporating the auxiliary data can be computed using either calibration or linear regression procedures (LR). We chose the LR approach. In the case of G population groups, the LR estimator is given by

$$\hat{Y}_{lr} = \hat{Y}_{\mathrm{HT}} + \sum_{g=1}^{G} (\mathbf{X}_g - \hat{\mathbf{X}}_{g,\mathrm{HT}})'\hat{\mathbf{B}}_g \qquad (1.1)$$

where $\hat{\mathbf{B}}_g = (\sum_{s_g} w_k\, \mathbf{x}_k\, \mathbf{x}_k'/c_k)^{-1} \sum_{s_g} w_k\, \mathbf{x}_k\, y_k/c_k$, and c_k are suitable positive constants. The use of auxiliary data in the domain context offers a wide range of choices for various levels at which auxiliary totals are used and regression models are constructed. To simplify matters, we assume that $g = 1$ (*e.g.*: a single group U), yielding the simple regression estimator $\hat{Y}_{lr} = \hat{Y}_{\mathrm{HT}} + (\mathbf{X} - \hat{\mathbf{X}}_{\mathrm{HT}})'\hat{\mathbf{B}}$, where $\hat{\mathbf{X}}_{\mathrm{HT}} = \sum_s w_k \mathbf{x}_k$.

We consider six estimators for estimating the domain population total Y_d. These estimators are based on whether we use the domain totals $\mathbf{X}_d$ or the population total $\mathbf{X}$, and whether we construct the regression estimator at the domain or at the population levels. The estimators are categorized into Horvitz-Thompson and "Hájek" types. We provide an example of the *ratio* estimator that is associated with each of these estimators.

2.1 Horvitz-Thompson Type Estimators

Case 1

We assume that the auxiliary information $\mathbf{x}_k$ is available at the population level U, $\mathbf{X} = \sum_U \mathbf{x}_k$ and that the domain specific y_{dk} variables are regressed on $\mathbf{x}_k$, $k \in U$. The resulting population regression parameter $\mathbf{B}_{1d} = (\sum_U \mathbf{x}_k\, \mathbf{x}_k'/c_k)^{-1} \sum_U \mathbf{x}_k\, y_{dk}/c_k$ is estimated by $\hat{\mathbf{B}}_{1d} = (\sum_s w_k \mathbf{x}_k\, \mathbf{x}_k'/c_k)^{-1} \sum_s w_k \mathbf{x}_k\, y_{dk}/c_k$ and the resulting estimator of the population total Y_d is

$$\hat{Y}_{d,lr_1} = \hat{Y}_{d,\mathrm{HT}} + (\mathbf{X} - \hat{\mathbf{X}}_{\mathrm{HT}})'\hat{\mathbf{B}}_{1d}. \qquad (2.1)$$

Example: The domain ratio estimator given by $\hat{Y}_{d,\mathrm{RAT}} = X\hat{R}_{1d}$, where $\hat{R}_{1d} = \hat{Y}_{d,\mathrm{HT}}/\hat{X}_{\mathrm{HT}}$. This estimator was first suggested by Hidiroglou (1991), and is discussed in more detail in Estevao *et al.* (1995).

If the auxiliary data totals are available at the domain level, $\mathbf{X}_d = \sum_{U_d} \mathbf{x}_k$, then two possible estimators of Y_d (cases 2 and 3) can be constructed, depending on how the population regression parameter is estimated.

Case 2

The population regression parameter

$$\mathbf{B}_{2d} = \left(\sum_{U_d} \mathbf{x}_k\, \mathbf{x}_k'/c_k\right)^{-1} \sum_{U_d} (\mathbf{x}_k\, y_k/c_k)$$

is estimated by regressing y_k on $\mathbf{x}_k$ for each domain U_d separately. Its estimator is given by

$$\hat{\mathbf{B}}_{2d} = \left(\sum_{s_d} w_k\, \mathbf{x}_k\, \mathbf{x}_k'/c_k\right)^{-1} \sum_{s_d} \left(w_k\, \mathbf{x}_k\, y_k/c_k\right),$$

and the resulting regression estimator of a domain total is

$$\hat{Y}_{d,lr_2} = \hat{Y}_{d,\mathrm{HT}} + (\mathbf{X}_d - \hat{\mathbf{X}}_{d,\mathrm{HT}})'\hat{\mathbf{B}}_{2d} \qquad (2.2)$$

where $\hat{\mathbf{X}}_d = \sum_s w_k\, \mathbf{x}_{dk}$ with $\mathbf{x}_{dk}$ defined similarly to y_{dk}.

Example: The Horvitz-Thompson post-stratified estimator given by $\hat{Y}_{d,\mathrm{POSTR}} = X_d \hat{R}_{2d}$, where $\hat{R}_{2d} = \hat{Y}_{d,\mathrm{HT}}/\hat{X}_{d,\mathrm{HT}}$.

Case 3

The population regression parameter

$$\mathbf{B}_3 = \left(\sum_U (\mathbf{x}_k \mathbf{x}_k'/c_k)\right)^{-1} \sum_U (\mathbf{x}_k y_k/c_k)$$

is estimated by regressing y_k on $\mathbf{x}_k$ using all units in U. The corresponding estimator is

$$\hat{\mathbf{B}}_3 = \left(\sum_s w_k \mathbf{x}_k \mathbf{x}_k'/c_k\right)^{-1} \sum_s (w_k \mathbf{x}_k y_k/c_k),$$

resulting in the regression estimator

$$\hat{Y}_{d,lr_3} = \hat{Y}_{d,\mathrm{HT}} + (\mathbf{X}_d - \hat{\mathbf{X}}_{d,\mathrm{HT}})'\hat{\mathbf{B}}_3. \qquad (2.3)$$

Example: The alternate ratio estimator given by $\hat{Y}_{d,\mathrm{ALTR}} = \hat{Y}_{d,\mathrm{HT}} + (X_d - \hat{X}_{d,\mathrm{HT}})\hat{R}_3$, where $\hat{R}_3 = \hat{Y}_{\mathrm{HT}} / \hat{X}_{\mathrm{HT}}$.

2.2 Hájek Type Estimators

Estimators (2.1)-(2.3) belong to the Horvitz-Thompson family. If the known population domain size N_d is also incorporated in the estimation, then we get the "Hájek" versions of the previously defined Horvitz-Thompson regression estimators. The Hájek regression estimators are obtained by replacing $\hat{Y}_{d,\mathrm{HT}}$, $\hat{\mathbf{X}}_{d,\mathrm{HT}}$, and $\hat{\mathbf{X}}_{\mathrm{HT}}$ by

$$\hat{Y}_{d,\mathrm{HA}} = \left(N_d/\hat{N}_d\right)\hat{Y}_{d,\mathrm{HT}}, \quad \hat{\mathbf{X}}_{d,\mathrm{HA}} = \left(N_d/\hat{N}_d\right)\hat{\mathbf{X}}_{d,}$$

and

$$\hat{\mathbf{X}}_{\mathrm{HA}} = \left(N/\hat{N}\right)\hat{\mathbf{X}}_{\mathrm{HT}},$$

where $\hat{N}_d = \sum_{s_d} w_k$ and $\hat{N} = \sum_s w_k$. The estimators are nearly conditionally unbiased for a given n_d, whereas their Horvitz-Thompson counterparts do not have this property. The "$\hat{\mathbf{B}}$"s contained within the Hájek regression estimators correspond exactly to their Horvitz-Thompson counterparts.

Case 4

$$\tilde{Y}_{d,lr_1} = \hat{Y}_{d,\mathrm{HA}} + (\mathbf{X} - \hat{\mathbf{X}}_{\mathrm{HA}})'\hat{\mathbf{B}}_{1d}. \qquad (2.4)$$

Example: The Hájek ratio estimator given by $\tilde{Y}_{d,\mathrm{RAT}} = \hat{Y}_{d,\mathrm{HA}} + (X - \hat{X}_{\mathrm{HA}})\hat{R}_{1d}$.

Case 5

$$\tilde{Y}_{d,lr_2} = \hat{Y}_{d,\mathrm{HA}} + (\mathbf{X}_d - \hat{\mathbf{X}}_{d,\mathrm{HA}})'\hat{\mathbf{B}}_{2d}. \qquad (2.5)$$

Example: The Hájek post-stratified ratio estimator given by $\tilde{Y}_{d,\mathrm{POSTR}} = \hat{Y}_{d,\mathrm{HA}} + (X_d - \hat{X}_{d,\mathrm{HA}})\hat{R}_{2d}$. This estimator is identical to the Horvitz-Thompson post-stratified estimator.

Case 6

$$\tilde{Y}_{d,lr_3} = \hat{Y}_{d,\mathrm{HA}} + (\mathbf{X}_d - \hat{\mathbf{X}}_{d,\mathrm{HA}})'\hat{\mathbf{B}}_3. \qquad (2.6)$$

Example: The Hájek alternate ratio estimator given by $\tilde{Y}_{d,\mathrm{ALTR}} = \hat{Y}_{d,\mathrm{HA}} + (X_d - \hat{X}_{d,\mathrm{HA}})\hat{R}_3$.

3. PROPERTIES OF THE DOMAIN ESTIMATORS

Estimators (2.1) - (2.6) may be expressed as:

$$\hat{Y}_{d,lr} = \sum_s w_k a_{dk} y_{dk} = \sum_s \tilde{w}_{dk} y_{dk} \qquad (2.7)$$

where a_{dk} is an adjustment factor that may or may not be domain dependent. The product of the design weight w_k and the adjustment factor a_{dk} is known as the regression weight (or calibration weight) $\tilde{w}_{dk}$. Tables 1 and 2 provide a summary of these factors, as well as the residuals required for unconditional variance estimation. The population and sample residuals are denoted as E_{dk} and e_{dk}. The indicator variable δ_{dk} is equal to one if $k \in U_d$ and zero otherwise.

The approximate population and corresponding estimated variances of the Horvitz-Thompson estimators $\hat{Y}_{d,lr_j}$ $(j = 1, 2, 3)$ are:

$$V\left(\hat{Y}_{d,lr_j}\right) = \sum\sum_U \Delta_{k\ell} \left(\frac{E_{dk}}{\pi_k}\right)\left(\frac{E_{d\ell}}{\pi_\ell}\right) \qquad (2.8)$$

and

$$v\left(\hat{Y}_{d,lr_j}\right) = \sum\sum_s \frac{\Delta_{k\ell}}{\pi_{k\ell}} \left(\frac{a_{dk}\, e_{dk}}{\pi_k}\right)\left(\frac{a_{d\ell}\, e_{d\ell}}{\pi_\ell}\right) \qquad (2.9)$$

where $\Delta_{k\ell} = \pi_{k\ell} - \pi_k \pi_\ell$; $\pi_{k\ell} = \Pr\{k, \ell \in s\}$ with the appropriate E_{dk}'s, e_{dk}'s, a_{dk}'s defined in Table 1.

The approximate unconditional population and correspondding estimated variances of the Hájek-type estimators $\tilde{Y}_{d,lr_j}$ $(j = 1, 2, 3)$ are:

$$V\left(\tilde{Y}_{d,lr_j}\right) = \begin{cases} \sum\sum_U \Delta_{k\ell} \left(\dfrac{E_{dk} - \left(\sum_U E_{dk}/N_d\right)\delta_{dk}}{\pi_k}\right) \times \\ \qquad \left(\dfrac{E_{d\ell} - \left(\sum_U E_{dk}/N_d\right)\delta_{d\ell}}{\pi_\ell}\right) \\ \qquad\qquad \text{for j = 1} \\ \sum\sum_{U_d} \Delta_{k\ell} \left(\dfrac{E_{dk} - \tilde{E}_{U_d}}{\pi_k}\right)\left(\dfrac{E_{d\ell} - \tilde{E}_{U_d}}{\pi_\ell}\right) \\ \qquad\qquad \text{for j = 2, 3} \end{cases} \qquad (2.10)$$

Table 1
Adjustment Factors and Residuals for Horvitz-Thompson Regression Estimators

Estimator	Domain Dependent	Adjustment Factor: a_{dk}	Residuals
$\hat{Y}_{d,\ell r_1}$	No	$1+(\mathbf{X}-\hat{\mathbf{X}}_{\mathrm{HT}})'\left(\sum_s \frac{w_k \mathbf{x}_k \mathbf{x}_k'}{c_k}\right)^{-1}\frac{\mathbf{x}_k}{c_k}$	$E_{dk}=y_{dk}-\mathbf{x}_k'\mathbf{B}_{1d}$ $e_{dk}=y_{dk}-\mathbf{x}_k'\hat{\mathbf{B}}_{1d}$
$\hat{Y}_{d,\ell r_2}$	Yes	$\delta_{dk}\left(1+(\mathbf{X}_d-\hat{\mathbf{X}}_{d,\mathrm{HT}})'\left(\sum_{s_d} \frac{w_k \mathbf{x}_k \mathbf{x}_k'}{c_k}\right)^{-1}\frac{\mathbf{x}_k}{c_k}\right)$	$E_{dk}=y_{dk}-\mathbf{x}_{dk}'\mathbf{B}_{2d}$ $e_{dk}=y_{dk}-\mathbf{x}_{dk}'\hat{\mathbf{B}}_{2d}$
$\hat{Y}_{d,\ell r_3}$	Yes	$\delta_{dk}+(\mathbf{X}_d-\hat{\mathbf{X}}_{d,\mathrm{HT}})'\left(\sum_s \frac{w_k \mathbf{x}_k \mathbf{x}_k'}{c_k}\right)^{-1}\frac{\mathbf{x}_k}{c_k}$	$E_{dk}=y_{dk}-\mathbf{x}_{dk}'\mathbf{B}_3$ $e_{dk}=y_{dk}-\mathbf{x}_{dk}'\hat{\mathbf{B}}_3$

Table 2
Adjustment Factors and Residuals for the Hájek-type Estimators

Estimator	Domain Dependent	Adjustment Factor: a_{dk}	Residuals
$\tilde{Y}_{d,\ell r_1}$	Yes	$\frac{N_d}{\hat{N}_d}+\left(\mathbf{X}-\hat{\mathbf{X}}_{\mathrm{HA}}\right)'\left(\sum_s \frac{w_k \mathbf{x}_k \mathbf{x}_k'}{c_k}\right)^{-1}\frac{\mathbf{x}_k}{c_k}$	$E_{dk}=y_{dk}-\mathbf{x}_k'\mathbf{B}_{1d}$ $e_{dk}=y_{dk}-\mathbf{x}_k'\hat{\mathbf{B}}_{1d}$
$\tilde{Y}_{d,\ell r_2}$	Yes	$\delta_{dk}\left(\frac{N_d}{\hat{N}_d}+\left(\mathbf{X}_d-\hat{\mathbf{X}}_{d,\mathrm{HA}}\right)'\left(\sum_{s_d} \frac{w_k \mathbf{x}_k \mathbf{x}_k'}{c_k}\right)^{-1}\frac{\mathbf{x}_k}{c_k}\right)$	$E_{dk}=y_{dk}-\mathbf{x}_{dk}'\mathbf{B}_{2d}$ $e_{dk}=y_{dk}-\mathbf{x}_{dk}'\hat{\mathbf{B}}_{2d}$
$\tilde{Y}_{d,\ell r_3}$	Yes	$\delta_{dk}\frac{N_d}{\hat{N}_d}+\left(\mathbf{X}_d-\hat{\mathbf{X}}_{d,\mathrm{HA}}\right)'\left(\sum_s \frac{w_k \mathbf{x}_k \mathbf{x}_k'}{c_k}\right)^{-1}\frac{\mathbf{x}_k}{c_k}$	$E_{dk}=y_{dk}-\mathbf{x}_{dk}'\mathbf{B}_3$ $e_{dk}=y_{dk}-\mathbf{x}_{dk}'\hat{\mathbf{B}}_3$

and

$$\mathrm{v}\left(\tilde{Y}_{d,\ell r_j}\right)=\sum\sum_s \frac{\Delta_{k\ell}}{\pi_{k\ell}}\left(\frac{a_{dk}e_{dk}}{\pi_k}\right)\left(\frac{a_{d\ell}e_{d\ell}}{\pi_\ell}\right)$$

for j = 1, 2, 3 (2.11)

where $\tilde{E}_{U_d}=\sum_{U_d} E_{dk}/N_d$. The appropriate E_{dk}'s, e_{dk}'s, and a_{dk}'s are defined in Table 2. Note that the form of the estimated unconditional variance is the same for both the Horvitz-Thompson and the Hájek-type estimators.

Result 3.1: The Hájek-type regression estimator can be obtained as a by-product of the regression of y_k on

$$(\mathbf{x}_k^\circ)'=\left(1,(\mathbf{x}_k-\bar{\mathbf{x}}_U)'\right),$$

where $\bar{\mathbf{x}}_U=N^{-1}\sum_U \mathbf{x}_k$. The resulting regression vector is

$$\hat{\mathbf{B}}^\circ=\left(\hat{B}_1^\circ,\hat{\mathbf{B}}_x'\right),$$

where

$$\hat{\mathbf{B}}_x=\left(\left(\sum_s w_k(\mathbf{x}_k-\tilde{\mathbf{x}}_s)(\mathbf{x}_k-\tilde{\mathbf{x}}_s)'/c_k\right)\right)^{-1}\times \sum_s\left(w_k(\mathbf{x}_k-\tilde{\mathbf{x}}_s)y_k/c_k\right)$$

and $\hat{B}_1^\circ=\tilde{y}_s+(\bar{\mathbf{x}}_U-\tilde{\mathbf{x}}_s)'\hat{\mathbf{B}}_x$, with $\tilde{y}_s=\hat{Y}_{\mathrm{HT}}/\hat{N}$ and $\tilde{\mathbf{x}}_s=\hat{\mathbf{X}}_{\mathrm{HT}}/\hat{N}$.

The regression estimator of total $\hat{Y}_{\ell r}=N\hat{B}_1^\circ$ is equal to the Hájek form $\tilde{Y}_{\ell r}=\hat{Y}_{\mathrm{HA}}+(\mathbf{X}-\hat{\mathbf{X}}_{\mathrm{HA}})'\hat{\mathbf{B}}_x$. The various Hájek-type domain regression estimators can be obtained using this approach. For instance, regressing y_{dk} on

$$(\mathbf{x}_k^\circ)'=\left(1,(\mathbf{x}_{dk}-\bar{\mathbf{x}}_{U_d})'\right)$$

yields $\tilde{Y}_{d,\ell r_1}$.

Proof. We first show how to arrive at the Hájek form of the regression estimator. Defining the auxiliary data vector $\mathbf{z}_k$ as $\mathbf{z}_k'=(x_{0k},\mathbf{x}_k')'$, the regression estimator is

$$\hat{Y}_{\ell r}=\hat{Y}_{\mathrm{HT}}+\left(\mathbf{Z}-\hat{\mathbf{Z}}_{\mathrm{HT}}\right)'\hat{\mathbf{B}}_{\mathbf{z}}$$

where

$$\hat{\mathbf{B}}_{\mathbf{z}}=\left(\sum_s w_k\mathbf{z}_k\mathbf{z}_k'/c_k\right)^{-1}\left(\sum_s w_k\mathbf{z}_k y_k/c_k\right),$$

$\mathbf{Z}=\sum_U \mathbf{z}_k$ and $\hat{\mathbf{Z}}_{\mathrm{HT}}=\sum_s w_k\mathbf{z}_k$.

If $x_{0k}=1$, $\hat{Y}_{\ell r}$ is exactly equivalent to $\hat{Y}_{\ell r}=\mathbf{Z}'\hat{\mathbf{B}}_{\mathbf{z}}$. Decomposing $\hat{\mathbf{B}}_{\mathbf{z}}$ as

$$\hat{\mathbf{B}}_{\mathbf{z}}'=\left(\hat{B}_0,\hat{\mathbf{B}}_x'\right),$$

we have that $\hat{Y}_{\ell r}=N\hat{B}_0+\sum_U \mathbf{x}_k'\hat{\mathbf{B}}_x$, where $\hat{B}_0=\tilde{y}_s-\tilde{\mathbf{x}}_s'\hat{\mathbf{B}}_x$ and

$$\hat{\mathbf{B}}_x=\left(\frac{\sum_s w_k(\mathbf{x}_k-\tilde{\mathbf{x}}_s)(\mathbf{x}_k-\tilde{\mathbf{x}}_s)'}{c_k}\right)^{-1}\times \frac{\sum_s w_k(\mathbf{x}_k-\tilde{\mathbf{x}}_s)(y_k-\tilde{y}_s)}{c_k}.$$

Hence, the Hájek form of the regression estimator is

$$\tilde{\mathbf{Y}}_{\ell r} = \hat{\mathrm{Y}}_{\mathrm{HA}} + \left(\mathbf{X}_U - \hat{\mathbf{X}}_{\mathrm{HA}}\right)' \hat{\mathbf{B}}_x.$$

Regressing y_k on

$$(\mathbf{x}_k^\circ)' = \left(1, (\mathbf{x}_k - \bar{\mathbf{x}}_U)'\right)$$

yields the estimated regression vector $\hat{\mathbf{B}}^{\circ\prime} = \left(\hat{B}_1^\circ, \hat{\mathbf{B}}_x'\right)$, where

$$\hat{\mathbf{B}}_x = \left(\frac{\sum_s w_k (\mathbf{x}_k - \tilde{\mathbf{x}}_s)(\mathbf{x}_k - \tilde{\mathbf{x}}_s)'}{c_k}\right)^{-1} \sum_s \frac{w_k (\mathbf{x}_k - \tilde{\mathbf{x}}_s) y_k}{c_k}$$

and $\hat{B}_1^\circ = \tilde{y}_s + (\bar{\mathbf{x}}_U - \tilde{\mathbf{x}}_s)' \hat{\mathbf{B}}_x$. Substituting $\hat{B}_1^\circ$ into $\hat{Y}_{\ell r} = N\hat{B}_1^0$ yields the Hájek form $\tilde{\mathrm{Y}}_{\ell r}$.

Remark 3.1: (Additivity). Suppose that the domains U_d are mutually exclusive ($U_{d_1} \cap U_{d_2} = \emptyset$ for $d_1 \neq d_2$) and exhaustive ($\bigcup_{d=1}^{D} U_d = U$). Additivity over such domains means that $\sum_{d=1}^{D} \hat{Y}_{d,\ell r_1} = \sum_{d=1}^{D} \hat{Y}_{d,\ell r_3} = \hat{Y}_{\ell r}$ where

$$\hat{Y}_{\ell r} = \hat{Y}_{\mathrm{HT}} + \left(\mathbf{X} - \hat{\mathbf{X}}_{\mathrm{HT}}\right)' \hat{\mathbf{B}}.$$

The additive property of $\hat{Y}_{d,\ell r_1}$ is desirable because a single set of calibration weights, $w_k\, a_{dk}$, can be used repeatedly to produce ad hoc domain estimates. Only two out of the six estimators, $\hat{Y}_{d,\ell r_1}$ and $\hat{Y}_{d,\ell r_3}$, are additive over all such domains.

Remark 3.2: (Calibrating on domain auxiliary data). Estevao *et al.* (1999) discussed some of the estimators provided in Tables 1 and 2 for the case of a single auxiliary variable x_k. They arrived at their estimators by controlling on domain information, either via auxiliary variables and /or control totals.

In what follows, we will assume that the sample s of size n has been selected using simple random sampling without replacement (SRSWOR) from a universe of size N. The estimated unconditional variance of the Horvitz-Thompson and Hájek-type estimators for this sampling plan is:

$$\mathrm{v}\left(\hat{Y}_{d,\ell r_j}\right) =$$

$$\mathrm{v}\left(\tilde{Y}_{d,\ell r_j}\right) = \sum_s \frac{N^2(1-f)}{n} \frac{\sum_s \left(a_{dk}\, e_{dk} - \overline{a_d\, e}\right)^2}{n-1} \qquad (2.12)$$

where $\overline{a_d\, e} = \sum_s (a_{dk}\, e_{dk}/n)$ and $f = n/N$ is the sampling fraction.

3.1 Unconditional Properties

The choice between the various regression estimators should be based on the level at which the auxiliary totals are available, as well as bias and variance. All the above estimators are asymptotically unconditionally unbiased; however, their variances differ. We compare the unconditional population variances of the six domain regression estimators (2.1) – (2.6) by distinguishing two cases: (i) an intercept term is included in the regression; and (ii) no intercept term is included in the regression.

Result 3.2: Assume that an intercept is included in the regression, $c_k = c$ for all $k \in U$, and $N > p$, where p refers to the number of auxiliary variables. The following inequalities hold for the population variances of the domain regression estimators (2.1) – (2.6):

(i) $V(\hat{Y}_{d,\ell r_2}) < V(\hat{Y}_{d,\ell r_1})$; $V(\hat{Y}_{d,\ell r_2}) \leq V(\hat{Y}_{d,\ell r_3})$; $V(\hat{Y}_{d,\ell r_3})$ may be smaller, equal or greater to $V(\hat{Y}_{d,\ell r_1})$.

(ii) $V(\tilde{Y}_{d,\ell r_2}) < V(\tilde{Y}_{d,\ell r_1})$ and $V(\tilde{Y}_{d,\ell r_2}) < V(\tilde{Y}_{d,\ell r_3})$; $V(\tilde{Y}_{d,\ell r_3})$ may be smaller, equal or greater to $V(\tilde{Y}_{d,\ell r_1})$.

Proof: In the case of simple random sampling without replacement, $V(\hat{Y}_{d,\ell r_j}) = A\sum_U (E_{dk} - \bar{E}_{U_d})^2$ for $\ell = 1, 2, 3$, where $A = N^2(1-f)/(n(N-1))$ and $\bar{E}_{U_d} = \sum_{U_d} E_{dk}/N$. Given that the regression contains an intercept, it follows that $\sum_U E_{dk} = 0$ or that $\sum_{U_d} E_{dk} = 0$, depending on which regression estimator we use. We only show that (i) holds: the proof for (ii) is similar. The population variances for $\hat{Y}_{d,\ell r_1}$ and $\hat{Y}_{d,\ell r_2}$ are respectively

$$V\left(\hat{Y}_{d,\ell r_1}\right) = A\sum_U (y_{dk} - \mathbf{x}_k' \mathbf{B}_{1d})^2$$

and

$$V\left(\hat{Y}_{d,\ell r_2}\right) = A\sum_U (y_{dk} - \mathbf{x}_k' \mathbf{B}_{2d})^2.$$

The population variance of $\hat{Y}_{d,\ell r_3}$ is

$$V\left(\hat{Y}_{d,\ell r_3}\right) = A\sum_U (E_{dk} - \tilde{E}_{U_d})^2,$$

where

$$\tilde{E}_{U_d} = N^{-1}\sum_U E_{dk} = \left(\frac{N_d}{N}\right)\left(\bar{y}_{U_d} - \bar{\mathbf{x}}_{U_d}' \mathbf{B}_3\right),$$

with $\bar{y}_{U_d} = N_d^{-1}\sum_{U_d} y_{dk}$ and $\bar{\mathbf{x}}_{U_d}'$ similarly defined.

We first show that $V(\hat{Y}_{d,\ell r_2}) < V(\hat{Y}_{d,\ell r_1})$. To this end, we decompose $\sum_U (y_{dk} - \mathbf{x}_k' \mathbf{B}_{1d})^2$ into its within domain U_d and outside domain $U_{\bar{d}}$ components, yielding

$$\sum_U (y_{dk} - \mathbf{x}_k' \mathbf{B}_{1d})^2 = \sum_{U_d} (y_{dk} - \mathbf{x}_k' \mathbf{B}_{1d})^2 + \sum_{U_{\bar{d}}} (y_{dk} - \mathbf{x}_k' \mathbf{B}_{1d})^2.$$

Since

$$\sum_{U_d} (y_{dk} - \mathbf{x}_k' \mathbf{B}_{1d})^2 = \sum_{U_d} (y_{dk} - \mathbf{x}_k' \mathbf{B}_{2d})^2 + \sum_{U_d} \left(\mathbf{x}_k' (\mathbf{B}_{2d} - \mathbf{B}_{1d})\right)^2,$$

it follows that $V(\hat{Y}_{d,\ell r_2}) < V(\hat{Y}_{d,\ell r_1})$.

Next, we show that

$$V\left(\hat{Y}_{d,\ell r_2}\right) \leq V\left(\hat{Y}_{d,\ell r_3}\right).$$

The variance $V(\hat{Y}_{d,\ell r_3})$ can be re-expressed as

$$V\left(\hat{Y}_{d,lr_3}\right)=\sum_{U_d}\left(\begin{array}{l}(y_k-\mathbf{x}_k'\mathbf{B}_3)^2\\ \quad-\left(\frac{N_d^2}{N}\right)\left(\bar{y}_{U_d}-\bar{\mathbf{x}}_{U_d}'\mathbf{B}_3\right)^2\end{array}\right).$$

The difference between $V(\hat{Y}_{d,lr_3})$ and $V(\hat{Y}_{d,lr_2})$ is:

$$\begin{aligned}&V\left(\hat{Y}_{d,lr_3}\right)-V\left(\hat{Y}_{d,lr_2}\right)\\&=A\left\{\begin{array}{l}\sum_{U_d}\left((y_k-\mathbf{x}_k'\mathbf{B}_3)^2\right)-\left(\frac{N_d^2}{N}\right)\left(\bar{y}_{U_d}-\bar{\mathbf{x}}_{U_d}'\mathbf{B}_3\right)^2\\ \qquad-\sum_{U_d}(y_k-\mathbf{x}_k'\mathbf{B}_{2d})^2\end{array}\right\}\\&=A\left\{\begin{array}{l}(\mathbf{B}_3-\mathbf{B}_{2d})'\left(\sum_{U_d}\mathbf{x}_k\mathbf{x}_k'\right)(\mathbf{B}_3-\mathbf{B}_{2d})\\ \quad-\left(\frac{N_d^2}{N}\right)\left(\bar{y}_{U_d}^2-2\mathbf{B}_3'\bar{\mathbf{x}}_{U_d}\bar{y}_{U_d}+\mathbf{B}_3'\bar{\mathbf{x}}_{U_d}\bar{\mathbf{x}}_{U_d}'\mathbf{B}_3\right)\end{array}\right\}\\&\geq A\left\{\begin{array}{l}(\mathbf{B}_3-\mathbf{B}_{2d})'\left(\sum_{U_d}\mathbf{x}_k\mathbf{x}_k'\right)(\mathbf{B}_3-\mathbf{B}_{2d})\\ \quad-N_d\left(\bar{y}_{U_d}^2-2\mathbf{B}_3'\bar{\mathbf{x}}_{U_d}\bar{y}_{U_d}+\mathbf{B}_3'\bar{\mathbf{x}}_{U_d}\bar{\mathbf{x}}_{U_d}'\mathbf{B}_3\right)\end{array}\right\}.\end{aligned}$$

Noting that $\bar{y}_{U_d}=\bar{\mathbf{x}}_{U_d}'\mathbf{B}_{2d}$ it follows that:

$$\begin{aligned}&\bar{y}_{U_d}^2-2\mathbf{B}_3'\bar{\mathbf{x}}_{U_d}'\bar{y}_{U_d}+\mathbf{B}_3'\bar{\mathbf{x}}_{U_d}\bar{\mathbf{x}}_{U_d}'\mathbf{B}_3\\&=\mathbf{B}_{2d}'\bar{\mathbf{x}}_{U_d}\bar{\mathbf{x}}_{U_d}'\mathbf{B}_{2d}-2\mathbf{B}_3'\bar{\mathbf{x}}_{U_d}\bar{\mathbf{x}}_{U_d}'\mathbf{B}_{2d}\\&\qquad+\mathbf{B}_3'\bar{\mathbf{x}}_{U_d}\bar{\mathbf{x}}_{U_d}'\mathbf{B}_3\\&=(\mathbf{B}_3-\mathbf{B}_{2d})'\bar{\mathbf{x}}_{U_d}\bar{\mathbf{x}}_{U_d}'(\mathbf{B}_3-\mathbf{B}_{2d})\end{aligned}$$

Since

$$\sum_{U_d}\mathbf{x}_k\mathbf{x}_k'-N_d\bar{\mathbf{x}}_{U_d}\bar{\mathbf{x}}_{U_d}'=\sum_{U_d}\left(\mathbf{x}_k-\bar{\mathbf{x}}_{U_d}\right)\left(\mathbf{x}_k-\bar{\mathbf{x}}_{U_d}\right)',$$

the difference $V(\hat{Y}_{d,lr_3})-V(\hat{Y}_{d,lr_2})$ can be expressed as:

$$\begin{aligned}&V\left(\hat{Y}_{d,lr_3}\right)-V\left(\hat{Y}_{d,lr_2}\right)\\&=A\left\{(\mathbf{B}_3-\mathbf{B}_{2d})'\left(\sum_{U_d}\mathbf{x}_k\mathbf{x}_k'-N_d\bar{\mathbf{x}}_{U_d}\bar{\mathbf{x}}_{U_d}'\right)(\mathbf{B}_3-\mathbf{B}_{2d})\right\}\\&=A\left\{(\mathbf{B}_3-\mathbf{B}_{2d})'\sum_{U_d}\left(\mathbf{x}_k-\bar{\mathbf{x}}_{U_d}\right)\left(\mathbf{x}_k-\bar{\mathbf{x}}_{U_d}\right)'(\mathbf{B}_3-\mathbf{B}_{2d})\right\}\\&\geq 0.\end{aligned}$$

Finally, we show that $V(\hat{Y}_{d,lr_3})$ may be smaller, equal or greater to $V(\hat{Y}_{d,lr_1})$ by constructing examples:

(i) $V(\hat{Y}_{d,lr_3}) < V(\hat{Y}_{d,lr_1})$, if $\mathbf{B}_3=\mathbf{B}_{2d}$;

(ii) $V(\hat{Y}_{d,lr_3}) = V(\hat{Y}_{d,lr_1})$, if $\mathbf{B}_3=\mathbf{B}_{1d}$;

(iii) $V(\hat{Y}_{d,lr_3}) > V(\hat{Y}_{d,lr_1})$, if the fit of y_k on $\mathbf{x}_k$ is much poorer than the fit y_{dk} on $\mathbf{x}_k$ for $k\in U$.

It can also be shown that $V(\tilde{Y}_{d,lr_2}) < V(\hat{Y}_{d,lr_2})$; $V(\tilde{Y}_{d,lr_3}) < V(\hat{Y}_{d,lr_3})$; and $V(\tilde{Y}_{d,lr_1}) < V(\hat{Y}_{d,lr_1})$. The estimator with the smallest variance is $\tilde{Y}_{d,lr_2}$. However, if it is assumed that the $\hat{\mathbf{B}}_{2d}$'s are similar across all domains, and that there are very few observations in s_d, it may be preferable to use $\tilde{Y}_{d,lr_3}$. The choice between $\tilde{Y}_{d,lr_2}$ and $\tilde{Y}_{d,lr_3}$ should not always be based on the asymptotic variance. If there are very few observations in s_d, this can cause significant bias in $\tilde{Y}_{d,lr_2}$ and also cause the exact variance of $\tilde{Y}_{d,lr_2}$ to be larger than that of $\tilde{Y}_{d,lr_3}$, so that the latter may be preferred.

Remark 3.3: If there is no intercept in the regression, then it does not necessarily follow that Result 3.2 holds.

