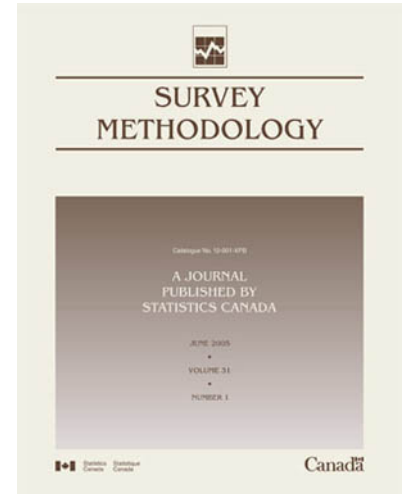


Catalogue no. 12-001-XIE  
ISSN: 1492-0921

Catalogue no. 12-001-XPB  
ISSN: 0714-0045

# Survey Methodology

December 2012



## How to obtain more information

For information about this product or the wide range of services and data available from Statistics Canada, visit our website, [www.statcan.gc.ca](http://www.statcan.gc.ca).

You can also contact us by

**email** at [infostats@statcan.gc.ca](mailto:infostats@statcan.gc.ca),

**telephone**, from Monday to Friday, 8:30 a.m. to 4:30 p.m., at the following toll-free numbers:

- Statistical Information Service 1-800-263-1136
- National telecommunications device for the hearing impaired 1-800-363-7629
- Fax line 1-877-287-4369

## Depository Services Program

- Inquiries line 1-800-635-7943
- Fax line 1-800-565-7757

## To access this product

This product, Catalogue no. 12-001-X, is available free in electronic format. To obtain a single issue, visit our website, [www.statcan.gc.ca](http://www.statcan.gc.ca), and browse by "Key resource" > "Publications."

This product is also available as a standard printed publication at a price of CAN\$30.00 per issue and CAN\$58.00 for a one-year subscription.

The following additional shipping charges apply for delivery outside Canada:

	<b>Single issue</b>	<b>Annual subscription</b>
United States	CAN\$6.00	CAN\$12.00
Other countries	CAN\$10.00	CAN\$20.00

All prices exclude sales taxes.

The printed version of this publication can be ordered as follows:

- Telephone (Canada and United States) 1-800-267-6677
  - Fax (Canada and United States) 1-877-287-4369
  - E-mail [infostats@statcan.gc.ca](mailto:infostats@statcan.gc.ca)
  - Mail  
Statistics Canada  
Finance  
R.H. Coats Bldg., 6th Floor  
150 Tunney's Pasture Driveway  
Ottawa, Ontario K1A 0T6
- In person from authorized agents and bookstores.

When notifying us of a change in your address, please provide both old and new addresses.

## Standards of service to the public

Statistics Canada is committed to serving its clients in a prompt, reliable and courteous manner. To this end, Statistics Canada has developed standards of service that its employees observe. To obtain a copy of these service standards, please contact Statistics Canada toll-free at 1-800-263-1136. The service standards are also published on [www.statcan.gc.ca](http://www.statcan.gc.ca) under "About us" > "The agency" > "Providing services to Canadians."

Published by authority of the Minister responsible for  
Statistics Canada

© Minister of Industry, 2012

All rights reserved. Use of this publication is governed by the  
Statistics Canada Open Licence Agreement ([http://www.  
statcan.gc.ca/reference/licence-eng.html](http://www.statcan.gc.ca/reference/licence-eng.html)).

Cette publication est aussi disponible en français.

### Note of appreciation

Canada owes the success of its statistical system to a long-standing partnership between Statistics Canada, the citizens of Canada, its businesses, governments and other institutions. Accurate and timely statistical information could not be produced without their continued co-operation and goodwill.

### Standard symbols

The following symbols are used in Statistics Canada publications:

- .
  - ..
  - ...
  - 0
  - 0<sup>s</sup>
  - P
  - r
  - X
  - E
  - F
  - \*
- not available for any reference period  
not available for a specific reference period  
not applicable  
true zero or a value rounded to zero  
value rounded to 0 (zero) where there is a meaningful distinction between true zero and the value that was rounded  
preliminary  
revised  
suppressed to meet the confidentiality requirements of the *Statistics Act*  
use with caution  
too unreliable to be published  
significantly different from reference category ( $p < 0.05$ )

The paper used in this publication meets the minimum requirements of American National Standard for Information Sciences – Permanence of Paper for Printed Library Materials, ANSI Z39.48 - 1984.



Le papier utilisé dans la présente publication répond aux exigences minimales de l'American National Standard for Information Sciences – “Permanence of Paper for Printed Library Materials”, ANSI Z39.48 - 1984.



# SURVEY METHODOLOGY

## A Journal Published by Statistics Canada

*Survey Methodology* is indexed in The ISI Web of knowledge (Web of science), 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. It is also covered by SCOPUS in the Elsevier Bibliographic Databases.

### MANAGEMENT BOARD

**Chairman** J. Kovar

**Past Chairmen** D. Royce (2006-2009)

G.J. Brackstone (1986-2005)

R. Platek (1975-1986)

**Members** G. Beaudoin

S. Fortier (Production Manager)

J. Gambino

M.A. Hidirolou

H. Mantel

### EDITORIAL BOARD

**Editor** M.A. Hidirolou, *Statistics Canada*

**Deputy Editor** H. Mantel, *Statistics Canada*

**Past Editor** J. Kovar (2006-2009)

M.P. Singh (1975-2005)

### Associate Editors

J.-F. Beaumont, *Statistics Canada*

J. van den Brakel, *Statistics Netherlands*

J.M. Brick, *Westat Inc.*

P. Cantwell, *U.S. Bureau of the Census*

R. Chambers, *Centre for Statistical and Survey Methodology*

J.L. Eltinge, *U.S. Bureau of Labor Statistics*

W.A. Fuller, *Iowa State University*

J. Gambino, *Statistics Canada*

D. Haziza, *Université de Montréal*

B. Hulliger, *University of Applied Sciences Northwestern Switzerland*

D. Judkins, *Westat Inc.*

D. Kasprzyk, *National Opinion Research Center*

J.K. Kim, *Iowa State University*

P.S. Kott, *RTI International*

P. Lahiri, *JPSM, University of Maryland*

P. Lavallée, *Statistics Canada*

P. Lynn, *University of Essex*

D.J. Malec, *National Center for Health Statistics*

G. Nathan, *Hebrew University*

J. Opsomer, *Colorado State University*

D. Pfeffermann, *Hebrew University*

N.G.N. Prasad, *University of Alberta*

J.N.K. Rao, *Carleton University*

J. Reiter, *Duke University*

L.-P. Rivest, *Université Laval*

F.J. Scheuren, *National Opinion Research Center*

P. do N. Silva, *Escola Nacional de Ciências Estatísticas*

P. Smith, *Office for National Statistics*

E. Stasny, *Ohio State University*

D. Steel, *University of Wollongong*

L. Stokes, *Southern Methodist University*

M. Thompson, *University of Waterloo*

V.J. Verma, *Università degli Studi di Siena*

K.M. Wolter, *National Opinion Research Center*

C. Wu, *University of Waterloo*

W. Yung, *Statistics Canada*

A. Zaslavsky, *Harvard University*

**Assistant Editors** C. Bocci, K. Bosa, G. Dubreuil, C. Leon, S. Matthews, Z. Patak, S. Rubin-Bleuer and Y. You, *Statistics Canada*

---

### 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, (smj@statcan.gc.ca, Statistics Canada, 150 Tunney's Pasture Driveway, Ottawa, Ontario, Canada, K1A 0T6). For formatting instructions, please see the guidelines provided in the journal and on the web site (www.statcan.gc.ca).

### Subscription Rates

The price of printed versions 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 \$20 (\$10 × 2 issues). 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. Electronic versions are available on Statistics Canada's web site: www.statcan.gc.ca.

**Survey Methodology**  
A Journal Published by Statistics Canada  
Volume 38, Number 2, December 2012

**Contents**

**Waksberg Invited Paper Series**

Lars Lyberg	
Survey Quality .....	107

**Regular Papers**

Jaqueline Garcia-Yi and Ulrike Grote	
Data collection: Experiences and lessons learned by asking sensitive questions in a remote coca growing region in Peru .....	131
Jun Shao, Martin Klein and Jing Xu	
Imputation for nonmonotone nonresponse in the survey of industrial research and development .....	143
Jae Kwang Kim and Minsun Kim	
Riddles Some theory for propensity-score-adjustment estimators in survey sampling .....	157
Ian Plewis, Sosthenes Ketende and Lisa Calderwood	
Assessing the accuracy of response propensity models in longitudinal studies.....	167
Sarat C. Dass, Tapabrata Maiti, Hao Ren and Samiran Sinha	
Confidence interval estimation of small area parameters shrinking both means and variances.....	173
Dan Liao and Richard Valliant	
Condition indexes and variance decompositions for diagnosing collinearity in linear model analysis of survey data.....	189
Qixuan Chen, Michael R. Elliott and Roderick J.A. Little	
Bayesian inference for finite population quantiles from unequal probability samples.....	203

**Short Notes**

Satkartar K. Kinney	
Multiple imputation with census data .....	215

<b>Notice</b> .....	219
<b>Corrigendum</b> .....	220
<b>Acknowledgements</b> .....	221
<b>Announcements</b> .....	223
<b>In Other Journals</b> .....	225

**ELECTRONIC  
PUBLICATIONS  
AVAILABLE AT**

**PUBLICATIONS  
ÉLECTRONIQUES  
DISPONIBLE À**

**[www.statcan.gc.ca](http://www.statcan.gc.ca)**

## Waksberg Invited Paper Series

The journal *Survey Methodology* has established an annual invited paper series in honour of Joseph Waksberg, who has made many important contributions to survey methodology. Each year a prominent survey researcher is chosen to author an article as part of the Waksberg Invited Paper Series. The paper reviews the development and current state of a significant topic within the field of survey methodology, and reflects the mixture of theory and practice that characterized Waksberg's work.

Please see the announcements at the end of the Journal for information about the nomination and selection process of the 2014 Waksberg Award.

This issue of *Survey Methodology* opens with the twelfth paper of the Waksberg Invited Paper Series. The editorial board would like to thank the members of the selection committee Elizabeth A. Martin (Chair), Mary E. Thompson, Steve Heeringa and J.N.K. Rao for having selected Lars Lyberg as the author of this year's Waksberg Award paper.

### 2012 Waksberg Invited Paper

#### Author: Lars Lyberg

Lars Lyberg, Ph.D., is former Head of the Research and Development Department at Statistics Sweden and currently Professor Emeritus at the Department of Statistics, Stockholm University. He is the founder of the Journal of Official Statistics (JOS) and served as Chief Editor for 25 years. He is chief editor of *Survey Measurement and Process Quality* (Wiley 1997) and co-editor of *Survey Methods in Multinational, Multiregional, and Multicultural Contexts* (Wiley 2010), *Telephone Survey Methodology* (Wiley 1988) and *Measurement Errors in Surveys* (Wiley 1991). He is co-author of *Introduction to Survey Quality* (Wiley 2003). He chaired the Leadership Group on Quality of the European Statistical System and chaired the Organizing Committee of the first European Conference on Quality in Official Statistics, Q2 001. He is former president of IASS and former chair of the ASA Survey Methods Section. He is a fellow of the American Statistical Association and the Royal Statistical Society.

**ELECTRONIC  
PUBLICATIONS  
AVAILABLE AT**

**PUBLICATIONS  
ÉLECTRONIQUES  
DISPONIBLE À**

**[www.statcan.gc.ca](http://www.statcan.gc.ca)**



# Survey Quality

Lars Lyberg<sup>1</sup>

## Abstract

Survey quality is a multi-faceted concept that originates from two different development paths. One path is the total survey error paradigm that rests on four pillars providing principles that guide survey design, survey implementation, survey evaluation, and survey data analysis. We should design surveys so that the mean squared error of an estimate is minimized given budget and other constraints. It is important to take all known error sources into account, to monitor major error sources during implementation, to periodically evaluate major error sources and combinations of these sources after the survey is completed, and to study the effects of errors on the survey analysis. In this context survey quality can be measured by the mean squared error and controlled by observations made during implementation and improved by evaluation studies. The paradigm has both strengths and weaknesses. One strength is that research can be defined by error sources and one weakness is that most total survey error assessments are incomplete in the sense that it is not possible to include the effects of all the error sources. The second path is influenced by ideas from the quality management sciences. These sciences concern business excellence in providing products and services with a focus on customers and competition from other providers. These ideas have had a great influence on many statistical organizations. One effect is the acceptance among data providers that product quality cannot be achieved without a sufficient underlying process quality and process quality cannot be achieved without a good organizational quality. These levels can be controlled and evaluated by service level agreements, customer surveys, process analysis using statistical process control, and organizational assessment using business excellence models or other sets of criteria. All levels can be improved by conducting improvement projects chosen by means of priority functions. The ultimate goal of improvement projects is that the processes involved should gradually approach a state where they are error-free. Of course, this might be an unattainable goal, albeit one to strive for. It is not realistic to hope for continuous measurements of the total survey error using the mean squared error. Instead one can hope that continuous quality improvement using management science ideas and statistical methods can minimize biases and other survey process problems so that the variance becomes an approximation of the mean squared error. If that can be achieved we have made the two development paths approximately coincide.

Key Words: Quality management; Total survey error; Quality framework; Mean squared error; Process variability; Statistical process control; Users of survey data.

## 1. Introduction

This article has been prepared in recognition of Joe Waksberg's unique contributions and leadership in survey methodology. My first encounter with Joe's work was his article on response errors in expenditure surveys written with John Neter (Neter and Waksberg 1964). Among other things that article introduced me to the cognitive phenomenon called telescoping. Later in life I had the opportunity to work with Joe on the first conference and monograph on telephone survey methodology where we were part of the editorial group (Groves, Biemer, Lyberg, Massey, Nicholls and Waksberg 1988). We also collaborated on the preparation of many of the Hansen Lectures that were published in the *Journal of Official Statistics* (JOS) during my term as its Chief Editor. Joe himself delivered the sixth lecture, which was published in JOS (Waksberg 1998). Joe was a fantastic leader and it is a great honor for me to have been invited to write this article on survey quality, a topic that occupied his mind a lot.

Many of my friends have conveyed their views or sent me materials in preparation of this article. Especially I want to thank Paul Biemer, Dan Kasprzyk, Fritz

Scheuren, Dennis Trewin, and Maria Bohata for helping me.

Survey quality is a vague, albeit intuitive, concept with many meanings. In this article I discuss some observations related to the development and treatment of the concept over the last 70 years and for some developments it is possible to trace roots that can be found even farther back. Most of my discussion, however, concerns current issues in government statistical organizations. It is within official statistics that most of my survey quality examples take place.

The article is organized as follows: In Section 2 I discuss the total survey error paradigm, including error typologies, treatment of the errors, and survey design taking all error sources into account. In section 3 I discuss quality management philosophies that have had a large impact on survey organizations since the early 1990's. This impact is manifested by methods and approaches like recognition of the user or the client, a discussion of costs and risks in survey research, and the need for organizations to continuously improve. Section 4 provides examples of quality initiatives in survey organizations. Section 5 deals with the difficulties in measuring quality, either

1. Lars Lyberg, Department of Statistics, Stockholm University, 10691 Stockholm, Sweden. E-mail: Lars.Lyberg@stat.su.se.

directly or indirectly via indicators. How these measures should be communicated to the users or clients is also covered. Section 6, finally, offers some thoughts about how survey practices *must* change to better serve the needs of the users. The last section contains references.

## 2. The total survey error paradigm

### 2.1 Some history of survey sampling

There are a number of papers describing the development of early survey sampling methodology. In that early development there is an implicit or explicit recognition of quality issues although they are hidden under labels such as errors and survey usefulness (Deming 1944). The historical overviews provided by, for instance, Kish (1995), Fienberg and Tanur (1996), and O’Muirheartaigh (1997) all emphasize the fact that the period up to 1950 is characterized by a full-bloom development of sampling theory. During the 1920s the International Statistical Institute agreed to promote ideas on representative sampling suggested by Kiear (1897) and Bowley (1913). In 1934 Neyman published his landmark paper on the representative method. Later Fisher’s (1935) randomization principle was used in agricultural sampling and Neyman (1938) developed cluster sampling, ratio estimation and two-phase sampling and introduced the concept of confidence interval. Neyman showed that the sampling error could actually be measured by calculating the variance of the estimator. Bill Cochran, Frank Yates, Ed Deming, Morris Hansen and many others further refined the concepts of sampling theory. Hansen led a research group at the U.S. Census Bureau where much of the applied work and new theory development was conducted in those days. One remarkable result of the Census Bureau efforts was the two-volume textbook on sampling theory and methods (Hansen, Hurwitz and Madow 1953). As a matter of fact the advances in sampling theory were so prominent at the time that Stephan (1948) found it worthwhile to write an article about the history of modern sampling methods.

It was early recognized that there could be survey errors other than those attributed to sampling. There are writings on the effects of question wording such as Muscio (1917). Research on questionnaire design was quite extensive in the 1940s. Problems with errors introduced by fieldworkers collecting agricultural data in India were addressed by Mahalanobis (1946), resulting in a method for estimating such errors. The method is called “interpenetration” and can be used to estimate, so called, correlated variances introduced by interviewers, editors, coders and those who supervise these groups. The most prominent error sources were certainly known around 1950. Deming had listed error sources (1944) that constitute the first published typology of

survey errors and Hansen and Hurwitz (1946) had discussed subsampling among nonrespondents in an attempt to provide unbiased estimates in a situation with an initial nonresponse. But the methodological emphasis, up to then, had been on developing sampling theory, which is quite understandable. It was very important to be able to show that surveys could be conducted on a sampling basis and in a variety of settings. By 1950 it had been demonstrated quite successfully that this was indeed possible. So it was time to move on to other issues and refinements.

In those early days the use of the word quality was confined to mainly quality control, sometimes as quality control of survey operations. It was common that the quality control was verification and/or estimation of error sizes for various operations. Statistics were known to be plagued by errors other than those stemming from sampling but the process quality issue of how to systematically reduce these errors and biases was still to be developed (Deming 1944; Hansen and Steinberg 1956).

The user 60 years ago was a somewhat obscure player, although not at all ignored by prominent survey methodology developers. For instance, Deming (1950) claimed that until the purpose is stated, there is no right or wrong way of going about a survey. Some other statisticians made similar statements. But the user was really hiding behind terms, such as subject-matter problem, study purpose or the key functions of a statistical system.

Even now survey and quality are vague concepts. As pointed out by Morganstein and Marker (1997) varying definitions of quality undermine improvement work so we should, at least, try to distinguish between different definitions to see what purposes they might serve. One of the most cited definitions is attributed to Joseph Juran, namely quality being a direct function of “fitness for use”. It turns out that Deming already in 1944 used the phrase “fitness for purpose”, not to define quality, but rather to explain what made a survey product work.

For a long time “good” quality was implicitly equivalent to a small mean squared error (MSE), *i.e.*, data should be accurate and accuracy of an estimate can be measured by MSE, which is the sum of the variance and the squared bias. We have noticed that survey statistics should also be useful, later denoted “relevant”. Many of today’s quality dimensions were not really an issue at the time. The users, too, were accustomed to the fact that surveys took time to carry out; timeliness was surely on the agenda but not as explicitly as it is today. A census took years to process. The users were accustomed to a technology that could only deliver relatively simple forms of accessibility. Hence, it was natural for users and producers to concentrate on making sure that the statistical problem coincided reasonably well with the subject-matter problem and that MSE was kept on a

decent level, where MSE many times was and still is equivalent with just the variance, without a squared bias term added.

Before proceeding any further, let us define “survey”. A *survey* is a statistical study designed to measure population characteristics so that population parameters can be estimated. Two examples of parameters are the proportion unemployed at a given time in a population of individuals, and the total revenue of a business or industry sector during a given time period. A survey can be defined as a list of prerequisites (Dalenius 1985a). According to Dalenius a study can be classified as a survey if the following prerequisites are satisfied:

1. The study concerns a set of objects comprising a population;
2. The population under study has one or more measurable properties;
3. The goal of the study is to describe the population by one or more parameters defined in terms of measurable properties, which requires observing (a sample of) the population;
4. To get observational access to the population a frame is needed;
5. A sample of objects is selected from the frame in accordance with a sampling design that specifies a probability mechanism and a sample size  $n$  (where  $n$  might equal  $N$ , the population size);
6. Observations are made on the sample in accordance with a measurement process (*i.e.*, a measurement method and a prescription as to its use);
7. Based on the measurements, an estimation process is applied to compute estimates of the parameters when making inference from the sample to the population under study.

This definition implicitly lists the specific error sources that are present in survey work. For each source there are a number of methods available that minimize the effects but also measure their sizes (Biemer and Lyberg 2003; Groves, Fowler, Couper, Lepkowski, Singer and Tourangeau 2009).

Deviations from the definition reflect quality flaws. Moreover such deviations are common. In some designs selection probabilities are unknown or the variance estimator chosen might not be the most suitable one, given the sample design applied. Whether such flaws are problematic or not depends on the purpose.

## 2.2 The components of the total survey error paradigm

The total survey error paradigm is a theoretical framework for optimizing surveys by minimizing the accumulated size of all error sources, given budgetary constraints. In

practice this means that we want to minimize the mean squared error for selected survey estimates, namely those that are considered most important by the main stakeholders. The mean squared error is the most common metric for survey work consisting of a sum of variances and squared bias terms from each known error source. Groves and Lyberg (2010) provide a summary of the status of the paradigm in the past and in today’s survey practice.

The idea that surveys should be designed taking all error sources into account stems from the early giants in the field. Morris Hansen, Bill Hurwitz, Joe Waksberg, Leon Pritzker, Ed Deming and others at the U.S. Census Bureau, Leslie Kish at the University of Michigan, P.C. Mahalanobis at the Indian Statistical Institute, and Tore Dalenius, Stockholm University were among those who took the lead in survey research, emphasizing errors and optimal design. They worried about the inherent limitations associated with sampling theory since non sampling errors could make the theory break down. They were very practical and thought a lot about balancing errors and the costs to deal with them. Some of them saw similarities between a factory assembly line (Deming and Geoffrey 1941) and the implementation of some of the survey processes and introduced control methods obtained from industrial applications.

Dalenius (1967) realized that there was as yet no “survey design formula” that could provide an optimal solution to the design problem. The approach taken by Dalenius and also Hansen, Hurwitz and Pritzker (1967) was a strategy of minimizing all biases and going for a minimum-variance scheme so that the variance became an approximation of the MSE. This was supposed to happen through intense verification schemes for ongoing productions and quite extensive evaluation studies for future productions. In 1969 Dalenius, inspired by Hansen, presented a paper on total survey design, where the word “total” reflected the thought about taking all error sources into account. Hansen, Hurwitz, Marks and Mauldin (1951), Hansen, Hurwitz and Bershad (1961), and Hansen, Hurwitz and Pritzker (1964) developed the U.S. Census Bureau Survey Model that reflected contributions from interviewers, coders, editors, and crewleaders and allowed the estimation of those contributions to the total survey error. These estimation schemes were elaborated on by Bailar and Dalenius (1969) and consisted of variations of replication and interpenetration. Bias estimation was assumed to be handled by comparing estimates obtained from the regular operations with those obtained from preferred procedures (that could not be used on a large scale due to financial, administrative or practical reasons). Today this kind of approach is called the “gold standard”.

It was stated that good survey design called for reasonably effective control of the total error by careful

specifications of the survey procedures, including adequate controls. Hansen, Deming and others did worry about control costs but although statistical process control and acceptance sampling had been implemented in a number of survey organizations, there was very little discussion about continuous process improvement. A lot of the quality work had to do with estimation of error rates, controlling error levels for individual operators and conducting large-scale evaluation studies that usually took a long time. Users were not directly involved in the design process but in the U.S. federal statistical system they had at least some influence on what should be collected and presented. Dalenius (1968) provides more than 200 references on users and user conferences associated with the products of the U.S. Federal statistical system.

While total survey design was first advocated by Hansen, Dalenius and others, users were seldom directly involved in the final determination of survey requirements. Quite often an official, administrator or statistician acted as a subject-matter specialist. Several decades ago this was the way we thought about users. Their opinions counted but they were not really involved in design decisions. Lurking in the back of our heads was the thought that this might not be a perfect model and in the late 1970's Statistics Sweden published an internal booklet called "What to do if a customer shows up on our doorstep".

The basic design approach suggested by Hansen, Dalenius and others contained a number of steps including:

- Specification of an ideal survey goal.
- Analysis of the survey situation regarding financial, methodological and information resources.
- Developing a small number of alternative designs.
- Evaluating the alternatives by reference to associated preliminary assessments of MSE and costs.
- Choosing one of the alternatives or a modification of one of them or deciding not to conduct a survey at all.
- Developing the administrative design including feasibility testing, a process signal system (currently called *paradata*), a design document, and a Plan B.

Kish (1965) had slightly different views on design. He liked the neo-Bayesian applications in survey sampling and psychometrics advocated by colleagues at the University of Michigan (Ericson 1969; Edwards, Lindman and Savage 1963). For instance, Kish liked the idea that judgment estimates of measurement biases might be combined with sampling variances to construct more realistic estimates of the total survey error. Regarding the optimization problem Kish thought that the multipurpose situation was economically favorable for surveys but that it could be difficult to decide on what to base the design on. If one principal

statistic can be identified then that alone can decide the design and if there are a small number of principal statistics a compromise design is possible but if statistics are too disparate a reasonable design might not exist. Kish also emphasized the need for design information obtained from pilot surveys and pretests to facilitate design decisions. Kish noted that survey design and measurement could vary greatly across environments while sampling did less so. That could be one reason that sampling can be easily placed among the traditional statistical theories and methods, while it is more difficult to place the survey process in one specific discipline (Frankel and King 1996 in their interview with Kish).

Kish, like the other giants, emphasized the importance of small biases but appreciated the fact that the reduction of one bias term might increase the total error. Kish was keen on getting a reasonable balance between different error sources and how error structures varied under different design alternatives. Like Hansen and colleagues Kish thought that relevant information should be contemporaneously recorded during implementation (again we see the parallel to *paradata*). Hansen and colleagues were really concerned about excessive but inadequate controls. They realized that some controls might have to be relaxed due to limited improvements and that degree of improvement in terms of affecting the estimates should be checked out before any relaxation could take place. They also suggested that one might have to compromise relevance to get controllable measurements or abstain from the survey. Both Hansen and colleagues and Kish were vigorously in favor of ending the practice that sampling error is the only survey error measured.

When we look at today's situation we can conclude that we still do not have a design formula for surveys. There is no planning manual to speak of and the literature on design is consequently very small, as is the literature on cost (Groves 1989 is an exception). And no design formula is in sight. Since the advent of the U.S. Census Bureau survey model a number of variants have appeared on the scene, some of them quite complicated (Groves and Lyberg 2010). A common characteristic is the fact that they tend to be incomplete, *i.e.*, they do not take all error sources into account. Most statistical attention is on variance components and especially on measurement error variance. There are a number of other weaknesses associated with the total survey error concept. Most notably a user perspective is missing and a vast majority of users are not in a position to question or even discuss accuracy. The complex error structures and interactions do not invite outside scrutiny and user contacts often tend to concern less technical issues such as timeliness, comparability and costs. Users are not really informed about real levels of accuracy and we know very

little about how users perceive information about errors and how to act on that.

As pointed out by Biemer (2001), in his discussion of Platek and Särndal (2001), there is a lack of routine measurements of MSE components in statistical organizations. There are good reasons for this state of affairs. Complexity has already been mentioned and to that we can add factors such as costs, the fact that it is almost impossible to publish such information at the time data are released, and that there is no measure of total error that would take all error sources into account, either because a lack of proper methodology or that some errors defy expression. Groves and Lyberg (2010) list some other weaknesses associated with the total survey error paradigm. For instance, we need to know more about the interplay between variances and biases. It is possible that an increase in simple response variance goes hand in hand with a reduction in response bias, say, when we compare interview mode with self-administrative alternatives. Recently, West and Olson (2010) showed that interviewer variance can occur not only from individual interviewers' effect on the responses within their assignments but also because individual interviewers successfully obtain cooperation from different groups of sample members.

Despite all its limitations, the strengths of the total survey error framework are quite convincing. The framework provides a taxonomic decomposition of errors, it separates variance from bias and observation from nonobservation, and it defines the different steps in the survey process. It serves as a conceptual foundation of the field of survey methodology, where subfields are defined by their associated error structures. Finally, it identifies the gaps in the research literature since any typology will show that some process steps are more "popular" than others. Just compare the respective sizes of the literatures on data collection and data processing.

It seems, however, as if the total survey error framework needs some expansion along lines some of which were identified half a century ago. We need some guidance on trade-offs between measuring error sizes and making processes more error-free. Spencer's (1985) question is: how much should we spend on measuring quality versus quality enhancement? We also need some guidance on how to integrate additional notions into the framework, so that it becomes a total survey quality framework rather than a total survey error framework (Biemer 2010). For instance, if "fitness for use" predominates as a conceptual base, how can we launch research that incorporates error variation associated with different uses? This aspect will be discussed in the next section.

### 3. Quality management philosophies in survey organizations

During the late 1980's and the early 1990's some statistical organizations were under severe financial pressure and in some cases simultaneously criticized for not being sufficiently attentive to user needs. Governments in Sweden, Australia, New Zealand and Canada as well as the Clinton administration in the U.S. were all keen on improving efficiency and user influence within their respective statistical systems. It was natural for these organizations to look for inspiration in management theories and methods (Drucker 1985) and specifically on what was called quality management (Juran and Gryna 1988). In that newer literature it was possible to study the role of the customer, leadership issues, the notion of continuous quality improvement, and various tools that could help the statistical organization improve. Especially influential to survey practitioners was work by Deming (1986), since he emphasized the role of statistics in quality improvement. He vigorously promoted the idea that improvement work should be led by statisticians, since they are trained in distinguishing between different kinds of process variation. He thought that there were too few statistical leaders advising top management in businesses and he wanted more proactive statisticians to become such leaders. He was especially keen on developing Shewhart's ideas about control charts as a means to distinguish between the different types of variation, namely common and special cause variation. Shewhart's improvement cycle Plan-Do-Check-Act was also part of Deming's thoughts on quality (Shewhart 1939).

Management principles have, of course, existed since ancient times. Juran (1995) provides lots of examples of what was in place in, for instance, the Roman empire. Craftsmanship and a guild system were basic building blocks. There were methods for choosing raw materials and suppliers. Processes were inspected and improved. Workers were trained and motivated and customers got warranties. All these features are found also in today's management systems. The more modern development includes quality frameworks or business excellence models such as Total Quality Management (TQM), International Organization for Standardization (ISO) standards, the Malcolm Baldrige quality award criteria, the European Foundation for Quality Management (EFQM), Six Sigma, Lean Six Sigma, and the Balanced Scorecard. These models are not totally different. They often share a common set of values and common criteria for excellence. Rather they represent a natural development that can be seen in all kinds of work.

Thus, there has been a gradual adoption of quality management models and quality strategies in statistical organizations and a merging with concepts and ideas already used in statistical organizations. My personal timeline for this development is the following (readers are invited to come up with different sets of events and dates):

- 1875 Taylor introduces what he called scientific management;
- 1900-1930 Taylor's ideas are used in, for instance, Ford's and Mercedes Benz's assembly lines;
- 1920's Fisher starts developing theories and methods for experimental design;
- 1924 Shewhart develops the control chart;
- 1940 The U.S. War Department develops a guide for analyzing process data;
- 1944 Deming presents the first typology of survey errors;
- 1944 Dodge and Romig present theory and tables for acceptance sampling;
- 1946 Deming goes to Japan;
- 1950 Ishikawa suggests the fishbone diagram as a tool for identifying factors that have a profound effect on the process outcome;
- 1954 Juran goes to Japan;
- 1960 Many businesses embark on a zero defects program;
- 1960 The U.S. Census Bureau quality control programs are developed;
- 1961 The U.S. Census Bureau survey model is launched;
- 1965-1966 Kish and Slobodan Zarkovich start talking about data quality rather than survey errors;
- 1970's Many statistical organizations provide quality guidelines;
- 1975 The Total Quality Management (TQM) framework is launched;
- 1976 The first quality framework in a statistical organization containing more dimensions than relevance and accuracy;
- 1987-1989 Launching of the ISO 9000, Malcolm Baldrige Award, Six Sigma and EFQM models;
- 1990's Many statistical organizations start working with quality improvement and excellence models;
- 1997 The Monograph on Survey Measurement and Process Quality;
- 1998 Mick Couper introduces the concept "paradata" as a subset of process data;

2001 The Eurostat leadership group on quality organizes the first conference on Quality Management in Official Statistics;

2007 Business architecture ideas enter the survey world.

From the mid 1990's and on quality management philosophies have had an enormous effect on many statistical organizations. The effect is not necessarily higher quality across the board (no one has checked that). But the philosophies have led to an awareness in most organizations of the importance of good contacts with users and clients, and an aspiration in many of them to become "the best" or "world class". Quality is on the agenda.

### 3.1 The concept of quality

During the last decades it has become obvious that accuracy and relevance are necessary but not sufficient when assessing survey quality. Other dimensions are also important to the users. The development of survey quality frameworks has taken place mainly within official statistics and has been triggered by the rapid technology development and other developments in society. These advanced technologies have created opportunities and user demands regarding potential quality dimensions such as accessibility, timeliness, and coherence that simply were not emphasized before. Decision-making in society has become more complex and global resulting in demands for harmonized and comparable statistics. Thus, there is a need for quality frameworks that can accommodate all these demands. Several frameworks of quality have been developed and they each consist of a number of quality dimensions. Accuracy and relevance are just two of these dimensions.

For instance, the framework developed by OECD (2011) has eight dimensions: relevance, accuracy, timeliness, credibility, accessibility, interpretability, coherence, and cost-efficiency (Table 1). Similar frameworks have been developed by Statistics Canada (Statistics Canada 2002; Brackstone 1999), and Statistics Sweden (Felme, Lyberg and Olsson 1976; Rosén and Elvers 1999). The Federal Statistical System of the U.S. has a strong tradition in emphasizing the accuracy component (U.S. Federal Committee on Statistical Methodology 2001) although it certainly appreciates other dimensions. Perhaps they are viewed as dimensions of a more nonstatistical nature that still need a share of the total survey budget. The International Monetary Fund (IMF) has developed a framework that differs from those of OECD, Australian Bureau of Statistics, Statistics Sweden, and Statistics Canada. IMF's framework consists of a set of prerequisites and five dimensions of quality: integrity, methodological soundness,

accuracy and reliability, serviceability, and accessibility (see Weisman, Balyozov and Venter 2010).

**Table 1**  
**OECD's quality framework**

Dimension	Description
Relevance	Statistics are relevant if users' needs are met.
Accuracy	Closeness between the value finally retained and the true, but unknown, population value.
Credibility	The degree of confidence that users place in data products based on their image of the data provider.
Timeliness	Time length between data availability and the event or phenomenon data describe.
Accessibility	How readily data can be located and accessed from within data holdings.
Interpretability	The ease with which the data user may understand and properly use and analyze the data.
Coherence	Reflects the degree to which data products are logically connected and mutually consistent.
Cost-efficiency	A measure of the costs and provider burden relative to the output.

Without sufficient accuracy, other dimensions are irrelevant but the opposite is also true. Very accurate data can be useless if they are released too late to affect important user decisions or if they are presented in ways that are difficult for the user to access or interpret. Furthermore, quality dimensions are often in conflict. Thus, providing a quality product is a balance act where informed users should be key players. Typical conflicts exist between timeliness and accuracy, since it takes time to get accurate data through, for instance, extensive nonresponse follow-up. Another conflict is the one between comparability and accuracy since application of new and more accurate methodology might disturb comparisons over time (Holt and Jones 1998).

Thus, many organizations have adopted a multi-faceted quality concept consisting not only of accuracy but also other dimensions. We might talk about a quality vector whose components vary slightly between organizations both in number and in contents. There are a number of problems associated with the quality vector approach.

First, the development has not been preceded by user contacts. Producers of statistics have believed that users are interested in a specific set of dimensions even though it is obvious that a vast majority of users think that error structures are too complicated to grasp and assume that the producer should be responsible for delivering the best possible accuracy. In cases where the user or client has specific accuracy requirements a more in-depth dialog can take place between the two. In the rare studies that have investigated user perceptions of information on quality it turns out that users are mostly interested in dimensions that are easily understood, such as timeliness and indicators that are seemingly straight forward, such as response rates. The

user wants the producing statistical organization to be credible, which translates into being capable of producing data with small or at least known errors and delivering them in a timely, reliable, and accessible fashion. The thought that it would be possible to produce a total quality measure based on weighted assessments of the different dimensions is not realistic, although Mirotschie (1993) argues to the contrary. In that paper Mirotschie makes a case for a standard set of quality indicators and provides a hypothetical illustration of indexing data quality indicators and computing an actual index (in this illustration the indicators are precision, nonresponse, reliability, timeliness and residuals). Even if a composite indicator in the form of an index were a possible development, the user would like to know which indicators contributed most to an index value. From a user's point of view the least favorable index value could still reflect a situation providing the highest quality. Rarely can a low accuracy be compensated by good ratings on other dimensions, not even in the case of election exit polls where timeliness is imperative. Accuracy is still necessary and there is wide agreement that all reputable organizations should meet accuracy standards (Scheuren 2001; Kalton 2001; Brackstone 2001). Phipps and Fricker (2011) provide an overview of quality frameworks and literature on total survey error. Thus, we can agree that survey quality is a multi-faceted concept involving multiple features of a statistical product or service.

### 3.2 The quality movement's impact on statistical organizations

Just extending the quality framework from one or two dimensions to several is not sufficient to create a quality environment. In the late 1980's and early 1990's many statistical organizations became interested in quality issues beyond traditional aspects of data quality. Issues concerning customer satisfaction, communicating with customers, competition, process variability, cost of poor quality, waste, business excellence models, core values, best practices, quality assurance, and continuous quality improvement were suddenly part of the everyday activities in many organizations.

Successful organizations know that continuous improvement (Kaizen) is necessary to stay in business and they have developed measures that help them change. This is true also for producers of statistics. Changes that are supposed to improve the statistical product are triggered by user demands, competition from other producers and from producer values that emphasize continuous improvement as part of the general business environment. The measures that can help a statistical organization improve are basically identical to those of other businesses. They can be built on business excellence models such as the European Foundation for

Quality Management (EFQM) (1999). The core values of the EFQM model include results orientation, customer focus, leadership and constancy of purpose, management by process measures and facts, personnel development and involvement, continuous learning, innovation and improvement, development of partnerships, and public responsibility. This model has been adopted by the European Statistical System (ESS) as a tool for national statistical institutes in Europe for achieving organizational quality. The thought is that good product quality, according to the dimensions mentioned (or some other product quality definition) cannot be achieved without good underlying processes used by the organization. It can also be argued that good product quality is achieved most efficiently and reliably by good process quality. If we view quality as a three-level concept it can be visualized as shown in Table 2.

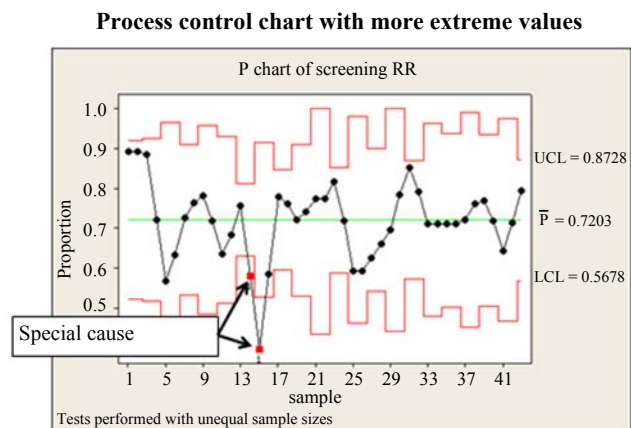
### 3.2.1 Product quality

The deliverables agreed upon are called the product. It can be one or several estimates, datasets, analyses, registers, standard processes or other survey materials such as frames and questionnaires. Product quality is the traditional quality concept used when informing users or clients about the quality of the product or service. It can be measured and controlled by means of degree of adherence to specifications and requirements for product characteristics adding up to quality dimensions of a framework. Measures of accuracy and margins of error belong here. Also observations whether service levels agreements established with the client have been accomplished are relevant. In line with quality management principles, it is also quite common to conduct user satisfaction surveys to find out what users think about the products and services that are provided.

### 3.2.2 Process quality

All processes have to be designed so that they deliver what they are supposed to. This means that we have to have some kind of quality assurance perspective when processes are defined. For instance, the process of interviewing implies that a number of elements must be in place for the

process to deliver what is expected. Examples of elements are an effective selection of interviewers and a training program, a compensation system as well as supervision and feedback activities. Thus we aim at building quality into the process via the quality assurance. Quality control efforts are only used to check if the process works as intended. It cannot by itself be used to build quality into the process. In Section 4.4 this process view is discussed in more detail. Process quality is measured and controlled via selection, observation and analyses of key process variables, so called process data or paradata (Morganstein and Marker 1997; Couper 1998; Lyberg and Couper 2005). Theory and methods imported from statistical process control can help the producer distinguish between the two types of variation, common and special cause. As long as all variation is contained within the upper and lower control limits associated with the control charts chosen, the process is said to be in statistical control and no process improvements are really possible by trying to adjust individual outcomes. If there are observations falling outside of the control limits, usually set at 3 sigma, then we have indications of special cause variation that should be taken care of so that the variation after adjustment is brought back to common cause variation. The following P-chart illustrates a possible situation:



**Table 2**  
Quality as a three-level concept\*

Quality level	Main stake-holders	Control instrument	Measures and indicators
Product	User, client	Product specs, SLA, evaluation studies, frameworks, standards	Frameworks, compliance, MSE, user surveys
Process	Survey designer	SPC, charts, acceptance sampling, risk analysis, CBM, SOP, paradata, checklists, verification	Variation via control charts, other paradata analyses, outcomes of periodic evaluation studies
Organization	Agency, owner, society	Excellence models, ISO, CoP, reviews, audits, self-assessments	Scores, strong and weak points, user surveys, staff surveys

\* SLA (Service Level Agreement), SPC (Statistical Process Control), CBM (Current Best Methods), SOP (Standard Operating Procedures), and COP (ESS Code of Practice).



Thus, the action sequence is the following. First the roots of the special causes are taken care of so that these variations are eliminated. After that the process displays common cause variation only. If that variation is deemed too large then the process has to change. The kinds of changes necessary are seldom obvious at the outset. Indeed perhaps several are necessary to decrease the process variation. Typically, a process improvement project is needed and the quality management literature has promoted a number of tools that are useful in such projects. Most of these tools are borrowed from statistics (control charts, experiments, regression analysis, Pareto diagrams, scatter plots, stratification) but there are also tools for identifying probable problem root causes (fishbone diagrams, process flow charts, brainstorming). The common thinking is that improvement projects should be “manned” by people working with the process or by people very much familiar with the process in other ways. Sometimes, we talk about forming an improvement team, where also the client or customer participates. In any improvement work suggested changes have to be tested. When Shewhart first developed his control charts he also suggested that improvement work should follow a sequence of operations, Plan-Do-Check-Act. What this sequence tells us is that any process changes suggested should be tested to see if they actually improve the process. If not, another change is made, and testing done again. Deming called this line of thinking the Shewhart cycle but since Deming spent a lot of time promoting it, many eventually called it the Deming cycle. The changes sought after could be decreased process variation, reduced costs, or increased customer satisfaction. The improvement project methodology is described in for instance Joiner (1994), Box and Friends (2006), Breyfogle (2003), and Deming (1986).

Another way of checking the process quality is to use acceptance sampling. Acceptance sampling (Schilling and Neubauer 2009) can be applied in situations where process elements can be grouped in batches. The batches are controlled and based on the outcome of that control it is decided whether the batch should be approved or reworked. Acceptance sampling plans guarantee an average outgoing quality in terms of, say, error rate, but there is no direct quality improvement involved. It is a control instrument that is suitable for operations such as coding, editing and scanning and then only when these processes are not really in statistical control. The method has been heavily criticized by Deming (1986) and others but can be the only control means available in situations where staff turnover is high and there is no time to wait for stable processes.

Global paradata (Scheuren 2001) are “error” rates of different kinds. Examples include nonresponse rates, coding error rates, scanning error rates, listing error rates, *etc.* In

some operations the error rates are calculated using verification, which means that the operation is repeated in some way. That is the case for the coding operation. In other operations the calculation can be based on a classification scheme, which is the case for nonresponse rate calculation. These global paradata tell us something about the process. They are process statistics, *i.e.*, summaries of data. A large nonresponse rate indicates problems with the data collection process and a high coding error rate indicates problems with the coding process. From these summaries it is sometimes possible to distinguish common and special cause variation and decide what action to take.

Some standardized processes can be controlled by means of simple checklists. Checklists are very effective when it is crucial that every process step is made and in the right order (Morganstein and Marker 1997). This is the case when airline pilots prepare for take-off. No matter how many times they have taken off, without a checklist the day will come when they forget an item. In statistics production sampling is such a process, albeit with less severe consequences if items are missed. It might very well be the case that a statistical organization has a standardized process for sample selection and a checklist that can be used as a combination of work instruction and control instrument.

There is a kind of checklist that can be used in more creative processes such as the overall survey design process. It is not possible to standardize the survey design process but it is possible to list a number of critical steps that always must be addressed. The list does not tell us how to address them. It just serves as a reminder that an individual step should not be omitted or forgotten. Morganstein and Marker (1997) discuss this kind of checklist and call them (and the simpler checklists) Current Best Methods (CBM). They describe the CBM development process and how the CBMs can be used to decrease the process variation in statistical organizations. For instance, an organization might have seven different imputation methods and systems in its toolbox. It is costly to maintain these seven systems. It is unlikely that they are equally efficient. If they are, it may not be economically feasible to keep them all. In this situation a CBM that describes fewer options to the organization seems like a good idea. This could be accomplished by forming an improvement team consisting of the imputation experts and some clients. CBMs are supposed to be revised when new knowledge is obtained, which implies that there is an expiration date associated with every CBM.

CBMs are of course “best practices” in some sense. Many organizations want best practices implemented and used. Morganstein and Marker offer a process for developing these best practices and keeping them current. It is beneficial for an organization if the variation in process

design can be kept at a minimum. It then becomes easier to train people and change the process when it becomes unstable or when new methods are developed. On the other hand, if CBMs and other standards are not vigorously enforced within an organization, they will not be widely used and the investment will not pay off.

### 3.2.3 Organizational quality

Management is responsible for quality in its widest sense. It is the organization that provides leadership, competence development, tools for good customer relations, investments, and funding. The quality management field has given us business excellence models that can help us evaluate our statistical organizations in the same way other businesses are evaluated. The two main business excellence models are the Baldrige National Quality Program and the European Foundation for Quality Management (EFQM).

These models consist of criteria to be checked when assessing an organization. The Malcolm Baldrige award uses seven main criteria: Leadership, strategic planning, customer and market focus, information and analysis, human resource focus, process management, and business results. Each criterion has a number of subcriteria. For instance, human resource focus consists of work systems, employee education, training and development, and employee well-being and satisfaction. The EFQM model has nine criteria: Leadership, strategy, people, partnerships & resources, processes, products & services, customer results, people results, society results, and key results. These models can be used for self-assessment or external assessment. The organization provides a description of what is in place regarding each criterion and the organization is scored based on that description. Typically self-assessments result in higher scores than external ones. It is very difficult to get a high score from external evaluators since the models are very demanding. For each criterion the organization is asked if it has a good approach in place somewhere in the organization. This is often the case. The next question is how wide-spread this good approach is within the organization. Many organizations lose momentum here, since there is very little truth in the mantra “the good examples are automatically spread throughout an organization”. Instead good approaches usually have to be vigorously promoted before they are accepted within the organization. The third question asks whether the approach is periodically evaluated to check if it achieves the results expected. This is where most organizations fail. Their usual strategy is to exhaust an approach until the problems are so great that the approach has to be replaced rather than adjusted. This strategy is, of course, disruptive and expensive and does not score highly in excellence assessments. The maximum number of points

that can be obtained using these models is 1000 and very rarely does a winner get more than 450-600 points, which is an indication that there is a lot of room for improvement even in world class organizations.

Some statistical organizations have used business excellence models for assessment. The Czech Statistical Office was announced Czech National Quality Award Winner for 2009 in the Public Sector category based on EFQM. The office got 464 points. Eurostat’s leadership group on quality recommended the European national statistical offices to use the EFQM as a model for their quality work and Finland and Sweden are among those that have done so. Since the leadership group released its report in 2001 (see Lyberg, Bergdahl, Blanc, Booleman, Grünwald, Haworth, Japec, Jones, Körner, Linden, Lundholm, Madaleno, Radermacher, Signore, Zilhao, Tzougas and van Brakel 2001) other frameworks and standards have been developed. The European Statistical System has launched its Code of Practice, which consists of a number of principles with associated indicators. Regarding some principles, however, the indicators are more like clarifications. The list of principles resembles other lists that have been developed by the UN and other organizations.

External assessments are probably more reliable than internal ones. There are a number of reasons for that. One is that it is difficult to criticize your peers since you have to interact with them in the future or if your own product or service will be assessed by those peers in the future. Experiences from Statistics Sweden and Statistics Canada show that self-assessments are limited in their capability of identifying serious weaknesses (see Section 5.3).

### 3.2.4 Some specific consequences for statistical organizations

Most statistical organizations have adopted quality management ideas to varying degrees and with varying success. As pointed out by Colledge and March (1993) it is possible to list a number of obstacles associated with such implementation. For a government agency it can be difficult to motivate its staff through monetary incentives, since there are restrictions on how tax money can be spent. The variety of users and products makes the dialog between the service provider and the user complicated and as mentioned neither the users, or for that matter the providers are totally familiar with all the biases and other quality problems that are present in statistics production. The effect of errors on the uses can vary and are often unknown. To complicate matters further, unlike most other businesses, suppliers are not very enthusiastic. In other businesses suppliers get paid while statistical organizations must motivate theirs, the respondents, who are seldom even given a cash incentive.

On the other hand statistical organizations have a great advantage when it comes to applying quality management principles. A statistical organization knows how to collect and analyse data that can guide improvement efforts. One of the cornerstones in quality management philosophies is that decisions should be based on data and businesses that do not have support from statisticians are often unaware of data quality problems, which can have consequences for their decision-making. By and large, though, a statistical organization is not different from any other business and it is quite possible to apply quality management ideas to improve all aspects of work.

#### 4. Examples of quality initiatives in statistical organizations

In this section we will provide some examples of initiatives that statistical organizations have engaged in as a result of a general interest in quality in society.

##### 4.1 The total survey error

Perhaps the most important thing to notice is that research and development in survey design, implementation, sampling and nonsampling errors, and the effect of errors on the data analysis continue to thrive. Data with small errors is the major goal for reputable organizations, which is indicated by the steady flow of textbooks on data collection, sampling, nonresponse, questionnaire design, measurement errors, and comparative studies. New textbooks are in progress covering gaps such as business surveys, translation of survey materials, and paradata. There are journals such as the *Journal of Official Statistics*, *Survey Methodology*, and *Survey Practice* that are entirely devoted to topics related to statistics production in a wide sense. Numerous other journals such as the *Public Opinion Quarterly*, the *Journal of the American Statistical Association*, and the *Journal of the Royal Statistical Society* devote much space to survey methods. The Wiley series on Survey Methodology and its associated conferences (on panel surveys, telephone survey methods (twice), measurement errors, process quality, business surveys, testing and evaluating questionnaires, computer assisted survey information collection, nonresponse, and comparative surveys) have been very successful and that is the case also for the continuing workshops on nonresponse and total survey error. Thus, there is no shortage of ideas regarding specific error sources and their treatment. Admittedly there are areas that are understudied such as specification errors, data processing errors and the impact of errors on the data analysis but by and large there is a healthy interest in knowing more about survey errors. The challenge lies in communicating this knowledge to people working in

statistical organizations and in developing design principles that can be used to improve statistics production. There is a noticeable gap between what is known through research and what is known and applied in the statistical organizations. Thus, staff capacity building seems to be a continuing need, especially since the common idea that good examples spread like ripples within and between organizations is a myth. If that indeed were the case quality would by now be fantastic everywhere. Since it is not, many organizations have developed extensive training programs (Lyberg 2002).

##### 4.2 Risk and risk management

One element of quality management that has entered the survey world is risk and risk management. Eltinge (2011) even talks about Total Survey Risk as an alternative to the total survey error paradigm. The identification and management of risks is an important part of modern internal auditing (Moeller 2005) and is perhaps the only major element that is missing in quality management frameworks such as EFQM. An error source can be seen as more risky than another and should, therefore, be handled with more care and resources than another less risky. For instance, not having an effective system for statistical disclosure control is seen as a very risky situation. Unlawful data disclosure is very rare historically, but when it happens it could potentially destroy future data collection attempts. Certain design decisions can be seen as risky. For instance, if we choose a data collection method that does not fit the survey topic we might get estimates that are so far from the truth that the results are useless. An example might be to study sensitive behaviors using face to face or telephone interviewing instead of a self-administered mode. There are also technical risks that need to be identified and assessed. For instance, the U.S. National Agricultural Statistical Service (Gleaton 2011) like many others has plans for disaster recovery. Groves (2011) and Dillman (1996) both discuss how the production culture and the research culture within a statistical organization might view risks in different ways. Change in statistical organizations is generally slow and there are so metimes good reasons for that. Change might result in failures such as unsuccessful implementation, large costs and decreased comparability. So in some sense both producers and users have a tendency to be hesitant toward changes suggested by researchers and innovators and that might be one reason why change takes a long time. It is very common to have parallel measurements for some time to handle risks associated with implementing a new method or system. According to Groves (2011) the production culture and the users have had the final say about any changes, at least up until now. At the same time innovation is badly needed in many production systems and there are examples of stove-pipe organizations that do not have much time left

(to remain unchanged) because the resources to maintain their systems are simply not there. So even though there is resistance against change, lack of resources and competition will make sure that statistical organizations become more process-oriented and efficient. Reducing the number of systems and applications and developing and using more standardization seem to be one road forward.

### 4.3 The client/customer/user

The advent of quality management ideas in statistical organizations has made the receivers of statistical products and services more visible. Commercial firms have always talked about the client or the customer while government organizations have tended to call them users. In any case the recognition of someone who is supposed to use the endproducts has not been obvious to some providers. Admittedly the user has been a speaking partner since the beginning of the survey industry. In the U.S., conferences for users were quite frequent already 50 years ago (Dalenius 1968; Hansen and Voight 1967). During six months 1965-66, for example, the U.S. Census Bureau organized 23 user conferences across the country and there were also advisory groups. The advisory nature of contacts with users has prevailed in many countries. The user conference format still exists but user input is now complemented by other means such as public discussions and internet forums. Rarely have users been directly involved in the planning and design of surveys. Even when it comes to discussions about the quality of data, producers have acted as stand-in users. The quality frameworks are a good example. The quality dimensions were defined with minimal consultation with users. The literature on how users perceive information about quality is extremely limited (Groves and Lyberg 2010). Also, we do not know if the information on quality that we provide is useful to them (Dalenius 1985b). In fact, an educated guess is that many times it is not. In many surveys the users are many and sometimes unknown and their information and analytical needs cannot be foreseen ahead of time. It is often possible to single out one or a few main users to communicate with, but many of the design and quality problems are so complicated that a vast majority of users expect the service provider to deliver a product with the smallest possible error. Hansen and Voight stated that accuracy should be sufficient to avoid interpretation problems. Today there seems to be consensus among many that what users are interested in are products and services that can be trusted, *i.e.*, the service provider should be credible. It is impossible for most users to check levels of accuracy. Aspects that an average user can discuss are issues such as timeliness, accessibility and relevance. Detailed discussions about technical matters and design trade-off

issues including accuracy and comparability are more difficult to have.

During recent decades the user has indeed become more prominent. Some organizations develop service level agreements together with a main user or client, where requirements of the final product or service are listed and can be checked at the time of delivery. Many organizations conducting business surveys have created units that continuously communicate with the largest businesses, since their participation and provision of accurate information is absolutely essential for the estimation process (Willimack, Nichols and Sudman 2002). The large businesses are not users in the strict sense. They are important suppliers often with an interest in the survey results. Another common communication tool is the customer satisfaction survey. The value of such surveys is limited due to the acquiescence phenomenon and problems finding a knowledgeable respondent who is also willing to respond. Also, many customer satisfaction surveys are based on self-selection resulting in zero inferential value. In those surveys the results can only be viewed as lists of issues and concerns that some customers convey. Such information can, of course, be very valuable but is not suitable for estimation purposes. Many survey organizations now conduct user surveys on a continuing basis (Ecochard, Hahn and Junker 2008).

### 4.4 The process view

Quality management has reemphasized the importance of having a process view in statistics production. To view the production process as a series of actions or steps towards achieving a particular end that satisfies a user, leads to a good product quality. Process quality is an assessment of how far each step meets defined requirements or specifications. One way of controlling the process quality is to collect process data that can vary with each repetition of the process. The interesting process variables to monitor are those that have a large effect on the process's end result. Thus to check a process for stability and variation we need mechanisms for identifying, collecting and analysing these key process variables. The quality management science has given us tools such as the Ishikawa fishbone diagram to identify candidates for key process variables. The statistical process control methodology has given us tools to distinguish between special and common cause variation and how to handle these two variation types. Usually we use control charts originally developed by Shewhart (Deming 1986; Mudryk, Burgess and Xiao 1996) to make those distinctions. Then, again, we use methods from quality management to adjust the process if necessary. Examples include flowcharts, Pareto diagrams, and other simple means for the production team to identify the root causes of problems (Juran 1988).

Process data have been used to check on processes used in statistics production since the 1940's, first within the U.S. Census Bureau and then at Statistics Canada and to some extent also in other agencies. Typical processes that were checked included coding, keying and printing and the process data were mainly error rates. Some of the process checks used at the U.S. Census Bureau were so complicated and expensive that their value was questioned (Lyberg 1981), especially since the associated feedback loops were inefficient and not always aiming for the root causes of the errors. It was common that operators were blamed for system problems and at the time there was no emphasis on continuous quality improvement. The thinking at the time was more directed toward verification and correction.

Morganstein and Marker (1997) developed a generic plan for process continuous improvement that can be used in statistics production. They had worked in many statistical organizations since the 1980's and observed that quality thinking was not really developed in most of them. Their generic plan was built on their first-hand experiences and the general quality management ideas laid out by *e.g.*, Juran (1988), Deming (1986), Box (1990), and Scholtes, Joiner and Streibel (1996). In essence the plan consists of seven steps:

- The critical product characteristics are identified together with the user, both broad and more single effort needs.
- A map of the process flow is developed by a team familiar with the process. The map should include the sequence of process steps, decision points and customers for each step.
- The key process variables are identified among a larger set of process variables.
- The measurement capability is evaluated. It is important that decisions are based on good data, not just data. Available data might be useless. This is an area where statistical organizations should have an advantage over other organizations. One should not reach conclusions about process stability without knowledge about measurement errors. Above all, data should allow quantification of improvement.
- The stability of the process is determined. The variability pattern of the process data is analyzed using control charts and other statistical tools.
- The system capability is determined. If stability is not achieved after special cause variation has been eliminated an improvement effort is called for. System changes must be made when the process variation is so large that it does not meet specifications, such as minimum error rates or production deadlines. Typical methods to reduce variation are the development and implementation of a new training program or the

enforcement of a standard operating procedure. The latter can be a process standard, a current best methods standard or a simple checklist.

- The final step of the improvement plan is to establish a system for continuous monitoring of the process. We cannot expect processes to remain stable over time. For many reasons they usually start drifting after some time. A monitoring system helps keeping track of new error structures, new customer requirements, and the potential of improved methods and technology and can suggest process improvements.

The Morganstein and Marker book chapter had a distinct effect on quality work and process thinking in many European statistical organizations. Interest in these issues increased and some organizations started their own quality management system where process improvement was central.

At the 1998 Joint Statistical Meetings Mick Couper presented an invited paper on measuring quality in a CASIC environment. He meant that the new technology generated lots of by-product data that could be used to improve the data collection process. He named those paradata, not in his paper but in his session presentation. This naming caught on very quickly in the survey community and it made sense to define the trilogy data, metadata, and paradata. Thus we had one term for data about the data (metadata) and another for data about the process (paradata). Obviously paradata are process data but for a long time paradata were confined to data about the data collection process, while the term used in many European statistical organizations was "process data" and took all survey processes into account (Aitken, Hörngren, Jones, Lewis and Zilhao 2004). Recently a renewed broadening of the meaning of the concept has taken place. Kennickell, Mulrow and Scheuren (2009) remind us about what they call macro paradata, global process data such as response rates, coverage rates, edit failure rates, and coding error rates that always have been indicators of process quality in statistical organizations. Lyberg and Couper (2005), Kreuter, Couper and Lyberg (2010), and Smith (2011) also use the more inclusive meaning of paradata where other processes than data collection are taken into account. There is a risk that paradata, like quality, becomes an overused concept. There are examples of discussions where all data, apart from the survey estimates, are considered paradata, which, of course, does not make sense.

Paradata is a great naming and they are necessary to judge process quality. However, a word of caution is in place. One should never collect paradata that are not related to process quality and it is important to know how to analyze them. Sometimes statistical process control methods

can be used but at other times other analytical techniques are needed. For instance, to be able to control interviewer falsification we might need to look at several processes simultaneously, but the theory and methodology for such analysis might not be readily available.

The expanded use of microdata that concern individual records, such as keystroke data and flagged imputed records, is an effect of using new technology. Modern data collection procedures generate enormous amounts of these kinds of paradata but so do systems for computer-assisted manual coding and systems for pure automated coding as well as systems for scanning of data. It makes no sense to confine the concept to data collection.

Quality management has taught us to prevent process problems rather than fix them when they appear, that it is important to distinguish between different types of process variation since they require different actions, that any process intervention or improvement should be based on good data and proper analysis methods, and that even stable processes eventually start drifting, which calls for continuous monitoring.

#### 4.5 Standardization and similar tools

One way of keeping process quality in control is to reduce variation by encouraging the use of standards and similar documents. Colledge and March (1997) discuss four classes of documents.

- A standard is a document that should be adhered to almost without exception. Deviations are not recommended and require approval of senior management. Corrective action should be taken when a standard is not fully met. An organization can become certified according to a standard. This is the case for ISO standards, where a few are relevant to statistical organizations.
- A policy should be applied without exceptions. For instance, an organization can have a policy regarding the use of incentives to boost response rates.
- Several organizations have developed guidelines for different aspects of the statistics production. Typically, guidelines can be skipped if there are “good” reasons to do so.
- A recommended practice is promoted but adherence is not mandatory.

Admittedly, the categories of this classification scheme are not mutually exclusive, especially if we also take language and cultural aspects into account. For instance, in the Swedish language policies and guidelines are very close conceptually. If we consult the unauthorized but consensus based Wikipedia it says that “policies describe standards while guidelines outline best practices for following these

guidelines”. This sentence contains three of the categories mentioned by Colledge and March. It is probably best to relate to these different kinds of documents in a similar fashion. They all attempt to improve quality by reducing various types of variation and we should not dwell too much on what they are called.

Although standards have been an important part of survey methodology for a long time they have gained momentum since statistical organizations became interested in quality management. Early standards such as Hansen *et al.* (1967) and U.S. Bureau of the Census (1974) concentrated on discussing the presentation of errors in data. At the U.S. Census Bureau all publications should inform users that data were subject to error, that analysis could be affected by those errors, and that estimated sampling errors are smaller than the total errors. For major surveys the nonsampling errors should be treated in more detail unlike in the past. Many other statistical organizations imported this line of thinking. For instance, the quality frameworks mentioned earlier are expansions including also other quality dimensions than accuracy. The European Statistical System has successively developed and launched what was first called Model Quality Reports and currently just Standard for Quality Reports (Eurostat 2009a). The standard provides recommendations to European National Institutes (notice the conceptual complexity) for preparation of quality reports for a “full” range of statistical processes and their outputs. The standard treats the basic quality dimensions relevance, accuracy, timeliness, accessibility, coherence and comparability.

Let us look at some examples. Regarding measurement error, which is part of the accuracy component, the standard says that the following information should be included in a quality report:

- Identification and general assessment of the main risks in terms of measurement error.
- If available, assessments based on comparisons with external data, reinterviews or experiments.
- Information on failure rates during data editing.
- The efforts made in questionnaire design and testing, information on interviewer training and other work on error reduction.
- Questionnaires used should be annexed in some form.

Regarding timeliness the standard says that the following information should be included:

- For annual or more frequent releases: the average production time for each release of data.
- For annual and more frequent releases: the percentage of releases delivered on time, based on scheduled release dates.
- The reasons for nonpunctual releases explained.

There are also sections on how to communicate information regarding trade-offs between quality dimensions, assessment of user needs and perceptions, performance and cost, respondent burden as well as confidentiality, transparency and security. Even though there is a section on user needs and perceptions, users have obviously not been involved in the preparation of the standard itself. We still know very little about how users perceive and use information about quality. The standard is backed by a much more detailed handbook for quality reports (Eurostat 2009b) and both documents are built around the 15 principles listed in the European Statistics Code of Practice, which is the basic quality framework for the European Statistical System. The Code of Practice principles concern professional independence, mandate for data collection, adequacy of resources, quality commitment, statistical confidentiality, impartiality and objectivity, sound methodology, appropriate statistical procedures, nonexcessive burden on respondents, cost-effectiveness, relevance, accuracy and reliability, timeliness and punctuality, coherence and comparability, and, finally, accessibility and clarity. Each principle is accompanied by a set of indicators that the individual organization can measure to establish whether it meets the Code or not. Some indicators are vague and very subjective in nature such as “the scope, detail and cost of statistics are commensurate with needs”, while others are more specific, such as “a standard daily time for the release of statistics is made public”. Peer reviews of compliance to a limited set of the principles have been conducted using an earlier version of the Code and, not surprisingly, many national statistical offices in Europe have problems living up to the Code (Eurostat 2011a). Therefore in order to assist the implementation of the Code a supporting framework has been developed, called the Quality Assurance Framework (QAF) that contains more specific guidance regarding methods and references (Eurostat 2011b). This seems to be a very useful document since its references are mainly summaries of the state-of-the-art in areas such as sampling, questionnaire design, editing and so on, which stimulates conformity to current best practices.

The Code of Practice has many similarities with the UN Fundamental Principles of Official Statistics (de Vries 1999). The latter promotes also the principle of international cooperation and coordination, which is, to a large extent, an element that is missing in today’s development of statistics production (Kotz 2005). Even neighbouring countries can have very different approaches and methodological competence levels and the differences are sometimes difficult to explain. Experience shows that global development collaboration is difficult to achieve. We meet, we talk, and we bring back ideas that might fit our own systems. It is harder to agree on common approaches. One global standard that

relates to statistics production is the ISO 20252 on market, opinion and social research (International Standards Organization 2006). This is a process standard with around 500 requirements concerning the research activities within an organization. It is a minimum standard for what to do rather than how to do things. It is suitable for organizations that conduct surveys and the organization can apply for certification. In April 2010 more than 300 organizations world-wide had been certified, most of them marketing firms. One national statistical office (Uruguay) was certified in 2009 and Statistics Sweden is planning a certification in 2013 but those are the only national offices that have chosen this path. The standard concerns the organization’s system for quality management, management of the executive elements of the research, data collection, data management and processing, and reporting on research projects (Blyth 2012).

The standards of the U.S. Federal Statistical System concentrate on the accuracy component. Although not formally a standard the U.S. Federal Committee on Statistical Methodology (2001) suggests various methods for measuring and reporting sources of error in surveys. In 2002 the U.S. Office of Management and Budget (OMB) issued information quality guidelines (OMB 2002) whose purpose was to ensure and maximize the quality, objectivity, utility, and integrity of information disseminated by federal agencies. OMB (2006 a) has also issued standards and guidelines for surveys. They are built in a standard fashion. First comes a standard such as “Response rates must be computed using standard formulas to measure the proportion of the eligible sample that is represented by the responding units in each study, as an indicator of potential nonresponse bias”. This standard is then followed by a number of guidelines on how to make the necessary calculations while the final guideline states that “If the overall nonresponse rate exceeds 20%, an analysis of the nonresponse bias should be conducted to see whether data are missing completely at random”. As in the case of the ESS standards, the OMB guidelines are complemented by a supporting document (OMB 2006b) that can facilitate adherence to the standards.