Proof. We illustrate this statement using the elementary ratio versions of cases 1 and 2. They are respectively the *Horvitz-Thompson ratio* estimator $\hat{Y}_{d,\mathrm{RAT}}=\hat{Y}_{d,\mathrm{HT}}(X/\hat{X}_{\mathrm{HT}})$ and the *Horvitz-Thompson post-stratified ratio* estimator $\hat{Y}_{d,\mathrm{POSTR}}=\hat{Y}_{d,\mathrm{HT}}(X_d/\hat{X}_{d,\mathrm{HT}})$. Also, suppose that the elements of the data vector (y_k,x_k) are positive for all $k\in U$. The population variances for $\hat{Y}_{d,\mathrm{RAT}}$ and $\hat{Y}_{d,\mathrm{POSTR}}$ are $V(\hat{Y}_{\mathrm{d,POSTR}})=A\sum_{U_d}(y_k-B_{2d}x_k)^2$ and $V(\hat{Y}_{d,\mathrm{RAT}})=A\sum_U(y_{dk}-B_{1d}x_k)^2$, where $B_{2d}=Y_d/X_d$, and $B_{1d}=Y_d/X$.

The difference $V(\hat{Y}_{d,\mathrm{RAT}})-V(\hat{Y}_{d,\mathrm{POSTR}})$ can be re-expressed as:

$$\begin{aligned}&A\sum_{U_d}(B_{1d}-B_{2d})^2x_k^2\\&\quad+2A(B_{1d}-B_{2d})\sum_{U_d}(y_k-B_{2d}x_k)x_k\\&\quad+A\sum_{U_{\bar{d}}}(y_{dk}-B_{1d}x_k)^2.\end{aligned}$$

Since the second term of this expression can be positive, negative or zero, the difference $V(\hat{Y}_{d,\mathrm{RAT}})-V(\hat{Y}_{d,\mathrm{POSTR}})$ can be negative.

3.2 Conditional Properties

For a given sample s, let n_d be the realized sample size of s_d. The following result can be used to evaluate the conditional bias of estimators (2.1) to (2.6).

Result 3.3: Let $\mathbf{z}_k$ be an arbitrary p-dimensional vector, that is $\mathbf{z}_k=(z_{k1},\ldots,z_{kp})'$, and suppose that $n_d\geq 1$. The conditional expectation of $\bar{\mathbf{z}}_s=n^{-1}\sum_s\mathbf{z}_k$ given n_d can be written as:

$$\begin{aligned}E(\bar{\mathbf{z}}_s\mid n_d)&=\frac{1}{n}\left[f_d\sum_{U_d}\mathbf{z}_k+f_{\bar{d}}\left(\sum_U\mathbf{z}_k-\sum_{U_d}\mathbf{z}_k\right)\right]\\&=\bar{\mathbf{z}}_U+\frac{w_d-W_d}{1-W_d}\left(\bar{\mathbf{z}}_{U_d}-\bar{\mathbf{z}}_U\right)\end{aligned}\qquad(3.1)$$

where $\bar{\mathbf{z}}_U=N^{-1}\sum_U\mathbf{z}_k$, $\bar{\mathbf{z}}_{U_d}=N_d^{-1}\sum_{U_d}\mathbf{z}_k$, $w_d=n_d/n$, $W_d=N_d/N$, $f_d=n_d/N_d$, $f_{\bar{d}}=n_{\bar{d}}/N_{\bar{d}}$ with $n_{\bar{d}}=n-n_d$, and $N_{\bar{d}}=N-N_d$.

Proof. Rewriting $\bar{\mathbf{z}}_s$ as

$$\frac{1}{n}\left(\sum_{s_d}\mathbf{z}_k+\sum_{s_{\bar{d}}}\mathbf{z}_k\right),$$

we have that

$$E(\bar{\mathbf{z}}_s\mid n_d)=\frac{1}{n}\left[\frac{n_d}{N_d}\sum_{U_d}\mathbf{z}_k+\frac{n-n_d}{N-N_d}\sum_{U_{\bar{d}}}\mathbf{z}_k\right]$$

where $s_{\bar{d}} = \{k \in s \text{ and } k \notin s_d\}$ and

$$U_{\bar{d}} = \{k \in U \text{ and } k \notin U_d\}.$$

Since $\sum_{U_{\bar{d}}} \mathbf{z}_k = \sum_U \mathbf{z}_k - \sum_{U_d} \mathbf{z}_k$, we obtain the required result, that is

$$E(\bar{\mathbf{z}}_s \mid n_d) = \bar{\mathbf{z}}_U + \frac{w_d - W_d}{1 - W_d}(\tilde{\mathbf{z}}_{U_d} - \bar{\mathbf{z}}_U).$$

Result 3.4: The conditional population variance of $\bar{\mathbf{z}}_s$ given n_d, can be written as

$$V(\bar{\mathbf{z}}_s \mid n_d) = \frac{w_d^2}{n_d}(1 - f_d)\mathbf{V}_{z_{U_d}} + \frac{w_{\bar{d}}^2}{n_{\bar{d}}}(1 - f_{\bar{d}})\mathbf{V}_{z_{U_{\bar{d}}}},$$

where

$$\mathbf{V}_{z_{U_d}} = \frac{1}{N_d - 1}\sum_{U_d}(\mathbf{z}_k - \tilde{\mathbf{z}}_{U_d})(\mathbf{z}_k - \tilde{\mathbf{z}}_{U_d})',$$

$$\mathbf{V}_{z_{U_{\bar{d}}}} = \frac{1}{N_{\bar{d}} - 1}\sum_{U_{\bar{d}}}(\mathbf{z}_k - \tilde{\mathbf{z}}_{U_{\bar{d}}})(\mathbf{z}_k - \tilde{\mathbf{z}}_{U_{\bar{d}}})',$$

with $\tilde{\mathbf{z}}_{U_{\bar{d}}} = N_{\bar{d}}^{-1}\sum_{U_{\bar{d}}} \mathbf{z}_k$, and $w_{\bar{d}} = 1 - w_d$.

The estimator of the conditional population variance $\mathbf{V}(\bar{\mathbf{z}}_s \mid n_d)$ is given by

$$\mathbf{v}(\bar{\mathbf{z}}_s \mid n_d) = \frac{w_d^2}{n_d} f_d \mathbf{v}_{z_{s_d}} + \frac{w_{\bar{d}}^2}{n_{\bar{d}}}(1 - f_{\bar{d}})\mathbf{v}_{z_{s_{\bar{d}}}},$$

where

$$\mathbf{v}_{z_{s_d}} = \frac{1}{n_d - 1}\sum_{s_d}(\mathbf{z}_k - \tilde{\mathbf{z}}_{s_d})(\mathbf{z}_k - \tilde{\mathbf{z}}_{s_d})'$$

and

$$\mathbf{v}_{z_{s_{\bar{d}}}} = \frac{1}{n_{\bar{d}} - 1}\sum_{s_{\bar{d}}}(\mathbf{z}_k - \tilde{\mathbf{z}}_{s_{\bar{d}}})(\mathbf{z}_k - \tilde{\mathbf{z}}_{s_{\bar{d}}})',$$

with $\tilde{\mathbf{z}}_{s_d} = n_d^{-1}\sum_{s_d} \mathbf{z}_k$, $\tilde{\mathbf{z}}_{s_{\bar{d}}} = n_{\bar{d}}^{-1}\sum_{s_{\bar{d}}} \mathbf{z}_k$.

Proof: It follows using arguments similar to those used in Result 3.3. We first illustrate how Result 3.3 can be used to obtain the conditional bias for the simpler estimators of domain totals. This includes the Horvitz-Thompson estimator $\hat{Y}_{d,\mathrm{HT}}$, as well as post-stratified ratio estimator $\hat{Y}_{d,\mathrm{POSTR}} = (X_d / \hat{X}_{d,\mathrm{HT}})\hat{Y}_{d,\mathrm{HT}}$. Let $\mathbf{z}_k$ be the domain variable y_{dk}. Using Result 3.3, we have that $E(\hat{Y}_{d,\mathrm{HT}} \mid n_d) = N w_d \tilde{y}_{U_d}$, where $\tilde{y}_{U_d} = Y_d / N_d$. The conditional bias of $\hat{Y}_{d,\mathrm{HT}}$ given n_d is therefore $\mathrm{Bias}(\hat{Y}_{d,\mathrm{HT}} \mid n_d) = N(w_d - W_d)\tilde{y}_{U_d}$. For the post-stratified ratio estimator, note that $\hat{Y}_{d,\mathrm{POSTR}} - Y_d \doteq \hat{Y}_{d,\mathrm{HT}} - (Y_d / X_d)\hat{X}_{d,\mathrm{HT}}$. Defining $\mathbf{z}_k$ as $y_{dk} - (Y_d / X_d)x_{dk}$, we obtain that $\mathrm{Bias}(\hat{Y}_{d,\mathrm{POSTR}} \mid n_d) \doteq 0$.

We next proceed to evaluate the conditional bias and variance of estimators (2.1) – (2.6). We only illustrate the procedure for the regression estimator $\hat{Y}_{d,lr_1}$, as the steps are similar for the other estimators. Conditional on n_d, the distribution of s_d is that of an SRSWOR. This means that, for each sample s_d, n_d can be considered as having been selected from N_d. We express $\hat{Y}_{d,lr_1}$ as $\hat{Y}_{d,lr_1} = \sum_U \hat{y}_k + N/n\sum_s e_{dk}$, where $e_{dk} = y_{dk} - \mathbf{x}_k' \hat{\mathbf{B}}_{1d}$ and $\hat{y}_k = \mathbf{x}_k' \hat{\mathbf{B}}_{1d}$.

Following Särndal and Hidiroglou (1989), we define the conditional regression vector $\mathbf{B}_{1d}^*$ as

$$\mathbf{B}_{1d}^* = \left[E\left(\sum_s \frac{\mathbf{x}_k \mathbf{x}_k'}{c_k}\right)\Big| n_d\right]^{-1} \times E\left[\left(\sum_s \frac{\mathbf{x}_k y_k}{c_k}\right)\Big| n_d\right]. \quad (3.2)$$

The estimated regression vector $\hat{\mathbf{B}}_{1d}$ will converge to $\mathbf{B}_{1d}^*$ (under appropriate conditions) in conditional design probability as n_d and N_d increase.

We have that

$$E\left[\sum_s \frac{\mathbf{x}_k \mathbf{x}_k'}{n c_k}\Big| n_d\right] = \sum_U \frac{\mathbf{x}_k \mathbf{x}_k'}{N c_k} + \mathbf{R}_c$$

and

$$E\left[\sum_s \frac{\mathbf{x}_k y_k}{n c_k}\Big| n_d\right] = \sum_U \frac{\mathbf{x}_k y_k}{N c_k} + \mathbf{r}_c,$$

where

$$\mathbf{R}_c = \frac{w_d - W_d}{1 - W_d}\left(\frac{1}{N_d}\sum_{U_d} \frac{\mathbf{x}_k \mathbf{x}_k'}{c_k} - \frac{1}{N}\sum_U \frac{\mathbf{x}_k \mathbf{x}_k'}{c_k}\right) \doteq \mathbf{0}$$

and

$$\mathbf{r}_c = \frac{w_d - W_d}{1 - W_d}\left(\frac{1}{N_d}\sum_{U_d} \frac{\mathbf{x}_k y_k}{c_k} - \frac{1}{N}\sum_U \frac{\mathbf{x}_k y_k}{c_k}\right) \doteq \mathbf{0}.$$

Consequently, using Result 3.3 and assuming that $(w_d - W_d)/(1 - W_d) \doteq 0$, we have that $\hat{\mathbf{B}}_{1d} \doteq \mathbf{B}_{1d}^*$.

Define the "conditional residual" for the k^{th} unit as

$$E_{dk}^* = y_{dk} - \mathbf{x}_k' \mathbf{B}_{1d}^*. \quad (3.3)$$

The deviation of $\hat{Y}_{d,lr_1}$ from the true value Y_d can be written as

$$\hat{Y}_{d,lr_1} - Y_d = -\sum_U E_{dk}^* + \frac{N}{n}\sum_s E_{dk}^* - \Delta_{1d}^* \quad (3.4)$$

where

$$\Delta_{1d}^* = \left(\frac{N}{n}\sum_s \mathbf{x}_k - \sum_U \mathbf{x}_k\right)'\left(\hat{\mathbf{B}}_{1d} - \mathbf{B}_{1d}^*\right).$$

In equation (3.4), Δ_{1d}^* is of lower order than $N/n\sum_s E_{dk}^*$. To see this, note that

$$E\left[\left(\frac{N}{n}\sum_s \mathbf{x}_k - \sum_U \mathbf{x}_k\right)\Big| n_d\right] = N\frac{w_d - W_d}{1 - W_d}\left(\tilde{\mathbf{x}}_{U_d} - \bar{\mathbf{x}}_U\right),$$

where $(w_d - W_d)/(1 - W_d)$ should be close to zero.

Also, as noted earlier, $\hat{\mathbf{B}}_{1d} - \mathbf{B}_{1d}^*$ is near the vector $\mathbf{0}$ in conditional design probability. Hence $E_{dk}^* = y_{dk} - \mathbf{x}_k' \mathbf{B}_{1d}^* \doteq y_{dk} - \mathbf{x}_k' \mathbf{B}_{1d} = E_{dk}$. This implies that we can write (3.4) as

$$\hat{Y}_{d,lr_1} - Y_d \doteq -\sum_U E_{dk} + \frac{N}{n}\sum_s E_{dk}. \quad (3.5)$$

The conditional expectation of $\hat{Y}_{d,lr_1} - Y_d$ is approximately:

$$E\left[(\hat{Y}_{d,lr_1} - Y_d)\,|\,n_d\right] = N\frac{w_d - W_d}{1 - W_d}\left(\tilde{E}_{U_d} - \bar{E}_{U_d}\right) \quad (3.6)$$

where $\tilde{E}_{U_d} = \sum_{U_d} E_{dk} / N_d$ and $\tilde{E}_{U_d} = \sum_{U} E_{dk} / N$. Since $\bar{y}_{U_d} = W_d\, \tilde{y}_{U_d}$, the conditional expectation (3.6) can be re-expressed as:

$$E\left[\left(\hat{Y}_{d,lr_1} - Y_d\right)\,|\,n_d\right] = N\frac{w_d - W_d}{1 - W_d}\left[\tilde{y}_{U_d}\left(1 - W_d\right) - \left(\tilde{\mathbf{x}}_{U_d} - \bar{\mathbf{x}}_U\right)' \mathbf{B}_{1d}\right]. \quad (3.7)$$

The term $\sum_U E_{dk}$ is constant in (3.5). Using Result 3.4, the conditional population variance of $\hat{Y}_{d,lr_1}$ and its estimated value are respectively

$$V\left(\hat{Y}_{d,lr_1}\,|n_d\right) = N^2\left[\frac{w_d^2}{n_d}\left(1 - f_d\right)V_{E_{U_d}} + \frac{w_{\bar{d}}^2}{n_{\bar{d}}}\left(1 - f_{\bar{d}}\right)V_{E_{U_{\bar{d}}}}\right]$$

and

$$v\left(\hat{Y}_{d,lr_1}\,|n_d\right) = N^2\left[\frac{w_d^2}{n_d}\left(1 - f_d\right)v_{e_{s_d}} + \frac{w_{\bar{d}}^2}{n_{\bar{d}}}\left(1 - f_{\bar{d}}\right)v_{e_{s_{\bar{d}}}}\right]$$

where $V_{E_{U_d}} = \left(N_d - 1\right)^{-1}\sum_{U_d}\left(E_{dk} - \tilde{E}_{U_d}\right)^2$,

$$V_{E_{U_{\bar{d}}}} = \left(N_{\bar{d}} - 1\right)^{-1}\sum_{U_{\bar{d}}}\left(E_{dk} - \tilde{E}_{U_{\bar{d}}}\right)^2,$$

$$v_{e_{s_d}} = \left(n_d - 1\right)^{-1}\sum_{s_d}\left(a_{dk}e_k - \sum_{s_d}\frac{a_{dk}\,e_k}{n_d}\right)^2,$$

and

$$v_{e_{s_{\bar{d}}}} = \left(n_{\bar{d}} - 1\right)^{-1}\sum_{s_{\bar{d}}}\left(a_{dk}e_{dk} - \sum_{s_{\bar{d}}}\frac{a_{dk}\,e_{dk}}{n_{\bar{d}}}\right)^2.$$

The conditional bias and variances of the remaining five estimators can be derived similarly. Table 3 presents a summary of these properties. The required adjustment factors a_{dk} and residual terms e_{dk} are given in Tables 1 and 2.

4. SIMULATION STUDY

A simulation study was carried out to illustrate the conditional and unconditional properties of the ratio version of estimators (2.1) – (2.6). We studied these properties using a population of 1,000 bivariate observations (y,x). This population resulted from the concatenation of two generated population domains: a large domain of size 900 and a small domain of size 100. The (y, x) observations were generated within each domain assuming a ratio model $y_k = \beta x_k + \varepsilon_k$ where $E(\varepsilon_k) = 0$ and $V(\varepsilon_k) = \sigma^2 x_k$. The β coefficients were 1.0 and 3.0 in the large and small domains. The auxiliary variable x was generated using a gamma distribution $\Gamma(a,b)$, where $a = 3$ and $b = 16$. The dependent variable y was also generated by a gamma distribution, $\Gamma(A,B)$ such that the parameters A and B satisfied $E(y_k) = \beta x_k = AB$ and $V(y_k) = \sigma^2 x_k = AB^2$. After solving for A and B, we obtained $A = \beta^2/\sigma^2$ and $B = \sigma^2/\beta$. The term σ^2 was chosen to satisfy a set correlation between x and y defined by

$$\rho_{X,Y} = \frac{\beta b}{\sqrt{\sigma^2 b + \beta^2 b^2}}.$$

The preceding equation yields the constant term

$$\sigma^2 = \beta^2 b\left(\frac{1}{\rho_{X,Y}^2} - 1\right)$$

of the error variance. Common correlation values $\rho_{x,y}$ were used for both domains, ranging from 0.1 to of 0.9 in steps of 0.1, resulting in nine different populations. Random samples (M = 10,000) of size 250 were then repeatedly selected from the populations. For each sample, estimates of domain totals were computed using the estimators given in Table 4. We do not include the Hájek post-stratified estimator, $\tilde{Y}_{d,lr_2}$, as it corresponds exactly to its Horvitz-Thompson analogue, $\hat{Y}_{d,lr_2}$.

Table 3
Conditional Bias and Variance of Estimators (2.1)-(2.6)

Estimator	Conditional Bias	Estimated Conditional Variance
$\hat{Y}_{d,lr_1}$	$N\left((w_d - W_d)/(1 - W_d)\right)\left(\tilde{y}_{U_d}(1 - W_d) - \left(\tilde{\mathbf{x}}_{U_d} - \bar{\mathbf{x}}_U\right)'\mathbf{B}_{1d}\right)$	$N^2\left[\left(w_d^2/n_d\right)\left(1 - f_d\right)v_{e_{s_d}} + \left(w_{\bar{d}}^2/n_{\bar{d}}\right)\left(1 - f_{\bar{d}}\right)v_{e_{s_{\bar{d}}}}\right]$
$\hat{Y}_{d,lr_2}$	Almost 0	$\left(N_d^2(1 - f_d)/n_d\right)\sum_{s_d}\left(\left(a_{dk}\,e_{dk} - \overline{a_d\,e}\right)^2/(n_d - 1)\right)$
$\hat{Y}_{d,lr_3}$	$N\left(w_d - W_d\right)\left(\tilde{y}_{U_d} - \tilde{\mathbf{x}}_{U_d}'\mathbf{B}_3\right)$	$\left((N_d w_d)^2(1 - f_d)/n_d\right)\sum_{s_d}\left(\left(a_{dk}\,e_{dk} - \overline{a_d\,e}\right)^2/(n_d - 1)\right)$
$\tilde{Y}_{d,lr_1}$	$N\left(w_d - W_d\right)\left(\left(\bar{\mathbf{x}}_U - \tilde{\mathbf{x}}_{U_d}\right)'\mathbf{B}_{1d}/(1 - W_d)\right)$	$(N/w_d)^2\left[\left(w_d^2/n_d\right)\left(1 - f_d\right)v_{e_{s_d}} + \left(w_{\bar{d}}^2/n_{\bar{d}}\right)\left(1 - f_{\bar{d}}\right)v_{e_{s_{\bar{d}}}}\right]$
$\tilde{Y}_{d,lr_2}$	Almost 0	$\left(N_d^2(1 - f_d)/n_d\right)\sum_{s_d}\left(\left(a_{dk}\,e_{dk} - \overline{a_d\,e}\right)^2/(n_d - 1)\right)$
$\tilde{Y}_{d,lr_3}$	Almost 0	$\left(N_d^2(1 - f_d)/n_d\right)\sum_{s_d}\left(\left(a_{dk}\,e_{dk} - \overline{a_d\,e}\right)^2/(n_d - 1)\right)$

Table 4
Estimators and Associated Error Terms

Estimator	Ratio Version	Error Term
HT ratio: $\hat{Y}_{d,\ell r_1}$	$\hat{Y}_{d,\mathrm{RAT}} = \hat{Y}_{d,\mathrm{HT}}\left(X/\hat{X}_{\mathrm{HT}}\right)$	$e_{dk} = y_{dk} - \hat{R}_{1d}\, x_k$, $\hat{R}_{1d} = \hat{Y}_{d,\mathrm{HT}}/\hat{X}_{\mathrm{HT}}$
HT post-stratified ratio: $\hat{Y}_{d,\ell r_2}$	$\hat{Y}_{d,\mathrm{POSTR}} = \hat{Y}_{d,\mathrm{HT}}\left(X_d/\hat{X}_{d,\mathrm{HT}}\right)$	$e_{dk} = y_{dk} - \hat{R}_{2d}\, x_{dk}$, $\hat{R}_{2d} = \hat{Y}_{d,\mathrm{HT}}/\hat{X}_{\mathrm{d,HT}}$
HT alternate ratio: $\hat{Y}_{d,\ell r_3}$	$\hat{Y}_{d,\mathrm{ALTR}} = \hat{Y}_{d,\mathrm{HT}} + (X_d - \hat{X}_{d,\mathrm{HT}})\left(\hat{Y}_{\mathrm{HT}}/\hat{X}_{\mathrm{HT}}\right)$	$e_{dk} = y_{dk} - \hat{R}_3 x_k$, $\hat{R}_3 = \hat{Y}_{\mathrm{HT}}/\hat{X}_{\mathrm{HT}}$
Hájek ratio: $\tilde{Y}_{d,\ell r_1}$	$\tilde{Y}_{d,\mathrm{RAT}} = \hat{Y}_{d,\mathrm{HA}} + (X - \hat{X}_{\mathrm{HA}})\left(\hat{Y}_{d,\mathrm{HT}}/\hat{X}_{\mathrm{HT}}\right)$.	$e_{dk} = y_{dk} - \hat{R}_{1d} x_k$, $\hat{R}_{1d} = \hat{Y}_{d,\mathrm{HT}}/\hat{X}_{\mathrm{HT}}$
Hájek alternate ratio: $\tilde{Y}_{d,\ell r_3}$	$\tilde{Y}_{d,\mathrm{ALTR}} = \hat{Y}_{d,\mathrm{HA}} + (X_d - \hat{X}_{d,\mathrm{HA}})\left(\hat{Y}_{\mathrm{HT}}/\hat{X}_{\mathrm{HT}}\right)$	$e_{dk} = y_{dk} - \hat{R}_3 x_k$, $\hat{R}_3 = \hat{Y}_{\mathrm{HT}}/\hat{X}_{\mathrm{HT}}$

4.1 Unconditional Results

The unconditional properties of the estimators were assessed using two performance measures: (i) root mean squared error (RMSE) and (ii) coverage rate (CR). They are:

i. The *RMSE* is defined as

$$\sqrt{\sum_{m=1}^{M}(\hat{Y}_d^{(m)} - Y_d)^2 / M} \ ,$$

where $\hat{Y}_d^{(m)}$ is the estimated total (either Horvitz-Thompson or Hájek type) based on sample m, and M is the total number of samples drawn for the simulation.

ii. The coverage rate *CR* for a given estimator $\hat{Y}$ is defined as the ratio of the number of times that the 95% confidence interval

$$\hat{Y}_d^{(m)} \pm 1.96\sqrt{v(\hat{Y}_d^{(m)})}$$

contains the true population total to the number of replicates. We used the unconditional variances given by (2.12), and the error terms in Table 4 to estimate the required variances.

The four graphs provided in Figures 1 and 2, summarize the unconditional analysis for small and large domains. Also shown is the impact of increasing $\rho_{x,y}$. The square root of the average mean squared error and coverage rates are used to compare the estimators.

In Figure 1, we note that the RMSE decreases substantially with increasing $\rho_{x,y}$. This can be attributed to the decreasing dispersion of the dependent variable conditional on the independent variable as the correlation between the two increases. We also note that the spread of the RMSE is narrower for the large domain than for the small domain. The ranking of the estimators in terms of RMSE from worst to best is as follows: (i) HT ratio (HT RAT), (ii) Hájek ratio (HA RAT), (iii) HT alternate ratio (HT ALTR), (iv) Hájek alternate ratio (HA ALTR), and (v) HT post-stratified ratio (HT POSTR). This ranking is in agreement with Result 3.2.

In Figure 2, we note that the unconditional coverage rates are similar across all the estimators regardless of the correlation $\rho_{x,y}$. For small domains the Horvitz-Thompson estimators exhibit a slight degradation in the coverage rate when $\rho_{x,y}$ is weak. But as the correlation increases, their coverage rate becomes comparable to the Hájek type estimators. The Hájek estimators have a better overall coverage rate than their Horvitz-Thompson counterparts.

4.2 Conditional Results

The conditional properties of the estimators were studied using: (i) average relative conditional bias and (ii) conditional coverage rates. They are defined as:

i. $\mathrm{ARB}_d = (100/M_d)\sum_{m=1}^{M_d}(\hat{Y}_d^{(m)} - Y_d)/Y_d$, where M_d is the number of samples of size n_d.

ii. The conditional coverage rate has the same definition as its unconditional counterpart. The associated variance is

$$v_d^2 = \frac{1}{M_d - 1}\sum_{m=1}^{M_d}(\hat{Y}_d^{(m)} - \overline{Y}_d)^2$$

where

$$\overline{Y}_d = \frac{1}{M_d}\sum_{m=1}^{M_d}\hat{Y}_d^{(m)} \ .$$

Table 5 summarizes the conditional biases of the ratio versions of estimators (2.1) – (2.4) and (2.6). They were obtained from Table 3 using a single auxiliary variable.

The relative conditional bias and coverage rates of the estimators are summarized in Figures 3, 4a, and 4b with respect to the realized sample size n_d for large and small domains, and for two correlations ($\rho_{X,Y} = 0.90$ and $\rho_{X,Y} = 0.60$).

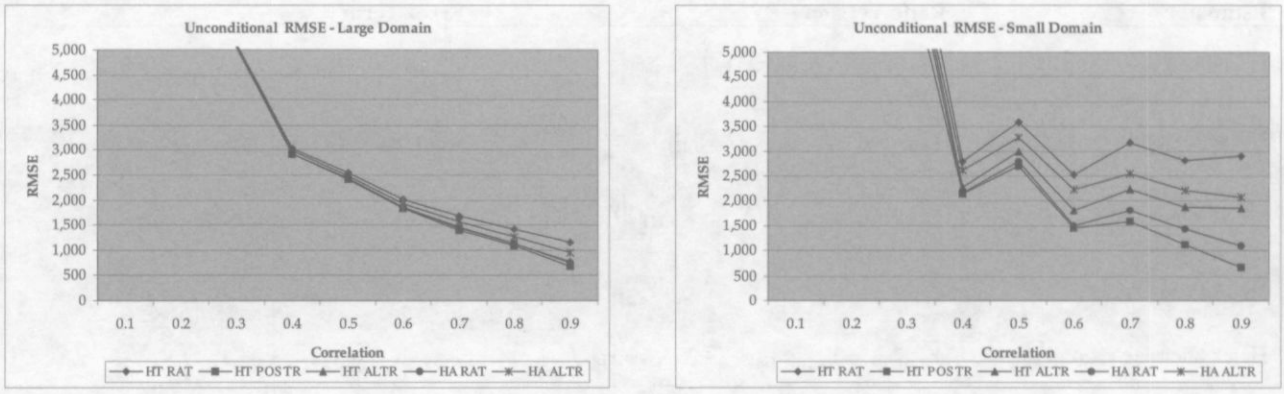


Figure 1. Unconditional RMSE

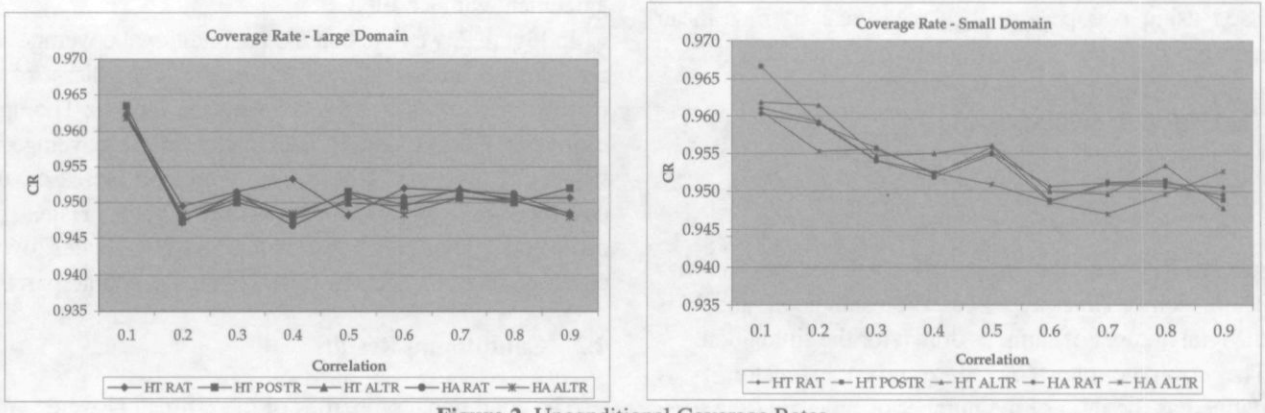


Figure 2. Unconditional Coverage Rates

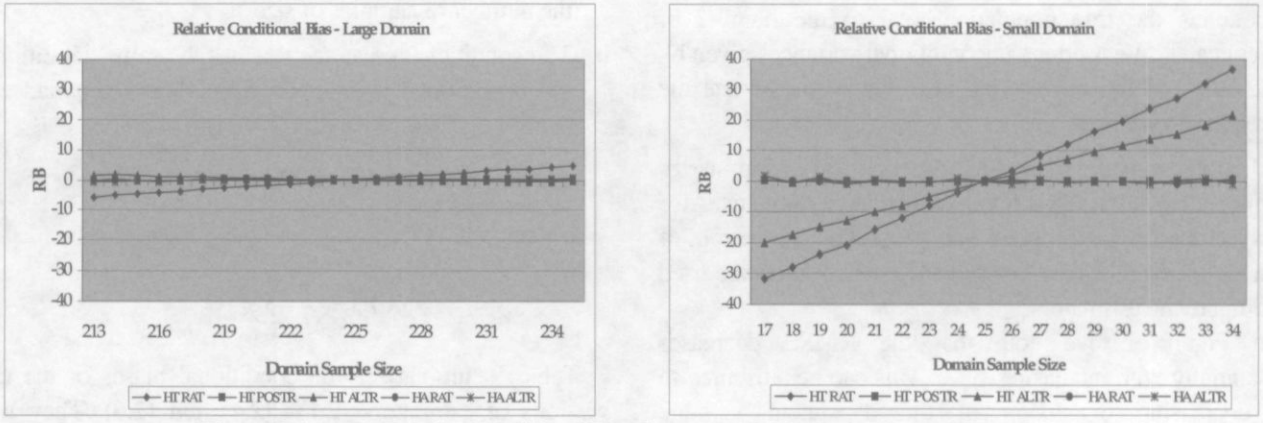


Figure 3. Average Relative Conditional Bias for $\rho_{X,Y} = 0.90$, $\beta_{d1} = 1.0$, and $\beta_{d2} = 3.0$.

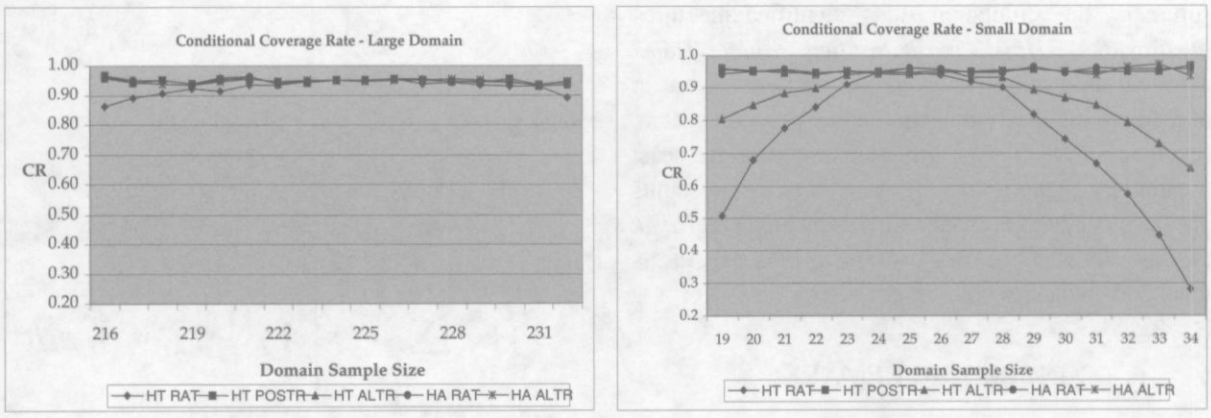


Figure 4a. Conditional Coverage Rates for $\rho_{X,Y} = 0.90$, $\beta_{d1} = 1.0$, and $\beta_{d2} = 3.0$.

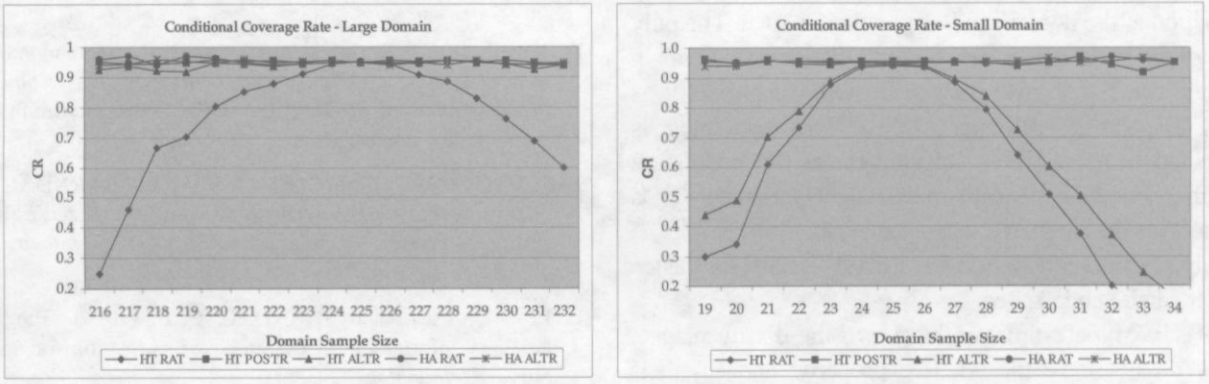


Figure 4b. Conditional Coverage Rates for $\rho_{X,Y} = 0.60$, $\beta_{d1} = 1.0$, and $\beta_{d2} = 3.0$.