Most agencies in the decentralized U.S. Federal Statistical System have documents in place that adapt the OMB guidelines. For instance, the U.S. Census Bureau has its own statistical quality standards that goes into more technical detail compared to the OMB documents. Each standard is described via requirements and sub-requirements and they often provide very specific examples of studies that can be conducted. Examples of other U.S. agencies that have standards related to the quality of information disseminated include the National Center for Health Statistics, National Center for Education Statistics, and the

Energy information Administration. All these standards can be downloaded from the agencies' websites.

Statistics Canada has issued quality guidelines since 1985. They are similar to the ESS guidelines since not just accuracy is emphasized. But they are much more detailed and contain lots of references. A special feature is that for some processes the guidelines prescribe the use of statistical process control. No other agency seems to be doing that. The latest edition of the guidelines is provided in Statistics Canada (2009).

Many other statistical organizations in the world have their own quality standards. They are sometimes described as guidelines or standards and sometimes as business support systems or quality assurance frameworks. In any case, the contents and style vary across organizations but the variation should be manageable. It should be possible to achieve higher degrees of standardization globally, since that has happened in other fields, such as air travel. Apter, Carruthers, Lee, Oehm and Yu (2011) discuss various ways to industrialize the statistical production process at the Australian Bureau of Statistics.

The question is whether international standards would benefit survey quality in general. Some areas where standards would be beneficial include computation of frequently used quality indicators such as error rates and design effects, as well as best practices for translation of survey materials, handling non-native language respondents, and weighting for nonresponse. One must bear in mind that once a standard is issued it has to be continually updated and it is well-known that they can be difficult to enforce. If they are comprehensive, standards can overwhelm the practitioner and, as a result, unless mandated and audited, they are largely ignored.

#### 4.6 Statistical business process models

During recent years concepts like the business process models and business architecture have become part of quality work in some statistical organizations. To make production processes more efficient and flexible they can be seen as part of a business architecture model (Reedman and Julien 2010). In statistics production a generic statistical process model is jointly developed by UNECE, Eurostat, and OECD. Any system redesign should be driven by customer demands, risk assessments and new developments. The architectural principles behind this thinking are summarized in Doherty (2010), which discusses architecture renewal at Statistics Canada.

Some of the principles are:

- Decision-making should be corporately optimal, which entails centralization of informatics, methodology support and processing.
- Use of corporate services such as collection, data capture and dissemination should be optimized.
- Reuse should be maximized by having the smallest possible number of distinct business processes and the smallest possible number of computer systems.
- The corporate toolkit should be minimized.
- There should be staff proficiency in tools and systems.
- Rework such as repeated editing should be eliminated.
- The focus should be on the core business and the work with support processes should be outsourced.
- Development should be separated from the on-going operations.
- Electronic data collection should be viewed as the initial mode.
- Structural obstacles, such as overlapping or unclear mandates should be removed.

These principles are very similar to those we identify when we apply quality management principles from the various frameworks and excellence models described previously. The principles represent a move from decentralization to more corporate level thinking. Many statistical organizations realize that stove-pipe thinking is a thing of the past and that a move to more centralization is necessary.

## 5. Measuring quality

Thus, quality is a multi-faceted concept and measuring it is a complicated task. We have noted that survey quality can be viewed as a three-dimensional concept associated with the final product, the underlying processes that lead to the product, and the organization that provides the means to carry out the processes and deliver the product or service in a successful way. There are basically two ways to measure quality. One is to directly estimate the total survey error or some components thereof. The other is to measure indicators of quality with the hope that they indeed reflect the concept itself.

### 5.1 Direct estimates of the total survey error

The existing decompositions of the mean squared error described in, for instance, Hansen *et al.* (1964), Fellegi (1964), Anderson, Kasper and Frankel (1979), Biemer and Lyberg (2003), Weisberg (2005), and Groves *et al.* (2009) are all incomplete in the sense that they do not reflect all error sources. It is seldom possible to compute the MSE directly in practical survey situations because this usually requires a parameter estimate that is essentially error free. However, it is possible to obtain a second best estimate of the true parameter value if there are resources available to collect data using some "gold standard" methodology that is not affordable or practical in a normal survey setting. This is



the standard evaluation methodology when the true parameter value can be uniquely defined. Gold standard methods are seldom error-free but they can to varying extents provide better estimates, and the difference between the regular estimate and the gold standard estimate can serve as an estimate of the bias, which is the methodology used in census post enumeration surveys (United Nations 2010). Often an evaluation concerns a specific error component such as census undercount, nonresponse bias, interviewer variance or simple response variance, since we want information not on total survey error per se but rather on the components' relative contribution to the total survey error so that root causes of problems can be identified and relevant processes improved. Large evaluation studies are very rare since they are so demanding and their value is sometimes questioned (United Nations 2010). Smaller regular evaluation studies, on the other hand, are necessary to get indications of process and methodological problems.

## 5.2 Indicators of quality

Continuing reporting of total survey error is a formidable task and no survey organization does that. Instead organizations provide indicators or statements regarding quality. For instance, according to Eurostat's (2009a) handbook for quality reports the following indicators should be measured:

- Coefficient of variation;
- Overcoverage rate;
- Edit failure rate;
- Unit response rate;
- Item response rates;
- Imputation rates;
- Number of mistakes;
- Average size of revisions.

The common theme here is that these paradata summary items are indicators that can be calculated without conducting special studies. The set of indicators that can be calculated directly from the survey data is by definition quite limited and their value questionable. For instance, to include overcoverage but not undercoverage just because only the former can be calculated directly from the available data does not make sense. It is undercoverage that poses the greatest coverage problem in surveys. Admittedly, the handbook prescribes the producer to assess the potential for bias (both sign and magnitude) but it is not clear how this should be accomplished. The producer is urged to include evaluation and quality control results, if such information exists as well. Level of effort measures for processes such as questionnaire design and coder training would be welcomed. There is no standard reporting format for such qualitative and quantitative information. In any case, the key

indicator list becomes severely limited when compared to the full list of main error sources and it is hard to see how they are perceived by the users and how they can be used by the producer to improve the process.

The producer needs a more complete list of indicators to be able to measure or assess various levels of quality to make sure that the design implementation is in control or to be able to mount a quality improvement project. The initial survey design must be modified or adapted during the implementation to control costs and maximize quality. Biemer (2010) discusses four strategies for reducing costs and errors in real time, *i.e.*, continuous quality improvement (CQI), responsive design (Groves and Heeringa 2006), Six Sigma (Breyfogle 2003), and adaptive total design and implementation.

When the continuous quality improvement strategy is used, key process variables are identified and so are process characteristics that are critical to quality (CTQ). For each CTQ, real-time, reliable metrics for the cost and quality are developed. The metrics are continuously monitored during the process and intervention is done to ensure that costs and quality are within acceptable limits. The responsive design strategy was developed to reduce nonresponse bias in face to face interviewing. It includes three phases. In the experimental phase a few design options are tested (*e.g.*, regarding incentive level). In the main data collection phase the option chosen in the experimental phase is implemented and the implementation continues until phase capacity is reached. In the nonresponse follow-up phase special methods are implemented to reduce nonresponse bias and control the data collection costs. Such methods include the Hansen-Hurwitz double sampling scheme, increased incentives, and using more experienced interviewers. Again the efforts continue until further reductions of the nonresponse bias are no longer cost-effective. Six Sigma is the most developed business excellence model since it relies so heavily on statistical methods. It contains a large set of techniques and tools that can be used to control and improve processes. Adaptive total design and implementation combines control features of CQI, responsive design and Six Sigma and does that so that it simultaneously monitors multiple error sources. Biemer and Lyberg (2012) give several examples of CTQs and metrics for various survey processes. For instance, regarding the measurement process attributes that are CTQs might include the abilities to identify and repair problematic survey questions, to detect and control response errors, and to minimize interviewer biases and variances. Corresponding metrics might include missing data item by question, refusal rate by size of business, results of replicate measurements, suspicious edits actually changed, and field work results by interviewer. The metrics can be analyzed using statistical process control or

analysis-of-variance methodologies. Different related metrics can be displayed together in a dashboard fashion. For instance if one CTQ is the ability to discover interviewer cheating we might want to have a dashboard showing the metrics average interview length by interviewer and the distribution of some sensitive sample characteristic, also by interviewer.

### 5.3 Self-assessments and audits

The quality management philosophy has introduced the concepts of self-assessment and audit into statistics production. We are anxious to know what users, clients, owners and other stakeholders think about the products and services provided by the statistical organization. There are a number of tools available for this kind of evaluation. We have already mentioned the customer satisfaction survey. Other tools include employee surveys, internal audits and external audits. Customer surveys can shed light on what users think about products and services provided. They can be used to determine user needs and to identify what product characteristics really matter to the users. Another line of questioning might concern the image of the organization and how it compares to the images of other organizations, be they competitors or not. The customer satisfaction survey is very common in society. Often it cannot be used to make inference to the target population of users due to its methodological and conceptual shortcomings. The abundance of satisfaction surveys in society, developed and implemented by people with no formal training in survey methods, contributes to lukewarm receptions in more serious settings resulting in nonresponse and measurement errors. For instance, the 2007 Eurostat User Satisfaction Survey consisted of two separate surveys. One was launched on the Eurostat webpage and the target population consisted of 3,800 registered users. Only those registered users that entered the website during the data collection period were exposed to the survey request and this led to a response rate around 5%. The second survey used email that was sent to a number of main users identified by Eurostat. This more controlled environment generated a response rate of 28%. These surveys also have problems identifying the most suitable respondent. If the “wrong” respondent is chosen within an organization this will most certainly lead to uninformed and misleading results.

The simplest type of self-assessment is the questionnaire or checklist that is filled out by the survey manager. An example is one from Statistics New Zealand. It is a checklist that consists of a number of indicators or assertions such as “information needs are regularly assessed through user consultation”, “good and accessible documentation”,






“indicators of accuracy regularly produced and monitored”, and “presentation standards met”. The manager is asked to answer yes or no to each assertion and make a comment if deemed necessary. Statistics Sweden had a similar system in place where one of the questions was “has overall quality of your product improved, declined or stayed the same compared to last year?” When results were compiled for these three categories for the entire organization, a very small proportion of the managers reported declining quality, a somewhat larger proportion reported improved quality, while a vast proportion reported status quo. The managers simply did not have the proper means to assess overall quality. Furthermore, vague quantifiers like “regularly”, “good”, and “meeting standards” invite generous assessments. Also most managers do not want to look bad and status quo becomes a perfect escape route. This system of self-assessment was eventually abandoned by Statistics Sweden. It is possible to increase the value of these assessments by asking additional questions concerning details about how and when quality work was conducted. Some organizations use internal teams that audit important products. Julien and Royce (2007) describe a quality audit of nine products at Statistics Canada, where the purposes were to identify weaknesses and their root causes as well as identifying best practices. Review teams of assistant managers were formed so that each reviewer reviewed three different programs. The main weakness with an approach like this is the internal feature itself. Every reviewer knows that sooner or later it is his or her turn to be reviewed and there is a risk that this fact might hold them back. It is also internal in the sense that users are not explicitly present in the review process. In its general audit program on data quality management, however, Statistics Canada puts great emphasis on its user liaison system (Julien and Born 2006), which is one of the five systems forming the agency’s quality assurance framework, the others being corporate planning, methods and standards, dissemination, and program reporting.

A further variant of self-assessment is when it precedes an external audit. Statistics Netherlands (1997) describes how the Department of Statistical Methods is assessed by its staff. The assessment resulted in a listing of weak and strong areas that were later examined by an external team. Typically an external audit uses some kind of benchmark like a set of rules, a standard, or a code of practice for assessment purposes. The audit then results in a number of recommendations for the organization or the individual product or service.

Recently a general system for evaluating the total survey error has been developed at Statistics Sweden. Sweden’s Ministry of Finance wants quality evaluation results to be able to monitor quality improvements over time. Survey

quality must be assessed for many surveys, administrative registers, and other programs within the agency so there is need for some indicators that can serve as proxies for actual measures of quality. At the same time, the assessment process must be thorough, the reporting simple and the results credible. For each of the error sources specification, frame, nonresponse, measurement, data processing, sampling, model/estimation, and revision eight key products were rated poor, fair, good, very good, and excellent regarding each of five criteria. The criteria were knowledge of risks, communication with users, compliance with standards and best practices, available expertise, and achievement toward risk mitigation and/or improvement plans. The rating guidelines varied by criterion. For knowledge of risks they were:

**An Example of the rating guidelines – Knowledge of risks**

Poor 	Fair 	Good 	Very Good 	Excellent 
Internal program documentation does not acknowledge the source of error as a potential factor for product accuracy.	Internal program documentation acknowledges error source as a potential factor in data quality.	Some work has been done to assess the potential impact of the error source on data quality.	Studies have estimated relevant bias and variance components associated with the error source and are well-documented.	There is an ongoing program of research to evaluate all the relevant MSE components associated with the error source and their implications for data analysis. The program is well-designed and appropriately focused, and provides the information required to address the risks from this error source.
	But: No or very little work has been done to assess these risks	But: Evaluations have only considered proxy measures (for example, error rates) of the impact with no evaluations of MSE components	But: Studies have not explored the implications of the errors on various types of data analysis including subgroup, trend, and multivariate analyses	

The evaluation process started with a self-assessment done by each of the eight key products. These reports and other relevant documents were studied by two external reviewers who then met with product owners and their staff to discuss the product processes. After that the reviewers presented detailed assessments and scored each product. The procedure identified important areas to improve within but also across products. In this first evaluation round measurement error turned out to be a problematic area for

almost all the key products. As any other approach at measuring or indicating total survey error this one does not really reflect total mean squared error. It requires thorough documentation of processes and improvements made and it is highly dependent on the skills and knowledge of the external reviewers. This study is reported in Biemer, Trewin, Japac, Bergdahl and Pettersson (2012).

**5.4 Quality profiles**

In continuing surveys there is an opportunity to develop quality profiles. Such documents contain all that is known about the quality of a continuing survey or other statistical product assembled over a number of years. Quality profiles exist for only a few major surveys, all, except one, conducted in the U.S., including the Current Population Survey (Brooks and Bailar 1978), the Survey of Income and Program Participation (Jabine, King and Petroni 1990; Kalton, Winglee and Jabine 1998), the Schools and Staffing Survey (Kalton, Winglee, Krawchuk and Levine 2000), and the American Housing Survey (Chakrabarty and Torres 1996). The exception is the British Household Panel Survey (Lynn 2003). The main problem with a quality profile is that it is not timely, since it compiles results from often time-consuming studies of quality. The goal of the quality profile is to identify areas where knowledge about errors is deficient so that improvements can be made. Kasprzyk and Kalton (2001) and Doyle and Clark (2001) review the use of quality profiles in the U.S.

**6. Where do we go from here?**

Quality management ideas have been influential in many survey organizations. Concepts such as leadership, quality culture, problem prevention, customer, competition, risk assessment, process thinking, improvement, business excellence, and business architecture are increasingly discussed by leaders of survey organizations, e.g., Trewin (2001), Pink (2010), Fellegi (1996), Brackstone (1999), de Vries (1999), Groves (2011), and Bohata (2011). It seems as if the survey community is moving in a direction where statistics production becomes more streamlined and cost-effective but the pace is slow. Some organizations have started using a quality management model for self-assessment and steering purposes. EFQM is the recommended model for national statistical institutes within the European Statistical System and a couple of institutes, the Czech Republic and Finland, have even applied for their respective national EFQM awards. Some marketing firms are certified according to the ISO 9001 quality management standard and others are certified according to the ISO 20252 standard for market, opinion, and social research. This development ought to result in quality improvements but we cannot be really sure

until we start collecting relevant data. One thing is sure, though. Some customers prefer service providers that are certified, have won awards or can show evidence that they are working according to some quality framework or model. Very few customers would think that this is a negative thing.

The margins of error that we associate with estimates are usually too short, since they do not include all sources of variation. Point estimates can be off due to biases. Ideally it would be good if we were able to produce estimates of the total survey error instead of what we produce today. Such a development is, however, not realistic. We are not in a position to produce such estimates, not even occasionally, for reasons that have to do with finances, timing and methodology. That leaves us with indicators of total survey error and its components. Such indicators are of limited value to the users. Users simply do not know what to do with information on nonresponse rates, response variance measured by reinterviews or edit failure rates. On the other hand, such indicators are very useful to the producers of surveys. For instance, reinterview studies can identify fabrication and survey questions with poor response consistency. A majority of users appreciate the service provider's credibility and part of the credibility is the ability to present accurate data. Another important part of credibility is the willingness of the providers to evaluate their own quality and to report the results of such evaluations. Even if these evaluations show problems, it is better for the provider to find the problems than if entities outside the provider's organization find them. Most users do not want to become involved in discussions about errors and trade-offs between errors and for good reasons. It is simply too technical and confusing. If we accept that a good process quality is a prerequisite for a good product quality, we should gradually improve the processes so that they approach ideal bias-free ones. In that way the variance of an estimate becomes a good approximation of the mean squared error.

Despite endless discussions and a myriad of survey quality initiatives, practices have not changed much (Lynn 2004; Pink, Borowik and Lee 2010; Groves 2011; Bohata 2011). Perhaps the lack of competence within survey organizations is one root cause of the slow pace. Many theories and methodologies including statistics, IT, management, communication, and behavioral sciences are needed in survey research. The behavioral sciences are needed to identify the root causes of nonsampling errors. If errors are just quantified no improvement can happen. Current training programs emphasize sampling, non-response, coverage and estimation in the presence of these. Other processes and error sources such as measurement and data processing are not dealt with to the same extent. This

leads to a situation where studies on measurement error and data processing error are rare compared to studies on, say, nonresponse. There is a considerable confusion regarding concepts and methods in both the producer and the user camps. Another cause of slow pace might be the consensus philosophy that rules in some organizations when it comes to decision-making regarding changes. This philosophy is one of compromise. Input from many stakeholders is gathered and a decision is usually based on the smallest common denominator, which is never a good standard. Furthermore, arriving at this compromise usually takes a long time and lots of resources. This approach is very far from Plan-Do-Check-Act.

Survey quality is not an absolute entity. Current quality reporting a la one-size-fits-all is not working since fitness for use is defined by each user. Quality dimensions such as timeliness, comparability and accessibility should be decided together with main users while best possible accuracy given various constraints is the responsibility of the service provider.

Have the survey quality discussion and the adoption of quality management strategies resulted in better data? We do not know. Survey quality has not been assessed in a before-after fashion. There is a tendency towards greater standardization and centralization, which should prove cost-efficient but when it comes to data quality some indicators point in the wrong direction. For instance, in many countries nonresponse rates are increasing and error properties of mixed-mode, translation of survey materials, and other design features are not fully known or are different across cultures. There is no design formula, which results in shaky trade-off decisions and problems deciding about intensities with which quality control should be applied. There is a persistent quest for best practices in survey organizations but implementation is difficult and scattered. There is definitely a great need for an upgrade in the competence level across the board. A structured international competence development program for service providers is necessary as is a systematic international collaboration on how to best design and implement surveys. We must serve our users better by providing data with small errors. We can do this by better combining our knowledge about statistics and cognitive phenomena with the principles of quality management. The great positive note is the overwhelming positive attitude toward quality improvement among statistical organizations around the world.

## References

- Aitken, A., Hörngren, J., Jones, N., Lewis, D. and Zilhao, M. (2004). *Handbook on improving quality by analysis of process variables*. Office for National Statistics, UK.

- Anderson, R., Kasper, J. and Frankel, F. (1979). *Total Survey Error: Applications to Improve Health Surveys*. San Francisco: Jossey-Bass.
- Apted, L., Carruthers, P., Lee, G., Oehm, D. and Yu, F. (2011). Industrialisation of statistical processes, methods and technologies. Paper presented at the International Statistical Institute Meeting, Dublin.
- Bailar, B., and Dalenius, T. (1969). Estimating the response variance components of the U.S. Bureau of the Census' Survey Model. *Sankhyā*, B, 341-360.
- Biemer, P. (2001). Comment on Platek and Särndal. *Journal of Official Statistics*, 17(1), 25-32.
- Biemer, P. (2010). Overview of design issues: Total survey error. In *Handbook of Survey Research*, (Eds., P. Mardsen and J. Wright), Second Edition. Emerald Group Publishing Limited.
- Biemer, P., and Lyberg, L. (2003). *Introduction to Survey Quality*. New York: John Wiley & Sons, Inc.
- Biemer, P., and Lyberg, L. (2012). Short course on Total Survey Error. The Joint Program in Survey Methodology (JPSM), April 16-17, Washington, DC.
- Biemer, P., Trewin, D., Japac, L., Bergdahl, H. and Pettersson, Å. (2012). A tool for managing product quality. Paper presented at the Q Conference, Athens.
- Blyth, B. (2012). ISO 20252; Turning frameworks into best practice. Paper presented at the Q Conference, Athens.
- Bohata, M. (2011). Fit-for-purpose statistics for evidence based policy making. Memo, Eurostat.
- Bowley, A.L. (1913). Working-class households in reading. *Journal of the Royal Statistical Society*, 76(7), 672-701.
- Box, G. (1990). Good quality costs less? How come? *Quality Engineering*, 3, 1, 85-90.
- Box, G., and Friends (2006). *Improving Almost Anything: Ideas and Essays*. New-York: John Wiley & Sons, Inc.
- Brackstone, G. (1999). Managing data quality in a statistical agency. *Survey Methodology*, 25, 2, 139-149.
- Brackstone, G. (2001). How important is accuracy? *Proceedings: Symposium 2001, Achieving Data Quality in a Statistical Agency: A Methodological Perspective*, Statistics Canada.
- Breyfogle, F. (2003). *Implementing Six Sigma*. Second Edition. New-York: John Wiley & Sons, Inc.
- Brooks, C., and Bailar, B. (1978). An error profile: Employment as measured by the Current Population Survey. Working paper 3, Office of Management and Budget, Washington, DC.
- Chakrabarty, R., and Torres, G. (1996). American Housing Survey: A Quality Profile. U.S. Department of Commerce, U.S. Bureau of the Census.
- Colledge, M., and March, M. (1993). Quality management: Development of a framework for a statistical agency. *Journal of Business and Economic Statistics*, 11, 157-165.
- Colledge, M., and March, M. (1997). Quality policies, standards, guidelines, and recommended practices. In *Survey Measurement and Process Quality*, (Eds., L. Lyberg, P. Biemer., M. Collins, E. De Leeuw, C. Dippo, N. Schwarz and D. Trewin), New-York: John Wiley & Sons, Inc.
- Couper, M. (1998). Measuring Survey Quality in a CASIC Environment. Paper presented at the Joint Statistical Meetings, American Statistical Association, Dallas, TX.
- Dalenius, T. (1967). Nonsampling Errors in Census and Sample Surveys. Report No. 5 in the research project Errors in Surveys. Stockholm University.
- Dalenius, T.E. (1968). Official statistics and their uses. *Review of the International Statistical Institute*, 26(2), 121-140.
- Dalenius, T. (1969). Designing descriptive sample surveys. In *New Developments in Survey Sampling*, (Eds., N.L. Johnson and H. Smith), New-York: John Wiley & Sons, Inc.
- Dalenius, T. (1985a). *Elements of Survey Sampling*. Swedish Agency for Research Cooperation with Developing Countries. Stockholm, Sweden.
- Dalenius, T. (1985b). Relevant official statistics. *Journal of Official Statistics*, 1(1), 21-33.
- Deming, E. (1944). On errors in surveys. *American Sociological Review*, 9, 359-369.
- Deming, E. (1950). *Some Theory of Sampling*. New-York: John Wiley & Sons, Inc.
- Deming, E. (1986). *Out of the Crisis*. MIT.
- Deming, W.E., and Geoffrey, L. (1941). On sample inspection in the processing of census returns. *Journal of the American Statistical Association*, 36, 215, 351-360.
- De Vries, W. (1999). Are we measuring up...? Questions on the performance of national systems. *International Statistical Review*, 67, 1, 63-77.
- Dillman, D. (1996). Why innovation is difficult in government surveys (with discussions). *Journal of Official Statistics*, 12, 2, 113-198.
- Doherty, K. (2010). How business architecture renewal is changing IT at Statistics Canada. Paper presented at the Meeting on the Management of Statistical Information Systems. Daejeon, South Korea, April 26-29.
- Doyle, P., and Clark, C. (2001). Quality profiles and data users. Paper presented at the International Conference on Quality in Official Statistics (Q), Stockholm.
- Drucker, P. (1985). *Management*. Harper Colophon.
- Ecohard, P., Hahn, M. and Junker, C. (2008). User satisfaction surveys in Eurostat and in the European Statistical System. Paper presented at the Q conference, Rome, Italy.
- Edwards, W., Lindman, H. and Savage, L.J. (1963). Bayesian statistical inference for psychological research. *Psychological Review*, 70, 193-242.
- Eltinge, J. (2011). Aggregate and systemic components of risk in total survey error models. Paper presented at ITSEW 2011, Quebec, Canada.
- Ericson, W. (1969). Subjective Bayesian models in sampling finite populations. *Journal of the Royal Statistical Society, Series B*, 195-233.

- European Foundation for Quality Management (1999). *The EFQM Excellence Model*. Van Haren.
- Eurostat (2009a). ESS Standard for Quality Reports. Eurostat.
- Eurostat (2009b). ESS handbook for Quality Reports. Eurostat.
- Eurostat (2011a). European statistics Code of Practice. Eurostat.
- Eurostat (2011b). Quality assurance framework (QAF). Eurostat.
- Fellegi, I. (1964). Response variance and its estimation. *Journal of the American Statistical Association*, 59, 1016-1041.
- Fellegi, I. (1996). Characteristics of an effective statistical system. *International Statistical Review*, 64, 2, 165-197.
- Felme, S., Lyberg, L. and Olsson, L. (1976). *Kvalitetsskydd av data*. (Protecting Data Quality.) Liber (in Swedish).
- Fienberg, S., and Tanur, J. (1996). Reconsidering the fundamental contributions of Fisher and Neyman on experimentation and sampling. *International Statistical Review*, 64, 237-253.
- Fisher, R. (1935). *The Design of Experiments*. New York: Hafner.
- Frankel, M., and King, B. (1996). A conversation with Leslie Kish. *Statistical Science*, 11, 1, 65-87.
- Gleaton, E. (2011). Centralizing LAN services. Memo, National Agricultural Statistics Service, U.S. Department of Agriculture.
- Groves, R. (1989). *Survey Errors and Survey Costs*. New York: John Wiley & Sons, Inc.
- Groves, R. (2011). The structure and activities of the U.S. Federal Statistical System: History and recurrent challenges. *The Annals of the American Academy of Political and Social Science*, 631, 163, Sage.
- Groves, R., Biemer, P., Lyberg, L., Massey, J., Nicholls, W. and Waksberg, J. (Eds.) (1988). *Telephone Survey Methodology*. New-York: John Wiley & Sons, Inc.
- Groves, R.M., Fowler, F.J., Couper, M.P., Lepkowski, J.M., Singer, E. and Tourangeau, R. (2009). *Survey Methodology*, Second Edition. New-York: John Wiley & Sons, Inc.
- Groves, R., and Heeringa, S. (2006). Responsive design for household surveys: Tools for actively controlling survey errors and costs. *Journal of the Royal Statistical Society, A*, 169, 439-457.
- Groves, R., and Lyberg, L. (2010). Total survey error: Past, present and future. *Public Opinion Quarterly*, 74, 5, 849-879.
- Hansen, M., and Hurwitz, W. (1946). The problem of nonresponse in sample surveys. *Journal of the American Statistical Association*, 517-529.
- Hansen, M., Hurwitz, W. and Bershada, M. (1961). Measurement errors in censuses and surveys. *Bulletin of the International Statistical Institute*, 32<sup>nd</sup> Session, 38, Part 2, 359-374.
- Hansen, M., Hurwitz, W. and Madow, W. (1953). *Sample Survey Methods and Theory*. Volumes I and II. New-York: John Wiley & Sons, Inc.
- Hansen, M., Hurwitz, W., Marks, E. and Mauldin, P. (1951). Response errors in surveys. *Journal of the American Statistical Association*, 46, 147-190.
- Hansen, M., Hurwitz, W. and Pritzker, L. (1964). The estimation and interpretation of gross differences and simple response variance. In *Contributions to Statistics*, (Ed., C. Rao). Oxford: Pergamon Press, 111-136.
- Hansen, M., Hurwitz, W. and Pritzker, L. (1967). Standardization of procedures for the evaluation of data: Measurement errors and statistical standards in the Bureau of the Census. Paper presented at the 36<sup>th</sup> session of the International Statistical Institute.
- Hansen, M., and Steinberg, J. (1956). Control of errors in surveys. *Biometrics*, 462-474.
- Hansen, M., and Voigt, R. (1967). Program guidance through the evaluation of uses of official Statistics in the United States Bureau of the Census. Paper presented at the International Statistical institute meeting, Canberra, Australia.
- Holt, T., and Jones, T. (1998). Quality work and conflicting policy objectives. *Proceedings of the 84<sup>th</sup> DGINS Conference*, May 28-29, Stockholm, Sweden. Eurostat.
- International Standards Organization (2006). Market, Opinion and Social Research. ISO Standard No. 20252.
- Jabine, T., King, K. and Petroni, R. (1990). Survey of Income and Program Participation (SIPP): Quality Profile. U.S. Department of Commerce, U.S. Bureau of the Census.
- Joiner, B. (1994). *Generation Management*. McGraw-Hill.
- Julien, C., and Born, A. (2006). Quality management assessment at Statistics Canada. *Proceedings of the Q Conference*, Cardiff, UK.
- Julien, C., and Royce, D. (2007). Quality review of key indicators at Statistics Canada. *Proceedings of the Third International Conference on Establishment Surveys (ICES-III)*, 1113-1120.
- Juran, J.M. (1988). *Juran on Planning for Quality*. New York: Free Press.
- Juran, J.M. (1995). *A History of Managing for Quality*. ASQC Quality Press.
- Juran, J., and Gryna, F. (Eds.) (1988). *Juran's Quality Control Handbook*, 4<sup>th</sup> Edition. McGraw-Hill.
- Kalton, G. (2001). How important is accuracy? *Proceedings: Symposium 2001, Achieving Data Quality in a Statistical Agency: A Methodological Perspective*, Statistics Canada.
- Kalton, G., Winglee, M. and Jabine, T. (1998). *SIPP Quality Profile*. U.S. Bureau of the Census, 3<sup>rd</sup> Edition.
- Kalton, G., Winglee, M., Krawchuk, S. and Levine, D. (2000). *Quality Profile for SASS Rounds 1-3: 1987-1995*. Washington, DC: U.S. Department of Education.
- Kasprzyk, D., and Kalton, G. (2001). Quality profiles in U.S. Statistical Agencies. *Proceedings of the International Conference on Quality in Official Statistics*, Stockholm 14-15 May 2001, CD-ROM.

- Kennickell, A., Mu lrow, E. and Scheuren, F. (2009). Paradata or process modeling for inference. Paper presented at the Conference on Modernization of Statistics Production, Stockholm, Sweden.
- Kiear, A.N. (1897). The representative method of statistical surveys. *Kristiania Videnskaps-selskabets Skrifter: Historik-filosofiske Klasse*, (in Norwegian), 4, 37-56.
- Kish, L. (1965). *Survey Sampling*. New York: John Wiley & Sons, Inc.
- Kish, L. (1995). *The Hundred Years' Wars of Survey Sampling*. Centennial Representative Sampling, Rome.
- Kotz, S. (2005). Reflections on early history of official statistics and a modest proposal for global coordination. *Journal of Official Statistics*, 21, 2, 139-144.
- Kreuter, F., Couper, M. and Lyberg, L. (2010). The use of paradata to monitor and manage survey data collection. *Proceedings of the Survey Research Methods Section*, American Statistical Association.
- Lyberg, L. (1981). *Control of the Coding Operation in Statistical Investigations: Some Contributions*. Ph.D. dissertation, Stockholm University.
- Lyberg, L. (2002). Training of survey statisticians in government agencies-A review. Invited paper presented at the Joint Statistical Meetings, American Statistical Association, New-York.
- Lyberg, L., Bergdahl, M., Blanc, M., Booleman, M., Grünewald, W., Haworth, M., Japac, L., Jones, L., Körner, T., Linden, H., Lundholm, G., Madaleno, M., Radermacher, W., Signore, M., Zilhao, M.J., Tzougas, I. and van Brakel, R. (2001). Summary report from the Leadership Group (LEG) on Quality. Eurostat.
- Lyberg, L., and Couper, M. (2005). The use of paradata in survey research. Invited paper, International Statistical Institute, Sydney, Australia.
- Lynn, P. (Ed.) (2003). *Quality Profile: British Household Panel Survey: Waves 1 to 10: 1991-2000*. Colchester: Institute for Social and Economic Research.
- Lynn, P. (2004). Editorial: Measuring and communicating survey quality. *Journal of the Royal Statistical Society*, Series A, 167.
- Mahalanobis, P.C. (1946). Recent experiments in statistical sampling in the Indian Statistical Institute. *Journal of the Royal Statistical Society*, 109, 325-378.
- Mirotschie, M. (1993). Data quality: A quest for standard indicators. *Proceedings of the International Conference on Establishment Surveys*, American Statistical Association, 729-734.
- Moeller, R. (2005). *Brink's Modern Internal Auditing*. Sixth Edition. New-York: John Wiley & Sons, Inc.
- Morganstein, D., and Marker, D. (1997). Continuous quality improvement in statistical agencies. In *Survey Measurement and Process Quality*, (Eds., L. Lyberg, P. Biemer, M. Collins, E. De Leeuw, C. Dippo, N. Schwarz and D. Trewin), New York: John Wiley & Sons, Inc., 475-500.
- Mudryk, W., Burgess, M.J. and Xiao, P. (1996). Quality control of CATI operations in Statistics Canada, Memo, Statistics Canada.
- Muscio, B. (1917). The influence of the form of a question. *The British Journal of Psychology*, 8, 351-389.
- Neter, J., and Waksberg, J. (1964). A study of response errors in expenditures data from household interviews. *Journal of the American Statistical Association*, 59, 305, 18-55.
- Neyman, J. (1934). On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society*, 97, 558-606.
- Neyman, J. (1938). *Lectures and Conferences on Mathematical Statistics and Probability*. U.S. Department of Agriculture, Washington, DC.
- OECD (2011). Quality dimensions, core values for OECD statistics and procedures for planning and evaluating statistical activities. OECD.
- O'Muircheartaigh, C. (1997). Measurement errors in surveys: A historical perspective. In *Survey Measurement and Process Quality*, (Eds., L. Lyberg, P. Biemer, M. Collins, E. De Leeuw, C. Dippo, N. Schwarz and D. Trewin), New York: John Wiley & Sons, Inc., 1-25.
- Phipps, P., and Fricker, S. (2011). Quality measures. Memo, Office of Survey Methods Research, U.S. Bureau of Labor Statistics.
- Pink, B., Borowik, J. and Lee, G. (2010). The case for an international statistical innovation program-Transforming national and international statistics systems. Paper presented at the Collaboration Leaders Workshop, April 19-23, Sydney, Australia.
- Platek, R., and Särndal, C.-E. (2001). Can a statistician deliver? *Journal of Official Statistics*, 17, 1, 1-20 and Discussion, 21-27.
- Reedman, L., and Julien, C. (2010). Current and future applications of the generic statistical business process model at Statistics Canada. Paper presented at the Q Conference, Helsinki.
- Rosén, B., and Elvers, E. (1999). Quality concept for official statistics. *Encyclopedia of Statistical Sciences*, New-York: John Wiley & Sons, Inc., update Volume 3, 621-629.
- Scheuren, F. (2001). How important is accuracy? *Proceedings: Symposium 2001, Achieving Data Quality in a Statistical Agency: A Methodological Perspective*, Statistics Canada.
- Schilling, E., and Neubauer, D. (2009). *Acceptance Sampling in Quality Control*, 2<sup>nd</sup> Ed. Chapman and Hall/CRC.
- Scholtes, P., Joiner, B. and Streibel, B. (1996). *The Team Handbook*. Joiner Associates Inc.
- Shewhart, W.A. (1939). *Statistical Methods from the Viewpoint of Quality Control*. U.S. Department of Agriculture, Washington, DC, U.S.A.
- Smith, T. (2011). Report on the International Workshop on using multi-level data from sample frames, auxiliary databases, paradata and related sources to detect and adjust for nonresponse bias in surveys. NORC/University of Chicago.
- Spencer, B. (1985). Optimal data quality. *Journal of the American Statistical Association*, 80, 564-573.
- Statistics Canada (2002). Statistics Canada's Quality Assurance Framework, Catalogue No.12-586-XIE, Ottawa.

- Statistics Canada (2009). Statistics Canada Quality Guidelines, fifth Edition, Ottawa.
- Statistics Netherlands (1997). A self assessment of the Department of Statistical Methods. Research paper No. 9747, Statistics Netherlands.
- Stephan, F.F. (1948). History of the uses of modern sampling procedures. *Journal of the American Statistical Association*, 43, 12-39.
- Trewin, D. (2001). The importance of a quality culture. *Proceedings: Symposium 2001, Achieving Data Quality in a Statistical Agency: A Methodological Perspective*, Statistics Canada.
- United Nations (2010). *Post Enumeration Surveys: Operational Guidelines*. Department of Economic and Social Affairs, Statistics Division.
- U.S. Bureau of the Census (1974). *Standards for Discussion and Presentation of Errors in Data*. U.S. Department of Commerce, Bureau of the Census.
- U.S. Federal Committee on Statistical Methodology (2001). *Measuring and Reporting Sources of Errors in Surveys*, Statistical Policy Working Paper 31, Washington, DC: U.S. Office of Management and Budget.
- U.S. Office of Management and Budget (2002). Guidelines for ensuring, and maximizing the quality, objectivity, utility, and integrity of information disseminated by Federal agencies. Federal register, 67, 36, February 22.
- U.S. Office of Management and Budget (2006a). *Standards and Guidelines for Statistical Surveys*. U.S. Office for Management and Budget.
- U.S. Office of Management and Budget (2006b). Questions and answers when designing surveys for information collection. U.S. Office for management and Budget.
- Waksberg, J. (1998). The Hansen era: Statistical research and its implementation at the Census Bureau, 1940-1970. *Journal of Official Statistics*, 14, 2, 119-137.
- Weisberg, H. (2005). *The Total Survey Error Approach*. The University of Chicago Press.
- Weisman, E., Balyozov, Z. and Venter, L. (2010). IMF's data quality assessment framework. Paper presented at the Conference on Data Quality for International Organizations, Helsinki, May 6-7.
- West, B., and Olson, K. (2010). How much of interviewer variance is really nonresponse error variance? *Public Opinion Quarterly*, 74, 5, 1004-1026.
- Willimack, D., Nichols, E. and Sudman, S. (2002). Understanding unit and item nonresponse in business surveys. In *Survey Nonresponse*, (Eds., R. Groves, D. Dillman, J. Eltinge and R. Little), 213-228.
- Zarkovich, S. (1966). *Quality of Statistical Data*. Food and Agricultural Organization of the United Nations: Rome, Italy.



## Data collection: Experiences and lessons learned by asking sensitive questions in a remote coca growing region in Peru

Jaqueline Garcia-Yi and Ulrike Grote<sup>1</sup>

### Abstract

Coca is a native bush from the Amazon rainforest from which cocaine, an illegal alkaloid, is extracted. Asking farmers about the extent of their coca cultivation areas is considered a sensitive question in remote coca growing regions in Peru. As a consequence, farmers tend not to participate in surveys, do not respond to the sensitive question(s), or underreport their individual coca cultivation areas. There is a political and policy concern in accurately and reliably measuring coca growing areas, therefore survey methodologists need to determine how to encourage response and truthful reporting of sensitive questions related to coca growing. Specific survey strategies applied in our case study included establishment of trust with farmers, confidentiality assurance, matching interviewer-respondent characteristics, changing the format of the sensitive question(s), and non enforcement of absolute isolation of respondents during the survey. The survey results were validated using satellite data. They suggest that farmers tend to underreport their coca areas to 35 to 40% of their true extent.

Key Words: Coca; Cocaine; Sensitive question; Misreporting; Nonresponse; Peru.

### 1. Introduction

Over the last 30 years, surveys have been increasingly used to explore sensitive topics (Tourangeau and Yan 2007). For example, data obtained from surveys have been used to investigate “socially undesirable” behaviors, such as the prevalence of illicit drug use (*e.g.*, Botvin, Griffin, Diaz, Scheier, Williams and Epstein 2000; Fergusson, Boden and Horwood 2008), illegal abortion (*e.g.*, Johnson-Hanks 2002; Varkey, Balakrishna, Prasad, Abraham and Joseph 2000), or alcohol consumption among adolescents (*e.g.*, Strunin 2001; Zufferey, Michaud, Jeannin, Berchtold, Chossis, van Melle and Suris 2007). Such surveys have been commonly utilized in academic research and policy analysis (Davis, Thake, and Vilhena 2009), even though asking sensitive questions has generally been seen as problematic. The responses have been considered to be prone to error and bias because respondents consistently underreport socially undesirable behaviors (Barnett 1998; Tourangeau and Yan 2007). Low response rates have been an additional concern. Those who are selected for a survey can simply refuse to take part in the survey or they can participate but refuse to answer the sensitive questions (Tourangeau and Yan 2007).

Recent surveys at the household level have incorporated sensitive questions related to the extent of coca growing areas (see *e.g.*, Ibanez and Carlsson 2010). Coca is a native bush from the Amazon rainforest in South America from the leaves of which cocaine is extracted. Colombia’s coca bush area represents 40%, Peru’s 40%, and Bolivia’s 20% of the total area under coca cultivation worldwide, amounting to

154,100 hectares (UNODC 2011). In Peru and Bolivia, the leaves of this bush have been traditionally used for many purposes from around 3000 B.C. (Rivera, Aufferdeide, Cartmell, Torres and Langsjoen 2005) until today. Those traditional uses mainly include coca chewing and coca tea drinking to overcome fatigue, hunger and thirst; and to relieve “altitude sickness” and stomach ache symptoms, respectively (Rospigliosi 2004). Since the 1970s, however, coca cultivation skyrocketed because of its use as the raw material for the production of cocaine (Caulkins, Reuter, Iguchi and Chiesa 2005). The cocaine content of the coca leaves is below 1%, and ranges from 0.13 to 0.86% (Holmstedt, Jaatmaa, Leander and Plowman 1977). Therefore narcotics traffickers need large quantities of coca leaves to obtain enough of the alkaloid for commercialization in the illegal market. In general, growing coca for the narcotics trafficking business is a profitable activity. In fact, the income of a coca growing farmer has been calculated to be 54% higher than the income of a non coca growing farmer (Davalos, Bejarano and Correa 2008).

Consequently, coca-related research has become oriented towards evaluating the profitability of coca versus other cash crops (see, *e.g.*, Gibson and Godoy 1993; Torrico, Pohlan and Janssens 2005). Different attempts were made to replace coca by other crops, but it has been generally established that crop substitution as an anti-drug policy has been a failure (UNODC 2001). Decision makers and researchers have recognized that there are relevant socio-economic determinants that lead to coca growing other than economic profitability. These include social capital (Thoumi 2003),

1. Jaqueline Garcia-Yi, chair of Agricultural and Food Economics Technical University of Munich Weihenstephaner Steig 22, 85350, Freising, Germany. E-mail: jaqueline.garcia-yi@tum.de; Ulrike Grote, Professor, Institute for Environmental Economics and World Trade, Leibniz University Hannover, Königsworther Platz 1, 30167 Hannover, Germany. E-mail: grote@iuw.uni-hannover.de.

saving account functions and financial reserve for large expenses (Bedoya 2003; Mansfield 2006). Comprehensive databases which include specific household-level information for coca growing areas are required to test those latter hypotheses.

Coca growing is not illegal *per se* in Peru (During the 1990s, the primary focus of the Peruvian Government was on “pacifying” the country by bringing terrorist groups under control. The Peruvian Government implemented what is currently known as the “Fujimori Doctrine”. The idea underlying this Doctrine was that the coca cultivation was not criminal in nature, but attributable to poverty. Consequently, the Fujimori Doctrine decriminalized all coca farmers, which diminished the farmers’ need for protection from terrorist associations, therefore making it easier for the Government to fight those violent groups (Obando 2006).), which partly reflects the social acceptance of traditional uses of coca in this country (UNODC 2001). Thus, the current legal framework seems to facilitate narcotics trafficking because coca used in illegal trade can be cultivated under the guise of traditional uses (INCB 2009; Durand 2005). Accordingly, Garcia and Antezana (2009) suggest that some farmers sell coca to those who purport to be traditional-use traders, but are actually narcotics traffickers who process coca leaves in different places, such as small towns at the border with Bolivia.

Even though coca farming is not illegal, coca-growing regions which are perceived to be supplying narcotics traffickers (*e.g.*, regions with large coca fields) can be targeted by the Government for the implementation of forced eradication programs (Obando 2006). After eradication, coca growers are likely to incur large economic losses, depending on the total extent of their individual coca cultivation areas. Thus, some of the farmers might be reluctant to provide information on whether or not they have any coca under cultivation. It should also be expected that some of the farmers who admit to cultivating coca, would not report the true extent of the area, given their fear that large coca fields could be more prone to eradication.

Since there are both political and policy concerns in accurately and reliably measuring coca growing areas, it is necessary for survey methodologists to determine how to encourage response and truthful reporting of answers to sensitive questions related to coca growing. This article suggests and evaluates a number of strategies to increase both the reporting and the reliability of household-level responses in a remote coca growing region in Peru.

Although the topic of this article is specifically related to coca growing, the lessons learned about survey design and implementation could be used as a reference for dealing with other sensitive topics such as health-related issues (*e.g.*, anti-conception and sexual behavior) or undesirable

behaviors (*e.g.*, illegal drug use) in other regions in different countries.

The structure of the article is as follows: Section 2 describes the community in Peru subject to study, the specific strategies to reduce non-response and misreporting as well as the lessons learned from data collection related to sensitive questions in the research area. Section 3 presents the coca growing-related survey results and their validation, while Section 4 is comprised of a summary of the main results followed by the conclusion.

## 2. Data collection in a coca-growing community in rural Peru

This section describes the coca-growing community, and the primary data collection strategies applied in our study and the lessons learned.

### 2.1 Description of the research area

The research area was located in the Upper Tambopata valley at the border with Bolivia, one of the most remote and difficult to access Amazon rainforest areas in Peru (UNODC Office in Peru 1999). This valley lies in the Vilcabamba-Amboro Biodiversity Corridor in close proximity to national protected areas (see Figure 1). The entire population of the upper Tambopata valley is composed of immigrants, especially descendants from the Aymara indigenous population. Aymara is a native ethnic group originally from the Andes and Altiplano regions of South America. During the 1950s, most of the farmers were seasonal immigrants who left their Altiplano subsistence plots for only three to six months every year, and made the 320 km journey to the upper Tambopata valley to cultivate coffee on their individually owned agricultural plots (Collins 1984). Over time, most farmers became permanent settlers in the upper Tambopata valley, and cultivate coffee as their main cash crop (*ibid*).

Before 1989, coca cultivation in the upper Tambopata valley was very minor. Small-scale coca production was limited to self-consumption or local markets for traditional uses such as coca chewing by Andean farmers and miners. After 1989, coca cultivation was intensified, primarily in the neighboring upper Inambari valley. The change did not appear to be in response to increases in local demand or external demand by traditional users (UNODC Office in Peru 1999). Coca from those valleys is considered as low quality due to its bitterness, and it is in less demand for traditional chewing than coca from Cuzco region (Caballero, Dietz, Taboada and Anduaga 1998). Those increases were therefore related to narcotic traffic demand. In recent years, large increases in coca cultivation in the upper Tambopata valley have been consistently reported by the

United Nations (UN), as observed in Table 1. The percentage variation per year in the upper Tambopata valley is above the annual change of around 4% at national level.

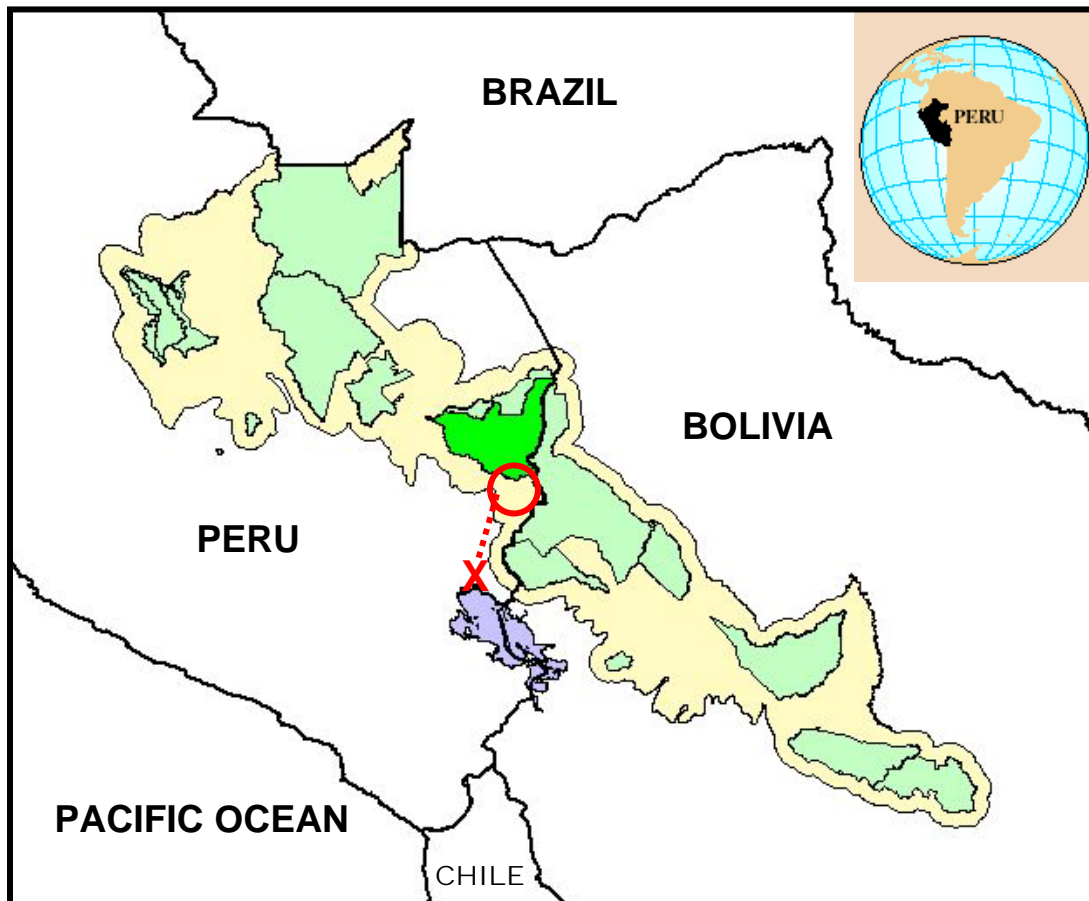
**Table 1**  
Coca cultivation in the upper Tambopata Valley (2005-2008)\*

Year	Hectares	Percentage of variation in relation to previous year
2005	253	-
2006	377	49.0
2007	863	128.9
2008	940	8.9

\* Since 2009 coca areas from the upper Tambopata valley are aggregated with coca areas from Inambari valley in UNODC reports. Therefore, it is not possible to estimate the percentage of variation in relation to previous year only for Tambopata valley during later years.

Source: Own calculation using data from UNODC (2009).

Coca provided by the upper Tambopata valley and upper Inambari valley seems to mainly supply cross border trade associations between Peruvian and Bolivian narcotics traffickers. Bolivia remains the world's third largest producer of cocaine, and it is a significant transit zone for cocaine of Peruvian-origin (U.S. Department of State 2009). Those valleys constitute a strategic coca production area for narcotics traffickers due to their proximity to an external exit route (UNODC Office in Peru 1999). Coca leaves are not always transformed into cocaine in the agricultural plots. Narcotics traffickers seem to take advantage of the large quantities of coca leaves transported to urban areas, ostensibly for traditional user markets. This coca is then purchased and processed at hidden facilities in urban areas near the Bolivian border. In this way the risk of being caught by authorities is reduced. From Bolivia the cocaine is dispatched to Brazil and Europe (Garcia and Antezana 2009).



Source: Own elaboration

**Map Description:**

- X Altiplano area
- Upper Tambopata Valley
- - Immigration Route

- Bahujaja Sonene National Park
- Other protected areas
- Vilcabamba-Amboro Biodiversity Corridor
- Titicaca Lake

**Figure 1** Map of the research area

Coca cultivation does not necessarily translate into better quality of life for the farmers in South America (Davalos, *et al.* 2008). According to the last population census, the living conditions in San Pedro de Putina Punco (SPPP), the district located in the heart of the Upper Tambopata valley, are difficult: 72% of the houses are rammed earth constructions, 88% have dirt floors, 16% have public electricity, 12% have public water, and only 9% have access to public sewage (INEI 2007). This situation is common in the major coca growing areas in Peru, where 70% of the inhabitants continue to live in poverty, and 42% in extreme poverty (Commission on Narcotic Drugs 2005).

## 2.2 Data collection strategies and lessons learned

A feasibility study to test if farmers would answer coca-related questions was conducted in December 2007. The pilot study for the designed questionnaire took place in May 2008, and the final survey was conducted between June and August 2008. The feasibility and pilot studies and the final survey were focused on the farmers located in San Pedro de Putina Punco (SPPP), a district in the upper Tambopata valley which is located in the deepest rainforest. All the farmers in the research area produce coffee as cash crop and some supplement their income with coca cultivation. There are five coffee co-operatives in SPPP. Farmers have to become a member of one of these co-operatives in order to be able to sell their coffee, because restrictions to coffee intermediaries are in place. The final survey was only conducted among the members of four of these co-operatives because most of the members of the remaining co-operative are based in San Juan del Oro, a district outside the research area.

The final survey consisted of a structured questionnaire which focused on agricultural production and social capital. The questionnaire was comprised of 15 sections:

1. General information about the farmer and household
2. General information about the agricultural plot and coffee area
3. Additional economic activities
4. Organic certification information
5. Cognitive social capital and identity
6. Information and communication
7. Personal aspirations and risk attitudes
8. Structural social capital
9. Covariant and idiosyncratic shocks
10. Human capital
11. Social networks
12. Coca use traditions
13. Detailed agricultural production costs
14. Labor access
15. Additional questions

The sensitive question related items of the survey are presented in the Appendix 1.

Asking farmers about their coca growing area is a sensitive question. Farmers who cultivate large areas of coca fear that the information provided could be accessed by authorities responsible for eradication programs. Thus, they might have concerns about the possible consequences of giving a truthful answer should the information become known to a third party. In these cases, the farmers need to be assured anonymity. Farmers could also be tempted to provide socially desirable answers to the interviewers. Coca has become an important focal symbol in the indigenous population's struggle for self-determination (Office of Technology Assessment 1993). Coca "yes", cocaine "no" constitutes the slogan of indigenous people (Henman 1990); the formulation tries to clearly separate traditional uses ("coca") from narcotics trafficking ("cocaine"). Hence, traditional uses such as coca chewing are ethnicity symbols (Allen 1981) and their persistence could be related to feelings of nationalism in Peru (Henman 1990). In this sense, it could be expected that farmers would not find it very problematic to indicate that they grow coca, as long as they can associate it with traditional uses. On the other hand, due to the association of larger production areas with illegal activities, coca growers may underreport the total extent of their coca production areas in an attempt to give the impression that they are growing only for traditional use.

Several strategies can help to reduce the potential biases associated with question sensitivity, item and unit nonresponse and deliberate misreporting. These strategies include: confidentiality assurances; careful selection of the data collection mode and setting of the sensitive question format; and tailoring interviewer characteristics and behavior (see Coutts and Jann 2008; Tourangeau and Yan 2007). Further information on the implementation of these strategies in our case study is provided below.

### *Establishing trust, and anonymity assurances*

Farmers in coca growing areas tend to distrust external people. In this particular area, we found out that they trust the coffee co-operative directors. One of the directors of the coffee co-operatives signed a letter of presentation authorizing our research related to agricultural cultivation. The letter was shown to the farmers prior to conducting the survey. A pilot test conducted with and without the presentation letter demonstrated that the letter was important to reduce survey participation refusals. In the survey introduction, it was also indicated by the interviewer that the co-operative director authorized the survey because the director expected the results to benefit co-operative members. In addition, farmers were clearly told at the beginning of the survey that the data collected would remain confidential, and the academic purpose of the questionnaire was

high-lighted (see Appendix 1a). This anonymity assurance was short and precise in order to minimize suspicion among farmers as suggested by Singer, Hippler and Schwarz (1992). Coca growing was treated as a common and ordinary behavior in the research region, and a long and elaborate confidentiality assurance might have aroused farmers' reservations instead of alleviating them. A brief reminder of the assurance of confidentiality was included in the middle of the questionnaire, before the questions related to traditional coca uses and prior to the sensitive question on the coca area. The reminder stated: "In this part of the survey, we will ask questions about coca uses and cultivation. Please remember that the survey is anonymous and that there are no correct or incorrect answers" (See Appendix 1b). This follows Willis (2005) who mentions that it is important to have warm-up questions and an announcement of the switching to the sensitive topic to reduce resistance to answer.

#### *Data collection mode*

Paper and pencil self-administration as data collection method was initially considered to try to reduce interviewer bias. However, during the feasibility study, it became evident that many farmers, even those with above elementary school education (52% of the population; INEI 2007), were not able to read effortlessly. Farmers work in their fields almost all day long and do not have many opportunities to practice their reading skills. Similarly, audio computer-assisted self-interviewing (ACASI) the method of choice for collecting data on sensitive topics in developed countries (Mensch, Hewett and Erulkar 2003), was out of the scope of this project due to the lack of equipment and power supply, and the computer illiteracy in the research area. The use of computers was likely to have increased the anxiety and suspicion about the survey as described in the African situation by Mensch, *et al.* (2003). Therefore, a face-to-face interview was the data collection mode selected and emphasis was placed on the selection of interviewers, their training and behavior.

#### *Selection of interviewers, training, and interviewers' behavior*

One problem with the selection of the interviewers was the lack of sufficiently educated professionals in the research area. Thus, a group of ten students from the nearest public university, located 16 hours away from the research area, was chosen as interviewers. All of the interviewers had Aymara or Quechua ethnic backgrounds; this was an attempt to partially match interviewer-respondent characteristics. It was thought that this could increase the likelihood of participation because the matching was likely to increase trust and sympathy between the interviewer and the respondent (Tourangeau and Yan 2007). The interviewers

presented themselves as students from the local university, and no additional information was given about any university or organization outside of the country financing the study to avoid potential misunderstandings and reduce distrust among the respondents. During the pilot study, some farmers had indicated concerns about externally financed coca eradication programs and therefore references to external institutions were minimized. As a result, only partial information was given to the respondents. This is unconventional, but under the specific circumstances of the study, there was no other alternative without facing potential security problems.

For training, the interviewers first attended a two-day workshop in Puno city, followed by a three-day workshop in the research area. The same group of interviewers also conducted the pilot study to test the questions and questionnaire with the objective of identifying comprehension, recall, judgement and acceptability issues in the survey, and allowing rephrasing, eliminating or adding questions. The pilot study also allowed assessment of the performance of the interviewers, and in some cases identified areas requiring tailored training based on the feedback on performance. For example, at the beginning one of the interviewers was hesitant about asking the coca-related question and that interviewer obtained a higher than average number of nonresponses to the sensitive question. After tailored training, the interviewer was able to modify their interviewing approach.

#### *Format of the sensitive question*

The question format presupposed the sensitive behavior under study, as suggested by Tourangeau and Yan (2007). Therefore, farmers were not first asked if they had any coca areas, and then asked for the total extent of their coca areas. Instead, all farmers were directly requested to state the total extent of their coca areas ("What is your coca growing area in meters or hectares?"). However, it was found during the pilot study that the farmers did not feel comfortable with this question format and they either skipped the question or simply withdrew from the survey. As a consequence, the question format was changed and a forgiving wording was used instead. Farmers were asked: "How many 'little bushes of coca' do you have in your agricultural plot?" Thus, the farmer could answer "Only a little, I have... coca bushes". Even though a difference was hardly perceptible, with the former question it was more difficult for the farmers to start their answers with "Only a little...". So, using the latter question, it was easier for the farmers to add apologetic explanations to their answers making them feel more relaxed. This latter sensitive question format also had the advantage of employing a familiar wording for the Aymara who commonly use diminutives in their daily conversations. On the other hand, this question format might indirectly

imply that the interviewer expected that the respondent had a small number of coca bushes likely resulting in underreporting. Consequently, while nonresponses were avoided using this latter question format, underreporting was still expected to some extent.

#### *Time period for conducting the survey and data collection setting*

The farmers' agricultural plots are scattered in the mountainous Amazon rainforest in Peru. It was difficult to reach individual farmers on their agricultural plots for the survey. Therefore, to conduct the survey, we mainly took advantage of the Saint Peter's Day celebration and the General Assembly meetings of the co-operatives in June and August 2008 respectively, when the farmers congregated in the town square. Attendance to the General Assembly meetings is mandatory for all co-operative members so all of the targeted respondents would have been accessible at those events. The only way to reach or exit the town square is through an unpaved road. To take advantage of this, the survey was conducted in a large tent that was erected on the unpaved road on those key days. This tent had ten divisions, one for each pair of interviewer and respondent. Absolute privacy was not enforced because during the pilot study, it was found that farmers did not feel comfortable being the "only one" who was being interviewed; they preferred to see others doing the same. However, farmers were not able to overhear other farmers' responses. Given that all farmers have to use the same unpaved road to reach the town square regardless of their specific geographic location, potential geographical biases, which in turn can be related to important variables such as farm size and income, were likely minimized in this research.

#### *Sampling representativeness*

A convenience sampling method was applied, but at the end of the survey, we asked the farmers for their co-operative registration number and used the co-operative registration lists to infer the sample's representativeness. The co-operative registration number provided by the farmer was written on separate piece of paper and was not attached to the respondent's questionnaire. Respondents were informed about this procedure and were able to witness the procedure.

The four co-operatives under study have 3,265 members in SPPP. Table 2 shows the number of respondents per co-operative. The number of collected questionnaires amounted to 508. In total, 12 respondents were excluded from the sample because their co-operative registration number was missing. In two cases, the farmers had refused to provide this information and in ten cases, the interviewers had forgotten to ask the respondents about their registration

number at the end of the interview. Therefore the absence of information was more associated with interviewer error than with the farmers' unwillingness to provide this information.

**Table 2**  
**Number of respondents per co-operative**

	<b>Total Number of Co-operative Members in SPPP</b>	<b>Survey's Sample Size</b>	<b>Percentage of Co-operative Members Interviewed (%)</b>
Co-operative 1	756	106	14
Co-operative 2	911	138	15
Co-operative 3	887	138	16
Co-operative 4	711	114	16
Total	3,265	496	15

Source: Own survey.

In order to test for representativeness of the sample, the distribution of the co-operative registration numbers obtained from the survey sample was compared with the distribution of the co-operative registration numbers from a simulated simple random sample without replacement obtained from co-operative lists. The co-operative lists were ordered by the registration number of the co-operative members and co-operative registration numbers are associated with the members' date of registration. Thus, most of the older farmers have lower registration numbers and the younger farmers have higher ones. Unfortunately, the co-operatives did not have other membership data available such as total land, coffee or coca hectares that might be used to select a stratified random sample. Two types of tests were used for comparison of the samples: a two-sample Wilcoxon rank-sum (Mann-Whitney) test and a two-sample Kolmogorov-Smirnov test for equality of distribution functions. The first test assesses how probable it is that the two groups come from the same distribution, and assumes that differences observed are caused by chance fluctuation. The second test is similar to the first one, but in addition it is sensitive to differences in both the location and shape of the empirical cumulative distribution functions of the two groups. The results of both tests failed to reject the null hypothesis of equality of distribution between the survey sample and the simulated simple random sample at a significance level of 0.05. Thus, the results suggest that the survey sample is equivalent to a simple random sample, and therefore representative of the population under study.

### **3. Survey results and validation issues**

#### **3.1 Survey results**

The survey response rate was around 90%, which is well above the minimum recommended response rate of 60% (Punch 2003). From the 496 completed questionnaires, 19

respondents (less than 4%) did not answer the coca-related question. When comparing the descriptive statistics of socio-economic, institutional, and coca-related variables, there were some significant differences between all the observations (without the non-respondents) and the 'sensitive question non-respondents' (see Appendix 2). The sensitive question non-respondents were all male, with a larger percentage of Aymara ethnic background, and more children. In addition, a larger percentage of them used coca as medicine. Interestingly, significantly more non-respondents are highly risk averse (73.7%) compared to all the other respondents (28.6%). This could indicate a potential fear of the 'sensitive question non-respondents' of interviewer disclosure of information to third parties. The setup of the risk aversion test followed by Binswanger (1980) is presented in Appendix 1c.

Basic comparative descriptive statistics of coca and non coca growers are presented in Table 3. The number of valid questionnaires was 477, if we do not account for the non respondents of the sensitive question. Of them, 64% indicated that they are coca growers.

There are no statistically significant differences with respect to general socio-economic characteristics (age, sex,

ethnic group, and number of children) between coca and non-coca growers. The only difference was observed in education. Non-coca growers have more years of schooling than coca growers. Coca growers have less total and primary forest areas, and more fallow land than non coca growers, although these differences are not statistically significant. Coca and non-coca growers have similar coffee and staple food areas. On the contrary, coca growers and non-coca growers show statistically significant differences in the social capital variables. More non-coca growers than coca growers find it important to obey national law. On the other hand, less non-coca growers than coca growers have experienced a negative change in trust towards their neighbors during the last five years, and have worked in community activities during the last year.

There is a statistically significant relationship between coca growing and traditional uses. A higher percentage of coca growers than non-coca growers chew coca and uses coca as medicine. More importantly, more coca growers find it easier to sell coca leaves than non-coca growers in the hypothetical case that they would cultivate coca for commercial purposes.

**Table 3**  
Comparative descriptive statistics between coca and non coca growers

Variable	Coca Growers	Non Coca Growers
Age	42.5 (12.7)	41.7 (12.5)
Male (%)	93.9	94.9
Aymara (%)	81.4	82.5
Number of Children	3.0 (2.0)	2.9 (2.1)
Years of schooling	8.2* (3.3)	8.7* (3.3)
Total area (ha)	7.9 (8.4)	8.0 (7.8)
Coffee area (ha)	2.2 (2.0)	2.2 (1.4)
Area secondary forest (fallow area)	1.6 (2.4)	1.4 (2.1)
Primary forest area (ha)	3.9 (7.5)	4.2 (7.0)
Staple food area (ha)	0.5 (0.7)	0.5 (0.6)
No other economic activities (%)	46.8	48.9
High risk aversion (%)	30.5	25.3
Important to obey national laws (%)	81.9**	88.6**
Negative change in trust in the last 5 years (%)	19.3**	12.5**
Have worked in community activities in 2007 (%)	92.0**	84.7**
Farmer chews coca (%)	76.0***	53.1***
Farmer uses coca as medicine (%)	81.7***	54.8***
Perception that it is easy to sell coca leaves (%)	26.4**	18.5**
Number of coca bushes	3,093 (6,710)	-
Number of Observations	305	172

Standard deviations are in parentheses for continuous variables.

Coca Growers and Non Coca Growers means are statistically different (T-test with unequal variances) at:

\* 0.1 significance level, \*\* 0.05 significance level, \*\*\* 0.01 significance level.

Source: Own calculations.

Finally, it is important to mention that the average number of coca bushes is relatively low, which could be due to underreporting of commercial coca growing areas or to coca cultivation only for self-consumption, or both. It is not possible to distinguish between those two scenarios, which makes it easier for commercial coca growers to disguise themselves as coca growers who produce for traditional uses.

### 3.2 Validation issues

The validity of individual responses cannot be verified directly because there is little prior empirical research on this topic, and there is an absence of other sources of confirming data. However, it is possible to provide a rough comparison between the survey data and the total area of coca production recounted by international organizations for the upper Tambopata valley using satellite data. The United Nations Office on Drugs and Crime (UNODC 2009) indicates that 940 hectares of coca were cultivated in the upper Tambopata valley in 2008. The conventional coca cultivation density for regions with traditional coca growers could be between 35,000 and 40,000 bushes per hectare (UNODC 2001) (During the 90s, the coca cultivation density was lower, between 20,000 and 25,000 bushes per hectare (UNODC 2009)). The coca cultivation density in the particular valley is relatively low because coca growers intercrop coca with coffee and staples, although the yields per bush have increased during the last years (UNODC 2009). Therefore, it is expected that the total number of coca bushes for this valley would be approximately from 32.9 to 37.6 million.