Table 5
Conditional Biases of Ratio Versions of Estimators (2.1)-(2.4) and (2.6)

Estimator	Conditional Bias
HT Ratio: $\hat{Y}_{d,\ell r_1}$	$N\,\tilde{y}_{U_d}(w_d - W_d)\dfrac{(\bar{x}_U - W_d\tilde{x}_{U_d})}{\bar{x}_U(1-W_d)}$
HT post-stratified ratio: $\hat{Y}_{d,\ell r_2}$	Almost 0
HT alternate ratio: $\hat{Y}_{d,\ell r_3}$	$N\,\tilde{y}_{U_d}(w_d - W_d)\dfrac{(\bar{x}_U - \tilde{x}_{U_d}\bar{y}_U / \tilde{y}_{U_d})}{\bar{x}_U}$
Hájek ratio: $\tilde{Y}_{d,\ell r_1}$	$N\,\tilde{y}_{U_d}(w_d - W_d)\dfrac{W_d(\bar{x}_U - \tilde{x}_{U_d})}{(1-W_d)\bar{x}_U}$
Hájek alternate ratio: $\tilde{Y}_{d,\ell r_3}$	Almost 0

The conditional bias presented in Figure 3 supports the theoretical results presented in Table 5. The three Hájek estimators are nearly conditionally unbiased. The magnitude of the conditional bias of both the HT ratio estimator and the HT alternate ratio estimator is in agreement with the theoretical conditional bias. But it should be noted that the conditional bias associated with the HT alternate ratio estimator is smaller than the one of the HT ratio estimator. Also, in larger domains, this conditional bias is less pronounced for the HT alternate ratio estimator.

The conditional coverage rates are given in Figures 4a and 4b. We note that the three Hájek estimators follow closely the nominal 95% coverage probability. The coverage rate of the HT alternate ratio estimator is reasonable in larger domains despite its being conditionally biased. But its coverage deteriorates substantially in smaller domains. The coverage rate of the HT ratio estimator is not acceptable. But it should be noted that the coverage rates of the conditionally biased estimators improve as the realized sample size n_d approaches the expected domain sample size $E(n_d)$.

In summary, the simulation study identified the three Hájek estimators, *Hájek post-stratified ratio*, *Hájek alternate ratio*, and *Hájek ratio* as the best estimators in terms of their conditional and unconditional properties. Note that even though the *Hájek ratio* estimator uses the least domain auxiliary data (it uses domain population counts N_d), its mean squared error is still reasonable. The *Hájek post-stratified ratio* is the best estimator in terms of its conditional and unconditional properties.

5. CONCLUDING REMARKS

We have studied six possible regression estimators of domain totals, each using various levels of auxiliary information at the domain and/or population level. The only estimator that has regression weights that are not domain dependent and that also have the additive property is Horvitz-Thompson estimator $\hat{Y}_{d,\ell_1}$. This estimator is constructed using auxiliary information at the population level: the domain dependent independent variable y_{dk} is regressed on the auxiliary vector $\mathbf{x}_k$. However, it can be seriously conditionally biased and the associated confidence intervals can be understated.

The Hájek-type estimators have two the disadvantages: (i) they do not have the additive property; and (ii) their associated regression weights are domain dependent. However, they have the best conditional properties. They are nearly conditionally unbiased, and the conditional confidence intervals associated with the estimators follow closely the nominal coverage rate. They also have the smaller unconditional MSE's. The Hájek estimator that uses the least auxiliary data at the domain level is $\tilde{Y}_{d,\ell_1}$. It requires domain population counts $N_d\,(d = 1, \ldots, D)$, and the population totals $\mathbf{X}$. Its conditional and unconditional properties are reasonable.

The best Hájek estimator, $\tilde{Y}_{d,\ell_2}$, uses auxiliary information at the domain level. The Hájek regression type estimator $\tilde{Y}_{d,\ell_1}$ can be made domain independent using a single set of regression weights as follows. Suppose that the most important domains are $U_g \subseteq U\ (g = 1, \ldots, G)$, and that these domains are mutually exclusive and exhaustive. The resulting Hájek estimator is

$$\tilde{Y}_{d,\ell_1} = \sum_{g=1}^{G} \left[\hat{Y}_{g,\mathrm{HA}} + (\mathbf{X}_g - \hat{\mathbf{X}}_{g,\mathrm{HA}})' \hat{\mathbf{B}}_{1g} \right]$$

where

$$\hat{Y}_{g,\mathrm{HA}} = \left(N_g / \hat{N}_g\right) \hat{Y}_{g,\mathrm{HT}},\ \hat{Y}_{g,\mathrm{HT}} = \sum_{s_g} w_k\ y_{dk}$$

and

$$\hat{\mathbf{B}}_{1g} = \left(\sum_{s_g} w_k\ \mathbf{x}_k\ \mathbf{x}_k' / c_k \right)^{-1} \sum_{s_g} w_k\ \mathbf{x}_k\ y_k / c_k .$$

REFERENCES

DURBIN, J. (1969). Inferential aspects of the randomness of sample size in survey sampling. In *New Developments in Survey Sampling*, (Eds. N.L. Johnson and H. Smith). New York: Wiley, Interscience.

ESTEVAO, V.M., HIDIROGLOU, M.A. and SÄRNDAL, C.-E. (1995). Methodological principles for a generalized estimation system at Statistics Canada. *Journal of Official Statistics*. 11, 2, 181-204.

ESTEVAO, V.M., and SÄRNDAL, C.-E. (1999). The use of auxiliary information in design-based estimation for domains. *Survey Methodology*. 213-231.

HARTLEY, H.O. (1959). *Analytic Studies of Survey Data*. Instituto di Statistica, Rome, Volume in honor of Corrado Gini.

HIDIROGLOU, M.A. (1991). Structure of the Generalized Estimation System (GES). Statistics Canada report, September, 1991.

HOLT, D., and SMITH, T.M.F. (1979). Post-stratification. *Journal of the Royal Statistical Society, Series A*. 142, 33-46.

RAO, J.N.K. (1985). Conditional Inferences in Survey Sampling. *Survey Methodology*. 15-32.

SÄRNDAL, C.-E., and HIDIROGLOU, M.A. (1989). Small Domain Estimation: A conditional analysis. *Journal of American Statistical Association*. 84, 405, 266-275.

SÄRNDAL, C.-E., SWENSSON B. and WRETMAN, J. (1992). *Model Assisted Survey Sampling*. Springer Series in Statistics.

Survey Methodology, June 2004
Vol. 30, No. 1, pp. 79-92
Statistics Canada

Prediction of Finite Population Totals Based on the Sample Distribution

MICHAIL SVERCHKOV and DANNY PFEFFERMANN[1]

ABSTRACT

This article studies the use of the sample distribution for the prediction of finite population totals under single-stage sampling. The proposed predictors employ the sample values of the target study variable, the sampling weights of the sample units and possibly known population values of auxiliary variables. The prediction problem is solved by estimating the expectation of the study values for units outside the sample as a function of the corresponding expectation under the sample distribution and the sampling weights. The prediction mean square error is estimated by a combination of an inverse sampling procedure and a re-sampling method. An interesting outcome of the present analysis is that several familiar estimators in common use are shown to be special cases of the proposed approach, thus providing them a new interpretation. The performance of the new and some old predictors in common use is evaluated and compared by a Monte Carlo simulation study using a real data set.

KEY WORDS: Bootstrap; Design consistency; Informative sampling; Sample-complement distribution.

1. INTRODUCTION

The sample distribution is the parametric distribution of the outcome values for units included in the sample. This distribution is different from the population distribution if the sample selection probabilities are correlated with the values of the study variable even when conditioning on the values of concomitant variables included in the population model. It is also different from the randomization (design) distribution that accounts for all the possible sample selections with the population values held fixed. The sample distribution is defined and discussed with examples in Pfeffermann, Krieger and Rinott (1998), and is further investigated in Pfeffermann and Sverchkov (1999) who use it for the estimation of linear regression models. Krieger and Pfeffermann (1997) use the sample distribution for testing population distribution functions and Pfeffermann and Sverchkov (2003a) discuss its use for fitting Generalized Linear Models. Chambers, Dorfman and Sverchkov (2003) utilize the sample distribution for nonparametric estimation of regression models, and Kim (2002) and Pfeffermann and Sverchkov (2003b) apply it for small area estimation problems.

In this article we study the use of the sample distribution for the prediction of finite population totals under single-stage sampling. It is assumed that the population outcome values (the y-values) are random realizations from some distribution that conditions on known values of auxiliary variables (the x-values). The problem considered is the prediction of the population total Y based on the sample y-values, the sampling weights for units in the sample and the population x-values. The use of the sample distribution permits conditioning on all these values, which is not possible under the randomization (design) distribution, and the prediction of Y is equivalent therefore to the prediction of the y-values for units outside the sample.

The prediction problem is solved by estimating the conditional expectation of the y-values (given the x-values) for units outside the sample as a function of the conditional sample expectation (the expectation under the sample distribution) and the sampling weights. The prediction mean square error is estimated by a combination of an inverse sampling procedure and a re-sampling method. As it turns out, several familiar estimators in common use and in particular, classical design based estimators are special cases of the proposed procedure, thus providing them a new interpretation. The performance of the new and old predictors is evaluated and compared by mean of a Monte Carlo simulation study using a real data set.

2. THE SAMPLE AND SAMPLE-COMPLEMENT DISTRIBUTIONS

2.1 The Sample Distribution

Suppose that the population values $\{\mathbf{y}, X\} = \{(y_1 \ldots y_N)', [\mathbf{x}_1 \ldots \mathbf{x}_N]'\}$ are random realizations with conditional probability density function (*pdf*) $f_p(y_i \mid \mathbf{x}_i)$ that may be discrete or continuous. The y-values are assumed to be scalars but the x-values can be vectors. We consider single stage sampling with sample inclusion probabilities $\pi_i = \Pr(i \in s) = g(\mathbf{y}, X, Z, i)$ for some function g, where Z defines the population values of design variables used for the sampling process. Note that the y-values are random and we also consider the design variables as random so that the

[1] Michail Sverchkov, The Bureau of Labor Statistics, Washington D.C. 20212, U.S.A.; Danny Pfeffermann, Hebrew University, Israël and University of Southampton, U.K.

g-values are random as well. Let $I_i = 1$ if $i \in s$ and $I_i = 0$, if $i \notin s$. The conditional marginal *sample pdf* is defined as,

$$f_s(y_i|\mathbf{x}_i) \overset{\text{def}}{=} f(y_i|\mathbf{x}_i, I_i = 1) = \frac{\Pr(I_i = 1|\, y_i, \mathbf{x}_i) f_p(y_i|\,\mathbf{x}_i)}{\Pr(I_i = 1|\,\mathbf{x}_i)} \quad (2.1)$$

with the second equality obtained by application of Bayes theorem. Note that $\Pr(I_i = 1|\, y_i, \mathbf{x}_i)$ is not necessarily the same as the actual sample selection probability $\pi_i = g(\mathbf{y}, X, Z, i)$ (see Remark 1 below). It follows from (2.1) that the population and sample *pdf*s are different, unless $\Pr(I_i = 1|\, y_i, \mathbf{x}_i) = \Pr(I_i = 1|\,\mathbf{x}_i)$ for all y_i. When the sample distribution differs from the population distribution it becomes *informative*, and the sampling scheme can not be ignored at the inference process.

Remark 1. It is important to emphasize that the definition and use of the sample distribution does not assume that the sample selection probabilities are function of only $(y_i, \mathbf{x}_i)$. As mentioned earlier and highlighted by expressing the selection probabilities as $\pi_i = g(\mathbf{y}, X, Z, i)$, the actual selection probabilities may depend on all the population values $(\mathbf{y}, X, Z)$. However, as shown in Pfeffermann and Sverchkov (1999), $E_p(\pi_i \mid y_i, \mathbf{x}_i) = \Pr(I_i = 1|\, y_i, \mathbf{x}_i)$. Thus, although the selection probabilities may depend on all the population values $(\mathbf{y}, X, Z)$, for given values $(y_i, \mathbf{x}_i)$ they equal $\Pr(I_i = 1|\, y_i, \mathbf{x}_i)$ 'on average'. In fact, π_i may not depend directly on $\mathbf{y}$ at all and only be a function of (X, Z), and still the expectation $E_p(\pi_i \mid y_i, x_i)$ equals $\Pr(I_i = 1|\, y_i, \mathbf{x}_i)$. The reason why the expectation may depend on y_i in this case is that Z may be correlated with $\mathbf{y}$. For example, the 1999 Canadian Workplace and Employee Survey uses a disproportionate stratified sample with the strata defined by region, activity, and the size of the workplace. The size information is obtained from tax records from 1998; see, Patak, Hidiroglou and Lavallée (2000) for details. When modeling the payrolls in 1999 against the number of employees, the sampling design is found to be informative, which is explained by the fact that the stratification is based in part on the size obtained from the tax records in the previous year, which are correlated with the payroll the year after. See Fuller (2003) for details of the analysis.

The discussion above should not be understood to mean that π_i is never a function of $(y_i, \mathbf{x}_i)$ only. A classical example for the latter case is retrospective sampling. Thus, in a case control study, the selection probabilities of the cases and controls usually only depend on the respective y and x values (and often just on the y values). In the empirical study of this paper we use a real data set where the sample was drawn by a disproportionate stratified sample with the strata boundaries defined by the values of the dependent variable.

In what follows we regard the probabilities π_i as random realizations of the random variable $g(\mathbf{y}, X, Z, i)$. Let $w_i = 1/\pi_i$ define the sampling weight of unit i. The following relationships, established in Pfeffermann and Sverchkov (1999) hold for general pairs of vector random variables $(\mathbf{u}_i, \mathbf{v}_i)$, with E_p and E_s defining expectations under the population and sample *pdf*s respectively. (As a special case, $\mathbf{u}_i = y_i$, $\mathbf{v}_i = \mathbf{x}_i$).

$$f_s(\mathbf{u}_i|\mathbf{v}_i) = \frac{E_p(\pi_i|\mathbf{u}_i, \mathbf{v}_i) f_p(\mathbf{u}_i|\mathbf{v}_i)}{E_p(\pi_i|\mathbf{v}_i)} \quad (2.2)$$

$$f_p(\mathbf{u}_i|\mathbf{v}_i) = \frac{E_s(w_i|\mathbf{u}_i, \mathbf{v}_i) f_s(\mathbf{u}_i|\mathbf{v}_i)}{E_s(w_i|\mathbf{v}_i)} \quad (2.3)$$

$$E_p(\mathbf{u}_i|\mathbf{v}_i) = \frac{E_s(w_i\mathbf{u}_i|\mathbf{v}_i)}{E_s(w_i|\mathbf{v}_i)}. \quad (2.4)$$

It follows from (2.4) that

a) $E_s(w_i|\mathbf{v}_i) = \dfrac{1}{E_p(\pi_i|\mathbf{v}_i)}$;

b) $E_p(\mathbf{u}_i) = \dfrac{E_s(w_i\mathbf{u}_i)}{E_s(w_i)}$;

c) $E_s(w_i) = \dfrac{1}{E_p(\pi_i)}$. (2.5)

For a detailed discussion of the sample distribution with illustrations, see Pfeffermann *et al.* (1998).

2.2 The Sample-Complement Distribution

Similar to (2.1), we define the conditional *pdf* for units outside the sample as,

$$f_c(y_i|\mathbf{x}_i) \overset{\text{def}}{=} f_p(y_i|\mathbf{x}_i, I_i = 0) = \frac{\Pr(I_i = 0|y_i, \mathbf{x}_i) f_p(y_i|\,\mathbf{x}_i)}{\Pr(I_i = 0|\mathbf{x}_i)}. \quad (2.6)$$

The relationships (2.2) – (2.5) and the equality $\Pr(I_i = 0 \mid \mathbf{u}_i, \mathbf{v}_i) = 1 - \Pr(I_i = 1|\, \mathbf{u}_i, \mathbf{v}_i) = 1 - E_p(\pi_i \mid \mathbf{u}_i, \mathbf{v}_i)$ imply the following representations of the sample-complement distribution for general pairs of vector random variables $(\mathbf{u}_i, \mathbf{v}_i)$.

$$f_c(\mathbf{u}_i|\mathbf{v}_i) = \frac{E_p[(1-\pi_i)|\mathbf{u}_i, \mathbf{v}_i] f_p(\mathbf{u}_i|\mathbf{v}_i)}{E_p[(1-\pi_i)|\mathbf{v}_i]}$$

$$= \frac{E_p[(1-\pi_i)|\mathbf{u}_i, \mathbf{v}_i]}{E_p[(1-\pi_i)|\mathbf{v}_i]} \frac{E_p[\pi_i|\mathbf{v}_i]}{E_p[\pi_i|\mathbf{u}_i, \mathbf{v}_i]} f_s(\mathbf{u}_i|\mathbf{v}_i) \quad (2.7)$$

$$f_c(\mathbf{u}_i|\mathbf{v}_i) = \frac{E_s[(w_i-1)|\mathbf{u}_i,\mathbf{v}_i]f_s(\mathbf{u}_i|\mathbf{v}_i)}{E_s[(w_i-1)|\mathbf{v}_i]}. \quad (2.8)$$

(Equation (2.8) follows by application of (2.5a) to the second expression in (2.7)). Also, by (2.8) and the first equation in (2.7),

$$E_c(\mathbf{u}_i|\mathbf{v}_i) = \frac{E_p[(1-\pi_i)\mathbf{u}_i|\mathbf{v}_i]}{E_p[(1-\pi_i)|\mathbf{v}_i]} = \frac{E_s[(w_i-1)\mathbf{u}_i|\mathbf{v}_i]}{E_s[(w_i-1)|\mathbf{v}_i]}. \quad (2.9)$$

Remark 2. In practical applications the sampling fraction is often very small and hence the sample selection probabilities are small for at least most of the population units. If $\pi_i < \delta$ with probability 1,

$$f_c(\mathbf{u}_i|\mathbf{v}_i) = \frac{E_p[(1-\pi_i)|\mathbf{u}_i,\mathbf{v}_i]f_p(\mathbf{u}_i|\mathbf{v}_i)}{E_p[(1-\pi_i)|\mathbf{v}_i]}$$

$$= f_p(\mathbf{u}_i|\mathbf{v}_i) + \frac{E_p\{[E_p(\pi_i|\mathbf{v}_i)-\pi_i]|\mathbf{u}_i,\mathbf{v}_i\}f_p(\mathbf{u}_i|\mathbf{v}_i)}{E_p[(1-\pi_i)|\mathbf{v}_i]}$$

$$= f_p(\mathbf{u}_i \mid \mathbf{v}_i)(1+\Delta) \quad (2.10)$$

where $-\delta < \Delta < \delta/(1-\delta)$. It follows from (2.10) that for δ sufficiently small, the difference between the population *pdf* and the sample-complement *pdf* is accordingly small, which is not surprising.

3. OPTIMAL PREDICTION OF FINITE POPULATION TOTALS

Let $Y = \sum_{i=1}^N y_i$ define the population total. The problem considered is how to predict Y based on the sample data and possibly population values of auxiliary variables. Denote the 'design information' available for prediction by $D_s = \{(y_i, w_i), i \in s;\ (\mathbf{x}_j, I_j), j = 1...N\}$ and let $\hat{Y} = \hat{Y}(D_s)$ define the predictor. The MSE of $\hat{Y}$ with respect to the population *pdf* given D_s is,

$$\begin{aligned}\mathrm{MSE}(\hat{Y}\mid D_s) &= E_p[(\hat{Y}-Y)^2|D_s] \\ &= E_p\{[\hat{Y}-E_p(Y|D_s)]^2|D_s\} + V_p(Y\mid D_s) \\ &= [\hat{Y}-E_p(Y|D_s)]^2 + V_p(Y|D_s) \quad (3.1)\end{aligned}$$

since $[\hat{Y}-E_p(Y\mid D_s)]$ is fixed given D_s. It follows from (3.1) that $\mathrm{MSE}(\hat{Y}\mid D_s)$ is minimized when $\hat{Y} = E_p(Y\mid D_s)$. The latter expectation can be decomposed as,

$$\begin{aligned}E_p(Y|D_s) &= \sum_{i=1}^N E_p(y_i|D_s) \\ &= \sum_{i\in s} E_p(y_i|D_s, I_i = 1) + \sum_{j\notin s} E_p(y_j|D_s, I_j = 0) \\ &= \sum_{i\in s} y_i + \sum_{j\notin s} E_c(y_j|D_s) \\ &= \sum_{i\in s} y_i + \sum_{j\notin s} E_c(y_j|\mathbf{x}_j) \quad (3.2)\end{aligned}$$

where in the last equality we assume that y_j for $j \notin s$ and D_s are uncorrelated given $\mathbf{x}_j$. The prediction problem reduces therefore to the estimation of the expectations $E_c(y_j \mid \mathbf{x}_j)$. In section 4 we consider semi-parametric estimation of these expectations.

4. SEMI-PARAMETRIC PREDICTION OF FINITE POPULATION TOTALS

Suppose that the sample-complement model takes the form,

$$\begin{aligned}y_j &= C_\beta(\mathbf{x}_j) + \varepsilon_j, \\ E_c(\varepsilon_j|\mathbf{x}_j) &= 0,\ E_c(\varepsilon_j^2|\mathbf{x}_j) = \sigma^2 v(\mathbf{x}_j), \\ E_c(\varepsilon_k\varepsilon_j|\mathbf{x}_k, \mathbf{x}_j) &= 0,\ k \neq j \quad (4.1)\end{aligned}$$

where $C_\beta(\mathrm{x})$ is a known (possibly nonlinear) function of $\mathbf{x}$ that depends on an unknown vector parameter β. The variances $\sigma^2\, v(\mathbf{x}_j)$ are assumed known except for σ^2.

Remark 3. In actual applications the model (4.1) can be identified by a two-step procedure, utilizing the equality $E_c(y_i|\mathbf{x}_i) = E_s(r_i y_i|\mathbf{x}_i)$ with $r_i = (w_i-1)/E_s[(w_i-1)|\mathbf{x}_i]$ (follows from Equation 2.9). First, estimate $E_s(w_i|\mathbf{x}_i)$ and hence r_i by regressing w_i against $\mathbf{x}_i$ using the sample data. Let $\hat{r}_i = (w_i-1)/[\hat{E}_s(w_i|\mathbf{x}_i)-1]$ and transform $y_i^* = \hat{r}_i y_i$. Second, study the relationship in the sample between y_i^* and $\mathbf{x}_i$ for identifying the form of $C_\beta(\mathbf{x}_i)$. See Pfeffermann and Sverchkov (1999, 2003a) for examples of estimating $E_s(w_i|\mathbf{x}_i)$. A similar procedure can be applied for identifying the variance function $v(\mathbf{x}_i)$, using the empirical residuals $\hat{\varepsilon}_i = y_i - \hat{E}_s(\hat{r}_i y_i|\mathbf{x}_i)$.

The function $C_\beta(\mathbf{x}_j)$ in (4.1) with the true vector parameter β satisfies for all $j \notin s$,

$$\begin{aligned}C_\beta(\mathbf{x}_j) &= \arg\min_{C_{\tilde{\beta}}(\mathbf{x}_j)} E_c\left(\frac{[y_j - C_{\tilde{\beta}}(\mathbf{x}_j)]^2}{v(\mathbf{x}_j)}\Big|\mathbf{x}_j\right) \\ &= \arg\min_{C_{\tilde{\beta}}(\mathbf{x}_j)} E_s\left(r_i\frac{[y_j - C_{\tilde{\beta}}(\mathbf{x}_j)]^2}{v(\mathbf{x}_j)}\right). \quad (4.2)\end{aligned}$$

(The second equality follows from (2.9)). Hence, by substituting the sample expectation outside the curved brackets by the sample mean (a straightforward application

of the method of moments) and estimating r_i by $\hat{r}_i$ (see Remark 3), the vector β can be estimated as,

$$\hat{\beta}_1 = \arg\min_{\tilde{\beta}} \sum_{i\in s}\left(\hat{r}_i \frac{[y_i - C_{\tilde{\beta}}(\mathbf{x}_i)]^2}{v(\mathbf{x}_i)}\right). \quad (4.3)$$

The predictor of the population total takes then the form,

$$\hat{Y}_1 = \sum_{i\in s} y_i + \sum_{j\notin s} C_{\hat{\beta}_1}(\mathbf{x}_j). \quad (4.4)$$

Alternatively, it follows from (4.1) that,

$$\begin{aligned} &E_c\left(\frac{[y_j - C_\beta(\mathbf{x}_j)]^2}{v(\mathbf{x}_j)}\Big|\mathbf{x}_j\right) \\ &= E_c\left(\frac{[y_j - C_\beta(\mathbf{x}_j)]^2}{v(\mathbf{x}_j)}\right) \\ &= E_s\left(\left[\frac{w_j - 1}{E_s(w_j) - 1}\right]\frac{[y_j - C_\beta(\mathbf{x}_j)]^2}{v(\mathbf{x}_j)}\right) \end{aligned} \quad (4.5)$$

where the right hand side expectation is with respect to the joint distribution of $(y_i, \mathbf{x}_j)$. Thus, β can be estimated as,

$$\hat{\beta}_2 = \arg\min_{\beta} \sum_{i\in s}(w_i - 1)\frac{[y_i - C_{\tilde{\beta}}(\mathbf{x}_i)]^2}{v(\mathbf{x}_i)} \quad (4.6)$$

since $E_s(w_i) = constant$. The predictor of Y with β estimated by $\hat{\beta}_2$ is therefore,

$$\hat{Y}_2 = \sum_{i\in s} y_i + \sum_{j\notin s} C_{\hat{\beta}_2}(\mathbf{x}_j). \quad (4.7)$$

Remark 4. A notable advantage of the use of the predictor $\hat{Y}_2$ over the use of the predictor $\hat{Y}_1$ is that it does not require the identification and estimation of the expectation $w(\mathbf{x}) = E_s(w \mid \mathbf{x})$. On the other hand, in situations where this expectation can be estimated properly, the predictor $\hat{Y}_1$ is likely to be more accurate since the weights $r_i = (w_i - 1)/[E_s(w_i \mid \mathbf{x}_i) - 1]$ will often be less variable than the weights $(w_i - 1)$. This is because the weights r_i only account for the net effect of the sampling process on the target conditional distribution $f_c(y_i \mid \mathbf{x}_i)$, whereas the weights $(w_i - 1)$ account for the effect of the sampling process on the joint distribution $f_c(y_i, \mathbf{x}_i)$. In particular, when w_i is a deterministic function of $\mathbf{x}_i$ such that $w_i = w(\mathbf{x}_i)$, the sampling process is noninformative and $f_c(y_i \mid \mathbf{x}_i) = f_s(y_i \mid \mathbf{x}_i) = f_p(y_i \mid \mathbf{x}_i)$. In this case the estimator $\hat{\beta}_1$ (but not $\hat{\beta}_2$) coincides with the optimal generalized least square (GLS) estimator of β since $r_i = 1$ and the model (4.1) holds for the sample data. (For the data analysed in section 7, the empirical variance of the weights r_i is 1.36, whereas the empirical variance of the weights w_i is 2.66). In contrast to this, when the sampling weights w_i are independent of $\mathbf{x}_i$, the estimates $\hat{\beta}_1$ and $\hat{\beta}_2$, and hence the predictors $\hat{Y}_1$ and $\hat{Y}_2$ are equal since $w(\mathbf{x}_i) = constant$.

An interesting special case of the predictor $\hat{Y}_2$ arises when the working model postulated for the sample-complement is linear with an intercept term and constant variance. Let $\mathbf{x}_i' = (1, \tilde{\mathbf{x}}_i')$. As easily verified, the estimator in this case takes the form,

$$\hat{Y}_{2,\,\mathrm{Reg}} = \sum_{i\in s} y_i + \hat{Y}_C + \tilde{B}_c'\left[\tilde{X}(c) - \hat{X}_C\right] \quad (4.8)$$

where $\tilde{X}(c) = \sum_{i\notin s}\tilde{\mathbf{x}}_i$, $(\hat{Y}_C, \hat{X}_C) = [(N-n)/\sum_{i\in s}(w_i - 1)]$ $[\sum_{i\in s}(w_i - 1)(y_i, \tilde{\mathbf{x}}_i)]$ and $\tilde{B}_c$ is the probability weighted estimator of the vector coefficient of $\tilde{\mathbf{x}}_i$ but with the weights $(w_i - 1)$ instead of w_i.

Remark 5. The predictor $\hat{Y}_{2,\,\mathrm{Reg}}$ can be obtained as a special case of the Cosmetic predictors proposed by Brewer (1999). It should be emphasized, however, that the development of the cosmetic predictors and the derivation of their MSE assumes explicitly *noninformative* sampling.

An important property of $\hat{Y}_{2,\,\mathrm{Reg}}$ is that under general conditions it is *design consistent* for Y, irrespective of the true sample-complement model (see Lemma 1 below). Many analysts view 'design consistency' as an essential requirement from any predictor; see the discussion in Hansen, Madow and Tepping (1983) and Särndal (1980). The following Lemma 1 defines conditions under which the more general predictor $\hat{Y}_2$ of (4.7) is design consistent for Y.

Lemma 1. The predictor $\hat{Y}_2$ is design consistent for Y if the working model used for the computation of $\hat{\beta}_2$ satisfies the conditions, *i*- $C_\beta(\mathbf{x})$ has an intercept term, *ii*- $C_\beta(\mathbf{x})$ is differentiable with respect to β in the neighborhood of $\hat{\beta}_2$ and *iii*- $v(\mathbf{x}) = \text{constant}$.

Proof: By (4.6) and condition *iii*, $\hat{\beta}_2 = \arg\min_{\tilde{\beta}}$ $\sum_{i\in s}(w_i - 1)[y_i - C_{\tilde{\beta}}(\mathbf{x}_i)]^2$ and by condition *i*, $C_\beta(\mathbf{x}) =$ $\beta_0 + C_{\beta_1,\ldots\beta_p}(\tilde{\mathbf{x}})$, so that by condition *ii*, $\partial/\partial\beta_0$ $\{\sum_{i\in s}(w_i - 1)[y_i - C_{\tilde{\beta}}(\mathbf{x}_i)]^2\}_{\beta=\hat{\beta}_2} = 0$, which implies $\sum_{i\in s}(w_i - 1)[y_i - C_{\hat{\beta}_2}(\mathbf{x}_i)] = 0$ or,

$$\sum_{i\in s} w_i y_i = \sum_{i\in s} y_i + \sum_{i\in s} w_i C_{\hat{\beta}_2}(\mathbf{x}_i) - \sum_{i\in s} C_{\hat{\beta}_2}(\mathbf{x}_i). \quad (4.9)$$

The proof is completed by noting that under mild regularity conditions $\sum_{i\in s} w_i y_i$ is design consistent for Y, and $\sum_{i\in s} w_i C_{\hat{\beta}_2}(\mathbf{x}_i)$ is design consistent for $\sum_{j=1}^{N} C_{\hat{\beta}_2}(\mathbf{x}_i)$. Thus, the right hand side of (4.9) converges in probability to $\hat{Y}_2$ while the left hand side converges in probability to Y.

It is important to emphasize again that the Lemma does not assume that the working model is the correct sample-complement model.

The use of the predictors $\hat{Y}_1$ and $\hat{Y}_2$ requires a specification of the sample-complement model. Next we develop another predictor that only requires the identification and estimation of the sample model. The approach leading to this predictor is a sample-complement analogue of the 'bias correction method' proposed by Chambers *et al.* (2003). The proposed predictor is based on the following relationship,

$$\begin{aligned}\sum_{j\in s} E_c(y_j|\mathbf{x}_j) &= \sum_{j\in s} E_s(y_j|\mathbf{x}_j) \\ &+ (N-n)\left\{\frac{1}{N-n}\sum_{j\in s} E_c\left\{\left[y_j - E_s(y_j|\mathbf{x}_j)\right]\middle|\mathbf{x}_j\right\}\right\} \\ &\equiv \sum_{j\in s} E_s(y_j|\mathbf{x}_j) \\ &+ (N-n)\left\{\frac{1}{N-n}\sum_{j\in s} E_c\left[y_j - E_s(y_j|\mathbf{x}_j)\right]\right\} \qquad (4.10)\end{aligned}$$

where in the second row we replaced the sample-complement average of the conditional expectations $E_c(y_j|\mathbf{x}_j)$ by its expectation over the sample-complement distribution of the **x**-values (n denotes the sample size). By (2.9),

$$\begin{aligned}&E_c[y_j - E_s(y_j|\mathbf{x}_j)] \\ &= E_s\left\{\frac{w_j - 1}{[E_s(w_j) - 1]}\left[y_j - E_s(y_j|\mathbf{x}_j)\right]\right\} \qquad (4.11)\end{aligned}$$

implying that the sample-complement mean in the second row of (4.10) can be estimated as $\hat{M}_c = 1/n \sum_{i\in s}\{[(w_i - 1)/(\bar{w}_s - 1)][y_i - \hat{E}_s(y_i|\mathbf{x}_i)]\}$, where $\bar{w}_s = \sum_{i\in s} w_i / n$. The proposed predictor therefore takes the form,

$$\hat{Y}_3 = \sum_{i\in s} y_i + \sum_{j\notin s} \hat{E}_s(y_j|\mathbf{x}_j) + (N-n)\hat{M}_c \qquad (4.12)$$

with $\hat{E}_s(y_j|\mathbf{x}_j)$ estimated from the sample data. The use of $\hat{Y}_3$ only requires the identification and estimation of the sample regression $E_s(y_j|\mathbf{x}_j)$, which can be carried out using conventional regression techniques. Moreover, under mild conditions $\hat{Y}_3$ is *design consistent* for Y even if the expectation $E_s(y_j|\mathbf{x}_j)$ is misspecified. This property follows from the fact that $\sum_{j\notin s}\hat{E}_s(y_j|\mathbf{x}_j)$ is design consistent for $\sum_{j\notin s} E_s(y_j|\mathbf{x}_j)$ and $(N-n)\hat{M}_c$ is design consistent for $M_c = \sum_{j\notin s}[y_j - E_s(y_j|\mathbf{x}_j)]$.

Remark 6. If the model fitted to the sample data is linear regression with an intercept and constant residual variance, the difference between the predictor $\hat{Y}_{2,\mathrm{Reg}}$ defined by (4.8) and the predictor $\hat{Y}_3$ is that $\hat{Y}_{2,\mathrm{Reg}}$ uses a consistent estimator for the regression coefficients defining the linear approximation to the model holding for the sample-complement, whereas in $\hat{Y}_3$ the regression coefficients are estimated by ordinary least squares (OLS), thus estimating the linear approximation to the sample model.