Our sample of 477 respondents (excluding farmers who did not report their co-operative registration number and non respondents to the sensitive question) reported a total of 960,000 coca bushes. This sample corresponds to 14.6% of a total of 3,265 co-operative members in SPPP. Thus, extrapolating for the total number of co-operative members located in the SPPP district would result in a total of 6.6 million coca bushes. In addition, we need to consider that the upper Tambopata valley also includes San Juan del Oro district which has around the same population as SPPP district (INEI 2007). Under the very strong assumption that farmers in SPPP behave similarly to the farmers in San Juan del Oro - at least in terms of coca cultivation - this would double the number of coca bushes for the entire upper Tambopata valley to around 13.2 million. This last estimate is between 35 and 40% of the 32.9 to 37.6 million obtained from UNODC satellite data. This result is in the expected range of reporting on sensitive issues. For reporting on abortion, this range is between 35 to 59% (Fu, Darroch, Henshaw and Kolb 1998), and for the use of opiates or cocaine between 30 to 70% (Tourangeau and Yan 2007).

## 4. Summary and conclusions

Coca, a raw material for the production of cocaine, is cultivated in Colombia, Peru and Bolivia. In the latter two countries, traditional uses of coca by indigenous populations date back to around 3000 B.C. (Riv era, *et al.* 2005). Nevertheless, asking farmers about the extent of their coca cultivation areas is considered a sensitive question. Coca growers are afraid of eradication programs even if they do not sell coca to the narcotics traffic business because it is difficult to distinguish between coca growers whose production is commercially oriented and those who produce only for self-consumption. Thus, farmers tend not to participate in surveys, not to answer any sensitive questions, or to underreport their coca cultivation areas in an attempt to minimize their identification for possible eradication.

Against this background, household-level data collection procedures need to consider and evaluate strategies to reduce nonresponses and misreporting. Most of the strategies used in our research area in Peru were based on best practices reported in the literature review. Some of the strategies that worked in our case were establishment of trust with the farmers using a presentation letter from a coffee co-operative director, confidentiality assurance at the beginning and in the middle of the questionnaire, matching of interviewer-respondent ethnic background characteristics, training of interviewers to reduce their hesitance to ask sensitive questions, changing the format of the sensitive question to a familiar and forgiving wording, and non enforcement of absolute privacy to prevent each farmer from feeling that they were the "only one" who was interviewed.

The validity of farmers' individual responses on their coca area extensions cannot be checked because the topic has produced little prior empirical research, and there is an absence of other sources of household-level confirming data. Thus, the extent of misreporting was evaluated using aggregate data. The results suggest that farmers only reported between 35 to 40% of their actual coca areas. Still, those values are between the ranges of what could be expected for answers to sensitive questions. In terms of survey nonresponse and sensitive question nonresponses, the results were more encouraging indicating values of 10% and of around 4%, respectively.

When conducting the survey, we mainly took advantage of celebrations and co-operative General Assemblies for which farmers congregated in town, since farmers are otherwise highly dispersed in the rainforest. The survey followed a convenience sampling method but it was possible to test the representativeness of this sample because all of the farmers are registered in one of the co-operatives in the research area. The obtained sample was compared with a simulated simple random sample without replacement



where each farmer had the same probability to be selected by chance from the co-operative member lists. There were no statistical differences in the distribution functions, so the sample is equivalent to a simple random one. The main drawback of this approach is that after the interview, we needed to ask the respondents for their co-operative member number. Even though the respondents were told that the co-operative identification number was not attached to their questionnaires, some farmers might have had doubts about it, and this could have had effects on confidentiality assurance credibility in following interviews due to word spreading.

On the other hand, comparing the characteristics of non-respondents to sensitive questions with the rest of respondents indicates that non-respondents were highly risk averse. Even though the number of non-respondents was small (less than 4% of the total sample), this could suggest that the main reason for item non-reporting is the fear of the consequences of the information leaking to third parties.

The coca areas reported by the farmers were on average very small. This could be an attempt by commercial coca growers to appear to be cultivating only for self-consumption. Coca growing for traditional uses does not have a negative connotation *per se* given that it is a symbol of ethnicity and the indigenous population's struggle for self-determination (Office of Technology Assessment 1993). It is not possible to distinguish farmers who underreported the extent of their coca cultivation areas from those who grow coca for self-consumption. Unfortunately, commercial coca growers can take advantage of this situation to continue growing coca under the guise of traditional uses.

### Acknowledgements

The research was funded by BMZ (the Federal Ministry for Economic Cooperation and Development, Germany) through the DAAD (German Academic Exchange Service), and by LACEEP (Latin American and Caribbean Environmental Economics Program).

## Appendix 1

### Relevant parts of the questionnaire

#### A) Presentation:

Good morning/afternoon/night. My name is \_\_\_\_\_. I am a student at \_\_\_\_\_. We are conducting a survey to identify the risks and vulnerabilities of coffee producers in your community. The coffee co-operative directives are aware of this survey and believe that the result could benefit the community. If you decide to answer our questionnaire, you may skip any questions or withdraw from this study at any time. The data collected in this survey will remain CONFIDENTIAL and will be used only for ACAD EMIC purposes. Your answers and opinions are extremely important for the co-operative and us. Would you be prepared to respond to some questions?

- a) Yes (proceed)  
b) No (thank the respondent, withdraw the survey, and indicate the characteristics of the person in format 1)

#### B) Coca Related Questions:

In this part, we will ask about coca uses and cultivation. Please, remember that this survey is anonymous and that there are no correct or incorrect answers.

- |   |        |       |
|---|--------|-------|
| Do you chew coca leaves?  | a) Yes | b) No |
| Do you use coca leaves as medicine?   | a) Yes | b) No |
| Do you feel obligated to offer coca leaves to your guests during ayni and minka activities? | a) Yes | b) No |
| Do you use coca leaves for rituals?   | a) Yes | b) No |
| Do you use coca leaves for payment to external workers?                                     | a) Yes | b) No |
| Do you use coca leaves as product exchange or as a gift for friends and relatives?          | a) Yes | b) No |
| How many little bushes of coca do you have in your agricultural plot?                       | _____  |       |

#### C) Risk Aversion Question:

This is a game. Before playing it, you need to choose one of the options displayed below. Then I toss a coin. If for example you have chosen option H, and I toss the coin and it is heads, you do not win any money at all; but if it is tails, you win \$1.200. On the other hand, if you have chosen option A, you receive \$1.50 regardless of if the tossed coin is heads or tails. Which option from all of the above would you choose before I toss the coin?

OPTION	If it is heads, you win:	If it is tails, you win:
A	50 soles	50 soles
B	45 soles	95 soles
C	40 soles	120 soles
D	35 soles	125 soles
E	30 soles	150 soles
F	20 soles	160 soles
G	10 soles	190 soles
H	0 soles	200 soles

## Appendix 2

## Comparative descriptive statistics between all observations and sensitive question non respondents

Variable	All Observations <sup>a</sup>	Sensitive Question Non Respondent
Age	42.2 (12.6)	45.9 (9.9)
Male (%)	94.3***	100***
Aymara (%)	81.8**	94.7**
Number of Children	3.0** (2.0)	4.1** (2.0)
Years of schooling	8.4 (3.3)	7.5 (2.9)
Total area (ha)	7.9 (8.3)	6.8 (3.2)
Coffee area (ha)	2.2 (1.8)	2.5 (1.2)
Area secondary forest (fallow area)	1.6 (2.3)	1.4 (1.1)
Primary forest area (ha)	4.0 (7.3)	2.9 (3.3)
Staple food area (ha)	0.5 (0.7)	0.6 (0.6)
No other economic activities (%)	47.5	57.9
High risk aversion (%)	28.6***	73.7***
Important to obey national laws (%)	84.3	89.5
Negative change in trust in the last 5 years (%)	16.8	26.3
Have worked in community activities in 2007 (%)	89.4	89.5
Farmer chews coca (%)	67.7	73.7
Farmer uses coca as medicine (%)	72.0*	84.2*
Easy to sell coca leaves (%)	23.6	27.8
Number of Observations	477	19

Standard deviations are in parentheses for continuous variables.

a) All observations without sensitive question non respondents.

Non respondent means are statistically different from the entire sample (T-test with unequal variances) at:

\* 0.1 significance level, \*\* 0.05 significance level, \*\*\* 0.01 significance level.

Source: Own calculations.

## References

- Allen, C. (1981). To be Quechua: The symbolism of coca chewing in highland Peru. *American Ethnologist*, 8, 1, 157-171.
- Barnett, J. (1998). Sensitive questions and response effects: An evaluation. *Journal of Managerial Psychology*, 13, 1/2, 63-67.
- Bedoya, E. (2003). Estrategias productivas y el riesgo entre los cocaleros del valle de los ríos apurímac y ene. In *Amazonía: Procesos Demográficos y Ambientales*, (Eds., C. Aramburu and E. Bedoya), Consorcio de Investigación Económica y Social. Lima, Peru.
- Binswanger, H. (1980). Attitude towards risk: Experimental measurement in rural India. *American Journal of Economics*, 62, 395-407.
- Botvin, G., Griffin, K., Diaz, T., Scheier, L., Williams, C. and Epstein, J. (2000). Preventing illicit drug use in adolescents: Long-term follow-up data from a randomized control trial of a school population. *Addictive Behaviors*, 25, 5, 769-774.
- Caballero, V., Dietz, E., Taboada, C. and Anduaga, J. (1998). Diagnostico Rural Participativo de las Cuencas Alto Inambari y Alto Tambopata Provincia de Sandia, Departamento de Puno. GTZ. Lima, Peru.
- Caulkins, J., Reuter, P., Iguchi, M. and Chiesa, J. (2005). How goes the War on Drugs? An Assessment of U.S. Drug Problems and Policy. RAND Drug Policy Research Center. U.S.
- Collins, J. (1984). The maintenance of peasant coffee production in a peruvian valley. *American Ethnologist*, 11, 3, 413-438.
- Commission on Narcotic Drugs (2005). Alternative Development: A Global Thematic Evaluation. Final Synthesis Report. Forty - Eight Session E/CN.7/2005/CRP.3. Austria.
- Coutts, E., and Jann, B. (2008). Sensitive Questions in Online Surveys: Experimental Results for the Randomized Response Technique (RRT) and the Unmatched Count Technique (UCT). ETH Zurich Sociology, Working Paper, No. 3.
- Davalos, L., Bejarano, A. and Correa, L. (2008). Disabusing cocaine: Pervasive myths and enduring realities of a globalised commodity. *International Journal of Drug Policy*, 20, 5, 381-386.

- Davis, C., Thake, J. and Vilhe na, N. (2 009). Social Desirability Biases in Self-Reported Alcohol Consumption and Harms. Addictive Behaviors. Article in Press.
- Durand, F. (2005). El Problema Cocalero y el Comercio Informal para Uso Tradicional. Debate Agrario 39. Lima, Peru.
- Fergusson, D., Boden, J. and Horwood, L. (2008). The developmental antecedents of illicit drug use: Evidence from a 25-Year longitudinal study. *Drug and Alcohol Dependence*, 96, 165-177.
- Fu, H., Darroch, J., Henshaw, S. and Kolb, E. (1998). Measuring the extent of abortion underreporting in the 1995 National Survey of Family Growth. *Family Planning Perspectives*, 30, 3, 128-138.
- Garcia, J., and Antezana, J. (2009). Diagnostico de la Situación del Desvío de IQ al Narcotráfico. ConsultAndes and DEVIDA. Lima, Peru.
- Gibson, B., and Godoy, R. (1993). Alternatives to coca production in Bolivia: A computable general equilibrium approach. *World Development*, 21, 6, 1007-1021.
- Henman, A. (1990). Tradicion y represion: Dos experiencias en america del sur. In *Coca, Cocaína y Narcotráfico. Laberinto en los Andes*, (Eds., Garcia – D. Sayan), Comision Andina de Juristas. Lima, Peru.
- Holmstedt, B., Jaatmaa, E., Leander, K. and Plowman, T. (1977). Determination of cocaine in some South American species of erythroxyllum using mass fragmentography. *Phytochemistry*, 16, 1753-1755.
- INCB (2009). Report on the International Narcotics Control Board for 2009. United Nations Publication. New York, U.S.A.
- Ibanez, M., and Carlsson, F. (2010). A survey-based choice experiment on coca cultivation. *Journal of Development Economics*, 93, 2, 249-263.
- INEI (2007). Censos Nacionales 2007: XI de Población y VI de Vivienda. Lima, Peru.
- Johnson-Hanks, J. (2002). The lesser shame: Abortion among educated women in southern Cameroon. *Social Science & Medicine*, 55, 8, 1337-1349.
- Mansfield, D. (2006). Development in Drug Environment: A Strategic Approach to Alternative Development. Discussion Paper. Development Oriented Drug Control Program. GTZ. Germany.
- Mensch, B., Hewett, P. and Erulkar, A. (2003). The reporting of sensitive behavior by adolescents: A methodological experiment in Kenya. *Demography*, 40, 2, 247-268.
- Obando, E. (2006). U.S. Policy toward Peru: At odds for twenty years. In *Addicted to Failure. U.S. Security Policy in Latin America and the Andean Region*, (Eds., B. Loveman). Rowman & Littlefield Publishers Inc. U.S.
- Office of Technology Assessment (1993). Alternative Coca Reduction Strategies in the Andean Region. U.S. Congress. OTA-F-556. Washington, U.S.
- Punch, K. (2003). Survey research. The basics. *Sage Publications, Inc.* U.K.
- Rivera, M., Aufderheide, A., Cartmell, L., Torres, C. and Langsjoen, O. (2005). Antiquity of coca – Leaf chewing in the south central Andes: A 3000 year archaeological record of coca - Leaf chewing from Northern Chile. *Journal of Psychoactive Drugs*, 37, 4, 455-458.
- Rospigliosi, F. (2004). Analisis de la Encuesta DEVIDA-INEI. In *El Consumo Tradicional de la Hoja de Coca en el Peru*, (Ed., F. Rospigliosi). Instituto de Estudios Peruanos. Lima, Peru.
- Singer, E., Hippler, H. and Schwarz, N. (1992). Confidentiality assurances in surveys: Reassurance or threat? *International Journal of Public Opinion Research*, 4, 3.
- Struin, L. (2001). Assessing alcohol consumption: developments from qualitative research methods. *Social Science & Medicine*, 53, 2, 215-226.
- Thoumi, F. (2003). Illegal Drugs, Economy, and Society in the Andes. Woodrow Wilson Center Press. Washington, U.S.
- Torrico, J., Pohlman, H. and Janssens, M. (2005). Alternatives for the transformation of drug production areas in the chapare region, Bolivia. *Journal of Food, Agriculture and Development*, 3, 3-4, 167-172.
- Tourangeau, R., and Yan, T. (2007). Sensitive questions in surveys. *Psychological Bulletin*, 133, 5, 859-883.
- UNODC (2001). Alternative Development in the Andean Area. The UNDCP Experience. Revised Edition. ODCCP Studies on Drugs and Crime. New York, U.S.
- UNODC (2009). Perú. Monitoreo de Cultivos de Coca 2008. Lima, Peru.
- UNODC (2011). Perú. Monitoreo de Cultivos de Coca 2010. Lima, Peru.
- UNODC Office in Peru (1999). Desarrollo Alternativo del Inambari y Tambopata. Documento de Proyecto AD/PER/99/D96. Availability: <http://www.onudd.org.pe/web/Html/Templates/proyectos.htm> (accessed on June 15, 2009).
- U.S. Department of State (2009). International Narcotics Control Strategy Report. Volume I: Drug and Chemical Control. Bureau for International Narcotics and Law Enforcement Affairs. U.S.
- Varkey, P., Balakrishna, P., Prasad, J., Abraham, S. and Joseph, A. (2000). The reality of unsafe abortion in a rural community in South India. *Reproductive Health Matters*, 8, 16, 83-91.
- Willis, G. (2005). Cognitive interviewing. A tool for improving questionnaire design. *Sage Publications, Inc.* U.S.
- Zufferey, A., Michaud, P., Jeannin, A., Berchtold, A., Chossis, I., van Melle, G. and Suris, J. (2007). Cumulative risk factors for adolescent alcohol misuse and its perceived consequences among 16 to 20 year old adolescents in Switzerland. *Preventive Medicine*, 45, 2-3, 233-239.

**ELECTRONIC  
PUBLICATIONS  
AVAILABLE AT**

**PUBLICATIONS  
ÉLECTRONIQUES  
DISPONIBLE À**

**[www.statcan.gc.ca](http://www.statcan.gc.ca)**

# Imputation for nonmonotone nonresponse in the survey of industrial research and development

Jun Shao, Martin Klein and Jing Xu<sup>1</sup>

## Abstract

Nonresponse in longitudinal studies often occurs in a nonmonotone pattern. In the Survey of Industrial Research and Development (SIRD), it is reasonable to assume that the nonresponse mechanism is past-value-dependent in the sense that the response propensity of a study variable at time point  $t$  depends on response status and observed or missing values of the same variable at time points prior to  $t$ . Since this nonresponse is nonignorable, the parametric likelihood approach is sensitive to the specification of parametric models on both the joint distribution of variables at different time points and the nonresponse mechanism. The nonmonotone nonresponse also limits the application of inverse propensity weighting methods. By discarding all observed data from a subject after its first missing value, one can create a dataset with a monotone ignorable nonresponse and then apply established methods for ignorable nonresponse. However, discarding observed data is not desirable and it may result in inefficient estimators when many observed data are discarded. We propose to impute nonrespondents through regression under imputation models carefully created under the past-value-dependent nonresponse mechanism. This method does not require any parametric model on the joint distribution of the variables across time points or the nonresponse mechanism. Performance of the estimated means based on the proposed imputation method is investigated through some simulation studies and empirical analysis of the SIRD data.

Key Words: Bootstrap; Imputation model; Kernel regression; Missing not at random; Longitudinal study; Past-value-dependent.

## 1. Introduction

Longitudinal studies, in which data are collected from every sampled subject at multiple time points, are very common in research areas such as medicine, population health, economics, social sciences, and sample surveys. The statistical analysis in a sample survey typically aims to estimate or make inference on the mean of a study variable at each time point. Nonresponse or missing data in the study variable is a serious impediment to performing a valid statistical analysis, because the response propensity (PSI) may directly or indirectly depend on the value of the study variable. Nonresponse is monotone if, whenever a value is missing at a time point  $t$ , all future values at  $s > t$  are missing. We focus on nonmonotone nonresponse, which often occurs in longitudinal surveys. In the Survey of Industrial Research and Development (SIRD) conducted jointly by the U.S. Census Bureau and the U.S. National Science Foundation (NSF), for example, a business may be a nonrespondent on research and development expenditures at year  $t - 1$  but a respondent at year  $t$ . For ease we refer to SIRD in the present tense throughout, but we note that as of 2008, it has been replaced by the Business R&D and Innovation Survey.

Some existing methods for handling nonmonotone nonresponse can be briefly described as follows. The parametric approach assumes parametric models for both the PSI and

the joint distribution of the study variable across time points (e.g., Troxel, Harrington and Lipsitz 1998, Troxel, Lipsitz and Harrington 1998). The validity of the parametric approach, however, depends on whether parametric models are correctly specified. Vansteelandt, Rotnitzky and Robins (2007) proposed some methods under some models of the PSI at time  $t$  conditional on observed past data. Xu, Shao, Palta and Wang (2008) derived an imputation procedure under the assumptions that (i) the PSI at  $t$  depends only on values of the study variable at time  $t - 1$  and (ii) the study variables over different time points is a Markov chain. Another approach, which will be referred to as censoring, is to create a dataset with “monotone nonresponse” by discarding all observed values of the study variable from a sampled subject after its first missing value. Methods appropriate for monotone nonresponse (e.g., Diggle and Kenward 1994, Robins and Rotnitzky 1995, Paik 1997) can then be applied to the reduced dataset. This approach may be inefficient when many observed data are discarded. Furthermore, in practical applications it is not desirable to throw away observed data.

The purpose of this article is to propose an imputation method for longitudinal data with nonmonotone nonresponse under the past-value-dependent PSI assumption described by Little (1995): at a time point  $t$ , the nonresponse propensity depends on values of the study variable at time points prior to  $t$ . This assumption on the PSI is weaker than

1. Jun Shao, Department of Statistics, University of Wisconsin, Madison, WI 53706. E-mail: shao@stat.wisc.edu; Martin Klein, Center for Statistical Research and Methodology, U.S. Census Bureau, Washington, D.C. 20233; Jing Xu, Department of Statistics, University of Wisconsin, Madison, WI 53706.

that in Xu *et al.* (2008) and is different from those in Vansteelandt *et al.* (2007). We consider imputation which does not require building a model for the PSI. Imputation is commonly used to compensate for missing values in survey problems (Kalton and Kasprzyk 1986). Once all missing values are imputed, estimates of parameters are computed using the estimated means for complete data by treating imputed values as observations. The proposed imputation and estimation methodology, including a bootstrap method for variance estimation, is introduced in Section 2. To examine the finite sample performance of the proposed method, we present some simulation results in Section 3. We also include an application of the proposed method to the SIRD. The last section contains some concluding remarks.

## 2. Methodology

We consider the model-assisted approach for survey data sampled from a finite population  $P$ . We assume that the population  $P$  is divided into a fixed number of imputation classes, which are typically unions of some strata. Within each imputation class, the study variable from a population unit follows a superpopulation. Let  $y_t$  be the study variable at time point  $t$ ,  $t = 1, \dots, T$ ,  $\mathbf{y} = (y_1, \dots, y_T)$ ,  $\delta_t$  be the indicator of whether  $y_t$  is observed, and  $\boldsymbol{\delta} = (\delta_1, \dots, \delta_T)$ . Since imputation is carried out independently within each imputation class, for simplicity of notation we assume in this section that there is only a single imputation class.

Throughout this paper, we consider nonmonotone nonresponse and assume that there is no nonresponse at baseline  $t = 1$ . The PSI is past-value-dependent if

$$P(\delta_t = 1 \mid \mathbf{y}, \delta_1, \dots, \delta_{t-1}, \delta_{t+1}, \dots, \delta_T) = P(\delta_t = 1 \mid y_1, \dots, y_{t-1}, \delta_1, \dots, \delta_{t-1}), \quad t = 2, \dots, T, \quad (1)$$

where  $P$  is with respect to the superpopulation. When nonresponse is monotone, the past-value-dependent PSI becomes ignorable (Little and Rubin 2002), since we either observe all past values or know with certainty that  $y_t$  is missing if it is missing at  $t - 1$ , and an imputation method using linear regression proposed by Paik (1997) can be used. When nonresponse is nonmonotone, however, the past-value-dependent PSI is nonignorable because the response indicator at time  $t$  is statistically dependent upon previous values of the study variable, some of which may not be observed. In this case Paik's method does not apply.

### 2.1 Imputation for subjects whose first missing is at $t$

Let  $t > 1$  be a fixed time point and  $r + 1$  be the time point at which the first missing value of  $\mathbf{y}$  occurs. When  $r + 1 = t$ , *i.e.*, a subject whose first missing value is at  $t$ ,

our proposed imputation procedure is the same as that for the case of monotone nonresponse (Paik 1997). However, we still need to provide a justification since we have a different PSI. It is shown in the Appendix that, under assumption (1),

$$E(y_t \mid y_1, \dots, y_{t-1}, \delta_1 = \dots = \delta_{t-1} = 1, \delta_t = 0) = E(y_t \mid y_1, \dots, y_{t-1}, \delta_1 = \dots = \delta_{t-1} = 1, \delta_t = 1) \quad t = 2, \dots, T, \quad (2)$$

where  $E$  is the expectation with respect to the superpopulation. Denote the quantity on the first line of (2) by  $\phi_{t,t-1}(y_1, \dots, y_{t-1})$ , which is the conditional expectation of a missing  $y_t$  given observed  $y_1, \dots, y_{t-1}$ . If  $\phi_{t,t-1}$  is known, then a natural imputed value for  $y_t$  is  $\phi_{t,t-1}(y_1, \dots, y_{t-1})$ . However,  $\phi_{t,t-1}$  is usually unknown. Since  $\phi_{t,t-1}$  cannot be estimated by regressing  $y_t$  on  $y_1, \dots, y_{t-1}$  based on data from subjects with missing  $y_t$  values, we need to use (2), *i.e.*, the fact that  $\phi_{t,t-1}$  is the same as the quantity on the second line of (2), which is the conditional expectation of an observed  $y_t$  given observed  $y_1, \dots, y_{t-1}$  and can be estimated by regressing  $y_t$  on  $y_1, \dots, y_{t-1}$ , using data from all subjects having observed  $y_t$  and observed  $y_1, \dots, y_{t-1}$ . Note that (2) is a counterpart of (5) in Xu *et al.* (2008) under the last-value-dependent assumption, which is stronger than the past-value-dependent assumption (1). Under a stronger assumption, we are able to utilize more data in regression fitting.

Suppose that a sample  $S$  is selected from  $P$  according to a given probability sampling plan. For each  $i \in S$ ,  $\boldsymbol{\delta}_i = (\delta_{i1}, \dots, \delta_{iT})$  is observed, the study variable  $y_{it}$  with  $\delta_{it} = 1$  is observed, and  $y_{it}$  with  $\delta_{it} = 0$  is not observed,  $t = 1, \dots, T$ . With respect to the superpopulation,  $(\mathbf{y}_i, \boldsymbol{\delta}_i)$  has the same distribution as  $(\mathbf{y}, \boldsymbol{\delta})$  and  $(\mathbf{y}_i, \boldsymbol{\delta}_i)$ 's are independent, where  $\mathbf{y}_i = (y_{i1}, \dots, y_{iT})$ . For  $t = 2, \dots, T$ , let  $\hat{\phi}_{t,t-1}$  be the regression estimator of  $\phi_{t,t-1}$  based on observations with  $\delta_{i1} = \dots = \delta_{i(t-1)} = 1$ . A missing  $y_{it}$  with observed  $y_{i1}, \dots, y_{i(t-1)}$  is then imputed by  $\tilde{y}_{it} = \hat{\phi}_{t,t-1}(y_{i1}, \dots, y_{i(t-1)})$ .

To illustrate, we consider the case of  $t = 3$  or 4. The horizontal direction in Table 1 corresponds to time points and the vertical direction corresponds to different missing patterns, where each pattern is represented by a vector of 0's and 1's with 0 indicating a missing value and 1 indicating an observed value. For  $t = 3$  and  $r = 2$ , as the first of the two steps, we consider missing data at time 3 with first missing at time 3, *i.e.*, pattern (1,1,0). According to imputation model (2), we fit a regression using data in pattern (1,1,1) indicated by + (used as predictors) and  $\times$  (used as responses). Then, imputed values (indicated by  $\circ$ ) are obtained from the fitted regression using data indicated by \* as predictors. For  $t = 4$  and  $r = 3$ , imputation in pattern (1,1,1,0) can be similarly done using data in pattern (1,1,1,1) for regression fitting.

**Table 1**  
**Illustration of imputation process**

Pattern	Step 1: $r = 2, t = 3$			Step 2: $r = 1, t = 3$		
	Time 1	Time 2	Time 3	Time 1	Time 2	Time 3
(1,0,0)				*		○
(1,1,0)	*	*	○	+		⊗
(1,1,1)	+	+	×			
(1,0,1)						

Pattern	Step 1: $r = 3, t = 4$				Step 2: $r = 2, t = 4$				Step 3: $r = 1, t = 4$			
	Time 1	Time 2	Time 3	Time 4	Time 1	Time 2	Time 3	Time 4	Time 1	Time 2	Time 3	Time 4
(1,0,0,0)									*			○
(1,1,0,0)					*	*		○	+			⊗
(1,1,1,0)	*	*	*	○	+	+		⊗	+			⊗
(1,0,1,0)									*			○
(1,0,0,1)												
(1,1,0,1)												
(1,0,1,1)												
(1,1,1,1)	+	+	+	×								

+ : observed data used in regression fitting as predictors.  
 × : observed data used in regression fitting as responses.  
 ⊗ : imputed data used in regression fitting as responses.  
 \* : observed data used as predictors in imputation.  
 ○ : imputed values.

What type of regression we can fit to obtain  $\hat{y}_{it}$ ? It is shown in the Appendix that, if (1) holds and  $E(y_t | y_1, \dots, y_{t-1})$  is linear in  $y_1, \dots, y_{t-1}$  for any  $t$  in the case of no nonresponse, then

$$E(y_t | y_1, \dots, y_{t-1}, \delta_1 = \dots = \delta_{t-1} = 1) \text{ is linear in } y_1, \dots, y_{t-1} \quad (3)$$

and, hence, linear regression under the model-assisted approach can be used to estimate  $\phi_{t,t-1}$ . If  $E(y_t | y_1, \dots, y_{t-1})$  is not linear, one of the methods described in Section 2.3 can be applied.

**2.2 Imputation for subjects whose first missing is at  $r + 1 < t$**

Imputation for a subject whose first missing value is at time  $r + 1 < t$  is more complicated and very different from that for the case of monotone nonresponse. This is because when  $r + 1 < t$  and nonresponse is monotone,

$$E(y_t | y_1, \dots, y_r, \delta_1 = \dots = \delta_r = 1, \delta_t = 0) = E(y_t | y_1, \dots, y_r, \delta_1 = \dots = \delta_r = 1, \delta_t = 1) \quad r = 1, \dots, t - 2, \quad t = 2, \dots, T, \quad (4)$$

whereas (4) does not hold when nonresponse is non-monotone (see the proof in the Appendix). Hence, we need to construct different models for subjects whose first missing value is at  $r + 1 < t$ . It is shown in the Appendix that, when  $r + 1 < t$ ,

$$E(y_t | y_1, \dots, y_r, \delta_1 = \dots = \delta_r = 1, \delta_{r+1} = 0, \delta_t = 0) = E(y_t | y_1, \dots, y_r, \delta_1 = \dots = \delta_r = 1, \delta_{r+1} = 1, \delta_t = 0) \quad r = 1, \dots, t - 2, \quad t = 2, \dots, T. \quad (5)$$

We now explain how to use (5) to impute missing values at a fixed time point  $t$ . Let  $\phi_{t,r}(y_1, \dots, y_r)$  be the quantity on the first line of (5). If  $\phi_{t,r}$  is known, then  $y_t$  can be imputed by  $\phi_{t,r}(y_1, \dots, y_r)$ . Otherwise, it needs to be estimated based on (5). Unlike in model (2) or (4), the conditional expectation on the second line of (5) is conditional on a missing  $y_t$  ( $\delta_t = 0$ ), although  $y_1, \dots, y_r$  are observed. If we carry out imputation sequentially according to  $r = t - 1, t - 2, \dots, 1$ , then, for a given  $r < t - 1$ , the missing  $y_t$  values from subjects whose first missing is at time point  $r + 2$  have already been imputed using the method in this section or Section 2.1. We can fit a regression between imputed  $y_t$  and observed  $y_1, \dots, y_r$  using data from all subjects having already imputed  $y_t$  (used as responses), observed  $y_1, \dots, y_r$  (used as predictors), and  $\delta_{r+1} = 1$ . Once an estimator  $\hat{\phi}_{t,r}$  is obtained, a missing  $y_{it}$  with first missing at  $r + 1$  is then imputed by  $\hat{y}_{it} = \hat{\phi}_{t,r}(y_{i1}, \dots, y_{ir})$ .

Consider again the case of  $t = 3$  or 4 and Table 1. Following the first step for  $t = 3$  discussed in Section 2.1, at the second step, we impute missing values with  $r = 1$  in pattern (1,0,0). According to imputation model (5), we fit a regression using data in pattern (1,1,0) indicated by + (used as predictors) and ⊗ (previously imputed values used as

responses). Then, imputed values (indicated by  $\circ$ ) are obtained from the fitted regression using data indicated by  $*$  as predictors. For  $t = 4$ , following the first step discussed in Section 2.1, at the second step ( $r = 2$ ) we fit a regression using data in pattern (1,1,1,0) indicated by  $+$  (used as predictors) and  $\otimes$  (previously imputed values used as responses). Then, imputed values (indicated by  $\circ$ ) at  $t = 4$  in pattern (1,1,0,0) are obtained from the fitted regression using data indicated by  $*$  as predictors. At step 3 for  $t = 4$ , we fit a regression using data in patterns (1,1,0,0) and (1,1,1,0) indicated by  $+$  (used as predictors) and  $\otimes$  (previously imputed values used as responses). Then, imputed values (indicated by  $\circ$ ) at  $t = 4$  in patterns (1,0,0,0) and (1,0,1,0) are obtained from the fitted regression using data indicated by  $*$  as predictors.

Although at time  $t$ , imputation has to be carried out sequentially as  $r = t - 1, \dots, 1$ , imputation for different time points can be done in any order. This can be seen from the illustration given by Table 1, where the imputed values at  $t = 3$  are not involved in the imputation process at  $t = 4$  or vice versa, although some observed data will be repeatedly used in regression fitting. When data come according to time, it is natural to impute nonrespondents in the order  $t = 2, \dots, T$ .

Why can we use previously imputed values as responses in the estimation of the regression function  $\phi_{t,r}$  when  $r < t - 1$ ? For given  $t$  and  $r < t - 1$ , a previously imputed value with first missing at  $s + 1 > r + 1$  is an estimator of

$$\begin{aligned} \tilde{y}_t &= E(y_t | y_1, \dots, y_s, \delta_1 = \dots = \delta_s = 1, \delta_{s+1} = 0, \delta_t = 0) \\ &= E(y_t | y_1, \dots, y_s, \delta_1 = \dots = \delta_{s+1} = 1, \delta_t = 0). \end{aligned}$$

By the property of conditional expectation and (5),

$$\begin{aligned} E[E(y_t | y_1, \dots, y_s, \delta_1 = \dots = \delta_{s+1} = 1, \delta_t = 0) | \\ y_1, \dots, y_r, \delta_1 = \dots = \delta_{r+1} = 1, \delta_t = 0] \\ &= E(y_t | y_1, \dots, y_r, \delta_1 = \dots = \delta_{r+1} = 1, \delta_t = 0) \\ &= E(y_t | y_1, \dots, y_r, \delta_1 = \dots = \delta_r = 1, \delta_{r+1} = 0, \delta_t = 0). \end{aligned} \quad (6)$$

This means that  $y_t$  and  $\tilde{y}_t$  have the same conditional expectation, given  $y_1, \dots, y_r, \delta_1 = \dots = \delta_r = 1, \delta_{r+1} = 0, \delta_t = 0$ . Therefore, using previously imputed values as responses in regression produces a valid estimator of  $\phi_{t,r}$ . Note that previously imputed values should not be used as predictors in regression, as equation (6) does not hold if some of  $y_1, \dots, y_s$  are imputed values.

Although all observed data at any time  $t$  are used for the estimation of  $E(y_t)$ , some but not all observed data at time  $< t$  are utilized in imputation to avoid biases under nonignorable nonresponse. This is different in the ignorable nonresponse case, where typically all past observed data can be used in regression imputation.

### 2.3 Regression for imputation

The conditional expectations in (5) depend not only on the distribution of  $y$ , but also on the PSI. Even if  $E(y_t | y_1, \dots, y_{t-1})$  is linear, conditional expectations in (5) are not necessarily linear, which is different from the case of  $r + 1 = t$  considered in Section 2.1. An example is given by result (10) in the Appendix.

When we do not have a suitable parametric model for  $\phi_{t,r}$ , the nonparametric kernel regression method given in Cheng (1994) may be applied to obtain  $\hat{\phi}_{t,r}$ . Since the regressor  $(y_{i1}, \dots, y_{ir})$  is multivariate when  $r \geq 2$ , however, kernel regression has a large variability unless the number of sampled subjects in the category defined by  $\delta_{i1} = \dots = \delta_{i(r+1)} = 1$  is very large. This issue is commonly referred to as the curse of dimensionality.

Thus, we consider the following alternatives under the additional assumption that the dependence of  $\delta_t$  on  $y_1, \dots, y_{t-1}$  is through a linear combination of  $y_1, \dots, y_{t-1}$ . That is,

$$P(\delta_t = 1 | y_1, \dots, y_{t-1}, \delta_1, \dots, \delta_{t-1}) = \Psi \left( \sum_{l=1}^{t-1} \gamma_l^{\delta_1, \dots, \delta_{t-1}} y_l \right), \quad (7)$$

where  $\gamma_l^{\delta_1, \dots, \delta_{t-1}}, l = 1, \dots, t - 1$ , are unknown parameters depending on  $\delta_1, \dots, \delta_{t-1}$  and  $\Psi$  is an unknown function with range  $[0, 1]$ . Under (7), it is shown in the Appendix that

$$\begin{aligned} E(y_t | z_r, \delta_1 = \dots = \delta_r = 1, \delta_{r+1} = 0, \delta_t = 0) \\ &= E(y_t | z_r, \delta_1 = \dots = \delta_r = 1, \delta_{r+1} = 1, \delta_t = 0) \\ &\quad r = 1, \dots, t - 2, t = 2, \dots, T, \end{aligned} \quad (8)$$

where  $z_r = \sum_{l=1}^r \gamma_{r,l} y_l$  and  $\gamma_{r,l} = \gamma_l^{\delta_1, \dots, \delta_r}$  with  $\delta_1 = \dots = \delta_r = 1$ . Hence, to impute nonrespondents, we can condition on the linear combination  $z_r$  and use (8), instead of conditioning on  $y_1, \dots, y_r$  and using (5).

Let  $\psi_{t,r}(z_r)$  be the function defined on the second line of (8). Note that  $\psi_{t,r}$  is not necessarily the same as  $\phi_{t,r}$ . If there is a strong linear relationship between  $y_t$  and  $y_1, \dots, y_r$ , then  $\psi_{t,r}$  may be approximately linear so that we can fit a linear regression to obtain an estimator  $\hat{\psi}_{t,r}$ . In theory, this method is biased when  $\psi_{t,r}$  is not linear. If  $\gamma_r = (\gamma_{r,1}, \dots, \gamma_{r,r})'$  is known, then we can apply a one-dimensional kernel regression to obtain an estimator  $\hat{\psi}_{t,r}$ , using the one-dimensional index  $z_r$ . Since  $\gamma_r$  is unknown, we first need to estimate it by  $\hat{\gamma}_r$  and then obtain  $\hat{\psi}_{t,r}$  by applying the one-dimensional kernel regression with  $\gamma_r$  replaced by  $\hat{\gamma}_r$ . For example, the sliced inverse regression (Duan and Li 1991) can be applied to obtain  $\hat{\gamma}_r$ . However, this type of nonparametric method may be inefficient. If there is a strong linear relationship between  $y_t$  and  $y_1, \dots, y_r$ , we may apply linear regression to obtain  $\hat{\gamma}_r$ . In any case, we use  $y_{i1}, \dots, y_{ir}$  with  $\delta_{i1} = \dots = \delta_{i(r+1)} = 1$  as predictors and imputed  $y_{it}$  values as responses in any type



of regression fitting. After  $\hat{\psi}_{t,r}$  and  $\hat{\gamma}_r = (\hat{\gamma}_{r,1}, \dots, \hat{\gamma}_{r,r})'$  are obtained, a missing  $y_{it}$  is imputed by  $\tilde{y}_{it} = \hat{\psi}_{t,r}(\hat{\gamma}_{r,1}y_{i1} + \dots + \hat{\gamma}_{r,r}y_{ir})$ .

We refer to the method of simply applying linear regression as the linear regression imputation method, and the method of applying kernel regression to the index  $z_r$  as the one-dimensional index kernel regression imputation method. An advantage of one-dimensional index kernel regression imputation over kernel regression imputation is that only a one-dimensional kernel regression is applied and, thus, it avoids the curse of dimensionality and has smaller variability.

These methods can also be applied to the case of  $r = t - 1$  if  $E(y_t | y_1, \dots, y_{t-1})$  is not linear.

In theory, estimators such as the estimated means based on kernel regression or one-dimensional index kernel regression imputation are asymptotically unbiased, but they may not be better than those based on linear regression imputation when the number of sampled subjects in each  $(t, r)$  category is not very large. The performances of the estimated means based on linear regression, kernel regression, and one-dimensional index kernel regression imputation are examined by simulation in Section 3.

## 2.4 Estimation

We consider the estimation of the finite population total or the mean of  $y_t$  at each fixed  $t$ , which is often the main purpose of a survey study. At any  $t$ , let  $\tilde{y}_{it} = y_{it}$  when  $\delta_{it} = 1$  and  $\tilde{y}_{it}$  be the imputed value using one of the methods in Section 2 when  $\delta_{it} = 0$ . The finite population total and the mean of  $y_t$  can be estimated by

$$\hat{Y}_t = \sum_{i \in S} w_i \tilde{y}_{it} \quad \text{and} \quad \bar{Y}_t = \sum_{i \in S} w_i \tilde{y}_{it} / \sum_{i \in S} w_i, \quad (9)$$

respectively, where  $w_i$  is the survey weight constructed such that, in the case of no nonresponse,  $\hat{Y}_t$  is an unbiased estimator of the finite population total at time  $t$  with respect to the probability sampling. The superpopulation mean of  $y_t$  can also be estimated by  $\bar{Y}_t$ . Note that  $\sum_{i \in S} w_i$  is an unbiased estimator of the finite population size  $N$  and, for some simple sampling designs, it is exactly equal to  $N$ .

The survey weights should also be used in the regression fitting for imputation. Under the same conditions given in Cheng (1994),  $\hat{Y}_t$  or  $\bar{Y}_t$  based on kernel regression or one-dimensional index kernel regression imputation is consistent and asymptotically normal as the sample size increases to  $\infty$ . The required conditions and proofs can be found in Xu (2007).

If we apply the linear regression imputation method as discussed in Section 2.3, then the resulting estimated mean at  $t$  may be asymptotically biased. This bias is small if the function  $\psi_{t,r}$  can be well approximated by a linear function in the range of the data values. On the other hand, kernel or

one-dimensional index kernel regression imputation may require a much larger sample size than that for linear regression imputation. Hence, the overall performance of the estimated mean based on linear regression imputation may still be better, as indicated by the simulation results in Section 3.

## 2.5 Variance estimation

For assessing statistical accuracy or inference such as constructing a confidence interval for the mean of  $y_t$  at  $t$ , we need variance estimators of  $\hat{Y}_t$  or  $\bar{Y}_t$  based on imputed data. Because of the complexity of the imputation procedure, it is difficult to obtain explicit formulas for variance of  $\hat{Y}_t$  or  $\bar{Y}_t$ . The bootstrap method (Efron 1979) is then considered. A correct bootstrap can be obtained by repeating the process of imputation in each of the bootstrap samples (Shao and Sitter 1996). Let  $\hat{\theta}$  be the estimator under consideration. A bootstrap procedure can be carried out as follows.

1. Draw a bootstrap sample as a simple random sample of the same size as  $S$  with replacement from the set of sampled subjects.
2. For units in the bootstrap sample, their survey weights, response indicators, and observed data from the original data set are used to form a bootstrap data set. Apply the proposed imputation procedure to the bootstrap data. Calculate the bootstrap analog  $\hat{\theta}^*$  of  $\hat{\theta}$ .
3. Independently repeat the previous steps  $B$  times to obtain  $\hat{\theta}^{*1}, \dots, \hat{\theta}^{*B}$ . The sample variance of  $\hat{\theta}^{*1}, \dots, \hat{\theta}^{*B}$  is the bootstrap variance estimator for  $\hat{\theta}$ .

In application, each  $\hat{\theta}^{*b}$  can be calculated using the  $b^{\text{th}}$  bootstrap data  $(\mathbf{y}_i, \delta_i, w_i^{*b})$ ,  $i \in S$ , where  $w_i^{*b} = w_i$  multiplied by the number of times unit  $i$  appears in the  $b^{\text{th}}$  bootstrap sample. Note that the same  $w_i^{*b}$  can be used for all variables of interest, not just  $y_t$ .

## 3. Empirical results

We study  $\hat{Y}_t$  or  $\bar{Y}_t$  in (9) based on the proposed imputation methods at each time point  $t$ . We first consider a simulation with a normal population for the  $y_t$ 's. An application to the SIRD data is presented next. To examine the performance of the proposed methods for the SIRD, a simulation with a population generated using the SIRD data is presented in the end. We have implemented the proposed imputation methods in R (R Development Core Team 2009). To fit the required nonparametric regressions, we use the R function *loess* with default settings, which fits a local polynomial surface in one or more regressor variables. The required linear regressions are easily fit in R using the

function  $lm$ . Our implementations of the proposed methods include error checking; (such as ensuring that there are sufficient points for regression fitting at each stage) which is particularly important in bootstrap and simulation settings where the imputation methods are replicated many times, and each iteration cannot be examined manually. We defaulted to an overall mean imputation in cases where there were not enough data points to fit a regression.

### 3.1 Simulation results from a normal population

A simulation study was conducted with normally distributed  $y_1, \dots, y_n$ ,  $n = 2,000$ , and  $T = 4$ . A single imputation class and simple random sampling with replacement was considered. In the simulation,  $y_i$ 's were independently generated from the multivariate normal distribution with mean vector (1.33, 1.94, 2.73, 3.67) and the covariance matrix having the AR(1) structure with correlation coefficient 0.7 and unit variance; all data at  $t = 1$  were observed; missing data at  $t = 2, 3, 4$  were generated according to

$$P(\delta_t = 1 \mid y_1, \dots, y_{t-1}, \delta_1, \dots, \delta_{t-1}) = 1 - \Phi\left(0.6\left(1 - \sum_{j=1}^{t-1} y_j \gamma_j^{\delta_1 \dots \delta_{t-1}}\right)\right)$$

where

$$\gamma_j^{\delta_1, \dots, \delta_{t-1}} = \frac{j + (1 - \delta_j)j}{\sum_{k=1}^{t-1} [k + (1 - \delta_k)k]}, \quad j = 1, \dots, t - 1,$$

and  $\Phi$  is the standard normal distribution function. The unconditional probabilities of nonresponse patterns are given in Table 2.

For comparison, we included a total of nine estimators of the mean of  $y_t$ : they are sample means based on (1) the complete data (used as the gold standard); (2) respondents with adjusted weights assuming the probability of response is the same within each imputation class; (3) censoring and linear regression imputation, which first discards all observations of a subject after the first missing value to create a dataset with “monotone nonresponse” and then applies linear regression imputation as described in Paik (1997); (4) the proposed kernel regression imputation; (5) the proposed linear regression imputation; (6) the proposed one-dimensional index kernel regression imputation using the sliced inverse regression to obtain  $\hat{\gamma}_r$ ; (7) the kernel regression imputation proposed in Xu *et al.* (2008) based on the last-value-dependent PSI; (8) the linear regression imputation based on a regression between respondents at time  $t$  and observed and imputed values at time points  $1, \dots, t - 1$  (treating imputed as observed); (9) the linear regression imputation based on a regression between respondents at time  $t$  and observed data from units with the same missing pattern at time points  $1, \dots, t - 1$ .

**Table 2**  
Probabilities of nonresponse patterns in the simulation study (Normal population)

	Pattern	Probability	
Monotone	(1, 0, 0, 0)	0.062	total = 0.181
	(1, 1, 0, 0)	0.043	
	(1, 1, 1, 0)	0.076	
Intermittent	(1, 0, 0, 1)	0.113	total = 0.494
	(1, 0, 1, 0)	0.071	
	(1, 0, 1, 1)	0.186	
	(1, 1, 0, 1)	0.124	
Complete	(1, 1, 1, 1)	0.325	

Method (2) simply ignores nonrespondents and, hence, is biased and inefficient. Under the PSI assumption (1) methods (7)-(9) are also biased for  $t \geq 3$ , because method (7) requires the last-value-dependent assumption that is stronger than (1), method (8) treats previously imputed values as observed in regression, and method (9) requires the following condition that is not true under (1):

$$E(y_t \mid y_1, \dots, y_{t-1}, \delta_1 = j_1, \dots, \delta_{t-1} = j_{t-1}, \delta_t = 0) = E(y_t \mid y_1, \dots, y_{t-1}, \delta_1 = j_1, \dots, \delta_{t-1} = j_{t-1}, \delta_t = 1)$$

where  $(j_1, \dots, j_{t-1})$  is a fixed missing pattern. Finally, as we discussed in Section 2.3, method (5) is also biased for  $t \geq 3$  since linear regression is not an exactly correct model. However, methods (5), (8), and (9) may still perform well when the biases are not substantial, because the use of a simpler model and more data in regression for imputation may compensate for the loss in biased imputation. Furthermore, any assumption on the PSI may hold only approximately and it is desired to empirically study various methods in any particular application.

For the case of  $r = t - 1$ , linear regression imputation is applied as discussed in Section 2.1. Hence, methods (3)-(6), (8)-(9) all give the same results when  $t = 2$ .

Table 3 reports (based on 1,000 simulation runs) the relative bias and standard deviation (SD) of the mean estimator, the mean of  $\widehat{SD}_{boot}$ , the bootstrap estimator of SD based on 200 bootstrap replications, and the coverage probability of the approximate 95% confidence interval (CI) obtained using point estimator  $\pm 1.96 \times \widehat{SD}_{boot}$ . The following is a summary of the results in Table 3.

1. The sample mean based on ignoring missing data is clearly biased. Although in the case of  $t = 4$  its relative bias is only 3.5%, it still leads to a very low coverage probability of the confidence interval, because the SD of the estimated mean is also very small.
2. The bootstrap estimator of standard deviation performs well in all cases, even when the mean estimator is biased.
3.  $\bar{Y}_t$  based on censoring and linear regression imputation has negligible bias so that the related

confidence interval has a coverage probability close to the nominal level 95%; but it has a large SD when  $t = 3$  or  $t = 4$ . The inefficiency of this method is obviously caused by discarding observed data from nearly 50% of sampled subjects who have intermittent nonresponse. Its performance becomes worse as  $t$  increases.

4.  $\bar{Y}_t$  based on the proposed kernel regression imputation has a relative bias between 0.0% and 0.5%, but the bias is large enough to result in a poor coverage performance of the related confidence interval at  $t = 4$ .

5.  $\bar{Y}_t$  based on the proposed linear regression imputation has negligible bias as well as a variance smaller than that of  $\bar{Y}_t$  based on kernel regression. The related confidence interval has a coverage probability close to the nominal level 95%.
6.  $\bar{Y}_t$  based on the proposed one-dimensional index kernel regression imputation is generally good but slightly worse than that based on the linear regression imputation.
7.  $\bar{Y}_t$  based on methods (7)-(9) has non-negligible bias when  $t = 3$  or  $t = 4$ , which results in poor performance of the related confidence interval.

**Table 3**  
Simulation results for mean estimation (Normal population)

Method	Quantity	$t = 2$	$t = 3$	$t = 4$
Complete data	relative bias	0%	0%	0%
	SD	0.0221	0.0223	0.0221
	$\widehat{SD}_{boot}$	0.0223	0.0223	0.0224
	CI coverage	94.9%	94.4%	95.4%
Respondents only	relative bias	12.8%	6.8%	3.5%
	SD	0.0282	0.0272	0.0248
	$\widehat{SD}_{boot}$	0.0285	0.0267	0.0252
	CI coverage	0.0%	0.0%	0.2%
Censoring and linear regression imputation	relative bias	0.0%	0.0%	-0.1%
	SD	0.0275	0.0358	0.0418
	$\widehat{SD}_{boot}$	0.0276	0.0354	0.0431
	CI coverage	95.1%	94.6%	95.6%
Proposed kernel regression imputation	relative bias	0.0%	0.4%	0.5%
	SD	0.0275	0.0288	0.0283
	$\widehat{SD}_{boot}$	0.0276	0.0288	0.0288
	CI coverage	95.1%	92.5%	88.6%
Proposed linear regression imputation	relative bias	0.0%	0.1%	0.0%
	SD	0.0275	0.0286	0.0279
	$\widehat{SD}_{boot}$	0.0276	0.0287	0.0293
	CI coverage	95.1%	93.8%	95.7%
Proposed 1-dimensional index kernel regression imputation	relative bias	0.0%	0.4%	0.4%
	SD	0.0275	0.0288	0.0279
	$\widehat{SD}_{boot}$	0.0276	0.0288	0.0288
	CI coverage	95.1%	92.5%	91.7%
Last-value-dependent kernel regression imputation	relative bias	0.6%	1.0%	0.6%
	SD	0.0284	0.0310	0.0257
	$\widehat{SD}_{boot}$	0.0288	0.0295	0.0263
	CI coverage	93.7%	84.2%	86.2%
Linear regression imputation treating previously imputed values as observed	relative bias	0.0%	1.6%	0.8%
	SD	0.0275	0.0261	0.0241
	$\widehat{SD}_{boot}$	0.0276	0.0260	0.0246
	CI coverage	95.1%	59.7%	76.0%
Linear regression imputation based on currently and previously observed data	relative bias	0.0%	1.6%	0.8%
	SD	0.0275	0.0261	0.0242
	$\widehat{SD}_{boot}$	0.0276	0.0261	0.0246
	CI coverage	95.1%	59.0%	76.1%

Although the kernel regression is asymptotically valid, in this simulation study the total number of subjects is 2,000 and, according to Table 2, the average numbers of data points used in kernel regression under patterns  $(t, r) = (4, 1)$  and  $(4, 2)$  are 238 and 152, respectively, which may not be enough for kernel regression and lead to some small biases in imputation. On the other hand, linear regression is more stable and works well with a sample size such as 152. Although linear regression imputation has a bias in theory, the bias may be small when  $E(y_t | y_1, \dots, y_{t-1})$  is linear.

### 3.2 Application to the SIRD

The SIRD is an annual survey of about 31,000 companies potentially involved in research and development. The NSF sponsors this survey as part of a mandate requiring that NSF collect, interpret, and analyze data on scientific and engineering resources in the United States. The survey is conducted jointly by the U.S. Census Bureau and NSF. The surveyed companies are asked to provide information related to their total research and development (RD) expenditure for the calendar year of the survey. The SIRD deterministically surveys some companies each year by placing them in a certainty stratum, since they account for a large percentage of the total RD dollar investment in the U.S. The remaining companies that appear in the survey are sampled each year using a stratified probability proportionate to size (PPS) sampling design. Longitudinal measurements are available on the core of companies that are sampled with certainty and on other companies that happen to be selected each year. For the purposes of illustrating our imputation methods, we restrict attention to only those companies that were selected for the survey in each of the years 2002 through 2005 ( $T = 4$ ), and companies that provided a response in 2002. For documentation on the SIRD and detailed statistical tables, we refer to the document titled *Research and Development in Industry: 2005*, available from <http://www.nsf.gov/statistics/nsf10319>. Additional information on the Business R&D and Innovation Survey is available online at <http://bhs.dev.econ.census.gov/bhs/brdis/> and <http://www.nsf.gov/statistics/srvyindustry/about/brdis/>.

We divide the data into two imputation classes. One class consists of all companies contained in a certainty stratum for each of the four years; the other consists of the rest of companies. Within each imputation class, the data take the form  $(\mathbf{y}_i, \boldsymbol{\delta}_i)$ ,  $i = 1, \dots, n$ , where  $y_{it}$  represents the total RD expenditure for company  $i$  at time  $t = 1$  (2002), 2 (2003), 3 (2004), 4 (2005). The sample size here is  $n = 2,309$  for the certainty strata class and  $n = 1,039$  for the non-certainty strata class. Missingness is nonmonotone and the missing percentages for the years 2003, 2004, and 2005 were 10.4%, 14.0%, and 18.8%, for the certainty strata

class, and 15.2%, 20.7%, and 26.0% for the non-certainty strata class.

Table 4 shows the estimated totals and standard errors obtained by using the methods (2)-(9) described in the simulation study in Section 3.1. As discussed in the end of Section 2.1, in each of the proposed imputation methods we use linear regression when  $r + 1 = t$ . The standard errors shown in Table 4 were computed using the bootstrap method. Table 4 also displays estimated totals obtained when missing data are filled in by the values that were put in place by the Census Bureau in order to produce the officially published data tables (officially published data tables are available from [http://www.nsf.gov/statistics/pubseri.cfm?seri\\_id=26](http://www.nsf.gov/statistics/pubseri.cfm?seri_id=26)). The method that was used by the Census Bureau to handle missing data when producing these published data tables (which we call the "current method") was ratio imputation for companies with prior year data using imputation cells formed by industry type; we refer to Bond (1994) for further details. Table 4 also presents the estimated RD totals obtained from respondents only with no weight adjustment which indicate that ignoring the missing data leads to biased estimates. Methods (3)-(9) give comparable results, which is likely due to the strong linear dependence in the data so that theoretically biased methods exhibit negligible bias. The estimated totals based on the current method are comparable to those based on the proposed methods for the certainty strata case, but are different in the non-certainty strata case. The method of censoring and linear regression has similar SD to the proposed methods because the number of data points discarded under censoring is not too large. In the certainty strata imputation class only 10% of the sample has an intermittent nonresponse pattern and the percentage of complete cases is 72%. In the non-certainty class, only 9% of the sample has an intermittent nonresponse pattern and the percentage of complete cases is 66%.

### 3.3 Simulation results based on the SIRD population

An additional simulation study was conducted using a population constructed from the SIRD data. The simulation was run independently for the certainty strata and non-certainty strata imputation classes. To construct the population, we begin with the SIRD data with missing values imputed using the current imputation method for the SIRD. Let  $\boldsymbol{\delta}_i$  be the observed response indicator vector for company  $i$  and  $\tilde{\mathbf{y}}_i$  be the vector of either the observed or imputed values of total RD expenditures for company  $i$  over time,  $i = 1, \dots, n$ . For the simulation, we sample from a population based on  $\{(\tilde{\mathbf{y}}_i, \boldsymbol{\delta}_i), i = 1, \dots, n\}$  as follows. We first draw a sample of size  $n$  with replacement from  $\tilde{\mathbf{y}}_1, \dots, \tilde{\mathbf{y}}_n$ , then we add independent normal random noise, with mean 0 and standard deviation 500, to each component

of each of the sampled vectors. Any resulting negative values are set to zero. We denote these simulated RD totals by  $y_1^*, \dots, y_n^*$ , where  $n$  is the same as that in Section 3.2. We denote the simulated response indicators by  $\delta_1^*, \dots, \delta_n^*$ . For all  $i$  and each  $t = 2, 3, 4$ ,  $\delta_{it}^*$ 's were binary random variables with

$$P(\delta_{it}^* = 1 \mid y_{i1}^*, \dots, y_{i,t-1}^*) = \frac{\exp(\beta_0^{(t)} + \beta_1^{(t)}y_{i,1}^* + \dots + \beta_{t-1}^{(t)}y_{i,t-1}^*)}{1 + \exp(\beta_0^{(t)} + \beta_1^{(t)}y_{i,1}^* + \dots + \beta_{t-1}^{(t)}y_{i,t-1}^*)}$$

The coefficients  $\beta_0^{(t)}, \beta_1^{(t)}, \dots, \beta_{t-1}^{(t)}$  are fixed throughout the simulation and they were obtained as the estimated coefficients from an initial fit of a logistic regression of  $\delta_{it}$  on  $(\tilde{y}_{i1}, \dots, \tilde{y}_{i,t-1})$  for  $i = 1, \dots, n$ .

Table 5 reports the simulation results for total estimators based on 1,000 runs and methods (1)-(9) described in Section 3.1, where the quantities appearing in the table are defined in Section 3.1. To compute the relative bias we obtain the true value of the total through a preliminary run of the simulation model. Several of the conclusions from the normal population simulation of Section 3.1 carry over to this setting. The following is a summary of some additional findings.

1. In contrast to the normal population simulation setting, the estimated total based on censoring and linear regression has SD that is comparable with the proposed imputation methods. This is because the number of data points discarded under censoring is small in this case. The probabilities of an intermittent response pattern are 17% and 19% for the certainty and non-certainty strata classes, respectively. In the normal population simulation these probabilities were nearly 50% as shown in Table 2.
2. All of the proposed imputation methods give relatively similar performance. As noted previously, linear regression imputation is generally biased in theory. However, the bias is small because of the strong linear dependence in data.
3. Method (7) does not have a good performance at  $t \geq 3$  for the non-certainty strata case, because the last-value-dependent PSI assumption does not hold.
4. Methods (8) and (9) perform well, again due to the strong linear dependence in data. Although these methods use more observed data in regression imputation, they are comparable with the proposed linear regression method.

**Table 4**  
RD total estimates (in thousands) from SIRD data based on years 2002 to 2005.  
Bootstrap standard error (in thousands) in parentheses<sup>1</sup>

Method	Certainty strata			Non-certainty strata		
	t = 2	t = 3	t = 4	t = 2	t = 3	t = 4
Current imputation	154,066	156,754	168,015	2,694	2,790	2,782
	-	-	-	-	-	-
Respondents only with no weight adjustment	149,502 (15,907)	148,300 (16,160)	159,822 (17,149)	2,448 (172)	2,553 (193)	2,419 (207)
Respondents only with adjusted weights	166,924 (17,728)	172,419 (18,720)	196,815 (21,045)	2,887 (199)	3,219 (237)	3,269 (273)
Censoring and linear regression imputation	154,824 (15,888)	159,206 (16,394)	172,631 (17,470)	2,843 (189)	3,079 (208)	3,257 (246)
Proposed kernel regression imputation	154,824 (15,888)	159,394 (16,414)	171,633 (17,603)	2,843 (189)	2,997 (199)	3,161 (290)
Proposed linear regression imputation	154,824 (15,888)	159,198 (16,383)	172,042 (17,247)	2,843 (189)	3,043 (203)	3,302 (250)
Proposed 1-dimensional index kernel regression imputation	154,824 (15,888)	159,394 (16,414)	171,494 (17,268)	2,843 (189)	2,997 (199)	3,254 (248)
Last-value-dependent kernel regression imputation	154,688 (15,900)	158,768 (16,286)	170,606 (17,234)	2,831 (188)	2,983 (197)	3,177 (240)
Linear regression imputation treating previously imputed values as observed	154,824 (15,888)	159,401 (16,390)	172,600 (17,306)	2,843 (189)	3,098 (208)	3,257 (236)
Linear regression imputation based on currently and previously observed data	154,824 (15,888)	160,205 (16,534)	172,452 (17,209)	2,843 (189)	3,168 (233)	3,273 (254)

<sup>1</sup> Disclaimer: The values in Table 4 do not necessarily represent national estimates because we have made some restrictions on the data to fit our framework.

**Table 5**  
Simulation results for total estimation (in thousands) SIRD based population

Method	Quantity	Certainty Strata			Non-Certainty Strata		
		$t = 2$	$t = 3$	$t = 4$	$t = 2$	$t = 3$	$t = 4$
Complete data	relative bias	0%	0.1%	0.1%	0.2%	0.0%	0.4%
	SD	15,541	16,045	16,947	184	203	224
	$\widehat{SD}_{boot}$	15,654	15,994	16,941	186	201	218
	CI coverage	94.0%	94.0%	94.3%	94.3%	93.7%	93.9%
Respondents only with adjusted weights	relative bias	5%	6.3%	11.6%	-1.1%	1.1%	-2.7%
	SD	16,870	17,858	20,032	191	220	244
	$\widehat{SD}_{boot}$	16,917	17,915	20,048	192	219	234
	CI coverage	94.8%	94.8%	87.3%	93.2%	94.5%	89.8%
Censoring and linear regression imputation	relative bias	0%	0.4%	0.5%	0.4%	0.1%	-0.4%
	SD	15,582	16,272	17,247	191	214	238
	$\widehat{SD}_{boot}$	15,654	16,145	17,195	194	214	236
	CI coverage	93.8%	93.5%	94.2%	94.8%	94.0%	93.7%
Proposed kernel regression imputation	relative bias	0%	0.2%	-0.1%	0.4%	-0.3%	-0.3%
	SD	15,582	16,130	17,098	191	205	246
	$\widehat{SD}_{boot}$	15,654	16,072	17,231	194	204	262
	CI coverage	93.8%	93.5%	94.2%	94.8%	93.4%	93.7%
Proposed linear regression imputation	relative bias	0%	0.2%	0.0%	0.4%	0.0%	-0.5%
	SD	15,582	16,130	16,955	191	206	229
	$\widehat{SD}_{boot}$	15,654	16,072	16,964	194	206	224
	CI coverage	93.8%	93.5%	94.2%	94.8%	94.0%	93.7%
Proposed 1-dimensional index kernel regression imputation	relative bias	0%	0.2%	-0.1%	0.4%	-0.3%	-0.9%
	SD	15,582	16,130	16,957	191	205	227
	$\widehat{SD}_{boot}$	15,654	16,072	16,965	194	204	220
	CI coverage	93.8%	93.5%	94.3%	94.8%	93.4%	93.1%
Last-value-dependent kernel regression imputation	relative bias	0%	0.1%	-0.3%	0.0%	-0.7%	-0.7%
	SD	15,565	16,019	16,990	184	204	242
	$\widehat{SD}_{boot}$	15,635	16,003	16,983	187	202	230
	CI coverage	93.8%	93.7%	94.0%	93.9%	92.7%	91.1%
Linear regression imputation treating previously imputed values as observed	relative bias	0%	0.2%	0.0%	0.4%	0.6%	-0.6%
	SD	15,582	16,120	16,952	191	210	231
	$\widehat{SD}_{boot}$	15,654	16,065	16,954	194	210	225
	CI coverage	93.8%	93.6%	94.3%	94.8%	93.8%	92.8%
Linear regression imputation based on currently and previously observed data	relative bias	0%	0.2%	0.0%	0.4%	0.6%	-0.6%
	SD	15,582	16,117	16,945	191	213	241
	$\widehat{SD}_{boot}$	15,654	16,062	16,954	194	211	254
	CI coverage	93.8%	93.5%	94.3%	94.8%	93.6%	93.7%

#### 4. Concluding remarks

We consider a longitudinal study variable having non-monotone nonresponse. Under the assumption that the PSI depends on past observed or unobserved values of the study variable, we propose several imputation methods that lead to unbiased or nearly unbiased estimators of the total or mean of the study variable at a given time point. Our methods do

not require any parametric model on the joint distribution of the variables across time points or the PSI. They are based on regression models under different nonresponse patterns derived from the past-data-dependent PSI. Three regression methods are adopted, linear regression, kernel regression, and one-dimensional index kernel regression. The imputation method based on the kernel type regression is asymptotically valid, but it requires a large number of

observations in each nonresponse pattern. The imputation method based on linear regression is asymptotically biased when the linear relationship does not hold, but it is more stable and, therefore, it may still out-perform methods based on kernel regression.

The method of censoring, which discards all observed data from a subject after its first missing value, may work well when the number of data discarded is small; otherwise it may be very inefficient especially when  $T$  is large. For the SIRD data analysis in Sections 3.2-3.3, censoring is comparable with the proposed linear regression imputation method. However, the results are based on four years of data only and censoring may lead to inefficient estimators when more years of data are considered. In applications, it may be a good idea to compare estimators based on censoring with those based on the proposed methods.

Estimators based on the linear regression imputation methods (8) and (9) described in Section 3.1 are asymptotically biased in general. Although they perform well in the simulation study based on the SIRD population, they have poor performance under the simulation setting in Section 3.1, while the proposed linear regression imputation performs well.

The results in Section 2 can be extended to the situation where each sample unit has an observed covariate  $\mathbf{x}_t$  at time  $t$  without missing values. Assumption (1) may be modified to include covariates:

$$P(\delta_t = 1 \mid \mathbf{y}, \mathbf{X}, \delta_1, \dots, \delta_{t-1}, \delta_{t+1}, \dots, \delta_T) = P(\delta_t = 1 \mid y_1, \dots, y_{t-1}, \mathbf{X}, \delta_1, \dots, \delta_{t-1}), \quad t = 2, \dots, T,$$

where  $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_T)$ . Missing components of  $\mathbf{y}_i$  can be imputed using one of the procedures in Sections 2.1-2.3 with  $(y_{i1}, \dots, y_{ir})$  replaced by  $(y_{i1}, \dots, y_{ir}, \mathbf{X}_i)$ . After all missing values are imputed, we can also estimate the relationship between  $\mathbf{y}$  and  $\mathbf{X}$  using some popular approaches such as the generalized estimation equation approach. Some details can be found in Xu (2007).

It is implicitly assumed throughout the paper that the  $y$ -values are continuous variables with no restriction. When  $y$ -values have a particular order or are integer valued, the proposed regression imputation methods are clearly not suitable. New methods for these situations have to be developed.

### Acknowledgements

We thank Katherine Jenny Thompson and David L. Kinyon, both of the U.S. Census Bureau, as well as two referees and the associate editor for providing many helpful comments on the paper. The research was partially supported by an NSF grant. This article is released to inform interested parties of ongoing research and to encourage

discussion. The views expressed are those of the authors and not necessarily those of the U.S. Census Bureau.

### Appendix

*Proof of (2) - (3).* Let  $L(\xi)$  denote the distribution of  $\xi$  and  $L(\xi \mid \zeta)$  denote the conditional distribution of  $\xi$  given  $\zeta$ . Let  $\mathbf{y}_t = (y_1, \dots, y_t)$  and  $\boldsymbol{\delta}_t = (\delta_1, \dots, \delta_t)$ . Then, both (2) and (3) follow from  $L(y_t \mid \mathbf{y}_{t-1}, \boldsymbol{\delta}_t) = L(y_t \mid \mathbf{y}_{t-1}, \boldsymbol{\delta}_{t-1}) = L(\mathbf{y}_t, \boldsymbol{\delta}_{t-1}) / L(\mathbf{y}_{t-1}, \boldsymbol{\delta}_{t-1}) = [L(\delta_{t-1} \mid \mathbf{y}_t, \boldsymbol{\delta}_{t-2}) / L(\delta_{t-1} \mid \mathbf{y}_{t-1}, \boldsymbol{\delta}_{t-2})] L(y_t \mid \mathbf{y}_{t-1}, \boldsymbol{\delta}_{t-2}) = L(y_t \mid \mathbf{y}_{t-1}, \boldsymbol{\delta}_{t-2}) = L(y_t \mid \mathbf{y}_{t-1}, \boldsymbol{\delta}_{t-3}) = \dots = L(y_t \mid \mathbf{y}_{t-1})$ , where the first and third equalities follow from assumption (1).

*Proof of (5).* Using the same notation as in the proof of (2) and letting  $\Delta_r = 1$  be the indicator of  $\delta_1 = \dots = \delta_r = 1$ , we have  $L(y_t \mid \mathbf{y}_r, \Delta_r = 1, \delta_{r+1} = 0, \delta_t = 0) = [L(\delta_{r+1} = 0 \mid y_t, \mathbf{y}_r, \Delta_r = 1, \delta_t = 0) / L(\delta_{r+1} = 0 \mid \mathbf{y}_r, \Delta_r = 1, \delta_t = 0)] L(y_t \mid \mathbf{y}_r, \Delta_r = 1, \delta_t = 0)$ , which is equal to  $L(y_t \mid \mathbf{y}_r, \Delta_r = 1, \delta_t = 0)$  by (1). Similarly, we can show that  $L(y_t \mid \mathbf{y}_r, \Delta_r = 1, \delta_{r+1} = 1, \delta_t = 0) = L(y_t \mid \mathbf{y}_r, \Delta_r = 1, \delta_t = 0)$ . Hence,  $L(y_t \mid \mathbf{y}_r, \Delta_r = 1, \delta_{r+1} = 0, \delta_t = 0) = L(y_t \mid \mathbf{y}_r, \Delta_r = 1, \delta_{r+1} = 1, \delta_t = 0)$  and result (5) follows.

*An example in which (4) does not hold.* To show that (4) does not hold in general, we only need to give a counterexample. Consider  $T = 3$ . Let  $(y_1, y_2, y_3)$  be jointly normal with  $E(y_t) = 0$ ,  $\text{var}(y_t) = 1$ ,  $t = 1, 2, 3$ ,  $\text{cov}(y_1, y_2) = \text{cov}(y_1, y_3) = \rho$ , and  $\text{cov}(y_2, y_3) = \rho^2$ , where  $\rho \neq 0$  is a parameter. Suppose that  $y_1$  is always observed and  $P(\delta_t = 0 \mid y_{t-1}) = \Phi(a_{t-1} + b_{t-1} y_{t-1})$ ,  $t = 2, 3$ , where  $a_t$  and  $b_t$  are parameters,  $\Phi$  is the cumulative distribution function of the standard normal distribution. Then,  $E(y_3 \mid y_2, y_1) = \rho y_2$ ,  $E(y_2 \mid y_1) = \rho y_1$ , and  $E(y_3 \mid y_1) = \rho^2 y_1$ . Note that

$$\begin{aligned} E(y_3 \mid y_1, \delta_3 = 0, \delta_2 = \delta_1 = 1) &= E(y_3 \mid y_1, \delta_3 = 0, \delta_2 = 1) \\ &= E(y_3 \mid y_1, \delta_3 = 0) \\ &= \int y_3 L(y_3 \mid y_1, \delta_3 = 0) dy_3 \\ &= \int y_3 \int L(y_3 \mid y_1, y_2, \delta_3 = 0) L(y_2 \mid y_1, \delta_3 = 0) dy_2 dy_3 \\ &= \iint y_3 L(y_3 \mid y_1, y_2) L(y_2 \mid y_1, \delta_3 = 0) dy_2 dy_3 \\ &= \int \left( \int y_3 L(y_3 \mid y_2) dy_3 \right) L(y_2 \mid y_1, \delta_3 = 0) dy_2 \\ &= \rho \int y_2 L(y_2 \mid y_1, \delta_3 = 0) dy_2 \\ &= \frac{\rho \int y_2 P(\delta_3 = 0 \mid y_2) L(y_2 \mid y_1) dy_2}{\int P(\delta_3 = 0 \mid y_2) L(y_2 \mid y_1) dy_2} \\ &= \frac{\rho \int y_2 \Phi(a_2 + b_2 y_2) L(y_2 \mid y_1) dy_2}{\int \Phi(a_2 + b_2 y_2) L(y_2 \mid y_1) dy_2}, \end{aligned}$$

where the first equality holds because  $y_1$  is always observed, the second equality holds because under (1),  $\delta_2$  and  $y_3$  are independent given  $y_1$ . The denominator of the previous expression is equal to

$$h(y_1) = \Phi\left(\frac{a_2 + b_2 \rho y_1}{\sqrt{1 + b_2^2(1 - \rho^2)}}\right).$$

Using integration by parts, we obtain that

$$\begin{aligned} g(y_1) &= \int (y_2 - \rho y_1) \Phi(a_2 + b_2 y_2) L(y_2 | y_1) dy_2 \\ &= b_2(1 - \rho^2) \int \Phi'(a_2 + b_2 y_2) L(y_2 | y_1) dy_2 \\ &= \frac{b_2^2(1 - \rho^2)}{2\pi\sqrt{1 - \rho^2}} \int \exp\left\{-\frac{(a_2 + b_2 \rho y_2)^2}{2} - \frac{(y_2 - \rho y_1)^2}{2(1 - \rho^2)}\right\} dy_2 \\ &= \frac{b_2(1 - \rho^2)}{2\pi[1 + b_2^2(1 - \rho^2)]} \exp\left\{-\frac{(a_2 + b_2 \rho y_1)^2}{2[1 + b_2^2(1 - \rho^2)]}\right\}. \end{aligned}$$

Thus,

$$E(y_3 | y_1, \delta_3 = 0, \delta_2 = \delta_1 = 1) = \rho^2 y_1 + \rho \frac{g(y_1)}{h(y_1)}. \tag{10}$$

However,

$$\begin{aligned} E(y_3 | y_1, \delta_1 = \delta_2 = 1) &= E(y_3 | y_1, \delta_1 = 1) \\ &= E(y_3 | y_1) = \rho^2 y_1. \end{aligned}$$

This shows that (4) does not hold in this special case.

*Proof of (8).* Using the notation in the proof of (2)-(3) and writing the  $(t - 2)$ -dimensional vector  $(y_1, \dots, y_{r-1}, y_{r+1}, \dots, y_{t-1})$  as  $\mathbf{u}_{t,r}$ , we obtain that

$$\begin{aligned} L(\delta_{r+1} = 1 | y_t, z_r, \Delta_r = 1, \delta_t = 0) &= \int L(\delta_{r+1} = 1 | y_t, z_r, \mathbf{u}_{t,r}, \Delta_r = 1, \delta_t = 0) \\ &\quad L(\mathbf{u}_{t,r} | y_t, z_r, \Delta_r = 1, \delta_t = 0) d\mathbf{u}_{t,r} \\ &= \int L(\delta_{r+1} = 1 | y_1, \dots, y_r, \Delta_r = 1) \\ &\quad L(\mathbf{u}_{t,r} | y_t, z_r, \Delta_r = 1, \delta_t = 0) d\mathbf{u}_{t,r} \\ &= \int L(\delta_{r+1} = 1 | z_r, \Delta_r = 1) \\ &\quad L(\mathbf{u}_{t,r} | y_t, z_r, \Delta_r = 1, \delta_t = 0) d\mathbf{u}_{t,r} \\ &= L(\delta_{r+1} = 1 | z_r, \Delta_r = 1) \\ &\quad \int L(\mathbf{u}_{t,r} | y_t, z_r, \Delta_r = 1, \delta_t = 0) d\mathbf{u}_{t,r} \\ &= L(\delta_{r+1} = 1 | z_r, \Delta_r = 1), \end{aligned}$$

where the second equality follows from assumption (1) and the fact that there is a one-to-one function between  $(z_r, \mathbf{u}_{t,r})$  and  $(y_1, \dots, y_{t-1})$ , and the third equality follows from assumption (7). Similarly,  $L(\delta_{r+1} = 1 | z_r, \Delta_r = 1, \delta_t = 0) = L(\delta_{r+1} = 1 | z_r, \Delta_r = 1)$  and, hence,  $L(\delta_{r+1} = 1 | y_t, z_r, \Delta_r = 1, \delta_t = 0) = L(\delta_{r+1} = 1 | z_r, \Delta_r = 1, \delta_t = 0)$ . Then,

$$\begin{aligned} L(y_t | z_r, \Delta_{r+1} = 1, \delta_t = 0) &= \frac{L(y_t, z_r, \Delta_{r+1} = 1, \delta_t = 0)}{L(z_r, \Delta_{r+1} = 1, \delta_t = 0)} \\ &= \frac{L(\delta_{r+1} = 1 | y_t, z_r, \Delta_r = 1, \delta_t = 0) L(y_t, z_r, \Delta_r = 1, \delta_t = 0)}{L(\delta_{r+1} = 1 | z_r, \Delta_r = 1, \delta_t = 0) L(z_r, \Delta_r = 1, \delta_t = 0)} \\ &= L(y_t | z_r, \Delta_r = 1, \delta_t = 0). \end{aligned}$$

Similarly,  $L(y_t | z_r, \Delta_r = 1, \delta_{r+1} = 0, \delta_t = 0) = L(y_t | z_r, \Delta_r = 1, \delta_t = 0)$ . Hence,  $L(y_t | z_r, \Delta_r = 1, \delta_{r+1} = 0, \delta_t = 0) = L(y_t | z_r, \Delta_{r+1} = 1, \delta_t = 0)$  and result (8) follows.

### References

Bond, D. (1994). An evaluation of imputation methods for the Survey of Industrial Research and Development. *U.S. Bureau of the Census, Economic Statistical Methods and Programming Division Report Series*. 9404. Washington, DC.

Cheng, P.E. (1994). Nonparametric estimation of mean functionals with data missing at random. *Journal of the American Statistical Association*, 89, 81-87.

Diggle, P., and Kenward, M.G. (1994). Informative drop-out in longitudinal data analysis. *Journal of the Royal Statistical Society, Series C (Applied Statistics)*, 43, 49-93.

Duan, N., and Li, K. C. (1991). Sliced regression: A link-free regression method. *The Annals of Statistics*, 19, 505-530.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Annals of Statistics*, 7, 1-26.

Kalton, G., and Kasprzyk, D. (1986). The treatment of missing survey data. *Survey Methodology*, 12, 1, 1-16.

Little, R.J. (1995). Modeling the dropout mechanism in repeated-measures studies. *Journal of the American Statistical Association*, 90, 1112-1121.

Little, R.J., and Rubin, D.B. (2002). *Statistical Analysis with Missing Data*, second edition. New York: John Wiley & Sons, Inc.

National Science Foundation, Division of Science Resources Statistics (2010). *Research and Development in Industry: 2005. Detailed Statistical Tables*. Available from <http://www.nsf.gov/statistics/nsf10319/>.

Paik, M.C. (1997). The generalized estimating equation approach when data are not missing completely at random. *Journal of the American Statistical Association*, 92, 1320-1329.



- R Development Core Team (2009). A language and environment for statistical computing. *R Foundation for Statistical Computing*, Vienna, Austria, ISBN 3-900051-07-0.
- Robins, J.M., and Rotnitzky, A. (1995). Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association*, 90, 122-129.
- Shao, J., and Sitter, R.R. (1996). Bootstrap for imputed survey data. *Journal of the American Statistical Association*, 91, 1278-1288.
- Troxel, A.B., Harrington, D.P. and Lipsitz, S.R. (1998). Analysis of longitudinal data with non-ignorable non-monotone missing values. *Applied Statistics*, 47, 425-438.
- Troxel, A.B., Lipsitz, S.R. and Harrington, D.P. (1998). Marginal models for the analysis of longitudinal measurements with non-ignorable non-monotone missing data. *Biometrika*, 85, 661-672.
- Vansteelandt, S., Rotnitzky, A. and Robins, J.M. (2007). Estimation of regression models for the mean of repeated outcomes under nonignorable nonmonotone nonresponse. *Biometrika*, 94, 841-860.
- Xu, J. (2007). Methods for intermittent missing responses in longitudinal data. Ph.D. Thesis, Department of Statistics, University of Wisconsin-Madison.
- Xu, J., Shao, J., Palta, M. and Wang, L. (2008). Imputation for nonmonotone last-value-dependent nonrespondents in longitudinal surveys. *Survey Methodology*, 34, 2, 153-162.

**ELECTRONIC  
PUBLICATIONS  
AVAILABLE AT**

**PUBLICATIONS  
ÉLECTRONIQUES  
DISPONIBLE À**

**[www.statcan.gc.ca](http://www.statcan.gc.ca)**

# Some theory for propensity-score-adjustment estimators in survey sampling

Jae Kwang Kim and Minsun Kim Riddles<sup>1</sup>

## Abstract

The propensity-scoring-adjustment approach is commonly used to handle selection bias in survey sampling applications, including unit nonresponse and undercoverage. The propensity score is computed using auxiliary variables observed throughout the sample. We discuss some asymptotic properties of propensity-score-adjusted estimators and derive optimal estimators based on a regression model for the finite population. An optimal propensity-score-adjusted estimator can be implemented using an augmented propensity model. Variance estimation is discussed and the results from two simulation studies are presented.

Key Words: Calibration; Missing data; Nonresponse; Weighting.

## 1. Introduction

Consider a finite population of size  $N$ , where  $N$  is known. For each unit  $i$ ,  $y_i$  is the study variable and  $\mathbf{x}_i$  is the  $q$ -dimensional vector of auxiliary variables. The parameter of interest is the finite population mean of the study variable,  $\theta = N^{-1} \sum_{i=1}^N y_i$ . The finite population  $\mathcal{F}_N = \{(\mathbf{x}'_1, y_1), (\mathbf{x}'_2, y_2), \dots, (\mathbf{x}'_N, y_N)\}$  is assumed to be a random sample of size  $N$  from a superpopulation distribution  $F(\mathbf{x}, y)$ . Suppose a sample of size  $n$  is drawn from the finite population according to a probability sampling design. Let  $w_i = \pi_i^{-1}$  be the design weight, where  $\pi_i$  is the first-order inclusion probability of unit  $i$  obtained from the probability sampling design. Under complete response, the finite population mean can be estimated by the Horvitz-Thompson (HT) estimator,  $\hat{\theta}_{HT} = N^{-1} \sum_{i \in A} w_i y_i$ , where  $A$  is the set of indices appearing in the sample.

In the presence of missing data, the HT estimator  $\hat{\theta}_{HT}$  cannot be computed. Let  $r$  be the response indicator variable that takes the value one if  $y$  is observed and takes the value zero otherwise. Conceptually, as discussed by Fay (1992), Shao and Steel (1999), and Kim and Rao (2009), the response indicator can be extended to the entire population as  $\mathcal{R}_N = \{r_1, r_2, \dots, r_N\}$ , where  $r_i$  is a realization of the random variable  $r$ . In this case, the complete-case (CC) estimator  $\hat{\theta}_{CC} = \sum_{i \in A} w_i r_i y_i / \sum_{i \in A} w_i r_i$  converges in probability to  $E(Y | r = 1)$ . Unless the response mechanism is missing completely at random in the sense that  $E(Y | r = 1) = E(Y)$ , the CC estimator is biased. To correct for the bias of the CC estimator, if the response probability

$$p(\mathbf{x}, y) = \Pr(r = 1 | \mathbf{x}, y) \quad (1)$$

is known, then the weighted CC estimator  $\hat{\theta}_{WCC} = N^{-1} \sum_{i \in A} w_i r_i y_i / p(\mathbf{x}_i, y_i)$  can be used to estimate  $\theta$ . Note that  $\hat{\theta}_{WCC}$  is unbiased because  $E\{\sum_{i \in A} w_i r_i y_i / p(\mathbf{x}_i, y_i) | \mathcal{F}_N\} = E\{\sum_{i=1}^N r_i y_i / p(\mathbf{x}_i, y_i) | \mathcal{F}_N\} = \sum_{i=1}^N y_i$ .

If the response probability (1) is unknown, one can postulate a parametric model for the response probability  $p(\mathbf{x}, y; \phi)$  indexed by  $\phi \in \Omega$  such that  $p(\mathbf{x}, y) = p(\mathbf{x}, y; \phi_0)$  for some  $\phi_0 \in \Omega$ . We assume that there exists a  $\sqrt{n}$ -consistent estimator  $\hat{\phi}$  of  $\phi_0$  such that

$$\sqrt{n}(\hat{\phi} - \phi_0) = O_p(1), \quad (2)$$

where  $g_n = O_p(1)$  indicates  $g_n$  is bounded in probability. Using  $\hat{\phi}$ , we can obtain the estimated response probability by  $\hat{p}_i = p(\mathbf{x}_i, y_i; \hat{\phi})$ , which is often called the propensity score (Rosenbaum and Rubin 1983). The propensity-score-adjusted (PSA) estimator can be constructed as

$$\hat{\theta}_{PSA} = \frac{1}{N} \sum_{i \in A} w_i \frac{r_i}{\hat{p}_i} y_i. \quad (3)$$

The PSA estimator (3) is widely used. Many surveys use the PSA estimator to reduce nonresponse bias (Fuller, Loughin and Baker 1994; Rizzo, Kalton and Brick 1996). Rosenbaum and Rubin (1983) and Rosenbaum (1987) proposed using the PSA approach to estimate the treatment effects in observational studies. Little (1988) reviewed the PSA methods for handling unit nonresponse in survey sampling. Duncan and Stasny (2001) used the PSA approach to control coverage bias in telephone surveys. Folsom (1991) and Iannacchione, Milne and Folsom (1991) used a logistic regression model for the response probability estimation. Lee (2006) applied the PSA method to a volunteer panel web survey. Durrant and Skinner (2006) used the PSA approach to address measurement error.

Despite the popularity of PSA estimators, asymptotic properties of PSA estimators have not received much attention in survey sampling literature. Kim and Kim (2007) used a Taylor expansion to obtain the asymptotic mean and variance of PSA estimators and discussed variance estimation. Da Silva and Opsomer (2006) and Da Silva and

1. Jae Kwang Kim and Minsun Kim Riddles, Department of Statistics, Iowa State University, Ames, IA, U.S.A. 50011. E-mail: jkim@iastate.edu.

Opsomer (2009) considered nonparametric methods to obtain PSA estimators.

In this paper, we discuss optimal PSA estimators in the class of PSA estimators of the form (3) that use a  $\sqrt{n}$ -consistent estimator  $\hat{\phi}$ . Such estimators are asymptotically unbiased for  $\theta$ . Finding minimum variance PSA estimators among this particular class of PSA estimators is a topic of major interest in this paper.

Section 2 presents the main results. An optimal PSA estimator using an augmented propensity score model is proposed in Section 3. In Section 4, variance estimation of the proposed estimator is discussed. Results from two simulation studies can be found in Section 5 and concluding remarks are made in Section 6.

## 2. Main results

In this section, we discuss some asymptotic properties of PSA estimators. We assume that the response mechanism does not depend on  $y$ . Thus, we assume that

$$\Pr(r = 1 | \mathbf{x}, y) = \Pr(r = 1 | \mathbf{x}) = p(\mathbf{x}; \phi_0) \quad (4)$$

for some unknown vector  $\phi_0$ . The first equality implies that the data are missing-at-random (MAR), as we always observe  $\mathbf{x}$  in the sample. Note that the MAR condition is assumed in the population model. In the second equality, we further assume that the response mechanism is known up to an unknown parameter  $\phi_0$ . The response mechanism is slightly different from that of Kim and Kim (2007), where the response mechanism is assumed to be under the classical two-phase sampling setup and depends on the realized sample:

$$\Pr(r = 1 | \mathbf{x}, y, I = 1) = \Pr(r = 1 | \mathbf{x}, I = 1) = p(\mathbf{x}; \phi_0^I). \quad (5)$$

Here,  $I$  is the sampling indicator function defined throughout the population. That is,  $I_i = 1$  if  $i \in A$  and  $I_i = 0$  otherwise. Unless the sampling design is non-informative in the sense that the sample selection probabilities are correlated with the response indicator even after conditioning on auxiliary variables (Pfeffermann, Krieger and Rinott 1998), the two response mechanisms, (4) and (5), are different. In survey sampling, assumption (4) is more appropriate because an individual's decision on whether or not to respond to a survey is at his or her own discretion. Here, the response indicator variable  $r_i$  is defined throughout the population, as discussed in Section 1.

We consider a class of  $\sqrt{n}$ -consistent estimators of  $\phi_0$  in (4). In particular, we consider a class of estimators which can be written as a solution to

$$\hat{\mathbf{U}}_h(\phi) \equiv \sum_{i \in A} w_i \{r_i - p_i(\phi)\} \mathbf{h}_i(\phi) = \mathbf{0}, \quad (6)$$

where  $p_i(\phi) = p(\mathbf{x}_i; \phi)$  for some function  $\mathbf{h}_i(\phi) = \mathbf{h}(\mathbf{x}_i; \phi)$ , a smooth function of  $\mathbf{x}_i$  and parameter  $\phi$ . Thus, the solution to (6) can be written as  $\hat{\phi}_h$ , which depends on the choice of  $\mathbf{h}_i(\phi)$ . Any solution  $\hat{\phi}_h$  to (6) is consistent for  $\phi_0$  in (4) because  $E\{\hat{\mathbf{U}}_h(\phi_0) | \mathcal{F}_N\} = E[\sum_{i=1}^N \{r_i - p_i(\phi_0)\} \mathbf{h}_i(\phi_0) | \mathcal{F}_N]$  is zero under the response mechanism in (4). If we drop the sampling weights  $w_i$  in (6), the estimated parameter  $\hat{\phi}_h$  is consistent for  $\phi_0^A$  in (5) and the resulting PSA estimator is consistent only when the sampling design is non-informative. The PSA estimators obtained from (6) using the sampling weights are consistent regardless of whether the sampling design is non-informative or not. According to Chamberlain (1987), any  $\sqrt{n}$ -consistent estimator of  $\phi_0$  in (4) can be written as a solution to (6). Thus, the choice of  $\mathbf{h}_i(\phi)$  in (6) determines the efficiency of the resulting PSA estimator.

Let  $\hat{\theta}_{\text{PSA},h}$  be the PSA estimator in (3) using  $\hat{p}_i = p_i(\hat{\phi}_h)$  with  $\hat{\phi}_h$  being the solution to (6). To discuss the asymptotic properties of  $\hat{\theta}_{\text{PSA},h}$ , assume a sequence of finite populations and samples, as in Isaki and Fuller (1982), such that  $\sum_{i \in A} w_i \mathbf{u}_i - \sum_{i=1}^N \mathbf{u}_i = O_p(n^{-1/2}N)$  for any population characteristics  $\mathbf{u}_i$  with bounded fourth moments. We also assume that the sampling weights are uniformly bounded. That is,  $K_1 < N^{-1}nw_i < K_2$  for all  $i$  uniformly in  $n$ , where  $K_1$  and  $K_2$  are fixed constants. In addition, we assume the following regularity conditions:

- [C1] The response mechanism satisfies (4), where  $p(\mathbf{x}; \phi)$  is continuous in  $\phi$  with continuous first and second derivatives in an open set containing  $\phi_0$ . The responses are independent in the sense that  $\text{Cov}(r_i, r_j | \mathbf{x}) = 0$  for  $i \neq j$ . Also,  $p(\mathbf{x}_i; \phi) > c$  for all  $i$  for some fixed constant  $c > 0$ .
- [C2] The solution to (6) exists and is unique almost everywhere. The function  $\mathbf{h}_i(\phi) = \mathbf{h}(\mathbf{x}_i; \phi)$  in (6) has a bounded fourth moment. Furthermore, the partial derivative  $\partial\{\hat{\mathbf{U}}_h(\phi)\}/\partial\phi$  is nonsingular for all  $n$ .
- [C3] The estimating function  $\hat{\mathbf{U}}_h(\phi)$  in (6) converges in probability to  $\mathbf{U}_h(\phi) = \sum_{i=1}^N \{r_i - p_i(\phi)\} \mathbf{h}_i(\phi)$  uniformly in  $\phi$ . Furthermore, the partial derivative  $\partial\{\hat{\mathbf{U}}_h(\phi)\}/\partial\phi$  converges in probability to  $\partial\{\mathbf{U}_h(\phi)\}/\partial\phi$  uniformly in  $\phi$ . The solution  $\phi_N$  to  $\mathbf{U}_h(\phi) = \mathbf{0}$  satisfies  $N^{1/2}(\phi_N - \phi_0) = O_p(1)$  under the response mechanism.

Condition [C1] states the regularity conditions for the response mechanism. Condition [C2] is the regularity condition for the solution  $\hat{\phi}_h$  to (6). In Condition [C3], some regularity conditions are imposed on the estimating function  $\hat{\mathbf{U}}_h(\phi)$  itself. By [C2] and [C3], we can establish the  $\sqrt{n}$ -consistency (2) of  $\hat{\phi}_h$ .

Now, the following theorem deals with some asymptotic properties of the PSA estimator  $\hat{\theta}_{\text{PSA},h}$ .

*Theorem 1* If conditions [C1] - [C3] hold, then under the joint distribution of the sampling mechanism and the response mechanism, the PSA estimator  $\hat{\theta}_{\text{PSA},h}$  satisfies

$$\sqrt{n}(\hat{\theta}_{\text{PSA},h} - \tilde{\theta}_{\text{PSA},h}) = o_p(1), \tag{7}$$

where

$$\tilde{\theta}_{\text{PSA},h} = \frac{1}{N} \sum_{i \in A} w_i \left\{ p_i \mathbf{h}'_i \gamma_h^* + \frac{r_i}{p_i} (y_i - p_i \mathbf{h}'_i \gamma_h^*) \right\}, \tag{8}$$

$\gamma_h^* = (\sum_{i=1}^N r_i \mathbf{z}_i p_i \mathbf{h}'_i)^{-1} (\sum_{i=1}^N r_i \mathbf{z}_i y_i)$ ,  $p_i = p(\mathbf{x}_i; \boldsymbol{\phi}_0)$ ,  $\mathbf{z}_i = \partial\{p^{-1}(\mathbf{x}_i; \boldsymbol{\phi}_0)\} / \partial\boldsymbol{\phi}$ , and  $\mathbf{h}_i = \mathbf{h}(\mathbf{x}_i; \boldsymbol{\phi}_0)$ . Moreover, if the finite population is a random sample from a superpopulation model, then

$$V(\tilde{\theta}_{\text{PSA},h}) \geq V_l \equiv V(\hat{\theta}_{\text{HT}}) + \frac{1}{N^2} E \left\{ \sum_{i \in A} w_i^2 \left( \frac{1}{p_i} - 1 \right) V(Y | \mathbf{x}_i) \right\}. \tag{9}$$

The equality in (9) holds when  $\hat{\boldsymbol{\phi}}_h$  satisfies

$$\sum_{i \in A} w_i \left\{ \frac{r_i}{p(\mathbf{x}_i; \hat{\boldsymbol{\phi}}_h)} - 1 \right\} E(Y | \mathbf{x}_i) = 0, \tag{10}$$

where  $E(Y | \mathbf{x}_i)$  is the conditional expectation under the superpopulation model.

*Proof.* Given  $p_i(\boldsymbol{\phi}) = p(\mathbf{x}_i; \boldsymbol{\phi})$  and  $\mathbf{h}_i(\boldsymbol{\phi}) = \mathbf{h}(\mathbf{x}_i; \boldsymbol{\phi})$ , define

$$\hat{\theta}(\boldsymbol{\phi}, \gamma) = N^{-1} \sum_{i \in A} w_i \left[ p_i(\boldsymbol{\phi}) \mathbf{h}'_i(\boldsymbol{\phi}) \gamma + \frac{r_i}{p_i(\boldsymbol{\phi})} \{y_i - p_i(\boldsymbol{\phi}) \mathbf{h}'_i(\boldsymbol{\phi}) \gamma\} \right].$$

Since  $\hat{\boldsymbol{\phi}}_h$  satisfies (6), we have  $\hat{\theta}_{\text{PSA}} = \hat{\theta}(\hat{\boldsymbol{\phi}}_h, \gamma)$  for any choice of  $\gamma$ . We now want to find a particular choice of  $\gamma$ , say  $\gamma^*$ , such that

$$\hat{\theta}(\hat{\boldsymbol{\phi}}_h, \gamma^*) = \hat{\theta}(\boldsymbol{\phi}_0, \gamma^*) + o_p(n^{-1/2}). \tag{11}$$

As  $\hat{\boldsymbol{\phi}}_h$  converges in probability to  $\boldsymbol{\phi}_0$ , the asymptotic equivalence (11) holds if

$$E \left\{ \frac{\partial}{\partial \boldsymbol{\phi}} \hat{\theta}(\boldsymbol{\phi}, \gamma^*) \mid \boldsymbol{\phi} = \boldsymbol{\phi}_0 \right\} = \mathbf{0}, \tag{12}$$

using the theory of Randles (1982). Condition (12) holds if  $\gamma^* = \gamma_h^*$ , where  $\gamma_h^*$  is defined in (8). Thus, (11) reduces to

$$\hat{\theta}_{\text{PSA},h} = \frac{1}{N} \sum_{i \in A} w_i \left\{ p_i \mathbf{h}'_i \gamma_h^* + \frac{r_i}{p_i} (y_i - p_i \mathbf{h}'_i \gamma_h^*) \right\} + o_p(n^{-1/2}), \tag{13}$$

which proves (7). The variance of  $\tilde{\theta}_{\text{PSA},h}$  can be derived as

$$\begin{aligned} V(\tilde{\theta}_{\text{PSA},h}) &= V(\hat{\theta}_{\text{HT}}) + \frac{1}{N^2} E \left\{ \sum_{i \in A} w_i^2 \left( \frac{1}{p_i} - 1 \right) (y_i - p_i \mathbf{h}'_i \gamma_h^*)^2 \right\} \\ &= V(\hat{\theta}_{\text{HT}}) + \frac{1}{N^2} E \left[ \sum_{i \in A} w_i^2 \left( \frac{1}{p_i} - 1 \right) \left\{ y_i - E(Y | \mathbf{x}_i) \right. \right. \\ &\quad \left. \left. + E(Y | \mathbf{x}_i) - p_i \mathbf{h}'_i \gamma_h^* \right\}^2 \right] \\ &= V(\hat{\theta}_{\text{HT}}) + \frac{1}{N^2} E \left\{ \sum_{i \in A} w_i^2 \left( \frac{1}{p_i} - 1 \right) V(Y | \mathbf{x}_i) \right\} \\ &\quad + \frac{1}{N^2} E \left[ \sum_{i \in A} w_i^2 \left( \frac{1}{p_i} - 1 \right) \left\{ E(Y | \mathbf{x}_i) - p_i \mathbf{h}'_i \gamma_h^* \right\}^2 \right], \tag{14} \end{aligned}$$

where the last equality follows because  $y_i$  is conditionally independent of  $E(Y | \mathbf{x}_i) - p_i \mathbf{h}'_i \gamma_h^*$ , conditioning on  $\mathbf{x}_i$ . Since the last term in (14) is non-negative, the inequality in (9) is established. Furthermore, if  $E(Y | \mathbf{x}_i) = p_i \mathbf{h}'_i \boldsymbol{\alpha}$  for some  $\boldsymbol{\alpha}$ , then (10) holds and  $E(\gamma_h^* | \mathbf{x}_i) = \boldsymbol{\alpha}$ , by the definition of  $\gamma_h^*$ . Thus,  $E(Y | \mathbf{x}_i) - p_i \mathbf{h}'_i \gamma_h^* = -p_i \mathbf{h}'_i \{ \gamma_h^* - E(\gamma_h^* | \mathbf{x}_i) \} = o_p(1)$ , implying that the last term in (14) is negligible.

In (9),  $V_l$  is the lower bound of the asymptotic variance of PSA estimators of the form (3) satisfying (6). Any PSA estimator that has the asymptotic variance  $V_l$  in (9) is optimal in the sense that it achieves the lower bound of the asymptotic variance among the class of PSA estimators with  $\hat{\boldsymbol{\phi}}$  satisfying (2). The asymptotic variance of optimal PSA estimators of  $\theta$  is equal to  $V_l$  in (9). The PSA estimator using the maximum likelihood estimator of  $\boldsymbol{\phi}_0$  does not necessarily achieve the lower bound of the asymptotic variance.

Condition (10) provides a way of constructing an optimal PSA estimator. First, we need an assumption for  $E(Y | \mathbf{x})$ , which is often called the outcome regression model. If the outcome regression model is a linear regression model of the form  $E(Y | \mathbf{x}) = \beta_0 + \boldsymbol{\beta}'_1 \mathbf{x}$ , an optimal PSA estimator of  $\theta$  can be obtained by solving

$$\sum_{i \in A} w_i \frac{r_i}{p_i(\boldsymbol{\phi})} (1, \mathbf{x}_i) = \sum_{i \in A} w_i (1, \mathbf{x}_i). \tag{15}$$

Condition (15) is appealing because it says that the PSA estimator applied to  $y = a + \mathbf{b}'\mathbf{x}$  leads to the original HT estimator. Condition (15) is called the calibration condition in survey sampling. The calibration condition applied to  $\mathbf{x}$  makes full use of the information contained in it if the study variable is well approximated by a linear function of  $\mathbf{x}$ . Condition (15) was also used in Nevo (2003) and Kott (2006) under the linear regression model.

If we explicitly use a regression model for  $E(Y | \mathbf{x})$ , it is possible to construct an estimator that has asymptotic variance (9) and is not necessarily a PSA estimator. For example, if we assume that

$$E(Y | \mathbf{x}) = m(\mathbf{x}; \boldsymbol{\beta}_0) \tag{16}$$

for some function  $m(\mathbf{x}; \cdot)$  known up to  $\boldsymbol{\beta}_0$ , we can use the model (16) directly to construct an optimal estimator of the form

$$\hat{\theta}_{\text{opt}} = \frac{1}{N} \sum_{i \in A} w_i \left[ m(\mathbf{x}_i; \hat{\boldsymbol{\beta}}) + \frac{r_i}{p_i(\hat{\boldsymbol{\phi}})} \{y_i - m(\mathbf{x}_i; \hat{\boldsymbol{\beta}})\} \right], \tag{17}$$

where  $\hat{\boldsymbol{\beta}}$  is a  $\sqrt{n}$ -consistent estimator of  $\boldsymbol{\beta}_0$  in the superpopulation model (16) and  $\hat{\boldsymbol{\phi}}$  is a  $\sqrt{n}$ -consistent estimator of  $\boldsymbol{\phi}_0$  computed by (6). The following theorem shows that the optimal estimator (17) achieves the lower bound in (9).

*Theorem 2* Let the conditions of Theorem 1 hold. Assume that  $\hat{\boldsymbol{\beta}}$  satisfies  $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}_0 + O_p(n^{-1/2})$ . Assume that, in the superpopulation model (16),  $m(\mathbf{x}; \boldsymbol{\beta})$  has continuous first-order partial derivatives in an open set containing  $\boldsymbol{\beta}_0$ . Under the joint distribution of the sampling mechanism, the response mechanism, and the superpopulation model (16), the estimator  $\hat{\theta}_{\text{opt}}$  in (17) satisfies

$$\sqrt{n} (\hat{\theta}_{\text{opt}} - \tilde{\theta}_{\text{opt}}^*) = o_p(1),$$

where

$$\tilde{\theta}_{\text{opt}}^* = N^{-1} \sum_{i \in A} w_i \left[ m(\mathbf{x}_i; \boldsymbol{\beta}_0) + \frac{r_i}{p_i} \{y_i - m(\mathbf{x}_i; \boldsymbol{\beta}_0)\} \right],$$

$p_i = p_i(\boldsymbol{\phi}_0)$ , and  $V(\tilde{\theta}_{\text{opt}}^*)$  is equal to  $V_l$  in (9).

*Proof.* Define  $\hat{\theta}_{\text{opt}}(\boldsymbol{\beta}, \boldsymbol{\phi}) = N^{-1} \sum_{i \in A} w_i [m(\mathbf{x}_i; \boldsymbol{\beta}) + r_i p_i^{-1}(\boldsymbol{\phi}) \{y_i - m(\mathbf{x}_i; \boldsymbol{\beta})\}]$ . Note that  $\hat{\theta}_{\text{opt}}$  in (17) can be written as  $\hat{\theta}_{\text{opt}} = \hat{\theta}_{\text{opt}}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\phi}})$ . Since

$$\frac{\partial}{\partial \boldsymbol{\beta}} \hat{\theta}_{\text{opt}}(\boldsymbol{\beta}, \boldsymbol{\phi}) = \frac{1}{N} \sum_{i \in A} w_i \left\{ \bar{m}(\mathbf{x}_i; \boldsymbol{\beta}) - \frac{r_i}{p_i(\boldsymbol{\phi})} \bar{m}(\mathbf{x}_i; \boldsymbol{\beta}) \right\},$$

where  $\bar{m}(\mathbf{x}_i; \boldsymbol{\beta}) = \partial m(\mathbf{x}_i; \boldsymbol{\beta}) / \partial \boldsymbol{\beta}$ , and

$$\frac{\partial}{\partial \boldsymbol{\phi}} \hat{\theta}_{\text{opt}}(\boldsymbol{\beta}, \boldsymbol{\phi}) = \frac{1}{N} \sum_{i \in A} w_i r_i \mathbf{z}_i(\boldsymbol{\phi}) \{y_i - m(\mathbf{x}_i; \boldsymbol{\beta})\},$$

where  $\mathbf{z}_i(\boldsymbol{\phi}) = \partial \{p_i^{-1}(\boldsymbol{\phi})\} / \partial \boldsymbol{\phi}$ , we have  $E[\partial \{\hat{\theta}_{\text{opt}}(\boldsymbol{\beta}, \boldsymbol{\phi})\} / \partial(\boldsymbol{\beta}, \boldsymbol{\phi}) | \boldsymbol{\beta} = \boldsymbol{\beta}_0, \boldsymbol{\phi} = \boldsymbol{\phi}_0] = \mathbf{0}$  and the condition of Randles (1982) is satisfied. Thus,

$$\hat{\theta}_{\text{opt}}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\phi}}) = \hat{\theta}_{\text{opt}}(\boldsymbol{\beta}_0, \boldsymbol{\phi}_0) + o_p(n^{-1/2}) = \tilde{\theta}_{\text{opt}}^* + o_p(n^{-1/2})$$

and the variance of  $\tilde{\theta}_{\text{opt}}^*$  is equal to  $V_l$ , the lower bound of the asymptotic variance.

The (asymptotic) optimality of the estimator in (17) is justified under the joint distribution of the response model (4) and the superpopulation model (16). When both models are correct,  $\hat{\theta}_{\text{opt}}$  is optimal and the choice of  $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\phi}})$  does not affect the efficiency of the  $\hat{\theta}_{\text{opt}}$  as long as  $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\phi}})$  is  $\sqrt{n}$ -consistent. Robins, Rotnitzky and Zhao (1994) also advocated using  $\hat{\theta}_{\text{opt}}$  in (17) under simple random sampling.

*Remark 1* When the response model is correct and the superpopulation model (16) is not necessarily correct, the choice of  $\hat{\boldsymbol{\beta}}$  does affect the efficiency of the optimal estimator. Cao, Tsiatis and Davidian (2009) considered optimal estimation when only the response model is correct. Using Taylor linearization, the optimal estimator in (17) with  $\hat{\boldsymbol{\phi}}$  satisfying (6) is asymptotically equivalent to

$$\tilde{\theta}(\boldsymbol{\beta}) = \sum_{i \in A} w_i \left[ m(\mathbf{x}_i; \boldsymbol{\beta}) + \frac{r_i}{p_i} \{y_i - m(\mathbf{x}_i; \boldsymbol{\beta})\} - \left( \frac{r_i}{p_i} - 1 \right) \mathbf{c}'_{\boldsymbol{\beta}} p_i \mathbf{h}_i \right],$$

where  $\mathbf{c}_{\boldsymbol{\beta}}$  is the probability limit of  $\hat{\mathbf{c}}_{\boldsymbol{\beta}} = \{\sum_{i \in A} w_i r_i \mathbf{z}_i(\hat{\boldsymbol{\phi}}) \hat{p}_i \mathbf{h}'_i(\hat{\boldsymbol{\phi}})\}^{-1} \sum_{i \in A} w_i r_i \mathbf{z}_i(\hat{\boldsymbol{\phi}}) \{y_i - m(\mathbf{x}_i; \boldsymbol{\beta})\}$  and  $\mathbf{z}_i(\boldsymbol{\phi}) = \partial \{p_i^{-1}(\boldsymbol{\phi})\} / \partial \boldsymbol{\phi}$ . The asymptotic variance is then equal to

$$V\{\tilde{\theta}(\boldsymbol{\beta})\} = V(\hat{\theta}_{\text{HT}}) + E \left[ \sum_{i \in A} w_i^2 \frac{1 - p_i}{p_i} \{y_i - m(\mathbf{x}_i; \boldsymbol{\beta}) - \mathbf{c}'_{\boldsymbol{\beta}} p_i \mathbf{h}_i\}^2 \right].$$

Thus, an optimal estimator of  $\boldsymbol{\beta}$  can be computed by finding  $\hat{\boldsymbol{\beta}}$  that minimizes

$$Q(\boldsymbol{\beta}) = \sum_{i \in A} w_i^2 r_i \frac{1 - \hat{p}_i}{\hat{p}_i^2} \{y_i - m(\mathbf{x}_i; \boldsymbol{\beta}) - \hat{\mathbf{c}}'_{\boldsymbol{\beta}} \hat{p}_i \mathbf{h}_i(\hat{\boldsymbol{\phi}})\}^2.$$

The resulting estimator is design-optimal in the sense that it minimizes the asymptotic variance under the response model.

### 3. Augmented propensity score model

In this section, we consider optimal PSA estimation. Note that the optimal estimator  $\hat{\theta}_{\text{opt}}$  in (17) is not necessarily written as a PSA estimator form in (3). It is in the PSA estimator form if it satisfies  $\sum_{i \in A} w_i r_i \hat{p}_i^{-1} m(\mathbf{x}_i; \hat{\boldsymbol{\beta}}) = \sum_{i \in A} w_i m(\mathbf{x}_i; \hat{\boldsymbol{\beta}})$ . Thus, we can construct an optimal PSA estimator by including  $m(\mathbf{x}_i; \hat{\boldsymbol{\beta}})$  in the model for the propensity score. Specifically, given  $\hat{m}_i = m(\mathbf{x}_i; \hat{\boldsymbol{\beta}})$ ,  $\hat{p}_i = p_i(\hat{\boldsymbol{\phi}})$  and  $\hat{\mathbf{h}}_i = \mathbf{h}_i(\hat{\boldsymbol{\phi}})$ , where  $\hat{\boldsymbol{\phi}}$  is obtained from (6), we augment the response model by

$$p_i^*(\hat{\boldsymbol{\phi}}, \boldsymbol{\lambda}) \equiv \frac{\hat{p}_i}{\hat{p}_i + (1 - \hat{p}_i) \exp(\lambda_0 + \lambda_1 \hat{m}_i)}, \tag{18}$$

where  $\lambda = (\lambda_0, \lambda_1)'$  is the Lagrange multiplier which is used to incorporate the additional constraint. If  $(\lambda_0, \lambda_1)' = \mathbf{0}$ , then  $p_i^*(\hat{\phi}, \lambda) = \hat{p}_i$ . The augmented response probability  $p_i^*(\hat{\phi}, \lambda)$  always takes values between 0 and 1. The augmented response probability model (18) can be derived by minimizing the Kullback-Leibler distance  $\sum_{i \in A} w_i r_i q_i^* \log(q_i^*/q_i)$ , where  $q_i^* = (1 - p_i^*)/p_i^*$  and  $q_i = (1 - \hat{p}_i)/\hat{p}_i$ , subject to the constraint  $\sum_{i \in A} w_i (r_i/p_i^*)(1, \hat{m}_i) = \sum_{i \in A} w_i (1, \hat{m}_i)$ .

Using (18), the optimal PSA estimator is computed by

$$\hat{\theta}_{\text{PSA}}^* = \frac{1}{N} \sum_{i \in A} w_i \frac{r_i}{p_i^*(\hat{\phi}, \hat{\lambda})} y_i, \quad (19)$$

where  $\hat{\lambda}$  satisfies

$$\sum_{i \in A} w_i \frac{r_i}{p_i^*(\hat{\phi}, \hat{\lambda})} (1, \hat{m}_i) = \sum_{i \in A} w_i (1, \hat{m}_i). \quad (20)$$

Under the response model (4), it can be shown that

$$\hat{\theta}_{\text{PSA}}^* = \frac{1}{N} \sum_{i \in A} w_i \left\{ \hat{b}_0 + \hat{b}_1 \hat{m}_i + \frac{r_i}{\hat{p}_i} (y_i - \hat{b}_0 - \hat{b}_1 \hat{m}_i) \right\} + o_p(n^{-1/2}),$$

where

$$\begin{pmatrix} \hat{b}_0 \\ \hat{b}_1 \end{pmatrix} = \left\{ \sum_{i \in A} w_i r_i \begin{pmatrix} 1 \\ \hat{p}_i \end{pmatrix} - 1 \right\}^{-1} \begin{pmatrix} 1 \\ \hat{m}_i \end{pmatrix} \begin{pmatrix} 1 \\ \hat{m}_i \end{pmatrix} \right\}^{-1} \sum_{i \in A} w_i r_i \begin{pmatrix} 1 \\ \hat{p}_i \end{pmatrix} - 1 \begin{pmatrix} 1 \\ \hat{m}_i \end{pmatrix} y_i. \quad (21)$$

Furthermore, by the argument for Theorem 1, we can establish that

$$\begin{aligned} \hat{\theta}_{\text{PSA}}^* = \frac{1}{N} \sum_{i \in A} w_i & \left\{ b_0 + b_1 \hat{m}_i + \gamma'_{h2} p_i \mathbf{h}_i \right. \\ & \left. + \frac{r_i}{p_i} (y_i - b_0 - b_1 \hat{m}_i - \gamma'_{h2} p_i \mathbf{h}_i) \right\} \\ & + o_p(n^{-1/2}), \end{aligned}$$

where  $(b_0, b_1, \gamma'_{h2})$  is the probability limit of  $(\hat{b}_0, \hat{b}_1, \hat{\gamma}'_{h2})$  with

$$\begin{aligned} \hat{\gamma}'_{h2} = & \left\{ \sum_{i \in A} w_i r_i \mathbf{z}_i(\hat{\phi}) \hat{p}_i \mathbf{h}'_i(\hat{\phi}) \right\}^{-1} \\ & \sum_{i \in A} w_i r_i \mathbf{z}_i(\hat{\phi}) (y_i - \hat{b}_0 - \hat{b}_1 \hat{m}_i) \end{aligned} \quad (22)$$

and the effect of estimating  $\phi_0$  in  $\hat{p}_i = p(\mathbf{x}_i; \hat{\phi})$  can be safely ignored.

Note that, under the response model (4),  $(\hat{\phi}, \hat{\lambda})$  in (19) converges in probability to  $(\phi_0, \mathbf{0})$ , where  $\phi_0$  is the true parameter in (4). Thus, the propensity score from the augmented model converges to the true response probability.

Because  $\hat{\lambda}$  converges to zero in probability, the choice of  $\hat{\beta}$  in  $\hat{m}_i = m(\mathbf{x}_i; \hat{\beta})$  does not play a role for the asymptotic unbiasedness of the PSA estimator. The asymptotic variances are changed for different choices of  $\hat{\beta}$ .

Under the superpopulation model (16),  $\hat{b}_0 + \hat{b}_1 \hat{m}_i \rightarrow E(Y | \mathbf{x}_i)$  in probability. Thus, the optimal PSA estimator in (19) is asymptotically equivalent to the optimal estimator in (17). Incorporating  $\hat{m}_i$  into the calibration equation to achieve optimality is close in spirit to the model-calibration method proposed by Wu and Sitter (2001).

#### 4. Variance estimation

We now discuss variance estimation of PSA estimators under the assumed response model. Singh and Folsom (2000) and Kott (2006) discussed variance estimation for certain types of PSA estimators. Kim and Kim (2007) discussed variance estimation when the PSA estimator is computed with the maximum likelihood method.

We consider variance estimation for the PSA estimator of the form (3) where  $\hat{p}_i = p_i(\hat{\phi})$  is constructed to satisfy (6) for some  $\mathbf{h}_i(\phi) = \mathbf{h}(\mathbf{x}_i; \phi, \beta)$ , where  $\beta$  is obtained using the postulated superpopulation model. Let  $\beta^*$  be the probability limit of  $\hat{\beta}$  under the response model. Note that  $\beta^*$  is not necessarily equal to  $\beta_0$  in (16) since we are not assuming that the postulated superpopulation model is correctly specified in this section.

Using the argument for the Taylor linearization (13) used in the proof of Theorem 1, the PSA estimator satisfies

$$\hat{\theta}_{\text{PSA}} = \frac{1}{N} \sum_{i \in A} w_i \eta_i(\phi_0, \beta^*) + o_p(n^{-1/2}), \quad (23)$$

where

$$\begin{aligned} \eta_i(\phi, \beta) = & p_i(\phi) \mathbf{h}'_i(\phi, \beta) \gamma_h^* \\ & + \frac{r_i}{p_i(\phi)} \{ y_i - p_i(\phi) \mathbf{h}'_i(\phi, \beta) \gamma_h^* \}, \end{aligned} \quad (24)$$

$\mathbf{h}_i(\phi, \beta) = \mathbf{h}(\mathbf{x}_i; \phi, \beta)$  and  $\gamma_h^*$  is defined as in (8) with  $\mathbf{h}_i$  replaced by  $\mathbf{h}_i(\phi_0, \beta^*)$ . Since  $p_i(\hat{\phi})$  satisfies (6) with  $\mathbf{h}_i(\phi) = \mathbf{h}(\mathbf{x}_i; \phi, \beta)$ ,  $\hat{\theta}_{\text{PSA}} = N^{-1} \sum_{i \in A} w_i \eta_i(\hat{\phi}, \hat{\beta})$  holds and the linearization in (23) can be expressed as  $N^{-1} \sum_{i \in A} w_i \eta_i(\hat{\phi}, \hat{\beta}) = N^{-1} \sum_{i \in A} w_i \eta_i(\phi_0, \beta^*) + o_p(n^{-1/2})$ . Thus, if  $(\mathbf{x}_i, y_i, r_i)$  are independent and identically distributed (IID), then  $\eta_i(\phi_0, \beta^*)$  are IID even though  $\eta_i(\hat{\phi}, \hat{\beta})$  are not necessarily IID. Because  $\eta_i(\phi_0, \beta^*)$  are IID, we can apply the standard complete sample method to estimate the variance of  $\hat{\eta}_{\text{HT}} = N^{-1} \sum_{i \in A} w_i \eta_i(\phi_0, \beta^*)$ , which is asymptotically equivalent to the variance of  $\hat{\theta}_{\text{PSA}} = N^{-1} \sum_{i \in A} w_i \eta_i(\hat{\phi}, \hat{\beta})$ . See Kim and Rao (2009).

To derive the variance estimator, we assume that the variance estimator  $\hat{V} = N^{-2} \sum_{i \in A} \sum_{j \in A} \Omega_{ij} g_i g_j$  satisfies

$\hat{V}/V(\hat{g}_{HT}|\mathcal{F}_N) = 1 + o_p(1)$  for some  $\Omega_{ij}$  related to the joint inclusion probability, where  $\hat{g}_{HT} = N^{-1}\sum_{i \in A} w_i g_i$  for any  $g$  with a finite second moment and  $V(g_{HT}|\mathcal{F}_N) = N^{-2}\sum_{i=1}^N \sum_{j=1}^N \Omega_{N-ij} g_i g_j$ , for some  $\Omega_{N-ij}$ . We also assume that

$$\sum_{i=1}^N |\Omega_{N-ij}| = O(n^{-1}N). \tag{25}$$

To obtain the total variance, the *reverse framework* of Fay (1992), Shao and Steel (1999), and Kim and Rao (2009) is considered. In this framework, the finite population is first divided into two groups, a population of respondents and a population of nonrespondents. Given the population, the sample  $A$  is selected according to a probability sampling design. Thus, selection of the population respondents from the whole finite population is treated as the first-phase sampling and the selection of the sample respondents from the population respondents is treated as the second-phase sampling in the reverse framework. The total variance of  $\hat{\eta}_{HT}$  can be written as

$$V(\hat{\eta}_{HT}|\mathcal{F}_N) = V_1 + V_2 = E\{V(\hat{\eta}_{HT}|\mathcal{F}_N, \mathcal{R}_N) | \mathcal{F}_N\} + V\{E(\hat{\eta}_{HT} | \mathcal{F}_N, \mathcal{R}_N) | \mathcal{F}_N\}. \tag{26}$$

The conditional variance term  $V(\hat{\eta}_{HT} | \mathcal{F}_N, \mathcal{R}_N)$  in (26) can be estimated by

$$\hat{V}_1 = N^{-2} \sum_{i \in A} \sum_{j \in A} \Omega_{ij} \hat{\eta}_i \hat{\eta}_j, \tag{27}$$

where  $\hat{\eta}_i = \eta_i(\hat{\phi}, \hat{\beta})$  is defined in (24) with  $\gamma_h^*$  replaced by a consistent estimator such as  $\hat{\gamma}_h^* = \{\sum_{i \in A} w_i r_i \mathbf{z}_i(\hat{\phi}) \hat{p}_i \hat{\mathbf{h}}_i'\}^{-1} \sum_{i \in A} w_i r_i \mathbf{z}_i(\hat{\phi}) y_i$ , and  $\hat{\mathbf{h}}_i = \mathbf{h}(\mathbf{x}_i; \hat{\phi}, \hat{\beta})$ . To show that  $\hat{V}_1$  is also consistent for  $V_1$  in (26), it suffices to show that  $V\{n \cdot V(\hat{\eta}_{HT} | \mathcal{F}_N, \mathcal{R}_N) | \mathcal{F}_N\} = o(1)$ , which follows by (25) and the existence of the fourth moment. See Kim, Navarro and Fuller (2006). The second term  $V_2$  in (26) is

$$V\{E(\hat{\eta}_{HT} | \mathcal{F}_N, \mathcal{R}_N) | \mathcal{F}_N\} = V\left(N^{-1} \sum_{i=1}^N \eta_i | \mathcal{F}_N\right) = \frac{1}{N^2} \sum_{i=1}^N \frac{1-p_i}{p_i} (y_i - p_i \mathbf{h}_i^* \gamma_h^*)^2,$$

where  $\mathbf{h}_i^* = \mathbf{h}(\mathbf{x}_i; \phi_0, \beta^*)$ . A consistent estimator of  $V_2$  can be derived as

$$\hat{V}_2 = \frac{1}{N^2} \sum_{i \in A} w_i r_i \frac{1-\hat{p}_i}{\hat{p}_i^2} (y_i - \hat{p}_i \hat{\mathbf{h}}_i' \hat{\gamma}_h^*)^2, \tag{28}$$

where  $\hat{\gamma}_h^*$  is defined after (27). Therefore,

$$\hat{V}(\hat{\theta}_{PSA}) = \hat{V}_1 + \hat{V}_2, \tag{29}$$

is consistent for the variance of the PSA estimator defined in (3) with  $\hat{p}_i = p_i(\hat{\phi})$  satisfying (6), where  $\hat{V}_1$  is in (27) and  $\hat{V}_2$  is in (28).

Note that the first term of the total variance is  $V_1 = O_p(n^{-1})$ , but the second term is  $V_2 = O_p(N^{-1})$ . Thus, when the sampling fraction  $nN^{-1}$  is negligible, that is,  $nN^{-1} = o(1)$ , the second term  $V_2$  can be ignored and  $\hat{V}_1$  is a consistent estimator of the total variance. Otherwise, the second term  $V_2$  should be taken into consideration, so that a consistent variance estimator can be constructed as in (29).

*Remark 2* The variance estimation of the optimal PSA estimator with augmented propensity model (18) with  $(\hat{\phi}, \hat{\lambda})$  satisfying (20) can be derived by (29) using  $\hat{\eta}_i = \hat{b}_0 + \hat{b}_1 \hat{m}_i + \hat{\gamma}'_{h2} \hat{p}_i \hat{\mathbf{h}}_i + r_i \hat{p}_i^{-1} (y_i - \hat{b}_0 - \hat{b}_1 \hat{m}_i - \hat{\gamma}'_{h2} \hat{p}_i \hat{\mathbf{h}}_i)$  where  $(\hat{b}_0, \hat{b}_1)$  and  $\hat{\gamma}_{h2}$  are defined in (21) and (22), respectively.

## 5. Simulation study

### 5.1 Study one

Two simulation studies were performed to investigate the properties of the proposed method. In the first simulation, we generated a finite population of size  $N = 10,000$  from the following multivariate normal distribution:

$$\begin{pmatrix} x_1 \\ x_2 \\ e \end{pmatrix} \sim N \left[ \begin{pmatrix} 2 \\ -1 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0.5 & 0 \\ 0.5 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \right].$$

The variable of interest  $y$  was constructed as  $y = 1 + x_1 + e$ . We also generated response indicator variables  $r_i$  independently from a Bernoulli distribution with probability

$$p_i = \frac{\exp(2 + x_{2i})}{1 + \exp(2 + x_{2i})}.$$

From the finite population, we used simple random sampling to select two samples of size,  $n = 100$  and  $n = 400$ , respectively. We used  $B = 5,000$  Monte Carlo samples in the simulation. The average response rate was about 69.6%.

To compute the propensity score, a response model of the form

$$p(\mathbf{x}; \phi) = \frac{\exp(\phi_0 + \phi_1 x_2)}{1 + \exp(\phi_0 + \phi_1 x_2)} \tag{30}$$

was postulated and an outcome regression model of the form

$$m(\mathbf{x}; \beta) = \beta_0 + \beta_1 x_1 \tag{31}$$

was postulated to obtain the optimal PSA estimators. Thus, both models are correctly specified. From each sample, we computed four estimators of  $\theta = N^{-1} \sum_{i=1}^N y_i$ :



1. (PSA-MLE): PSA estimator in (3) with  $\hat{p}_i = p_i(\hat{\phi})$  and  $\hat{\phi}$  being the maximum likelihood estimator of  $\phi$ .
2. (PSA-CAL): PSA estimator in (3) with  $\hat{p}_i$  satisfying the calibration constraint (15) on  $(1, x_{2i})$ .
3. (AUG): Augmented PSA estimator in (19).
4. (OPT): Optimal estimator in (17).

In the augmented PSA estimators,  $\hat{\phi}$  was computed by the maximum likelihood method. Under model (30), the maximum likelihood estimator of  $\phi = (\phi_0, \phi_1)'$  was computed by solving (6) with  $\mathbf{h}_i(\phi) = (1, x_{2i})'$ . Parameter  $(\beta_0, \beta_1)$  for the outcome regression model was computed using ordinary least squares, regressing  $y$  on  $x_i$ . In addition to the point estimators, we also computed the variance estimators of the point estimators. The variance estimators of the PSA estimators were computed using the pseudo-values in (24) and the  $\mathbf{h}_i(\phi)$  corresponding to each estimator. For the augmented PSA estimators, the pseudo-values were computed by the method in Remark 2.

Table 1 presents the Monte Carlo biases, variances, and mean square errors of the four point estimators and the Monte Carlo percent relative biases and  $t$ -statistics of the variance estimators of the estimators. The percent relative bias of a variance estimator  $\hat{V}(\hat{\theta})$  is calculated as  $100 \times \{V_{MC}(\hat{\theta})\}^{-1} [E_{MC}\{\hat{V}(\hat{\theta})\} - V_{MC}(\hat{\theta})]$ , where  $E_{MC}(\cdot)$  and  $V_{MC}(\cdot)$  denote the Monte Carlo expectation and the Monte Carlo variance, respectively. The  $t$ -statistic in Table 1 is the test statistic for testing the zero bias of the variance estimator. See Kim (2004).

Based on the simulation results in Table 1, we have the following conclusions.

1. All of the PSA estimators are asymptotically unbiased because the response model (30) is correctly specified. The PSA estimator using the calibration method is slightly more efficient than the PSA estimator using the maximum likelihood estimator, because the last term of (14) is smaller for the calibration method as the predictor for  $E(Y | \mathbf{x}_i) = \beta_0 + \beta_1 x_{1i}$  is better approximated by a linear function of  $(1, x_{2i})$  than by a linear function of  $(\hat{p}_i, \hat{p}_i x_{2i})$ .

2. The augmented PSA estimator is more efficient than the direct PSA estimator (3). The augmented PSA estimator is constructed by using the correctly specified regression model (31) and so it is asymptotically equivalent to the optimal PSA estimator in (17).
3. Variance estimators are all approximately unbiased. There are some modest biases in the variance estimators of the PSA estimators when the sample size is small ( $n = 100$ ).

### 5.2 Study two

In the second simulation study, we further investigated the PSA estimators with a non-linear outcome regression model under an unequal probability sampling design. We generated two stratified finite populations of  $(x, y)$  with four strata ( $h = 1, 2, 3, 4$ ), where  $x_{hi}$  were independently generated from a normal distribution  $N(1, 1)$  and  $y_{hi}$  were dichotomous variables that take values of 1 or 0 from a Bernoulli distribution with probability  $p_{1yhi}$  or  $p_{2yhi}$ . Two different probabilities were used for two populations, respectively:

1. Population 1 (Pop1):

$$p_{1yhi} = 1 / \{1 + \exp(0.5 - 2x)\}.$$

2. Population 2 (Pop2):

$$p_{2yhi} = 1 / [1 + \exp\{0.25(x - 1.5)^2 - 1.5\}].$$

In addition to  $x_{hi}$  and  $y_{hi}$ , the response indicator variables  $r_{hi}$  were generated from a Bernoulli distribution with probability  $p_{hi} = 1 / \{1 + \exp(-1.5 + 0.7x_{hi})\}$ . The sizes of the four strata were  $N_1 = 1,000$ ,  $N_2 = 2,000$ ,  $N_3 = 3,000$ , and  $N_4 = 4,000$ , respectively. In each of the two sets of finite population, a stratified sample of size  $n = 400$  was independently generated without replacement, where a simple random sample of size  $n_h = 100$  was selected from each stratum. We used  $B = 5,000$  Monte Carlo samples in this simulation. The average response rate was about 67%.

**Table 1**  
**Monte Carlo bias, variance and mean square error(MSE) of the four point estimators and percent relative biases (R.B.) and  $t$ -statistics( $t$ -stat) of the variance estimators based on 5,000 Monte Carlo samples**

n	Method	$\hat{\theta}$			$V(\hat{\theta})$	
		Bias	Variance	MSE	R.B. (%)	$t$ -stat
100	(PSA-MLE)	-0.01	0.0315	0.0317	-2.34	-1.12
	(PSA-CAL)	-0.01	0.0308	0.0309	-3.56	-1.70
	(AUG)	0.00	0.0252	0.0252	-0.61	-0.30
	(OPT)	0.00	0.0252	0.0252	-0.21	-0.10
400	(PSA-MLE)	-0.01	0.00737	0.00746	0.35	0.17
	(PSA-CAL)	-0.01	0.00724	0.00728	0.29	0.14
	(AUG)	0.00	0.00612	0.00612	0.07	0.03
	(OPT)	0.00	0.00612	0.00612	-0.14	-0.07

To compute the propensity score, a response model of the form

$$p(x; \phi) = \frac{\exp(\phi_0 + \phi_1 x)}{1 + \exp(\phi_0 + \phi_1 x)}$$

was postulated for parameter estimation. To obtain the augmented PSA estimator, a model for the variable of interest of the form

$$m(x; \beta) = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)} \tag{32}$$

was postulated. Thus, model (32) is a true model under (Pop1), but it is not a true model under (Pop2).

We computed four estimators:

1. (PSA-MLE): PSA estimator in (3) using the maximum likelihood estimator of  $\phi$ .
2. (PSA-CAL): PSA estimator in (3) with  $\hat{p}_i$  satisfying the calibration constraint (15) on  $(1, x)$ .
3. (AUG-1): Augmented PSA estimator  $\hat{\theta}_{PSA}^*$  in (19) with  $\hat{\beta}$  computed by the maximum likelihood method.
4. (AUG-2): Augmented PSA estimator  $\hat{\theta}_{PSA}^*$  in (19) with  $\hat{\beta}$  computed by the method of Cao *et al.* (2009) discussed in Remark 1.

We considered the augmented PSA estimator in (19) with  $\hat{p}_i = p_i(\hat{\phi})$ , where  $\hat{\phi}$  is the maximum likelihood estimator of  $\phi$ . The first augmented PSA estimator (AUG-1) used  $\hat{m}_i = m(x_i; \hat{\beta})$  with  $\hat{\beta}$  found by solving  $\sum_{h=1}^4 \sum_{i \in A_h} w_{hi} r_{hi} \{y_{hi} - m(x_{hi}; \beta)\} (1, x_{hi}) = \mathbf{0}$ , where  $A_h$  is the set of indices appearing in the sample for stratum  $h$  and  $w_{hi}$  is the sampling weight of unit  $i$  for stratum  $h$ .

Table 2 presents the simulation results for each method. In each population, the augmented PSA estimator shows some improvement comparing to the PSA estimator using the maximum likelihood estimator of  $\phi$  or the calibration estimator of  $\phi$  in terms of variance. Under (Pop1), since model (32) is true, there is essentially no difference between

the augmented PSA estimators using different methods of estimating  $\beta$ . However, under (Pop2), where the assumed outcome regression model (32) is incorrect, the augmented PSA estimator with  $\hat{\beta}$  computed by the method of Cao *et al.* (2009) results in slightly better efficiency, which is consistent with the theory in Remark 1. Variance estimates are approximately unbiased in all cases in the simulation study.

## 6. Conclusion

We have considered the problem of estimating the finite population mean of  $y$  under nonresponse using the propensity score method. The propensity score is computed from a parametric model for the response probability, and so the asymptotic properties of PSA estimators are discussed. In particular, the optimal PSA estimator is derived with an additional assumption for the distribution of  $y$ . The propensity score for the optimal PSA estimator can be implemented by the augmented propensity model presented in Section 3. The resulting estimator is still consistent even when the assumed outcome regression model fails to hold.

We have restricted our attention to missing-at-random mechanisms in which the response probability depends only on the always-observed  $x$ . If the response mechanism also depends on  $y$ , PSA estimation becomes more challenging. PSA estimation when missingness is not at random is beyond the scope of this article and will be a topic of future research.

## Acknowledgements

The research was partially supported by a Cooperative Agreement between the US Department of Agriculture Natural Resources Conservation Service and Iowa State University. The authors wish to thank F. Jay Breidt, three anonymous referees, and the associate editor for their helpful comments.