Finally, rather than only predicting the sample-complement values as with the previous predictors, one could instead predict all the population values by their estimated expectations under the population model. Assuming that the latter model is linear regression with an intercept term and constant residual variance, application of (2.5b) yields,

$$\begin{aligned}\beta &= \arg\min_{\tilde{\beta}} E_p(y_k - \mathbf{x}_k'\tilde{\beta})^2 \\ &= \arg\min_{\tilde{\beta}} \frac{E_s[w_k(y_k - \mathbf{x}_k'\tilde{\beta})^2]}{E_s(w_k)}. \qquad (4.13)\end{aligned}$$

Estimating the sample expectation in the numerator of (4.13) by the corresponding sample mean (application of the method of moments) and minimizing the sample mean with respect to β yields the familiar probability weighted estimator $\hat{B}_{pw} = (X_{[s]}' W_s X_{[s]})^{-1}(X_{[s]}' W_s Y_s)$, where $(X_{[s]}, Y_s) = \{[\mathbf{x}_1 \ldots \mathbf{x}_n]', (y_1 \ldots y_n)'\}$ and $W_s = \mathrm{Diag}[w_1 \ldots w_n]$. Let $\mathbf{x}_i' = (1, \tilde{\mathbf{x}}_i')$. Estimating $\hat{E}_p(y_k|\mathbf{x}_k) = \mathbf{x}_k'\hat{B}_{pw} = \hat{B}_0 + \tilde{\mathbf{x}}_k'\tilde{B}_{pw}$ and summing over all the population values yields the familiar generalized regression (GREG) estimator (Särndal 1980),

$$\hat{Y}_{\mathrm{GREG}} = N\frac{\sum_{i\in s} w_i y_i}{\sum_{i\in s} w_i} + \tilde{B}_{pw}'\left[\tilde{X}(p) - N\frac{\sum_{i\in s} w_i \mathbf{x}_i}{\sum_{i\in s} w_i}\right];$$

$$\tilde{X}(p) = \sum_{k=1}^{N} \tilde{\mathbf{x}}_k. \qquad (4.14)$$

Remark 7. By considering the estimation of Y as a prediction problem, the use of the predictor $\hat{Y}_{2,\mathrm{Reg}}$ in (4.8) requires the prediction of $(N - n)$ values whereas the use of the GREG requires the prediction of N values. Hence, in situations where both the sample-complement model and the population model can be approximated fairly well by linear regression models with intercept terms (but possibly with different vectors of coefficients for the two models), one expects that for sufficiently large sampling fractions n/N the predictor $\hat{Y}_{2,\mathrm{Reg}}$ will be superior (see the empirical results in section 7).

5. EXAMPLES

5.1 Prediction with No Concomitant Variables

Let $\mathbf{x}_i = 1$ for all i. By (3.2),

$$\hat{Y} = \sum_{i\in s} y_i + \sum_{j\notin s} \hat{E}_c(y_j) = \sum_{i\in s} y_i + (N-n)\hat{E}_s\left(\frac{w_j - 1}{\hat{E}_s(w_j) - 1} y_j\right). \quad (5.1)$$

Estimating the two sample expectations in the right hand side of (5.1) by the respective sample means yields the estimator,

$$\hat{Y}_{EI} = \sum_{i\in s} y_i + (N-n)\frac{1}{n}\sum_{i\in s}\frac{w_i - 1}{\bar{w}_s - 1} y_i = \sum_{i\in s} y_i + \frac{(N-n)}{\sum_{i\in s}(w_i - 1)}\sum_{i\in s}(w_i - 1)y_i. \quad (5.2)$$

In (5.2), $\sum_{i\in s}(w_i - 1)y_i$ is a 'Horvitz-Thompson estimator' of $\sum_{j\notin s} y_j$. The multiplier $(N-n)/\sum_{i\in s}(w_i - 1)$ is a 'Hajek type correction' for controlling the variability of the sampling weights. Notice that $\hat{Y}_{EI}$ is a special case of the predictor $\hat{Y}_{2,\mathrm{Reg}}$ defined in (4.8), obtained by setting $\mathbf{x}_i = 1$ for all i. It is also a special case of the predictor $\hat{Y}_3$ if one estimates $\hat{E}_s(y_j) = \bar{y} = \sum_{i\in s} y_i/n$. For sampling designs such that $\sum_{i\in s} w_i = N$ for all s, or if one estimates $\hat{E}_s(w_i) = N/n$, the predictor $\hat{Y}_{EI}$ reduces to the familiar Horvitz-Thompson estimator of the population total, $\hat{Y}_{\mathrm{H-T}} = \sum_{i\in s} w_i y_i$.

As with the GREG estimator considered in section 4, rather than predicting the sample-complement total $Y_c = \sum_{j\notin s} y_j$ and using the predictor $\hat{Y}_{EI}$, one could predict all the population y-values by estimating their expectations under the population model. By (2.5b), $E_p(y_i) = E_s(w_i y_i)/E_s(w_i)$. Estimating the two sample expectations by the corresponding sample means yields the familiar Hajek estimator,

$$\hat{Y}_{\mathrm{Hajek}} = \sum_{k=1}^{N}\hat{E}_p(y_k) = N\hat{E}_s\left(\frac{w_i y_i}{\hat{E}_s(w_i)}\right) = \frac{N}{\sum_{i\in s} w_i}\sum_{i\in s} w_i y_i. \quad (5.3)$$

Here again, we anticipate $\hat{Y}_{EI}$ to be more precise than $\hat{Y}_{\mathrm{Hajek}}$ as the sampling fraction increases (see also the empirical results in section 7). Note that $\hat{Y}_{EI}$ and $\hat{Y}_{\mathrm{Hajek}}$ are the same and coincide with the Horvitz-Thompson estimator for sampling designs satisfying $\sum_{i\in s} w_i = N$.

5.2 Optimal Prediction with Concomitant Variables, Comparison with Optimal Predictors Under Noninformative Sampling

Let the population model be,

$$y_i = H_\beta(\mathbf{x}_i) + \varepsilon_i,\ E_p(\varepsilon_i|\,\mathbf{x}_i) = 0,\ E_p(\varepsilon_i^2|\,\mathbf{x}_i) = v(\mathbf{x}_i),\ E_p(\varepsilon_i\varepsilon_j|\,\mathbf{x}_i, \mathbf{x}_j) = 0,\ i \neq j \quad (5.4)$$

and suppose that the sample inclusion probabilities can be modeled as,

$$\pi_i = K\times[\,y_i\ g(\mathbf{x}_i) + \delta_i\,],\ E_p(\delta_i|\,\mathbf{x}_i, y_i) = 0 \quad (5.5)$$

where $H_\beta(\mathbf{x})$, $v(\mathbf{x})$ and $g(\mathbf{x})$ are positive functions and K is a normalizing constant. (Below we consider the special case of 'regression through the origin'). This sampling scheme is considered for illustration only, although in section 2 we mention several practical situations where the sample selection probabilities depend directly on the y and $\mathbf{x}$-values. In particular, this is the case with the data set analysed in section 7. Under (5.4) and (5.5), $\pi(\mathbf{x}_i) = E_p(\pi_i|\,\mathbf{x}_i) = KH_\beta(\mathbf{x}_i)g(\mathbf{x}_i)$. Hence, by (2.9), (5.4) and (5.5),

$$E_c(y_j|\,\mathbf{x}_j) = E_p\left(\frac{1-\pi_j}{1-\pi(\mathbf{x}_j)} y_j|\,\mathbf{x}_j\right) = E_p\left(\frac{1-\pi(\mathbf{x}_j) - K\varepsilon_j g(\mathbf{x}_j) - K\delta_j}{1-\pi(\mathbf{x}_j)} y_j|\,\mathbf{x}_j\right) = E_p(y_j|\,\mathbf{x}_j) - \frac{Kg(\mathbf{x}_j)v(\mathbf{x}_j)}{1-\pi(\mathbf{x}_j)}. \quad (5.6)$$

The last expression in (5.6) shows that $E_c(y_j|\,\mathbf{x}_j) < E_p(y_j|\,\mathbf{x}_j) = H_\beta(\mathbf{x}_j)$, which is clear since for the inclusion probabilities defined by (5.5), the sample-complement tends to include the units with the smaller y-values for any given $\mathbf{x}$-values. Note, however, that as $n/N \to 0$, $K \to 0$ and $E_p(y_j|\,\mathbf{x}_j) - E_c(y_j|\,\mathbf{x}_j) \to 0$ (see Remark 2).

As a special case of (5.4), consider the case of a single auxiliary variable x and let $H_\beta(x) = x\beta$ and $v(x) = \sigma^2 x$ ('regression through the origin with variance proportional to x'). For *noninformative* sampling and known β, the optimal unbiased predictor of Y minimizing $E_p[(\hat{Y} - Y)^2|\,D_s]$ is in this case, $\hat{Y} = \sum_{i\in s} y_i + \beta\sum_{j\notin s} x_j$. In the practical case of unknown β, the optimal unbiased predictor of Y is the familiar Ratio estimator $\hat{Y}_R = N\bar{y}(\bar{X}/\bar{x})$ with $\bar{y}$ denoting the sample mean of Y and $(\bar{x}, \bar{X})$ denoting the sample and population means of x (Brewer 1963, Royall 1970).

Now let $g(x) = 1$ in (5.5) for all x, so that $\pi_i = n(y_i + \delta_i)/\sum_{j=1}^{N}(y_j + \delta_j)$. For sufficiently large N, we can approximate $\pi_i \approx n(y_i + \delta_i)/(N\beta\bar{X})$, implying that $\pi(x_i) = E_p(\pi_i|x_i) \approx nx_i/(N\bar{X})$. By (5.6), $E_c(y_j|\,x_j) = x_j\beta - \sigma^2 x_j/[\,\beta(f^{-1}\bar{X} - x_j)]$ where $f = n/N$ is the sampling fraction, so that for known β and σ^2 the optimal predictor of Y is,

$$\hat{Y}_{E,\mathrm{Reg}} = \sum_{i\in s} y_i + \beta\sum_{j\notin s} x_j - \frac{\sigma^2}{\beta}\sum_{j\notin s}\frac{x_j}{f^{-1}\bar{X} - x_j}. \quad (5.7)$$

Lemma 2: Let the population model be defined by (5.4) with $H_\beta(x) = x\beta$ and $v(x) = \sigma^2 x$. Assume also

$E_p(\varepsilon_i^3 \mid x_i) = 0$. Suppose that the sample units are selected independently with probabilities defined as in (5.5), with $g(x) = 1$. Then,

$$\mathrm{MSE}_p\left(\hat{Y}_{E,\,\mathrm{Reg}} \mid D_s\right) = \sigma^2 \sum_{j \notin s} x_j - (\sigma^2/\beta)^2 \sum_{j \notin s} [x_j^2/(f^{-1}\overline{X} - x_j)^2]. \quad (5.8)$$

Proof: By the independence of the population values and of the sample selections,

$$\begin{aligned}\mathrm{MSE}_p(\hat{Y}_{E,\,\mathrm{Reg}} \mid D_s) &= E_p[(\hat{Y}_{E,\,\mathrm{Reg}} - Y)^2 \mid D_s] \\ &= \sum_{j \notin s} E_c\{[y_j - E_c(y_j \mid x_j)]^2 \mid x_j\}.\end{aligned}$$

By (5.6), $[y_j - E_c(y_j \mid x_j)]^2 = \{\varepsilon_j + x_j^* / [1 - \pi(x_j)]\}^2$ where $x_j^* = K\sigma^2 x_j$, $K = n/\beta N\overline{X}$ and $\pi(x_j) = E_p(\pi_j \mid x_j) \approx nx_j/(N\overline{X})$. Hence,

$$\begin{aligned}E_c\{[y_j - E_c(y_j \mid x_j)]^2 \mid x_j\} &= E_c(\varepsilon_j^2 \mid x_j) + 2x_j^*/(1-\pi(x_j))E_c(\varepsilon_j \mid x_j) \\ &\quad + [x_j^*/(1-\pi(x_j))]^2.\end{aligned}$$

Now,

$$\begin{aligned}E_c(\varepsilon_j^2 \mid x_j) &= E_p[1 - \pi_j/(1-\pi(x_j))\varepsilon_j^2 \mid x_j] \\ &= E_p[1 - \pi(x_j) - K\varepsilon_j - K\delta_j/(1-\pi(x_j))\,\varepsilon_j^2 \mid x_j] \\ &= E_p(\varepsilon_j^2 \mid x_j) = \sigma^2 x_j\end{aligned}$$

and

$$\begin{aligned}E_c(\varepsilon_j \mid x_j) &= E_p[1 - \pi(x_j) - K\varepsilon_j - K\delta_j/(1-\pi(x_j))\,\varepsilon_j \mid x_j] \\ &= -x_j^*/(1-\pi(x_j)).\end{aligned}$$

It follows therefore that $\mathrm{MSE}_p(\hat{Y}_{E,\,\mathrm{Reg}} \mid D_s) = \sigma^2 \sum_{j \notin s} x_j - \sum_{j \notin s} [x_j^*/(1-\pi(x_j))]^2$. Q.E.D.

Remark 8: For *noninformative* sampling and with known β, the prediction MSE of the optimal predictor $\hat{Y} = \sum_{i \in s} y_i + \beta \sum_{j \notin s} x_j$ is, $E_p[(\hat{Y} - Y)^2 \mid D_s] = \sigma^2 \sum_{j \notin s} x_j$. This MSE is larger than the MSE obtained under the informative sampling scheme defined by the Lemma, which is obvious since the latter scheme tends to sample the units with the larger y-values and hence also with the larger x-values and the larger standard deviations.

6. MEAN SQUARE ERROR ESTIMATION

Estimating $\mathrm{MSE}(\hat{Y} \mid D_s) = E_p[(\hat{Y} - Y)^2 \mid D_s]$ for the predictors $\hat{Y}$ considered in section 4 requires strict model assumptions that could be hard to validate. This is largely due to the conditioning on the design information D_s. In order to deal with this problem, we propose to estimate instead the unconditional MSE, $\mathrm{MSE}(\hat{Y}) = E[(\hat{Y} - Y)^2] = E_{D_s}\{E_p[(\hat{Y} - Y)^2 \mid D_s]\}$, where $E_{D_s} = E_D E_s$ defines the expectation over the sample distribution (given the selected sample) and over all possible sample selections. Notice that $E_p[(\hat{Y} - Y)^2 \mid D_s]$ can be viewed as a random variable $u(D_s)$, so that $\mathrm{MSE}(\hat{Y}) = E_{D_s}[u(D_s)]$ defines its 'best predictor' with respect to the mean square loss function under the distribution f_{D_s} over which the expectation E_{D_s} is taken. By changing the order of the expectations, the unconditional MSE can be expressed as,

$$\begin{aligned}\mathrm{MSE}(\hat{Y}) &= E_s E_p E_D[(\hat{Y} - Y)^2 \mid y] \\ &= E_p E_D[(\hat{Y} - Y)^2 \mid y] \qquad (6.1)\end{aligned}$$

where $y = \{y_i;\ i \in U\}$. Estimating the unconditional MSE of any of the predictors $\hat{Y}$ can be carried out therefore by estimating its randomization MSE, see Pfeffermann (1993) for further discussion. Estimation of the randomization MSE of the various predictors has the additional advantage of allowing their use under the design based approach.

Estimation of randomization variances of design based estimators is considered extensively in the literature and many diverse methods are in routine use. However, in view of the complicated structure of some of the predictors considered in this study and in order not to restrict to particular sampling schemes, we propose below the use of a two-step procedure that combines an inverse sampling process (Step 1) and what can be viewed as a bootstrap resampling algorithm (Step 2). A notable advantage of this procedure is that it is general and applies 'equally' to all the predictors. Also, unlike other variance estimation methods in common use, it does not require knowledge of the pair wise joint selection probabilities $\pi_{ij} = \Pr(i, j \in s)$. As discussed later, a valid application of the first step requires sufficiently large samples. The two steps of the proposed procedure are as follows:

Step 1- Generate a single 'pseudo population' by selecting *with replacement* N units from the original sample with probabilities proportional to $w_i = 1/\pi_i$, where N is the population size. The justification for this step is given below, see also Remark 10. Denote by Y_{pp} the sum of the y-values in the pseudo population.

Step 2- Select independently a large number B of bootstrap samples from the pseudo population generated in Step 1, using the same sampling scheme as used for the selection of the original sample, and re-estimate the population total.

Let $\hat{Y}$ represent any of the predictors and denote the predictor obtained for bootstrap sample b by $\hat{Y}_{pp}^b$. Estimate,

$$\hat{E}_D(\hat{Y} - Y)^2 = \frac{1}{B}\sum_{b=1}^{B}(\hat{Y}_{pp}^b - Y_{pp})^2. \qquad (6.2)$$

The performance of the estimator (6.2) in estimating the randomization MSE depends obviously on the 'closeness' of the pseudo population generated in Step 1 to the actual population from which the original sample was drawn. The closeness of the two populations can be verified in part by noting that the marginal distribution of $y_i \mid \mathbf{x}_i$ in the pseudo population is the same as in the original population. To see this, note that the pseudo population generated in Step 1 is a 'sample with replacement' from the original sample with selection probabilities Cw_i on each draw, where $C = 1/\sum_{i=1}^{n} w_i$. Denoting by $f_{pp}(y_i \mid \mathbf{x}_i)$ the marginal pseudo population distribution we find using (2.2) and (2.5a),

$$f_{pp}(y_i|\mathbf{x}_i) = \frac{E_s(Cw_i|y_i,\mathbf{x}_i) f_s(y_i|\mathbf{x}_i)}{E_s(Cw_i|\mathbf{x}_i)}$$

$$= \frac{E_p(\pi_i|\mathbf{x}_i) f_s(y_i|\mathbf{x}_i)}{E_p(\pi_i|y_i,\mathbf{x}_i)} = f_p(y_i|\mathbf{x}_i). \quad (6.3)$$

Remark 9. Equation (6.3) only refers to the marginal distribution of $y_i \mid \mathbf{x}_i$. Like with the standard bootstrap method, a successful application of the proposed procedure requires that the original sample size is sufficiently large and that the sample measurements are approximately independent. Pfeffermann *et al.* (1998) establish conditions under which for independent population measurements the sample measurement are 'asymptotically independent' under commonly used sampling schemes with unequal selection probabilities.

Remark 10. Step 1 is similar and asymptotically equivalent to duplicating sample unit i w_i times. Notice, however, that the use of this duplication procedure does not yield pseudo populations of size N unless $\sum_{i=1}^{n} w_i = N$. It is also not clear how to establish the relationship (6.3) when using this procedure.

7. EMPIRICAL ILLUSTRATIONS

7.1 Description of Empirical Study

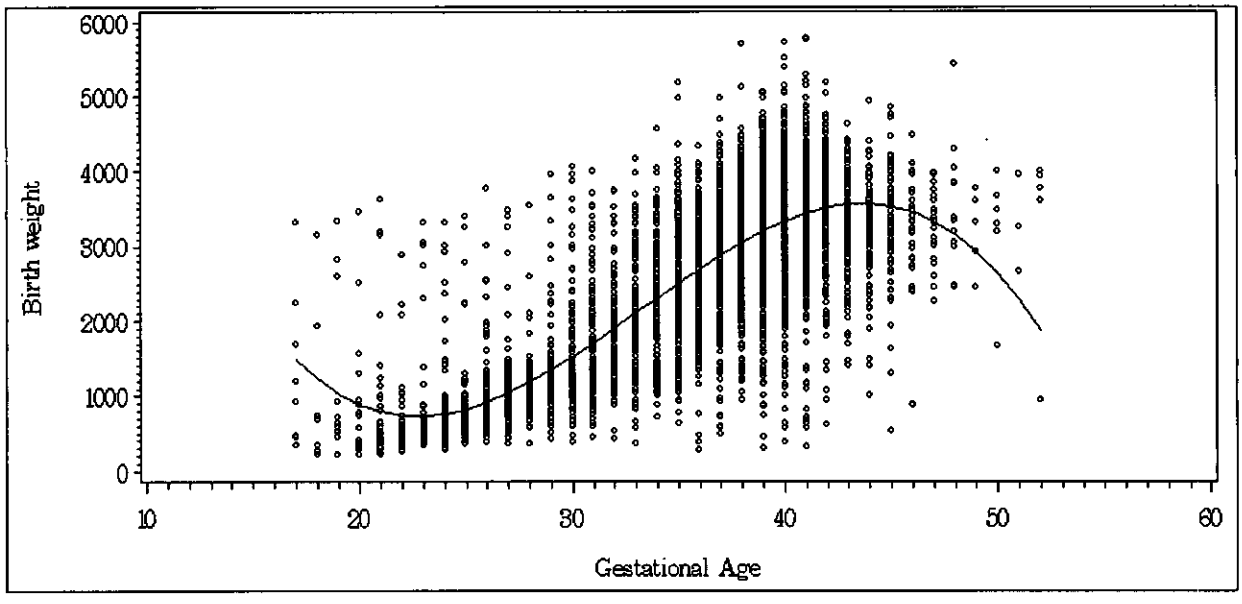
In order to illustrate the performance of the predictors and the associated MSE estimates discussed in previous sections we use a real data set, collected as part of the 1988 U.S. National Maternal and Infant Health Survey. The survey uses a disproportionate stratified random sample of vital records with the strata defined by *mother's race* and *child's birth weight;* see Korn and Graubard (1995) for details. For the empirical study in this section we considered the sample data as 'population' and selected independently 1,000 samples with probabilities proportional to the inverse of the original sampling weights, using a systematic PPS sampling scheme. The list of 'population units' was randomly ordered before every sample selection. For each sample we predicted the population total of *birth weight* (measured in *grams*, divided by 10,000 in the present study), using *gestational age* as the auxiliary variable (measured in *weeks*). The sample inclusion probabilities depend therefore on the values of the study variable that defines the original strata. Notice that although the original sample was supposedly a stratified random sample, the sampling weights actually vary within the strata, which is why we used systematic PPS sampling for the simulation study. We considered three different sample sizes, $n = 232$, 1,145, 2,429. The 'population' (original sample) size is $N = 9{,}948$. (For $n = 232$, $0.002 < \pi_i = \Pr(i \in s) < 0.15$. For $n = 1{,}145$, $0.01 < \pi_i < 0.73$. For $n = 2{,}429$, $0.03 < \pi_i < 0.99$ with mean $\bar{\pi} = 0.26$ and standard deviation $Std(\pi_i) = 0.29$. In the latter case some of the units were drawn almost with certainty).

Some of the predictors considered for this study (see below) require the specification of either the sample model or the sample-complement model. We assumed for both models the third order polynomial regression,

$$y_k = \beta_0 + \beta_1 x_k + \beta_2 x_k^2 + \beta_3 x_k^3 + \varepsilon_k \quad (7.1)$$

with independent residuals and constant variance. This model was found by Pfeffermann and Sverchkov (1999) to give a good fit to the 'population' (original sample) data with $R^2 = 0.61$ (see Figure 1), and it was found also to fit fairly well the sample data (with different coefficients) for several samples selected from this 'population'. Notice, on the other hand, that with this strongly informative sampling scheme, it is unlikely that the sample model, the population model and the sample-complement model are all from the same family even if with different parameters. The present study enables therefore studying the performance of the various predictors when some or all of the three models are misspecified. This important robustness question is further examined by fitting simple regression models instead of the third order polynomial regressions that is, by omitting the second and third powers of the auxiliary variable. The only exception is the model dependent predictor $\hat{Y}_1$ (Equation 4.4) where no coherent estimator for the expectation $E_s(w_j \mid x_j)$ could be found when restricting to simple regression. (The method considered in Pfeffermann and Sverchkov (1999) for the estimation of this expectation assumes normality of the population model residuals. This is a valid assumption when fitting the third order polynomial regression model but is clearly violated when dropping the second and third powers of the auxiliary variable).

U.S. National Maternal and Infant Health Survey,1988.



Model Fitted: $y_i = 17886 - 1827.7x_i + 61.2x_i^2 - 0.61x_i^3 + \varepsilon_i$

$\text{Var}(\varepsilon_i) = 603.2,\ R^2 = 0.61$

Figure 1. Scatterplot of Birth Weight against Gestational Age in 'Population' (original Sample), and Predicted Values Under 3rd Order Polynomial Regression.

The predictors considered for this study divide therefore into three groups. The first group consists of predictors that only use the sample y-values and the sampling weights. Included in this group are the Horvitz-Thompson estimator $\hat{Y}_{H-T} = \sum_{i\in s} w_i y_i$, the predictor $\hat{Y}_{EI}$ defined by (5.2) and Hajek's estimator $\hat{Y}_{Hajek}$ defined by (5.3). The second group consists of predictors that use the working model defined by (7.1). Included in this group are the two regression predictors $\hat{Y}_1$ and $\hat{Y}_{2,Reg}$ defined by (4.4) and (4.8) respectively, the bias corrected predictor $\hat{Y}_3$ defined by (4.12) and the GREG estimator defined by (4.14). The third group contains the same predictors as the second group (except for $\hat{Y}_1$, see above), but based on the simple regression model (only the first power of x).

The MSEs of all the predictors considered in this study have been estimated by use of the two-step procedure described in section 6. However, because of computing time limitations, the MSE estimators were only computed for a random selection of 200 out of the 1,000 samples and are based on only 200 bootstrap samples from each pseudo population. For assessing the performance of the MSE estimators we computed the corresponding empirical MSEs based on the 1,000 samples selected from the study population. Thus, the 'true' MSE of a generic predictor $\hat{Y}$ was computed as,

$$\text{MSE}(\hat{Y}) = \frac{1}{1{,}000}\sum_{r=1}^{1{,}000}(\hat{Y}_{(r)} - Y)^2 \qquad (7.2)$$

where $\hat{Y}_{(r)}$ denotes the predictor computed from the r^{th} sample. Notice that since the population values are fixed, the MSE in (7.2) is the randomization MSE over all possible sample selections, which is what the estimator (6.2) is intended to estimate.

7.2 Results of Empirical Study

The main results of this study are exhibited in Tables 1.1 – 1.3 (one table for each sample size). The third column of each table shows for every predictor $\hat{Y}$ the empirical bias, $[(\sum_{r=1}^{R}\hat{Y}_{(r)}/R) - Y]$, and the standard deviation (*Std*) of the empirical bias, computed as $[\sum_{r=1}^{R}(\hat{Y}_{(r)} - \overline{Y}_R)^2/R^2]^{1/2}$; $\overline{Y}_R = \sum_{r=1}^{R}\hat{Y}_{(r)}/R$, $R = 1{,}000$. The next two columns show respectively the 'true' (empirical) RMSE (square root of Equation 7.2), and the square root of the mean of the corresponding Bootstrap estimators defined by (6.2).

The main conclusions from Tables 1.1 – 1.3 are as follows:

1- All the predictors considered for this study are virtually design unbiased with all three sample sizes, irrespective of the underlying working model. The predictor $\hat{Y}_1$ has a statistically significant bias when tested by use of the conventional t-statistic but the actual bias is negligible when compared to the true population total. (The predictor $\hat{Y}_1$ is the only predictor considered in this study that is not design consistent).

The next three comments refer to the RMSE of the various predictors.

2- The predictors in Groups 2 and 3 that use the auxiliary values perform much better than the predictors in Group 1, particularly for the smaller sample sizes. The predictors in Group 2 that employ the 3rd order polynomial regression model (7.1) perform better than the corresponding predictors in Group 3 that employ the simple regression model as the working model, but the differences diminish as the sample size increases.

3- An important result emerging from this study is that the predictors $\hat{Y}_{2,\text{Reg}}$ and $\hat{Y}_{EI}$ (and also $\hat{Y}_3$ for the larger sample sizes), that only predict the y-values for units outside the sample indeed perform better than the other predictors in their respective groups (see also below). As surmised in Remark 7, this holds particularly with the larger sample sizes. Notice that the differences between $\hat{Y}_{2,\text{Reg}}$ and the GREG estimator for n=1,145 and n=2,250 are smaller under the polynomial model (Group 2) than under the simple regression model (Group 3), which is explained by the tight relationship between the study variable and auxiliary variables under the polynomial model. The predictor $\hat{Y}_3$ is less stable than $\hat{Y}_{2,\text{Reg}}$ for $n = 232$ but for the other two sample sizes the two predictors perform similarly.

4- The predictor $\hat{Y}_{2,\text{Reg}}$ performs somewhat better than the model dependent predictor $\hat{Y}_1$ that employs the expectations $E(w_i \mid x_i)$ to adjust the sampling weights. We have no clear explanation for this result because as illustrated in Pfeffermann and Sverchkov (1999) using the same data, adjusting the sampling weights improves the estimation of the regression coefficients very significantly.

Next consider the MSE estimators.

5- The MSE estimators developed in section 6 perform very well for all the predictors and with all the sample sizes. For the sample size $n = 232$ there is a systematic under-estimation of the RMSE by up to 3%, which is explained by the fact that the pseudo population is in this case less variable than the actual study population (see Remark 9). The MSE estimators are almost unbiased for the other sample sizes with the largest difference between the estimated and true RMSE being again in the magnitude of 3%.

Another way of assessing the bias of the various predictors and their MSE estimation is by studying the coverage properties of confidence intervals defined by these predictors. Tables 2.1 – 2.3 compare the empirical percentage coverage of the standard confidence intervals $\hat{Y} \pm Z_{1-\alpha/2}\sqrt{\hat{\text{MSE}}}$ with the corresponding nominal percentages for selected values of α (one table for each sample size). The empirical percentages are somewhat erratic with $n = 232$ sample units but they stabilize as the sample size increases, particularly with the use of the predictors in the second and third group. The empirical percentages are close to the nominal percentages with all the predictors when n = 2,250.

Table 1.1
Bias, RMSE and Square Root of Mean of MSE Estimators, n = 232

Group	Predictor	Bias (Std)	RMSE	$\sqrt{\hat{\text{MSE}}}$
1 No x-values	$\hat{Y}_{\text{H-T}}$	-4.5 (11.6)	365.1	355.0
	$\hat{Y}_{EI}$	1.5 (2.9)	91.1	89.8
	$\hat{Y}_{\text{Hajek}}$	1.7 (2.9)	93.0	91.6
2 3rd order polynomial regression	$\hat{Y}_1$	4.4 (2.0)	64.0	63.0
	$\hat{Y}_{2,\text{Reg}}$	3.5 (2.0)	63.4	62.4
	$\hat{Y}_3$	-0.3 (2.1)	65.4	65.0
	$\hat{Y}_{\text{GREG}}$	3.4 (2.1)	63.6	62.6
3 Simple Regression	$\hat{Y}_{2,\text{Reg}}$	-2.3 (2.2)	68.0	66.2
	$\hat{Y}_3$	-0.3 (2.2)	68.6	67.4
	$\hat{Y}_{\text{GREG}}$	-2.3 (2.2)	68.3	66.5

True 'population' total= 2710.7

Table 1.2
Bias, RMSE and Square Root of Mean of MSE Estimators, n = 1,145

Group	Predictor	Bias (Std)	RMSE	$\sqrt{\hat{\text{MSE}}}$
1 No x-values	$\hat{Y}_{\text{H-T}}$	-9.1 (5.0)	157.1	156.1
	$\hat{Y}_{EI}$	0.0 (1.1)	35.2	34.9
	$\hat{Y}_{\text{Hajek}}$	-0.1 (1.3)	39.5	39.3
2 3rd order polynomial regression	$\hat{Y}_{\text{I}}$	3.0 (0.9)	27.6	28.1
	$\hat{Y}_{2,\text{Reg}}$	2.0 (0.9)	27.4	27.3
	$\hat{Y}_{3}$	0.5 (0.9)	27.4	27.7
	$\hat{Y}_{\text{GREG}}$	1.7 (0.9)	27.8	27.8
3 Simple Regression	$\hat{Y}_{2,\text{Reg}}$	0.0 (1.0)	28.3	28.7
	$\hat{Y}_{3}$	0.1 (1.0)	28.2	28.9
	$\hat{Y}_{\text{GREG}}$	0.0 (2.0)	29.1	29.6

True 'population' total= 2710.7

Table 1.3
Bias, RMSE and Square Root of Mean of MSE Estimators, n=2,250

Group	Predictor	Bias (Std)	RMSE	$\sqrt{\hat{\text{MSE}}}$
1 No x-values	$\hat{Y}_{\text{H-T}}$	1.3 (2.7)	82.7	80.4
	$\hat{Y}_{EI}$	-0.2 (0.6)	18.5	18.8
	$\hat{Y}_{\text{Hajek}}$	0.1 (0.7)	23.5	23.8
2 3rd order polynomial regression	$\hat{Y}_{\text{I}}$	1.3 (0.5)	17.5	17.3
	$\hat{Y}_{2,\text{Reg}}$	0.6 (0.5)	16.9	16.3
	$\hat{Y}_{3}$	-0.3 (0.5)	17.1	16.5
	$\hat{Y}_{\text{GREG}}$	0.5 (0.5)	17.9	18.3
3 Simple Regression	$\hat{Y}_{2,\text{Reg}}$	-0.3 (0.5)	17.3	16.8
	$\hat{Y}_{3}$	-0.3 (0.5)	17.7	17.3
	$\hat{Y}_{\text{GREG}}$	-0.2 (0.6)	18.8	18.3

True 'population' total= 2710.7

Table 2.1
Nominal and Empirical Percentage Coverage of Confidence Intervals, n = 232

Group	Predictor	1.0	2.5	5.0	10.0	90.0	95.0	97.5	99.0
1 No x-values	$\hat{Y}_{\text{H-T}}$	2.5	3.5	5.5	10.0	90.0	97.0	99.0	99.5
	$\hat{Y}_{EI}$	0.5	2.0	4.0	8.0	88.5	91.5	95.5	98.0
	$\hat{Y}_{\text{Hajek}}$	0.5	2.0	4.0	8.0	88.5	91.5	95.5	98.0
2 3rd order polynomial regression	$\hat{Y}_{\text{I}}$	0.0	0.0	1.5	6.5	86.0	90.5	92.5	97.5
	$\hat{Y}_{2,\text{Reg}}$	0.0	0.0	2.0	7.0	85.0	90.5	93.5	98.0
	$\hat{Y}_{3}$	0.0	0.5	2.5	6.5	87.5	91.0	95.0	98.5
	$\hat{Y}_{\text{GREG}}$	0.0	0.0	2.0	7.0	85.0	90.5	93.5	98.0
3 Simple Regression	$\hat{Y}_{2,\text{Reg}}$	0.0	1.0	2.5	7.0	87.0	91.5	97.5	98.0
	$\hat{Y}_{3}$	0.0	1.0	2.5	7.0	86.0	91.5	96.5	98.0
	$\hat{Y}_{\text{GREG}}$	0.0	1.0	2.5	7.0	86.5	91.5	97.0	98.0

Table 2.2
Nominal and Empirical Percentage Coverage of Confidence Intervals, $n = 1{,}145$

Group	Predictor	1	2.5	5.0	10.0	90.0	95.0	97.5	99.0
	$\hat{Y}_{\text{H-T}}$	4.0	7.0	9.0	13.5	95.5	98.0	98.5	99.5
1	$\hat{Y}_{EI}$	3.0	5.0	8.0	12.5	92.5	95.5	99.5	100.0
No x-values	$\hat{Y}_{\text{Hajek}}$	3.5	5.0	9.5	12.5	92.5	96.0	99.5	100.0
	$\hat{Y}_{1}$	0.5	2.0	5.0	7.5	86.5	93.5	96.0	97.0
2	$\hat{Y}_{2,\text{Reg}}$	0.5	3.0	6.0	9.0	86.5	94.5	96.5	97.0
3rd order polynomial	$\hat{Y}_{3}$	0.5	2.0	6.0	9.5	88.0	94.0	97.0	98.0
regression	$\hat{Y}_{\text{GREG}}$	0.5	3.0	5.0	9.0	86.5	94.0	96.5	98.0
	$\hat{Y}_{2,\text{Reg}}$	0.5	3.0	6.0	11.0	90.0	93.0	97.0	99.5
3	$\hat{Y}_{3}$	0.5	2.5	5.5	10.5	90.0	94.0	97.0	99.5
Simple Regression	$\hat{Y}_{\text{GREG}}$	1.0	3.0	6.0	11.0	90.5	94.0	97.5	99.0

Table 2.3
Nominal and Empirical Percentage Coverage of Confidence Intervals, $n = 2{,}250$

Group	Predictor	1.0	2.5	5.0	10.0	90.0	95.0	97.5	99.0
	$\hat{Y}_{\text{H-T}}$	0.5	1.0	5.5	11.0	95.0	97.5	99.0	99.5
1	$\hat{Y}_{EI}$	1.0	3.0	5.5	9.0	91.5	96.0	99.0	99.5
No x-values	$\hat{Y}_{\text{Hajek}}$	1.0	2.5	5.5	9.0	93.0	97.0	98.5	99.5
	$\hat{Y}_{1}$	0.5	2.0	5.0	9.0	91.0	94.5	96.5	97.5
2	$\hat{Y}_{2,\text{Reg}}$	0.5	2.5	6.5	10.5	90.5	94.5	96.5	98.0
3rd order polynomial	$\hat{Y}_{3}$	0.5	2.0	7.5	12.5	91.5	95.5	96.5	97.5
regression	$\hat{Y}_{\text{GREG}}$	0.5	2.0	6.0	11.0	91.0	94.5	96.0	98.0
	$\hat{Y}_{2,\text{Reg}}$	1.0	3.0	6.0	11.0	91.0	95.0	97.5	99.0
3	$\hat{Y}_{3}$	1.0	2.0	6.0	12.0	90.0	95.0	97.5	98.0
Simple Regression	$\hat{Y}_{\text{GREG}}$	0.0	1.5	5.0	11.5	91.5	95.0	97.5	99.0

As implied by the theoretical developments of this article and illustrated in the empirical study, predicting only the y-values for units outside the sample employing the sample-complement model yields better predictors for the population total than predicting all the population values by use of the population model, as implicitly implemented when using the GREG or Hajek's estimators. Clearly, the differences are only appreciable when the sampling fractions are not negligible.