**Table 2**  
**Monte Carlo bias, variance and mean square error of the four point estimators and percent relative biases (R.B.) and  $t$ -statistics of the variance estimators, based on 5,000 Monte Carlo samples**

Population	Method	$\hat{\theta}_{PSA}$			$V(\hat{\theta}_{PSA})$	
		Bias	Variance	MSE	R.B. (%)	$t$ -stat
Pop1	(PSA-MLE)	0.00	0.000750	0.000762	-1.13	-0.57
	(PSA-CAL)	0.00	0.000762	0.000769	-1.45	-0.72
	(AUG-1)	0.00	0.000745	0.000757	-1.73	-0.86
	(AUG-2)	0.00	0.000745	0.000757	-1.83	-0.91
Pop2	(PSA-MLE)	0.00	0.000824	0.000826	0.29	0.14
	(PSA-CAL)	0.00	0.000829	0.000835	-0.94	-0.46
	(AUG-1)	0.00	0.000822	0.000823	-0.71	-0.35
	(AUG-2)	0.00	0.000820	0.000821	-0.61	-0.30

## References

- Cao, W., Tsiatis, A.A. and Davidian, M. (2009). Improving efficiency and robustness of the doubly robust estimator for a population mean with incomplete data. *Biometrika*, 96, 723-734.
- Chamberlain, G. (1987). Asymptotic efficiency in estimation with conditional moment restrictions. *Journal of Econometrics*, 34, 305-334.
- Da Silva, D.N., and Opsomer, J.D. (2006). A kernel smoothing method of adjusting for unit non-response in sample surveys. *Canadian Journal of Statistics*, 34, 563-579.
- Da Silva, D.N., and Opsomer, J.D. (2009). Nonparametric propensity weighting for survey nonresponse through local polynomial regression. *Survey Methodology*, 35, 2, 165-176.
- Duncan, K.B., and Stasny, E.A. (2001). Using propensity scores to control coverage bias in telephone surveys. *Survey Methodology*, 27, 2, 121-130.
- Durrant, G.B., and Skinner, C. (2006). Using missing data methods to correct for measurement error in a distribution function. *Survey Methodology*, 32, 1, 25-36.
- Fay, R.E. (1992). When are inferences from multiple imputation valid? *Proceedings of the Survey Research Methods Section*, American Statistical Association, 227-232.
- Folsom, R.E. (1991). Exponential and logistic weight adjustments for sampling and nonresponse error reduction. *Proceedings of the Social Statistics Section*, American Statistical Association, 197-202.
- Fuller, W.A., Loughin, 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, 1, 75-85.
- Iannacchione, V.G., Milne, J.G. and Folsom, R.E. (1991). Response probability weight adjustments using logistic regression. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 637-642.
- Isaki, C.T., and Fuller, W.A. (1982). Survey design under the regression superpopulation model. *Journal of the American Statistical Association*, 77, 89-96.
- Kim, J.K. (2004). Finite sample properties of multiple imputation estimators. *The Annals of Statistics*, 32, 766-783.
- Kim, J.K., and Kim, J.J. (2007). Nonresponse weighting adjustment using estimated response probability. *Canadian Journal of Statistics*, 35, 501-514.
- Kim, J.K., Navarro, A. and Fuller, W.A. (2006). Replication variance estimation for two-phase stratified sampling. *Journal of the American Statistical Association*, 101, 312-320.
- Kim, J.K., and Rao, J.N.K. (2009). A unified approach to linearization variance estimation from survey data after imputation for item nonresponse. *Biometrika*, 96, 917-932.
- Kott, P.S. (2006). Using calibration weighting to adjust for nonresponse and coverage errors. *Survey Methodology*, 32, 2, 133-142.
- Lee, S. (2006). Propensity score adjustment as a weighting scheme for volunteer panel web surveys. *Journal of Official Statistics*, 22, 329-349.
- Little, R.J.A. (1988). Missing-data adjustments in large surveys. *Journal of Business and Economic Statistics*, 6, 287-296.
- Nevo, A. (2003). Using weights to adjust for sample selection when auxiliary information is available. *Journal of Business and Economic Statistics*, 21, 43-52.
- Pfeffermann, D., Krieger, A.M. and Rinott, Y. (1998). Parametric distributions of complex survey data under informative probability sampling. *Statistica Sinica*, 8, 1087-1114.
- Randles, R.H. (1982). On the asymptotic normality of statistics with estimated parameters. *The Annals of Statistics*, 10, 462-474.
- Rizzo, L., Kalton, G. and Brick, J.M. (1996). A comparison of some weighting adjustment methods for panel nonresponse. *Survey Methodology*, 22, 1, 43-53.
- Robins, J.M., Rotnitzky, A. and Zhao, L.P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89, 846-866.
- Rosenbaum, P.R. (1987). Model-based direct adjustment. *Journal of the American Statistical Association*, 82, 387-394.
- Rosenbaum, P.R., and Rubin, D.B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70, 41-55.
- Shao, J., and Steel, P. (1999). Variance estimation for survey data with composite imputation and nonnegligible sampling fractions. *Journal of the American Statistical Association*, 94, 254-265.
- Singh, A.C., and Folsom, R.E. (2000). Bias corrected estimating function approach for variance estimation adjusted for poststratification. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 610-615.
- Wu, C., and Sitter, R.R. (2001). A model-calibration approach to using complete auxiliary information from survey data. *Journal of the American Statistical Association*, 96, 185-193.

**ELECTRONIC  
PUBLICATIONS  
AVAILABLE AT**

**PUBLICATIONS  
ÉLECTRONIQUES  
DISPONIBLE À**

**[www.statcan.gc.ca](http://www.statcan.gc.ca)**

# Assessing the accuracy of response propensity models in longitudinal studies

Ian Plewis, Sosthenes Ketende and Lisa Calderwood<sup>1</sup>

## Abstract

Non-response in longitudinal studies is addressed by assessing the accuracy of response propensity models constructed to discriminate between and predict different types of non-response. Particular attention is paid to summary measures derived from receiver operating characteristic (ROC) curves and logit rank plots. The ideas are applied to data from the UK Millennium Cohort Study. The results suggest that the ability to discriminate between and predict non-respondents is not high. Weights generated from the response propensity models lead to only small adjustments in employment transitions. Conclusions are drawn in terms of the potential of interventions to prevent non-response.

Key Words: Longitudinal studies; Missing data; Weighting; Propensity scores; ROC curves; Millennium Cohort Study.

## 1. Introduction

Examples of studies that have modelled the predictors of different kinds of, and different reasons for the non-response that affect longitudinal studies are plentiful, stimulated by being able to draw on auxiliary variables obtained from sample members before (and after) the occasions at which they are non-respondents. See, for example, Lepkowski and Couper (2002) for an analysis that separates refusals from not being located or contacted; Hawkes and Plewis (2006) who separate wave non-respondents from attrition cases in the UK National Child Development Study; and Plewis (2007a) and Plewis, Ketende, Joshi and Hughes (2008) who consider non-response in the first two waves of the UK Millennium Cohort Study. The focus of this paper is on how we can assess the accuracy of these response propensity models (Little and Rubin 2002). The paper is built around a framework that is widely used in epidemiology (Pepe 2003) and criminology (Copas 1999) to evaluate risk scores but has not, to our knowledge, been used in survey research before. Response propensity models can be used to construct weights intended to remove biases from estimates, to inform imputations, and to predict potential non-respondents at future waves thereby directing fieldwork resources to those respondents who might otherwise be lost. The accuracy of response propensity models has not, however, been given the amount of attention it warrants in terms of their ability to discriminate between respondents and non-respondents, and to predict future non-response. Good estimates of accuracy can be used to compare the efficacy of different weighting methods, and to help to determine the allocation of scarce fieldwork resources in order to reduce non-response.

The paper is organised as follows. The framework for assessing accuracy is set out in the next section. Section 3 introduces the UK Millennium Cohort Study and the methods are illustrated using data from this study in Section 4. Section 5 concludes.

## 2. Models for predicting non-response

A typical response propensity model for a binary outcome (*e.g.*, Hawkes and Plewis 2006) is:

$$f(\pi_{it}) = \sum_p \beta_p x_{pi} + \sum_q \sum_k \gamma_{qk} x_{qi,t-k}^* + \sum_r \sum_k \delta_{rk} z_{ri,t-k} \quad (1)$$

where

- $\pi_{it} = E(r_{it})$  is the probability of not responding for subject  $i$  at wave  $t$ ;  $r_{it} = 0$  for a response and 1 for non-response;  $f$  is an appropriate function such as logit or probit.
- $i = 1, \dots, n$  where  $n$  is the observed sample size at wave one.
- $t = 1, \dots, T_i$  where  $T_i$  is the number of waves for which  $r_{it}$  is recorded for subject  $i$ .
- $x_{pi}$  are fixed characteristics of subject  $i$  measured at wave one,  $p = 0, \dots, P$ ;  $x_0 = 1$  for all  $i$ .
- $x_{qi,t-k}^*$  are time-varying characteristics of subject  $i$ , measured at waves  $t - k$ ,  $q = 1, \dots, Q$ ,  $k = 1, 2, \dots$ , often  $k$  will be 1.
- $z_{ri,t-k}$  are time-varying characteristics of the data collection process, measured for subject  $i$  at waves  $t - k$ ,  $r = 1, \dots, R$ ,  $k = 0, 1, \dots$ , often  $k$  will be 1 but can be 0 for variables such as number of contacts before a response is obtained.

1. Ian Plewis, Social Statistics, University of Manchester, Manchester M13 9PL, U.K. E-mail: ian.plewis@manchester.ac.uk; Sosthenes Ketende and Lisa Calderwood, Centre for Longitudinal Studies, Institute of Education, London WC1H 0AL, U.K.

Model (1) can easily be extended to more than two response categories such as {response, wave non-response, attrition}. Other approaches are also possible. For example, it is often more convenient to model the probability of not responding just at wave  $t = t^*$  in terms of variables measured at earlier waves  $t^* - k, k \geq 1$  or, when there is no wave non-response so that non-response has a monotonic rather than an arbitrary pattern, to model time to attrition as a survival process.

The estimated response probabilities  $p_i$ , for  $t = t^*$ , are derived from the estimated non-response probabilities in (1) and they can be used to generate inverse probability weights  $g_i (= 1/p_i)$ . These are widely applied (see Section 4.2 for an example) to adjust for biases arising from non-response under the assumption that data are missing at random (MAR) as defined by Little and Rubin (2002).

### 2.1 Assessing the accuracy of predictions

A widely used method of assessing the accuracy of models like (1) is to estimate their goodness-of-fit by using one of several possible pseudo- $R^2$  statistics. Estimates of pseudo- $R^2$  are not especially useful in this context, partly because they are difficult to compare across datasets but also because they assess the overall fit of the model and do not, therefore, distinguish between the accuracy of the model for the respondents and non-respondents separately.

As Pepe (2003) emphasises, there are two related components of accuracy: discrimination (or classification) and prediction. Discrimination refers to the conditional probabilities of having a propensity score ( $s$ : the linear predictor from (1)) above a chosen threshold ( $c$ ) given that a person either is or is not a non-respondent. Prediction, on the other hand, refers to the conditional probabilities of being or

becoming a non-respondent given a propensity score above or below the threshold.

More formally, let  $D$  and  $\bar{D}$  refer to the presence and absence of the poor outcome (*i.e.*, non-response) and define  $+$  ( $s > c$ ) and  $-$  ( $s \leq c$ ) as positive and negative tests derived from the propensity score and its threshold. Then, for discrimination, we are interested in  $P(+|D)$ , the true positive fraction (TPF) or sensitivity of the test, and  $P(-|\bar{D})$  its specificity, equal to one minus the false positive fraction ( $1 - \text{FPF}$ ). For prediction, however, we are interested in  $P(D|+)$ , the positive predictive value (PPV) and  $P(\bar{D}| -)$ , the negative predictive value (NPV). If the probability of a positive test ( $P(+)=\tau$ ) is the same as the prevalence of the poor outcome ( $P(D)=\rho$ ) then inferences about discrimination and prediction are essentially the same: sensitivity equals PPV and specificity equals NPV. Generally, however,  $\{\text{TPF}, \text{FPF}, \rho\}$  and  $\{\text{PPV}, \text{NPV}, \tau\}$  convey different pieces of information. TPF can be plotted against FPF for any risk score threshold  $c$ . This is the receiver operating characteristic (ROC) curve (Figure 1). Krzanowski and Hand (2009) give a detailed discussion of how to estimate ROC curves. The AUC – the area enclosed by the ROC curve and the x-axis in Figure 1 – is of particular interest and can vary from 1 (perfect discrimination) down to 0.5, the area below the diagonal (implying no discrimination). The AUC can be interpreted as the probability of assigning a pair of cases, one respondent and one non-respondent, to their correct categories, bearing in mind that guessing would correspond to a probability of 0.5. A linear transformation of AUC ( $= 2 * \text{AUC} - 1$ ) – sometimes referred to as a Gini coefficient and equivalent to Somer’s D rank correlation index (Harrell, Lee and Mark 1996) – is commonly used as a more natural measure than AUC because it varies from 0 to 1.

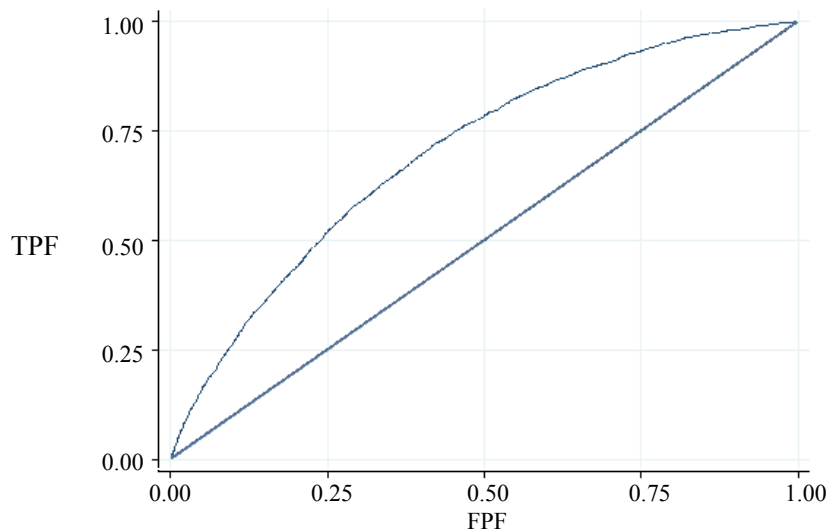


Figure 1 ROC curve

Copas (1999) proposes the logit rank plot as an alternative to the ROC as a means of assessing the predictiveness of a propensity score. If the propensity score is derived from a logistic regression then a logit rank plot is just a plot of the linear predictor from the model against the logistic transformation of the proportional rank of the propensity scores. More generally, it is a plot of  $\text{logit}(p_i)$  where  $p_i$  is the estimated probability from any form of (1) *i.e.*,  $p(D|x, x^*, z)$ , against the logits of the proportional ranks ( $r/n$ ) where  $r$  is the rank position of case  $i$  ( $i = 1, \dots, n$ ) on the propensity score. This relation is usually close to being linear and its slope – which can vary from zero to one – is a measure of the predictive strength of the propensity score. Copas argues that the slope is more sensitive to changes in the specification of the propensity model, and to changes in the prevalence of the outcome, than the Gini coefficient is. A good estimate of the slope can be obtained by calculating quantiles of the variables on the  $y$  and  $x$  axes and then fitting a simple regression model.

The extent to which propensity scores discriminate between respondents and non-respondents is one indicator of the effectiveness of any statistical adjustments for missingness. A lack of discrimination suggests either that there are important predictors absent from the propensity score or that a substantial part of the process that drives the missingness is essentially random. The extent to which propensity scores predict whether a case will be a non-respondent in subsequent waves – and what kind of non-respondent they will be – is an indication of whether any intervention to reduce non-response will be successful.

### 3. The Millennium Cohort Study

The wave one sample of the UK Millennium Cohort Study (MCS) includes 18,552 families born over a 12-month period during the years 2000 and 2001, and living in selected UK electoral wards at age nine months. The initial response rate was 72%. Areas with high proportions of

Black and Asian families, disadvantaged areas and the three smaller UK countries are all over-represented in the sample which is disproportionately stratified and clustered as described in Plewis (2007b). The first four waves took place when the cohort members were (approximately) nine months, 3, 5 and 7 years old. At wave two, 19% of the target sample – which excludes child deaths and emigrants – were unproductive. The unproductive cases were equally divided between wave non-response and attrition, and between refusals and other non-productives (not located, not contacted *etc.*).

## 4. Analyses of non-response

### 4.1 Accuracy of discrimination and prediction

Plewis (2007a) and Plewis *et al.* (2008) show that variables measured at wave one of the MCS that are associated with attrition at wave two are not necessarily associated with wave non-response then (and vice-versa). The same is true for correlates of refusal and other non-productives. Table 1 gives the accuracy estimates from the response propensity models. The estimate of the Gini coefficient for overall non-response (0.38) is relatively low: it corresponds to an AUC of 0.69 which is the probability of correctly assigning (based on their predicted probabilities) a pair of cases (one respondent, one non-respondent), indicating that discrimination between non-respondents and respondents from the propensity score is not especially good. Discrimination is slightly better for wave non-respondents than it is for attrition and notably better for other non-productive than it is for refusal. These estimates were obtained from pairwise comparisons of each non-response category with being a respondent. A similar picture emerges when we look at the slopes of the logit rank plots although these bring out more clearly the differences in predictiveness for the different types of, and reasons for non-response.

**Table 1**  
Accuracy estimates from response propensity models, MCS wave two

Accuracy measure	Overall non-response <sup>(2)</sup>	Non-response type <sup>(2)</sup>		Non-response reason <sup>(2)</sup>	
		Wave non-response	Attrition	Refusal	Other non-productive
AUC <sup>(1)</sup>	0.69	0.71	0.69	0.68	0.77
Gini <sup>(1)</sup>	0.38	0.42	0.39	0.37	0.53
Logit rank plot: slope <sup>(1)</sup>	0.45	0.51	0.44	0.40	0.63
Sample size	18,230	16,210	16,821	16,543	16,513

<sup>(1)</sup> AUC estimated under the binormal assumption (Krzanowski and Hand 2009); 95% confidence limits for (a) AUC not more than  $\pm 0.015$ , (b) Gini coefficient and logit rank plot slope not more than  $\pm 0.03$ .

<sup>(2)</sup> Based on a logistic regression, allowing for the survey design using the `svy` commands in STATA with the sample size based on the sum of the productive and relevant non-response category.

The correct specification of models for explaining non-response can be difficult to achieve. New candidates for inclusion in a model can appear after the model and the corresponding inverse probability weights have been estimated, others remain unknown. How much effect on measures of accuracy might the inclusion of new variables have? Here we examine the effects of adding three new variables to the MCS models: (i) whether or not respondents gave consent to having their survey records linked to health records at wave one; (ii) a neighbourhood conditions score derived from interviewer observations at wave two; and (iii) whether, at wave one, the main respondent reported voting at the last UK general election. The first two of these variables were not available for the analyses summarised in Table 1: refusing consent at wave  $t$  might be followed by overall refusal at wave  $t + 1$ , and non-response might be greater in poorer neighbourhoods. The voting variable is an indicator of social engagement that might be related to the probability of responding. As the neighbourhood conditions score could not be obtained for cases that were not located, we use this variable just in the model that compares refusals with productives.

Table 2 presents the results using the same methods of estimation as for Table 1 with corresponding levels of precision. We see (from the notes) that each of the three variables is associated with at least one kind of non-response. The increase in accuracy of the AUC is more than would be expected by chance ( $p < 0.001$  apart from wave non-response:  $p > 0.06$ ) but is small except for refusal where the inclusion of the three new variables does make a difference: the estimate of the Gini coefficient increases

from 0.37 to 0.41 and the slope of the logit rank plot increases from 0.40 to 0.45 (although missing data for the neighbourhood conditions score does reduce the sample size).

#### 4.2 Using weights to adjust for non-response

Although non-response at wave two of MCS is systematically related to a number of variables measured at or after wave one, we have seen that the models' ability to discriminate between and predict categories of non-response is not high. We now consider what effect the weights generated from the response propensity models have on a longitudinal estimate of interest. We focus on transitions between not working and working across the two waves. As Groves (2006) argues, the keys to unlocking missingness problems of bias are to find those variables that predict whether a piece of data is missing, and which of those variables that predict missingness are also related to the variable of interest. We find that all the variables that predict overall non-response are also related to whether or not the main respondent works at wave two, conditional on whether she was working at wave one so we might expect the application of non-response weights to reduce bias. The results are presented in Table 3 and show that, compared with just using the survey weights, the introduction of the non-response weights based on the model underpinning Table 1 leads to small adjustments in the estimated transition probabilities. The consent and vote variables have no additional effect, however, and this is consistent with the marginal increases in accuracy reported in Table 2.

**Table 2**  
Accuracy estimates for enhanced response propensity models, MCS wave two

Accuracy measure	Overall non-response <sup>(1)</sup>	Non-response type		Non-response reason	
		Wave non-response <sup>(2)</sup>	Attrition <sup>(3)</sup>	Refusal <sup>(4)</sup>	Other non-productive <sup>(5)</sup>
AUC	0.70	0.72	0.71	0.70	0.77
Gini	0.41	0.44	0.41	0.41	0.54
Logit rank plot: slope	0.47	0.52	0.46	0.45	0.65
Sample size	18,148	16,177	16,745	15,656	16,443

<sup>(1)</sup> Includes consent (odds ratio (OR) = 2.1, s.e. = 0.20) and vote (OR = 1.4, s.e. = 0.08).  
<sup>(2)</sup> Includes vote only (OR = 1.4, s.e. = 0.11), consent not important ( $t = 1.33$ ;  $p > 0.18$ ).  
<sup>(3)</sup> Includes consent (OR = 2.7, s.e. = 0.26) and vote (OR = 1.4, s.e. = 0.09).  
<sup>(4)</sup> Includes consent (OR = 2.6, s.e. = 0.32), vote (OR = 1.3, s.e. = 0.10) and neighbourhood score (OR = 1.02, s.e. = 0.014).  
<sup>(5)</sup> Includes consent (OR = 1.6, s.e. = 0.20) and vote (OR = 1.5, s.e. = 0.11).

**Table 3**  
Weighted employment transitions (standard errors), MCS wave two

Variable	Survey weights only	Overall weight <sup>(1)</sup>	Overall weight <sup>(2)</sup>
No change	0.30 (0.0053)	0.30 (0.0056)	0.31 (0.0056)
Working → not working	0.34 (0.0059)	0.35 (0.0059)	0.35 (0.0060)
Not working → working	0.37 (0.0073)	0.35 (0.0073)	0.35 (0.0073)
Weight range <sup>(3)</sup>	0.23 – 2.0	0.19 – 4.1	0.19 – 6.3
Sample size	14,891	14,796	14,733

<sup>(1)</sup> Based on the product of the survey weights and the non-response weights using the model underpinning Table 1.  
<sup>(2)</sup> Non-response weights based on a model that includes consent and vote.  
<sup>(3)</sup> All weights standardised to have mean of one.



## 5. Discussion

Survey methodologists working with longitudinal data have long been exercised by the problem of non-response. Nearly all longitudinal studies suffer from accumulating non-response over time and it is common even for well-conducted mature studies to obtain data for less than half the target sample. On the other hand, a lot can be learnt about the correlates of different types of non-response by drawing on auxiliary variables from earlier waves. The main purpose of this paper has been to introduce a different way of thinking about the utility of the approaches that rely on general linear models both to construct inverse probability weights and to inform imputations. Treating the linear predictors from the regression models as response propensity scores and then generating ROCs enables methods for summarising the information in these scores to be used to assess the accuracy of discrimination and prediction for different kinds of non-response.

The application of this approach to the Millennium Cohort Study has shown that, despite using a wide range of explanatory variables, discrimination is rather low. One implication of this finding is that some non-response is generated by circumstantial factors, none of them important on their own, which can reasonably be regarded as chance. There is some support for this hypothesis in that the accuracy of the models for overall non-response, wave non-response and other non-productive (the latter two being related) were little changed by the introduction of the voting and consent variables. On the other hand, these variables (and the neighbourhood conditions score) did improve the discrimination between productives, and attrition cases and refusals (which are also related). Nevertheless, discrimination for these two categories remained lower than for the other types of non-response. A second possible implication is that the models do not discriminate well because data are not missing at random (NMAR) in Little and Rubin's (2002) sense. In other words, it might be changes in circumstances after the previous wave that influences non-response at the current wave.

The implications of our findings for prediction are that it might be difficult to predict which cases will become non-respondents with a high degree of accuracy. If interventions to prevent non-response in longitudinal studies are to be effective then they need to be targeted at those cases least likely to respond because these cases are probably the most different from the respondents and therefore the major source of bias. This is where the ROC approach can be especially useful because, as Swets, Dawes and Monahan (2000) show, it is possible to determine the optimum threshold for the response propensity score based on the costs and benefits of intervening according to the true and

false positive rates implied by the threshold. A more detailed assessment of these issues is beyond the scope of this paper but would include considering interventions to prevent different kinds of non-response, and the benefits of potential reductions in bias and variability arising from a sample that is both larger and closer in its characteristics to the target sample.

## Acknowledgements

This research was funded by the U.K. Economic and Social Research Council under its Survey Design and Measurement Initiative (ref. RES-175-25-0010).

## References

- Copas, J. (1999). The effectiveness of risk scores: The logit rank plot. *Applied Statistics*, 48, 165-183.
- Groves, R.M. (2006). Nonresponse rates and non-response bias in household surveys. *Public Opinion Quarterly*, 70, 646-675.
- Harrell, F.E. Jr., Lee, K.L. and Mark, D.B. (1996). Multivariable prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine*, 15, 361-387.
- Hawkes, D., and Plewis, I. (2006). Modelling non-response in the National Child Development Study. *Journal of the Royal Statistical Society A*, 169, 479-491.
- Krzanowski, W.J., and Hand, D.J. (2009). *ROC Curves for Continuous Data*. Boca Raton, FL: Chapman and Hall/CRC.
- Lepkowski, J.M., and Couper, M.P. (2002). Nonresponse in the second wave of longitudinal household surveys. In *Survey Nonresponse*, (Eds., R.M. Groves *et al.*). New York: John Wiley & Sons, Inc.
- Little, R.J.A., and Rubin, D.B. (2002). *Statistical Analysis with Missing Data* (2<sup>nd</sup> Ed.). New York: John Wiley & Sons, Inc.
- Pepe, M.S. (2003). *The Statistical Evaluation of Medical Tests for Classification and Prediction*. Oxford: OUP.
- Plewis, I. (2007a). Non-response in a birth cohort study: The case of the Millennium Cohort Study. *International Journal of Social Research Methodology*, 10, 325-334.
- Plewis, I. (Ed.) (2007b). *The Millennium Cohort Study: Technical Report on Sampling* (4<sup>th</sup> Ed.). London: Institute of Education, University of London.
- Plewis, I., Ketende, S.C., Joshi, H. and Hughes, G. (2008). The contribution of residential mobility to sample loss in a birth cohort study: Evidence from the first two waves of the Millennium Cohort Study. *Journal of Official Statistics*, 24, 365-385.
- Swets, J.A., Dawes, R.M. and Monahan, J. (2000). Psychological science can improve diagnostic decisions. *Psychological Sciences in the Public Interest*, 1, 1-26.

**ELECTRONIC  
PUBLICATIONS  
AVAILABLE AT**

**PUBLICATIONS  
ÉLECTRONIQUES  
DISPONIBLE À**

**[www.statcan.gc.ca](http://www.statcan.gc.ca)**

# Confidence interval estimation of small area parameters shrinking both means and variances

Sarat C. Dass, Tapabrata Maiti, Hao Ren and Samiran Sinha<sup>1</sup>

## Abstract

We propose a new approach to small area estimation based on joint modelling of means and variances. The proposed model and methodology not only improve small area estimators but also yield “smoothed” estimators of the true sampling variances. Maximum likelihood estimation of model parameters is carried out using EM algorithm due to the non-standard form of the likelihood function. Confidence intervals of small area parameters are derived using a more general decision theory approach, unlike the traditional way based on minimizing the squared error loss. Numerical properties of the proposed method are investigated via simulation studies and compared with other competitive methods in the literature. Theoretical justification for the effective performance of the resulting estimators and confidence intervals is also provided.

Key Words: EM algorithm; Empirical Bayes; Hierarchical models; Rejection sampling; Sampling variance; Small area estimation.

## 1. Introduction

Small area estimation and related statistical techniques have become a topic of growing importance in recent years. The need for reliable small area estimates is felt by many agencies, both public and private, for making useful policy decisions. An example where small area techniques are used in practice is in the monitoring of socio-economic and health conditions of different age-sex-race groups where the patterns are observed over small geographical areas.

It is now widely recognized that direct survey estimates for small areas are usually unreliable due to their typically large standard errors and coefficients of variation. Hence, it becomes necessary to obtain improved estimates with higher precision. Model-based approaches, either explicit or implicit, are elicited to connect the small areas and improved precision is achieved by “borrowing strength” from similar areas. The estimation technique is also known as shrinkage estimation since the direct survey estimates are shrunk towards the overall mean. The survey based direct estimates and sample variances are the main ingredients for building aggregate level small area models. The typical modeling strategy assumes that the sampling variances are known while a suitable linear regression model is assumed for the means. For details of these developments, we refer to reader to Ghosh and Rao (1994), Pfeffermann (2002) and Rao (2003). The typical area level models are subject to two main criticisms. First, in practice, the sampling variances are estimated quantities, and hence, are subject to substantial errors. This is because they are often based on equivalent sample sizes from which the direct estimates are calculated. Second, the assumption of known and fixed sampling variances of typical small area models does not take into

account the uncertainty in the variance estimation into the overall inference strategy.

Previous attempts have been made to model only the sampling variances; see, for example, Maples, Bell and Huang (2009), Gershunskaya and Lahiri (2005), Huff, Eltinge and Gershunskaya (2002), Cho, Eltinge, Gershunskaya and Huff (2002), Valliant (1987) and Otto and Bell (1995). The articles Wang and Fuller (2003) and Rivest and Vandal (2003) extended the asymptotic mean square error (MSE) estimation of small area estimators when the sampling variances are estimated as opposed to the standard assumption of known variances. Additionally, You and Chapman (2006) considered the modelling of the sampling variances with inference using full Bayesian estimation techniques.

The necessity of variance modelling has been felt by many practitioners. The latest developments in this area are nicely summarized in a recent article by William Bell of the United States Census Bureau 2008. He carefully examined the consequences of these issues in the context of MSE estimation of model based small area estimators. He also provided numerical evidence of MSE estimation for Fay-Herriot models (given in Equation 1) when sampling variances are assumed to be known. The developments in the small area literature so far can be “loosely” viewed as (i) smoothing the direct sampling error variances to obtain more stable variance estimates with low bias and (ii) (partial) accounting of the uncertainty in sampling variances by extending the Fay-Herriot model.

As evident, lesser or no attention has been given to account for the sampling variances effectively while modeling the mean compared to the volume of research that has been done for modeling and inferring the means. There is a lack of systematic development in the small area literature that

1. Sarat C. Dass and Tapabrata Maiti, Department of Statistics & Probability, Michigan State University. E-mail: maiti@stt.msu.edu; Hao Ren, CTB/McGraw-Hill, 20 Ryan Ranch Rd, Monterey, CA 93940; Samiran Sinha, Department of Statistics, Texas A & M University.

includes “shrinking” both means and variances. In other words, we like to exploit the technique of “borrowing strength” from other small areas to “improve” variance estimates as we do to “improve” the small area mean estimates. We propose a hierarchical model which uses both the direct survey and sampling variance estimates to infer all model parameters that determine the stochastic system. Our methodological goal is to develop the dual “shrinkage” estimation for both the small area means and variances, exploiting the structure of the mean-variance joint modelling so that the final estimators are more precise. Numerical evidence shows the effectiveness of dual shrinkage on small area estimates of the mean in terms of the MSE criteria.

Another major contribution of this article is to obtain confidence intervals of small area means. The small area literature is dominated by point estimates and their associated standard errors; it is well known that the standard practice of [point estimate  $\pm q \times$  standard error], where  $q$  is the  $Z$  (standard normal) or  $t$  cut-off point, does not produce accurate coverage probabilities of the intervals; see Hall and Maiti (2006) and Chatterjee, Lahiri and Li (2008) for more details. Previous work is based on the bootstrap procedure and has limited use due to the repeated estimation of model parameters. We produce confidence intervals for the means from a decision theory perspective. The construction of confidence intervals is easy to implement in practice.

The rest of the article is organized as follows. The proposed hierarchical model for the sample means and variances is developed in Section 2. The estimation of model parameters via the EM algorithm is developed in Section 3. Theoretical justification for the proposed confidence interval and coverage properties are presented in Section 4. Sections 5 and 6 present a simulation study and a real data example, respectively. Some discussion and concluding remarks are presented in Section 7. An alternative model formulation for small area as well as mathematical details are provided in the Appendix.

## 2. Proposed model

Suppose  $n$  small areas are in consideration. For the  $i^{\text{th}}$  small area, let  $(X_i, S_i^2)$  be the pair of direct survey estimate and sampling variance, for  $i = 1, 2, \dots, n$ . Let  $\mathbf{Z}_i = (Z_{i1}, \dots, Z_{ip})^T$  be the vector of  $p$  covariates available at the estimation stage for the  $i^{\text{th}}$  small area. We propose the following hierarchical model:

$$\left. \begin{aligned} X_i | \theta_i, \sigma_i^2 &\sim \text{Normal}(\theta_i, \sigma_i^2) \\ \theta_i &\sim \text{Normal}(\mathbf{Z}_i^T \boldsymbol{\beta}, \tau^2) \end{aligned} \right\} \quad (1)$$

$$\left. \begin{aligned} \frac{(n_i - 1)S_i^2}{\sigma_i^2} &\sim \chi_{n_i - 1}^2 \\ \sigma_i^{-2} &\sim \text{Gamma}(a, b), \end{aligned} \right\} \quad (2)$$

independently for  $i = 1, 2, \dots, n$ . In the model elicitation,  $n_i$  is the sample size for a simple random sample (SRS) from the  $i^{\text{th}}$  area,  $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)^T$  is the  $p \times 1$  vector of regression coefficients, and  $\mathbf{B} \equiv (a, b, \boldsymbol{\beta}, \tau^2)^T$  is the collection of all unknown parameters in the model. Also,  $\text{Gamma}(a, b)$  is the Gamma density function with positive shape and scale parameters  $a$  and  $b$ , respectively, defined as  $f(x) = \{b^a \Gamma(a)\}^{-1} e^{-x/b} x^{a-1}$  for  $x > 0$ , and 0 otherwise. The unknown  $\sigma_i^2$  is the true variance of  $X_i$  and is usually estimated by the sample variance  $S_i^2$ . Although  $S_i^2$ 's are assumed to follow a chi-square distribution with  $(n_i - 1)$  degrees of freedom (as a result of normality and SRS), we note that for complex survey designs, the degree of freedom needs to be determined carefully [e.g., Maples *et al.* 2009]. More importantly, the role of the sample sizes in shrinkage estimation of  $\sigma_i^2$  is as follows: For low values of  $n_i$ , the estimate of  $\sigma_i^{-2}$  is shrunk more towards the overall mean ( $ab$ ) compared to higher  $n_i$  values. Thus, for variances, sample sizes play the same role as precision in shrinkage estimation of the small area mean estimates. We note that You and Chapman (2006) also considered the second level of the sampling variance modelling. However, the hyperparameters related to prior of  $\sigma_i^2$  are not data driven, they are rather chosen in such a way that the prior will be vague. Thus, their model can be viewed as the Bayesian version of the models considered in Rivest and Vandal (2003) and Wang and Fuller (2003). The second level modelling of  $\sigma_i^{-2}$  in (2) can be further extended to  $\sigma_i^{-2} \sim \text{Gamma}(b, \exp(\mathbf{Z}_i^T \boldsymbol{\beta}_2)/b)$  so that  $E(\sigma_i^{-2}) = \exp(\mathbf{Z}_i^T \boldsymbol{\beta}_2)$  for another set of  $p$  regression coefficients  $\boldsymbol{\beta}_2$  to accommodate covariate information in the variance modeling.

Although our model is motivated by Hwang, Qiu and Zhao (2009), we like to mention that Hwang *et al.* (2009) considered shrinking means and variances in the context of microarray data where they prescribed an important solution by plugging in a shrinkage estimator of variance into the mean estimator. The shrinkage estimator of the variance in Hwang *et al.* (2009) is a function of  $S_i^2$  only, and not of both  $X_i$  and  $S_i^2$ ; see Remarks 2 and 3 in Section 2. Thus, inference of the mean does not take into account the full uncertainty in the variance estimation. Further, their model does not include any covariate information. The simulation study described subsequently indicate that our method of estimation performed better than Hwang *et al.* (2009).

In the above model formulation, inference for the small area mean parameter  $\theta_i$  can be made based on the conditional distribution of  $\theta_i$  given all of the data  $\{(X_i, S_i^2, \mathbf{Z}_i), i = 1, \dots, n\}$ . Under our model set up, the conditional

distribution of  $\theta_i$  is a non-standard distribution and does not have a closed form, thus requiring numerical methods, such as Monte Carlo and the EM algorithm, for inference, and the details are provided in the next section.

### 3. Inference methodology

#### 3.1 Estimation of unknown parameters via EM algorithm

In practice,  $\mathbf{B} \equiv (a, b, \boldsymbol{\beta}, \tau^2)^T$  is unknown and has to be estimated from the data  $\{(X_i, S_i^2, \mathbf{Z}_i), i = 1, 2, \dots, n\}$ . Our proposal is to estimate  $\mathbf{B}$  by the marginal maximum likelihood method: Estimate  $\mathbf{B}$  by  $\hat{\mathbf{B}}$  where  $\hat{\mathbf{B}}$  maximizes the marginal likelihood  $L_M(\mathbf{B}) = \prod_{i=1}^n L_{M,i}(\mathbf{B})$ , where

$$L_{M,i} \propto \frac{\Gamma(n_i/2+a)}{\tau \Gamma(a) b^a} \int \exp\left\{-\frac{(\theta_i - \mathbf{Z}_i^T \boldsymbol{\beta})^2}{2\tau^2}\right\} \psi_i^{-(n_i/2+a)} d\theta_i, \quad (3)$$

and

$$\psi_i \equiv \left\{0.5(X_i - \theta_i)^2 + 0.5(n_i - 1)S_i^2 + \frac{1}{b}\right\}. \quad (4)$$

The marginal likelihood  $L_M$  involves integrals that cannot be evaluated in closed-form, and hence, one has to resort to numerical methods for its maximization. One such algorithm is the EM (Expectation-Maximization) iterative procedure which is used when such integrals are present. The EM algorithm involves augmenting the observed likelihood  $L_M(\mathbf{B})$  with missing data; in our case, the variables of the integration,  $\theta_i, i = 1, 2, \dots, n$ , constitute this missing information. Given  $\boldsymbol{\theta} \equiv \{\theta_1, \theta_2, \dots, \theta_n\}$ , the complete data log likelihood ( $\ell_c$ ) can be written as

$$\ell_c(\mathbf{B}, \boldsymbol{\theta}) = \sum_{i=1}^n \left[ \log\{\Gamma(n_i/2+a)\} - \log\{\Gamma(a)\} - a \log(b) - 0.5 \log(\tau^2) - \frac{(\theta_i - \mathbf{Z}_i^T \boldsymbol{\beta})^2}{2\tau^2} - (n_i/2+a) \log(\psi_i) \right],$$

where the expression of  $\psi_i$  is given in Equation (4). Starting from an initial value of  $\mathbf{B}, \mathbf{B}^{(0)}$  say, the EM algorithm iteratively performs a maximization with respect to  $\mathbf{B}$ . At the  $t^{\text{th}}$  step the objective function maximized is

$$\begin{aligned} Q(\mathbf{B} | \mathbf{B}^{(t-1)}) &= E(\ell_c(\mathbf{B}, \boldsymbol{\theta})) \\ &= \sum_{i=1}^n \left[ \log\{\Gamma(n_i/2+a)\} - \log\{\Gamma(a)\} - a \log(b) - 0.5 \log(\tau^2) - \frac{E(\theta_i - \mathbf{Z}_i^T \boldsymbol{\beta})^2}{2\tau^2} - (n_i/2+a) E\{\log(\psi_i)\} \right]. \end{aligned}$$

The expectation in  $Q(\mathbf{B} | \mathbf{B}^{(t-1)})$  is taken with respect to the conditional distribution of each  $\theta_i$  given the data,  $\pi(\theta_i | X_i, S_i^2, \mathbf{Z}_i, \mathbf{B}^{(t-1)})$ , which is

$$\pi(\theta_i | X_i, S_i^2, \mathbf{Z}_i, \mathbf{B}) \propto \exp\{-0.5(\theta_i - \mathbf{Z}_i^T \boldsymbol{\beta})^2 / \tau^2\} \psi_i^{-(n_i/2+a)}. \quad (5)$$

One challenge here is that the expectations are not available in closed form. Thus, we resort to a Monte Carlo method for evaluating the expressions. Suppose that  $R$  iid samples of  $\theta_i$  are available, say  $\theta_{i,1}, \theta_{i,2}, \dots, \theta_{i,R}$ . Then, each expectation of the form  $E\{h(\theta_i)\}$  can be approximated by the Monte Carlo mean

$$E\{h(\theta_i)\} \approx \frac{1}{R} \sum_{r=1}^R h(\theta_{i,k}). \quad (6)$$

However, drawing random numbers from the conditional distribution  $\pi(\theta_i | X_i, S_i^2, \mathbf{Z}_i, \mathbf{B}^{(t-1)})$  is also not straightforward since this is not a standard density. Samples are drawn using the accept-reject procedure (Robert and Casella 2004): For a sample from the target density  $f$ , sample  $x$  from the proposal density  $g$ , and accept the sample as a sample from  $f$  with probability  $f(x)/\{M^*g(x)\}$  where  $M^* = \sup_x \{f(x)/g(x)\}$ . One advantage of the accept-reject method is that the target density  $f$  only needs to be known upto a constant of proportionality which is the case for  $\pi(\theta_i | X_i, S_i^2, \mathbf{Z}_i, \mathbf{B}^{(t-1)})$  in (5); due to the non-standard form of the density, the normalizing constant cannot be found in a closed form. For the accept-reject algorithm, we used the normal density  $g(\theta_i) \propto \exp\{-0.5(\theta_i - \mathbf{Z}_i^T \boldsymbol{\beta})^2 / \tau^2\}$  as the proposal density. The acceptance probability is calculated to be  $[\{1/b + 0.5(n_i - 1)S_i^2\} / \{1/b + 0.5(n_i - 1)S_i^2 + 0.5(\theta_i - X_i)^2\}]^{n_i/2+a}$ . One can choose a better proposal distribution to increase acceptance probability or different algorithm (such as the adaptive rejection sampling or envelope accept-reject algorithms) but our chosen proposal worked satisfactorily in the studies we conducted.

The maximizer of  $Q(\mathbf{B} | \mathbf{B}^{(t-1)})$  at the  $t^{\text{th}}$  step can be described explicitly. The solutions for  $\boldsymbol{\beta}$  and  $\tau^2$  are available in closed form as

$$\boldsymbol{\beta}^{(t)} = \left( \sum_{i=1}^n \mathbf{Z}_i \mathbf{Z}_i^T \right)^{-1} \left( \sum_{i=1}^n \mathbf{Z}_i E(\theta_i) \right)$$

and

$$(\tau^2)^{(t)} = \frac{1}{n} \sum_{i=1}^n E(\theta_i - \mathbf{Z}_i^T \boldsymbol{\beta})^2,$$

respectively. Also,  $a^{(t)}$  and  $b^{(t)}$  are obtained by solving  $S_a = \partial Q(\mathbf{B} | \mathbf{B}^{(t-1)}) / \partial a = 0$  and  $S_b = \partial Q(\mathbf{B} | \mathbf{B}^{(t-1)}) / \partial b = 0$  using the Newton-Raphson method where

$$S_a = \sum_{i=1}^n \frac{\partial}{\partial a} \log\{\Gamma(n_i/2 + a)\} - n \left\{ \frac{\partial}{\partial a} \log\{\Gamma(a)\} \right\} - n \log(b) - \sum_{i=1}^n E\{\log(\psi_i)\}$$

and

$$S_b = -\frac{na}{b} + \sum_{i=1}^n \frac{(n_i/2 + a)}{b^2} E(\psi_i^{-1}).$$

We set  $\mathbf{B}^{(t)} = (a^{(t)}, b^{(t)}, \boldsymbol{\beta}^{(t)}, (\tau^{(t)})^2)$  and proceed to the  $(t + 1)$ -st step. This maximization procedure is repeated until the estimate  $\mathbf{B}^{(t)}$  converges. The MLE of  $\mathbf{B}$ ,  $\hat{\mathbf{B}} = \mathbf{B}^{(\infty)}$ , once convergence is established.

### 3.2 Point estimate and confidence interval for $\theta_i$

Following the standard technique, the small area estimator of  $\theta_i$  is taken to be

$$\hat{\theta}_i = E(\theta_i | X_i, S_i^2, \mathbf{Z}_i, \mathbf{B}) \Big|_{\mathbf{B}=\hat{\mathbf{B}}}, \tag{7}$$

the expectation of  $\theta_i$  with respect to the conditional density  $\pi(\theta_i | X_i, S_i^2, \mathbf{Z}_i, \mathbf{B})$  with the maximum likelihood estimate  $\hat{\mathbf{B}}$  plugged in for  $\mathbf{B}$ . The estimate  $\hat{\theta}_i$  is calculated numerically using the Monte Carlo procedure (6) described in the previous section. Subsequently, all quantities involving the unknown  $\mathbf{B}$  will be plugged in by  $\hat{\mathbf{B}}$  although we still keep using the notation  $\mathbf{B}$  for simplicity.

Further, we develop a confidence interval for  $\theta_i$  based on a decision theory approach. Following Joshi (1969), Casella and Hwang (1991), Hwang *et al.* (2009), consider the loss function associated with the confidence interval  $C$  given by  $(k/\sigma)L(C) - I_C(\theta)$  where  $k$  is a tuning parameter independent of the model parameters,  $L(C)$  is the length of  $C$  and  $I_C(\theta)$  is the indicator function taking values 1 or 0 depending on whether  $\theta \in C$  or not. Note that this loss function takes into account both the coverage probability as well as the length of the interval; the positive quantity  $(k/\sigma)$  serves as the relative weight of the length compared to the coverage probability of the confidence interval. If  $k = 0$ , the length of the interval is not under consideration, which leads to the optimal  $C$  to be  $(-\infty, \infty)$  with coverage probability 1. On the other hand, if  $k = \infty$ , then the coverage probability is 0, leading to optimal  $C$  to be a point set. The Bayes confidence interval for  $\theta_i$  is obtained by minimizing the risk function (the expected loss)  $E\{[(k/\sigma)L(C) - I_C(\theta)] | X_i, S_i^2, \mathbf{Z}_i, \mathbf{B}\}$ . The optimal choice of  $C$  is given by

$$C_i(\mathbf{B}) = \{\theta_i: kE(\sigma_i^{-1} | X_i, S_i^2, \mathbf{Z}_i, \mathbf{B}) < \pi(\theta_i | X_i, S_i^2, \mathbf{Z}_i, \mathbf{B})\}. \tag{8}$$

Since  $C_i(\mathbf{B})$  is obtained by minimizing the posterior risk, one may like to interpret this as a Bayesian credible set. However, following Casella and Berger (1990, page 470), we will continue naming  $C_i(\mathbf{B})$  as a confidence interval. From an empirical Bayes perspective also, this terminology is more appropriate. How the tuning parameter  $k$  determines the confidence level of  $C_i(\mathbf{B})$  will be shown explicitly in Section 3.3.

Assuming  $k$  is known for the moment, we follow the steps below to calculate  $C_i(\mathbf{B})$ . The conditional densities of  $\sigma_i^2$  and  $\theta_i$  are given by

$$\pi(\sigma_i^2 | X_i, S_i^2, \mathbf{Z}_i, \mathbf{B}) \propto \frac{\exp\left[ \frac{-0.5(X_i - \mathbf{Z}_i^T \boldsymbol{\beta})^2}{(\sigma_i^2 + \tau^2) - \left\{0.5(n_i - 1)S_i^2 + \frac{1}{b}\right\} \left(\frac{1}{\sigma_i^2}\right)} \right]}{(\sigma_i^2)^{(n_i-1)/2+a+1} (\sigma_i^2 + \tau^2)^{1/2}} \tag{9}$$

and (5), respectively, which as mentioned before, are not available in closed form. Thus, similar to the case of  $\theta_i$ ,  $E(\sigma_i^{-1} | X_i, S_i^2, \mathbf{Z}_i, \mathbf{B})$  is computed numerically using the Monte Carlo method by approximating the expected value with the mean  $1/N \sum_{k=1}^N 1/\sigma_{i,k}$  where  $\sigma_{i,r}^2$ ,  $r = 1, 2, \dots, R$  are  $R$  samples from the conditional density  $\pi(\sigma_i^2 | X_i, S_i^2, \mathbf{Z}_i, \mathbf{B})$ . The accept reject procedure is used to draw random numbers from  $\pi(\sigma_i^2 | X_i, S_i^2, \mathbf{Z}_i, \mathbf{B})$  with a proposal density given by the inverse Gamma

$$\frac{\exp\left[ -\left\{0.5(n_i - 1)S_i^2 + \frac{1}{b}\right\} \left(\frac{1}{\sigma_i^2}\right) \right]}{(\sigma_i^2)^{(n_i-1)/2+a+1}},$$

and the acceptance probability

$$\frac{\exp\left\{ \frac{-0.5(X_i - \mathbf{Z}_i^T \boldsymbol{\beta})^2}{(\sigma_i^2 + \tau^2)} \right\}}{(\sigma_i^2 + \tau^2)^{1/2}} \times \exp(0.5) \times |X_i - \mathbf{Z}_i^T \boldsymbol{\beta}|.$$

The next step is to determine the boundary values of  $C_i(\mathbf{B})$  by finding two  $\theta_i$  values that satisfy the equation  $kE(\sigma_i^{-1} | X_i, S_i^2, \mathbf{Z}_i, \mathbf{B}) - \pi(\theta_i | X_i, S_i^2, \mathbf{Z}_i, \mathbf{B}) = 0$ . This requires the normalizing constant in (5)

$$D_i = \int_{-\infty}^{\infty} \exp\{-0.5(\theta_i - \mathbf{Z}_i^T \boldsymbol{\beta})^2 / \tau^2\} \psi_i^{-(n_i/2+a)} d\theta_i$$

to be evaluated numerically. This is obtained using the Gauss-Hermite integration with 20 nodes.

### 3.3 Choice of $k$

The choice of the tuning parameter  $k$  in (8) is taken to be

$$k = k(\mathbf{B}) = u_{i,0} \phi \left( t_{\alpha/2} \sqrt{\frac{n_i + 2a + 2}{n_i - 1}} \right) \quad (10)$$

where  $\phi$  is the standard normal distribution,  $t_{\alpha/2}$  is  $(1 - \alpha / 2)^{\text{th}}$  percentile of  $t$  distribution with  $(n_i - 1)$  degrees of freedom, and  $u_{i,0} = \sqrt{1 + \sigma_i^2 / \tau^2}$ . Since  $u_{i,0}$  involves  $\sigma_i^2$  which is unknown, an estimated version  $\hat{u}_{i,0}$  is obtained by plugging in the maximum a posteriori estimate

$$\hat{\sigma}_i^2 = \hat{\sigma}_i^2(\hat{\mathbf{B}}) = \arg \max_{\sigma_i^2} \pi(\sigma_i^2 | X_i, S_i^2, \mathbf{Z}_i, \mathbf{B}) \Big|_{\mathbf{B}=\hat{\mathbf{B}}} \quad (11)$$

in place of  $\sigma_i^2$ . Also,  $\mathbf{B}$  is replaced by  $\hat{\mathbf{B}}$  in (11). We demonstrate that the coverage probability of  $C_i(\hat{\mathbf{B}})$  with this choice of  $k$  is close to  $1 - \alpha$ . Theoretical justifications are provided in Section 4.

### 3.4 Other related methods for comparison

Our method will be denoted as Method I. Three other methods to be compared are briefly described below.

*Method II:* Wang and Fuller (2003) considered the Fay-Herriot small area estimation model given by (1). Their primary contribution is the construction of the mean squared error estimation formulae for small area estimators with estimated sampling variances. In the process, they had constructed two formulae denoted by  $\widehat{\text{MSE}}_1$  and  $\widehat{\text{MSE}}_2$ . We use  $\widehat{\text{MSE}}_1$  for our comparisons, which was derived following the bias correction approach of Prasad and Rao (1990). The basic difference with our approach is that they did not smooth the sampling variances, only taking the uncertainty into account while making inference on the small area parameters. The method of parameter estimation, which is moment based for all the model parameters, is also different from ours.

*Method III:* Hwang *et al.* (2009) considered the log-normal and inverse Gamma models for  $\sigma_i^2$  in (2) for microarray data analysis. Their simulation study showed improved performance of confidence intervals for small area estimators under the log-normal model compared to the inverse gamma. We thus modified their log-normal model to add covariates and for unequal sample sizes  $n_i$  as follows:

$$\left. \begin{aligned} X_i | \theta_i, \sigma_i^2 &\sim \text{Normal}(\theta_i, \sigma_i^2) \\ \theta_i &\sim \text{Normal}(\mathbf{Z}_i^T \boldsymbol{\beta}, \tau^2); \end{aligned} \right\} \quad (12)$$

$$\left. \begin{aligned} \log S_i^2 &= \log(\sigma_i^2) + \delta_i; \delta_i \sim N(m_i, \sigma_{ch,i}^2) \\ \log(\sigma_i^2) &\sim N(\mu_v, \tau_v^2), \end{aligned} \right\} \quad (13)$$

independently for  $i = 1, 2, \dots, n$ . Note that the model for the means in (12) is identical to (1). The quantities  $\tau^2$ ,  $m_i$  and  $\sigma_{ch,i}^2$  are assumed to be known and are given by  $m_i = E[\log(\chi_{n_i-1}^2 / (n_i - 1))]$  and  $\sigma_{ch,i}^2 = \text{Var}[\log(\chi_{n_i-1}^2 / (n_i - 1))]$ .

Thus, the sample size  $n_i$ 's determine the shape of the  $\chi^2$  distribution via its degrees of freedom parameter. More importantly, as mentioned earlier, the different sample sizes account for different degrees of shrinkage for the corresponding true variance parameter. Similar to their estimation approach, the unknown model parameters  $\mu_v$  and  $\tau_v^2$  are estimated using a moment based approach in an empirical Bayes framework giving  $\hat{\mu}_v$  and  $\hat{\tau}_v^2$ , respectively. Note that in Hwang *et al.* (2009), these estimates are obtained based on the hierarchical model for  $\sigma_i^2$  of (13) *only* without regard to the modelling (1) of the mean. We refer to the Section 5 of their paper for details of the estimation of the hyper-parameters. We follow the same procedure using only (13) to estimate  $\mu_v$  and  $\tau_v^2$  in the case of unequal sample sizes.

The Bayes estimate of  $\sigma_i^2$  is derived to be

$$\begin{aligned} \hat{\sigma}_{i,B}^2 &= \exp \left[ E \{ \ln(\sigma_i^2) \mid \ln(S_i^2) \} \right] \\ &= \left\{ \frac{S_i^2}{\exp(m_i)} \right\}^{M_{v,i}} \exp \{ \mu_v (1 - M_{v,i}) \} \end{aligned}$$

where  $M_{v,i} = \tau_v^2 / (\tau_v^2 + \sigma_{ch,i}^2)$  and with estimates plugged in for the unknown quantities. The conditional distribution of  $\theta_i$  given  $(X_i, S_i^2)$ , is

$$\pi(\theta_i | X_i, S_i^2) = \int_0^\infty \pi(\theta_i | X_i, S_i^2, \sigma_i^2) \pi(\sigma_i^2 | X_i, S_i^2) d\sigma_i^2,$$

is approximated as  $\pi(\theta_i | X_i, S_i^2) \approx \int_0^\infty \pi(\theta_i | X_i, S_i^2, \hat{\sigma}_{i,B}^2) \pi(\sigma_i^2 | X_i, S_i^2) d\sigma_i^2 = \pi(\theta_i | X_i, S_i^2, \hat{\sigma}_{i,B}^2)$ . This suggests the approximate Bayes estimator of the small area parameters given by

$$\hat{\theta}_i = E(\theta_i | X_i, \hat{\sigma}_{i,B}^2) = \hat{M}_i X_i + (1 - \hat{M}_i) \mathbf{Z}_i^T \hat{\boldsymbol{\beta}}, \quad (14)$$

where  $\hat{M}_i = \hat{\tau}_v^2 / (\hat{\tau}_v^2 + \hat{\sigma}_{i,B}^2)$ . The confidence interval for  $\theta_i$  is obtained as

$$C_i^H = \left\{ \theta_i : \frac{|\theta_i - \hat{\theta}_i|}{\hat{M}_i \hat{\sigma}_{i,B}^2} < -2 \ln \{ k \sqrt{2\pi} \} - \ln(\hat{M}_i) \right\}. \quad (15)$$

In Section 3 of Hwang *et al.* (2009) pages 269-271, the interval  $C_i^H$  is matched with the  $100(1 - \alpha)\%$   $t$ -interval  $[|\theta_i - X_i| < t S_i]$  to obtain the expression of  $k$  as  $k \equiv k_i = \exp\{-t^2/2\} \exp\{m_i/2\} / (\sqrt{2\pi})$ .

*Method IV:* This method comprises of a special case of the Fay-Herriot model in (1) but with the estimation of model parameters adopted from Qiu and Hwang (2007). Qiu and Hwang (2007) considered the model

$$\left. \begin{aligned} X_i | \theta_i, \sigma^2 &\sim \text{Normal}(\theta_i, \sigma^2) \\ \theta_i &\sim \text{Normal}(0, \tau^2), \end{aligned} \right\} \quad (16)$$

independently for  $i = 1, 2, \dots, n$ , for analyzing microarray experimental data. When model parameters are known, they proposed the point estimator  $\hat{\theta}_i = \hat{M}X_i$ ,  $\hat{M} = (1 - ((n - 2)\sigma^2 / |X|^2))_+$  where  $a_+$  denotes  $\max(0, a)$  for any number  $a$  and  $|X| = (\sum_{i=1}^n X_i^2)^{1/2}$ . The confidence interval for  $\theta_i$  is  $\hat{\theta}_i \pm v_1(\hat{M})$ , where  $v_1(\hat{M}) = \sigma^2 \hat{M}(q_1 - \ln(\hat{M}))$  with  $q_1$  denoting the standard normal cut-off point corresponding to desired level of confidence coefficient and  $v_1(0) \equiv 0$ . Here For the purpose of comparisons with our method, the first level of the hierarchical model in (16) is modified as follows:

$$X_i = \mathbf{Z}_i^T \boldsymbol{\beta} + v_i + e_i$$

where  $v_i \sim \text{Normal}(0, \tau^2)$  and  $e_i \sim \text{Normal}(0, S_i^2)$  independently for  $i = 1, 2, \dots, n$ , and  $S_i^2$  is treated as known. Following Qiu and Hwang (2007),  $\tau^2$  is estimated by

$$\hat{\tau}^2 = \frac{1}{n - p} \left[ \sum_i \hat{u}_i^2 - \sum_i S_i^2 \left\{ 1 - \mathbf{Z}_i^T \left( \sum_{i=1}^n \mathbf{Z}_i \mathbf{Z}_i^T \right)^{-1} \mathbf{Z}_i \right\} \right]$$

and  $\hat{\tau}^2 = \max(\hat{\tau}^2, 1/n)$  where  $\hat{u}_i = X_i - \mathbf{Z}_i^T \hat{\boldsymbol{\beta}}$  and  $\hat{\boldsymbol{\beta}} = (\sum_{i=1}^n \mathbf{Z}_i \mathbf{Z}_i^T)^{-1} (\sum_{i=1}^n \mathbf{Z}_i X_i)$ . Next, define  $\hat{M}_{0i} = \hat{\tau}^2 / (\hat{\tau}^2 + S_i^2)$  and  $\hat{M}_i = \max(\hat{M}_{0i}, M_i)$  where in the latter expression,  $\hat{M}_{0i}$  is truncated by  $M_{li} = 1 - Q_\alpha / (n_i - 2)$ , and  $Q_\alpha$  is the  $\alpha^{\text{th}}$  quantile of a chi-squared distribution with  $n_i$  degrees of freedom. This  $\hat{M}_i$  is used in the formula of the confidence interval for  $\theta_i$  given earlier. When applying this method in our simulation study and real data analysis, we modified the model to accommodate such unequal sample sizes and covariate information mentioned earlier.

*Remark 1.* Hwang *et al.* (2009) choose  $k$  by equating (15) to the  $t$  interval based on only  $X_i$  for the small area parameters  $\theta_i$ . Note that  $X_i$  is the direct survey estimator. Consequently, this choice of  $k$  does not have any direct control over the coverage probability of the interval constructed under *shrinkage estimation*. On the other hand, our proposed choice of  $k$  has been derived to maintain nominal coverage under, specifically, shrinkage estimation.

*Remark 2.* Note that without any hierarchical modelling assumption,  $S_i$  and  $X_i$  are independent as  $S_i^2$  and  $X_i$  are, respectively, ancillary and the complete sufficient statistics for  $\theta_i$ . However, under models (1) and (2) the conditional distribution of  $\sigma_i^2$  and  $\theta_i$  involve both  $X_i$  and  $S_i^2$  which is seen from (5) and (9).

*Remark 3.* In Hwang *et al.* (2009), the shrinkage estimator for  $\sigma_i^2$  is based only on the information on  $S_i^2$ , and not of both  $X_i$  and  $S_i^2$ . The Bayes estimator of  $\sigma_i^2$  is plugged into the expression for the Bayes estimator of small area parameters. Thus, Hwang *et al.*'s small area estimator is written as  $E(\theta_i | X_i, \hat{\sigma}_{i,B}^2)$  in (14) where  $\hat{\sigma}_{i,B}^2$  is the Bayes

estimator of  $\sigma_i^2$ . Due to equation (9), the shrinkage estimator of  $\sigma_i^2$  depends on  $(X_i - \mathbf{Z}_i^T \boldsymbol{\beta})^2$  in addition to  $S_i^2$  in contrast to Hwang *et al.* (2009). We believe this could be the reason for improved performance of our method compared to Hwang *et al.* (2009).

*Remark 4.* As mentioned previously, the degree of freedom associated with the  $\chi^2$  distribution for the sampling variance need not to be simply  $n_i - 1$ ,  $n_i$  being the sample size for  $i^{\text{th}}$  area. There is no sound theoretical result for determining the degree of freedom when the survey design is complex. The article Wang and Fuller (2003) approximated the  $\chi^2$  with a normal based on the Wilson-Hilferty approximation. If one knows the exact sampling design then the simulation based guideline of Maples *et al.* (2009) could be useful. For county level estimation using the American Community Survey, Maples *et al.* (2009) suggested the estimated degrees of freedom of  $0.36 \times \sqrt{n_i}$ .

### 4. Theoretical justification

Theoretical justification for the choice of  $k$  according to equation (10) is presented in this section. As in Hwang *et al.* (2009), the conditional distribution of  $\theta_i$  given  $X_i$  and  $S_i^2$  can be approximated as  $\pi(\theta_i | X_i, S_i^2, \mathbf{B}) \approx \pi(\theta_i | X_i, S_i^2, \mathbf{B}, \hat{\sigma}_i^2)$ , where  $\hat{\sigma}_i^2$  as defined in (11). In a similar way, approximate  $E(\sigma_i^{-1} | X_i, S_i^2, \mathbf{B})$  by  $E(\sigma_i^{-1} | X_i, S_i^2, \mathbf{B}) \approx \hat{\sigma}_i^{-1}$ . Based on these approximations, we have  $C_i(\mathbf{B}) \approx \tilde{C}_i(\mathbf{B})$  where  $\tilde{C}_i(\mathbf{B})$  is the confidence interval for  $\theta_i$  given by  $\tilde{C}_i(\mathbf{B}) = \{\theta_i: \pi(\theta_i | X_i, S_i^2, \mathbf{B}, \hat{\sigma}_i^2) \geq k \hat{\sigma}_i^{-1}\}$ . From (1), it follows that the conditional density  $\pi(\theta_i | X_i, S_i^2, \mathbf{B}, \sigma_i^2)$  is a normal with mean  $\mu_i$  and variance  $v_i$ , where  $\mu_i$  and  $v_i$  are given by the expressions

$$\begin{aligned} \mu_i &= w_i X_i + (1 - w_i) \mathbf{Z}_i^T \boldsymbol{\beta}, \\ v_i &= \left( \frac{1}{\sigma_i^2} + \frac{1}{\tau^2} \right)^{-1} = \sigma_i^2 \left( 1 + \frac{\sigma_i^2}{\tau^2} \right)^{-1}, \end{aligned} \tag{17}$$

and

$$w_i = \frac{1 / \sigma_i^2}{(1 / \sigma_i^2 + 1 / \tau^2)}.$$

Now, choosing

$$k = \hat{u}_0 \phi \left( t_{\alpha/2} \sqrt{\frac{n_i + 2a + 2}{n_i - 1}} \right)$$

as discussed, the confidence interval  $\tilde{C}_i(\mathbf{B})$  becomes

$$\tilde{C}_i(\mathbf{B}) = \left\{ \theta_i: \hat{u}_{0i} \frac{|\theta_i - \hat{\mu}_i|}{\hat{\sigma}_i} \leq t_{\alpha/2} \sqrt{\frac{n_i + 2a + 2}{n_i - 1}} \right\}, \tag{18}$$



where  $\hat{\mu}_i$  is the expression for  $\mu_i$  in (17) with  $\sigma_i^2$  replaced by  $\hat{\sigma}_i^2$ . Now consider the behavior of  $\hat{\sigma}_i^2 \equiv \hat{\sigma}_i^2(\mathbf{B})$  as  $\tau^2$  ranges between 0 and  $\infty$ . When  $\tau^2 \rightarrow \infty$ ,  $\hat{\sigma}_i^2$  converges to

$$\hat{\sigma}_i^2(\infty) \equiv \hat{\sigma}_i^2(a, b, \boldsymbol{\beta}, \infty) = \frac{\frac{(n_i-1)S_i^2 + \frac{1}{b}}{\frac{n_i-1}{2} + a+1}}{\frac{(n_i-1)S_i^2 + \frac{2}{b}}{n_i + 2a + 1}}.$$

Similarly, when  $\tau^2 \rightarrow 0$ ,  $\hat{\sigma}_i^2$  converges to

$$\hat{\sigma}_i^2(0) \equiv \hat{\sigma}_i^2(a, b, \boldsymbol{\beta}, 0) = \frac{(X_i - \mathbf{Z}_i^T \boldsymbol{\beta})^2 + (n_i - 1)S_i^2 + \frac{2}{b}}{n_i + 2a + 2}.$$

For all intermediate values of  $\tau^2$ , we have  $\min\{\hat{\sigma}_i^2(0), \hat{\sigma}_i^2(\infty)\} \leq \hat{\sigma}_i^2 \leq \max\{\hat{\sigma}_i^2(0), \hat{\sigma}_i^2(\infty)\}$ . Therefore, it is sufficient to consider the following two cases: (i)  $\hat{\sigma}_i^2 \geq \hat{\sigma}_i^2(\infty)$ , where it follows that  $(n_i + 2a + 2)\hat{\sigma}_i^2 = (n_i + 2a + 1)\hat{\sigma}_i^2 + \hat{\sigma}_i^2 \geq (n_i - 1)S_i^2 + 2/b + \hat{\sigma}_i^2 \geq (n_i - 1)S_i^2$ , and (ii)  $\hat{\sigma}_i^2 \leq \hat{\sigma}_i^2(0)$ , where it follows that  $(n_i + 2a + 2)\hat{\sigma}_i^2 = (X_i - \mathbf{Z}_i^T \boldsymbol{\beta})^2 + (n_i - 1)S_i^2 + 2/b \geq (n_i - 1)S_i^2$ . So, in both cases (i) and (ii),

$$(n_i + 2a + 2) \hat{\sigma}_i^2 \geq (n_i - 1)S_i^2. \tag{19}$$

Since  $\theta_i - \mu_i \sim N(0, \sigma_i^2 \tau^2 / (\sigma_i^2 + \tau^2))$  and  $(n_i - 1)S_i^2 / \sigma_i^2 \sim \chi_{n_i-1}^2$ , the confidence interval

$$D_i = \left\{ \theta_i : u_{0i} \frac{|\theta_i - \mu_i|}{S_i} \leq t_{\alpha/2} \right\} \tag{20}$$

has coverage probability  $1 - \alpha$ . Thus, if  $u_0$  and  $\mu_i$  are replaced by  $\hat{u}_0$  and  $\hat{\mu}_i$ , it is expected that the resulting confidence interval  $\tilde{D}_i$ , say, will have coverage probability of approximately  $1 - \alpha$ . From (19), we have

$$P\{\tilde{C}_i(\mathbf{B})\} \geq P(\tilde{D}_i) \approx 1 - \alpha, \tag{21}$$

establishing an approximate lower bound of  $1 - \alpha$  for the confidence level of  $\tilde{C}_i(\mathbf{B})$ .

In (21),  $\mathbf{B}$  was assumed to be fixed and known. When  $\mathbf{B}$  is unknown, we replace  $\mathbf{B}$  by its marginal maximum likelihood estimate  $\hat{\mathbf{B}}$ . Since (21) holds regardless of the true value of  $\mathbf{B}$ , substituting  $\hat{\mathbf{B}}$  for  $\mathbf{B}$  in (21) will involve an order  $O(1/\sqrt{N})$  of error where  $N = \sum_{i=1}^n n_i$ . Compared to each single  $n_i$ , this pooling of  $n_i$ 's is expected to reduce the error significantly so that  $\tilde{C}_i(\hat{\mathbf{B}})$  is sufficiently close to  $\tilde{C}_i(\mathbf{B})$  to satisfy the lower bound of  $1 - \alpha$  in (21).

## 5. A simulation study

### 5.1 Simulation setup

We considered a simulation setting using a subset of parameter configurations from Wang and Fuller (2003).

Each sample in the simulation study was generated from the following steps: First, generate observations using the model

$$X_{ij} = \beta + u_i + e_{ij},$$

where  $u_i \sim N(0, \tau^2)$  and  $e_{ij} \sim N(0, n_i \sigma_i^2)$ , independently for  $j = 1, \dots, n_i$  and  $i = 1, \dots, n$ . Then, the random effects model for the small area mean,  $X_i$ , is

$$X_i = \beta + u_i + e_i, \text{ independently for } i = 1, \dots, n,$$

where  $X_i \equiv \bar{X}_i \equiv n_i^{-1} \sum_{j=1}^{n_i} X_{ij}$  and  $e_i \equiv \bar{e}_i \equiv n_i^{-1} \sum_{j=1}^{n_i} e_{ij}$ . Therefore,  $X_i \sim N(\theta_i, \sigma_i^2)$  where  $\theta_i = \beta + u_i$ ,  $\theta_i \sim N(\beta, \tau^2)$  and  $e_i \sim N(0, \sigma_i^2)$ . We estimated  $\sigma_i^2$  with the unbiased estimator

$$S_i^2 = (n_i - 1)^{-1} n_i^{-1} \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)^2,$$

and it follows that  $(n_i - 1)S_i^2/\sigma_i^2 \sim \chi_{n_i-1}^2$ , independently for  $i = 1, 2, \dots, n$ . Note that the simulation layout has ignored the second level modeling of sampling variances in (2). Thus, our result will indicate robustness with respect to the variance model misspecification.

The above steps produced the data  $(X_i, S_i^2), i = 1, \dots, n$ . To simplify the simulation, we do not choose any covariate information  $\mathbf{Z}_i$ . Similar to Wang and Fuller (2003), we set all  $n_i$ 's equal to  $m$  to ease programming efforts. However, the true sampling variances are still chosen to be unequal: One-third of the  $\sigma_i^2$  are set to 1, another one-third are set to 4, and the remaining one-third are set to 16. We take  $\beta = 10$  and three different choices of  $\tau^2 = 0.25, 1$  and 4. These parameter values are chosen from Qiu and Hwang (2007). For each of  $\tau^2$ , we generated 200 samples for the two combinations  $(m, n) = (9, 36)$  and  $(18, 180)$ .

In the simulation study, we compare the proposed method with the methods of Wang and Fuller (2003), Hwang *et al.* (2009) and Qiu and Hwang (2007) which are referred to as Methods I, II, III, and IV, respectively, based on bias, mean squared error (MSE), coverage probability (CP) of the confidence intervals and the length of the confidence intervals (ALCI). Table 1 contains the parameter estimates for  $a, b, \beta$  and  $\tau^2$ . The numerical results indicate good performance of the maximum likelihood estimates for the model parameters; the estimated values of  $\beta$  and  $\tau^2$  are close to the true values indicating good robustness properties with respect to distributional misspecification in the second level of (2). Statistically significant estimates for both  $a$  and  $b$  indicate that ‘‘shrunk’’ sampling variances are incorporated in the proposed method. Tables 2, 3 and 4 provide numerical results averaged over areas within each group having the same true sampling variances. The results in the Tables are based on 200 replications.

**Table 1**  
Simulation results for the model parameters,  $a$  (top left panel),  $b$  (top right panel),  $\beta$  (bottom left panel) and  $\tau^2$  (bottom right panel). Here SD represents the standard deviation over 200 replicates. We took  $\beta = 10$  and  $\tau^2 = 0.25, 1$  and  $4$

$\tau^2$	$n = 36, m = 9$		$n = 180, m = 18$		$\tau^2$	$n = 36, m = 9$		$n = 180, m = 18$	
	Mean	SD	Mean	SD		Mean	SD	Mean	SD
	$a$					$b$			
0.25	1.0959	0.1540	1.0328	0.0442	0.25	0.3992	0.0983	0.4249	0.0323
1	1.0937	0.1555	1.0325	0.0445	1	0.4030	0.1012	0.4253	0.0326
4	1.0996	0.1577	1.0339	0.0450	4	0.3999	0.1017	0.4245	0.0328
	$\beta$					$\tau^2$			
0.25	10.0071	0.3618	9.9951	0.1853	0.25	0.2558	0.0605	0.2575	0.0097
1	10.0142	0.3311	9.9970	0.1743	1	0.9418	0.3333	1.0426	0.1264
4	10.0282	0.4639	10.0048	0.2254	4	3.5592	1.3316	4.0817	0.5551

**Table 2**  
Simulation results for prediction when  $\tau^2 = 0.25$ . Here MSE, ALCI, CP represent the mean squared error, average confidence interval width, and coverage probability, respectively

	$\sigma_i^2$	$n = 36, m = 9$				$n = 180, m = 18$			
		Method				Method			
		I	II	III	IV	I	II	III	IV
Relative bias	1	0.0048	0.0198	0.0272	0.0018	-0.0051	-0.0086	-0.0112	-0.0111
	4	-0.0033	-0.0061	-0.0145	-0.0158	-0.0130	-0.0109	-0.0065	-0.0116
	16	0.0126	0.0370	0.0369	0.0096	-0.0046	-0.0045	-0.0080	-0.0061
MSE	1	0.3066	0.3890	0.6861	0.3805	0.2258	0.2680	0.4470	0.2922
	4	0.3281	0.5430	1.3778	0.7285	0.2595	0.3000	0.5805	0.3748
	16	0.3715	0.5240	1.6749	1.9316	0.2815	0.2850	0.4856	0.6383
ALCI	1	2.1393	2.5485	4.4906	3.0528	1.9220	1.6006	3.6466	2.4811
	4	2.2632	3.9574	6.8887	5.6842	2.0557	2.1524	5.2472	4.2160
	16	2.3221	4.5619	9.3335	11.1363	2.1046	2.3308	6.5273	7.8492
CP	1	0.9468	0.9770	0.9771	0.9708	0.9564	0.9710	0.9851	0.9631
	4	0.9468	0.9710	0.9829	0.9917	0.9555	0.9660	0.9967	0.9967
	16	0.9365	0.9660	0.9933	0.9975	0.9529	0.9610	0.9998	0.9999

**Table 3**  
Simulation results for prediction when  $\tau^2 = 1$ . Here MSE, ALCI, CP represent the mean squared error, average confidence interval width and coverage probability, respectively

	$\sigma_i^2$	$n = 36, m = 9$				$n = 180, m = 18$			
		Method				Method			
		I	II	III	IV	I	II	III	IV
Relative bias	1	-0.0152	0.0205	0.0255	0.0051	-0.0064	-0.0085	-0.0111	-0.0101
	4	-0.0167	-0.0164	-0.0151	-0.0219	-0.0151	-0.0121	-0.0133	-0.0164
	16	-0.0323	0.0508	0.0515	0.0216	-0.0028	-0.0017	-0.0073	-0.0039
MSE	1	0.5645	0.6330	0.7238	0.6260	0.5288	0.5430	0.5673	0.6336
	4	0.8566	1.1100	1.5396	1.0992	0.8159	0.8770	0.9415	0.8948
	16	1.0482	1.3100	2.1059	2.3156	0.9786	1.0000	1.1024	1.1878
ALCI	1	3.4550	3.1822	4.4938	3.2117	3.1088	2.5094	3.6763	2.8676
	4	4.0321	5.8733	6.8984	5.7909	3.7844	4.2908	5.3323	4.5543
	16	4.4082	7.4286	9.3555	11.1555	4.1187	5.1590	6.6785	7.8937
CP	1	0.9704	0.9640	0.9762	0.9275	0.9660	0.9650	0.9786	0.8879
	4	0.9633	0.9560	0.9812	0.9808	0.9627	0.9680	0.9918	0.9740
	16	0.9533	0.9490	0.9912	0.9938	0.9613	0.9680	0.9974	0.9979

**Table 4**  
**Simulation results for prediction when  $\tau^2 = 4$ . Here MSE, ALCI, CP represent the mean squared error, average confidence interval length and the coverage probability, respectively**

	$\sigma_i^2$	$n = 36, m = 9$				$n = 180, m = 18$			
		Method				Method			
		I	II	III	IV	I	II	III	IV
Relative bias	1	-0.0024	0.0248	0.0229	0.0180	-0.0084	-0.0098	-0.0122	-0.0106
	4	-0.0343	-0.0310	-0.0210	-0.0340	-0.0110	-0.0092	-0.0174	-0.0132
	16	-0.0147	0.0702	0.0767	0.0467	0.0016	0.0024	-0.0059	0.0012
MSE	1	0.8822	0.8590	0.8579	1.0559	0.8359	0.8180	0.8541	0.8605
	4	2.0577	2.2900	2.1818	2.2422	2.0424	2.1000	2.0935	2.1130
	16	3.4516	3.7600	3.9267	3.8981	3.3153	3.3500	3.3939	3.3631
ALCI	1	4.6318	4.1936	4.5369	3.7677	4.0256	3.5346	3.9626	3.7499
	4	6.2015	10.9093	7.0376	6.4314	5.9000	9.0913	6.2217	6.1540
	16	7.7221	18.0039	9.6718	11.3341	7.4430	14.6665	8.3908	8.7537
CP	1	0.9791	0.9670	0.9733	0.9029	0.9674	0.9570	0.9600	0.9468
	4	0.9556	0.9670	0.9725	0.9496	0.9592	0.9610	0.9633	0.9573
	16	0.9510	0.9670	0.9796	0.9858	0.9573	0.9650	0.9718	0.9776

*Bias Comparisons:* In most cases, the bias of the four methods are comparable. There is no clear evidence of significant differences between them in terms of the bias. High sampling variance gives more weight to the population mean by construction that makes the estimator closer to the mean at the second level. On the other hand, Methods I - III use shrinkage estimators of the sampling variances which would be less than the maximum of all sampling variances. Thus, Methods I - III tend to have little more bias. However, due to shrinkage in sampling variances, one may expect a gain in the variance of the estimators which, in turn, makes the MSE smaller. Among Methods I - III, Method I performed better compared to Methods II and III, which were quite similar to each other. The maximum gain using Method I compared to Method II is 99%.

*MSE Comparisons:* In terms of the MSE, Method I performed consistently better than the other three in all cases except when the ratio of  $\sigma_i^2$  to  $\tau^2$  is the lowest:  $(\sigma_i^2 = 1) / (\tau^2 = 4) = 0.25$ . In this case, the variance between small areas (model variance) is much higher than the variance within the areas (sampling variance). When using our method to estimate  $\theta_i$ , the information “borrowed” from other areas may misdirect the estimation: The estimated mean of the Gamma distribution for  $\sigma_i^{-2}$  from the second level in (2) is  $\hat{a}\hat{b}$  which equals 0.44 approximately for both the  $(m, n)$  combinations of (9, 36) and (18, 180) (the true value is  $ab = 0.4$ ). Thus,  $E(\sigma_i^{-2} | X_i, S_i^2, \hat{B})$  is significantly smaller than 1 due to shrinkage towards the mean for the group which has the true value of  $\sigma_i^2 = 1$ . Also, since  $\sigma_i^2$  is smaller than  $\tau^2$ , the weight of  $X_i$  should be much more compared to  $\beta$ , the overall mean. However,

due to underestimation of  $\sigma_i^{-2}$  in this case, the resulting estimator puts less weight on  $X_i$  which leads to higher MSE. However, this underestimation will decrease for large sample sizes due to the consistency of Bayes estimators. This fact is actually observed when the sample size increases from  $n = 36$  to  $n = 180$  for the case  $\sigma_i^2 = 1$  and  $\tau^2 = 4$ . Compared to Method II, Method I shows gains in most of the simulation cases; the maximum gain is 30% while the only loss is 9% for the combination  $\sigma_i^2 = 1$  and  $\tau^2 = 4$  for  $n = 36$  and  $m = 9$ . Similarly, for Method III, the maximum gain of Method I is 77% and the only loss of 11% is for the same parameter and sample size specifications.

*ACP Comparisons:* We obtained confidence intervals with confidence level 95%. Methods I and III do not indicate any under-coverage. This is expected from their optimal confidence interval construction. Method I meets the nominal coverage rate more frequently than any other methods. Method II has some under coverage and can go as low as 82%.

*ALCI Comparisons:* Method I produced considerably shorter confidence intervals in general. Method IV produced comparable lengths as the other methods in all cases except when  $\sigma_i^2$  was high, in which case, the lengths were considerably higher. The confidence interval proposed in Qiu and Hwang (2007) does not have good finite sample properties, particularly for small  $\tau^2$ . To avoid low coverage, they proposed to truncate  $M_0 = \tau^2 / (\tau^2 + \sigma_i^2)$  with a positive number  $M_1 = 1 - Q_\alpha / (v - 2)$  for known  $\sigma_i^2$  where  $Q_\alpha$  is the  $\alpha^{\text{th}}$ -quantile of a chi-squared distribution with  $v$  degrees of freedom. When the ratio of sampling

variance to model variance,  $\sigma_i^2 / \tau^2$ , is high,  $M_1$  tends to be higher than  $M_0$ . This results in a nominal coverage but with larger interval lengths. For example, in case of  $(\sigma_i^2, \tau^2) = (16, 0.25)$ , the ALCI is 11.13 for Method IV whereas ALCI is only 2.78 and 4.56 for Methods I and II.

### 5.2 Robustness study

In order to study the robustness of the proposed method with respect to departures from the normality assumption in the errors, we conducted the following simulation study. Data was generated as before but with  $e_{ij}$ 's drawn from a double-exponential (Laplace) and an uniform distribution. The estimators from Methods II and III had little effect. This is perhaps due to the fact that these methods used moment based estimation for model parameter estimation. Method IV resulted in larger relative bias, MSE and ALCI, and lower coverage probability. The MSE from Method I is always lower than that from Method II. For  $\tau^2 = 0.25$  and 1, ALCI is smaller for Method I compared to Method II for  $(n = 36, m = 9)$  but the results are opposite when  $(n = 180, m = 18)$ . In terms of CP, Method II has some under coverage (lowest is 80%). However, Method I did not have any under-coverage. In order to save space we only provide

the results for parameters  $a, b, \beta$  and  $\tau^2$  under the Laplace errors (see Table 5).

## 6. Real data analysis

We illustrate our methodology based on a widely studied example. The data set is from the U.S. Department of Agriculture and was first analyzed by Battese (1988). The data set is on corn and soybeans productions in 12 Iowa counties. The sample sizes for these areas are small, ranging from 1 to 5. We shall consider corn only to save space. For the proposed model, the sample sizes  $n_i > 1$  necessarily. Therefore, modified data from You and Chapman (2006) with  $n_i \geq 2$  are used. The mean reported crop hectares for corn ( $X_i$ ) are the direct survey estimates and are given in Table 6. Table 6 also gives the sample variances which are calculated based on the original data assuming simple random sampling. The sample standard deviation varies widely, ranging from 5.704 to 53.999 (the coefficient of variation varies from 0.036 to 0.423). Two covariates are considered in Table 6:  $Z_{i1}$ , the mean of pixels of corn, and  $Z_{i2}$ , the mean of pixels of soybean, from the LANDSAT satellite data.

**Table 5**  
Simulation results for the model parameters,  $a$  (top left panel),  $b$  (top right panel),  $\beta$  (bottom left panel) and  $\tau^2$  (bottom right panel) when the errors follow a laplace distribution. Here SD represents the standard deviation over 200 replicates. We took  $\beta = 10$  and  $\tau^2 = 0.25, 1$  and  $4$

$\tau^2$	$n = 36, m = 9$		$n = 180, m = 18$		$\tau^2$	$n = 36, m = 9$		$n = 180, m = 18$	
	Mean	SD	Mean	SD		Mean	SD	Mean	SD
$a$					$b$				
0.25	0.9624	0.1632	0.9471	0.0498	0.25	0.5793	0.1733	0.5279	0.0501
1	0.9628	0.1657	0.9476	0.0497	1	0.5816	0.1777	0.5275	0.0503
4	0.9689	0.1694	0.9487	0.0499	4	0.5758	0.1796	0.5263	0.0503
$\beta$					$\tau^2$				
0.25	9.9736	0.3775	9.9800	0.1773	0.25	0.2696	0.0882	0.2565	0.0074
1	9.9753	0.3709	9.9836	0.1662	1	1.0508	0.2501	1.0403	0.0668
4	9.9736	0.4835	9.9855	0.2161	4	3.9624	1.1719	4.1256	0.4201

**Table 6**  
Corn data from You and Chapman (2006)

County	$n_i$	$X_i$	$Z_{i1}$	$Z_{i2}$	$\sqrt{S_i^2}$
Franklin	3	158.623	318.21	188.06	5.704
Pocahontas	3	102.523	257.17	247.13	43.406
Winnebago	3	112.773	291.77	185.37	30.547
Wright	3	144.297	301.26	221.36	53.999
Webster	4	117.595	262.17	247.09	21.298
Hancock	5	109.382	314.28	198.66	15.661
Kossuth	5	110.252	298.65	204.61	12.112
Hardin	5	120.054	325.99	177.05	36.807

The estimates of  $B$  are as follows:  $a = 1.707$ ,  $b = 0.00135$ ,  $\tau^2 = 90.58$  and  $\beta = (-186.0, 0.7505, 0.4100)$ . The estimated prior mean of  $1/\sigma_i^2$  which is the mean of the Gamma distribution with parameters  $a$  and  $b$  is  $ab = 0.002295$  with a square root of 0.048 (note that  $1/0.048 = 20.85$  consistent with the range of the sample standard deviations between 5.704 and 53.999). The small area estimates and their confidence intervals are summarized in Table 7 and Figure 1. Point estimates of all 4 methods are comparable: the summary measures comprising of the mean, median, and range of the small area parameter estimates for Methods I, II, III, and IV are (121.9, 124.1, 122.2, 122.6), (125.2, 120.4, 115.0, 114.5) and (23.1, 53.0, 58.4, 56.6), respectively. The distribution of  $\hat{\theta}_i$  (plotted based on considering all the  $i$ 's) are summarized in Figure 2 which shows that there is a significant difference in their variability. Method I has the lowest variability and is superior in this sense. Further, smoothing sampling variances has strong implication in measuring uncertainty and hence in the interval estimation. The proposed method has the shortest confidence interval on an average compared to all other methods. Methods II and III provide intervals with negative lower limits. This seems unrealistic because the direct average of area under corn is positive and large for all the 12 counties (the crude confidence intervals  $(x_i \pm t_{0.025} S_i)$  do not contain zero for any of the areas either). Note that Method II does not have any theoretical support on its confidence intervals. Methods II and III produce wider confidence intervals when the sampling variance is high. For example, the sample size for both Franklin county

and Pocahontas county is three, but sampling standard deviations are 5.704 and 43.406. Although the confidence interval under Method I is comparable, they are wide apart for Methods II and III. This is because although these methods consider the uncertainty in sampling variance estimates, the smoothing did not use the information from direct survey estimates, resulted the underlying sampling variance estimates remain highly variable (due to small sample size). In effect, the variance of the variance estimator (of the point estimates) is bigger compared to that in method I. This is further confirmed by the fact that the intuitive standard deviations of the “smoothed” small area estimates (one fourth of the interval) are smaller and less variable under method I compared to the others. Another noticeable aspect of our method is that the interval widths are similar for counties with same sample size. This could be an indication of obtaining equ-efficient estimators for equivalent sample sizes.

*Model selection:* For choosing the best fitting model, we used the Bayesian Information Criteria (BIC) which takes into account both the likelihood as well as the complexity of the fitted models. We calculated BICs for the models used in Methods I and III (Hwang *et al.* 2009). These two models have the same numbers of parameters with a difference in only the way the parameters are estimated. The model BIC for Method I is 210.025 and that for Method III is 227.372. This indicates superiority of our model. We could not compute the BIC for Wang and Fuller (2003) since they did not use any explicit likelihood.

**Table 7**  
**Results of the corn data analysis. Here CI and LCI represent the confidence interval and the length of the confidence interval, respectively**

County	$\hat{\theta}_i$	CI	LCI	$\hat{\theta}_i$	CI	LCI
		I: Proposed method			II: Wang and Fuller (2003)	
Franklin	131.8106	104.085, 159.372	55.287	155.4338	124.151, 193.094	68.943
Pocahontas	108.7305	80.900, 136.436	55.536	102.3682	-38.973, 244.019	282.993
Winnebago	109.0559	81.430, 136.646	55.216	115.9093	-53.768, 279.314	333.083
Wright	131.6113	103.736, 159.564	55.828	131.0674	8.330, 280.263	271.932
Webster	113.1484	92.805, 133.348	40.543	109.4795	32.514, 202.675	170.161
Hancock	129.4279	111.781, 147.193	35.412	124.1028	56.750, 162.013	105.262
Kossuth	121.0071	103.451, 138.626	35.175	116.7147	68.049, 152.454	84.405
Hardin	130.2520	112.373, 148.114	35.741	137.7983	51.734, 188.373	136.638
		III: Hwang <i>et al.</i> (2009)			IV: Qiu and Hwang (2007)	
Franklin	158.4677	128.564, 188.370	59.805	157.7383	146.999, 168.477	21.478
Pocahontas	100.1276	-44.039, 244.295	288.334	101.1661	19.444, 182.887	163.442
Winnebago	114.1473	0.065, 228.228	228.163	113.7746	56.263, 171.286	115.022
Wright	140.3717	-24.119, 304.862	328.982	143.2244	41.559, 244.889	203.330
Webster	115.7865	50.297, 181.275	130.978	115.2224	75.124, 155.320	80.196
Hancock	111.3087	66.213, 156.403	90.189	113.1766	83.691, 142.661	58.970
Kossuth	110.9585	74.366, 147.550	73.184	112.3239	89.520, 135.127	45.607
Hardin	126.6093	40.040, 213.178	173.137	123.9049	54.607, 193.202	138.594

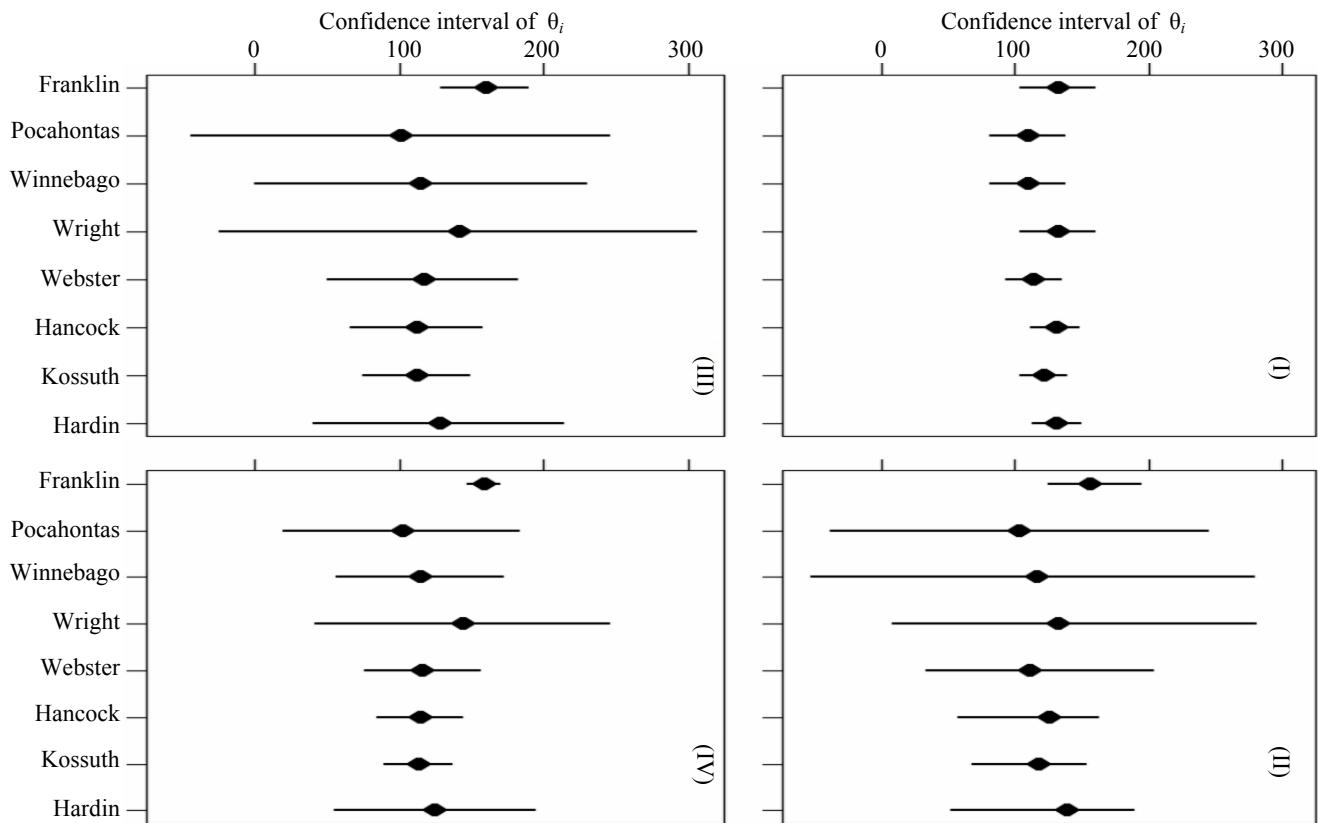


Figure 1 Corn hectares estimation. The horizontal line for each county displays the confidence interval of  $\hat{\theta}_i$ , with  $\hat{\theta}_i$  marked by the circle, for (I) Proposed method, (II) Wang and Fuller (2003), (III) Hwang *et al.* (2009) and (IV) Qiu and Hwang (2007)

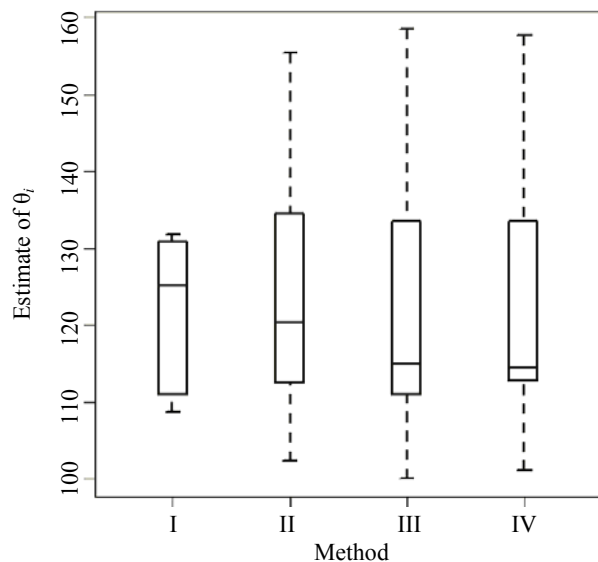


Figure 2 Boxplot of estimates of corn hectares for each county. (I) to (IV) are the 4 methods corresponding to Figure 1

### 7. Conclusion

In this paper, joint area level modeling of means and variances is developed for small area estimation. The resulting small area estimators are shown to be more efficient than the traditional estimators obtained using Fay-Herriot models which only shrink the means. Although our model is same as one considered in Hwang *et al.* (2009), our method of estimation is different in two ways: In the determination of the tuning parameter  $k$  and the use of  $\pi(\sigma_i^2 | X_i, S_i^2, \mathbf{Z}_i)$  (which depends additionally on  $X_i$ ), instead of  $\pi(\sigma_i^2 | S_i^2, \mathbf{Z}_i)$ , for constructing the conditional distribution of the small area parameters  $\theta_i$ . We demonstrated robustness properties of the model when the assumption that  $\sigma_i^2$  arise from an inverse Gamma distribution is violated. The borrowing of  $X_i$  information when estimating  $\sigma_i^2$  as well as the robustness with respect to prior elicitation demonstrate the superiority of our proposed method. The parameter values chosen in the simulation study are different than in the real data analysis. The real data analysis given here is merely for illustration purposes. Our main aim was to

develop the methodology of mean-variance modeling and contrast with some closely related methods to show its effectiveness. For this reason, we chose parameter settings in the simulation to be the same as in the well-known small area estimation article Wang and Fuller (2003).

Obtaining improved sampling variance estimators is a byproduct of the proposed approach. We have provided an innovative estimation technique which is theoretically justified and user friendly. Computationally, the method is much simpler compared to other competitive methods such as Bayesian MCMC procedures or bootstrap resampling methods. We need sampling from posterior distribution only once during the model parameter estimation, and the sampled values can be used subsequently for all other purposes. The software is available from the authors upon request.

### Acknowledgements

The authors like to thank two referees and the Associate Editor for their constructive comments that have led to a significantly improved version of this article. The research is partially supported by NSF grants SES 0961649, 0961618 and DMS 1106450.

### Appendix

#### A. Derivation of the conditional distributions

From Equation (1) and (2), the conditional joint distribution of  $\{X_i, S_i^2, \theta_i, \sigma_i^2\}$ ,  $\pi(X_i, S_i^2, \theta_i, \sigma_i^2 | a, b, \boldsymbol{\beta}, \tau^2)$ , is

$$\begin{aligned} \pi(X_i, S_i^2, \theta_i, \sigma_i^2 | \mathbf{Z}_i, \mathbf{B}) &= \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left\{-\frac{(X_i - \theta_i)^2}{2\sigma_i^2}\right\} \frac{1}{\Gamma\left(\frac{n_i-1}{2}\right) 2^{\frac{n_i-1}{2}}} \\ &\times \left\{(n_i-1) \frac{S_i^2}{\sigma_i^2}\right\}^{\frac{n_i-1}{2}-1} \exp\left\{-\frac{(n_i-1)S_i^2}{2\sigma_i^2}\right\} \\ &\times \left(\frac{n_i-1}{\sigma_i^2}\right) \frac{1}{\sqrt{2\pi\tau^2}} \exp\left\{-\frac{(\theta_i - \mathbf{Z}_i^T \boldsymbol{\beta})^2}{2\tau^2}\right\} \\ &\times \frac{1}{\Gamma(a)b^a} \left(\frac{1}{\sigma_i^2}\right)^{a+1} \exp\left(-\frac{1}{b\sigma_i^2}\right) \\ &\propto \exp\left[-\frac{(\theta_i - \mathbf{Z}_i^T \boldsymbol{\beta})^2}{2\tau^2} - \left\{\frac{(X_i - \theta_i)^2}{2} + \frac{(n_i-1)S_i^2}{2} + \frac{1}{b}\right\} \frac{1}{\sigma_i^2}\right] \\ &\times \left(\frac{1}{\sigma_i^2}\right)^{\frac{n_i+a+1}{2}} \left(\frac{1}{\tau^2}\right)^{\frac{1}{2}} \frac{1}{\Gamma(a)b^a}. \end{aligned}$$

Therefore the conditional distribution of  $\sigma_i^2$  and  $\theta_i$  given the data and  $\mathbf{B}$  are

$$\begin{aligned} \pi(\sigma_i^2 | X_i, S_i^2, \mathbf{Z}_i, \mathbf{B}) &= \int \pi(X_i, S_i^2, \theta_i, \sigma_i^2 | \mathbf{Z}_i, \mathbf{B}) d\theta_i \propto \frac{1}{(\sigma_i^2)^{(n_i-1)/2+a+1} (\sigma_i^2 + \tau^2)^{1/2}} \\ &\exp\left[-\frac{(X_i - \mathbf{Z}_i^T \boldsymbol{\beta})^2}{2(\sigma_i^2 + \tau^2)} - \left\{\frac{1}{2}(n_i-1)S_i^2 + \frac{1}{b}\right\} \left(\frac{1}{\sigma_i^2}\right)\right], \end{aligned}$$

$$\begin{aligned} \pi(\theta_i | X_i, S_i^2, \mathbf{Z}_i, \mathbf{B}) &= \int \pi(X_i, S_i^2, \theta_i, \sigma_i^2 | \mathbf{Z}_i, \mathbf{B}) d\sigma_i^2 \\ &\propto \exp\left\{-\frac{(\theta_i - \mathbf{Z}_i^T \boldsymbol{\beta})^2}{2\tau^2}\right\} \psi_i^{-\left(\frac{n_i}{2}+a\right)} \end{aligned}$$

where  $\psi_i$  is defined in Equation (4).

#### B. Details of the EM algorithm

The maximization of  $Q(\mathbf{B} | \mathbf{B}^{(t-1)})$  is done by setting the partial derivatives with respect to  $\mathbf{B}$  to be zero, that is,

$$\frac{\partial Q(\mathbf{B} | \mathbf{B}^{(t-1)})}{\partial \mathbf{B}} = 0. \tag{B.1}$$

From the expression of  $Q(\mathbf{B} | \mathbf{B}^{(t-1)})$  in the text, we give explicit expressions for the partial derivatives with respect to each component of  $\mathbf{B}$ . The partial derivative corresponding to  $\boldsymbol{\beta}$  is

$$\begin{aligned} \frac{\partial Q(\mathbf{B} | \mathbf{B}^{(t-1)})}{\partial \boldsymbol{\beta}} &= \sum_{i=1}^n \frac{\int \mathbf{Z}_i \left(\frac{\theta_i - \mathbf{Z}_i^T \boldsymbol{\beta}}{\tau^2}\right) \exp\left\{-\frac{(\theta_i - \mathbf{Z}_i^T \boldsymbol{\beta})^2}{2\tau^2}\right\} \psi_i^{-\left(\frac{n_i}{2}+a\right)} d\theta_i}{\int \exp\left\{-\frac{(\theta_i - \mathbf{Z}_i^T \boldsymbol{\beta})^2}{2\tau^2}\right\} \psi_i^{-\left(\frac{n_i}{2}+a\right)} d\theta_i} \\ &= \sum_{i=1}^n E\left\{\mathbf{Z}_i \left(\frac{\theta_i - \mathbf{Z}_i^T \boldsymbol{\beta}}{\tau^2}\right)\right\} \end{aligned}$$

where the expectation is with respect to the conditional distribution of  $\theta_i$ ,  $\pi(\theta_i | X_i, S_i^2, \mathbf{B})$ . The expression of the partial derivative corresponding to  $\tau^2$  is:

$$\begin{aligned} \frac{\partial Q(\mathbf{B} | \mathbf{B}^{(t-1)})}{\partial \tau^2} &= -\frac{n}{2\tau^2} + \sum_{i=1}^n \frac{\int \frac{(\theta_i - \mathbf{Z}_i^T \boldsymbol{\beta})^2}{2(\tau^2)^2} \exp\left\{-\frac{(\theta_i - \mathbf{Z}_i^T \boldsymbol{\beta})^2}{2\tau^2}\right\} \psi_i^{-\left(\frac{n_i}{2}+a\right)} d\theta_i}{\int \exp\left\{-\frac{(\theta_i - \mathbf{Z}_i^T \boldsymbol{\beta})^2}{2\tau^2}\right\} \psi_i^{-\left(\frac{n_i}{2}+a\right)} d\theta_i} \\ &= -\frac{n}{2\tau^2} + \sum_{i=1}^n E\left\{\frac{(\theta_i - \mathbf{Z}_i^T \boldsymbol{\beta})^2}{2(\tau^2)^2}\right\}. \end{aligned}$$

Similarly for  $a$  and  $b$ , we get the solutions by setting  $S_a = 0$  and  $S_b = 0$  where  $S_a$  and  $S_b$  are, respectively, the partial derivatives of  $Q(\mathbf{B} | \mathbf{B}^{(t-1)})$  with respect to  $a$  and  $b$  with expressions given in the main text. These equations are solved using the Newton-Raphson method which requires the matrix of second derivatives with respect to  $a$  and  $b$ . These are given by the following expressions:

$$S_{aa} = \sum_{i=1}^n \left[ \log'' \left\{ \Gamma \left( \frac{n_i}{2} + a \right) \right\} - \log'' \{ \Gamma(a) \} + \text{Var} \{ \log(\psi_i) \} \right]$$

$$S_{ab} = \sum_{i=1}^n \left[ -\frac{1}{b} + \frac{1}{b^2} E \left( \frac{1}{\psi_i} \right) - \left( \frac{n_i}{2} + a \right) \frac{1}{b^2} \text{Cov} \left\{ \frac{1}{\psi_i}, \log(\psi_i) \right\} \right],$$

and

$$S_{bb} = \sum_{i=1}^n \left\{ \frac{a}{b^2} - (n_i + 2a) \frac{1}{b^3} E \left( \frac{1}{\psi_i} \right) + \left( \frac{n_i}{2} + a \right) \frac{1}{b^4} E \left( \frac{1}{\psi_i^2} \right) + \left( \frac{n_i}{2} + a \right)^2 \frac{1}{b^4} \text{Var} \left( \frac{1}{\psi_i} \right) \right\}$$

with  $S_{ba} = S_{ab}$ . At the  $u^{\text{th}}$  step, the update of  $a$  and  $b$  are given by

$$\begin{bmatrix} a^{(u)} \\ b^{(u)} \end{bmatrix} = \begin{bmatrix} a^{(u-1)} \\ b^{(u-1)} \end{bmatrix} - \begin{bmatrix} S_{aa}^{(u-1)} & S_{ab}^{(u-1)} \\ S_{ba}^{(u-1)} & S_{bb}^{(u-1)} \end{bmatrix}^{-1} \begin{bmatrix} S_a^{(u-1)} \\ S_b^{(u-1)} \end{bmatrix}, \tag{B.3}$$

where the superscript  $(u - 1)$  on  $S_{aa}$ ,  $S_{ab}$ ,  $S_{ba}$ ,  $S_{bb}$ ,  $S_a$  and  $S_b$  denote these quantities evaluated at the values of  $a$  and  $b$  at the  $(u - 1)^{\text{th}}$  iteration. Once the Newton Raphson procedure converges, the value of  $a$  and  $b$  at the  $t^{\text{th}}$  step of the EM algorithm is set as  $a^{(t)} = a^{(\infty)}$  and  $b^{(t)} = b^{(\infty)}$ .

**C. An alternative small area model formulation**

It is possible to reduce the width of the confidence interval  $\tilde{C}(\mathbf{B})$  based on an alternative hierarchical model for small area estimation which has some mathematical elegance. The constant term  $n_i + 2a + 2$  in (19) becomes  $n_i + 2a$  in this alternative model formulation. The model is given by

$$X_i | \theta_i, \sigma_i^2 \sim N(\theta_i, \sigma_i^2), \tag{C.1}$$

$$\theta_i | \sigma_i^2 \sim N(\mathbf{Z}_i \boldsymbol{\beta}, \lambda \sigma_i^2), \tag{C.2}$$

$$\frac{(n_i - 1) S_i^2}{\sigma_i^2} \Big| \sigma_i^2 \sim \chi_{n_i - 1}^2, \tag{C.3}$$

$$\sigma_i^2 \sim \text{Inverse - Gamma}(a, b), \tag{C.4}$$

independently for  $i = 1, 2, \dots, n$ . Note that in the above formulation, it is assumed that the conditional variance of  $\theta_i$  is proportional to  $\sigma_i^2$  whereas the marginal variance is constant (by integrating out  $\sigma_i^2$  using (C.4)). In (1) and (2), the variance of  $\theta_i$  is a constant,  $\tau^2$ , independent of  $\sigma_i^2$ , and there is no conditional structure for  $\theta_i$  depending on  $\sigma_i^2$ . The set of all unknown parameters in the current hierarchical model is  $\mathbf{B} = (a, b, \boldsymbol{\beta}, \lambda)$ . The inference procedure for this model is given subsequently. The model essentially assumes that the true small area effects are not identically distributed even after eliminating the known variations.

**C.1 Inference methodology**

By re-parameterizing the variance as in (C.2), some analytical simplifications are obtained in the derivation of the posteriors of  $\theta_i$  and  $\sigma_i$  given  $X_i, S_i^2$  and  $\mathbf{B}$ . We have

$$\begin{aligned} \pi(\sigma_i^2 | X_i, S_i^2, \mathbf{B}) &= IG \left( \frac{n_i}{2} + a, \left[ \frac{(n_i - 1) S_i^2}{2} + \frac{(X_i - \mathbf{Z}_i \boldsymbol{\beta})^2}{2(1 + \lambda)} + \frac{1}{b} \right]^{-1} \right) \end{aligned}$$

where  $IG(a, b)$  stands for the inverse Gamma distribution with shape and scale parameters  $a$  and  $b$ , respectively. Given  $\mathbf{B}$  and  $\sigma_i^2$ , the conditional distribution of  $\theta_i$  is

$$\pi(\theta_i | X_i, \sigma_i^2, \mathbf{B}) = \text{Normal} \left( \mathbf{Z}_i^T \boldsymbol{\beta}, \frac{\lambda \sigma_i^2}{1 + \lambda} \right).$$

Integrating out  $\sigma_i^2$ , one obtains the conditional distribution of  $\theta_i$  given  $X_i, S_i^2$  and  $\mathbf{B}$ ,

$$\begin{aligned} \pi(\theta_i | X_i, S_i^2, \mathbf{B}) &= \int_0^\infty \pi(\theta_i | X_i, \sigma_i^2, \mathbf{B}) \pi(\sigma_i^2 | X_i, S_i^2, \mathbf{B}) d\sigma_i^2 \\ &\propto \left\{ \frac{(1 + \lambda)}{2\lambda} (\theta_i - \mathbf{Z}_i^T \boldsymbol{\beta})^2 + \frac{\delta^2}{2} \right\}^{-(n_i + 2a + 1)/2}, \tag{C.5} \end{aligned}$$

where  $\delta^2 = (n_i - 1) S_i^2 + (X_i - \mathbf{Z}_i \boldsymbol{\beta})^2 / (1 + \lambda) + 2/b$ . We can rewrite (C.5) as

$$\begin{aligned} \pi(\theta_i | X_i, S_i^2, \mathbf{B}) &= \frac{\Gamma((n_i + 1)/2 + a) \sqrt{1 + \lambda}}{\delta^* \Gamma(n_i/2 + a) \sqrt{(n_i + 2a) \lambda \pi}} \\ &\quad \left\{ 1 + \frac{(\theta_i - \mu_i)^2}{(n_i + 2a) \delta^{*2} \lambda / (1 + \lambda)} \right\}^{-(n_i + 2a + 1)/2} \end{aligned}$$

which can be seen to be a scaled t-distribution with  $n_i + 2a$  degrees of freedom and scale parameter  $\delta^* \sqrt{\lambda / (1 + \lambda)}$  with  $\delta^{*2} = \delta^2 / (n_i + 2a)$ . Hence,

$$\begin{aligned} E(\sigma_i^{-1} | X_i, S_i^2, \mathbf{B}) &= \frac{\Gamma((n_i + 1)/2 + a) (\delta^2 / 2)^{-\{(n_i + 1)/2 + a\}}}{\Gamma(n_i/2 + a) (\delta^2 / 2)^{-(n_i/2 + a)}} \\ &= \frac{\Gamma((n_i + 1)/2 + a)}{\Gamma(n_i/2 + a)} \frac{\sqrt{2}}{\delta^* \sqrt{n_i + 2a}}. \end{aligned}$$



In this context, choosing

$$k = k(\mathbf{B}) = \left\{ 1 + \frac{t_{\alpha/2}^2}{n_i - 1} \right\}^{- (n_i + 2a + 1)/2} \sqrt{\frac{1 + \lambda}{\lambda}} \frac{1}{\sqrt{2\pi}},$$

the confidence interval in (8) simplifies to

$$C_i(\mathbf{B}) \equiv \left\{ \theta_i : \frac{|\theta_i - \mu_i|}{\sqrt{\frac{\lambda}{1 + \lambda} \frac{(n_i + 2a)\delta^{*2}}{n_i - 1}}} \leq t_{\alpha/2} \right\}. \quad (\text{C.6})$$

Using the similar arguments as before and noting that  $(n_i + 2a)\delta^{*2} \geq (n_i - 1)S_i^2$ , we have  $P\{C_i(\mathbf{B})\} \geq P(D_i) = 1 - \alpha$  where  $D_i$  is the confidence interval in (20). When  $\mathbf{B}$  is unknown, we replace  $\mathbf{B}$  by its marginal maximum likelihood estimate  $\hat{\mathbf{B}}$ . It is expected that the pooling technique will result in an error small enough so that  $P\{C_i(\hat{\mathbf{B}})\} \approx P\{C_i(\mathbf{B})\} \geq 1 - \alpha$ .

## References

- Battese, G.E., Harter, R.M. and Fuller, W.A. (1988). An error component model for prediction of county crop areas using survey and satellite data. *Journal of the American Statistical Association*, 95, 28-36.
- Bell, W. (2008). Examining sensitivity of small area inferences to uncertainty about sampling error variances. Technical Report U.S. Census Bureau.
- Casella, G., and Hwang, J. (1991). Evaluating confidence sets using loss functions. *Statistica Sinica*, 1, 159-173.
- Chatterjee, S., Lahiri, P. and Li, H. (2008). Parametric bootstrap approximation to the distribution of EBLUP and related prediction intervals in linear mixed models. *Annals of Statistics*, 36, 1221-1245.
- Cho, M., Eltinge, J., Gershunskaya, J. and Huff, L. (2002). Evaluation of generalized variance function estimators for the U.S. current employment survey. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 534-539.
- Fay, R., and Herriot, R. (1979). Estimates of income for small places: An application of James-Stein procedures to census data. *Journal of the American Statistical Association*, 74, 269-277.
- Gershunskaya, J., and Lahiri, P. (2005). Variance estimation for domains in the U.S. current employment statistics program. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 3044-3051.
- Ghosh, M., and Rao, J. (1994). Small area estimation: An appraisal. *Statistical Science*, 9, 54-76.
- Hall, P., and Maiti, T. (2006). Nonparametric estimation of mean squared prediction error in nested-error regression models. *Annals of Statistics*, 34, 1733-1750.
- Huff, L., Eltinge, J. and Gershunskaya, J. (2002). Exploratory analysis of generalized variance function models for the U.S. current employment survey. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 1519-1524.
- Hwang, J., Qiu, J. and Zhao, Z. (2009). Empirical Bayes confidence intervals shrinking both mean and variances. *Journal of the Royal Statistical Society*, B, 71, 265-285.
- Joshi, V. (1969). Admissibility of the usual confidence sets for the mean of a univariate or bivariate normal population. *The Annals of Mathematical Statistics*, 40, 1042-1067.
- Maples, J., Bell, W. and Huang, E. (2009). Small area variance modeling with application to county poverty estimates from the American community survey. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 5056-5067.
- Otto, M., and Bell, W. (1995). Sampling error modelling of poverty and income statistics for states. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 160-165.
- Pfeffermann, D. (2002). Small area estimation - New developments and directions. *International Statistical Review*, 70, 125-143.
- Prasad, N., and Rao, J. (1990). The estimation of the mean squared error of small-area estimators. *Journal of the American Statistical Association*, 85, 163-171.
- Qiu, J., and Hwang, J. (2007). Sharp simultaneous intervals for the means of selected populations with application to microarray data analysis. *Biometrics*, 63, 767-776.
- Rao, J. (2003). Some new developments in small area estimation. *Journal of the Iranian Statistical Society*, 2, 145-169.
- Rivest, L.-P., and Vandal, N. (2003). Mean squared error estimation for small areas when the small area variances are estimated. *Proceedings of the International Conference on Recent Advances in Survey Sampling*.
- Robert, C., and Casella, G. (2004). *Monte Carlo Statistical Methods* (Second edition).
- Valliant, R. (1987). Generalized variance functions in stratified two-stage sampling. *Journal of the American Statistical Association*, 82, 499-508.
- Wang, J., and Fuller, W. (2003). The mean squared error of small area predictors constructed with estimated error variances. *Journal of the American Statistical Association*, 98, 716-723.
- You, Y., and Chapman, B. (2006). Small area estimation using area level models and estimated sampling variances. *Survey Methodology*, 32, 1, 97-103.

**ELECTRONIC  
PUBLICATIONS  
AVAILABLE AT**

**PUBLICATIONS  
ÉLECTRONIQUES  
DISPONIBLE À**

**[www.statcan.gc.ca](http://www.statcan.gc.ca)**

# Condition indexes and variance decompositions for diagnosing collinearity in linear model analysis of survey data

Dan Liao and Richard Valliant<sup>1</sup>

## Abstract

Collinearities among explanatory variables in linear regression models affect estimates from survey data just as they do in non-survey data. Undesirable effects are unnecessarily inflated standard errors, spuriously low or high  $t$ -statistics, and parameter estimates with illogical signs. The available collinearity diagnostics are not generally appropriate for survey data because the variance estimators they incorporate do not properly account for stratification, clustering, and survey weights. In this article, we derive condition indexes and variance decompositions to diagnose collinearity problems in complex survey data. The adapted diagnostics are illustrated with data based on a survey of health characteristics.

Key Words: Diagnostics for survey data; Multicollinearity; Singular value decomposition; Variance inflation.

## 1. Introduction

When predictor variables in a regression model are correlated with each other, this condition is referred to as collinearity. Undesirable side effects of collinearity are unnecessarily high standard errors, spuriously low or high  $t$ -statistics, and parameter estimates with illogical signs or ones that are overly sensitive to small changes in data values. In experimental design, it may be possible to create situations where the explanatory variables are orthogonal to each other, but this is not true with observational data. Belsley (1991) noted that: "... in nonexperimental sciences, ..., collinearity is a natural law in the data set resulting from the uncontrollable operations of the data-generating mechanism and is simply a painful and unavoidable fact of life." In many surveys, variables that are substantially correlated are collected for analysis. Few analysts of survey data have escaped the problem of collinearity in regression estimation, and the presence of this problem encumbers precise statistical explanation of the relationships between predictors and responses.

Although many regression diagnostics have been developed for non-survey data, there are considerably fewer for survey data. The few articles that are available concentrate on identifying influential points and influential groups with abnormal data values or survey weights. Elliot (2007) developed Bayesian methods for weight trimming of linear and generalized linear regression estimators in unequal probability-of-inclusion designs. Li (2007a, b) and Li and Valliant (2009, 2011) extended a series of traditional diagnostic techniques to regression on complex survey data. Their papers cover residuals and leverages, several diagnostics based on case-deletion (DFBETA, DFBETAS, DFFIT, DFFITS, and Cook's Distance), and the forward search approach. Although an extensive literature in applied

statistics provides valuable suggestions and guidelines for data analysts to diagnose the presence of collinearity (e.g., Belsley, Kuh and Welsch 1980; Belsley 1991; Farrar and Glauber 1967; Fox 1986; Theil 1971), almost none of this research touches upon diagnostics for collinearity when fitting models with survey data. One prior, survey-related paper on collinearity problems is (Liao and Valliant 2012) which adapted variance inflation factors for linear models fitted with survey data.

Suppose the underlying structural model in the super-population is  $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$ . The matrix  $\mathbf{X}$  is an  $n \times p$  matrix of predictors with  $n$  being the sample size;  $\boldsymbol{\beta}$  is a  $p \times 1$  vector of parameters. The error terms in the model have a general variance structure  $\mathbf{e} \sim (0, \sigma^2 \mathbf{R})$  where  $\sigma^2$  is an unknown constant and  $\mathbf{R}$  is a unknown  $n \times n$  covariance matrix. Define  $\mathbf{W}$  to be the diagonal matrix of survey weights. We assume throughout that the survey weights are constructed in such a way that they can be used for estimating finite population totals. The survey weighted least squares (SWLS) estimator is

$$\hat{\boldsymbol{\beta}}_{\text{SW}} = (\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W} \mathbf{Y} \equiv \mathbf{A}^{-1} \mathbf{X}^T \mathbf{W} \mathbf{Y},$$

assuming  $\mathbf{A} = \mathbf{X}^T \mathbf{W}^{-1} \mathbf{X}$  is invertible. Fuller (2002) describes the properties of this estimator. The estimator  $\hat{\boldsymbol{\beta}}_{\text{SW}}$  is model unbiased for  $\boldsymbol{\beta}$  under the model  $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$  regardless of whether  $\text{Var}_M(\mathbf{e}) = \sigma^2 \mathbf{R}$  is specified correctly or not, and is approximately design-unbiased for the census parameter  $\mathbf{B}_U = (\mathbf{X}_U^T \mathbf{X}_U)^{-1} \mathbf{X}_U^T \mathbf{Y}_U$ , in the finite population  $U$  of  $N$  units. The finite population values of the response vector and matrix of predictors are  $\mathbf{Y}_U = (Y_1, \dots, Y_N)^T$ , and  $\mathbf{X}_U = (\mathbf{X}_1, \dots, \mathbf{X}_p)$  with  $\mathbf{X}_k$  being the  $N \times 1$  vector of values for covariate  $k$ .

The remainder of the paper is organized as follows. Section 2 reviews results on condition numbers and variance

1. Dan Liao, RTI International, 701 13<sup>th</sup> Street, N.W., Suite 750, Washington DC, 20005. E-mail: dliao@rti.org; Richard Valliant, University of Michigan and University of Maryland, Joint Program in Survey Methodology, 1218 Lefrak Hall, College Park, MD, 20742.

decompositions for ordinary least squares. These are extended to be appropriate for survey estimation in section 3. The fourth section gives some numerical illustrations of the techniques. Section 5 is a conclusion. In most derivations, we use model-based calculations since the forms of the model-variances are useful for understanding the effects of collinearity. However, when presenting variance decompositions, we use estimators that have both model- and design-based justifications.

## 2. Condition indexes and variance decompositions in ordinary least squares estimation

In this section we briefly review techniques for diagnosing collinearity in ordinary least squares (OLS) estimation based on condition indexes and variance decompositions. These methods will be extended in section 3 to cover complex survey data.

### 2.1 Eigenvalues and eigenvectors of $\mathbf{X}^T\mathbf{X}$

When there is an exact (perfect) collinear relation in the  $n \times p$  data matrix  $\mathbf{X}$ , we can find a set of values,  $\mathbf{v} = (v_1, \dots, v_p)$ , not all zero, such that

$$v_1\mathbf{X}_1 + \dots + v_p\mathbf{X}_p = \mathbf{0}, \text{ or } \mathbf{X}\mathbf{v} = \mathbf{0}. \quad (1)$$

However, in practice, when there exists no exact collinearity but some near dependencies in the data matrix, it may be possible to find one or more non-zero vectors  $\mathbf{v}$  such that  $\mathbf{X}\mathbf{v} = \mathbf{a}$  with  $\mathbf{a} \neq \mathbf{0}$  but close to  $\mathbf{0}$ . Alternatively, we might say that a near dependency exists if the length of vector  $\mathbf{a}$ ,  $\|\mathbf{a}\|$ , is small. To normalize the problem of finding the set of  $\mathbf{v}$ 's that makes  $\|\mathbf{a}\|$  small, we consider only  $\mathbf{v}$  with unit length, that is, with  $\|\mathbf{v}\| = 1$ . Belsley (1991) discusses the connection of the eigenvalues and eigenvectors of  $\mathbf{X}^T\mathbf{X}$  with the normalized vector  $\mathbf{v}$  and  $\|\mathbf{a}\|$ . The minimum length  $\|\mathbf{a}\|$  is simply the positive square root of the smallest eigenvalue of  $\mathbf{X}^T\mathbf{X}$ . The  $\mathbf{v}$  that produces the  $\mathbf{a}$  with minimum length must be the eigenvector of  $\mathbf{X}^T\mathbf{X}$  that corresponds to the smallest eigenvalue. As discussed in the next section, the eigenvalues and eigenvectors of  $\mathbf{X}$  are related to those of  $\mathbf{X}^T\mathbf{X}$  and have some advantages when examining collinearity.

### 2.2 Singular-value decomposition, condition number and condition indexes

The singular-value decomposition (SVD) of matrix  $\mathbf{X}$  is very closely allied to the eigensystem of  $\mathbf{X}^T\mathbf{X}$ , but with its own advantages. The  $n \times p$  matrix  $\mathbf{X}$  can be decomposed as  $\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^T$ , where  $\mathbf{U}^T\mathbf{U} = \mathbf{V}^T\mathbf{V} = \mathbf{I}_p$  and  $\mathbf{D} = \text{diag}(\mu_1, \dots, \mu_p)$  is the diagonal matrix of singular values

(or eigenvalues) of  $\mathbf{X}$ . Here, the three components in the decomposition are matrices with very special, highly exploitable properties:  $\mathbf{U}$  is  $n \times p$  (the same size as  $\mathbf{X}$ ) and is column orthogonal;  $\mathbf{V}$  is  $p \times p$  and both row and column orthogonal;  $\mathbf{D}$  is  $p \times p$ , nonnegative and diagonal. Belsley *et al.* (1980) felt that the SVD of  $\mathbf{X}$  has several advantages over the eigen system of  $\mathbf{X}^T\mathbf{X}$ , for the sake of both statistical usages and computational complexity. For prediction,  $\mathbf{X}$  is the focus not the cross-product matrix  $\mathbf{X}^T\mathbf{X}$  since  $\hat{Y} = \mathbf{X}\hat{\beta}$ . In addition, the lengths  $\|\mathbf{a}\|$  of the linear combinations (1) of  $\mathbf{X}$  that relate to collinearity are properly defined in terms of the square roots of the eigenvalues of  $\mathbf{X}^T\mathbf{X}$ , which are the singular values of  $\mathbf{X}$ . A secondary consideration, given current computing power, is that the singular value decomposition of  $\mathbf{X}$  avoids the additional computational burden of forming  $\mathbf{X}^T\mathbf{X}$ , an operation involving  $np^2$  unneeded sums and products, which may lead to unnecessary truncation error.

The condition number of  $\mathbf{X}$  is defined as  $\kappa(\mathbf{X}) = \mu_{\max} / \mu_{\min}$ , where  $\mu_{\max}$  and  $\mu_{\min}$  are the maximum and minimum singular values of  $\mathbf{X}$ . Condition indexes are defined as  $\eta_k = \mu_{\max} / \mu_k$ . The closer that  $\mu_{\min}$  is to zero, the nearer  $\mathbf{X}^T\mathbf{X}$  is to being singular. Empirically, if a value of  $\kappa$  or  $\eta$  exceeds a cutoff value of, say, 10 to 30, two or more columns of  $\mathbf{X}$  have moderate or strong relations. The simultaneous occurrence of several large  $\eta_k$ 's is always remarkable for the existence of more than one near dependency.

One issue with the SVD is whether the  $\mathbf{X}$ 's should be centered around their means. Marquardt (1980) maintained that the centering of observations removes nonessential ill conditioning. In contrast, Belsley (1984) argues that mean-centering typically masks the role of the constant term in any underlying near-dependencies. A typical case is a regression with dummy variables. For example, if gender is one of the independent variables in a regression and most of the cases are male (or female), then the dummy for gender can be strongly collinear with the intercept. The discussions following Belsley (1984) illustrate the differences of opinion that occur among practitioners (Wood 1984; Snee and Marquardt 1984; Cook 1984). Moreover, in linear regression analysis, Wissmann, Toutenburg and Shalabh (2007) found that the degree of multicollinearity with dummy variables may be influenced by the choice of reference category. In this article, we do not center the  $\mathbf{X}$ 's but will illustrate the effect of the choice of reference category in section 4.

Another problem with the condition number is that it is affected by the scale of the  $x$  measurements (Steward 1987). By scaling down any column of  $\mathbf{X}$ , the condition number can be made arbitrarily large. This situation is known as *artificial ill-conditioning*. Belsley (1991) suggests

scaling each column of the design matrix  $\mathbf{X}$  using the Euclidean norm of each column before computing the condition number. This method is implemented in SAS and the package *perturb* of the statistical software R (Hendrickx 2010). Both use the root mean square of each column for scaling as its standard procedure. The condition number and condition indexes of the scaled matrix  $\mathbf{X}$  are referred to as the *scaled condition number* and *scaled condition indexes* of the matrix  $\mathbf{X}$ . Similarly, the variance decomposition proportions relevant to the scaled  $\mathbf{X}$  (which will be discussed in next section) will be called the *scaled variance decomposition proportions*.

### 2.3 Variance decomposition method

To assess the extent to which near dependencies (*i.e.*, having high condition indexes of  $\mathbf{X}$  and  $\mathbf{X}^T\mathbf{X}$ ) degrade the estimated variance of each regression coefficient, Belsley *et al.* (1980) reinterpreted and extended the work of Silvey (1969) by decomposing a coefficient variance into a sum of terms each of which is associated with a singular value. In the remainder of this section, we review the results of ordinary least squares (OLS) under the model  $E_M(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta}$  and  $\text{Var}_M(\mathbf{Y}) = \sigma^2\mathbf{I}_n$  where  $\mathbf{I}_n$  is the  $n \times n$  identity matrix. These results will be extended to survey weighted least squares in section 3. Recall that the model variance-covariance matrix of the OLS estimator  $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$  is  $\text{Var}_M(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$ . Using the SVD,  $\mathbf{X} = \mathbf{UDV}^T$ ,  $\text{Var}_M(\hat{\boldsymbol{\beta}})$  can be written as:

$$\text{Var}_M(\hat{\boldsymbol{\beta}}) = \sigma^2[(\mathbf{UDV}^T)^T(\mathbf{UDV}^T)]^{-1} = \sigma^2\mathbf{VD}^{-2}\mathbf{V}^T \quad (2)$$

and the  $k^{\text{th}}$  diagonal element in  $\text{Var}_M(\hat{\boldsymbol{\beta}})$  is the estimated variance for the  $k^{\text{th}}$  coefficient,  $\hat{\beta}_k$ . Using (2),  $\text{Var}_M(\hat{\beta}_k)$  can be expressed as:

$$\text{Var}_M(\hat{\beta}_k) = \sigma^2 \sum_{j=1}^p \frac{v_{kj}^2}{\mu_j^2} \quad (3)$$

where  $\mathbf{V} = (v_{kj})_{p \times p}$ . Let  $\phi_{kj} = v_{kj}^2 / \mu_j^2$ ,  $\phi_k = \sum_{j=1}^p \phi_{kj}$  and  $\mathbf{Q} = (\phi_{kj})_{p \times p} = (\mathbf{VD}^{-1}) \cdot (\mathbf{VD}^{-1})$ , where  $\cdot$  is the Hadamard (elementwise) product. The variance-decomposition proportions are  $\pi_{jk} = \phi_{jk} / \phi_k$ , which is the proportion of the variance of the  $k^{\text{th}}$  regression coefficient associated with the  $j^{\text{th}}$  component of its decomposition in (3). Denote the *variance decomposition proportion matrix* as  $\boldsymbol{\Pi} = (\pi_{jk})_{p \times p} = \mathbf{Q}^T\bar{\mathbf{Q}}^{-1}$ , where  $\bar{\mathbf{Q}}$  is the diagonal matrix with the row sums of  $\mathbf{Q}$  on the main diagonal and 0 elsewhere.

If the model is  $E_M(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta}$ ,  $\text{Var}_M(\mathbf{Y}) = \sigma^2\mathbf{W}^{-1}$  and weighted least squares is used, then  $\hat{\boldsymbol{\beta}}_{\text{WLS}} = (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\mathbf{Y}$  and  $\text{Var}_M(\hat{\boldsymbol{\beta}}_{\text{WLS}}) = \sigma^2(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}$ . The decomposition in (3) holds with  $\tilde{\mathbf{X}} = \mathbf{W}^{1/2}\mathbf{X}$  being decomposed

as  $\tilde{\mathbf{X}} = \mathbf{UDV}^T$ . However, in survey applications, it will virtually never be the case that the covariance matrix of  $\mathbf{Y}$  is  $\sigma^2\mathbf{W}^{-1}$  if  $\mathbf{W}$  is the matrix of survey weights. Section 3 covers the more realistic case.

In the variance decomposition (3), other things being equal, a small singular value  $\mu_j$  can lead to a large component of  $\text{Var}(\hat{\beta}_k)$ . However, if  $v_{kj}$  is small too, then  $\text{Var}(\hat{\beta}_k)$  may not be affected by a small  $\mu_j$ . One extreme case is when  $v_{kj} = 0$ . Suppose the  $k^{\text{th}}$  and  $j^{\text{th}}$  columns of  $\mathbf{X}$  belong to separate orthogonal blocks. Let  $\mathbf{X} \equiv [\mathbf{X}_1, \mathbf{X}_2]$  with  $\mathbf{X}_1^T\mathbf{X}_2 = \mathbf{0}$  and let the singular-value decompositions of  $\mathbf{X}_1$  and  $\mathbf{X}_2$  be given, respectively, as  $\mathbf{X}_1 = \mathbf{U}_1\mathbf{D}_{11}\mathbf{V}_{11}^T$  and  $\mathbf{X}_2 = \mathbf{U}_2\mathbf{D}_{22}\mathbf{V}_{22}^T$ . Since  $\mathbf{U}_1$  and  $\mathbf{U}_2$  are the orthogonal bases for the space spanned by the columns of  $\mathbf{X}_1$  and  $\mathbf{X}_2$  respectively,  $\mathbf{X}_1^T\mathbf{X}_2 = \mathbf{0}$  implies  $\mathbf{U}_1^T\mathbf{U}_2 = \mathbf{0}$  and  $\mathbf{U} \equiv [\mathbf{U}_1, \mathbf{U}_2]$  is column orthogonal. The singular value decomposition of  $\mathbf{X}$  is simply  $\mathbf{X} = \mathbf{UDU}_2^T$ , with:

$$\mathbf{D} = \begin{bmatrix} \mathbf{D}_{11} & \mathbf{0} \\ \mathbf{0} & \mathbf{D}_{22} \end{bmatrix} \quad (4)$$

and

$$\mathbf{V} = \begin{bmatrix} \mathbf{V}_{11} & \mathbf{0} \\ \mathbf{0} & \mathbf{V}_{22} \end{bmatrix}. \quad (5)$$

Thus  $\mathbf{V}_{12} = \mathbf{0}$ . An analogous result clearly applies to any number of mutually orthogonal subgroups. Hence, if all the columns in  $\mathbf{X}$  are orthogonal, all the  $v_{kj} = 0$  when  $k \neq j$  and  $\pi_{kj} = 0$  likewise. When  $v_{kj}$  is nonzero, this is a signal that predictors  $k$  and  $j$  are not orthogonal.

Since at least one  $v_{kj}$  must be nonzero in (3), this implies that a high proportion of any variance can be associated with a large singular value even when there is no collinearity. The standard approach is to check a high condition index associated with a large proportion of the variance of two or more coefficients when diagnosing collinearity, since there must be two or more columns of  $\mathbf{X}$  involved to make a near dependency. Belsley *et al.* (1980) suggested showing the matrix  $\boldsymbol{\Pi}$  and condition indexes of  $\mathbf{X}$  in a variance decomposition table as below. If two or more elements in the  $j^{\text{th}}$  row of matrix  $\boldsymbol{\Pi}$  are relatively large and its associated condition index  $\eta_j$  is large too, it signals that near dependencies are influencing regression estimates.

Condition Index	Proportions of variance			
	$\text{Var}_M(\hat{\beta}_1)$	$\text{Var}_M(\hat{\beta}_2)$	...	$\text{Var}_M(\hat{\beta}_p)$
$\eta_1$	$\pi_{11}$	$\pi_{12}$	...	$\pi_{1p}$
$\eta_2$	$\pi_{21}$	$\pi_{22}$	...	$\pi_{2p}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$
$\eta_p$	$\pi_{p1}$	$\pi_{p2}$	...	$\pi_{pp}$

### 3. Adaptation in survey-weighted least squares

#### 3.1 Condition indexes and variance decomposition proportions

In survey-weighted least squares (SWLS), we are more interested in the collinear relations among the columns in the matrix  $\tilde{\mathbf{X}} = \mathbf{W}^{1/2}\mathbf{X}$  instead of  $\mathbf{X}$ , since  $\hat{\boldsymbol{\beta}}_{\text{SW}} = (\tilde{\mathbf{X}}^T \tilde{\mathbf{X}})^{-1} \tilde{\mathbf{X}}^T \tilde{\mathbf{Y}}$ . Define the singular value decomposition of  $\tilde{\mathbf{X}}$  to be  $\tilde{\mathbf{X}} = \mathbf{U}\mathbf{D}\mathbf{V}^T$ , where  $\mathbf{U}$ ,  $\mathbf{V}$ , and  $\mathbf{D}$  are usually different from the ones of  $\mathbf{X}$ , due to the unequal survey weights.