In order to highlight this point further, we present in Table 3 the mean prediction error (mpe) in the original scale (grams) over the 1,000 samples when predicting the sample-complement values;

$$\text{mpe} = \sum_{r=1}^{1{,}000}\left[\sum_{j\notin S_r}(\hat{y}_j - y_j)/(N-n)\right]\Big/1{,}000$$

where S_r defines the r^{th} selected sample. The mpe's are shown for three predictors, all utilizing the working model (7.1) and thus having the general form, $\hat{y}_j = \hat{\beta}_0 + \hat{\beta}_1 x_j + \hat{\beta}_2 x_j^2 + \hat{\beta}_3 x_j^3$, $j \notin s$. For the first predictor the vector $\beta = (\beta_0, \beta_1, \beta_2, \beta_3)'$ is estimated by OLS, which corresponds to the use of the sample model; for the second predictor β is estimated by the probability weighted estimator $\hat{B}_{pw}$, that corresponds to the use of the population model whereas for the third predictor β is estimated by the estimator $\hat{B}_c$ which is computed similarly to $\hat{B}_{pw}$ but with weights $(w_i - 1)$, that corresponds to the use of the sample-complement model.

Table 3
Mean Prediction Errors and Std of Means (in brackets) Under Three Prediction Models

Sample size	Sample Model	Population model	Sample-Complement model
232	329.0 (2.2)	10.3 (2.3)	4.3 (2.3)
1,145	375.0 (0.9)	37.7 (1.1)	2.4 (1.1)
2,250	387.5 (0.6)	85.8 (0.7)	0.9 (0.8)

The clear conclusion emerging from Table 3 is that the use of either the population model or the model holding for units in the sample for the prediction of y-values of units outside the sample can result in appreciable biases. Notice that the bias induced by use of the population model increases as the sampling fraction increases, which agrees with the previous discussion asserting that the difference between the sample and sample-complement models only shows up with relatively large sample sizes (see Comment 2).

8. CONCLUDING REMARKS

In this article we use the sample and sample-complement distributions for developing *design consistent* predictors of finite population totals. Known predictors in common use are shown to be special cases of the present theory. The MSEs of the new predictors are estimated by a combination of an inverse sampling algorithm and a resampling method. As supported by theory and illustrated in the empirical study, predictors of finite population totals that only require the prediction of the outcome values for units outside the sample perform better than predictors in common use even under a design based framework, unless the sampling fractions are very small. The MSE estimators are shown to perform well both in terms of bias and when used for the computation of confidence intervals for the population totals. Further experimentation with this kind of predictors and MSE estimation is therefore highly recommended.

ACKNOWLEDGEMENT

The authors would like to thank the associate editor and two referees for very constructive comments.

REFERENCES

BREWER, K.R.W. (1963). Ratio estimationand finite populations: some results deducible from the assumptions of an underlying stochastic process. *Australian Journal of Statistics*. 5, 93-105.

BREWER, K.R.W. (1999). Cosmetic calibration with unequal probability sampling. *Survey Methodology*. 25, 205-212.

CHAMBERS, R.L., DORFMAN, A. and SVERCHKOV, M. (2003). Nonparametric regression with complex survey data. In, *Analysis of Survey Data*, (Eds. C. Skinner and R. Chambers). New York: John Wiley & Sons, Inc. 151-174.

FULLER, W. (2003). Statistical analysis from complex survey data. Tutorial presented at the International Statistical Institute meeting, Berlin, Germany. Slides of the Tutorial appear in http://cssm.iastate.edu/academic/ staff/fuller.html.

HANSEN, M.H., MADOW, W.G. and TEPPING, B.J. (1983). An evaluation of model-dependent and probability-sampling inferences in sample surveys (with discussion). *Journal of the American Statistical Association.*. 78, 776-807.

KIM, D.H. (2002). Bayesian and empirical Bayesian analysis under informative sampling. *Sankhyā B*. 64, 267-288.

KORN, E.L., and GRAUBARD, B.I. (1995). Examples of differing weighted and unweighted estimates from a sample survey. *The American Statistician*. 49, 291-295.

PATAK, Z., HIDIROGLOU, M. and LAVALLÉE, P. (2000). The methodology of the Workplace and Employee Survey. *Proceedings of the Second International Conference on Establishment Surveys*, June 17-21, 2000, Buffalo, New York, American Statistical Association. 223-232.

PFEFFERMANN, D. (1993). The role of sampling weights when modeling survey data. *International Statistical Review*. 61, 317-337.

PFEFFERMANN, D., and KRIEGER, A.M. (1997). Testing of distribution functions from complex sample surveys. *Journal of Official Statistics*. 13, 123-142.

PFEFFERMANN, D., KRIEGER, A.M. and RINOTT, Y. (1998). Parametric distributions of complex survey data under informative probability sampling. *Statistica Sinica*. 8, 1087-1114.

PFEFFERMANN, D., and SVERCHKOV, M. (1999). Parametric and semi-parametric estimation of regression models fitted to survey data. *Sankhyā*, Series B. 61. 166-186.

PFEFFERMANN, D., and SVERCHKOV, M. (2003a). Fitting generalized linear models under informative probability sampling. In *Analysis of survey Data*, (Eds. C. Skinner and R. Chambers). New York: John Wiley & Sons, Inc. 175-195.

PFEFFERMANN, D., and SVERCHKOV, M. (2003b). Small area estimation under informative sampling. *Proceedings of the Section on Survey Research Methods*, American Statistical Association (to appear).

ROYALL, R.M. (1970). On finite population sampling theory under certain linear regression models. *Biometrika*. 57, 377-387.

SÄRNDAL, C.E. (1980). On π-inverse weighting versus best linear unbiased weighting in probability sampling. *Biometrika*. 67, 639-650.

Survey Methodology, June 2004
Vol. 30, n° 1, pp. 93-103
Statistics Canada

Weighted Estimation in Multilevel Ordinal and Binary Models in the Presence of Informative Sampling Designs

LEONARDO GRILLI and MONICA PRATESI[1]

ABSTRACT

Multilevel models are often fitted to survey data gathered with a complex multistage sampling design. However, if such a design is informative, in the sense that the inclusion probabilities depend on the response variable even after conditioning on the covariates, then standard maximum likelihood estimators are biased. In this paper, following the Pseudo Maximum Likelihood (PML) approach of Skinner (1989), we propose a probability-weighted estimation procedure for multilevel ordinal and binary models which eliminates the bias generated by the informativeness of the design. The reciprocals of the inclusion probabilities at each sampling stage are used to weight the log-likelihood function and the weighted estimators obtained in this way are tested by means of a simulation study for the simple case of a binary random intercept model with and without covariates. The variance estimators are obtained by a bootstrap procedure. The maximization of the weighted log-likelihood of the model is done by the NLMIXED procedure of the SAS, which is based on adaptive Gaussian quadrature. Also the bootstrap estimation of variances is implemented in the SAS environment.

KEY WORDS: Informative design; Multilevel ordinal model; Multistage sampling; Pseudo Maximum Likelihood; Weighting.

1. INTRODUCTION

Multilevel models for ordinal responses, including binary responses as a special case, are frequently used in many areas of research for modelling hierarchically clustered populations. In fact, both in human and biological sciences, the status or the response of a subject may often be classified in two categories or in a set of ordered categories (ordinal or graded scale). At the same time, subjects are observed clustered in groups (*e.g.*, schools, firms, clinics, geographical areas). The hierarchical population structure is often also employed to design multistage sampling schemes, with unequal selection probabilities at some or all the stages of the sampling process. In the multilevel analysis of survey data, complex sampling schemes are often ignored even if they may cause the violation of the basic assumptions underlying multilevel models. In fact, in complex sampling designs both the subjects and the clusters at all levels could be selected with probabilities that, even conditionally on the covariates, do depend on the response variable; in other words, the sampling design might be informative.

For data that are clustered and obtained by multistage informative designs, proposals for fitting multilevel models have been formulated mainly for the case of continuous response variables. In particular, Pfeffermann, Skinner, Holmes, Goldstein and Rasbash (1998) propose probability-weighting procedures of first and second level units that adjust for the effect of an informative design on the estimation in two-level models with a continuous response variable. The method, known as Pseudo Maximum Likelihood (PML), consists in writing down a closed form expression for the census likelihood, estimating the log-likelihood function and then maximizing the estimated function numerically. The method needs the sampling weights for the sampled elements and clusters at all levels. The authors also develop appropriate 'sandwich' estimators for the variances of the estimators.

The work of Pfeffermann *et al.* (1998) is mainly concerned with the implementation of the PML principle in the IGLS (Iterative Generalised Least Squares) algorithm (Goldstein 1986), which is suitable for linear multilevel models. The probability-weighted IGLS algorithm is available in the widespread package MLwiN (Rasbash, Browne, Goldstein, Yang, Plewis, Healy, Woodhouse and Draper 1999). However, the extension to nonlinear models is not trivial. For the nonlinear case the developers of MLwiN implemented a weighting procedure that parallels the one used for linear models with some ad hoc solution for the level 1 variation: for example, for binary responses the subject-level weights are included in the binomial denominator. The proposed method is straightforward to implement, but its properties have not been investigated yet. Moreover Renard and Molenberghs (2002) report the case of an application where the aforementioned algorithm for weighting in multilevel binary models did non converge or yielded implausible results.

[1] Leonardo Grilli, Dipartimento di Statistica, Università di Firenze. E-mail: grilli@ds.unifi.it; Monica Pratesi, Dipartimento di Statistica e Matematica applicata all'Economia, Università di Pisa. E-mail: m.pratesi@ec.unipi.it.

The simulation study which we will use to judge the performance of the PML estimators will closely follow the lines of Pfeffermann *et al.* (1998), since they use a similar approach for the linear model, so that some interesting comparisons are possible. However, when making the comparisons it should always kept in mind that, while in the two-level linear model the two variance components can be estimated separately, in the two-level binary model only a ratio of the two variance components is estimable, as discussed further on.

A recent paper which deals with the estimation of variance components is Korn and Graubard (2003), whose work is motivated by the substantial bias showed in small samples by several weighted estimators of variance components proposed to adjust for informative designs (Graubard and Korn 1996). Though the topic is same, the work of Korn and Graubard is different from ours in many respects: a) As Pfeffermann *et al.* (1998), they consider only the linear multilevel model. b) In the context of the linear multilevel model, they focus on unbiased estimation of the variance components in small samples: in fact they propose some estimators for the variance components and only sketch how to derive similar estimators for the linear model with covariates, but without testing their performance. Anyway, the extension to nonlinear multilevel models is not trivial. c) The main estimators proposed by Korn and Graubard (2003), which are in closed form, showed good performance even in small samples. However they rely on the pairwise joint inclusion probabilities. When such probabilities are not available, which is often the case in practice, the authors propose a variant whose bias is substantial when the number of sampled clusters is moderate (33 in their simulation plan). In contrast, the PML method adopted in our work do not require joint inclusion probabilities. d) The informative design used by Korn and Graubard (2003) for their simulation study is quite different from ours: in fact, in their design the undersampling of the units depends on whether the model's random errors are greater than a certain threshold in absolute value, while in our design the criterion depends on whether the random errors are high or low. Therefore a comparison of the results is difficult.

The wide use of nonlinear multilevel models in many fields of application urges for a general and reliable weighted estimation method, which should be both effective and simple to implement, preferably in the framework of a standard statistical software. The present paper represents a contribution in this direction.

It is worth to note that the PML method we exploit is quite general, so it can be applied to a wide range of models. In the paper the focus is on models for ordinal and binary responses, since they are very common and can be represented as a linear model for the latent response endowed with a set of thresholds (see section 2), facilitating the comparison with the existing results for the linear model. However the description of the PML approach is absolutely general and the estimation technique based on the NLMIXED procedure of SAS (reported in Appendix A) is easy to generalize.

The structure of the paper is as follows. Basic definitions for the multilevel ordinal model are set out in section 2, while in section 3 the general PML approach is described, along with some details for fitting the model using SAS NLMIXED. In section 4 the properties of the various estimators for the random intercept binary model are evaluated by a simulation study. Section 5 concludes with some final remarks.

2. THE MULTILEVEL ORDINAL MODEL

In order to ease the comparison with the results concerning the linear model (Pfeffermann *et al.* 1998; Korn and Graubard 2003), it is useful to write the ordinal model in terms of a latent linear model endowed with a set of thresholds. Suppose that an observed ordinal response variable Y, with $k = 1,2,...,K$ levels, is generated, through a set of thresholds, by a latent continuous variable $\tilde{Y}$ following a variance component model (Hedeker and Gibbons 1994):

$$\tilde{Y}_{ij} = \boldsymbol{\beta}'\mathbf{x}_{ij} + \omega u_j + \varepsilon_{ij}, \qquad (1)$$

with $i = 1,2,...,N_j$ elementary units (subjects) for the j-th cluster ($j = 1,2,...,M$). In (1) $\mathbf{x}_{ij}$ is a covariate vector and β is the corresponding vector of slopes; the random variables ε_{ij} and u_j are the disturbances, respectively at the first (subject) and second (cluster) level; and ω^2 is the second level variance component.

For the disturbances of model (1) we make the standard assumptions, *i.e.*, a) the ε_{ij}'s are iid with zero mean and unknown variance σ^2; b) the u_j's are Gaussian iid with zero mean and unit variance; c) the ε_{ij}'s and u_j's are mutually independent.

Note that model (1) leads to the simplest case of a multilevel ordinal model, with just two levels and a single random effect on the intercept; the extension to three or more levels and to multiple random effects is straightforward in principle (Gibbons and Hedeker 1997), but the complications in the formulae suggest to consider only the simplest case, which is sufficient for the discussion of the main conceptual issues.

The observed ordinal variable Y is linked to the latent one $\tilde{Y}$ through the following relationship:

$$\{Y_{ij} = k\} \Leftrightarrow \{\gamma_{k-1} < \tilde{Y}_{ij} \leq \gamma_k\},$$

where the thresholds satisfy $-\infty = \gamma_0 \le \gamma_1 \le \ldots \le \gamma_{K-1} \le \gamma_K = +\infty$. Therefore, conditional on u_j, the model probability for subject i of cluster j is

$$P(Y_{ij} = k \mid u_j) = P(\gamma_{k-1} < \tilde{Y}_{ij} \le \gamma_k \mid u_j)$$

$$= P(\tilde{Y}_{ij} \le \gamma_k \mid u_j) - P(\tilde{Y}_{ij} \le \gamma_{k-1} \mid u_j), \quad (2)$$

with

$$P(\tilde{Y}_{ij} \le \gamma_k \mid u_j) = P(\varepsilon_{ij} \le \gamma_k - [\beta' \mathbf{x}_{ij} + \omega u_j] \mid u_j)$$

$$= F\left(\frac{\gamma_k}{\sigma} - \left[\frac{1}{\sigma} \beta' \mathbf{x}_{ij} + \frac{\omega}{\sigma} u_j \right] \right)$$

$$= F\left(\gamma_{\sigma,k} - \left[\beta_\sigma' \mathbf{x}_{ij} + \omega_\sigma u_j \right]\right), \quad (3)$$

where $F(\cdot)$ is the distribution function of the standardized first level error term ε_{ij}/σ. All the model parameters are defined in terms of the unknown σ, the standard deviation of the first level error term, so only the ratios of the model parameters to the standard deviation of the first level error term are identifiable; we use the notation ψ_σ to indicate that the latent model parameter ψ is in σ units, *i.e.*, $\psi_\sigma \equiv \psi/\sigma$. Note that $F(\cdot)$ is also the inverse of the link function of the ordinal model: for example, the standard Gaussian distribution function yields the ordinal probit model.

As for identification, if β_σ includes the intercept, the estimable thresholds are $K-2$; so it is customary to set $\gamma_{\sigma,1} = 0$. Alternatively, if the intercept is fixed to zero all the $K-1$ thresholds are estimable.

Now let θ denote the vector of all estimable parameters, which include $\beta_\sigma, \omega_\sigma$ and $K-2$ thresholds $\{\gamma_{\sigma,k}: k = 2,\ldots,K-1\}$ ($\gamma_{\sigma,1}$ is fixed to zero to insure identifiability). The conditional likelihood for subject i of cluster j is

$$L_{ij}(\theta \mid u_j) = \prod_{k=1}^{K} \left[P(Y_{ij} = k \mid u_j) \right]^{d_{ijk}}, \quad (4)$$

where $P(Y_{ij} = k \mid u_j)$ is defined by (2) and (3), while d_{ijk} is the indicator function of the event $\{Y_{ij} = k\}$. Then the marginal likelihood for cluster j is

$$L_j(\theta) = \int_{-\infty}^{+\infty} \prod_{i=1}^{N_j} L_{ij}(\theta \mid u) \varphi(u) du,$$

where φ is the standard Gaussian density function. Finally, the overall marginal likelihood is

$$L(\theta) = \prod_{j=1}^{M} L_j(\theta). \quad (5)$$

3. PROBABILITY-WEIGHTED ESTIMATION

3.1 Pseudo Maximum Likelihood (PML) Estimators

Suppose that the whole population of M clusters (level 2 units) with N_j elementary units (subjects or level 1 units) per cluster is not observed; instead the following two-stage sampling scheme is used:

- first stage: m clusters are selected with inclusion probabilities $\pi_j (j = 1,\ldots,M)$;
- second stage: n_j elementary units are selected within the j-th selected cluster with probabilities $\pi_{i|j} (i = 1,\ldots,N_j)$.

The unconditional sample inclusion probabilities are then $\pi_{ij} = \pi_{i|j} \pi_j$.

When the sampling mechanism is informative, *i.e.*, the π_j and/or the $\pi_{i|j}$ depend on the model disturbances and hence on the response variable, the maximum likelihood estimator of the parameters of the multilevel ordinal model defined in section 2 may be seriously biased.

A standard solution to this problem is provided by the Pseudo Maximum Likelihood (PML) approach (Skinner 1989). However in the context of multilevel models the implementation of the PML approach is complicated by the fact that the population log-likelihood is not a simple sum of elementary unit contributions, but rather a function of sums across level 2 and level 1 units. This can be seen by writing the logarithm of the likelihood (5) as follows:

$$\log L(\theta) = \sum_{j=1}^{M} \log \int_{-\infty}^{+\infty} \left[\exp\left\{ \sum_{i=1}^{N_j} \log L_{ij}(\theta \mid u) \right\} \right] \varphi(u) du. \quad (6)$$

A design consistent estimate of the population log-likelihood (6) can be obtained applying the Horvitz-Thompson principle, *i.e.*, replacing each sum over the level 2 population units j by a sample sum weighted by $w_j \equiv 1/\pi_j$ and each sum over the level 1 units i by a sample sum weighted by $w_{i|j} \equiv 1/\pi_{i|j}$:

$$\log \hat{L}(\theta) =$$

$$\sum_{j}^{s} w_j \log \int_{-\infty}^{+\infty} \left[\exp\left\{ \sum_{i}^{s} w_{i|j} \log L_{ij}(\theta \mid u) \right\} \right] \varphi(u) du, \quad (7)$$

where Σ^s denotes a sum over sample units.

Note that inserting the weights in the log-likelihood implies the use of a design consistent estimator of the population score function. In fact, the population score function $U(\theta) \equiv \partial/\partial\theta \log L(\theta)$ can be written as

$$\sum_{j=1}^{M} \frac{\int_{-\infty}^{+\infty}\left[\exp\left\{\sum_{i=1}^{N_j}\log L_{ij}\right\}\right]\cdot\left\{\sum_{i=1}^{N_j}\frac{\partial}{\partial\theta}\log L_{ij}\right\}\varphi(u)du}{\int_{-\infty}^{+\infty}\exp\left\{\sum_{i=1}^{N_j}\log L_{ij}\right\}\varphi(u)du}, \quad (8)$$

where $L_{ij} = L_{ij}(\theta|u)$, whose corresponding Horvitz-Thompson estimator $\hat{U}(\theta)$ is

$$\sum_{j}^{s} w_j \frac{\int_{-\infty}^{+\infty}\left[\exp\left\{\sum_{i}^{s} w_{i|j}\log L_{ij}\right\}\right]\cdot\left\{\sum_{i}^{s} w_{i|j}\frac{\partial}{\partial\theta}\log L_{ij}\right\}\varphi(u)du}{\int_{-\infty}^{+\infty}\exp\left\{\sum_{i}^{s} w_{i|j}\log L_{ij}\right\}\varphi(u)du}, \quad (9)$$

which equals the score obtained by differentiating the probability-weighted loglikelihood (7).

Under mild conditions, the solution $\hat{\theta}_{PML}$ to the estimating equations $\hat{U}(\theta)=0$ is design consistent for the finite population maximum likelihood estimator $\hat{\theta}$ which, in turn, is model-consistent for the super-population parameter θ: therefore $\hat{\theta}_{PML}$ is a consistent estimator of θ with respect to the mixed design-model distribution (Pfeffermann 1993).

Note that general probability-weighted estimators for nonlinear multilevel models can also be devised by weighting suitable estimating functions, as in the work of Singh, Folsom and Vaish (2002) in the context of small area estimation.

The implementation of the PML approach requires the knowledge of the inclusion probabilities at both levels. Using only second level weights or only first level weights may be insufficient or may even worsen the situation, as shown by our simulations.

3.2 Scaling the Weights

A controversial issue discussed in Pfeffermann *et al.* (1998) and Korn and Graubard (2003) is the scaling of the weights to obtain estimators with little bias even in small samples. Obviously, scaling is not relevant for the level 2 weights, since from (7) and (9) it is clear that multiplying the w_j's by a constant does not change the PML estimates (it simply inflates the information matrix by that constant). On the contrary, scaling the level 1 weights may have important effects on the small sample behavior of the PML estimator. In the simulation study discussed in section 4 we present the results for the following type of scaling (named 'scaling method 2' in Pfeffermann *et al.* 1998):

$$w_{i|j}^{\text{scaled}} = \frac{w_{i|j}}{\bar{w}_j}, \quad (10)$$

where $\bar{w}_j = (\sum_i^s w_{i|j})/n_j$, so that, for the j-th cluster, the sum of the scaled weights equals the cluster sample size n_j. In the present paper we do not wish to discuss the relative merits of the various scaling methods, so we limit our simulations to scaled weights (10), which have an intuitive meaning and showed good performance in the study of Pfeffermann *et al.* (1998), although they may yield a substantial bias with certain designs, as discussed in Korn and Graubard (2003). The topic will be broached again in section 4.

3.3 Estimation Technique

The maximization of the weighted log-likelihood (7) involves the computation of several integrals which do not have a closed-form solution, so a numerical approximation technique is required. When the dimensionality of the integrals is low, a simple and very accurate technique is Gaussian quadrature, which is based on a summation over an appropriate set of points. The NLMIXED procedure of SAS (SAS Institute 1999) is a general procedure for fitting nonlinear random effects models using adaptive Gaussian quadrature. Various optimization techniques are available to carry out the maximization; the default, used in the simulations of section 4, is a dual quasi–Newton algorithm, where dual means that the upgrading concerns the Cholesky factor of an approximate Hessian (SAS Institute 1999).

Though the NLMIXED procedure does not include an option for PML estimation, it is still possible to insert the weights in the likelihood, using different tricks for level 1 and level 2 weights, as explained in Appendix A.

3.4 Variance Estimation

In standard maximum likelihood the estimation of the covariance matrix of the estimators is obtained by inverting the information matrix. However this conventional estimator is not appropriate when using the PML method since it does not take into account the variability stemming from the sampling design. To get a more reliable covariance matrix Skinner (1989) proposed the use of a robust 'sandwich' estimator, which is employed also by Pfeffermann *et al.* (1998).

As noted in section 3.3, the NLMIXED procedure of SAS allows to fit the model with the PML approach, but the estimated covariance matrix, which is obtained by inverting the information matrix, is likely to be misleading in order to appreciate the actual variability of PML estimators. In the SAS framework the derivation of 'sandwich' estimators is not trivial. However, a simple and effective solution, requiring a bit of programming, is to empirically estimate the variance through the bootstrap technique for finite populations (Särndal, Swensson and Wretman 1992), which consists of the following steps: a) using the sample data, an artificial finite population is constructed, assumed to mimic

the real population; b) a series of independent bootstrap samples is drawn from the artificial finite population and for each bootstrap sample an estimate of the target parameter is calculated; c) the bootstrap variance estimate is obtained as the variance of the observed distribution of the bootstrap estimates.

The artificial finite population can be generated in the following way: i) for the j-th sampled cluster, each of the n_j sampled elementary units is replicated $w_{i|j}$ times, rounding the weight to the nearest integer, obtaining an artificial cluster of about N_j elementary units; ii) each of the m artificial clusters is replicated w_j times, rounding the weight to the nearest integer, obtaining an artificial population of about M clusters. Then the samples are selected from the artificial population in the following way: i) m clusters are resampled with probability proportional to π_j; ii) for the j-th resampled cluster, n_j elementary units are resampled with probability proportional to $\pi_{i|j}$.

When the sampling fraction m/M is low, most of the variance is due to the sampling of the clusters, so the bootstrap procedure described above could be simplified by omitting the steps concerning the elementary units, *i.e.*, step i) in the construction of the artificial population and step ii) in the resampling process.

A simpler resampling technique for variance estimation, considered by Korn and Graubard (2003), is the jackknife. In the case of clustered designs the technique entails the calculation of the variance from the set of point estimates obtained by deleting one cluster at a time, though the performance of the jackknife with correlated data is not always satisfactory (Shao and Tu 1995). In our simulation study the jackknife variance estimator seems unreliable, so it is not used. Further research is needed to fully evaluate the potentialities of the jackknife by testing some suitable modifications of the technique.

4. SIMULATION STUDY

4.1 Design of Experiment

The experiment reflects the two-stage scheme assumed for the observed variables: first, the finite population values are generated from the adequate superpopulation model (stage I) and then an informative or non-informative sample is selected from the finite population (stage II), with one sample per population. The two-stage selection scheme was repeated 1,000 times for each combination of sample size and type of informativeness. In order to compare our results with the ones obtained for the multilevel linear model, the experiment has been designed following the example of Pfeffermann *et al.* (1998, section 7).

The simulation study focussed on a simple instance of the model defined in section 2, namely the random intercept probit binary model, which has only two categories for the response variable (*i.e.*, $K=2$) and one cluster-level Gaussian random error. To parallel the study of Pfeffermann *et al.* (1998) the main simulation plan refers to the model without covariates, but some additional simulations are conducted to assess the performance of the estimators in the model with one cluster-level covariate and one subject-level covariate.

The values of the binary response variable Y_{ij} were generated using the following two-stage scheme which parallels the one of Pfeffermann *et al.* (1998):

- Stage I. Finite population values Y_{ij} ($j=1,...,M; i=1,...,N_j$) were obtained by first generating a value from the superpopulation latent model $\tilde{Y}_{ij} = \beta + u_j + \varepsilon_{ij}$, with $u_j \sim N(0, \omega^2)$ and $\varepsilon_{ij} \sim N(0, \sigma^2)$, and then putting $Y_{ij} = 0$ if $\tilde{Y}_{ij} \leq 0$ or $Y_{ij} = 1$ if $\tilde{Y}_{ij} > 0$ (recall that the binary model has only one threshold which is set to zero to guarantee identifiability). The latent model parameter values employed in the simulation are $\beta = 0$, $\omega^2 = 0.2$ and $\sigma^2 = 0.5$, so that the parameters estimable from the binary model are $\beta_\sigma \equiv \beta/\sigma = 0$ and $\omega_\sigma \equiv \omega/\sigma = 0.632$ (see expression (3)). The hierarchical structure of the population comprises $M = 300$ clusters, while the cluster sizes N_j were determined by $N_j = 75\exp(\tilde{u}_j)$, with $\tilde{u}_j$ generated from $N(0, \omega^2)$, truncated below by -1.5ω and above by 1.5ω. As a result, in our population N_j lies in the range [38, 147] with mean around 80.
- Stage II. Once the finite population values were obtained, we adopted one of the following sampling schemes:

  (a) *Informative at both levels*: first, m clusters were selected with probability proportional to a 'measure of size' X_j, *i.e.*, $\pi_j = mX_j / \sum_{j=1}^{M} X_j$; the measure X_j was determined in the same way as N_j but with $\tilde{u}_j$ replaced by u_j, the random effect at level 2. The elementary units in the j-th sampled cluster were then partitioned into two strata according to whether $\varepsilon_{ij} > 0$ or $\varepsilon_{ij} \leq 0$ and simple random samples of sizes $0.25n_j$ and $0.75n_j$ were selected from the respective strata. The sizes n_j were either fixed, $n_j = n_0$, or proportional to N_j.

  (b) *Informative only at level 2*: the scheme is the same as the previous one, except that simple random sampling was employed for the selection of level 1 units within each sampled cluster.

(c) *Non-informative*: the scheme is the same as the previous one, except that the size measure X_j was set equal to N_j.

The simulation study included samples with $m = 35$ clusters and varying numbers of elementary units: large samples with fixed size $n_j = n_0 = 38$ and proportional allocation $n_j = 0.4N_j$, and small samples with fixed size $n_j = n_0 = 9$ and proportional allocation $n_j = 0.1N_j$ (mean of about 9).

The simulation study was carried out entirely within the SAS System (SAS Institute 1999), writing specific code with the macro language. The models were fitted with the NLMIXED procedure (see Appendix A), using 10-point adaptive Gaussian quadrature with a dual quasi-Newton algorithm, which reached convergence in a few iterations. As explained in Appendix A, to avoid gross rounding errors the level 2 weights were pre-multiplied by a factor $k = 10{,}000$ and the estimated covariance matrix was then multiplied by the same factor.

4.2 Results

The results of the simulations are shown in Tables 1 and 2. For each sampling design the behavior of the point estimators of the intercept β_σ and the second level standard deviation ω_σ is summarized by the mean and standard deviation of their Monte Carlo sampling distribution. The point estimators under study are the standard maximum likelihood unweighted estimator and the following three weighted versions of it: a) *cluster-level weighted*: the weights are only at level 2 (*i.e.*, varying w_j's and constant $w_{i|j}$'s); b) *unscaled fully weighted*: the weights are at both levels and the level 1 weights are unscaled; c) *scaled fully weighted*: the weights are at both levels and the level 1 weights are scaled according to (10), *i.e.*, 'scaling method 2' of Pfeffermann *et al.* (1998).

Our results are shown and discussed according to the following three scenarios: 1) *Base scenario*: the sampling design is non-informative. In this situation all the basic assumptions underlying the random intercept binary model are fulfilled, so this case can be assumed as a benchmark for judging the subsequent results. 2) *Informative/ Unweighted scenario*: the sampling design is informative, while the estimator is unweighted. In this situation the basic assumptions underlying the random intercept binary model are violated because of the informativeness of the design and no adjustment is used. 3) *Informative/Weighted scenario*: the sampling design is informative and the estimator is weighted. Also in this case the basic assumptions underlying the random intercept binar model are violated, but the weights are introduced as a tentative adjustment for the bias of the standard estimator.

4.2.1 Base Scenario

When the sampling design is non-informative the standard maximum likelihood unweighted estimator is asymptotically unbiased (Tables 1 and 2: rows 9-12, column 1). However for small samples ($n_j = 9$ and $n_j 0.1N_j$) there is an appreciable negative bias in the estimation of ω_σ.

If the weights are introduced when there is no need to adjust for the effect of the design (Tables 1 and 2: rows 9-12, columns 2-4), we face a slight increase in the variability of the estimators, which is more pronounced when the unscaled fully weighted estimator is used in small samples. Note that, still in small samples, the unscaled fully weighted estimator of ω_σ is upward biased.

4.2.2 Informative/Unweighted Scenario

The informativeness of the sampling design produces biased and unstable estimates. The bias is still evident for large samples (Tables 1 and 2: rows 1-8, column 1). The conclusions are the same for both types of informative designs, though the bias tends to have a different sign. Moreover the informativeness of the design inflates the variability of the standard estimator with respect to the base scenario: in particular, when the design is informative at both levels the standard error of the estimator of β_σ is doubled.

4.2.3 Informative/Weighted Scenario

Estimation of β_σ.

The results in Table 1 show that, when the design is informative, the weighted-based adjustment is effective in removing the bias in the estimation of β_σ.

Particularly, when the design is informative only at level 2 (Table 1: rows 5-8, columns 2-4) and the weights are introduced only at this level (cluster-level weighted estimator), the bias in the estimation is corrected with no important increase in the sampling variance. The result is valid also for fully weighted estimators (unscaled or scaled). The bias correction works for small samples too.

When the design is informative at both levels (Table 1: rows 1-4, columns 2-4) and the weights are introduced at both levels (fully weighted estimators), the bias in the estimation of β_σ is corrected. Moreover, the fully weighted estimators have smaller sampling variance than the unweighted counterpart, except for the unscaled version in small samples. The scaled version is preferable especially in small samples, since it allows to achieve an unbiased estimator with a substantial lower sampling variance. It should be noted that when the design is informative at both levels, the cluster-level weighted estimator is worse than the standard unweighted estimator.

Table 1
Simulation Means and Standard Deviations (in parenthesis) of Point Estimators of the Intercept (true value 0, number of replicates 1,000)

Sampling design	Unweighted estimator	Weighted estimators		
		Cluster-level weighted	Unscaled fully weighted	Scaled fully weighted
Informative at both levels				
Fixed size $n_j = 38$	-0.120 (0.212)	-0.411 (0.202)	0.014 (0.193)	0.015 (0.188)
Prop. size $n_j = 0.4N_j$	-0.163 (0.212)	-0.453 (0.200)	0.018 (0.190)	0.021 (0.183)
Fixed size $n_j = 9$	-0.214 (0.204)	-0.512 (0.190)	-0.062 (0.258)	0.000 (0.185)
Prop. size $n_j = 0.1N_j$	-0.164 (0.220)	-0.450 (0.209)	-0.074 (0.294)	0.008 (0.203)
Informative only at cluster level (level 2)				
Fixed size $n_j = 38$	0.281 (0.169)	0.018 (0.168)	0.017 (0.170)	0.017 (0.169)
Prop. size $n_j = 0.4N_j$	0.274 (0.169)	0.014 (0.178)	0.014 (0.182)	0.014 (0.181)
Fixed size $n_j = 9$	0.274 (0.187)	0.010 (0.195)	0.010 (0.212)	0.009 (0.196)
Prop. size $n_j = 0.1N_j$	0.269 (0.179)	0.007 (0.179)	0.007 (0.203)	0.006 (0.182)
Non-informative				
Fixed size $n_j = 38$	0.000 (0.108)	0.000 (0.114)	0.001 (0.115)	0.001 (0.115)
Prop. size $n_j = 0.4N_j$	0.003 (0.113)	0.004 (0.120)	0.003 (0.123)	0.003 (0.122)
Fixed size $n_j = 9$	-0.007 (0.108)	-0.009 (0.115)	-0.010 (0.125)	-0.010 (0.117)
Prop. size $n_j = 0.1N_j$	-0.002 (0.110)	-0.002 (0.114)	-0.004 (0.132)	-0.003 (0.117)

Table 2
Simulation Means and Standard Deviations (in parenthesis) of Point Estimators of the Second Level Standard Deviation (true value 0.632, number of replicates 1,000)

Sampling design	Unweighted estimator	Weighted estimators		
		Cluster-level weighted	Unscaled fully weighted	Scaled fully weighted
Informative at both levels				
Fixed size $n_j = 38$	0.671 (0.106)	0.638 (0.112)	0.637 (0.137)	0.604 (0.128)
Prop. size $n_j = 0.4N_j$	0.673 (0.108)	0.636 (0.112)	0.645 (0.142)	0.592 (0.130)
Fixed size $n_j = 9$	0.644 (0.145)	0.584 (0.172)	0.920 (0.289)	0.536 (0.222)
Prop. size $n_j = 0.1N_j$	0.598 (0.164)	0.546 (0.183)	1.002 (0.317)	0.498 (0.242)
Informative only at cluster level (level 2)				
Fixed size $n_j = 38$	0.595 (0.100)	0.596 (0.110)	0.605 (0.111)	0.601 (0.111)
Prop. size $n_j = 0.4N_j$	0.582 (0.096)	0.582 (0.115)	0.603 (0.113)	0.596 (0.113)
Fixed size $n_j = 9$	0.547 (0.121)	0.548 (0.135)	0.671 (0.144)	0.563 (0.133)
Prop. size $n_j = 0.1N_j$	0.538 (0.122)	0.535 (0.142)	0.696 (0.158)	0.551 (0.139)
Non-informative				
Fixed size $n_j = 38$	0.611 (0.086)	0.612 (0.092)	0.621 (0.090)	0.617 (0.091)
Prop. size $n_j = 0.4N_j$	0.609 (0.084)	0.606 (0.088)	0.626 (0.088)	0.618 (0.088)
Fixed size $n_j = 9$	0.561 (0.105)	0.561 (0.112)	0.685 (0.119)	0.575 (0.111)
Prop. size $n_j = 0.1N_j$	0.551 (0.109)	0.546 (0.113)	0.703 (0.134)	0.559 (0.112)

Estimation of ω_σ.

The results in Table 2, concerning ω_σ, are more difficult to interpret (Table 2: rows 1-8, columns 2-4). First note that also in the base scenario the estimation of ω_σ is biased, especially in small samples. Therefore the weight-based adjustment should be judged as effective if it is able to reproduce the same bias which is observed in the base scenario. On these grounds the behavior of the scaled fully weighted estimator is satisfactory in nearly all situations, with the exception of the small samples when the design is informative at both levels. In that case there is also a not negligible number of replications which yielded a zero estimate for ω_σ (4.5% for the design with fixed size and 2% for the design with proportional size). The unscaled fully

weighted estimator does not suffer from the problem of null estimates, but, apart from having a larger variance than the scaled version, tends to overestimate ω_σ, showing a relative bias of about 50% in small samples when the design is informative at both levels. Note also that the scaled fully weighted estimator outperforms the cluster-level weighted estimator even when the design is informative only at level 2.

4.2.4 Additional Simulations Using the Model with Covariates

Some additional simulations were conducted to assess the performance of the scaled fully weighted estimator in the model with one cluster-level covariate and one subject-level covariate. The model is the same used in the main simulation plan, except for the inclusion of a covariate at each hierarchical level. For each covariate the values are generated from a standard Gaussian distribution, while the corresponding regression coefficient is fixed to 0.1.

As shown by Tables 3 and 4, the scaled fully weighted estimator is effective in removing the bias induced by the informative design. Relative to the unweighted estimator the sampling variance is higher, especially for the subject-level regression coeffcient. Overall, the performance of the weighted estimator is satisfactory.

Table 3
Simulation Means and Standard Deviations (in parenthesis) of Point Estimators of the Regression Coefficient of the Subject-Level Covariate (true value 0.1, number of replicates 1,000)

Sampling design	Non informative	Informative at both levels	
	Unweighted estimator	Unweighted estimator	Scaled fully weighted estimator
Fixed size $n_j = 38$	0.101 (0.028)	0.117 (0.040)	0.098 (0.050)
Prop. size $n_j = 0.4N_j$	0.099 (0.026)	0.117 (0.043)	0.098 (0.052)
Fixed size $n_j = 9$	0.099 (0.055)	0.119 (0.083)	0.100 (0.104)
Prop. size $n_j = 0.1N_j$	0.098 (0.056)	0.116 (0.089)	0.098 (0.107)

Table 4
Simulation Means and Standard Deviations (in parenthesis) of Point Estimators of the Regression Coefficient of the Cluster-Level Covariate (true value 0.1, number of replicates 1,000)

Sampling design	Non informative	Informative at both levels	
	Unweighted estimator	Unweighted estimator	Scaled fully weighted estimator
Fixed size $n_j = 38$	0.096 (0.119)	0.117 (0.130)	0.102 (0.142)
Prop. size $n_j = 0.4N_j$	0.102 (0.110)	0.106 (0.133)	0.106 (0.142)
Fixed size $n_j = 9$	0.094 (0.117)	0.116 (0.141)	0.105 (0.150)
Prop. size $n_j = 0.1N_j$	0.094 (0.119)	0.115 (0.144)	0.095 (0.158)

4.2.5 General Remarks

Our simulations showed that the PML approach is, in most cases, a simple and effective strategy to deal with informative sampling designs. The only requirement is the knowledge of the inclusion probabilities at every stage of the sampling process (except when the informativeness does not concern all the levels).

As for the regression parameters, the scaled version of the fully weighted estimator showed good performance in our simulations, achieving a low bias with a modest increase in the sampling variance (in some cases the variance even diminished). Even when weighting is superfluous, the loss of efficiency due to the inclusion of scaled weights is very low.

While for the estimation of the regression parameters weighting seems to be always effective, for the variance component ω_σ attention should be paid to the sample size: in fact, weighting leads to satisfactory results only when the cluster size is high, *i.e.*, when it allows a good representation of the complex variance structure. However the sample size is crucial in the estimation of ω_σ also when all the basic assumptions of the multilevel ordinal model are satisfied.

The differences induced by the type of clusters in the sample, fixed or variable size, are minimal, with equal sized clusters leading to slightly better estimators; however, as already noted, the important differences are largely due to the average size of the clusters in the sample.

The results of our simulation study confirm the findings of Pfeffermann *et al.* (1998) on the random intercept linear model: probability-weighted estimators are good for the intercept, while some relevant bias remains in the estimation of the variance components when the sample is small. As was to be expected, when passing from a linear to a nonlinear model the performance of the estimators slightly worsen, but the direction and importance of the bias in the various cases are similar. Also the advantages of scaling are confirmed.

The rise in the sampling variance due to the inclusion of the weights often has a magnitude which is in line with the results of Pfeffermann *et al.* (1998), though in some cases we found a reduction in the sampling variance, notably for the intercept when the weights are scaled and the design is informative at both levels. An interesting difference with respect to Pfeffermann *et al.* (1998) is the role of scaling in reducing the sampling variance: in this respect, scaling seems to be more effective in the binary model than in the linear model.

As already noted, the critical point in the random intercept binary model is the estimation of the cluster-level variance ω_σ, which represents a difficult task also when the

design is non-informative. Using the threshold formulation outlined in section 2, ω_σ is defined as ω/σ, so estimation of ω_σ involves the problems observed in the linear model associated with estimation of the two variance components. The simulations showed that the performance of the scaled weighted estimator of ω_σ is not entirely satisfactory in the case of small sample sizes. A possible way to improve the performance of the estimator is the adoption of a different scaling method. Korn and Graubard (2003) investigated the issue of scaling in the context of the linear model and warned that the scaling method here adopted ('scaling method 2' of Pfeffermann *et al.* 1998) may be badly biased under some designs, even if the sample size of clusters and sample sizes within the clusters are large. To get an idea of the extent of the bias we performed a short simulation study under the unfavorable scenario outlined by Korn and Graubard (2003), namely a simple random sample of clusters whose population sizes are all equal, and a simple random sample of individuals within each sampled cluster that is of size $2m$ or $m/2$ for a fixed m, depending on whether the observed variability of the individuals within the clusters tends to be large or small, respectively. In this case the scaled weights at subject level are all equal to 1, so weighting becomes ineffective. As a consequence, in the linear variance component model the within variance will be biased high. To see how this behavior extends to the random intercept binary model we simulated 1,000 datasets with 80 clusters and cluster sizes of 36 or 9 depending on whether the binomial variance of the responses of the cluster is over or under the median, respectively. Under the same superpopulation model as in the main simulations, the simulation means (and standard deviations) are -0.003 (0.098) for β_σ and 0.451 (0.144) for ω_σ. The cluster-level variance is heavily underestimated, though its value is not so far from the worst case of the main simulations (0.498 under the informative design with $n_j = 0.1N_j$). Therefore, it seems unlikely to encounter situations where the bias is much greater than already shown by our simulations. Obviously, if estimation of the variance components is of primary interest it is important to improve the method, but this requires further research.

4.2.6 Bootstrap Variance Estimation

The estimated covariance matrix of the parameter estimates obtained by inversion of the information matrix, yielded by default by the NLMIXED procedure, is not reliable when using the weighted estimators to adjust for an informative design. For example, the estimated standard error of the scaled fully weighted estimator under the design informative at both levels with $n_j = 0.4N_j$ is 0.109 for β_σ (compared with a Monte Carlo value of 0.183) and 0.089 for ω_σ (compared with a Monte Carlo value of 0.130). For the other sampling sizes similar downward biases arise, so an alternative variance estimator should be devised.

The bootstrap procedure described in section 3.4 has been applied to estimate the sampling standard deviations of the weighted point estimators of β_σ and ω_σ. We limited the analysis to the scaled fully weighted estimator and to designs that are informative at both levels. To save computational resources we implemented a bootstrap procedure which omits the steps concerning the elementary units, *i.e.*, only the clusters are resampled. This procedure is expected to produce sufficiently accurate results, given the low sampling fraction (35/300) of the clusters (see section 3.4). Each simulation comprises 1,000 replications. For every replication the values of the response variable are generated through the two-stage scheme described in section 4.1 and 200 bootstrap samples are selected. Table 5 reports, for each parameter, the Monte Carlo standard error of the sampling distribution of the scaled weighted estimator on 1,000 replications of the complex design (see Tables 1 and 2), the corresponding average bootstrap estimate and the relative bias.

Table 5
Simulation Standard Deviations of the Scaled Weighted Point Estimators of the Intercept and of the Second Level Standard Deviation and Corresponding Bootstrap Estimates (with 200 Bootstrap Samples Each) for Designs Informative at Both Levels (1,000 Replicates for Each Design)

Sampling design Inform. Both levels	β_σ			ω_σ		
	Simul. s.d.	Boot. Estim.	Relative error	Simul. s.d.	Boot. Estim.	Relative error
Fixed size $n_j = 38$	0.185	0.175	-5.4%	0.124	0.106	-14.5%
Prop. size $n_j = 0.4N_j$	0.183	0.173	-5.5%	0.140	0.129	-7.9%
Fixed size $n_j = 9$	0.200	0.167	-16.5%	0.234	0.599	156.0%
Prop. size $n_j = 0.1N_j$	0.195	0.173	-11.3%	0.247	0.538	117.8%

Due to the extremely long computational time, we limited our experiment to a specific bootstrap procedure based on only 200 bootstrap samples. Further work is needed to calibrate the number of bootstrap samples and to explore possible variants of the method. Nonetheless, the entries of Table 5 give some hints about the behavior of bootstrap estimators.

The performance is better for the estimation of the sampling standard deviation of the estimator of β_σ, rather than of ω_σ. Especially for ω_σ the sample size is the critical factor: for small cluster sizes ($n_j = 9$ and $n_j = 0.1N_j$) the bootstrap estimate is completely unreliable. On the contrary with large cluster sizes ($n_j = 38$ and $n_j = 0.4N_j$) the results are quite good, since for both β_σ and ω_σ the bootstrap produces a slight underestimation of the true variance. Note, however, that the bad performance of the variance

estimator for ω_σ is not as critical since Wald tests for variance parameters are not generally recommended in ordinary situations anyway.

5. FINAL REMARKS

The wide use of multilevel ordinal and binary models in many fields of application has motivated our study on the effects of complex sampling designs on the fitting of such models. In the paper we showed, by means of simulations, the bias induced by a two-stage complex sampling design on the fitting of a simple random intercept binary model when the clusters and/or the subjects are selected with probabilities that depend on the model's random terms. The simulation study also showed that in such situations the bias can be reduced in an effective manner by the probability-weighted estimation procedure (PML) described in the paper, which is easily implemented in the SAS environment. In particular, the scaled version of the weighted estimator achieved, for both fixed and random parameters, a low bias with a modest increase in the sampling variance. Even when weighting is superfluous, the loss of efficiency due to the inclusion of scaled weights seems to be very low.

The application of the proposed methodology to real life examples requires an operational strategy which depends on the extent of the available information on the sampling design. Two extreme cases can be envisaged: a) for each stage of the sampling plan, the probabilities of inclusion and the adjustments for poststratification and nonresponse are exactly known; b) the information is limited to the final overall weights, which also include adjustments for post-stratification and nonresponse.

In case a) the weights can be calculated at each sampling stage as the reciprocals of the product of sample selection probabilities and response probabilities given the sample selection, with a further correction for a possible poststratification. This is the idea behind the real life application presented in Pfefferman *et al.* (1998).

In case b) the lack of information is critical, since, even in the absence of nonresponse and poststratification, it is not possible to disentangle the cluster-level and the (conditional) subject-level weights, at least without strong assumptions. As a result, weighted estimation cannot be performed.

Between the two extreme cases just outlined there are many possible intermediate situations which require *ad hoc* solutions. For example, a common case arises when the researcher has access to the cluster-level inclusion probabilities (π_j) and to the final overall subject-level weights (w_{ij}), which also include adjustments for poststratification and nonresponse. When the poststratification and nonresponse affect only the subject level, then the subject-level (conditional) weights can be calculated as $w^*_{i|j} = w_{ij} \cdot \pi_j$. Another more complex situation is described by Korn and Graubard (2003).

A drawback of probability-weighted estimation is the need for special procedures to estimate the variability of the estimators. In our application we adopted a bootstrap technique, which is conceptually simple and easy to program, but requires some computational effort. Our limited simulation study suggests that its performance is good only for large sample cluster sizes; however more simulations would be needed to fully understand the behavior of the bootstrap estimator.

Another open question is the choice of the most effective scaling method for reducing the bias of the estimator of the variance components when the sample size is small.

The PML approach described in the paper is absolutely general and the estimation technique based on the NLMIXED procedure of SAS is easy to generalize to other nonlinear models. Therefore it would be of interest to assess the performance of the method in models other than the random intercept binary model here considered.

APPENDIX A

We report the SAS code used for implementing the probability-weighted (PML) estimators described in the paper. The essential part of the code is the NLMIXED procedure of SAS, which is a general procedure for fitting nonlinear random effects models using adaptive Gaussian quadrature. Though the NLMIXED procedure does not include an option for PML estimation, it is still possible to insert the weights in the likelihood, using different tricks for level 1 and level 2 weights. To insert level 1 weights it is necessary to exploit the option which allows to write down the expression for the conditional likelihood of the model: then one should simply translate in SAS programming statements the expression $w_{i|j} \log L_{ij}(\boldsymbol{\theta}|u)$ (see section 3.1). On the other hand, level 2 weights can be inserted in the likelihood through the `replicate` statement. Unfortunately, this statement is limited to integer weights, so to avoid gross approximations it is advisable to proceed as follows: a) inflate all the level 2 weights by an arbitrary constant k (equal to 10,000 in our application); b) insert the integer part of the inflated weights in the likelihood through the `replicate` statement; c) multiply the estimated covariance matrix by k by means of the `cfactor` option. This trick relies on the fact that multiplying the level 2 weights by a constant has the only effect of inflating the information matrix by that constant, leaving the estimates unchanged. Anyway, when using the weighted estimation

method to adjust for an informative design the estimated covariance matrix of the parameter estimates is not reliable.

In the following the SAS code is reported, where the symbols /* and */ include the comments:

```
proc nlmixed data=dataname qpoints=10
cfactor=10,000;
/* cfactor is a constant multiplying the
estimated covariance matrix of the parameter
estimates */
parms b0=0 sd=0.5; /* initial values */
bounds sd >= 0;
eta=b0+randeff*sd;
if (yobs=1) then z=probnorm(eta);
else if (yobs=0) then z=1-probnorm(eta);
if (z >1e-8) then ll=log(z); else ll=-1e100;
/*to avoid numerical problems if z becomes
too small*/
ll=ll*w1_2; /* inclusion of level 1 weights
*/
model yobs~general(ll);
random randeff ~normal(0,1) subject=j;
/* j is the cluster identifier */
replicate w2; /* inclusion of level 2
weights (only integers) */
ods output ParameterEstimates=pe
ConvergenceStatus=cs;
run;
```

ACKNOWLEDGEMENTS

We wish to thank two anonymous referees for their suggestions which contributed to substantially improve the paper.

REFERENCES

GIBBONS, R.D., and HEDEKER, D. (1997). Random effects probit and logistic regression models for three-level data. *Biometrics*. 53, 1527-1537.

GOLDSTEIN, H. (1986). Multilevel mixed linear model analysis using iterative generalised least squares. *Biometrika*. 73, 43-56.

GRAUBARD, B.I., and KORN, E.L. (1996). Modelling the sampling design in the analysis of health surveys. *Statistical Methods in Medical Research*. 5, 263-281.

HEDEKER, D., and GIBBONS, R.D. (1994). A random-effects ordinal regression model for multilevel analysis. *Biometrics*. 50, 933-944.

KORN, E.L., and GRAUBARD, B.I. (2003). Estimating variance components by using survey data. *Journal of the Royal Statistical Society B*. 65(1), 175-190.

PFEFFERMANN, D. (1993). The role of sampling weights when modeling survey data. *International Statistical Review*. 61(2), 317-337

PFEFFERMANN, D., SKINNER, C.J., HOLMES, D.J., GOLDSTEIN, H. and RASBASH, J. (1998). Weighting for unequal selection probabilities in multilevel models. *Journal of the Royal Statistical Society B*. 60(1), 23-40.

RASBASH, J., BROWNE, W., GOLDSTEIN, H., YANG, M., PLEWIS, I., HEALY, M., WOODHOUSE, G. and DRAPER, D. (1999). A users guide to MLwiN. London: Multilevel models project, Institute of Education, University of London.

RENARD, D., and MOLENBERGHS, G. (2002). Multilevel modeling of complex survey data. In *Topics in Modelling of Clustered Data* (Ed. M. Aerts M). London: Chapman and Hall. 263-272.

SÄRNDAL, C.-E., SWENSSON, B. and WRETMAN, J. (1992). Model Assisted Survey Sampling. New York: Springer-Verlag.

SAS INSTITUTE (1999). SAS/STAT User's Guide Version 8. Cary: SAS Institute Inc.

SHAO, J., and TU, D. (1995). *The Jackknife and the Boostrap*. New York: Springer-Verlag.

SINGH, A.C., FOLSOM, R.E. and VAISH, A.K. (2002). A hierarchical Bayes generalization of the Fay-Herriot method to unit level nonlinear mixed models for small area estimation. *Proceedings of the Section on Survey Research Methods*, American Statistical Association. New York, August 10-13. 3258-3263.

SKINNER, C.J. (1989). Domain means, regression and multivariate analysis. In *Analysis of Complex Surveys*, (Eds. C.J. Skinner, D. Holt and T.M.F. Smith). Chichester: Wiley. 59-87.

Survey Methodology, June 2004
Vol. 30, No. 1, pp. 105-114
Statistics Canada

Longitudinal Analysis of Labour Force Survey Data

GEOFF ROWE and HUAN NGUYEN[1]

ABSTRACT

The Canadian Labour Force Survey (LFS) was not designed to be a longitudinal survey. However, given that respondent households typically remain in the sample for six consecutive months, it is possible to reconstruct six-month fragments of longitudinal data from the monthly records of household members. Such longitudinal micro-data – altogether consisting of millions of person-months of individual and family level data – is useful for analyses of monthly labour market dynamics over relatively long periods of time, 25 years and more.

We make use of these data to estimate hazard functions describing transitions among the labour market states: *self-employed, paid employee* and *not employed.* Data on job tenure, for employed respondents, and on the date last worked, for those not employed – together with the date of survey responses – allow the construction of models that include terms reflecting seasonality and macro-economic cycles as well as the duration dependence of each type of transition. In addition, the LFS data permits spouse labour market activity and family composition variables to be included in the hazard models as time-varying covariates. The estimated hazard equations have been incorporated in the LifePaths microsimulation model. In that setting, the equations have been used to simulate lifetime employment activity from past, present and future birth cohorts. Simulation results have been validated by comparison with the age profiles of LFS employment/population ratios for the period 1976 to 2001.

KEY WORDS: Microsimulation; Censoring; Truncation; Employment dynamics.

1. INTRODUCTION

In recent years, there has been increased recognition of the importance of studying labour market dynamics using individual level (micro-) data. For this purpose, new panel surveys have been developed, for example, the Survey of Income and Labour Dynamics (SLID) (Statistics Canada 1998). But, existing LFS data (Statistics Canada 2002) provides a virtually untapped historical resource, in the form of many fragmentary event histories. From a conventional standpoint, the data currently comprises a time series of more than 300 cross-sectional surveys that were conducted monthly over more than 25 years. However, from a longitudinal perspective, those same data consist of about 6.5 million fragmentary event histories covering overlapping time intervals within the past quarter century and totalling over 34 million person-months of observation.

The analysis referred to in this paper was specifically directed towards development of hazard models to be incorporated in LifePaths (Statistics Canada 2001) – a micro-simulation model of the Canadian population. Further details on the LifePaths model are available from the Statistics Canada website at www.statcan.ca/english/spsd/index.htm.

The paper is organized in the following way. In section 2, we discuss some features of LFS data when reorganized as longitudinal records and we present three examples comparing estimates derived from the resulting longitudinal file with corresponding estimates from other sources. In section 3, we focus on the use of the data to model employment activity for LifePaths. There, we discuss the use of LFS micro-data in estimating hazard equations that describe employment dynamics. Finally, we present some illustrations of estimation results and a validation of LifePaths simulations that make use of the hazard equations.

2. LONGITUDINAL LFS DATA: DISTINGUISHING FEATURES AND PROOF-OF-CONCEPT

A longitudinal version of the LFS data was constructed by concatenating the monthly records of individual respondents into a file containing one record per respondent. Since an LFS respondent normally remains in the LFS sample for six consecutive months, we can obtain six-month histories for most respondents. These histories are not, by themselves, long enough for most longitudinal analyses. However, given the overlapping rotation groups that are part of the LFS design, these six-month fragments may be used in analysis of the experiences of employment cohorts over decades. (In line with the focus of the analysis below, we use the term "cohort" to refer to a relatively homogeneous group for all of whom a specified initial event has occurred. Thus, an "employment cohort" might refer to all persons who started a new job within a specified time period or, more narrowly, to all of those who started their third job

[1] Geoff Rowe and Huan Nguyen, Socio-Economic Analysis and Modeling Division, Analysis and Development Branch, Statistics Canada, Ottawa, Ontario, Canada, K1A 0T6.

within a specified time period. The data available from the LFS determines how narrowly such a cohort can be defined here).

Figure 1, which illustrates some characteristics of the LFS data after they are formed into longitudinal records, focuses on changes in employment status for the employment cohort who started a job in January 1976. Respondents who were members of this cohort and who entered the sample through rotation 1 contribute data on the first six months, from January 1976, when the job started, to June 1976, when they left the LFS sample. For respondents from rotation 2, the six-month longitudinal data window shifts right one month (starting and ending one month later than those given by rotation 1). The overlapping data windows of respondents from subsequent rotations evolve similarly. Thus, the longitudinal LFS data can be seen as a combination of overlapping sets of panel data, in which respondents from the same rotation constitute a conventional data panel.

Successive six-month fragments of longitudinal LFS data can be combined to provide successive estimates of cumulative attrition from an initial employment cohort and, further, to identify new cohorts defined in terms either of a new job or of a period without employment. Thus, over the long term (currently up to 25 years), many different samples of individuals can contribute information about the same employment cohort observed at different points in time. Even so, month-to-month changes are observed largely from the same sample of individuals. The two shaded areas in Figure 1 illustrate this. The respondents from each of the rotations 2-5 contribute data for both the May-June and the June-July intervals.

This is not the first attempt to use LFS data longitudinally. Stasny (1986) and Lemaître (1988) studied errors in the estimation of "gross flows" between labour force states (*employed*, *unemployed* and *not in the labour force*) over intervals of one month. Lemaître found that problems arose both because of response errors and because "Labour Force Survey concepts, designed for cross-sectional purposes, tend to "create" flows when consecutive months' responses are linked". (Examples include the treatment of *on-call workers* and of the *self-employed without a business).* Nevertheless, he concluded, "Administrative data have shown that not all sub-groups of status changers are seriously overestimated". Kinack (1991) examined the longitudinal consistency of responses to questions on job search activity that were used to distinguish between the categories *unemployed* and *not in the labour force*. He found substantial inconsistency, particularly when associated with proxy responses from different proxy respondents. These studies have shown that focusing on transitions between the categories *employed* and *not employed* (*i.e.*, without distinguishing between *unemployed* and *not in the labour force*) could help reduce the impact of response error.

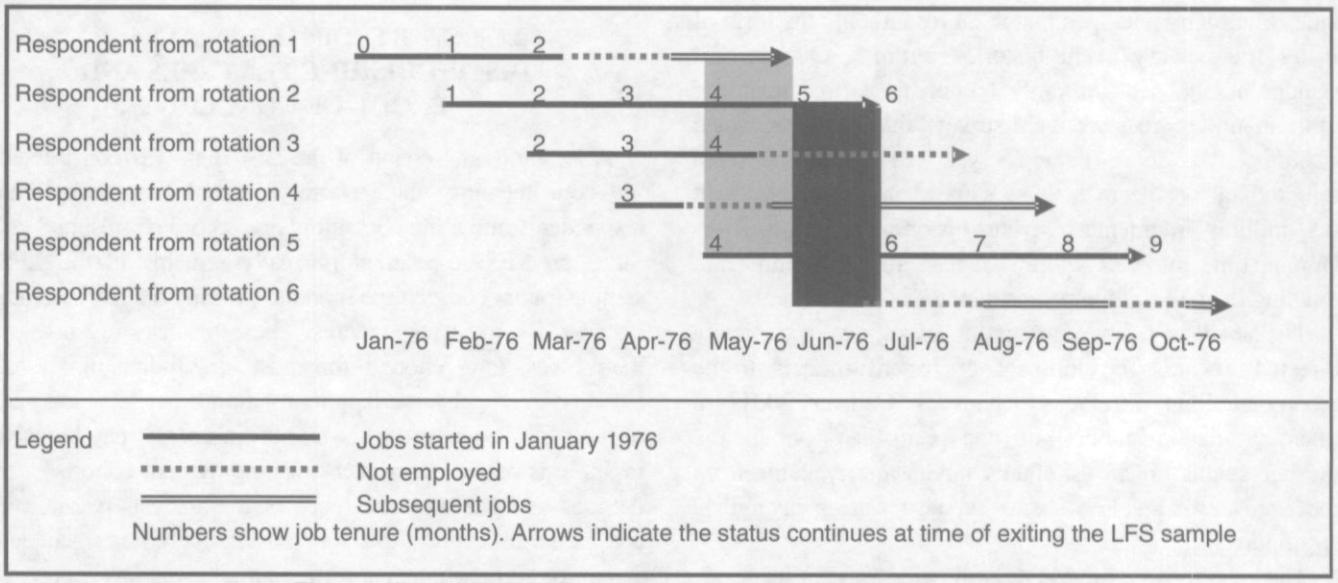


Figure 1. Illustration of LFS fragmentary data on cohort starting jobs in January 1976

Cross-sectional LFS data have previously been used to estimate frequencies of job hiring and job separation over monthly intervals (Lemaître, Picot and Murray 1992). In that case, hiring was directly observed from the frequency of reported job-tenures of one month or less, while separation was determined residually using aggregate estimates of employment change together with the estimates of hiring. Cross-sectional LFS data have also been used to calculate and compare duration statistics for synthetic-cohorts. For example, Corak and Heisz (1995) use retention rates from a single time interval to represent a hypothetical cohort's experience. Synthetic-cohort retention rates were obtained using the numbers of employed LFS respondents reporting job tenure "t" in month "m" together with those reporting tenure "$t+1$" the next month. Such uses of cross-sectional data have certain limitations. In particular, because the movement of individuals is not directly observed, destination states are unknown. (Although we may estimate the proportion that separated from a job, we can not estimate the proportion of those that became unemployed rather than dropping out of the labour force or beginning another job immediately). Nevertheless, a time series of synthetic-cohort statistics – for example, the proportions of jobs that might last a certain duration – can serve as an index that is sensitive to changing labour market conditions.

2.1 Proof-of-Concept: Selected Examples of Longitudinal Data Validation

The LFS data were not intended to be used longitudinally and problems can arise with such use (Stasny 1986; Lemaître 1988; Kinack 1991). Consequently, it is important to verify, for each analysis individually, that valid estimates can be obtained by month-to-month comparison of longitudinal responses. We present three examples of the verification of LFS longitudinal estimates below. In Figure 2, we compare estimates of the annual number of job separations in Canada from 1976 to 1995 (separations of all types, permanent and temporary) based on LFS data and on administrative data. The latter are based on Records of Employment (ROE) issued by employers at the time of job separation for Employment Insurance purposes (Statistics Canada 1998).

As may be seen, the number of transitions determined by month-to-month comparison of LFS data corresponds closely to the number from ROE data. Still, there are differences between the two series. Some of these differences could arise because of differences in coverage between the LFS and administrative data, as well as periodic changes in the LFS design or questionnaire. Another source of difference could arise because our counts based on LFS data neglect job separations of multiple job holders who remained employed in at least one job (*i.e.*, we counted only main-job changes). Nevertheless, we regard the degree of agreement between the LFS and administrative data as close enough to justify further analysis of the LFS micro-data. Both data sources imply that the annual rate of job separations was high: based on ROE data between 1978 and 1995, the average annual job separation rate for males was over 38 percent of annual person-jobs. Further analysis of the LFS micro-data can shed light on these dynamics.

Figure 3 goes further in the validation of employment dynamics, comparing "job survival" probabilities for males and females who started a job in 1993, as estimated from the LFS data and from SLID. (Note that 1993 corresponds to the first year of SLID data).

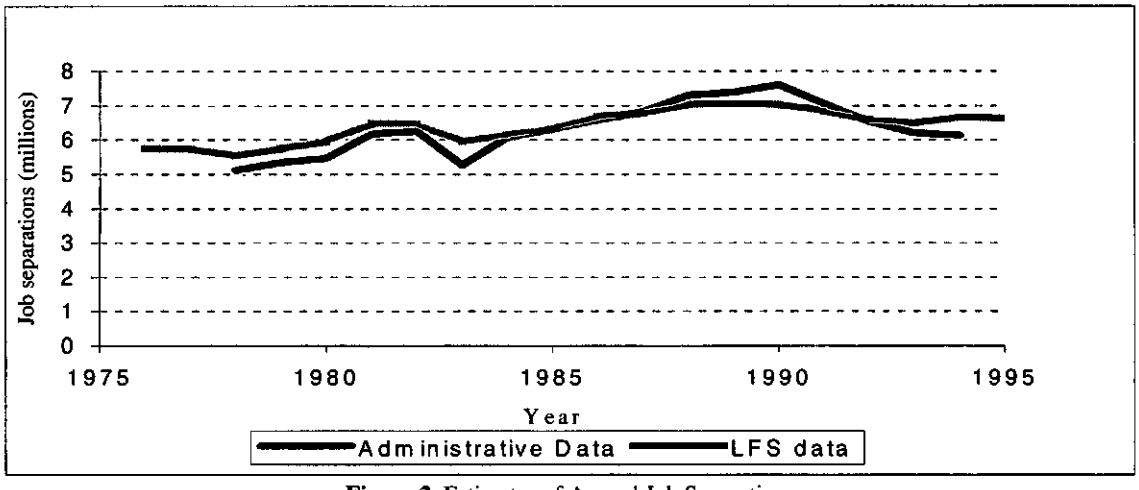


Figure 2. Estimates of Annual Job Separations

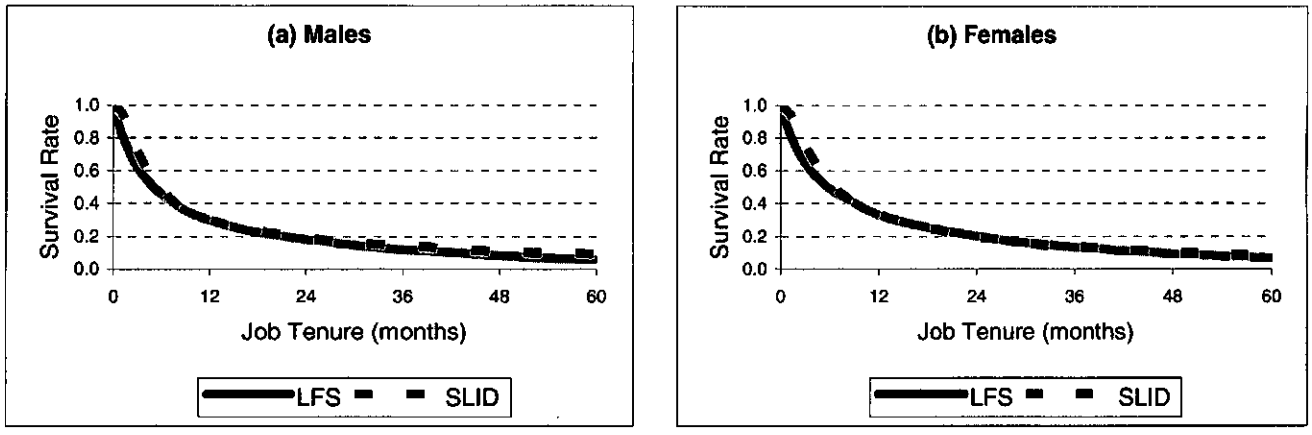


Figure 3. 'Job Survival' Probabilities of the Cohort Starting Jobs in 1993: Comparison of Estimates Based on LFS and SLID

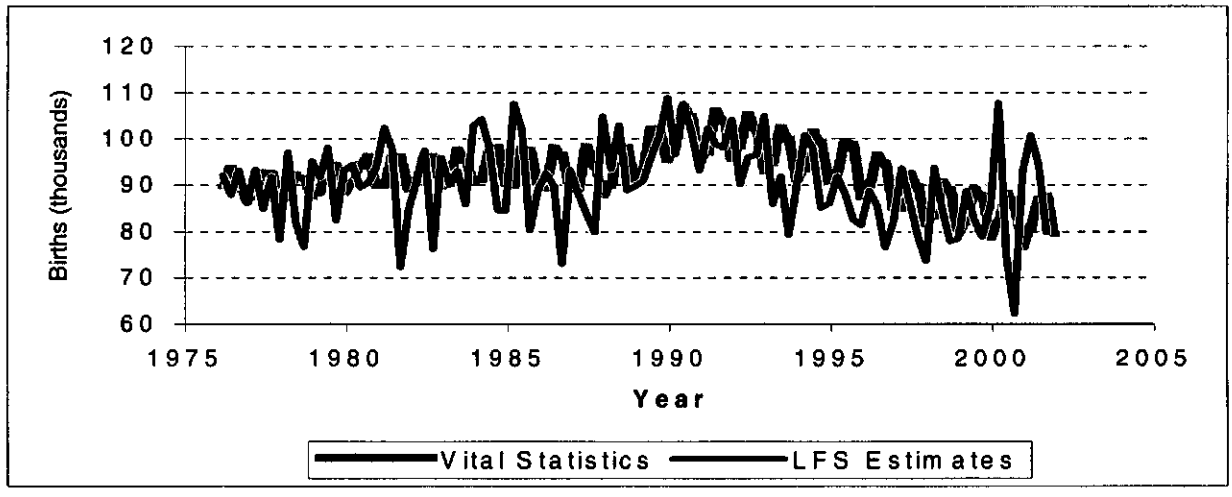


Figure 4. Estimates of Births in Canada by Quarter, 1976 - 2001

The "job survival" probabilities were estimated from LFS data by the chained product of average retention rates derived from monthly main-job separation rates over the period 1993 to 1998. Survival probabilities from the SLID data were estimated in a similar manner using the reported job tenure and dates of job end. Both survival curves display the same characteristic shape; showing relatively high attrition for jobs of duration less than a year, but with much lower attrition rates at job tenures of one to five years. There are discrepancies between the estimates for durations of about six months or less, which may be related to the one-year recall period of SLID interviews and to the restriction of LFS job-tenure data to main-jobs. However, over periods as long as five years, the LFS and SLID provide very similar estimates. And, with the available LFS data, we can track some employment cohorts for as long as 25 years after the employment spell began.