The condition number of  $\tilde{\mathbf{X}}$  is defined as  $\kappa(\tilde{\mathbf{X}}) = \mu_{\max} / \mu_{\min}$ , where  $\mu_{\max}$  and  $\mu_{\min}$  are maximum and minimum singular values of  $\tilde{\mathbf{X}}$ . The condition number of  $\tilde{\mathbf{X}}$  is also usually different from the condition number of the data matrix  $\mathbf{X}$  due to unequal survey weights. Condition indexes are defined as

$$\eta_k = \mu_{\max} / \mu_k, \quad k = 1, \dots, p \tag{6}$$

where  $\mu_k$  is one of the singular values of  $\tilde{\mathbf{X}}$ . The scaled condition indexes and condition numbers are the condition indexes and condition numbers of the scaled  $\tilde{\mathbf{X}}$ .

Based on the extrema of the ratio of quadratic forms (Lin 1984), the condition number  $\kappa(\tilde{\mathbf{X}})$  is bounded in the range of:

$$\frac{w_{\min}^{1/2}}{w_{\max}^{1/2}} \kappa(\mathbf{X}) \leq \kappa(\tilde{\mathbf{X}}) \leq \frac{w_{\max}^{1/2}}{w_{\min}^{1/2}} \kappa(\mathbf{X}), \tag{7}$$

where  $w_{\min}$  and  $w_{\max}$  are the minimum and maximum survey weights. This expression indicates that if the survey weights do not vary too much, the condition number in SWLS resembles the one in OLS. However, in a sample with a wide range of survey weights, the condition number can be very different between SWLS and OLS. When SWLS has a large condition number, OLS might not. In the case of exact linear dependence among the columns of  $\mathbf{X}$ , the columns of  $\tilde{\mathbf{X}}$  will also be linearly dependent. In this extreme case at least one eigenvalue of  $\mathbf{X}$  will be zero, and both  $\kappa(\mathbf{X})$  and  $\kappa(\tilde{\mathbf{X}})$  will be infinite. As in OLS, large values of  $\kappa$  or of the  $\eta_k$ 's of 10 or more signal that two or more columns of  $\mathbf{X}$  have moderate to strong dependencies.

The model variance of the SWLS parameter estimator under a model with  $\text{Var}_M(\mathbf{e}) = \sigma^2\mathbf{R}$  is:

$$\text{Var}_M(\hat{\boldsymbol{\beta}}_{\text{SW}}) = \sigma^2(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\mathbf{R}\mathbf{W}\mathbf{X}(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1} = \sigma^2(\tilde{\mathbf{X}}^T\tilde{\mathbf{X}})^{-1}\mathbf{G}, \tag{8}$$

where

$$\mathbf{G} = (\mathbf{g}_{ij})_{p \times p} = \mathbf{X}^T\mathbf{W}\mathbf{R}\mathbf{W}\mathbf{X}(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1} \tag{9}$$

is the *misspecification effect* (MEFF) that represents the inflation factor needed to correct standard results for the effect of intracluster correlation in clustered survey data and for the fact that  $\text{Var}_M(\mathbf{e}) = \sigma^2\mathbf{R}$  and not  $\sigma^2\mathbf{W}^{-1}$  (Scott and Holt 1982).

Using the SVD of  $\tilde{\mathbf{X}}$ , we can rewrite  $\text{Var}_M(\hat{\boldsymbol{\beta}}_{\text{SW}})$  as

$$\text{Var}_M(\hat{\boldsymbol{\beta}}_{\text{SW}}) = \sigma^2\mathbf{V}\mathbf{D}^{-2}\mathbf{V}^T\mathbf{G}. \tag{10}$$

The  $k^{\text{th}}$  diagonal element in  $\text{Var}_M(\hat{\boldsymbol{\beta}}_{\text{SW}})$  is the estimated variance for the  $k^{\text{th}}$  coefficient,  $\hat{\beta}_k$ . Using (10),  $\text{Var}_M(\hat{\beta}_k)$  can be expressed as:

$$\text{Var}_M(\hat{\beta}_k) = \sigma^2 \sum_{j=1}^p \frac{v_{kj}}{\mu_j^2} \lambda_{kj} \tag{11}$$

where  $\lambda_{kj} = \sum_{i=1}^p v_{ij}g_{ik}$ . if  $\mathbf{R} = \mathbf{W}^{-1}$ , then  $\mathbf{G} = \mathbf{I}_p$ ,  $\lambda_{kj} = v_{kj}$ , and (11) reduces to (3). However, the situation is more complicated when  $\mathbf{G}$  is not the identity matrix, *i.e.*, when the complex design affects the variance of an estimated regression coefficient. If predictors  $k$  and  $j$  are orthogonal,  $v_{kj} = 0$  for  $k \neq j$  and the variance in (11) depends only on the  $k^{\text{th}}$  singular value and is unaffected by  $g_{ij}$ 's that are non-zero. If predictor  $k$  and several  $j$ 's are not orthogonal, then  $\lambda_{kj}$  has contributions from all of those eigenvectors and from the off-diagonal elements of the MEFF matrix  $\mathbf{G}$ . The term  $\lambda_{kj}$  then measures both non-orthogonality of  $x$ 's and effects of the complex design.

Consequently, we can define variance decomposition proportions analogous to those for OLS but their interpretation is less straightforward. Let  $\phi_{kj} = v_{kj} \lambda_{kj} / \mu_j^2$ ,  $\phi_k = \sum_{j=1}^p \phi_{kj}$  and  $\mathbf{Q} = (\phi_{kj})_{p \times p} = (\mathbf{V}\mathbf{D}^{-2}) \cdot (\mathbf{V}^T\mathbf{G})^T$ . The variance-decomposition proportions are  $\pi_{jk} = \phi_{jk} / \phi_k$ , which is the proportion of the variance of the  $k^{\text{th}}$  regression coefficient associated with the  $j^{\text{th}}$  component of its decomposition in (11). Denote the variance decomposition proportion matrix as

$$\boldsymbol{\Pi} = (\pi_{jk})_{p \times p} = \mathbf{Q}^T\bar{\mathbf{Q}}^{-1}, \tag{12}$$

where  $\bar{\mathbf{Q}}$  is the diagonal matrix with the row sums of  $\mathbf{Q}$  on the main diagonal and 0 elsewhere. The interpretation of the proportions in (12) is not as clear-cut as for OLS because the effect of the MEFF matrix. Section 3.2 discusses the interpretation in more detail in the context of stratified cluster sampling.

Analogous to the method for OLS regression, a variance decomposition table can be formed like the one at the end of section 2. When two or more independent variables are collinear (or “nearly dependent”), one singular value should make a large contribution to the variance of the parameter estimates associated with those variables. For example, if the proportions  $\pi_{31}$  and  $\pi_{32}$  for the variances of  $\hat{\boldsymbol{\beta}}_{\text{SW1}}$  and

$\hat{\beta}_{SW2}$  are large, this would say that the third singular value makes a large contribution to both variances and that the first and second predictors in the regression are, to some extent, collinear. As shown in section 2.3, when the  $k^{th}$  and  $j^{th}$  columns in  $\mathbf{X}$  are orthogonal,  $v_{kj} = 0$  and the  $j^{th}$  singular value's decomposition proportion  $\pi_{jk}$  on  $\text{Var}(\hat{\beta}_k)$  will be 0.

Several special cases are worth noting. If  $\mathbf{R} = \mathbf{W}^{-1}$  as assumed in WLS, then  $\mathbf{G} = \mathbf{I}$ . The variance decomposition in (11) has the same form as (2) in OLS. However, having  $\mathbf{R} = \mathbf{W}^{-1}$  in survey data would be unusual since survey weights are not typically computed based on the variance structure of a model. Note that  $\mathbf{V}$  is still different from the one in OLS and is one component of the SVD of  $\tilde{\mathbf{X}}$  instead of  $\mathbf{X}$ . Another special case here is when  $\mathbf{R} = \mathbf{I}$  and the survey weights are equal, in which case the OLS results can be used. However, when the survey weights are unequal, even when  $\mathbf{R} = \mathbf{I}$ , the variance decomposition in (11) is different from (2) in OLS since  $\mathbf{G} \neq \mathbf{I}$ . In the next section, we will consider some special models that take the population features such as clusters and strata into account when estimating this variance decomposition.

### 3.2 Variance decomposition for a model with stratified clustering

The model variance of  $\hat{\beta}_{SW}$  in (8) contains the unknown  $\mathbf{R}$  that must be estimated. In this section, we present an estimator for  $\hat{\beta}_{SW}$  that is appropriate for a model with stratified clustering. The variance estimator has both model-based and design-based justification. Suppose that in a stratified multistage sampling design, there are strata  $h = 1, \dots, H$  in the population, clusters  $i = 1, \dots, N_h$  in stratum  $h$  and units  $t = 1, \dots, M_{hi}$  in cluster  $hi$ . We select clusters  $i = 1, \dots, n_h$  in stratum  $h$  and units  $t = 1, \dots, m_{hi}$  in cluster  $hi$ . Denote the set of sample clusters in stratum  $h$  by  $s_h$  and the sample of units in cluster  $hi$  as  $s_{hi}$ . The total number of sample units in stratum  $h$  is  $m_h = \sum_{i \in s_h} m_{hi}$ , and the total in the sample is  $m = \sum_{h=1}^H m_h$ . Assume that clusters are selected with varying probabilities and with replacement within strata and independently between strata. The model we consider is:

$$E_M(Y_{hit}) = \mathbf{x}_{hit}^T \boldsymbol{\beta}$$

$$h = 1, \dots, H, i = 1, \dots, N_h, t = 1, \dots, M_{hi}$$

$$\text{Cov}_M(\varepsilon_{hit}, \varepsilon_{hi't'}) = 0$$

where  $\varepsilon_{hit} = Y_{hit} - \mathbf{x}_{hit}^T \boldsymbol{\beta}, i \neq i'$

$$\text{Cov}_M(\varepsilon_{hit}, \varepsilon_{hi't'}) = 0 \quad h \neq h'. \tag{13}$$

Units within each cluster are assumed to be correlated but the particular form of the covariances does not have to be

specified for this analysis. The estimator  $\hat{\beta}_{SW}$  of the regression parameter can be written as:

$$\hat{\beta}_{SW} = \sum_{h=1}^H \sum_{i \in s_h} (\tilde{\mathbf{X}}^T \tilde{\mathbf{X}})^{-1} \mathbf{X}_{hi}^T \mathbf{W}_{hi} \mathbf{Y}_{hi} \tag{14}$$

where  $\mathbf{X}_{hi}$  is the  $m_{hi} \times p$  matrix of covariates for sample units in cluster  $hi$ ,  $\mathbf{W}_{hi} = \text{diag}(w_t), t \in s_{hi}$ , is the diagonal matrix of survey weights for units in cluster  $hi$  and  $\mathbf{Y}_{hi}$  is the  $m_{hi} \times 1$  vector of response variables in cluster  $hi$ . The model variance of  $\hat{\beta}_{SW}$  is:

$$\text{Var}_M(\hat{\beta}_{SW}) = (\tilde{\mathbf{X}}^T \tilde{\mathbf{X}})^{-1} \mathbf{G}_{st} \tag{15}$$

where

$$\mathbf{G}_{st} = \left[ \sum_{h=1}^H \sum_{i \in s_h} \mathbf{X}_{hi}^T \mathbf{W}_{hi} \mathbf{R}_{hi} \mathbf{W}_{hi} \mathbf{X}_{hi} \right] (\tilde{\mathbf{X}}^T \tilde{\mathbf{X}})^{-1}$$

$$= \left[ \sum_{h=1}^H \mathbf{X}_h^T \mathbf{W}_h \mathbf{R}_h \mathbf{W}_h \mathbf{X}_h \right] (\tilde{\mathbf{X}}^T \tilde{\mathbf{X}})^{-1} \tag{16}$$

with  $\mathbf{R}_{hi} = \text{Var}_M(\mathbf{Y}_{hi}), \mathbf{W}_h = \text{diag}(\mathbf{W}_{hi}),$  and  $\mathbf{R}_h = \text{Blkdiag}(\mathbf{R}_{hi}), \mathbf{W}_h = \text{diag}(\mathbf{W}_{hi}), \mathbf{X}_h^T = (\mathbf{X}_{h1}^T, \mathbf{X}_{h2}^T, \dots, \mathbf{X}_{hn_h}^T), i \in s_h.$  Expression (16) is a special case of (9) with  $\mathbf{X}^T = (\mathbf{X}_1^T, \mathbf{X}_2^T, \dots, \mathbf{X}_H^T),$  where  $\mathbf{X}_h$  is the  $m_h \times p$  matrix of covariates for sample units in stratum  $h, \mathbf{W} = \text{diag}(\mathbf{W}_{hi}),$  for  $h = 1, \dots, H$  and  $i \in s_h$  and  $\mathbf{R} = \text{Blkdiag}(\mathbf{R}_h).$

Based on the development in Scott and Holt (1982, section 4), the MEFF matrix  $\mathbf{G}_{st}$  can be rewritten for a special case of  $\mathbf{R}_h$  in a way that will make the decomposition proportions in (12) more understandable. Consider the special case of (13) with

$$\text{Cov}_M(\mathbf{e}_{hi}) = \sigma^2(1 - \rho) \mathbf{I}_{m_{hi}} + \sigma^2 \rho \mathbf{1}_{m_{hi}} \mathbf{1}_{m_{hi}}^T$$

where  $\mathbf{I}_{m_{hi}}$  is the  $m_{hi} \times m_{hi}$  identity matrix and  $\mathbf{1}_{m_{hi}}$  is a vector of  $m_{hi}$  1's. In that case,

$$\mathbf{X}_h^T \mathbf{W}_h \mathbf{R}_h \mathbf{W}_h \mathbf{X}_h = (1 - \rho) \mathbf{X}_h^T \mathbf{W}_h^2 \mathbf{X}_h$$

$$+ \rho \sum_{i \in s_h} m_{hi} \mathbf{X}_{Bhi}^T \mathbf{W}_{hi}^2 \mathbf{X}_{Bhi}$$

where  $\mathbf{X}_{Bhi} = m_{hi}^{-1} \mathbf{1}_{m_{hi}} \mathbf{1}_{m_{hi}}^T \mathbf{X}_{hi}.$  Suppose that the sample is self-weighting so that  $\mathbf{W}_{hi} = w \mathbf{I}_{m_{hi}}.$  After some simplification, it follows that

$$\mathbf{G}_{st} = w[\mathbf{I}_p + (\mathbf{M} - \mathbf{I}_p) \rho]$$

where  $\mathbf{I}_p$  is the  $p \times p$  identity matrix and  $\mathbf{M} = (\sum_{h=1}^H \sum_{i \in s_h} m_{hi} \mathbf{X}_{Bhi}^T \mathbf{X}_{Bhi}) (\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1}.$  Thus, if the sample is self-weighting and  $\rho$  is very small, then  $\mathbf{G}_{st} \approx w \mathbf{I}_p$  and

$\text{Var}_M(\hat{\beta}_{SW})$  in (15) will be approximately the same as the OLS variance. If so, the SWLS variance decomposition proportions will be similar to the OLS proportions. In regression problems,  $\rho$  often is small since it is the correlation of the errors,  $\varepsilon_{hit} = Y_{hit} - \mathbf{x}_{hit}^T \boldsymbol{\beta}$ , for different units rather than for  $\mathbf{Y}_{hit}$ 's. This is related to the phenomenon that design effects for regression coefficients are often smaller than for means—a fact first noted by Kish and Frankel (1974). In applications where  $\rho$  is larger, the variance decomposition proportions in (12) will still be useful in identifying collinearity although they will be affected by departures of the model errors from independence.

Denote the cluster-level residuals as a vector,  $\mathbf{e}_{hi} = \mathbf{Y}_{hi} - \mathbf{X}_{hi} \hat{\beta}_{SW}$ . The estimator of (15) that we consider was originally derived from design-based considerations. A linearization estimator, appropriate when clusters are selected with replacement, is:

$$\text{var}_L(\hat{\beta}_{SW}) = (\tilde{\mathbf{X}}^T \tilde{\mathbf{X}})^{-1} \hat{\mathbf{G}}_L \quad (17)$$

with the estimated misspecification effect as

$$\hat{\mathbf{G}}_L = (\hat{g}_{ij})_{p \times p} = \left[ \sum_{h=1}^H \frac{n_h}{n_h - 1} \sum_{i \in s_h} (\mathbf{z}_{hi}^* - \bar{\mathbf{z}}_h^*)(\mathbf{z}_{hi}^* - \bar{\mathbf{z}}_h^*)^T \right] (\tilde{\mathbf{X}}^T \tilde{\mathbf{X}})^{-1}, \quad (18)$$

where  $\bar{\mathbf{z}}_h^* = 1/n_h \sum_{i \in s_h} \mathbf{z}_{hi}^*$  and  $\mathbf{z}_{hi}^* = \mathbf{X}_{hi}^T \mathbf{W}_{hi} \mathbf{e}_{hi}$  with  $\mathbf{e}_{hi} = \mathbf{Y}_{hi} - \mathbf{X}_{hi} \hat{\beta}_{SW}$ , and the variance-covariance matrix  $\mathbf{R}$  can be estimated by

$$\hat{\mathbf{R}} = \sum_{h=1}^H \frac{n_h}{n_h - 1} \left[ \text{Blkdiag}(\mathbf{e}_{hi} \mathbf{e}_{hi}^T) - \frac{1}{n_h} \mathbf{e}_h \mathbf{e}_h^T \right].$$

Expression (17) is used by the Stata and SUDAAN packages, among others. The estimator  $\text{var}_L(\hat{\beta}_{SW})$  is consistent and approximately design-unbiased under a design where clusters are selected with replacement (Fuller 2002). The estimator in (17) is also an approximately model-unbiased estimator of (15) (see Liao 2010). Since the estimator  $\text{var}_L(\hat{\beta}_{SW})$  is also currently available in software packages, we will use it in the empirical work in section 4.

Using (12) derived in section 2, the variance decomposition proportion matrix  $\boldsymbol{\Pi}$  for  $\text{var}_L(\hat{\beta}_{SW})$  can then be written as

$$\boldsymbol{\Pi} = (\pi_{jk})_{p \times p} = \mathbf{Q}_L^T \bar{\mathbf{Q}}_L^{-1} \quad (19)$$

with  $\mathbf{Q}_L = (\phi_{kj})_{p \times p} = (\mathbf{V} \mathbf{D}^{-2}) \cdot (\mathbf{V}^T \hat{\mathbf{G}}_L)^T$  and  $\bar{\mathbf{Q}}_L$  is the diagonal matrix with the row sums of  $\mathbf{Q}_L$  on the main diagonal and 0 elsewhere.

## 4. Numerical illustrations

In this section, we will illustrate the collinearity measures described in section 3 and investigate their behaviors using the dietary intake data from 2007-2008 National Health and Nutrition Examination Survey (NHANES).

### 4.1 Description of the data

The dietary intake data are used to estimate the types and amounts of foods and beverages consumed during the 24-hour period prior to the interview (midnight to midnight), and to estimate intakes of energy, nutrients, and other food components from those foods and beverages. NHANES uses a complex, multistage, probability sampling design; oversampling of certain population subgroups is done to increase the reliability and precision of health status indicator estimates for these groups. Among the respondents who received the in-person interview in the mobile examination center (MEC), around 94% provided complete dietary intakes. The survey weights were constructed by taking MEC sample weights and further adjusting for the additional nonresponse and the differential allocation by day of the week for the dietary intake data collection. These weights are more variable than the MEC weights. The data set used in our study is a subset of 2007-2008 data composed of female respondents aged 26 to 40. Observations with missing values in the selected variables are excluded from the sample which finally contains 672 complete respondents. The final weights in our sample range from 6,028 to 330,067, with a ratio of 55:1. The U.S. National Center for Health Statistics recommends that the design of the sample is approximated by the stratified selection with replacement of 32 PSUs from 16 strata, with 2 PSUs within each stratum.

### 4.2 Study one: Correlated covariates

In the first empirical study, a linear regression model of respondent's body mass index (BMI) was considered. The explanatory variables considered included two demographic variables, respondent's age and race (Black/Non-black), four dummy variables for whether the respondent is on a special diet of any kind, on a low-calorie diet, on a low-fat diet, and on a low-carbohydrate diet (when he/she is on diet, value equals 1, otherwise 0), and ten daily total nutrition intake variables, consisting of total calories (100kcal), protein (100gm), carbohydrate (100gm), sugar (100gm), dietary fiber (100gm), alcohol (100gm), total fat (100gm), total saturated fatty acids (100gm), total monounsaturated fatty acids (100gm), and total polyunsaturated fatty acids (100gm). The correlation coefficients among these variables are displayed in Table 2. Note that the correlations among the daily total nutrition intake variables are often high. For



example, the correlations of the total fat intakes with total saturated fatty acids, total monounsaturated fatty acids and total polyunsaturated fatty acids are 0.85, 0.97 and 0.93.

Three types of regressions were fitted for the selected sample to demonstrate different diagnostics. More details about these three regression types and their diagnostic statistics are displayed in Table 1.

TYPE1: OLS regression with estimated  $\sigma^2$ ; the diagnostic statistics are obtained using the standard methods reviewed in section 2;

TYPE2: WLS regression with estimated  $\sigma^2$  and assuming  $\mathbf{R} = \mathbf{W}^{-1}$ ; the scaled condition indexes are estimated using (6) and the scaled variance decomposition proportions are estimated using (12). With  $\mathbf{R} = \mathbf{W}^{-1}$ , these are the variance decompositions that will be produced by standard software using WLS and specifying the weights to be the survey weights;

TYPE3: SWLS with estimated  $\hat{\mathbf{R}}$ ; the scaled condition indexes are estimated using (6); the scaled variance decomposition proportions are estimated using (12).

Their diagnostic statistics, including the scaled condition indexes and variance decomposition proportions are reported in Tables 3, 4 and 5, respectively. To make the

tables more readable, only the proportions that are larger than 0.3 are shown. Proportions that are less than 0.3 are shown as dots. Note that some terms in decomposition (12) can be negative. This leads to the possibility of some “proportions” being greater than 1. This occurs in five cases in Table 5. Belsley *et al.* (1980) suggest that a condition index of 10 signals that collinearity has a moderate effect on standard errors; an index of 100 would indicate a serious effect. In this study, we consider a scaled condition index greater than 10 to be relatively large, and ones greater than 30 as large and remarkable. Furthermore, the large scaled variance-decomposition proportions (greater than 0.3) associated with each large scaled condition index will be used to identify those variates that are involved in a near dependency. The intracluster correlation of the residuals is shown in the last row of Table 6 under the column labeled “Original Model”. In the model used for Tables 3-5,  $\rho = 0.0366$  as estimated from a model with random effects for clusters. As noted in section 3.2, when  $\rho$  is small and the sample is self-weighting, the SWLS decomposition proportions can be interpreted in the same way as those of OLS. Although the NHANES sample does not have equal weights,  $\rho$  is small in this example and the decomposition proportions should still provide useful information.

**Table 1**  
Regression models and their collinearity diagnostic statistics used in this experimental study

Type	Regression Method	Weight matrix $\mathbf{W}^a$	$\text{var}(\hat{\boldsymbol{\beta}})$	$\text{var}(\hat{\boldsymbol{\beta}}_k)$	Matrix for Condition Indexes <sup>b</sup>	Variance Decomposition Proportion $\pi_{jk}$
TYPE1	OLS	$\mathbf{I}$	$\hat{\sigma}^2(\mathbf{X}^T\mathbf{X})^{-1}$	$\hat{\sigma}^2 \sum_{j=1}^p \frac{u_{2kj}^2}{\mu_j^2}$ <sup>c</sup>	$\mathbf{X}^T\mathbf{X}$	$\frac{u_{2kj}^2}{\mu_j^2} / \sum_{j=1}^p \frac{u_{2kj}^2}{\mu_j^2}$
TYPE2	WLS	$\mathbf{W}$	$\hat{\sigma}^2(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}$	$\hat{\sigma}^2 \sum_{j=1}^p \frac{u_{2kj}^2}{\mu_j^2}$ <sup>d</sup>	$\mathbf{X}^T\mathbf{W}\mathbf{X}$	$\frac{u_{2kj}^2}{\mu_j^2} / \sum_{j=1}^p \frac{u_{2kj}^2}{\mu_j^2}$
TYPE3	SWLS	$\mathbf{W}$	$\hat{\sigma}^2(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\hat{\mathbf{R}}\mathbf{W}\mathbf{X}(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}$	$\hat{\sigma}^2 \sum_{j=1}^p \frac{u_{2kj} \sum_{i=1}^p \hat{g}_{ik} u_{2ij}}{\mu_j^2}$ <sup>e</sup>	$\mathbf{X}^T\mathbf{W}\mathbf{X}$	$\frac{u_{2kj} \sum_{i=1}^p \hat{g}_{ik} u_{2ij}}{\mu_j^2} / \sum_{j=1}^p \frac{u_{2kj} \sum_{i=1}^p \hat{g}_{ik} u_{2ij}}{\mu_j^2}$

$$\hat{\mathbf{R}} = \sum_{h=1}^H \frac{n_h}{n_h - 1} \left[ \text{Blkdiag}(\mathbf{e}_h \mathbf{e}_h^T) - \frac{1}{n_h} \mathbf{e}_h \mathbf{e}_h^T \right]$$

<sup>a</sup> In all the regression models, the parameters are estimated by:  $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\mathbf{Y}$ .  
<sup>b</sup> The eigenvalues of this matrix will be used to compute the Condition Indexes for the corresponding regression model.  
<sup>c</sup> The terms  $u_{2kj}$  and  $\mu_j$  are from the singular value decomposition of the data matrix  $\mathbf{X}$ .  
<sup>d</sup> The terms  $u_{2kj}$  and  $\mu_j$  are from the singular value decomposition of the weighted data matrix  $\tilde{\mathbf{X}} = \mathbf{W}^{1/2}\mathbf{X}$ .  
<sup>e</sup> The terms  $u_{2kj}$  and  $\mu_j$  are from the singular value decomposition (SVD) of the weighted data matrix  $\tilde{\mathbf{X}}$ . The term  $\hat{g}_{ik}$  is the unit element of misspecification effect matrix  $\hat{\mathbf{G}}$ .

In Tables 3, 4 and 5, the weighted regression methods, WLS and SWLS, used the survey-weighted data matrix  $\tilde{\mathbf{X}}$  to obtain the condition indexes while the unweighted regression method, OLS, used the data matrix  $\mathbf{X}$ . The largest scaled condition index in WLS and SWLS is 566, which is slightly smaller than the one in OLS, 581. Both of these values are much larger than 30 and, thus, signal a severe near-dependency among the predictors in all three regression models. Such large condition numbers imply that the inverse of the design matrix,  $\mathbf{X}^T\mathbf{W}\mathbf{X}$ , may be numerically unstable, *i.e.*, small changes in the  $x$  data could make large changes in the elements of the inverse.

The values of the decomposition proportions for OLS and WLS are very similar and lead to the same predictors being identified as potentially collinear. Results for SWLS are somewhat different as sketched below. In OLS and WLS, six daily total nutrition intake variables—calorie, protein, carbohydrate, alcohol, dietary fiber and total fat—are involved in the dominant near-dependency that is associated with the largest scaled condition index. Four daily fat intake variables, total fat, total saturated fatty acids, total monounsaturated fatty acids and total polyunsaturated fatty acids, are involved in the secondary near-dependency that is associated with the second largest scaled condition index. A moderate near-dependency between intercept and age is also shown in all three tables. The associated scaled condition index is equal to 38 in OLS and 37 in WLS and SWLS. However, when SWLS is used, sugar, total saturated fatty acids and total polyunsaturated fatty acids also appear to be involved in the dominant near-dependency as shown in Table 5. While, only three daily fat intake variables, total saturated fatty acids, total monounsaturated fatty acids and total polyunsaturated fatty acids, are involved in the secondary near-dependency that is associated with the second largest scaled condition index. Thus, when OLS or WLS is used, the impact of near-dependency among sugar, total saturated fatty acids, total polyunsaturated fatty acids and the six daily total nutrition intake variables is not as strong as the ones in SWLS. If conventional OLS or WLS diagnostics are used for SWLS, this near-dependency might be overlooked.

Rather than using the scaled condition indexes and variance decomposition method (in Tables 3, 4 and 5), an analyst might attempt to identify collinearities by examining the unweighted correlation coefficient matrix in Table 2. Although the correlation coefficient matrix shows that almost all the daily total nutrition intake variables are highly or moderately pairwise correlated, it cannot be used to

reliably identify the near-dependencies among these variables when used in a regression. For example, the correlation coefficient between “on any diet” and “on low-calorie diet” is relatively large (0.73). This near dependency is associated with a scaled condition index equal to 11 (larger than 10, but less than the cutoff of 30) in OLS and WLS (shown in Table 3 and 4) and is associated with a scaled condition index equal to 2 (less than 10) in SWLS (shown in Table 5). The impact of this near dependency appears to be not very harmful no matter which regression method is used. On the other hand, alcohol is weakly correlated with all the daily total nutrition intake variables but is highly involved in the dominant near-dependency shown in the last row of Tables 3-5.

After the collinearity patterns are diagnosed, the common corrective action would be to drop the correlated variables, refit the model and reexamine standard errors, collinearity measures and other diagnostics. Omitting  $\mathbf{X}$ 's one at a time may be a divisible because of the potentially complex interplay of explanatory variables. In this example, if the total fat intake is one of the key variables that an analyst feels must be kept, sugar might be dropped first followed by protein, calorie, alcohol, carbohydrate, total fat, dietary fiber, total monounsaturated fatty acids, total polyunsaturated fatty acids and total saturated fatty acids. Other remedies for collinearity could be to transform the data or use some specialized techniques such as ridge regression and mixed Bayesian modeling, which require extra (prior) information beyond the scope of most research and evaluations.

To demonstrate how the collinearity diagnostics can improve the regression results in this example, Table 6 presents the SWLS regression analysis output of the original models with all the explanatory variables and a reduced model with fewer explanatory variables. In the reduced model, all of the dietary intake variables are eliminated except total fat intake. After the number of correlated offending variables is reduced, the standard error of total fat intake is only the one forty-sixth of its standard error in the original model. The total fat intake becomes significant in the reduced model. The reduction of correlated variables appears to have substantially improved the accuracy of estimating the impact of total fat intake on BMI. Note that the collinearity diagnostics do not provide a unique path toward a final model. Different analysts may make different choices about whether particular predictors should be dropped or retained.

**Table 2**  
Correlation coefficient matrix of the data matrix X

	age	black	on any diet	on low-calorie diet	on low-fat diet	on low-carb diet <sup>a</sup>	calorie	protein	Carbo-hydrate	sugar	fiber	alcohol	total. fat	sat. fat	mono. fat	poly. fat
age	1															
black	<i>.<sup>b</sup></i>	1														
on any diet	.	.	1													
on low-calorie diet	.	.	<i>0.87<sup>c</sup></i>	1												
on low-fat diet	.	.	.	.	1											
one low-carb diet	.	.	.	.	.	1										
calorie	.	.	.	.	.	.	1									
protein	.	.	.	.	.	.	<i>0.75</i>	1								
carb	.	.	.	.	.	.	<i>0.84</i>	<i>0.45</i>	1							
sugar	.	.	.	.	.	.	<i>0.58</i>	.	<i>0.84</i>	1						
fiber	.	.	.	.	.	.	<i>0.57</i>	<i>0.52</i>	<i>0.54</i>	.	1					
alcohol	.	.	.	.	.	.	.	.	.	.	.	1				
total.fat	.	.	.	.	.	.	<i>0.86</i>	<i>0.72</i>	<i>0.54</i>	.	<i>0.48</i>	.	1			
sat.fat <sup>d</sup>	.	.	.	.	.	.	<i>0.74</i>	<i>0.56</i>	<i>0.47</i>	.	<i>0.46</i>	.	<i>0.85</i>	1		
mono.fat <sup>e</sup>	.	.	.	.	.	.	<i>0.83</i>	<i>0.68</i>	<i>0.51</i>	.	<i>0.46</i>	.	<i>0.97</i>	<i>0.82</i>	1	
poly.fat <sup>f</sup>	.	.	.	.	.	.	<i>0.81</i>	<i>0.71</i>	<i>0.51</i>	.	<i>0.43</i>	.	<i>0.93</i>	<i>0.63</i>	<i>0.87</i>	1

<sup>a</sup> The term “carb” stands for carbohydrate.

<sup>b</sup> Correlation coefficients less than 0.3 are omitted in this table.

<sup>c</sup> Correlation coefficients larger than 0.3 are italicized in this table.

<sup>d</sup> Total Saturated Fatty Acids.

<sup>e</sup> Total Monounsaturated Fatty Acids.

<sup>f</sup> Total Polyunsaturated Fatty Acids.

**Table 3**  
Scaled condition indexes and variance decomposition proportions: Using TYPE1: OLS

Scaled Condition Index	Scaled Proportion of the Variance of									
	Intercept	Age	Black	on any Diet	on Low-Calorie Diet	on Low-fat Diet	on Low-carb Diet	Calorie	Protein	
1	<i>.<sup>a</sup></i>	.	.	.	.	.	.	.	.	.
2	.	.	.	.	.	.	.	.	.	.
3	.	.	.	.	.	.	<i>0.574</i>	.	.	.
3	.	.	.	.	.	<i>0.379</i>	.	.	.	.
4	.	.	<i>0.794</i>	.	.	.	.	.	.	.
5	.	.	.	.	.	.	.	.	.	.
6	.	.	.	.	.	.	.	.	.	.
8	.	.	.	.	.	.	.	.	.	.
9	.	.	.	.	.	.	.	.	.	.
11	.	.	.	<i>0.842</i>	<i>0.820</i>	.	.	.	.	.
12	.	.	.	.	.	.	.	.	.	.
22	.	.	.	.	.	.	.	.	.	.
26	.	.	.	.	.	.	.	.	.	.
38	<i>0.970</i>	<i>0.960</i>	.	.	.	.	.	.	.	.
157	.	.	.	.	.	.	.	.	.	.
581	.	.	.	.	.	.	.	<i>0.993</i>	<i>0.966</i>	.
Scaled Condition Index	Carbohydrate	Sugar	Dietary Fiber	Alcohol	Total Fat	Sat.fat <sup>b</sup>	Mono.fat <sup>c</sup>	Poly.fat <sup>d</sup>		
1	.	.	.	.	.	.	.	.	.	.
2	.	.	.	.	.	.	.	.	.	.
3	.	.	.	.	.	.	.	.	.	.
3	.	.	.	.	.	.	.	.	.	.
4	.	.	.	.	.	.	.	.	.	.
5	.	.	.	.	.	.	.	.	.	.
6	.	.	.	.	.	.	.	.	.	.
8	.	.	.	.	.	.	.	.	.	.
9	.	.	.	.	.	.	.	.	.	.
11	.	.	.	.	.	.	.	.	.	.
12	.	.	.	.	.	.	.	.	.	.
22	.	.	.	.	.	.	.	.	.	.
26	.	<i>0.633</i>	.	.	.	.	.	.	.	.
38	.	.	.	.	.	.	.	.	.	.
157	.	.	.	.	<i>0.304</i>	<i>0.866</i>	<i>0.890</i>	<i>0.904</i>	.	.
581	<i>0.988</i>	.	<i>0.482</i>	<i>0.986</i>	<i>0.696</i>	.	.	.	.	.

<sup>a</sup> The scaled variance decomposition proportions smaller than 0.3 are omitted in this table.

<sup>b</sup> Total Saturated Fatty Acids.

<sup>c</sup> Total Monounsaturated Fatty Acids.

<sup>d</sup> Total Polyunsaturated Fatty Acids.

**Table 4**  
Scaled condition indexes and variance decomposition proportions: Using TYPE2: WLS

Scaled Condition Index	Scaled Proportion of the Variance of								
	Intercept	Age	Black	on any Diet	on Low-Calorie Diet	on Low-fat Diet	on Low-carb Diet	Calorie	Protein
1	<sup>a</sup>	.	.	.	.	.	.	.	.
2	.	.	.	.	.	.	.	.	.
3	.	.	.	.	.	.	0.609	.	.
3	.	.	.	.	.	0.347	.	.	.
4	.	.	0.711	.	.	.	.	.	.
5	.	.	.	.	.	.	.	.	.
7	.	.	.	.	.	.	.	.	.
8	.	.	.	.	.	.	.	.	.
10	.	.	.	.	.	.	.	.	.
11	.	.	.	0.902	0.878	.	.	.	.
13	.	.	.	.	.	.	.	.	.
21	.	.	.	.	.	.	.	.	.
26	.	.	.	.	.	.	.	.	.
37	0.959	0.940	.	.	.	.	.	.	.
165	.	.	.	.	.	.	.	.	.
566	.	.	.	.	.	.	.	0.992	0.963
Scaled Condition Index	Carbohydrate	Sugar	Dietary Fiber	Alcohol	Total Fat	Sat.fat <sup>b</sup>	Mono.fat <sup>c</sup>	Poly.fat <sup>d</sup>	
1	.	.	.	.	.	.	.	.	.
2	.	.	.	.	.	.	.	.	.
3	.	.	.	.	.	.	.	.	.
3	.	.	.	.	.	.	.	.	.
4	.	.	.	.	.	.	.	.	.
5	.	.	.	.	.	.	.	.	.
7	.	.	.	.	.	.	.	.	.
8	.	.	.	.	.	.	.	.	.
10	.	.	.	.	.	.	.	.	.
11	.	.	.	.	.	.	.	.	.
13	.	.	.	.	.	.	.	.	.
21	.	.	.	.	.	.	.	.	.
26	.	0.630	.	.	.	.	.	.	.
37	.	.	.	.	.	.	.	.	.
165	.	.	.	.	0.342	0.871	0.909	0.919	.
566	0.987	.	0.486	0.981	0.658	.	.	.	.

<sup>a</sup> The scaled variance decomposition proportions smaller than 0.3 are omitted in this table.  
<sup>b</sup> Total Saturated Fatty Acids.  
<sup>c</sup> Total Monounsaturated Fatty Acids.  
<sup>d</sup> Total Polyunsaturated Fatty Acids.

**Table 5**  
Scaled condition indexes and variance decomposition proportions: Using TYPE3: SWLS

Scaled Condition Index	Scaled Proportion of the Variance of								
	Intercept	Age	Black	on any Diet	on Low-Calorie Diet	on Low-fat Diet	on Low-carb Diet	Calorie	Protein
1	<sup>a</sup>	.	.	.	.	.	.	.	.
2	.	.	.	0.717	1.278	0.553	.	.	.
3	.	.	.	.	.	.	0.697	.	.
3	.	.	.	.	.	.	.	.	.
3	.	.	.	.	.	.	.	.	.
4	.	.	.	.	.	.	.	.	.
5	.	.	.	.	.	.	.	.	.
7	0.766	1.686	0.461	.	.	.	.	.	.
8	.	.	.	.	.	.	.	.	.
10	.	.	.	.	.	.	.	.	.
11	.	.	.	.	.	.	.	.	.
13	.	.	.	.	.	.	.	.	.
21	.	.	.	.	.	.	.	.	.
26	.	.	.	.	.	.	.	.	.
37	.	.	.	.	.	.	.	.	.
165	.	.	.	.	.	.	.	.	.
566	0.318	.	.	.	.	.	.	1.095	1.190

<sup>a</sup> The scaled variance decomposition proportions smaller than 0.3 are omitted in this table.  
<sup>b</sup> Total Saturated Fatty Acids.  
<sup>c</sup> Total Monounsaturated Fatty Acids.  
<sup>d</sup> Total Polyunsaturated Fatty Acids.

**Table 5 (continued)**  
**Scaled condition indexes and variance decomposition proportions: Using TYPE3: SWLS**

Scaled Condition Index	Scaled Proportion of the Variance of							
	Carbohydrate	Sugar	Dietary Fiber	Alcohol	Total Fat	Sat.fat <sup>b</sup>	Mono.fat <sup>c</sup>	Poly.fat <sup>d</sup>
1	.	.	.	.	.	.	.	.
2	.	.	.	.	.	.	.	.
3	.	.	.	.	.	.	.	.
3	.	.	.	.	.	.	.	.
4	.	.	.	.	.	.	.	.
5	.	.	.	.	.	.	.	.
7	.	.	.	.	.	.	.	.
8	.	.	.	.	.	.	.	.
10	.	.	.	.	.	.	.	.
11	.	.	.	.	.	.	.	.
13	.	.	.	.	.	.	.	.
21	.	.	.	.	.	.	.	.
26	.	0.379	.	.	.	.	.	.
37	.	.	.	.	.	.	.	.
165	.	.	.	.	.	0.651	0.749	0.615
566	1.008	1.509	0.740	1.036	0.805	0.486	.	0.390

<sup>a</sup> The scaled variance decomposition proportions smaller than 0.3 are omitted in this table.

<sup>b</sup> Total Saturated Fatty Acids.

<sup>c</sup> Total Monounsaturated Fatty Acids.

<sup>d</sup> Total Polyunsaturated Fatty Acids.

**Table 6**  
**Regression analysis output using TYPE3: SWLS**

Variable	Original Model		Reduced Model	
	Coefficient	SE <sup>a</sup>	Coefficient	SE
Intercept	24.14*** <sup>b</sup>	2.77	24.20***	2.69
Age	0.06	0.08	0.06	0.08
Black	3.19***	1.04	3.67***	0.98
on any Diet <sup>c</sup>	1.79	1.52	1.28	1.80
on Low-calorie Diet	4.09**	1.50	4.59**	1.69
on Low-fat Diet	3.67	2.86	3.87	3.76
on Low-carb Diet	0.46	3.51	0.87	3.86
Calorie	-0.88	2.36		
Protein	7.05	9.59		
Carbohydrate	3.69	9.62		
Sugar	-0.31	1.11		
Dietary Fiber	-14.52*	5.89		
Alcohol	2.09	16.47		
Total Fat	29.34	31.37	1.47*	0.68
Total Saturated Fatty Acids	-15.90	20.18		
Total Monounsaturated Fatty Acids	-22.40	23.01		
Total Polyunsaturated Fatty Acids	-27.69	21.10		
Intracluster Coefficient $\rho$	0.0366		0.0396	

<sup>a</sup> standard error.

<sup>b</sup> p-value: \*, 0.05; \*\*, 0.01; \*\*\*, 0.005.

<sup>c</sup> The reference category is “not being on diet” for all the on-diet variables here.

### 4.3 Study two: Reference level for categorical variables

As noted earlier, using non-survey data, dummy variables can also play an important role as a possible source for collinearity. The choice of reference level for a categorical variable may affect the degree of collinearity in the data. To be more specific, choosing a category that has a low frequency as the reference and omitting that level in order to

fit the model may give rise to collinearity with the intercept term. This phenomenon carries over to survey data analysis as we now illustrate.

We employed the four on-diet dummy variables used in the previous study, which we denote this section as “on any diet” (DIET), “on low-calorie diet” (CALDIET), “on low-fat diet” (FATDIET) and “one low-carbohydrate diet” (CARBDIET). The model considered here is:

$$\begin{aligned}
 \text{BMI}_{hit} = & \beta_0 + \beta_{\text{black}} * \text{black}_{hit} \\
 & + \beta_{\text{TOTAL.FAT}} * \text{TOTAL.FAT}_{hit} \\
 & + \beta_{\text{DIET}} * \text{DIET}_{hit} \\
 & + \beta_{\text{CALDIET}} * \text{CALDIET}_{hit} \\
 & + \beta_{\text{FATDIET}} * \text{FATDIET}_{hit} \\
 & + \beta_{\text{CARBDIET}} * \text{CARBDIET}_{hit} + \varepsilon_{hit} \quad (20)
 \end{aligned}$$

where subscript *hit* stands for the *t*<sup>th</sup> unit in the selected PSU *hi*, *black* is the dummy variable of black (*black* = 1 and non-black = 0), and *TOTAL.FAT* is the variable of daily total fat intake. According to the survey-weighted frequency table, 15.04% of the respondents are “on an y diet”, 11.43% of them are “on low-calorie diet”, 1.33% of them are “on low-fat diet” and 0.47% of them are “on low-carbohydrate diet”. Being on a diet is, then, relatively rare in this example. If we choose the majority level, “not being on the diet”, as the reference category for all the four on-diet dummy variables, we expect no severe collinearity between dummy variables and the intercept, because most of values in the dummy variables will be zero. However, when fitting model (20), assume that an analyst is interested to see the impact of “not on any diet” on respondent’s BMI and reverses the reference level of variable *DIET* in model (20) into “being on the diet”. This change may cause a near dependency in the model because the column in **X** for variable *DIET* will nearly equal the column of ones for the

intercept. The following empirical study will illustrate the impact of this change on the regression coefficient estimation and how we should diagnose the severity of the resulting collinearity.

Table 7 and 8 present the regression analysis output of the model in (20) using the three regression types, OLS, WLS and SWLS, listed in Table 1. Table 7 is modeling the effects of on-diet factors on BMI by treating “not being on the diet” as the reference category for all the four on-diet variables. While Table 8 changes the reference level of variable *DIET* from “not on any diet” into “On any diet” and models the effect of “not on any diet” on BMI. The choice of reference level effects the sign of the estimated coefficient for variable *DIET* but not its absolute value or standard error. The size of the estimated intercept and its SE are different in Tables 7 and 8, but the estimable functions, like predictions, will of course, be the same with either set of reference levels. The SE of the intercept is about three times larger when “on any diet” is the reference level for variable *DIET* (Table 8) than when it is not (Table 7).

When choosing “not being on any diet” as the reference category for *DIET* in Table 9, the scaled condition indexes are relatively small and do not signify any remarkable near-dependency regardless of the type of regression. Only the last row for the largest condition index is printed in Tables 9 and 10. Often, the reference category for a categorical predictor will be chosen to be analytically meaningful. In this example, using “not being on any diet” would be logical.

**Table 7**  
Regression analysis output: When “not on any diet” is the reference category for *DIET* variable in the model

Regression Type	Intercept	black	total.fat	on any diet	on low-calorie diet	on low-fat diet	on low-carb diet
TYPE1	27.22*** <sup>a</sup>	3.20***	0.95	3.03	1.75	2.75	-1.48
OLS	(0.61) <sup>b</sup>	(0.70)	(0.72)	(1.94)	(2.03)	(2.72)	(3.66)
TYPE2	26.13***	3.65***	1.44*	1.39	4.46*	3.86	0.94
WLS	(0.58)	(0.82)	(0.67)	(1.67)	(1.79)	(2.59)	(4.22)
TYPE3	26.13***	3.65***	1.44*	1.39	4.46**	3.86	0.94
SWLS	(0.64)	(0.99)	(0.63)	(1.80)	(1.70)	(3.73)	(3.87)

<sup>a</sup> p-value: \*, 0.05; \*\*, 0.01; \*\*\*, 0.005.

<sup>b</sup> Standard errors are in parentheses under parameter estimates.

**Table 8**  
Regression analysis output: When “on any diet” is the reference category for *DIET* variable in the model

Regression Type	Intercept	black	total.fat	not on any diet	on low-calorie diet	on low-fat diet	on low-carb diet
TYPE1	30.25*** <sup>a</sup>	3.20***	0.95	-3.03	1.75	2.75	-1.48
OLS	(2.00) <sup>b</sup>	(0.70)	(0.72)	(1.94)	(2.03)	(2.72)	(3.66)
TYPE2	27.52***	3.65***	1.44*	-1.39	4.46*	3.86	0.94
WLS	(1.71)	(0.82)	(0.67)	(1.67)	(1.79)	(2.59)	(4.22)
TYPE3	27.52***	3.65***	1.44*	-1.39	4.46**	3.86	0.94
SWLS	(1.75)	(0.99)	(0.63)	(1.80)	(1.70)	(3.73)	(3.87)

<sup>a</sup> p-value: \*, 0.05; \*\*, 0.01; \*\*\*, 0.005.

<sup>b</sup> Standard errors are in parentheses under parameter estimates.

In Table 10, when “on any diet” is chosen as the reference category for variable DIET, the scaled condition indexes are increased and show a moderate degree of collinearity (condition index larger than 10) between the on-diet dummy variables and the intercept. Using the table of scaled variance decomposition proportions, in OLS and WLS, dummy variable for “not on any diet” and “on low-calorie diet” are involved in the dominant near-dependency with the intercept; however, in SWLS, only the dummy variable for “not on any diet” is involved in the dominant near-dependency with the intercept and the other three on-diet variables are much less worrisome.

### 5. Conclusion

Dependence between predictors in a linear regression model fitted with survey data affects the properties of parameter estimators. The problems are the same as for non-survey data: standard errors of slope estimators can be inflated and slope estimates can have illogical signs. In the extreme case when one column of the design matrix is exactly a linear combination of others, the estimating equations cannot be solved. The more interesting cases are ones where predictors are related but the dependence is not exact. The collinearity diagnostics that are available in standard software routines are not entirely appropriate for survey data. Any diagnostic that involves variance estimation needs

modification to account for sample features like stratification, clustering, and unequal weighting. This paper adapts condition numbers and variance decompositions, which can be used to identify cases of less than exact dependence, to be applicable for survey analysis.

A condition number of a survey-weighted design matrix  $\mathbf{W}^{1/2}\mathbf{X}$  is the ratio of the maximum to the minimum eigenvalue of the matrix. The larger the condition number the more nearly singular is  $\mathbf{X}^T\mathbf{W}\mathbf{X}$ , the matrix which must be inverted when fitting a linear model. Large condition numbers are a symptom of some of the numerical problems associated with collinearity. The terms in the decomposition also involve “misspecification effects” if the model errors are not independent as would be the case in a sample with clustering. The variance of an estimator of a regression parameter can also be written as a sum of terms that involve the eigenvalues of  $\mathbf{W}^{1/2}\mathbf{X}$ . The variance decompositions for different parameter estimators can be used to identify predictors that are correlated with each other. After identifying which predictors are collinear, an analyst can decide whether the collinearity has serious enough effects on a fitted model that action should be taken. The simplest step is to drop one or more predictors, refit the model, and observe how estimates change. The tools we provide here allow this to be done in a way appropriate for survey-weighted regression models.

**Table 9**  
Largest scaled condition indexes and its associated variance decomposition proportions: When “not on any diet” is the reference category for variable DIET in the model

Scaled Condition Index	Intercept	gender	total.fat	Scaled Proportion of the Variance of			
				on any diet	on low-calorie diet	on low-fat diet	on low-carb diet
TYPE1: OLS							
6	0.005	0.000	0.016	0.949	0.932	0.157	0.200
TYPE2: WLS							
6	0.013	0.008	0.020	0.938	0.926	0.189	0.175
TYPE3: SWLS							
6	0.006	0.007	0.013	0.686	0.741	0.027	0.061

**Table 10**  
Largest scaled condition indexes and its associated variance decomposition proportions: When “on any diet” is the reference category for variable DIET in the model

Scaled Condition Index	Intercept	gender	total.fat	Scaled Proportion of the Variance of			
				not on any diet	on low-calorie diet	on low-fat diet	on low-carb diet
TYPE1: OLS							
17	0.982	0.001	0.034	0.968	0.831	0.155	0.186
TYPE2: WLS							
17	0.982	0.011	0.029	0.968	0.820	0.182	0.160
TYPE3: SWLS							
17	0.897	0.018	-0.006	0.971	0.318	0.014	-0.019

## Acknowledgements

The authors thank the associate editor and referees whose comments led to important improvements. This work was partially supported by the U.S. National Science Foundation under Grant No. 0617081. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

## References

- Belsley, D.A. (1984). Demeaning conditioning diagnostics through centering. *The American Statistician*, 38(2), 73-77.
- Belsley, D.A. (1991). *Conditioning Diagnostics, Collinearity and Weak Data in Regression*. New York: John Wiley & Sons, Inc.
- Belsley, D.A., Kuh, E. and Welsch, R.E. (1980). *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. Wiley Series in Probability and Statistics. New York: Wiley Interscience.
- Cook, R.D. (1984). Comment on demeaning conditioning diagnostics through centering. *The American Statistician*, 2, 78-79.
- Elliot, M.R. (2007). Bayesian weight trimming for generalized linear regression models. *Survey Methodology*, 33, 1, 23-34.
- Farrar, D.E., and Glauber, R.R. (1967). Multicollinearity in regression analysis. *Review of Economics and Statistics*, 49, 92-107.
- Fox, J. (1986). *Linear Statistical Models and Related Methods, with Applications to Social Research*. New York: John Wiley & Sons, Inc.
- Fuller, W.A. (2002). Regression estimation for survey samples. *Survey Methodology*, 28, 1, 5-23.
- Hendrickx, J. (2010). *perturb: Tools for evaluating collinearity*. R package version 2.04. URL <http://CRAN.R-project.org/package=perturb>.
- Kish, L., and Frankel, M. (1974). Inference from complex samples. *Journal of the Royal Statistical Society B*, 36(1), 1-37.
- Li, J. (2007a). Linear regression diagnostics in cluster samples. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 3341-3348.
- Li, J. (2007b). Regression diagnostics for complex survey data. Unpublished doctoral dissertation, University of Maryland.
- Li, J., and Valliant, R. (2009). Survey weighted hat matrix and leverages. *Survey Methodology*, 35, 1, 15-24.
- Li, J., and Valliant, R. (2011). Detecting groups of influential observations in linear regression using survey data-adapting the forward search method. Festschrift for Ken Brewer. *Pakistan Journal of Statistics*, 27, 507-528.
- Liao, D. (2010). *Collinearity Diagnostics for Complex Survey Data*. Ph.D. thesis, University of Maryland.
- Liao, D., and Valliant, R. (2012). Variance inflation factors in the analysis of complex survey data. *Survey Methodology*, 38, 1, 53-62.
- Lin, C. (1984). Extrema of quadratic forms and statistical applications. *Communications in Statistics-Theory and Methods*, 13, 1517-1520.
- Marquardt, D.W. (1980). Comment on "A critique on some ridge regression methods" by G. Smith and F. Campbell: "You should standardize the predictor variables in your regression models". *Journal of the American Statistical Association*, 75(369), 87-91.
- Scott, A.J., and Holt, D. (1982). The effect of two-stage sampling on ordinary least squares methods. *Journal of the American Statistical Association*, 77(380), 848-854.
- Silvey, S.D. (1969). Multicollinearity and imprecise estimation. *Journal of the Royal Statistical Society*, 31(3), 539-552.
- Snee, R.D., and Marquardt, D.W. (1984). Collinearity diagnostics depend on the domain of prediction, and model, and the data. *The American Statistician*, 2, 83-87.
- Steward, G.W. (1987). Collinearity and least squares regression. *Statistical Science*, 2(1), 68-84.
- Theil, H. (1971). *Principles of Econometrics*. New York: John Wiley & Sons, Inc.
- Wissmann, M., Toutenburg, H. and Shalabh (2007). Role of categorical variables in multicollinearity in the linear regression model. Technical Report Number 008, Department of Statistics, University of Munich. Available at [http://epub.ub.uni-muenchen.de/2081/1/report008\\_statistics.pdf](http://epub.ub.uni-muenchen.de/2081/1/report008_statistics.pdf).
- Wood, F.S. (1984). Effect of centering on collinearity and interpretation of the constant. *The American Statistician*, 2, 88-90.



# Bayesian inference for finite population quantiles from unequal probability samples

Qixuan Chen, Michael R. Elliott and Roderick J.A. Little<sup>1</sup>

## Abstract

This paper develops two Bayesian methods for inference about finite population quantiles of continuous survey variables from unequal probability sampling. The first method estimates cumulative distribution functions of the continuous survey variable by fitting a number of probit penalized spline regression models on the inclusion probabilities. The finite population quantiles are then obtained by inverting the estimated distribution function. This method is quite computationally demanding. The second method predicts non-sampled values by assuming a smoothly-varying relationship between the continuous survey variable and the probability of inclusion, by modeling both the mean function and the variance function using splines. The two Bayesian spline-model-based estimators yield a desirable balance between robustness and efficiency. Simulation studies show that both methods yield smaller root mean squared errors than the sample-weighted estimator and the ratio and difference estimators described by Rao, Kovar, and Mantel (RKM 1990), and are more robust to model misspecification than the regression through the origin model-based estimator described in Chambers and Dunstan (1986). When the sample size is small, the 95% credible intervals of the two new methods have closer to nominal confidence coverage than the sample-weighted estimator.

Key Words: Bayesian analysis; Cumulative distribution function; Heteroscedastic errors; Penalized spline regression; Survey samples.

## 1. Introduction

We consider inference for finite population quantiles of a continuous variable from a sample survey with unequal inclusion probabilities. The finite-population quantiles are usually estimated by the sample-weighted quantiles, a Horvitz-Thompson type estimator. Often in sample surveys the design variable (here, the inclusion probability) or a correlated auxiliary variable is measured on the non-sampled units, and this information can be used to improve the efficiency of the sample-weighted estimators (Zheng and Little 2003; Chen, Elliott, and Little 2010).

Methods for using auxiliary information in estimating finite-population distribution functions have been extensively studied. Chambers and Dunstan (1986) proposed a model-based method, illustrating their approach for a zero intercept linear regression superpopulation model. We refer to this estimator from now on as the CD estimator. Dorfman and Hall (1993) applied the CD approach, replacing the linear regression model with a non-parametric model. Lombardía, González-Manteiga, and Prada-Sánchez (2003, 2004) proposed a bootstrap approximation to these estimators based on resampling a smoothed version of the empirical distribution of the residuals. Kuk and Welsh (2001) also modified the CD approach to address departures from the model by estimating the conditional distribution of residuals as a function of the auxiliary variable. Rao, Kovar, and Mantel (RKM 1990) demonstrated advantages of

design-based ratio and difference estimators over the CD estimator when the model is misspecified. Wang and Dorfman (1996) suggested a weighted average of the CD and the RKM estimators. Kuk (1993) proposed a kernel-based estimator that combines the known distribution of the auxiliary variable with a kernel estimate of the conditional distribution of the survey variable given the value of the auxiliary variable. Chambers, Dorfman, and Wehrly (1993) proposed a kernel-smoothed model-based estimator, and Wu and Sitter (2001) and Harms and Duchesne (2006) proposed calibration type estimators.

Research on using auxiliary information for inference about finite population quantiles (defined as the inverse of the distribution function) is more limited. Chambers and Dunstan (1986) discussed estimation by inverting the CD estimator of the distribution function, but did not compare the performance of this quantile estimator with alternatives. Rao *et al.* (1990) proposed simple ratio and difference quantile estimators that were considerably more efficient than the sample-weighted estimator when the survey outcome was approximately proportional to the auxiliary variable.

We assume here unequal probability sampling with inclusion probabilities that are known for all the units in the population. We develop two Bayesian spline-model-based estimators of finite population quantiles that incorporate the inclusion probabilities. The first method is to estimate the distribution function at a number of sample values using

1. Qixuan Chen is a Assistant Professor, Department of Biostatistics, Columbia University Mailman School of Public Health, 722 West 168 Street, New York, NY 10032. E-mail: qc2138@columbia.edu; Michael R. Elliott and Roderick J.A. Little are professors, Department of Biostatistics, University of Michigan School of Public Health, 1420 Washington Heights, Ann Arbor, MI 48109. E-mail: mreliott@umich.edu and rlittle@umich.edu.

Bayesian penalized spline predictive estimators (Chen *et al.* 2010). The finite population quantiles are then estimated by inverting the predictive distribution function. The second method is a Bayesian two-moment penalized spline predictive estimator, which predicts the values of non-sampled units based on a normal model, with mean and variance both modeled with penalized splines on the inclusion probabilities. We compare the performance of these two new methods with the sample-weighted estimator, the CD estimator, and the RKM's ratio and difference estimators, using simulation studies on artificially generated data and farm survey data.

## 2. Estimators of the quantiles

Let  $s$  denote an unequal probability random sample of size  $n$ , drawn from the finite population of  $N$  identifiable units according to inclusion probabilities  $\{\pi_i, i = 1, \dots, N\}$ , which are assumed to be known for all the units before a sample is drawn. Let  $Y$  denote a continuous survey variable, with values  $\{y_1, y_2, \dots, y_n\}$  observed in the random sample  $s$ . The finite-population  $\alpha$ -quantile of  $Y$  is defined as:

$$\theta(\alpha) = \inf \left\{ t; N^{-1} \sum_{i=1}^N \Delta(t - y_i) \geq \alpha \right\}, \quad (1)$$

where  $\Delta(u) = 1$  when  $u \geq 0$  and  $\Delta(u) = 0$  elsewhere. The  $\theta(\alpha)$  is often estimated using the sample-weighted  $\alpha$ -quantile  $\hat{\theta}(\alpha) = \inf \{t, \hat{F}_w(t) \geq \alpha\}$ , where  $\hat{F}_w(t)$  is the sample-weighted distribution function given by

$$\hat{F}_w(t) = \frac{\sum_{i \in s} \pi_i^{-1} \Delta(t - y_i)}{\sum_{i \in s} \pi_i^{-1}}.$$

Woodruff (1952) proposed a method of calculating confidence limits for the sample weighted  $\alpha$ -quantile. First, a pseudo-population is obtained by weighting each sample item by its sampling weight; the standard deviation of the percentage of items less than the estimated  $\alpha$ -quantile is estimated; and the estimated standard deviation is multiplied by the appropriate  $z$  percentile and is added to and subtracted from  $\alpha$  to construct the confidence limits for the percentage of items less than the estimated  $\alpha$ -quantile. Finally, the values of the survey variable corresponding to the confidence limits of the percentage of items less than the estimated  $\alpha$ -quantile are read-off the weighted pseudo-population arrayed in order of size. Variance estimation of the percentage of items in the pseudo-population less than the estimated  $\alpha$ -quantile is discussed in Woodruff (1952). Sitter and Wu (2001) showed that the Woodruff intervals perform well even in moderate to extreme tail regions of the distribution function. An alternative variance estimate was derived by Francisco and Fuller (1991) using a smoothed version of the large-sample test inversion.

## 2.1 Bayesian model-based approach, inverting the estimated CDF

The finite population quantile function is the inverse of the finite population cumulative distribution function (CDF), defined as  $F(t) = N^{-1} \sum_{i=1}^N \Delta(t - y_i)$ , where  $\Delta(x) = 1$  when  $x \geq 0$  and  $\Delta(x) = 0$  elsewhere. We can estimate the finite population quantiles by first building a continuous and strictly monotonic predictive estimate of  $F(t)$ , by treating  $\Delta(t - y)$  as a binary outcome variable and applying methods for estimating finite population proportions.

In particular, Chen *et al.* (2010) proposed a Bayesian penalized spline predictive (BPSP) estimator for finite population proportions in unequal probability sampling. They regress the binary survey variable  $z$  on the inclusion probabilities in the sample, using the following probit penalized spline regression model (2) with  $m$  pre-selected fixed knots:

$$\Phi^{-1}(E(z_i | \beta, b, \pi_i)) = \beta_0 + \sum_{k=1}^p \beta_k \pi_i^k + \sum_{l=1}^m b_l (\pi_i - k_l)_+^p, \\ b_l \sim N(0, \tau^2). \quad (2)$$

Self-representing units are included by setting  $\pi_i = 1$ . Assuming non-informative prior distributions for  $\beta$  and  $\tau^2$ , they simulated draws of  $z$  for the non-sampled units from their posterior predictive distribution. A draw from the posterior distribution of the finite population proportion is then obtained by averaging the observed sample units and the draws of the non-sample units. This is repeated many times to simulate the posterior distribution of the finite population proportion. Simulation studies indicated that the BPSP estimator is more efficient than the sample-weighted and generalized regression estimators of the finite population proportion, with confidence coverage closer to nominal levels.

We employ the BPSP approach  $n$  times to estimate  $F(t)$  at each of the sampled values of  $y$ ,  $t = \{y_1, y_2, \dots, y_n\}$ . This estimator does not take into account the fact that we are estimating a whole distribution function, and is not necessarily a monotonic function. In addition, linear interpolation of the  $n$  estimated distribution functions may lead to a poorly-estimated CDF. To overcome these two problems, we fit a smooth cubic regression curve to the  $n$  estimated distribution functions with monotonicity constraints (Wood 1994). We denote the resulting estimated distribution function as  $\hat{F}(t)$ . The Bayesian model-based estimator of  $\theta(\alpha)$ , obtained by inverting the estimated CDF, is then defined as follows:

$$\hat{\theta}_{\text{inv-CDF}}(\alpha) = \inf \{t, \hat{F}(t) \geq \alpha\}. \quad (3)$$

We also fit two other monotonic smooth regression curves to the upper and lower limits of the 95% credible intervals (CI) of these estimated distribution functions, denoted as  $\hat{F}_U(t)$  and  $\hat{F}_L(t)$ . To reduce computation time in our simulation studies, we only estimate the CDF at  $k < n$  pre-selected sample points.

The basic idea behind this approach is shown graphically in Figure 1. Suppose a sample of size 100 is drawn from a finite population. We pick 20 observations from the sample and estimate their corresponding distribution functions and associated 95% CI using the BPSP estimator. In Figure 1(a) we plot the BPSP estimates of these 20 points with black dots and the upper and lower limits of 95% CI with “-” signs, and connect the upper and lower limits with solid lines. In Figure 1(b) we add three monotonic smooth predictive curves using black solid curve for the point estimate and black dash curves for the upper and lower limits of the 95% CI.

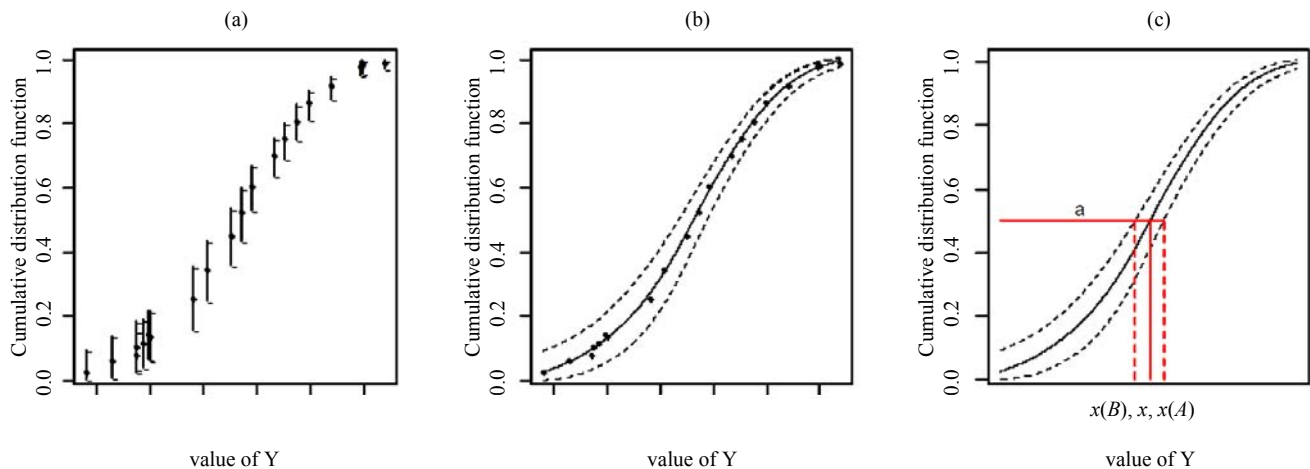
In Figure 1(c) we draw a horizontal line across the graph with  $\alpha$  as the y-axis value. We read  $x_A$ ,  $x$ , and  $x_B$  respectively from the x-axis such that  $\hat{F}_L(x_A) = \alpha$ ,  $\hat{F}(x) = \alpha$ , and  $\hat{F}_U(x_B) = \alpha$ . Then  $x$  is the inverse-CDF Bayesian estimate of  $\theta(\alpha)$ . If the 95% CI of the distribution function  $F(\cdot)$  is formed by splitting the tail areas of the posterior distribution equally, the interval formed by  $x_A$  and  $x_B$  is a 95% CI of  $\theta(\alpha)$ . The proof is as follows: If  $\alpha$  is the lower limit of the 95% CI of  $F(x_A)$ , only 2.5 percent of the draws of  $F(x_A)$  in the posterior distribution are smaller than  $\alpha$ . That is,

$$\Pr(F^{-1}(\alpha) > F^{-1}(F(x_A))) \equiv \Pr(\theta(\alpha) > x_A) = 0.025.$$

Similarly with  $\alpha$  as the upper limit of the 95% CI of  $F(x_B)$ ,  $\Pr(\theta(\alpha) < x_B) = 0.975$ . Therefore, there is 95% probability that  $\theta(\alpha)$  is within  $x_A$  and  $x_B$  in the posterior distribution, given the sample.