A final illustration of effective longitudinal use of LFS data involves month-to-month comparison of the number of children aged less than one year as reported by female economic family heads or by the spouse of a male head. A infant child that is newly reported by a woman aged between 15 and 50 likely signifies the birth of a child. In order to make direct comparisons between these LFS estimates and vital statistics, we made some straight-forward adjustments to account for the proportion of births occurring to other women living in economic families (*e.g.*, teen lone parents living with their parents) and for births in the Yukon, NWT and Nunavut. A comparison of the resulting LFS monthly estimates of births with the corresponding counts of births registered in vital statistics (Figure 4) demonstrates that the LFS estimates follow secular trends in fertility as well as capturing some of the month-to-month fluctuation in births. Taken together, these three examples indicate that – with careful attention to survey coverage, survey concepts and the possibility of response error – the LFS can provide useful longitudinal micro-data.

3. USING LONGITUDINAL LFS MICRO-DATA FOR MODELING EMPLOYMENT ACTIVITY IN LIFEPATHS

This section focuses on the use of the LFS data to simulate employment activity in LifePaths. Currently, LifePaths uses a 3-category classification of employment status – employee (E), self-employed (SE), and not employed (NE). We have not analyzed transitions involving unemployment. (Unemployment is a complex state requiring additional questions to ascertain and so, as noted above, unemployment transitions are particularly subject to response error).

There are six transitions that can result in a change in employment status (as represented in Figure 5). LifePaths models all of these transitions. In addition, job changes that do not appear to involve an interruption of employment are also modeled by LifePaths (denoted here as $E \Rightarrow E$). The LFS micro-data were used to estimate hazard equations for each of these seven transitions. The estimated coefficients of these equations became parameters in the LifePaths "Career Work" module. Below we discuss some technical issues that arise due to the limitations of the LFS data, followed by an illustration of the estimation results and then of a simulation outcome.

The fragmentary nature of these data poses a challenge for analysis. An important question is whether there are unavoidable biases that result from their fragmentary nature. In general, the answer is that the limitations of these data can be accounted for and potential sources of bias can be avoided with careful analysis.

3.1 Censoring and/or Truncation of Event Histories

One source of concern for an analyst of these data is the absence of retrospective employment information other than the length of the *current* employment spell. We might think of individual employment histories as consisting of a (largely unobserved) succession of contingent employment states (illustrated in Figure 6) with transitions among these states reflecting the process of career development. Thus, given only the transitions observable within the LFS window, the transition rates that can be estimated will inevitably involve pooling data from respondents who have had markedly different prior careers. In contrast, panel surveys like SLID, collect retrospective data at the first interview that, although limited, at least permits some experience rating of respondents in terms of previous extended work interruptions or periods of part-time work.

Another concern, illustrated in Figure 6, is that LFS employment spell durations may be left-truncated and/or right-censored. Right-censoring refers to the circumstance in which a spell ceases to be observed or a respondent ceases to be at risk without a transition occurring of the type being studied. This happens either (1) because the respondent's household "rotated out" of the LFS sample before any transition occurred, or (2) because another transition occurred that was not of the type under active study. Similarly, these data are frequently left-truncated. This refers to the circumstance in which the beginning of a spell is unobserved, because it happened before the respondent's household "rotated in" to the LFS sample. (These data are left-truncated rather than left-censored, because respondents provide the information necessary to determine the elapsed duration of the current spell at the time of the first interview). Since both truncation and censoring are generally independent of employment event processes, neither should lead to bias in the estimation of transition probabilities, if properly accounted for in the likelihood function.

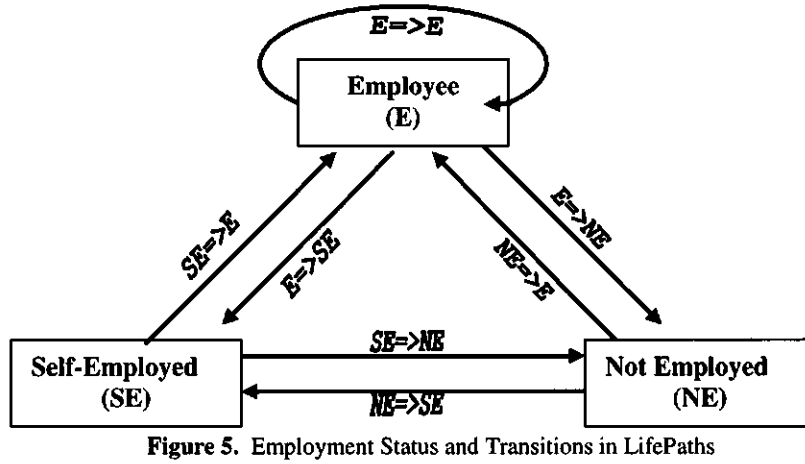


Figure 5. Employment Status and Transitions in LifePaths

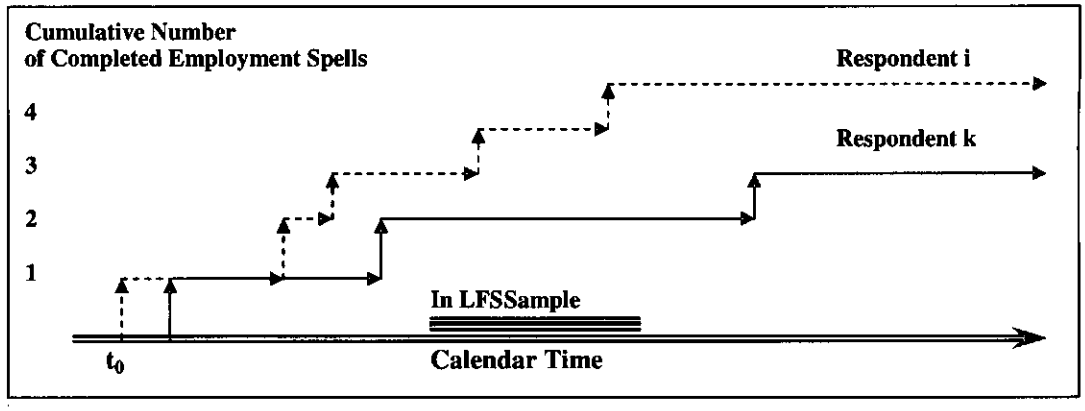


Figure 6. Recurrent Events and Employment Spell Durations Observable within the LFS Sample Window

The combination of full and partial information provided by left-truncated and right-censored data can be represented in a conditional likelihood (Wang 1991). In a competing risks framework, the likelihood of an employment transition type j involving respondent i may be expressed in terms of the spell duration observed k months after i was first observed to be at risk of transition j. Let t_i denote the year and month of the LFS interview in which i's current employment state was first observed (*i.e.*, often the first interview). Based on information collected at each interview, we can determine the length of the current spell of employment or spell not employed (m_{t_i}). Then $m_{t_i+k} = m_{t_i} + k$ would denote the elapsed spell duration in the state as assessed k months after the first observation – assuming no intervening events – and the likelihood of a transition of type j (*i.e.*, L_{j,t_i+k}) can be expressed in terms of m_{t_i+k}. Terms in the likelihood function comprise: the probability density of durations leading up to transitions of type j ($f_j(m_{t_i+k})$), the corresponding cumulative probability ($F_j(m_{t_i+k})$), a binary variable indicating whether or not censoring has occurred (C_{j,t_i+k}), and a further binary variable indicating whether or not the current spell was left-truncated ($LT_{i,j}$). Note that, in the competing risks framework, the density $f_j(m_{t_i+k})$ relates to a latent variable – the waiting time leading specifically to transition j – and that we must assume there is one such density for each competing event. In principle, the completed spell duration (observed when a transition occurs) will correspond to the minimum of competing, latent waiting times.

To account for left truncation, the likelihood is expressed in terms of conditional probabilities given the spell duration first observed (m_{t_i}): these probabilities take the form either of conditional probabilities evaluated at the time of an observed transition ($f_j(m_{t_i+k} | m_{t_i})$) or of conditional probabilities of surviving – without the occurrence specifically of transition j – to the observed duration ($1 - F_j(m_{t_i+k} | m_{t_i})$), depending on whether or not censoring has occurred.

$$L_{i,j,t+k} = f_j\left(m_{t_i+k} \middle| m_{t_i}\right)^{1-C_{j,t_i+k}} \left(1 - F_j\left(m_{t_i+k} \middle| m_{t_i}\right)\right)^{C_{j,t_i+k}}$$

$$= \frac{f_j\left(m_{t_i+k}\right)^{1-C_{j,t_i+k}} \left(1 - F_j\left(m_{t_i+k}\right)\right)^{C_{j,t_i+k}}}{\left(1 - F_j\left(m_{t_i}\right)\right)^{LT_{i,j}}}. \qquad (1)$$

This likelihood accounts for all of the information we have regarding the specific risk of transition j and can incorporate the effect of other competing risks by treating them as censoring events that are in addition to censorship by "rotating out" of the sample. Competing risks problems are commonly formulated in terms of such latent waiting times, especially in epidemiology and biostatistics, but also in economics (*e.g.*, Heckman and Honoré 1989). However, while providing a mathematically convenient motivation for the likelihood, the approach has been criticized "on the basis of unwarranted assumptions, lack of physical interpretation and identifiability problems" (Prentice, Kalbfleisch, Peterson, Flournoy, Farewell and Breslow 1978).

The conditional likelihood (1) can be approximated by a Poisson likelihood (Holford 1980; Laird and Olivier 1981), thereby also acknowledging the discreteness of the data (*i.e.*, transitions are generally "observed" in the one month interval between successive interviews). Equation (1) can be re-expressed in terms of a binary variable (Y_{j,t_i+k}) that represents occurrence or non-occurrence of a transition in a particular time interval (note that $Y_{j,t_i+k} = 1 - C_{j,t_i+k}$). Then, Y_{j,t_i+k} is treated as a Poisson random variable having an expected value equal to the hazard "h_{j,t_i+k}" which is assumed piecewise constant. Under this model, the contribution from i to the log-likelihood over n periods (using $h_{j,t_i+k} = f_j(m_{t_i+k}) / (1 - F_j(m_{t_i+k})) = -\partial \ln(1 - F_j(m_{t_i+k})) / \partial m_{t_i+k}$ together with (1)) is approximately:

$$\ln(L_{ij}) \approx \sum_{k=1}^{n} \left[Y_{j,t_i+k} \ln(\hat{h}_{j,t_i+k}) - \hat{h}_{j,t_i+k} \right]. \quad (2)$$

It is common practice to account for a complex survey design by means of a "pseudo" likelihood that incorporates the survey weight. Maximizing the "pseudo" likelihood corresponds to minimization of a weighted sum of deviance terms (*i.e.*, terms representing the difference between estimated likelihood contributions and their maximum possible values). Thus, the full-sample, conditional log-likelihood for transition j may be transformed into a weighted deviance D_j (note that W is derived from the survey weights and, since transitions are typically identified by comparing employment states between interviews, we use averages of consecutive cross-sectional survey weights to obtain W):

$$D_j \approx -2 \left(\begin{array}{l} \sum_i \left[\sum_{k=1}^{n_i} W_{t_i+k} Y_{j,t_i+k} \ln\left(\hat{h}_{j,t_i+k}\right) \right] \\ + \sum_i \left[\sum_{k=1}^{n_i} W_{t_i+k} \left[Y_{j,t_i+k} - \hat{h}_{j,t_i+k} \right] \right] \end{array} \right). \quad (3)$$

In the analysis of each transition type j, we treat other events (*i.e.*, non-j events occurring to the same population-at-risk) as censoring, and so the deviance for a set of such events will be the sum of component deviances (*i.e.*, if the overall hazard is the sum of competing hazards, then the competing risks may be treated as independent (Prentice *et al.* 1978)).

A more direct motivation of the same deviance takes Poisson processes as its starting point (Borgan 1984; Andersen 1985; Andersen and Borgan 1985; Lawless 1987), rather than starting with postulated event-specific, latent, duration densities like $f_j(m_{t_i+k})$. In this case, we can model sampled multivariate counting processes that represent the number of occurrences of each specific transition in time intervals $[t_0, t)$. Sample counting processes, represented by the step functions in Figure 6, are observable counterparts of cumulative hazard functions. The assumption that the underlying hazard functions are approximately piecewise constant leads directly to the Poisson deviance as an approximation (Lindsey 1995). To limit bias, the principal concerns are that the population-at-risk can be identified, that censoring or truncation mechanisms are conditionally independent of the underlying employment processes and that the intervals over which hazards are assumed constant are not too large.

It is possible to obtain simple averaged estimates of employment hazard functions (such as those displayed in Figure 3) by implicitly splicing together all available information on members of a defined cohort from the longitudinal LFS samples. (That is, maximizing likelihood (1), but without considering any covariates). Making allowance for censoring and truncation in this way is a relatively simple example of such problems compared with the more complex observation schemes considered by Alioum and Commenges (1996). This implicit splicing of information is apparent in the deviance (3) which has two components: the first component is non-zero only at observed transitions, while the second component reflects the weighted differences between cumulative events and cumulative hazards (accumulated over all durations prior to the events or to censoring times). To the extent that the LFS cross-sections are representative samples for each reference week, then – taken together – they will provide an accurate estimate of the numbers of events occurring over the "life" of an employment cohort. Similarly, within samples from employment cohorts, we can expect to find left-truncated and right-censored respondent spells that might fill-in the missing prior histories of those left-truncated spells that terminate with a transition. As such, the first component of the deviance will accurately reflect whether hazard estimates tend to be large over periods where observed events are frequent. And the second component, summed over all respondent-months, may have a value similar to that which we might have obtained had there been no left-truncation. So, for data as extensive as these, the conditional likelihood may be almost equivalent to an unconditional likelihood.

3.2 Estimating Employment Transition Hazard Equations

Patterns of employment transition differ significantly among different demographic groups. For example, full-time students are most active in the labour market during their summer break, whereas the maternity leave that an employed pregnant woman takes may be largely determined by Employment Insurance regulations. Accordingly, LifePaths distinguishes among the following groups and models their employment activities separately:

- Those who are full-time students;
- Those who have just graduated or left school and are in a transition to an after-school job;
- Pregnant women for whom a maternity-leave may apply;
- Those who are in prime ages of employment; and
- Older workers in transition to retirement.

We discuss here only the estimation for the fourth group, comprising individuals who are in what is referred to in LifePaths as their "career employment" phase (the most important phase in terms of impact on the economy). Particulars for the other groups are available from the Statistics Canada website noted above.

For implementation in LifePaths, our hazard model uses a log-linear form of regression equation – one equation for each of the 7 transitions and for each sex separately, giving a total of 14 equations:

$$E\left(Y_{j,t_i+k}\right) \approx \hat{h}_{j,t_i+k} = \exp\begin{pmatrix} \hat{g}\left(m_{j,t_i+k}\right) \\ + X_{j,t_i+k}\,\hat{\beta} \end{pmatrix} \qquad (4)$$

where $E(\,)$ is the expectation operator, $g(m)$ is a log-linear spell duration spline, X is a vector of time-varying covariates and β is a vector of regression coefficients. The term $g(m)$ corresponds to a piecewise Weibull baseline hazard, which, in our specification, distinguishes employment transition risks at durations of less than a year from risks at durations of more than a year. The covariates, X, include variables representing individual age, education, province of residence, presence of children by age group, spouse's employment status, calendar month and calendar year, as well as interactions among some of these factors. Final estimates of β and $g(m)$ minimize the deviance (3).

The only example of detailed results that we present here involves the mutual influence of husband's and wife's employment status on each other's respective transition hazards. Figure 7 compares coefficient estimates from the seven equations that correspond to the seven transitions we specified. The two panels correspond to the separate sets of equations for males and females. The category "no spouse present" was treated as the reference category and the spouse's employment status was classified into "with paid employment", "self-employed", and "not employed". The estimated coefficients are presented here in terms of risk relative to the reference group. Thus, with other covariates controlled, the hazard of becoming self-employed for female employees whose husbands are self-employed is about 2.5 times higher than the hazard of their counterparts who do not have a spouse (see tallest bar in the top panel).

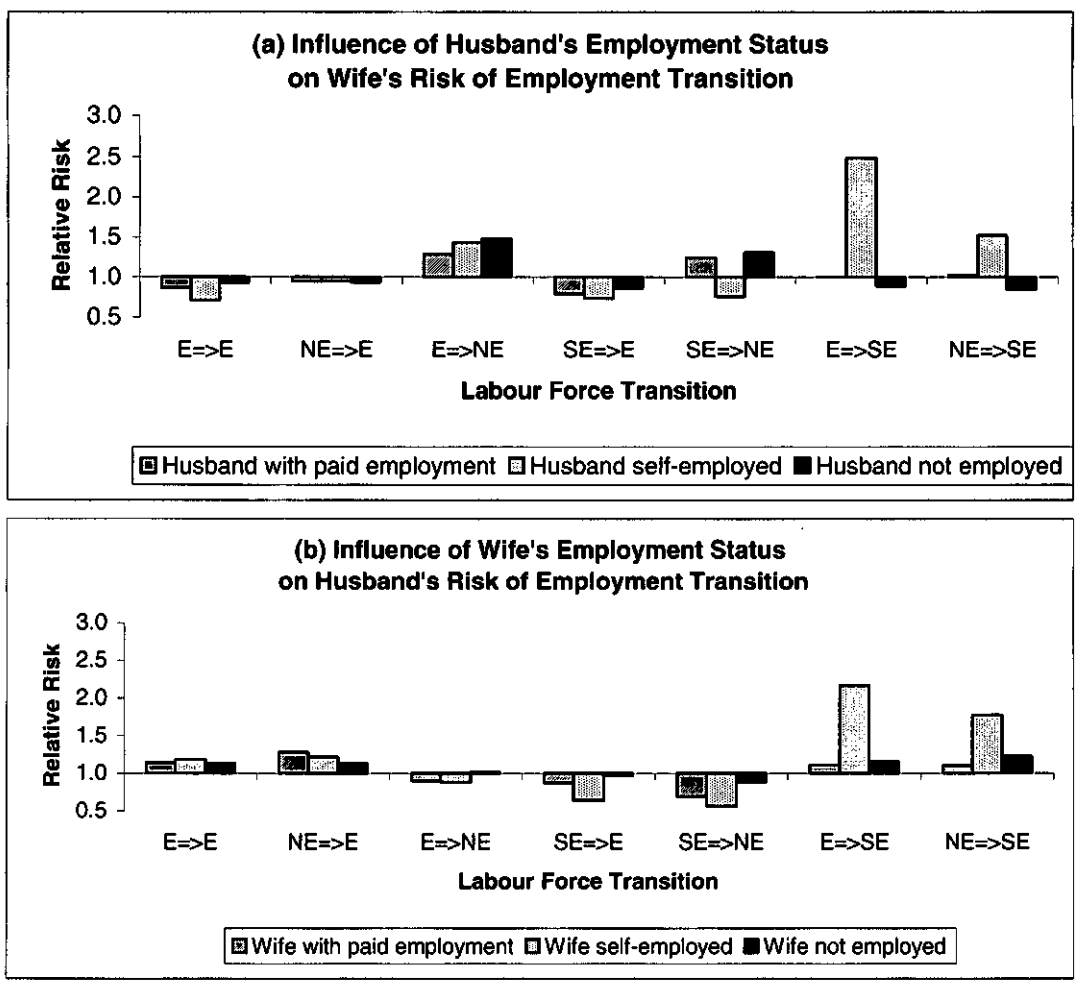


Figure 7. Impact of Spouse's Employment Status on Employment Transition Risks

Figure 7 shows that the very presence of a spouse can work in opposite directions for males and females. The most frequent transitions for both sexes are $E \Rightarrow E$, $NE \Rightarrow E$ and $E \Rightarrow NE$. For females, the first two of those transitions are less likely to occur to married women than to single women, while the transition to "not employed" is more likely. (The presence of children is not the reason for this, as their presence is accounted for by other terms in the equation). For males, the pattern is reversed. Thus, these results appear consistent with conventional gender roles. However, taking account of the magnitudes of these relative risks, we are not given the impression that gender roles have a particularly strong influence after the influence of other variables is credited.

Figure 7 reveals another conspicuous pattern. First, the relative risks of a transition into self-employment, for spouses with husbands/wives in self-employment, stand out as the highest among all other transitions. In addition, spouses with husbands/wives in self-employment have the lowest relative risks of a transition out of self-employment. Thus, self-employment status seems to be mutually reinforcing within families. These observations are consistent with forms of joint self-employment involving a family business (*e.g.*, a corner store) or involving endogamy among professionals (*e.g.*, lawyers marrying other lawyers).

4. FROM ESTIMATED PARAMETERS TO THE SIMULATION RESULTS: AN ILLUSTRATION

Our example of the role of spouse's employment status points to the need for family context in the simulation of employment activities. It is a challenge for LifePaths to integrate these relationships into the simulation process. For example, if individual education progression or the effects of education on employment transitions are not modeled appropriately and accurately, then the consequences will cascade from direct education-employment relationships to a chain of indirect impacts, involving relationships between education and marriage, fertility, interprovincial migration, *etc.* These impacts would then spill over to the simulated spouse, as indicated above. It is not difficult to see that, unless these relationships are specified appropriately and the parameters are estimated with reasonable accuracy, bias would be spread over a wide range of simulated outcomes.

An overall validation of the LifePaths employment hazard equations was obtained by comparing simulated annual average employment/population ratios with direct cross-sectional estimates from the LFS. The simulated employment/population ratios were obtained from a synthetic population whose members were exposed appropriately to one or other of the seven types of employment hazards over the course of each simulated year. The simulated employment/population ratios were calculated from the resulting annual person-years of employment in the synthetic population: that is, these ratios are an outcome of simulated flows into and out of employment. The simulations necessarily involved generating appropriate distributions of covariates that in turn determine the distributions of employment transition hazards. As may be seen in Figure 8, LifePaths accurately reflects the age patterns of female employment in both 1976 and 2001 and correspondingly accounts for the dramatic change observed in those age patterns over the past quarter century.

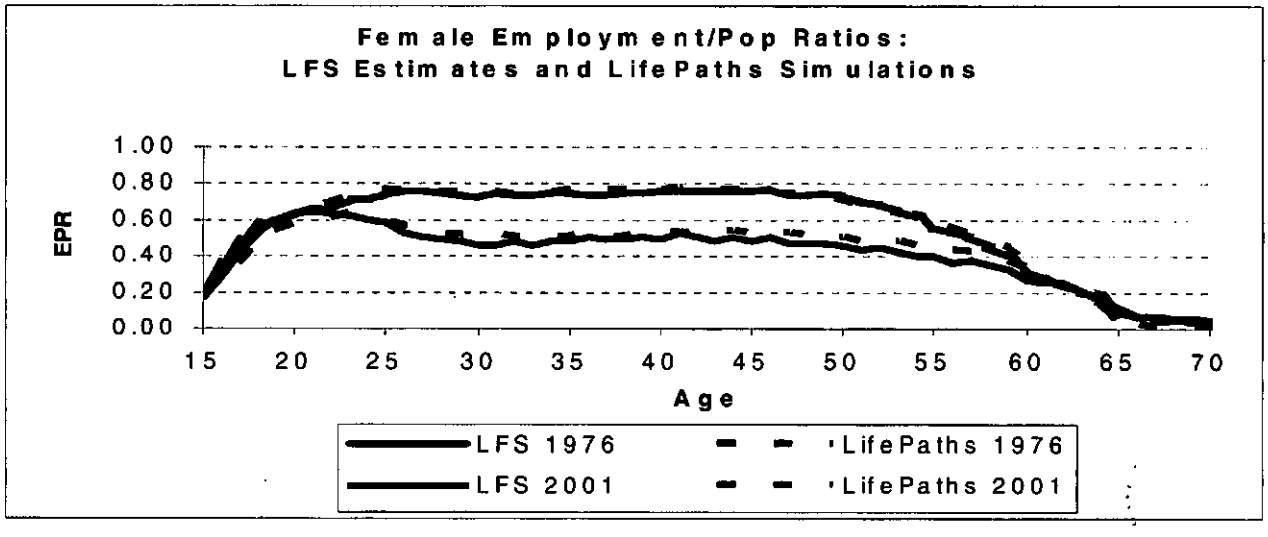


Figure 8. Validating hazard equations using LifePaths

5. CONCLUSIONS

We have demonstrated that the LFS data – when organized into the fragmentary event histories collected over the six-month periods that most respondents spend in the sample – represents a significant longitudinal micro-data asset. There is sufficient sample and breadth of content to provide for important analysis of labour market dynamics and, conceivably, of demographic processes such as fertility. Moreover, the data is monthly and spans more than a quarter century, so that analysis based on it has uninterrupted time depth that is unique in Canada.

In our main application (employment transitions), other results (not reported here) appear to confirm the influence of a range of explanatory variables on an individual's chances of an employment transition. These covariates include age, job tenure (or duration not employed), educational attainment, presence of young children (especially for women), province of residence, seasonality, and business cycles. However, this work is still in its initial stage and, to date, our approach to inference has been informal. Future work will involve extending and refining our models and establishing a more rigorous basis for evaluation of the models.

REFERENCES

ALIOUM, A., and COMMENGES, D. (1996). A proportional hazards model for arbitrarily censored and truncated data. *Biometrics*. 52, 512-524.

ANDERSEN, P. (1985). Statistical models for longitudinal labor market data based on counting processes. In *Longitudinal Analysis of Labor Market Data*. (Eds. James J. Heckman and Burton Singer). Cambridge University Press, Cambridge.

ANDERSEN, P., and BORGAN, Ø. (1985). Counting process models for life history data: A review. *Scandinavian Journal of Statistics*. 12, 97-158.

BORGAN, Ø. (1984). Maximum likelihood estimation in parametric counting process models, with applications to censored failure time data. *Scandinavian Journal of Statistics*. 11, 1-16.

CORAK, M., and HEISZ, A. (1995). The duration of unemployment: A user guide. Research Paper Series No. 84, Analytical Studies Branch, Statistics Canada.

HECKMAN, J.J., and HONORÉ, BO.E. (1989). The identifiability of the competing risks model. *Biometrika*. 76(2), 325-330.

HOLFORD, T.R. (1980). The analysis of rates and survivorship using log-linear models. *Biometrics*. 36, 299-305.

KINACK, M. (1991). Measuring data quality with longitudinal data. *1991 Proceedings of the Section on Survey Research Methods*, American Statistical Association. 514-519.

LAIRD, N., and OLIVIER, D. (1981). Covariance analysis of censored survival data using log-linear analysis techniques. *Journal of the American Statistical Association*. 76, 231-240.

LAWLESS, J.F. (1987). Regression methods for Poisson process data. *Journal of the American Statistical Association*. 82, 808-815.

LEMAÎTRE, G. (1988). The measurement and analysis of gross flows. Working Paper, Labour and Household Surveys Analysis Division, Statistics Canada.

LEMAÎTRE, G., PICOT, G. and MURRAY, S. (1992). Workers on the move: An overview of labour turnover. *Perspectives on Labour and Income*. 4(2), Statistics Canada.

LINDSEY, J.K. (1995). Fitting parametric counting processes by using log-linear models. *Applied Statistics*. 44, 201-212.

PRENTICE, R.L., KALBFLEISCH, J.D., PETERSON, A.V., FLOURNOY, JR. N., FAREWELL, V.T. and BRESLOW, N.E. (1978). The analysis of failure times in the presence of competing risks. *Biometrics*. 34, 541-554.

STASNY, E.A. (1986). Estimating gross flows using panel data with nonresponse: An example from the Canadian Labour Force Survey. *Journal of the American Statistical Association*. 81, 42-47.

STATISTICS CANADA (1998). Permanent layoffs, quits and hirings in the Canadian Economy: 1978-1995. Catalogue # 71-539-XIB.

STATISTICS CANADA (1998). Survey of Income and Labour Dynamics: A Survey Overview. Catalogue # 75F0011XPB, http://www.statcan.ca/english/freepub/75F0011XIE/free.htm.

STATISTICS CANADA (2001). The LifePaths Microsimulation Model: An Overview. http://statcan.ca/english/spsd/LifePaths.htm.

STATISTICS CANADA (2002). Guide to the Labour Force Survey. Catalogue # 71-543-GIE, http://www.statcan.ca/english/IPS/Data/71-543-GIE.htm.

WANG, M.-C. (1991). Nonparametric estimation from cross-sectional survival data. *Journal of the American Statistical Association*. 86, 130-143.

Survey Methodology, June 2004
Vol. 30, n° 1, pp. 115-124
Statistics Canada

Contact and Cooperation in the Belgian Fertility and Family Survey

MARC CALLENS and CHRISTOPHE CROUX[1]

ABSTRACT

Combining response data from the Belgian Fertility and Family Survey with individual level and municipality level data from the 1991 Census for both nonrespondents and respondents, multilevel logistic regression models for contact and cooperation propensity are estimated. The covariates introduced are a selection of indirect features, all out of the researchers' direct control. Contrary to previous research, Socio Economic Status is found to be positively related to cooperation. Another unexpected result is the absence of any considerable impact of ecological correlates such as urbanity.

KEY WORDS: Nonresponse; Multilevel analysis; Fertility and Family Survey.

1. INTRODUCTION

The aim of this paper is to empirically assess the relative importance of correlates of contact and cooperation rates in the Belgian Fertility and Family Survey (FFS Belgium 1991).

The conceptual and theoretical nonresponse framework used in this paper has been proposed by Groves and Couper (G&C 1998). In their view, nonresponse arising from noncontact is directly influenced by survey design features such as the number and the timing of calls. Conditionally on these survey design features, other important features such as physical impediments of the housing units and accessible-at-home patterns of the would-be respondents, which are indirectly measured by various social environmental and socio-demographic attributes, also play an important role. The decision to cooperate or to refuse is primarily regarded as a direct function of a dynamic social communicative process between the interviewer and the interviewee. Survey design, main interviewer, sample person and social environment characteristics are considered to have only an indirect influence on cooperation rates.

We use both individual level and municipality level data from the 1991 Census data, matched to the fieldwork outcome variable for nonrespondents and respondents of the 1991 Belgian FFS. In this survey, individuals are the sampling units. It is a face-to-face survey with low noncontact (4%) and moderate refusal rates (22%). We consider our data to be hierarchically nested with sample units at the lower and municipalities at the higher level. Including covariates at both levels, multilevel logistic regression models for contact and cooperation propensity are estimated. The covariates are a selection of indirect features, all out of the researchers' direct control.

Some intriguing results are: (1) Socio Economic Status indicators like education are positively related to cooperation and (2) ecological factors including urbanicity are not correlated with nonresponse. This is in contrast with findings from previous US-based research.

2. A THEORY FOR CONTACTABILITY AND COOPERATION

The process of realising an interview consists of two major components: the process of contacting a sample person and dependent on contact, the process of cooperation with a survey request. An attractive multi-level theoretical framework for studying contactability and cooperation has been proposed by Groves and Couper (G&C 1998).

2.1 Contactability

Chronologically, the process of contacting a sample person comes first. Some sample persons are never contacted by interviewers and hence never make a decision about their survey cooperation. Relative to the process of cooperation, the process of contacting a sample person is quite simple.

G&C (1998) consider contactability to be a function of three factors: (1) whether there are any physical impediments that prevent interviewers to get in touch with the sample person, (2) when sample persons are at home and (3) when and how many times the interviewer tries to contact the sample person. The number and timing of calls

[1] Marc Callens, Population and Family Study Centre (CBGS), Markiesstraat 1, B-1000 Brussels, Belgium & Dept. of Applied Economics, Katholieke Universiteit Leuven, Naamsestraat 69, B-3000 Leuven, Belgium. E-mail: Marc.Callens@cbgs.be; Christophe Croux, Dept. of Applied Economics, Katholieke Universiteit Leuven, Naamsestraat 69, B-3000 Leuven, Belgium. E-mail: Christophe.Croux@econ.kuleuven.ac.be.

by the interviewer and the accessible-at-home patterns of the sample persons are the proximate causes of contactability. The accessible-at-home patterns of the sample person are affected by the presence of physical impediments (*e.g.*, telephone presence), socio-demographic attributes (*e.g.*, commuting times) and social environmental attributes (*e.g.*, crime). Also survey design features such as the length of the data collection period and the interviewer workload might have an influence on contact rates.

2.2 Cooperation

The central question in the survey stage following contact is why sample persons do or do not cooperate with the interviewer request. In the Groves-Couper model to study cooperation, the proximate causes of the decision to cooperate or to refuse lie at the level of the householder and his or her interaction with the interviewer. Another component in the theoretical framework of G&C (1998) is the set of survey design features, such as: the agency of data collection, advance warning of the survey request, topic saliency, *etc.*

G&C (1998) consider also two factors that are out of the control of the survey designer: influences of the sample person and social environmental influences. These variables are not considered to be direct causal influences on cooperation, but indirect measures of what are essentially social psychological constructs. Important theoretical constructs in this respect are: opportunity costs, social exchange and social isolation.

2.2.1 Opportunity Costs

The notion of opportunity costs implies that sample persons weigh the opportunity costs in agreeing to spend their time responding to a survey interview. An important ingredient in the opportunity costs theory is the amount of discretionary time for the sample person available to complete the survey. Those with less discretionary time are less likely to feel free to participate in a survey. Some indirect indicators for the amount of discretionary time are: the inverse of the number of adults in a household and (the amount) of labour force participation. Of course, there are also obligations away from employment tasks such as commitments to friends and relatives that also might raise the opportunity costs of a survey.

2.2.2 Social Exchange

Social exchange theory considers the perceived value of equity of long-term associations between persons or between a person and societal institutions (Blau 1964). Central to all conceptualisations of social exchange is the notion that, unlike economic exchange, all social commodities are part of an intuitive bookkeeping system in which debts (*e.g.*, obligations) and credits (*e.g.*, expectations) are taken into account (G&C 1998). The social exchange perspective can be applied whenever there is an ongoing relationship between the survey organisation and the sample person (*e.g.*, government surveys).

Those receiving fewer services from the government may – in considering the cumulative effect of multiple government contacts – feel less need to cooperate. Since government services are disproportional across socio-economic strata, indicators of Socio-Economic Status (SES) should reflect exchange influences on survey participation. However, a major problem with social exchange theory is that two alternative hypotheses between SES and cooperation might be deduced from it (G&C 1998). First, one can argue that lower SES groups may have the greatest indebtedness to the government for the public assistance they may receive. Higher SES groups feel far less that they owe any sort of repayment. In this perspective, the relationship between socio-economic status and cooperation propensity is a negative one. Alternatively, a curvilinear relationship between SES and cooperation may be hypothesised. The lowest SES groups may believe that they are disadvantaged routinely compared to more fortunate people. The highest SES groups feel themselves repeatedly targeted in terms of time and money but receive little in return. In such a hypothesis, both the highest and the lowest SES feel relatively deprived in the relationship with large-scale social institutions and tend to refuse survey cooperation.

2.2.3 Social Isolation

Closely related to the social exchange hypothesis is the social isolation hypothesis. Social isolates are out of touch with the mainstream culture of a society: they tend to behave in accordance with subcultural norms or in explicit rejection of those of the dominant culture. They are believed to be less likely to participate in a variety of social and political activities, including responding to surveys (Couper, Singer and Kulka 1997). In terms of SES, social isolation theory implies a positive relationship between SES and cooperation: lower SES groups are resentful of their dependence on the government, whereas higher SES groups have a greater sense of civic obligation. Such a positive relationship between SES and social isolation is opposite to the relationships predicted by social exchange theory.

Demographic indicators of social isolation are race, ethnicity, age and gender; with minorities, elderly and men in the role of the relatively isolated. Indicators of social isolation at the micro-level include whether the sample person lives in a single-person household, whether the sample person has any children, whether the sample person has moved recently and whether the sample person lives in a large multiunit structure.

2.2.4 Urbanicity

At the community level contextual factors such as urbanicity, population density, crime rates and lack of social cohesion are hypothesised to influence survey cooperation. Residents of rural areas tend to cooperate at a higher level compared to residents in towns. However, it is not clear which mechanism is responsible for this urbanicity effect which might be explained in terms of greater population density, higher crime rates and higher social disorganisation that are associated with life in urban areas. Population density is hypothesised to reduce cooperation through the experience of crowding. Fear of crime may produce an unwillingness to provide information to strangers. Finally, urban life is associated with social disorganisation, characterised by weakened local kinship and friendship networks and reduced participation in local affairs.

3. DATA AND METHOD

3.1 Data

In this study we make use of both aggregated and micro-level data of the Belgian 1991 Census linked to the response status for respondents and nonrespondents from the Belgian Fertility and Family Survey (FFS-Belgium 1991) held shortly after the Census operations.

3.1.1 The FFS Survey (1991)

The Fertility and Family Survey in Belgium was organised by the Population and Family Study Centre (CBGS), a Scientific Institute from the Flemish Government. This survey was carried out between April and October 1991, which is very close to the decennial census date: April 1 in the same year. The main focus of the FFS-project is on reproductive behaviour, to be seen however in the broader context of partnership and family history, and the interaction between employment and reproduction (Cliquet and Callens 1993; Callens 1995). The target population consists of men and women of Belgian nationality, born in the period 1951-1970 and with main residence in the Flemish Region of Belgium.

A two-stage cluster sampling design was used for men and women separately. In a first stage, municipalities were selected from various socio-economic strata (Vanneste 1989). In each selected municipality, individuals were selected at random. In this way 2,975 women and 1,989 men were selected to take part in the survey. A fieldwork method was used to compensate for non-response: stratified random substitution of nonrespondents of the target sample by persons selected from a reserve sample (Chapman 1983; Vehovar 1999).

The final sample size, *i.e.*, including the substitution operation, equals 4,776 persons (2,897 women and 1,879 men). In this study we make use of respondents and nonrespondent cases of both the initial target sample and the fieldwork substitution operation (N= 6,847).

Among both men and women, the nonresponse can be ascribed in 7 out of 10 cases to a refusal to participate in the survey. In 2 out of 10 cases, nonresponse is due to the fact that the persons selected could not be contacted, and in 1 out of 10 cases, an interview was impossible because of sickness, language difficulties or some other reason.

3.1.2 Matching 1991 Census Person-Level Data (1991)

Our primary source of information on both respondent and nonrespondent cases is provided by the 1991 Census.

In an effort to reconcile privacy concerns and scientific interests, we used a simple technique to make the matching of person-level Census data and survey data anonymous. We provided a dataset to the National Institute of Statistics (NIS) containing only the national identification number and the response status for each respondent and nonrespondent case. As a result of the matching operation by the NIS, we received a selection of the 1991 Census data enriched with only two survey variables: the response status variable and an indicator whether a sample person belongs to the base or substitute sample.

The 1991 Census individual level data we have at our disposal are: the individual form and the house unit form. The individual form contains information about: the place of residence, the nationality, the labour force activity status, the first marriage, the birth year of the children, education and professional activities. The house unit form includes information on the housing unit of the household such as: the type of housing unit, the number of housing units in the building, ownership, building period, the number of rooms and corresponding squared meters, the presence of a telephone and comfort indicators such as the number of bath rooms.

3.1.3 Contactability and its Determinants

To study the process of contactability, we ideally need data on the outcomes of all successive attempts to contact sample persons. In this study however, we do not have such detailed information at our disposal: we only know the final outcome of each survey request. Therefore, we can only study the probability of ever making contact with the sample person (coded 1 = contact and coded 0 = non-contact) and not whether it was easy or difficult to make contact. Sample persons that are known not to reside (anymore) on the sample address we do consider contacted. At 241 out of 6,847 sample units (3.52%), all contact attempts failed.

The data we use are measured at two levels: the individual level (n= 6,847) and the municipality level (n=123). At the sample person level, we consider three types of variables: physical impediments to contact sample persons, reasons for sample persons to be present in their homes and control variables.

As there are no direct interviewer observations of physical impediments available to us, we have to rely solely on indicators for physical impediments available in the Census data. Three variables are used: whether the housing unit is a single-family structure or not, whether the housing unit is large (more than 10 units) or not and whether the sample person has a telephone or not.

Determinants of at-home patterns in this study are: civil status (unmarried, married and divorced), age (20-24, 25-29, 30-34 and 35-39 years) and activity status (inactive *vs.* other). For women only, we also consider the number of children (0, 1, 2 and 3+). For those in the labour force we have also detailed information about: working part-time *vs.* working full-time, the number of weekly working hours (<21, 21-35, 36-42, >42 hours), employment status (employee *vs.* own-account), having a second job or not and working at home or not.

We also use two control variables: substitution (whether a sample person originates from the base target sample or from the substitution sample) and gender (whether a sample person comes from the female sample or from the male sample).

At the municipality level (n=123), we use five variables: population density (persons per square km for the residence of the sample person), urban status (the cities of Antwerp and Gent *vs.* other municipalities), percentage multi-unit structures (in quartile format: <7.13, 7.13-15.14, 15.14-27 and >27), percentage homes owner-occupied (in quartile format: <64.5, 64.5-71, 71-77.7 and >77.7) and percentage persons of minority race (in quartile format: <0.90, 0.9-2.22, 2.22-5.29 and >5.29).

3.1.4 Cooperation and its Determinants

We are interested in the probability of ever getting cooperation (coded 1 = cooperation and coded 0 = non-cooperation) conditionally on contact; not whether it was easy or difficult to get cooperation from the sample person. For 1,399 out of 6,606 contacted sample persons (21.18%), all attempts to get cooperation failed.

Again, the data we use are measured at two levels: the individual level and the municipality level. At the sample person level, we have indicators for the opportunity costs hypothesis, the exchange hypothesis and the isolation hypothesis. Substitution is used as a control variable.

Indicators for the opportunity costs hypothesis are: activity status (inactive *vs.* other), working part-time *vs.* working full-time, the number of weekly working hours (<21, 21-35, 36-42, >42 hours) and employment status (employee *vs.* own-account).

Indicators for Socio-Economic Status in our study are: the surface of the living rooms (in squared meters: <65, 65-84, 85-104, 105-124 and >125), the number of bathrooms (0, 1 and 2+) and educational level (primary, secondary – first stage, secondary – second stage, high – non-university and high – university level). Other exchange hypothesis indicators are: whether one receives a replacement income from the government or not and whether the house is owner-occupied or not.

Indicators for the social isolation hypothesis are: gender, civil status (unmarried, married and divorced), age (20-24, 25-29, 30-34 and 35-39 years), single-family structure of the housing unit and for women only: the number of children (0, 1, 2 and 3+) and the presence of children under the age of five years. Finally, substitution is included as a control variable.

At the municipality level, we use the same five variables as in section 3.1.3: urban status, population density, percentage multiunit structures, percentage owner-occupied and percentage persons of minority race.

3.2 Method of Analysis

3.2.1 Bivariate χ^2 -Test

In a first exploratory series of analyses of the correlates of contactability and cooperation, we calculate percentages for two-way contingency tables and include the results for the χ^2-test of independence against association. Such a χ^2-test, like any significance test, indicates the degree of evidence for the existence of an association, not the strength of an association. When at least one variable is ordinal, more powerful tests of independence than the χ^2-test such as the linear trend test do exist, but for reasons of simplicity of presentation, we do not use them in this paper.

3.2.2 Multilevel Logistic Regression

In a second series of analyses, we use multilevel logistic regression to simultaneously estimate the impact of the various determinants (Snijders and Boskers 1999). We opt for the use of a multilevel method, because we regard our data as hierarchically nested with individuals at the lower level (level 1) and municipalities at the higher level (level 2).

Let p_{ij} be the probability that an individual i belonging to municipality j is contacted (or cooperates). We will consider four different models for explaining this probability: the null random model, two versions of the random intercept model and the standard logistic regression model.

The empty or unconditional model does not take explanatory variables into account. We specify the model such that

logit transformed probabilities p_{ij} have a normal distribution:

$$\text{logit}\ (p_{ij}) = 1/(1+\exp(p_{ij})) = \gamma_0 + \boldsymbol{u_{0j}}$$

where γ_0 is the population average and $\boldsymbol{u_{0j}}$ the random deviation from this average for group j. These deviations $\boldsymbol{u_{0j}}$ are assumed to be independent normally distributed random variables with mean zero and variance τ_0^2.

When there are r variables at the individual level that are potentially explicative for the observed outcomes, then they are incorporated as a linear function in the random intercept model:

$$\text{logit}\ (p_{ij}) = \gamma_0 + \sum_{h=1}^{r} \gamma_h x_{hij} + \boldsymbol{u_{0j}}$$

where $\gamma_1, ..., \gamma_r$ are the slope parameters measuring the effect of the explicative variables.

If we would drop the random effects $\boldsymbol{u_{0j}}$ then we obtain a standard logistic regression model:

$$\text{logit}\ (p_{ij}) = \gamma_0 + \sum_{h=1}^{r} \gamma_h x_{hij}.$$

By also including s variables at the community level, we get an intercept model with both level-1 and level-2 covariates:

$$\text{logit}\ (p_{ij}) = \gamma_0 + \sum_{h=1}^{r} \gamma_h x_{hij} + \sum_{k=1}^{s} \gamma_k x_{kij} + \boldsymbol{u_{0j}}.$$

We use SAS Proc Nlmixed (SAS Institute 1999) to actually estimate the parameters. In SAS Proc Nlmixed an adaptive version of Gauss-Hermite Quadrature (numerical integration) is used to solve the maximum likelihood estimation problem. To test if a specific parameter equals zero, a Likelihood Ratio χ^2-test is used.

4. RESULTS

4.1 Contactability

Table 1 presents the bivariate results by the χ^2-test of the percentage never contacted by various indicators of physical impediments. One strong correlate is whether the housing unit is a single-family structure or not, the latter having much higher noncontact rates (8.1%) than other units (2.4%). Also, sample persons living in large multiunit housing structures tend to have higher noncontact rates (11%) than those not living in large multiunit housing structures (3.1%). Another strong correlate is the presence of a telephone: 9.7% of those with no telephone were never contacted.

Table 1
Percentage Never-Contacted by 'Physical Impediments' Attributes

Physical impediments attributes	Percentage never contacted	χ^2	*df*	*p*
Single-Family Structure		97.6	1	<0.0001
No	8.1			
Yes	2.4			
Large multi-unit structure (>10)		38.4	1	<0.0001
No	3.1			
Yes	11.0			
Telephone		88.9	1	<0.0001
No	9.7			
Yes	2.7			

Table 2 shows the bivariate results for contactability by 'reasons to be present at home' attributes. Relatively more unmarried (4.4%) and divorced (6.9%) sample persons than married (2.9%) sample persons are never contacted. There are much lower rates of noncontacts among those that are inactive (0.9%) compared to other persons (3.5%). Having at least 3 or more children (0.9%) leads to low noncontact rates, compared to having two children (2.6%) or at most 1 child (4%). Those working at home (1.5%) and those being an independent worker (1.9%) show modestly lower noncontact rates than those working elsewhere (3.6%) or those working as an employee respectively (3.6%). Age, the number of weekly working hours, working part-time *vs.* full-time and having a second job or not have no significant influence on contactability.

Table 2
Percentage Never-Contacted by 'Reasons to be Present at Home' Attributes

Reasons to be present at home	Percentage never contacted	χ^2	*df*	*p*
Civil status		19.4	2	<0.0001
Unmarried	4.4			
Married	2.9			
Divorced	6.9			
Inactive *vs.* other		4.0	1	0.04
Inactive	0.9			
Other	3.5			
Number of children[a]		14.5	3	0.0023
0	4.3			
1	4.0			
2	2.6			
3+	0.9			
Employment place[b]		4.6	1	0.03
At home	1.5			
Elsewhere	3.6			
Employment status[b]		4.0	1	0.05
Employee	3.6			
Own-account	1.9			

[a] subsample of women only (n=4,098)
[b] subsample of active persons only (n=5,368)

In addition, substitution is associated with higher noncontact rates (5.9%) compared to the base sample (2.6%). No significant difference has been found for the male and the female subsample.

In a multiple logistic regression model of the combined effects of those individual-level indicators that have some marginal bivariate effect on contactability only single-family structure ($\chi^2 = 35.75, p = <0.0001$), telephone ($\chi^2 = 52.63$, $p=<0.0001$) and substitution ($\chi^2 =28.59$, $p= <0.0001$) remain significant.

In Table 3, noncontact rates for various environmental attributes are presented. Cities (6.6%) have higher non-contact rates compared to nonurban areas (3.1%). The percentage never contacted is higher for high-density areas (5.4%) than low-density areas (1.7%). The presence of multiunit structures and the presence of persons of other nationalities tend to increase non-contact rates. Finally, the percentage of owner-occupied houses shows a negative association with noncontact rates.

Table 3
Percentage Never-Contacted by 'Environmental' Attributes

Environmental attribute	Percentage never contacted	χ^2	*df*	*p*
Urban status		24.0	1	<0.0001
Cities	6.6			
Other	3.1			
Population density		34.4	3	<0.0001
Lowest quartile	1.7			
Second quartile	3.2			
Third quartile	3.8			
Highest quartile	5.4			
% Multi-unit structures		50.4	3	<0.0001
Lowest quartile	2.0			
Second quartile	2.2			
Third quartile	4.0			
Highest quartile	5.9			
% Persons of other nationalities		23.1	3	<0.0001
Lowest quartile	2.5			
Second quartile	2.3			
Third quartile	4.3			
Highest quartile	4.8			
% Homes owner-occupied		64.4	3	<0.0001
Lowest quartile	6.4			
Second quartile	3.6			
Third quartile	1.6			
Highest quartile	2.7			

We complement now the bivariate analysis with a multivariate analysis. In Table 4 four models for modelling contact relative to noncontact are presented. Model 1 is the null random model at the municipality level. Model 2 is a multiple logistic regression model. In this model, we have included the person-level effects that remained significant in a multivariate context (*i.e.*, single-family structure, telephone and substitution) and the variable activity status because of its theoretical importance. Model 3 is a random intercept version of model 2. In Model 4, we have extended Model 3 with the municipality level variable multi-units structures (in %) only.

Table 4
Results of (Multilevel) Logistic Regression Models of Contactability

Results	Model 1: Null Random	Model 2: Logistic Regression	Model 3: Random Intercept Level 1	Model 4: Random Intercept Level 1&2
Intercept	4.01***	3.08***	3.68***	4.15***
	(0.16)	(0.73)	(0.77)	(0.79)
Individual Characteristics				
Single-family structure		1.16***	1.02***	0.92***
		(0.15)	(0.17)	(0.17)
Telephone		1.19***	1.25***	1.26***
		(0.16)	(0.17)	(0.17)
Inactive *vs.* other		-1.23	-1.34	-1.33
		(0.72)	(0.75)	(0.74)
Substitution sample		-0.78***	-0.64***	-0.62***
		(0.14)	(0.15)	(0.15)
Municipality Characteristics				-0.02*
Multi-unit structures (%)				(0.01)
Estimated variances				
Var(Intercept)	1.03		0.82	0.79
Goodness of fit				
Deviance	1,720	1,658	1,606	1,599

Notes: Standard errors in parentheses. *$p<0.05$, **$p<0.01$, ***$p<0.001$, one-tailed tests.

The effects of the person-level covariates in Models 2, 3 and 4 are in accordance with the findings of the bivariate analysis. Single-family structure and the presence of a telephone have a positive influence on contactability, while the effect of activity status is not significant. The impact of field substitution is negative. We also notice a (rather small) reduction of the regression coefficient for single-family structure and substitution in the multilevel models 3 and 4. Models 3 and 4 have one variance component for the intercept. To test the null hypothesis that the random intercept variance equals zero, we use the Likelihood Ratio test and compare the conventional logit model (Model 2) with the random intercept model (Model 3). The difference in deviance between both models is large (52). So, there might be some variance in the intercept to explain by municipality level covariates. By introducing municipality characteristics one at a time, we can test for significant effects by calculating deviance differences between Model

4 and Model 3. The only deviance difference of importance noted is the case of the variable 'multi-unit structures' (7 units difference). No differences in deviances are found for the introduction of the other level-two variables (urban status, percentage owner occupied, population density and persons of other nationalities).

We consider Model 3 and Model 4 as the better models. According to these multilevel models, noncontact rates vary considerably across municipalities. However, the municipality level covariates in our study are not able to explain much of this variation.

4.2 Cooperation

In Table 5, we present the bivariate results for the opportunity costs hypothesis indicators. Being inactive or not does not seem to have an effect on the cooperation rate. However, when we use indicators of discretionary time, such as working part-time versus working full-time or the weekly number of working hours, the predicted negative relationship does show up in the bivariate results. In addition, self-employed sample persons have lower cooperation rates compared to employees.

Table 5
Percentage Cooperation by 'Opportunity Cost Hypothesis' Indicators

Opportunity cost indicators	Percentage cooperated	χ^2	df	p
Inactive *vs.* other		0.41	1	0.52
Inactive	77.0			
Other	78.9			
Part-time *vs.* Full-time[b]		10.04	1	0.001
Part-time	82.3			
Full-time	77.4			
Number of working hours[b]		15.3	3	0.0016
<20	80.1			
21-35	84.7			
36-42	77.6			
> 43	75.7			
Employment status[b]		4.2	1	0.04
Employee	78.7			
Own-account	74.6			

[b] subsample of active persons only (n=5,180)

The predictions of the exchange hypothesis theory do not show up in the bivariate results presented in Table 6. SES indicators like the surface of the living room and the number of bathrooms are not negatively, but positively related to cooperation. Of course, these measures are not ideal, because we are not able to control for household size. Another indication of a positive relationship between cooperation and SES is the case of educational level. Whether one receives a replacement income or not and whether the house is owner-occupied or not has no impact on cooperation rates.

In a multiple logistic regression model of the combined effects of those social exchange indicators that have some marginal bivariate effect on cooperation, only the effects of educational level (χ^2 =39.35, df=4, p <0.0001) and surface of the living room (χ^2 =13.4, df=4, p=0.0095) remain significant.

Table 6
Percentage Cooperation by 'Exchange Hypothesis' Indicators

Exchange indicators	Percentage cooperated	χ^2	df	p
Surface living rooms (m^2)		26.8	4	<0.0001
< 65	74.8			
65 - 84	77.6			
85 - 104	78.6			
105 - 124	79.9			
> 125	83.1			
Number of bathrooms		7.9	2	0.02
0	74.2			
1	78.6			
2	83.5			
Educational level		46.7	4	<0.0001
Primary	76.6			
Secondary, first stage	74.5			
Secondary, second stage	78.7			
High, non-university	85.1			
High, university	82.2			
Replacement income		0.3	1	0.58
No	78.7			
Yes	79.5			
Owner occupied		3.4	1	0.06
No	77.4			
Yes	79.4			

In the section for the exchange hypotheses, we have found support for the notion that those with low SES, cooperate less with surveys than those in the high SES groups. Such a positive relationship between SES is predicted by the social isolation hypothesis. Demographic indicators of social isolation theory are gender, civil status and age (See Table 7). No effects are found for gender, civil status (however, divorced sample persons are probably less cooperative) and single-family structure. Age seems to have a negative effect on cooperation. For women only, we have also data on the presence of children. We find that the number of children has a positive effect on cooperation rates. The age of the children is also important: the presence of young children is associated with higher cooperation.

The control variable substitution has a slightly negative effect on cooperation (χ^2=4.24, p=0.039) with lower

cooperation rates for the substitution sample (77.3%), compared to the base sample (79.5%).

Table 7
Percentage Cooperation by 'Social isolation Hypothesis' Indicators

Social isolation indicators	Percentage cooperated	χ^2	df	p
Gender		1.56	1	0.21
Male	78.1			
Female	79.3			
Civil status		3.11	2	0.21
Unmarried	79.8			
Married	78.6			
Divorced	75.4			
Single-family structure		0.76	1	0.38
No	78.9			
Yes	77.7			
Age		17.5	3	0.0006
20 - 24	80.8			
25 - 29	80.7			
30 - 34	78.3			
35 - 39	75.5			
Number of children[a]		18.2	3	0.0004
0	77.9			
1	76.3			
2	81.7			
3+	84.9			
Presence of young children[a]		12.3	1	0.0005
No	77.8			
Yes	82.8			

[a] subsample of women only (n=3,955)

Table 8 contains the bivariate results for social environmental differences in cooperation. Population density has a curvilinear effect on cooperation. Being a resident in a large metropolitan area has no effect. Thus, the evidence for the literature that crowding and high levels of stimulus input are negatively associated with cooperation is of a mixed nature.

The effect of indicators for social cohesion is not clear. Only the variable percentage owner-occupied has a (curvilinear) effect. The variables percentage persons of other nationalities and percentage multi-unit structures seem to have no effect.

Finally, we present in Table 9 a series of regression models for cooperation similar to those in section 4.1. In these models, we have included four individual level covariates: surface of the living room (<84 , >84 m²), education (up to secondary -second stage *vs.* high level), age (20-29, 30-39 years) and substitution sample. Surface of the living room and education have been selected as the only significant exchange hypothesis indicators in the previously described multiple logistic regression model. Age was the only significant effect in the bivariate analysis on the social isolation hypothesis. Finally, substitution is introduced to control for possible fieldwork effects. The slightly negative effect of substitution in Model 2 might indicate that fieldwork substitution negatively influences cooperation. However, this effect disappears completely when a random intercept is introduced (Models 3 and 4). The effects of the other individual level covariates are in accordance with the findings of the bivariate analysis and do not change across Models 2 to 4. SES indicators like education and surface of the living room have a positive effect and age has a negative effect on cooperation. These effects rather confirm the social isolation hypothesis than the exchange hypothesis.

Table 8
Percentage Cooperation by 'Environmental' Attributes

Environmental attribute	Percentage cooperated	χ^2	df	p
Urban status		0.84	1	0.36
Cities	80.1			
Other	78.7			
Population density		10.7	3	0.014
Lowest quartile	80.0			
Second quartile	79.9			
Third quartile	76.0			
Highest quartile	79.4			
% Multiunit structures		3.1	3	0.38
Lowest quartile	80.1			
Second quartile	79.2			
Third quartile	77.9			
Highest quartile	78.1			
% Homes owner-occupied		12.3	3	0.0063
Lowest quartile	79.7			
Second quartile	76.2			
Third quartile	78.5			
Highest quartile	80.9			
% Persons of other nationalities		5.2	3	0.16
Lowest quartile	77.7			
Second quartile	77.6			
Third quartile	79.6			
Highest quartile	80.2			

The only level two variable of (modest) importance is multi-unit structures (in %) and has been kept in Model 4. The Likelihood Ratio test for introducing this variable gives a difference of two units in deviance terms. The introduction of one or more other second level variables gives Likelihood Ratio tests differences close to zero in deviance terms. We consider Model 3 and 4 as the most suitable models. The difference in deviance terms between model 3 and model 2 is 8 units, which is significant. The variance for the intercept term is moderate (0.21). The introduction of second level covariates (including multi-unit structures)

leaves this variance term practically unchanged. Therefore, we may state that environmental attributes like urbanicity are not important for explaining cooperation.

Table 9
Results of (Multilevel) Logistic Regression Models of Cooperation

Results	Model 1: Null Random	Model 2: Logistic Regression	Model 3: Random Intercept Level 1	Model 4: Random Intercepts Level 1&2
Intercept	1.41***	1.24***	1.30***	1.39***
	(0.06)	(0.06)	(0.08)	(0.10)
Individual Characteristics				
Substitution sample		-0.15*	-0.03	-0.02
		(0.07)	(0.07)	(0.07)
Surface living rooms		0.23***	0.24***	0.24***
		(0.06)	(0.06)	(0.06)
Educational level		0.45***	0.47***	0.47***
		(0.08)	(0.08)	(0.08)
Age		-0.23***	-0.23***	-0.23***
		(0.06)	(0.06)	(0.06)
Municipality Characteristics				
Multi-unit structures (%)				-0.006
				(0.004)
Estimated variances				
Var(Intercept)	0.21		0.21	0.21
Goodness of fit				
Deviance	6,664	6,664	6,596	6,594

Notes: Standard errors in parentheses. * $p<0.05$, ** $p<0.01$, *** $p<0.001$, one-tailed tests.

5. DISCUSSION

In this paper, we have used 1991 individual and municipality level Census data matched to the response status variable of the Belgian Fertility and Family Survey to analyse the relative importance of correlates of contact and cooperation.

We have organised our analysis according to the Groves-Couper conceptual framework. In the bivariate analysis stage, we have found essentially the same kind of correlates as was predicted and actually found in an US-based multi-survey analysis (G&C 1998). One important difference between the present study and the US-results seems to be the nature of the effect of SES indicators (*e.g.*, education) on cooperation. In the present study, we find a positive relationship; in the US-study the inverse relationship is found. We can imagine two alternative explanations for these conflicting findings. A first one is based on survey design effects such as topic saliency. The FFS-survey in Belgium might be atypical in being disproportionally attractive to the higher educated because of the specific content of the survey. Replicating the present analysis for surveys about varying topics can easily test such a hypothesis. Another possible hypothesis is that effects of education on survey cooperation do vary across societies. Then the challenge is to find out why this relationship varies across countries. Such a hypothesis is far less easy to test in real, as data for several countries are needed.

In the multilevel logistic regression analysis stage, the impact of all but one contextual factor completely vanished. Only the impact of the variable percentage of multi-unit structures shows, however only weakly, some resistance against ecological randomness present in the random intercept models. To us, this is a very intriguing result. Random ecological variation at the municipality level seems to dominate largely even the urban-rural dichotomy. A possible explanation is that the variation at the community level is dominated by interviewer effects, not by ecological factors.

ACKNOWLEDGEMENTS

This paper grew out of our activities in the design and organisation of surveys at the Population and Family Study Centre (CBGS), a scientific Institute of the Flemish Government in Belgium. We are grateful to the Belgian National Statistical Institute, the local authorities and the National Register for their cooperation. We also thank the Editor and two anonymous referees for their valuable comments and suggestions. Finally, this research was supported by a FWO grant Bijzondere doctoraatsbeurs 2002-2003.

REFERENCES

BLAU, P.M. (1964). *Exchange and Power in Social Life*. New York: John Wiley & Sons, Inc.

CALLENS, M. (1995). *De 'Fertility and Family Survey' in Vlaanderen (Nego V, 1991), De gegevensverzameling*. Brussel: CBGS-document. 1995, 4.

CHAPMAN, D.W. (1983). The impact of substitution on survey estimates. In *Incomplete Data in Sample Surveys, Proceedings of the Symposium*, (Eds. W.G. Madow and I. Olkin). New York: Academic. 2, 45-61.

CLIQUET, R.L., and CALLENS, M. (Ed.) (1993). *Gezinsvorming in Vlaanderen. Hoe en Wanneer?* Brussel: CBGS Monografie 1.

COUPER, M.P., SINGER, E. and KULKA, R.A. (1997). Participation in the decennial census: Politics, privacy, pressures. *American Politics Quarterly*. 26, 59-80.

GROVES, R.M., and COUPER, M. (1998). *Nonresponse in Household Interview Surveys*. New York: John Wiley & Sons, Inc.

SAS INSTITUTE INC. (1999). *SAS/STAT® User's Guide*. Version 8, Cary, NC: SAS Institute Inc.

SNIJDERS, T.A.B., and BOSKERS, R.J. (1999). *Multilevel Analysis, an Introduction to Basic and Advanced Multilevel Modeling*. London: Sage Publications.

VANNESTE, D. (1989). *Economische Typering Van de Belgische Gemeenten*. Leuven: Leuvense Geografische Papers.

VEHOVAR, V. (1999). Field substitution and unit nonresponse, *Journal of Official Statistics*. 15, 2, 335-350.

JOURNAL OF OFFICIAL STATISTICS

An International Review Published by Statistics Sweden

JOS is a scholarly quarterly that specializes in statistical methodology and applications. Survey methodology and other issues pertinent to the production of statistics at national offices and other statistical organizations are emphasized. All manuscripts are rigorously reviewed by independent referees and members of the Editorial Board.

Contents
Volume 20, No. 1, 2004

The Effect of Multiple Weighting Steps on Variance Estimation
Richard Valliant 1

Combining Link-Tracing Sampling and Cluster Sampling to Estimate the Size of Hidden Populations
Martin H. Félix-Medina and Steven K. Thompson 19

Nonresponse in Time: Some Lessons from Major Finnish Surveys
Kari Djerf 39

Towards a Social Statistical Database and Unified Estimates at Statistics Netherlands
Marianne Houbiers 55

A Simultaneous Analysis of Interviewer Effects on Various Data Quality Indicators with Identification of Exceptional Interviewers
Jan Pickery and Geert Loosveldt 77

Frequency Reports Across Age Groups
Bärbel Knäuper, Norbert Schwarz, and Denise Park 91

Extracting Confidential Information from Public Documents: The 2000 Department of Justice Report on the Federal Use of the Death Penalty in the United States
David J. Algranati and Joseph B. Kadane 97

American Fact Finder: Disclosure Limitation for the Advanced Query System
Sam Hawala, Laura Zayatz, and Sandra Rowland 115

Keys to Successful Implementation of Continuous Quality Improvement in a Statistical Agency
David A. Marker and David R. Morganstein 125

In Other Journals 137

Corrigendum 139

All inquires about submissions and subscriptions should be directed to jos@scb.se

CONTENTS TABLE DES MATIÈRES

Volume 31, No. 4, December/Décembre 2003, 355-490

New Editor for *The Canadian Journal of Statistics* 355
Un nouveau rédacteur en chef pour *La revue canadienne de statistique* 356

Belkacem ABDOUS, Kilani GHOUDI & Bruno RÉMILLARD:
Nonparametric weighted symmetry tests 357

Sharon L. LOHR & N.G.N. PRASAD:
Small-area estimation with auxiliary survey data 383

Peilin SHI, Jane J. YE & Julie ZHOU:
Minimax robust designs for misspecified regression models 397

Brajendra C. SUTRADHAR & R. Prabhakar RAO:
On quasi-likelihood inference in generalized linear mixed models with two components of dispersion 415

Denis LAROCQUE:
An affine-invariant multivariate sign test for cluster correlated data 437

Zhide FANG:
Extrapolation designs with constraints 457

Giovanni PETRIS & Luca TARDELLA:
A geometric approach to transdimensional Markov chain Monte Carlo 469

Correction:
Yann GUÉDON & Christiane COCOZZA-THIVENT:
Nonparametric estimation of renewal processes from count data 483

Forthcoming papers/Articles à paraître 484

Index: Volume 31 (2003) 485

Volume 32 (2004): Subscription rates 489
Frais d'abonnement 490

GUIDELINES FOR MANUSCRIPTS

Before having a manuscript typed for submission, please examine a recent issue of *Survey Methodology* (Vol. 19, No. 1 and onward) as a guide and note particularly the points below. Articles must be submitted in machine-readable form, preferably in Word. A paper copy may be required for formulas and figures.

1. Layout

1.1 Manuscripts should be typed on white bond paper of standard size (8½ x 11 inch), one side only, entirely double spaced with margins of at least 1½ inches on all sides.
1.2 The manuscripts should be divided into numbered sections with suitable verbal titles.
1.3 The name and address of each author should be given as a footnote on the first page of the manuscript.
1.4 Acknowledgements should appear at the end of the text.
1.5 Any appendix should be placed after the acknowledgements but before the list of references.

2. Abstract

The manuscript should begin with an abstract consisting of one paragraph followed by three to six key words. Avoid mathematical expressions in the abstract.

3. Style

3.1 Avoid footnotes, abbreviations, and acronyms.
3.2 Mathematical symbols will be italicized unless specified otherwise except for functional symbols such as "exp(·)" and "log(·)", etc.
3.3 Short formulae should be left in the text but everything in the text should fit in single spacing. Long and important equations should be separated from the text and numbered consecutively with arabic numerals on the right if they are to be referred to later.
3.4 Write fractions in the text using a solidus.
3.5 Distinguish between ambiguous characters, (e.g., w, ; o, O, 0; l, 1).
3.6 Italics are used for emphasis. Indicate italics by underlining on the manuscript.

4. Figures and Tables

4.1 All figures and tables should be numbered consecutively with arabic numerals, with titles which are as nearly self explanatory as possible, at the bottom for figures and at the top for tables.
4.2 They should be put on separate pages with an indication of their appropriate placement in the text. (Normally they should appear near where they are first referred to).

5. References

5.1 References in the text should be cited with authors' names and the date of publication. If part of a reference is cited, indicate after the reference, e.g., Cochran (1977, p. 164).
5.2 The list of references at the end of the manuscript should be arranged alphabetically and for the same author chronologically. Distinguish publications of the same author in the same year by attaching a, b, c to the year of publication. Journal titles should not be abbreviated. Follow the same format used in recent issues.