This inverse-CDF Bayesian model-based approach avoids strong modeling assumptions, and can be applied to normal or skewed distributions. Estimating the distribution function at all  $n$  sample units makes full use of the sample information, but is computationally intensive; estimating the distribution function at  $k < n$  values reduces computation time at the expense of some loss of efficiency. In the traditional approach, the population quantiles are estimated by inverting the unsmoothed empirical CDF. We recommend fitting a smooth cubic regression curve to the estimated distribution functions before inverting the estimated CDF. The resulting quantile estimates are more efficient, because the smooth curve exploits information from all the data. Simulations not shown here suggest that the estimated CDF distribution function curve estimated based on a well-chosen subset of the  $k$  sample units is similar to the curve estimated based on all sample units, but the computation time is significantly reduced.

We suggest choosing the subset of  $k$  data points at evenly spaced intervals in the middle of the distribution, and more frequent intervals in the extremes to improve the estimate of the CDF in the tails. For instance, in our simulation study with a sample size of 100, we estimated the distribution functions at 20 points: the 3 smallest, the 3 largest, and 14 other equally spaced points in the middle of the ordered sample.



**Figure 1** Inverse-CDF Bayesian model-based approach in estimating finite population distribution functions and associated quantiles illustrated using a sample of size 100 drawn from a finite population. (a) BPSP method is used to estimate the finite population distribution functions at 20 sample points; the dots denote BPSP estimators and the minus signs denote the upper and lower limits of the 95% CI. (b) Three monotonic smooth cubic regression models are fit on the BPSP estimators, upper limits, and lower limits; the solid curve is the predictive continuous distribution functions and the two dash curves are the 95% CI of the distribution functions. (c) The point estimate and 95% CI of population  $\alpha$ -quantile are obtained by inverting the estimated CDF;  $x$  is the point estimate, and  $x(B)$  and  $x(A)$  are the lower and upper limits of the 95% CI

### 2.2 Bayesian two-moment penalized spline predictive approach

We consider alternative estimators of finite population quantiles of the form:

$$\tilde{\theta}(\alpha) = \inf \left\{ t; N^{-1} \left( \sum_{i \in S} \Delta(t - y_i) + \sum_{j \notin S} \Delta(t - \hat{y}_j) \right) \geq \alpha \right\}, \quad (4)$$

where  $\hat{y}_j$  is the predicted value of the  $j^{\text{th}}$  non-sample unit based on a regression on the inclusion probabilities  $\{\pi_i\}$ . A basic normal model for a continuous outcome assumes a mean function that is linear in  $\{\pi_i\}$ , that is:

$$Y_i \stackrel{\text{iid}}{\sim} N(\beta_0 + \beta_1 \pi_i, c_i \sigma^2), \quad (5)$$

with known constants  $c_i$  to model non-constant variance. This leads to a biased estimate of  $\theta(\alpha)$  when the relationship is not linear. For estimating finite population totals, Zheng and Little (2003, 2005) replaced the linear mean function in (5) with a penalized spline, and assumed  $c_i = \pi_i^{2k}$  with some known value of  $k$ . Simulations suggested that their model-based estimator of the finite population total outperforms the sample-weighted estimator, even when the variance structure is misspecified.

For estimation of quantiles rather than the total, correct specification of the variance structure is important in order to avoid bias. Therefore, we extend the penalized spline model in Zheng and Little (2003) by modeling both the mean and the variance using penalized splines. The two-moment penalized spline model can be written as (Ruppert, Wand, and Carroll 2003, page 264):

$$Y_i \stackrel{\text{iid}}{\sim} N(\text{SPL}_1(\pi_i, k), \exp(\text{SPL}_2(\pi_i, k'))),$$

$$\text{SPL}_1(\pi_i, k) = \beta_0 + \sum_{k=1}^p \beta_k \pi_i^k + \sum_{l=1}^{m_1} b_l (\pi_i - k_l)_+^p,$$

$$b_l \stackrel{\text{iid}}{\sim} N(0, \tau_b^2),$$

$$\text{SPL}_2(\pi_i, k') = \alpha_0 + \sum_{k=1}^p \alpha_k \pi_i^k + \sum_{l=1}^{m_2} v_l (\pi_i - k'_l)_+^p,$$

$$v_l \stackrel{\text{iid}}{\sim} N(0, \tau_v^2). \quad (6)$$

In (6), the mean and the logarithm of the variance are modeled as penalized splines ( $\text{SPL}_1$ ) and ( $\text{SPL}_2$ ) on  $\{\pi_i\}$ . Modeling the logarithm of the variance ensures positive estimates of the variance. We allow different numbers ( $m_1, m_2$ ) and locations ( $k, k'$ ) of the knots for the two splines.

Ruppert *et al.* (2003) suggested an iterative approach to estimate the parameters in (6). They first assumed that  $\text{SPL}_2$  was known and fitted a linear mixed model to estimate the parameters in  $\text{SPL}_1$ . They calculated the square of the difference between  $Y$  and  $\text{SPL}_1$ , which followed a Gamma distribution with the shape parameter as  $1/2$  and the scale parameter of  $2\text{SPL}_2$ . They then fitted a generalized linear mixed model for the squared differences to estimate the parameters in  $\text{SPL}_2$ . They iterated the above procedures until the parameter estimates converged. This iterative approach is simple to implement. However, our goal here is not to estimate the parameters but to obtain Bayesian predictions of  $Y$  for the non-sample units so that we can use (4) to estimate the quantiles.

Crainiceanu, Ruppert, Carroll, Joshi, and Goodner (2007) developed Bayesian inferential methodology for (6). They noted that the implementation of MCMC using multivariate Metropolis-Hastings steps is unstable with poor mixing properties. They suggested adding error terms to the second spline to make computations feasible, replacing sampling from complex full conditionals by simple univariate Metropolis-Hastings steps. This idea can be expressed as

$$Y_i \stackrel{\text{iid}}{\sim} N(\text{SPL}_1(\pi_i, k), \sigma_\epsilon^2(\pi_i)),$$

$$\log(\sigma_\epsilon^2(\pi_i)) \stackrel{\text{iid}}{\sim} N(\text{SPL}_2(\pi_i, k'), \sigma_A^2).$$

We used a prior distribution  $N(0, 10^6)$  for the fixed effects parameters  $\beta$  and  $\alpha$ , and a proper inverse-gamma prior distribution  $\text{IGamma}(10^{-6}, 10^{-6})$  for the variance components  $\tau_b^2$  and  $\tau_v^2$ . We fixed the values of  $\sigma_A^2 = 0.1$ . The full conditionals of the posterior are detailed in Crainiceanu *et al.* (2007).

The posterior distribution of the finite population  $\alpha$ -quantile is simulated by generating a large number  $D$  of draws and using the predictive estimator form

$$\tilde{\theta}^{(d)}(\alpha) = \inf \left\{ t; N^{-1} \left( \sum_{i \in S} \Delta(t - y_i) + \sum_{j \notin S} \Delta(t - \hat{y}_j^{(d)}) \right) \geq \alpha \right\},$$

where  $\hat{y}_j^{(d)}$  is a draw from the posterior predictive distribution of the  $j^{\text{th}}$  non-sampled unit of the continuous outcome. The average of these draws simulates the Bayesian two-moment penalized spline predictive (B2PSP) estimator of the finite population  $\alpha$ -quantile,

$$\hat{\theta}_{\text{B2PSP}}(\alpha) = D^{-1} \sum_{d=1}^D \tilde{\theta}^{(d)}(\alpha).$$

The Bayesian 95% credible interval for the population  $\alpha$ -quantile in the simulations is formed by splitting the tail area equally between the upper and lower endpoints.

### 3. Simulation study

#### 3.1 Simulation study with artificial data

We first simulated a super-population of size  $M = 20,000$ . The size variable  $X$  in the super-population takes 20,000 consecutive integer values from 710 to 20,709. A finite population of size  $N = 2,000$  was then selected from the super-population using systematic probability proportional to size (pps) sampling with the probability proportional to the inverse of the size variable. Consequently, the size variable in the finite population has a right skewed distribution. The survey outcome  $Y$  was drawn from a normal distribution with mean  $f(\pi)$  and error variance equal to 0.04 (homoscedastic error) or  $\pi$  (heteroscedastic error). Three different mean structures  $f(\pi)$  were simulated: no association between  $Y$  and  $\pi$  (NULL)  $f(\pi) = 0.5$ , a linear association (LINUP)  $f(\pi) = 6\pi$ , and a nonlinear association (EXP)  $f(\pi) = \exp(-4.64 + 52\pi)$ . For each of the six simulation conditions, one thousand replicate finite populations were generated, and a systematic pps sample ( $n = 100$ ) was drawn from each population with  $x$  as the size variable; thus  $\pi_i = nx_i / \sum_{j=1}^N x_j$ . Scatter plots of  $Y$  versus  $\pi$  for these six populations are displayed in Figure 2.

We compared the performance of the Bayesian inverse-CDF and the B2PSP estimators with five alternative approaches:

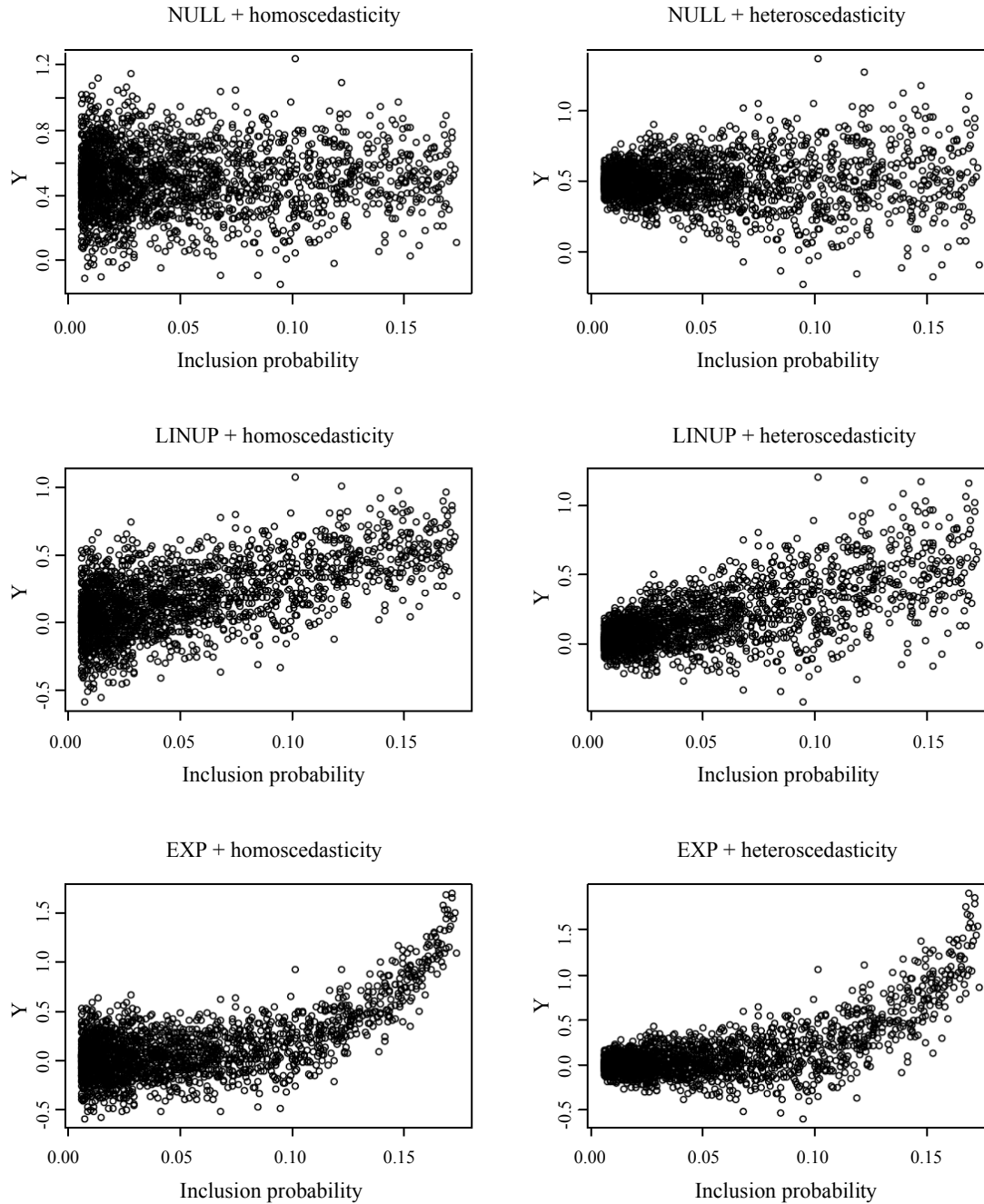
- SW, the sample-weighted estimator defined by inverting  $\hat{F}_w$ .
- Smooth-SW, the smooth sample-weighted estimator. A smooth cubic regression curve was fit to  $\hat{F}_w$ , and denoted as  $\tilde{F}_w$ . The smooth sample-weighted estimator is then defined as  $\hat{\theta}_w = \inf\{t; \tilde{F}_w \geq \alpha\}$ .
- CD, the Chambers and Dunstan estimator (1986), by assuming the following model:  $Y_i = \beta\pi_i + \sqrt{\pi_i}U_i$ , where  $U_i$  is an independent and identically distributed random variable with zero mean.
- Ratio, the RKM's ratio estimator (1990) given by  $\{\hat{\theta}_y(\alpha) / \hat{\theta}_x(\alpha)\} \times \theta_x(\alpha)$ , where  $\hat{\theta}_y(\alpha)$  and  $\hat{\theta}_x(\alpha)$  denotes respectively the sample-weighted estimates for  $Y$  and the size variable  $X$ , and  $\theta_x(\alpha)$  is the known population quantile of  $X$ .
- Diff, the RKM's difference estimator (1990) given by  $\hat{\theta}_y(\alpha) + \hat{R} \times \{\theta_x(\alpha) - \hat{\theta}_x(\alpha)\}$ , where  $\hat{R}$  is the sample-weighted estimate of  $Y/X$ .

The seven estimators for the finite-population 10<sup>th</sup>, 25<sup>th</sup>, 50<sup>th</sup>, 75<sup>th</sup>, and 90<sup>th</sup> percentiles were compared in terms of

empirical bias and root mean squared error (RMSE). Because of the complexity in the variance estimation for the CD and RKM's estimators, we only compared the average width and the non-coverage rate of the 95% confidence/credible interval (CI) for the two Bayesian model-based estimators and the sample-weighted estimator. For the 95% CI, we used Woodruff's method for the sample-weighted estimator, the method illustrated in Figure 1(c) for the inverse-CDF Bayesian estimator, and the 95% posterior probability of the quantile with equal tails for the B2PSP estimator. We used cubic splines with 15 equally spaced knots.

Tables 1 and 2 show the empirical bias and RMSE for the three normal distributions with homoscedastic errors and with heteroscedastic errors, respectively. Overall, the empirical bias in estimating the five quantiles is similar using the two Bayesian estimators, the two sample-weighted estimators, and the RKM's two design-based estimators. In contrast, the CD estimator has large bias and RMSE in all scenarios except for LINUP with heteroscedastic error, where its underlying model is correctly specified. The two Bayesian model-based estimators yield smaller root mean squared errors than the other estimators, and this improvement in efficiency is substantial in some scenarios, especially using the B2PSP estimator. By applying a smooth cubic regression curve on the estimated empirical sample-weighted CDF, the smooth-sample-weighted estimator gains some efficiency over the conventional sample-weighted estimators, but the RMSE is still larger than the Bayesian Inverse-CDF estimator. Comparisons of the three design-based estimators suggest that none of the three estimators uniformly dominates the other two. Specifically, the sample-weighted estimator has smaller RMSE than the RKM difference and ratio estimators for all five quantiles in the NULL and for the lower quantiles in the LINUP and EXP populations; on the other hand, the RKM estimators have smaller RMSE at the upper quantiles in the LINUP and EXP populations.

Table 3 shows the average width and non-coverage rate of 95% CI for the two Bayesian model-based estimators and the sample-weighted estimator. Overall, the two Bayesian model-based estimators yield shorter average 95% CI widths than the sample-weighted estimator. The coverage rate of the 95% CI is similar among the three estimators, except that when  $\alpha$  is equal to 0.1, where the 95% CI of the B2PSP estimator has the shortest average width and very good coverage, while the sample-weighted estimator has serious under-coverage. This happens because the Woodruff method for estimating the variance of the sample-weighted estimator is based on a large sample assumption, but here the pps sampling leads to only a small number of cases being sampled in the lower tail.



**Figure 2** Scatter plots of  $Y$  versus the inclusion probabilities for the six artificial finite populations of size equal to 2,000

Although the sample-weighted estimator performs similarly with the two Bayesian spline-model-based estimators in terms of overall empirical bias, the conditional bias of estimates varies largely as the sample mean of the inclusion probability increases. Following Royall and Cumberland (1981), the estimates from the 1,000 samples were ordered according to the sample mean of the inclusion probabilities and were split into 20 groups of 50 each, and then the empirical bias was calculated for each group. Figure 3

displays the conditional bias of the two Bayesian estimators and the sample-weighted estimator for the 90<sup>th</sup> percentile in the “EXP + homoscedastic error” case. Figure 3 shows that there is a linear trend for the bias in the sample-weighted estimator as the sample mean of the inclusion probabilities increases, while the grouped bias of the two Bayesian spline-model-based estimators is less affected by the sample mean of inclusion probabilities. Similar findings are also seen in other scenarios.

**Table 1**  
**Comparisons of empirical bias and root mean squared errors  $\times 10^3$  of  $\theta(\alpha)$  for  $\alpha = 0.1, 0.25, 0.5, 0.75,$  and  $0.9$ : Scenarios with homoscedastic errors**

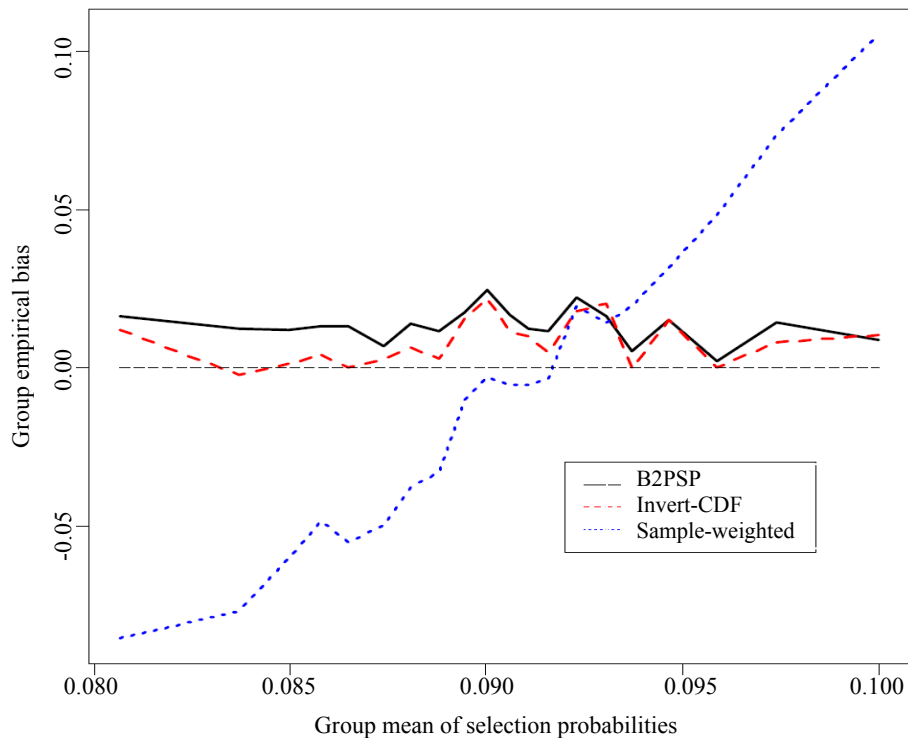
	Empirical bias					Empirical RMSE				
	0.1	0.25	0.5	0.75	0.9	0.1	0.25	0.5	0.75	0.9
<i>NULL</i>										
Inverse-CDF	-6	-3	-1	-1	-5	46	37	36	37	45
B2PSP	-5	-1	1	2	6	41	33	31	34	42
SW	-5	-3	-1	-4	-6	54	41	39	41	50
Smooth-SW	-7	-4	-1	-2	-5	50	39	37	38	47
CD	-197	-272	-265	-108	168	203	274	266	115	189
RKM's Ratio	3	25	33	16	6	77	125	159	112	79
RKM's Diff	-5	-1	6	14	14	58	58	94	122	113
<i>LINUP</i>										
Inverse-CDF	-15	-3	-2	-1	-2	70	49	39	34	33
B2PSP	-3	-1	1	4	7	56	43	35	31	29
SW	-15	-3	-3	-2	-6	77	57	48	44	42
Smooth-SW	-14	-5	-2	-1	-4	72	53	45	42	41
CD	101	35	-37	-49	1	104	38	39	53	31
RKM's Ratio	-23	-9	2	5	-0.2	95	67	53	51	40
RKM's Diff	-15	-4	-4	-0.2	-2	77	55	45	43	38
<i>EXP</i>										
Inverse-CDF	-8	0.4	4	7	4	60	45	41	43	49
B2PSP	-10	-6	-3	0.3	13	52	40	35	36	36
SW	-9	-3	-2	-2	-8	65	49	46	50	72
Smooth-SW	-12	-5	-2	-1	-2	62	47	43	46	68
CD	92	54	14	19	61	96	57	21	31	75
RKM's Ratio	-17	-11	1	3	-5	87	65	50	53	55
RKM's Diff	-9	-4	-2	-2	-7	65	49	47	47	59

**Table 2**  
**Comparisons of empirical bias and root mean squared errors  $\times 10^3$  of  $\theta(\alpha)$  for  $\alpha = 0.1, 0.25, 0.5, 0.75,$  and  $0.9$ : Scenarios with heteroscedastic errors**

	Empirical bias					Empirical RMSE				
	0.1	0.25	0.5	0.75	0.9	0.1	0.25	0.5	0.75	0.9
<i>NULL</i>										
Inverse-CDF	-9	-8	-2	4	1	30	24	22	24	31
B2PSP	-6	-6	1	7	7	25	21	19	23	27
SW	-4	-3	-2	-1	-5	34	26	23	26	35
Smooth-SW	-4	-5	-2	1	-4	34	26	23	26	35
CD	-298	-325	-253	-46	270	302	327	255	60	288
RKM's Ratio	8	31	32	16	5	81	143	154	94	57
RKM's Diff	-5	-1	6	17	16	44	54	87	113	97
<i>LINUP</i>										
Inverse-CDF	-11	-1	5	2	-3	32	24	24	29	35
B2PSP	-10	-1	7	3	1	29	22	22	24	30
SW	-5	-1	-0.1	-1	-4	31	28	33	45	51
Smooth-SW	-11	-3	2	-0.4	-5	32	26	30	44	50
CD	10	7	6	7	11	20	13	13	20	32
RKM's Ratio	-7	-3	2	3	1	36	29	30	35	41
RKM's Diff	-5	-2	-1	1	-0.2	32	27	28	33	41
<i>EXP</i>										
Inverse-CDF	-8	-3	5	7	-3	30	23	23	30	48
B2PSP	-11	-7	2	6	7	28	23	20	25	36
SW	-3	-3	-2	1	-2	30	26	26	41	84
Smooth-SW	-8	-5	1	2	-5	30	23	24	39	86
CD	18	16	35	84	68	27	21	38	88	81
RKM's Ratio	-5	-6	-1	2	-0.1	36	31	27	32	62
RKM's Diff	-3	-3	-2	1	-0.1	32	28	28	31	67

**Table 3**  
**Comparisons of average width and non-coverage rate of 95% CI  $\times 10^3$  of  $\theta(\alpha)$  for  $\alpha = 0.1, 0.25, 0.5, 0.75,$  and  $0.9$**

	Average width of 95% CI					Non-coverage rate of 95% CI				
	0.1	0.25	0.5	0.75	0.9	0.1	0.25	0.5	0.75	0.9
<i>Homoscedastic errors</i>										
<i>NULL</i>										
Inverse-CDF	199	156	141	152	184	46	35	44	38	67
B2PSP	178	134	118	134	177	52	55	61	59	50
SW	195	164	151	167	237	112	65	46	40	38
<i>LINUP</i>										
Inverse-CDF	257	207	157	139	141	61	45	37	46	52
B2PSP	230	167	134	123	121	58	54	44	57	59
SW	248	231	188	179	187	119	60	42	41	39
<i>EXP</i>										
Inverse-CDF	234	184	163	177	234	59	44	47	40	42
B2PSP	217	157	132	144	156	54	59	55	53	60
SW	231	199	175	210	402	106	64	47	40	40
<i>Heteroscedastic errors</i>										
<i>NULL</i>										
Inverse-CDF	146	104	90	101	137	42	43	38	38	47
B2PSP	107	89	79	89	107	38	49	37	68	65
SW	146	101	91	113	169	80	60	51	37	42
<i>LINUP</i>										
Inverse-CDF	131	107	104	124	154	70	31	36	42	40
B2PSP	125	97	87	93	116	47	35	50	58	52
SW	141	110	133	184	219	138	69	41	50	42
<i>EXP</i>										
Inverse-CDF	131	99	99	134	242	63	49	34	40	41
B2PSP	116	92	84	98	139	57	55	40	63	59
SW	135	100	106	186	378	111	65	46	45	34



**Figure 3** Variation of empirical bias of the three estimators for 90<sup>th</sup> percentile from the “EXP + homoscedasticity” case



### 3.2 Simulation study with the broadacre farm survey data

The B2PSP estimator assumes the outcome has a normal distribution, after conditioning on the inclusion probabilities. Since the inverse-CDF Bayesian model-based approach does not assume normality, we might expect it to out-perform the B2PSP when the normality assumption is violated. This motivates a comparison of the sample-weighted and the inverse-CDF Bayesian estimators for non-normal data.

The population considered here is defined by 398 broadacre farms (farms involved in the production of cereal crops, beef, sheep and wool) with 6,000 or less hectares that participated in the 1982 Australian Agricultural and Grazing Industries Survey carried out by the Australian Bureau of Agricultural and Resource Economics (ABARE 2003). The  $Y$  variable is the total farm cash receipts. One thousand systematic pps samples of size equal to 100 were drawn with the farm area,  $X$ , as the size variable, that is, larger farms are more likely to be selected into the sample. Figure 4 is the scatter plot of  $Y$  versus the size variable  $X$  for these

farms, with filled circles representing a selected pps sample. This shows that the variation of  $Y$  increases as  $X$  increases. Moreover,  $Y$  is right-skewed given  $X$ . A simulation study using this broadacre farms data was conducted to compare the two Bayesian spline-model-based estimators with the sample-weighted estimator.

Table 4 shows the simulation results. The inverse-CDF Bayesian approach yields smaller empirical bias and RMSE, and shorter average length of 95% CI than the sample-weighted estimator in general. The 95% CI of the inverse-CDF Bayesian approach also have closer to nominal level confidence coverage than the sample-weighted estimator when  $\alpha$  is 0.1 and 0.25. However, in the upper tail with  $\alpha = 0.90$ , the non-coverage rate of the inverse-CDF Bayesian approach is higher than the nominal level 0.05, while the Woodruff CI of the sample-weighted estimator does well. This is consistent with the findings of Sitter and Wu (2001) that the Woodruff intervals perform well even in the moderate to extreme tail regions of the distribution function. Since the conditional normality assumption is not reasonable here, the B2PSP estimator is biased and the 95% CI has poor confidence coverage.

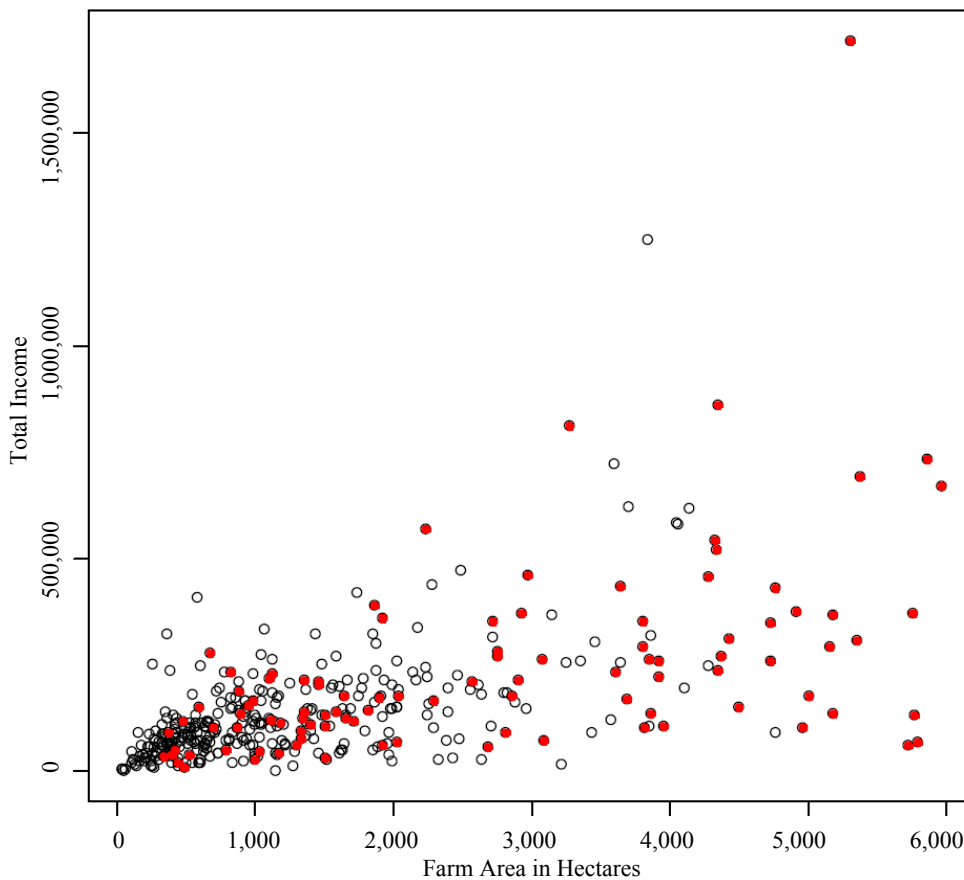


Figure 4 Scatter plot of the broadacre farm data with the filled circles representing a pps sample

**Table 4**  
**Empirical bias  $\times 10^{-2}$ , root mean squared errors  $\times 10^{-2}$ , average width of 95% CI  $\times 10^{-2}$ , and non-coverage rate of 95% CI  $\times 10^{-3}$  of  $\theta(\alpha)$  for  $\alpha = 0.1, 0.25, 0.5, 0.75,$  and  $0.9$ : The broadacre farm data**

	0.1	0.25	0.5	0.75	0.9
<i>Empirical bias</i>					
Inverse-CDF	8	14	10	-22	-60
B2PSP	-110	-125	-63	-12	88
SW	20	-19	-17	-21	-61
<i>Empirical RMSE</i>					
Inverse-CDF	117	117	108	164	256
B2PSP	113	141	124	140	206
SW	132	173	167	226	350
<i>Average width of 95% CI</i>					
Inverse-CDF	402	443	501	697	906
B2PSP	170	327	539	726	964
SW	285	468	615	864	1,589
<i>Non-coverage rate of 95% CI</i>					
Inverse-CDF	96	53	26	52	90
B2PSP	670	258	42	8	17
SW	220	121	68	42	44

#### 4. Discussion

Sample-weighted estimators for finite population quantiles are widely used in survey practice. Although the sample-weighted estimators with Woodruff's confidence intervals are easy to compute and can provide valid large-sample inferences, they may be inefficient and confidence coverage can be poor in small-to-moderate-sized samples. Model-based estimators can improve the efficiency of the estimates when the model is correctly specified, but lead to biased estimates when the model is misspecified. To achieve the balance between robustness and efficiency, we considered spline-model-based estimators. For the quantile estimation of a continuous survey variable, we can either estimate the model-based distribution functions and invert the distribution functions to obtain quantiles, or model the survey outcome on the inclusion probabilities directly. In this paper, we proposed two Bayesian spline-model-based quantile estimators. The first method is the Bayesian inverse-CDF estimator, obtained by inverting the spline-model-based estimates of distribution functions. The second method is the B2PSP estimator, estimated by assuming a normal distribution for the continuous survey outcome, with the mean function and the variance function both modeled using splines.

The simulations suggest that the two Bayesian spline-model-based estimators outperform the sample-weighted estimator, the design-based ratio and difference estimators, as well as the CD model-based estimator when its assumed model is incorrect. Both new methods yield smaller root

mean squared errors whether there is no association, a linear association, or a nonlinear association between the survey outcome and the inclusion probability. In some scenarios, the improvement in efficiency using the two Bayesian methods is substantial. When the normality assumption of the survey outcome given the inclusion probabilities is true, the B2PSP estimator has smaller RMSE and shorter credible interval than the inverse-CDF approach. Moreover, the two Bayesian model-based estimators are robust to the misspecification in both the mean and variance functions. In contrast, the CD model-based estimator is biased and inefficient when either the mean function or the variance function is misspecified. Finally, the Bayesian model-based methods have the advantage of easier calculation of the 95% CI and inference based on the posterior distributions of parameters. This is appealing, because variance estimation for the alternative design-based estimators can be complicated. Woodruff's variance estimation method for sample-weighted estimator performs well when a large fraction of the data is selected from the finite population, even in the moderate to extreme tail regions of the distribution function. However, when data from the population is sparse, the Woodruff's method tends to underestimate the confidence coverage, whereas both Bayesian methods have closer to nominal level confidence coverages.

All the three design-based estimators have comparable overall empirical bias to the two Bayesian spline-model-based estimators. However, there is a linear trend in the variation of bias for the sample-weighted estimator as the sample mean of inclusion probabilities increases. When

there is no association between the survey outcome and the inclusion probability, the ratio and difference estimators have relatively larger bias and RMSE than the sample-weighted estimator. However, in some simulation scenarios, the ratio and difference estimators achieve smaller RMSE than the sample-weighted estimator. The comparison between the conventional sample-weighted estimator and the smooth sample-weighted estimator suggests that fitting a smooth cubic curve to the sample-weighted CDF can improve the efficiency, but the smooth sample-weighted estimator still has larger RMSE than the Bayesian inverse-CDF estimator.

For normally distributed data, we recommend the use of the B2PSP estimator over the other estimators, because of smaller bias, smaller RMSE, and better confidence coverage with shorter interval length. The B2PSP estimator and its 95% posterior probability interval are easy to obtain using the algorithm proposed by Crainiceanu *et al.* (2007), which also has the advantage of relatively short computation time.

The B2PSP estimator is potentially biased when the conditional normal assumption does not hold. One possibility here is to transform the survey outcome to make the conditional normality assumption more reasonable. The B2PSP estimator can be applied to the transformed data, and the draws from the posterior distributions of the non-sampled units are transformed back to the original scale before estimating the quantiles of interest.

In our simulations with non-normal data, the inverse-CDF Bayesian approach was still more efficient than the sample-weighted estimator. Improvement in the confidence coverage was restricted to situations where the sample size is small, with Woodruff's CI method performing well when the large sample assumption holds. Thus for non-normal data where there no clear transformation to improve normality, we do not recommend the inverse-CDF Bayesian approach when the sample size is large. Given the good properties of the B2PSP estimator in the normal setting, one extension for future work is to relax the normality assumption in our proposed approaches.

We use the probability of inclusion as the auxiliary variable here. When there is only one relevant auxiliary variable, it does not matter whether the inclusion probability or the auxiliary variable is modeled. However, if there is more than one relevant auxiliary variable, the inclusion probability is the key auxiliary variable that needs to be modeled corrected, since misspecification of the model relating the survey outcome to the inclusion probability leads to bias. When other auxiliary variables are observed for all the units in the finite population, both of our Bayesian estimators can be easily extended to include additional auxiliary covariates by adding linear terms for these variables in the corresponding penalized spline model.

One reviewer suggested an alternative weighted Dirichlet approach, which is simple to calculate but it does not utilize the known auxiliary variables in the non-sampled units. Another possibility is to re-define the CD estimator by using the spline model we have used to define the B2PSP. Specifically, instead of assuming a regression model through the origin, a spline model is fitted to the first and second order moments of the conditional distribution of survey outcome given the inclusion probability. The spline-based CD estimator should perform similarly to the B2PSP estimator, and its variance can be estimated using resampling methods.

In the official statistics context, the methods in this article illustrate the potential benefits of a paradigm shift from design-based methods towards Bayesian modeling that is geared to yielding inferences with good frequentist properties. Design-based statistical colleagues raise two principal objections to this viewpoint.

First, the idea of an overtly model-based - even worse, Bayesian - approach to probability surveys is not well received, although our emphasis here is on Bayesian methods with good randomization properties. We believe that classical design-based methods do not provide the comprehensive approach needed for the complex problems that increasingly arise in official statistics. Judicious choices of well-calibrated models are needed to tackle such problems. Attention to design features and objective priors can yield Bayesian inferences that avoid subjectivity, and modeling assumptions are explicit, and hence capable of criticism and refinement. See Little (2004, 2012) for more discussion of these points.

The second objection is that Bayesian methods are too complex computationally for the official statistics world, where large number of routine statistics need to be computed correctly and created in a timely fashion. It is true that current Bayesian computation may seem forbidding to statisticians familiar with simple weighted statistics and replicate variance methods. Sedransk (2008), in an article strongly supportive of Bayesian approaches, points to the practical computational challenges as an inhibiting feature. We agree that work remains to meet this objection, but we do not view it insuperable. Research on Bayesian computation methods has exploded in recent decades, as have our computational capabilities. Bayesian models have been fitted to very large and complex problems, in some cases much more complex than those typically faced in the official statistics world.

### Acknowledgements

We thank Dr. Philip Kokic in the Commonwealth Scientific and Industrial Research Organisation for providing us the broadacre form data. We also thank an associate editor

and referees for their helpful comments on the original version of this paper.

## References

- ABARE (2003). Australian farm surveys report 2003. Canberra.
- Chambers, R.L., Dorfman, A.H. and Wehrly, T.E. (1993). Bias robust estimation in finite populations using nonparametric calibration. *Journal of American Statistical Association*, 88, 268-277.
- Chambers, R.L., and Dunstan, R. (1986). Estimating distribution functions from survey data. *Biometrika*, 73, 597-604.
- Chen, Q., Elliott, M.R. and Little, R.J.A. (2010). Bayesian penalized spline model-based inference for finite population proportion in unequal probability sampling. *Survey Methodology*, 36, 1, 23-34.
- Crainiceanu, C.M., Ruppert, D., Carroll, R.J., Joshi, A. and Goodner, B. (2007). Spatially adaptive Bayesian penalized splines with heteroscedastic error. *Journal of Computational and Graphical Statistics*, 16, 265-288.
- Dorfman, H., and Hall, P. (1993). Estimators of the finite population distribution function using nonparametric regression. *Annals of Statistics*, 21, 1452-1474.
- Francisco, C.A., and Fuller, W.A. (1991). Quantile estimation with a complex survey design. *The Annals of Statistics*, 19, 454-469.
- Harms, T., and Duchesne, P. (2006). On calibration estimation for quantiles. *Survey Methodology*, 32, 1, 37-52.
- Kuk, A.Y.C. (1993). A kernel method for estimating finite population functions using auxiliary information. *Biometrika*, 80, 385-392.
- Kuk, A.Y.C., and Welsh, A.H. (2001). Robust estimation for finite populations based on a working model. *Journal of the Royal Statistical Society, Series B*, 63, 277-292.
- Little, R.J. (2004). To model or not to model? Competing modes of inference for finite population sampling. *Journal of the American Statistical Association*, DOI: 10.1198/016214504000000467, 99, 546-556.
- Little, R.J. (2012). Calibrated Bayes: An alternative inferential paradigm for official statistics (with discussion and rejoinder). *Journal of Official Statistics*, 28, 309-334.
- Lombardía, M.J., González-Manteiga, W. and Prada-Sánchez, J.M. (2003). Bootstrapping the Chambers-Dunstan estimate of a finite population distribution function. *Journal of Statistical Planning and Inference*, 116, 367-388.
- Lombardía, M.J., González-Manteiga, W. and Prada-Sánchez, J.M. (2004). Bootstrapping the Dorfman-Hall-Chambers-Dunstan estimate of a finite population distribution function. *Journal of Nonparametric Statistics*, 16, 63-90.
- Rao, J.N.K., Kovar, J.G. and Mantel, H.J. (1990). On estimating distribution function and quantile from survey data using auxiliary information. *Biometrika*, 77, 365-375.
- Royall, R.M., and Cumberland, W.G. (1981). The finite-population linear regression estimator and estimators of its variance - An empirical study. *Journal of the American Statistical Association*, 76, 924-930.
- Ruppert, D., Wand, M.P. and Carroll, R.J. (2003). *Semiparametric Regression*. Cambridge, UK: Cambridge University Press.
- Sedransk, J. (2008). Assessing the value of Bayesian methods for inference about finite population quantities. *Journal of Official Statistics*, 24, 495-506.
- Sitter, R.R., and Wu, C. (2001). A note on Woodruff confidence intervals for quantiles. *Statistics and Probability Letters*, 52, 353-358.
- Wang, S., and Dorfman, A.H. (1996). A new estimator for the finite population distribution function. *Biometrika*, 83, 639-652.
- Wood, S.N. (1994). Monotonic smoothing splines fitted by cross validation SIAM. *Journal on Scientific Computing*, 15, 1126-1133.
- Woodruff, R. (1952). Confidence intervals for medians and other position measures. *Journal of the American Statistical Association*, 47, 635-646.
- Wu, C., and Sitter, R.R. (2001). A model-calibration approach to using complex auxiliary information from survey data. *Journal of the American Statistical Association*, 96, 185-193.
- Zheng, H., and Little, R.J.A. (2003). Penalized spline model-based estimation of finite population total from probability-proportional-to-size samples. *Journal of Official Statistics*, 19, 99-117.
- Zheng, H., and Little, R.J.A. (2005). Inference for the population total from probability-proportional-to-size samples based on predictions from a penalized spline nonparametric model. *Journal of Official Statistics*, 21, 1-20.

# Multiple imputation with census data

Satkartar K. Kinney <sup>1</sup>

## Abstract

A benefit of multiple imputation is that it allows users to make valid inferences using standard methods with simple combining rules. Existing combining rules for multivariate hypothesis tests fail when the sampling error is zero. This paper proposes modified tests for use with finite population analyses of multiply imputed census data for the applications of disclosure limitation and missing data and evaluates their frequentist properties through simulation.

Key Words: Finite Populations; Missing data; Significance testing; Synthetic data.

## 1. Introduction

Multiple imputation was first proposed for handling non-response in large complex surveys (Rubin 1987). Several other uses for multiple imputation have since been proposed, including statistical disclosure limitation and measurement error. An appeal of multiple imputation is that standard methods can be applied to each imputed dataset and then simple combining rules applied, which vary between applications. See Reiter and Raghunathan (2007) for a detailed overview of the different rules and applications. Existing multiple imputation combining rules were developed for use with random samples and superpopulation models (Deming and Stephan 1941). In finite population analyses of census data, where the sampling variance is zero, the combining rules for univariate estimands can still be applied as a special case; however, hypothesis tests for multivariate estimands break down.

Motivated by the use of multiple imputation to generate partially synthetic data (Rubin 1993; Little 1993) for the U.S. Census Bureau's Longitudinal Business Database (Kinney, Reiter, Reznek, Miranda, Jarmin and Abowd 2011), an economic census, this paper derives a multivariate test for finite populations for use with partially synthetic data and extends it to the application of missing data. Extensions to other multiple imputation applications are expected to be straightforward.

The remainder of this paper is organized as follows. Section 2 describes the case of partially synthetic data and Section 3 presents the extension to missing data. Simulations in Section 4 evaluate the combining rules for both the missing data and partially synthetic data cases.

## 2. Partially synthetic data

Partially synthetic datasets are constructed by replacing selected values in the confidential data with  $m$  independent draws from their posterior predictive distribution. For a

finite population of size  $N$ , let  $Z_j = 1, j = 1, \dots, N$  indicate that unit  $j$  has been selected to have any observed values replaced with imputations. Imputations should only be made from the posterior predictive distribution of those units with  $Z_j = 1$ . For simplicity, in this paper, we assume  $Z_j = 1, j = 1, \dots, N$ . Let  $Y = (y_1, \dots, y_d)$  be the matrix of confidential variables that will be replaced with imputations and  $X$  the matrix of variables that will not be replaced. Let  $D_{\text{cen}} = (X, Y)$  represent a census of all  $N$  units containing confidential data and assume that all units are fully observed, *i.e.*, no missing values are present. Let  $Y_{\text{rep}}^{(i)}, i = 1, \dots, m$  be the  $i^{\text{th}}$  imputation of  $Y$ , and let  $D_{\text{syn}}^{(i)} = (X, Y_{\text{rep}}^{(i)})$ . The set  $D_{\text{syn}} = \{D_{\text{syn}}^{(i)}, i = 1, \dots, m\}$  is what is released to the public.

Any proper imputation procedure from the broad literature on multiple imputation may be used to generate  $D_{\text{syn}}$  from  $D_{\text{cen}}$ . The finite population methods proposed here can be used regardless of whether a finite population was assumed in the generation of  $D_{\text{syn}}$ . Under a finite population assumption, since the data are a fully observed census the imputation model parameters would be considered known and fixed. See Reiter and Kinney (2012) for an illustration of how valid inferences are obtained from partially synthetic random samples generated with both fixed and random imputation model parameters. Simulations (not shown) confirm the same is true in the finite population case.

An analyst with access to  $D_{\text{syn}}$  but not  $D_{\text{cen}}$  can obtain valid inferences for a scalar or vector estimand  $Q$  using the following quantities:

$$\bar{Q}_m = \frac{1}{m} \sum_{i=1}^m Q^{(i)} \quad (2.1)$$

$$\bar{U}_m = \frac{1}{m} \sum_{i=1}^m U^{(i)} \quad (2.2)$$

$$B_m = \frac{1}{m-1} \sum_{i=1}^m (Q^{(i)} - \bar{Q}_m) (Q^{(i)} - \bar{Q}_m)' \quad (2.3)$$

1. Satkartar K. Kinney, National Institute of Statistical Sciences, Research Triangle Park, NC 27709, U.S.A. E-mail: saki@niss.org.

where  $Q^{(i)}$ ,  $i = 1, \dots, m$ , is the point estimate of  $Q$  obtained from  $D_{\text{syn}}^{(i)}$ ,  $U^{(i)}$  is the estimated variance of  $Q$ , and  $B_m$  is the sample variance of the  $Q^{(i)}$ ,  $i = 1, \dots, m$ .

When there is no sampling variance the combining rules for scalar  $Q$  derived by Reiter (2003) can be applied as a special case where  $\bar{U}_m = 0$ . The resulting simplification means the approximations of Reiter (2003) are not needed and the exact posterior under multivariate normal theory is  $(Q | D_{\text{syn}}) \sim t_{m-1}(\bar{Q}_m, B_m / m)$ . For a vector  $Q$ , however, the hypothesis test of Reiter (2005) relies on the assumption that  $B_\infty$  is proportional to  $\bar{U}_\infty$ , i.e., the proportion of information replaced with imputations is the same across components of  $Q$ , so a different assumption is needed for the case  $\bar{U}_\infty = 0$ .

### 2.1 Proposed multivariate test

In this section an alternate test is derived based on the stronger assumption that  $B_\infty = r_\infty I$ , for a scalar quantity  $r_\infty$  and  $k$ -dimensional identity matrix  $I$ . In other words, the between-imputation variance is constant across components of  $Q$ , and  $B_\infty$  is assumed to be diagonal. In both the Reiter (2005) test and the proposed test, one averages across variance components so the test is moderately robust to this assumption; however, the randomization validity declines when the estimates of  $Q$ ,  $\bar{Q}^{(i)}$ ,  $i = 1, \dots, m$ , are highly correlated. This is evaluated with simulations in Section 4.3. Comparable tests based on the assumption  $B_\infty \propto \bar{U}_\infty$  are known to lose power when the assumption is not met (Li et al. 1991).

The proposed test for the hypothesis  $H_0: Q = Q_0$  is conducted by referring the test statistic

$$S_c = \frac{(Q_0 - \bar{Q}_m)'(Q_0 - \bar{Q}_m)}{kr_c}$$

to an  $F_{k, k(m-1)}$  distribution, where  $r_c = 1 / m \text{tr}(B_m) / k$ .

Under the assumption  $B_\infty = r_\infty I$ , the Bayesian  $p$ -value is given by

$$\begin{aligned} & \int P(\chi_k^2 > (Q_0 - \bar{Q})' T_\infty^{-1} (Q_0 - \bar{Q}) | D_{\text{syn}}, B_\infty) \\ & \quad P(B_\infty | D_{\text{syn}}) dB_\infty \quad (2.4) \\ & = \int P\left(\chi_k^2 > \frac{(Q_0 - \bar{Q})' I (Q_0 - \bar{Q})}{r_\infty / m} \mid D_{\text{syn}}, r_\infty\right) \\ & \quad P(r_\infty | D_{\text{syn}}) dr_\infty \\ & = \int P\left(\frac{\chi_k^2}{k} \cdot \frac{r_\infty}{mr_c} > S_c \mid D_{\text{syn}}, r_\infty\right) \\ & \quad P(r_\infty | D_{\text{syn}}) dr_\infty. \quad (2.5) \end{aligned}$$

Thus the proportionality assumption reduces the number of variance parameters to be estimated from  $k(k-1)/2$  to 1 and allows for the closed-form approximation of the integral in (2.4). As  $\bar{U}_\infty = 0$ , the derivation is simplified from Reiter (2005). To complete the integration, we need the distribution of  $(r_\infty | D_{\text{syn}})$ . Extending the scalar case in Reiter (2003), the sampling distribution of  $Q^{(i)}$ , the estimate of  $Q$  obtained from  $D_{\text{syn}}^{(i)}$ , is given by  $(Q^{(i)} | Q_{\text{cen}}, B_\infty) \sim N(Q_{\text{cen}}, B_\infty)$ . Under the proportionality assumption, this becomes  $(Q^{(i)} | Q_{\text{cen}}, r_\infty) \sim N(Q_{\text{cen}}, r_\infty I)$ . With diffuse priors and standard multivariate normal theory for sample covariance matrices, we obtain

$$(m-1) \frac{\sum_{i=1}^m (Q^{(i)} - \bar{Q}_m)(Q^{(i)} - \bar{Q}_m)'}{(m-1)r_\infty} \mid D_{\text{syn}} \sim \text{Wish}(m-1, I).$$

Taking the trace of each side and integrating over  $r_\infty$  in (2.5) yields a Bayesian  $p$ -value of

$$P\left(\frac{\chi_k^2}{k} \frac{k(m-1)}{\chi_{k(m-1)}^2} > S_c \mid D_{\text{syn}}\right) = P(F_{k, k(m-1)} > S_c \mid D_{\text{syn}}).$$

### 3. Missing data

The extension to missing data is straightforward. When  $\bar{U}_\infty = 0$ , the combining rules (Rubin 1987) for scalar estimands  $q$  simplify so that  $(q | D_{\text{com}}) \sim N(\bar{q}_m, (1 + 1/m)B_m)$ , where  $D_{\text{com}}$  is the set of  $m$  completed datasets. Similar to Section 2, the tests of Rubin (1987) and Li, Raghunathan and Rubin (1991) for multivariate components rely on the assumption that  $B_\infty \propto \bar{U}_\infty$ , and so when  $\bar{U}_\infty = 0$  we derive a test under the assumption  $B_\infty = r_\infty I$ .

Following derivation procedures similar to that of Section 2.1, the Bayesian  $p$ -value for testing  $H: Q = Q_0$  with  $k$ -dimensional  $Q$  is found to be  $P(F_{k, k(m-1)} > S_q | D_{\text{com}})$  where

$$S_q = \frac{(Q_0 - \bar{Q}_m)'(Q_0 - \bar{Q}_m)}{kr_q},$$

and  $r_q = (1 + 1/m) \text{tr}(B_m) / k$ .

### 4. Simulation study

In this section, simple simulation examples illustrate the analytic validity of the proposed combining rules, first for the case of partially synthetic data, and then for the case missing data. Lastly, the robustness of the tests to the proportionality assumption is evaluated.

For a population of  $N = 50,000$ ,  $X = (X_1, \dots, X_{20})$  is drawn from a multivariate normal distribution with mean zero and covariance matrix with 1 in each diagonal element and 0.5 in each off-diagonal element.  $Y$  is drawn from a standard normal distribution. For each of 5,000 iterations, a new finite population is generated and  $m$  imputations are drawn for  $m \in \{2, 5, 10\}$ . The proposed hypothesis tests are conducted for  $H_0: Q = Q_0$ , where  $Q$  is the vector of regression coefficients, excluding the intercept, of the regression of  $Y$  on  $X$  and has dimension  $k$ ,  $k \in \{2, 5, 20\}$ , and  $Q_0$  is the true value of  $Q$  determined from the finite population  $(X, Y)$ . Since  $H_0$  is true by design,  $H_0$  should be rejected  $100\alpha\%$  of the time, for significance level  $\alpha = 0.05$ .

Random sampling scenarios are also simulated for comparison purposes. At each iteration, a random sample of size  $s = 50,000$  from an infinite population is generated from the distributions described above, prior to generating the  $m$  missing data and synthetic imputations. The same hypothesis  $H_0: Q = Q_0$  is tested where  $Q_0$  is the vector of true population values. The combining rules for the hypothesis tests are those of Reiter (2005) in the synthetic data case and Li *et al.* (1991) and Rubin (1987) in the missing data case.

**4.1 Partially synthetic data imputations**

Let  $Y$  be a confidential response variable and  $X$  be unreplaced predictors. Then  $Y_{syn}$  is generated by taking  $m$  independent draws from the posterior predictive distribution  $f(Y | X)$  assuming a normal linear model, using all available data.

Table 1 gives the nominal 5% rejection rate for the proposed hypothesis test for multicomponent estimands, which are seen to be close to the significance level 0.05, and close to the random sampling results. From these results it appears that the proposed combining rules for population data have good frequentist properties. Not shown are the rejection rates when the rules from random samples (Reiter 2005) were applied to finite populations, which were observed to be quite high, typically 1, in the simulations conducted.

**Table 1**  
Comparison of nominal 5% rejection rates for tests on partially synthetic data

	$k = 2$	$k = 5$	$k = 20$
Census data			
$m = 2$	0.048	0.065	0.052
$m = 5$	0.048	0.061	0.057
$m = 10$	0.051	0.067	0.055
Random sampling			
$m = 2$	0.067	0.062	0.060
$m = 5$	0.054	0.052	0.050
$m = 10$	0.047	0.049	0.049

**4.2 Missing data**

Simulations analogous to the synthetic data simulations were conducted for the missing data case. The missing values of  $Y$  are imputed from the posterior predictive distribution  $f(Y_{obs} | X)$  assuming a normal linear model. Missingness is simulated to be completely at random, with  $P(R_l = 1) = 0.3$ ,  $l = 1, \dots, s$ , where  $R$  is an indicator variable for missingness.

Table 2 gives the nominal 5% rejection rate for the proposed hypothesis test for multicomponent estimands, which are seen to be close to 0.05, and to the random sampling results. From these results it appears that the proposed combining rules for population data yield valid inferences.

**Table 2**  
Comparison of nominal 5% rejection rates for tests using completed census data

	$k = 2$	$k = 5$	$k = 20$
Census data			
$m = 2$	0.052	0.061	0.053
$m = 5$	0.048	0.063	0.051
$m = 10$	0.048	0.058	0.054
Random sampling			
$m = 2$	0.061	0.056	0.053
$m = 5$	0.056	0.052	0.052
$m = 10$	0.048	0.050	0.051

**4.3 Robustness**

The assumption that  $B_\infty \propto r_\infty I$  is striking at first glance, and is unlikely to be exactly true. In this section we evaluate the effect of strong correlations across components of  $Q$ . While moderately strong correlations were present in the previous simulations, here we increase the magnitude of the between-imputation variance, increasing the magnitude of the differences across the diagonal of  $B$  as well as the distance from zero of the off-diagonal elements of  $B$ .

These simulations are set up as before, for the finite population case, with  $k = 5$  and  $m = 5$ . The population in each iteration is generated in the same way as before, except that we let  $Y = (1, 2, 5, 10, 20, 0, \dots, 0) (X_1, X_2, \dots, X_{20})' + \eta$ ,  $\eta \sim N(0, 100)$  and  $X_2 = c \cdot X_1 + \varepsilon$ ,  $c \in \{0.5, 1, 5\}$  and  $\varepsilon \sim N(0, 1)$ . Increasing values of  $c$  yields increasingly higher correlations. The large variance for  $\eta$  induces larger and more variable values for elements of  $B$ .

The results in Table 3 indicate that while the tests have good properties even with moderately high violations of the proportionality assumption, their performance declines with increasingly large correlations. Continuing our assumption that  $Q$  represents a vector of regression coefficients, presence of such large correlation may also be indicative of multicollinearity in the model at hand, so analysts faced with high correlation across  $\bar{Q}^{(i)}$  might take steps to reduce multicollinearity before applying the proposed tests. If

variables are of substantially differing magnitude, standardization to rescale them will reduce differences across  $Q$ .

**Table 3**  
Evaluation of tests under assumption violations,  $k = 5, m = 5$

	$c = 0.5$	$c = 1$	$c = 5$
Synthetic Data	0.059	0.083	0.145
Missing Data	0.051	0.083	0.136

### Acknowledgements

A portion of this work was conducted while the author was a student at Duke University, supported by NSF grant ITR-0427889 and under the guidance of Jerry Reiter, whose assistance is greatly appreciated. In addition, the comments of anonymous reviewers were quite helpful.

### References

Deming, W.E., and Stephan, F.F. (1941). On the interpretation of censuses as samples. *Journal of the American Statistical Association*, 36, 213, 45-49.

Kinney, S.K., Reiter, J.P., Reznick, A.P., Miranda, J., Jarmin, R.S. and Abowd, J.M. (2011). Toward unrestricted public-use business microdata: The Longitudinal Business Database. *International Statistical Review*, 79, 3, 362-384.

Li, K.H., Raghunathan, T.E. and Rubin, D.B. (1991). Large-sample significance levels from multiply-imputed data using moment-based statistics and an  $F$  reference distribution. *Journal of the American Statistical Association*, 86, 1065-1073.

Little, R.J.A. (1993). Statistical analysis of masked data. *Journal of Official Statistics*, 9, 407-426.

Reiter, J.P. (2003). Inference for partially synthetic, public use microdata sets. *Survey Methodology*, 29, 2, 181-188.

Reiter, J.P. (2005). Significance tests for multi-component estimands from multiply-imputed, synthetic microdata. *Journal of Statistical Planning and Inference*, 131, 365-377.

Reiter, J.P., and Kinney, S.K. (2012). Inferentially valid, partially synthetic data: Generating from posterior predictive distributions not necessary. Technical report, National Institute of Statistical Sciences.

Reiter, J.P., and Raghunathan, T.E. (2007). The multiple adaptations of multiple imputation. *Journal of the American Statistical Association*, 102, 1462-1471.



**NOTICE**

Statistics Canada will be discontinuing its practice to print *Survey Methodology*. This current issue (December 2012 – volume 38 number 2) will be the last version available in print form. Please note that the electronic version of *Survey Methodology* will continue to be available free of charge on the Statistics Canada website, [www.statcan.gc.ca](http://www.statcan.gc.ca).

The next issue is to be published in June 2013 in electronic format and will maintain our high standard of content.

You may subscribe to “My Account” on Statistics Canada website to receive email notifications when new issues of the journal are released.

**CORRIGENDUM**

James Chipperfield and John Preston

“Efficient bootstrap for business surveys”, vol. 33, no. 2 (December 2007), 167-172.

In Section 4.2 of this paper, under the equation

$$\text{Var}(\hat{v}_{\text{boot}}) = \text{Var}_s(E_*[\hat{v}_{\text{boot}}|s]) + E_s(\text{Var}_*[\hat{v}_{\text{boot}}|s]),$$

there are five references to the term

$$\text{Var}_s(E_*[\hat{v}_{\text{boot}}|s]).$$

To be correct, these five referenced terms should be replaced by

$$E_s(\text{Var}_*[\hat{v}_{\text{boot}}|s]).$$

## ACKNOWLEDGEMENTS

*Survey Methodology* wishes to thank the following people who have provided help or served as referees for one or more papers during 2012.

S.R. Amer, *RTI International*  
 T. Asparouhov, *Mplus*  
 M. Barron, *NORC*  
 W. Bell, *U.S. Census Bureau*  
 E. Berg, *National Agricultural Statistical Services*  
 P. Biemer, *RTI*  
 I. Bilgen, *NORC*  
 C. Bocci, *Statistics Canada*  
 J. van den Brakel, *Statistics Netherlands*  
 M. Brick, *Westat, Inc.*  
 R. Bruni, *University of Rome, La Sapienza*  
 C.-T. Chao, *National Cheng-Kung University, Taiwan*  
 G. Chauvet, *CREST-ENSAI*  
 J. Chipperfield, *Australian Bureau of Statistics*  
 G. Datta, *University of Georgia*  
 M. Davern, *NORC*  
 T. DeWaal, *Statistics Netherlands*  
 D. Dolson, *Statistics Canada*  
 S. Eckman, *Institute for Employment Research, Germany*  
 S. Er, *Istanbul University*  
 E. Escobar, *University of Southampton*  
 V. Estevao, *Statistics Canada*  
 O.P. Fischer, *U.S. Census Bureau*  
 J. Gambino, *Statistics Canada*  
 N. Ganesh, *NORC at University of Chicago*  
 T.I. Garner, *U.S. Bureau of Labor Statistics*  
 J. Garrett, *Knowledge Networks, Inc.*  
 C. Goga, *Université de Bourgogne*  
 M. Graf, *Office fédéral de la Statistique, Suisse*  
 B. Hulliger, *University of Applied Sciences Northwestern Switzerland*  
 D. Kasprzyk, *NORC at the University of Chicago*  
 C. Kennedy, *Abt SRBI*  
 M.G.M. Khan, *University of the South Pacific, Fiji*  
 J.-K. Kim, *Iowa State University*  
 P. Kott, *RTI*  
 P. Lavallée, *Statistics Canada*  
 F. Li, *Duke University*  
 J. Li, *Westat Inc.*  
 P. Lugtig, *Utrecht University*  
 P. Lynn, *University of Essex*  
 D. Malec, *National Center for Health Statistics*  
 H. Mantel, *Statistics Canada*  
 I. Molina, *Universidad Carlos III de Madrid*  
 R. Münnich, *Economic and Social Statistics Dept. Univ. of Trier, Germany*  
 J. Oleson, *University of Iowa*  
 A.J. O'Malley, *Harvard Medical School*  
 J. Opsomer, *Colorado State University*  
 V. Parsons, *National Center for Health Statistics*  
 D. Pfeffermann, *Hebrew University*  
 F. van de Pol, *Statistics Netherlands*  
 N.G.N. Prasad, *University of Alberta*  
 L. Qualité, *Université de Neuchâtel*  
 T. Raghunathan, *University of Michigan*  
 J.N.K. Rao, *Carleton University*  
 J. Reiter, *Duke University*  
 L.-P. Rivest, *Université Laval*  
 R. Rodriguez, *U.S. Census Bureau*  
 K. Rust, *Westat, Inc.*  
 E. Saleh, *Carleton University*  
 F. Scheuren, *NORC*  
 A. Scott, *University of Auckland*  
 J. Sedransk, *Case Western Reserve University & University of Maryland*  
 P. do N. Silva, *Escola Nacional de Ciências Estatísticas*  
 R. Sigman, *Westat Inc.*  
 A. Singh, *NORC*  
 C. Skinner, *London School of Economics*  
 P.A. Smith, *Office for National Statistics*  
 P.W.F. Smith, *University of Southampton*  
 N. Thomas, *Pfizer*  
 R. Thomas, *Carleton University*  
 K.J. Thompson, *U.S. Census Bureau*  
 M. Thompson, *University of Waterloo*  
 Y. Tillé, *Université de Neuchâtel*  
 V. Toepoel, *Tilburg University*  
 M. Torabi, *University of Manitoba*  
 V. Vehovar, *University of Ljubljana*  
 J. Vermunt, *Tilburg School of Social and Behavioral Sciences*  
 M. de Toledo Vieira, *Universidade Federal de Juiz de Fora, Brazil*  
 J. Wagner, *University of Michigan*  
 K. Wolter, *NORC*  
 C. Wu, *University of Waterloo*  
 C. Yu, *Iowa State University*  
 W. Yung, *Statistics Canada*  
 E. Zanutto, *National Analysts Worldwide*

Acknowledgements are also due to those who assisted during the production of the 2012 issues: Céline Ethier of Statistical Research and Innovation Division, Christine Cousineau of Household Survey Methods Division, Nick Budko and Annette Everett of Business Survey Methods Division, Anne-Marie Fleury of Operations and Integration Division, Roberto Guido, Liliane Lanoie, Darquise Pellerin, Joseph Prince, Jacqueline Luffman, Suzanne Bélair, Janice Burt, Jeff Campbell, Kathy Charbonneau and Fadi Salibi of Dissemination Division.

**ELECTRONIC  
PUBLICATIONS  
AVAILABLE AT**

**PUBLICATIONS  
ÉLECTRONIQUES  
DISPONIBLE À**

**[www.statcan.gc.ca](http://www.statcan.gc.ca)**

## ANNOUNCEMENTS

### Nominations Sought for the 2014 Waksberg Award

The journal *Survey Methodology* has established an annual invited paper series in honour of Joseph Waksberg to recognize his contributions to survey methodology. Each year a prominent survey statistician is chosen to write a paper that reviews the development and current state of an important topic in the field of survey methodology. The paper reflects the mixture of theory and practice that characterized Joseph Waksberg's work.

The recipient of the Waksberg Award will receive an honorarium and give the 2014 Waksberg Invited Address at the Statistics Canada Symposium to be held in the autumn of 2014. The paper will be published in a future issue of *Survey Methodology* (targeted for December 2014).

The author of the 2014 Waksberg paper will be selected by a four-person committee appointed by *Survey Methodology* and the American Statistical Association. Nomination of individuals to be considered as authors or suggestions for topics should be sent before February 28, 2013 to the chair of the committee, Steve Heeringa (sheering@isr.umich.edu).

Previous Waksberg Award honorees and their invited papers are:

- 2001 Gad **Nathan**, "Telesurvey methodologies for household surveys – A review and some thoughts for the future?". *Survey Methodology*, vol. 27, 1, 7-31.
- 2002 Wayne A. **Fuller**, "Regression estimation for survey samples". *Survey Methodology*, vol. 28, 1, 5-23.
- 2003 David **Holt**, "Methodological issues in the development and use of statistical indicators for international comparisons". *Survey Methodology*, vol. 29, 1, 5-17.
- 2004 Norman M. **Bradburn**, "Understanding the question-answer process". *Survey Methodology*, vol. 30, 1, 5-15.
- 2005 J.N.K. **Rao**, "Interplay between sample survey theory and practice: An appraisal". *Survey Methodology*, vol. 31, 2, 117-138.
- 2006 Alastair **Scott**, "Population-based case control studies". *Survey Methodology*, vol. 32, 2, 123-132.
- 2007 Carl-Erik **Särndal**, "The calibration approach in survey theory and practice". *Survey Methodology*, vol. 33, 2, 99-119.
- 2008 Mary E. **Thompson**, "International surveys: Motives and methodologies". *Survey Methodology*, vol. 34, 2, 131-141.
- 2009 Graham **Kalton**, "Methods for oversampling rare subpopulations in social surveys". *Survey Methodology*, vol. 35, 2, 125-141.
- 2010 Ivan P. **Fellegi**, "The organisation of statistical methodology and methodological research in national statistical offices". *Survey Methodology*, vol. 36, 2, 123-130.
- 2011 Danny **Pfeffermann**, "Modelling of complex survey data: Why model? Why is it a problem? How can we approach it?". *Survey Methodology*, vol. 37, 2, 115-136.
- 2012 Lars **Lyberg**, "Survey Quality". *Survey Methodology*, vol. 38, 2, 107-130.
- 2013 Ken **Brewer**, Manuscript topic under consideration.

**Members of the Waksberg Paper Selection Committee (2012-2013)**

Steve Heeringa, *University of Michigan* (Chair)

Cynthia Clark, *USDA*

Louis-Paul Rivest, *Université de Laval*

J.N.K. Rao, *Carleton University*

**Past Chairs:**

Graham Kalton (1999 - 2001)

Chris Skinner (2001 - 2002)

David A. Binder (2002 - 2003)

J. Michael Brick (2003 - 2004)

David R. Bellhouse (2004 - 2005)

Gordon Brackstone (2005 - 2006)

Sharon Lohr (2006 - 2007)

Robert Groves (2007 - 2008)

Leyla Mojadjer (2008 - 2009)

Daniel Kasprzyk (2009 - 2010)

Elizabeth A. Martin (2010 - 2011)

Mary E. Thompson (2011 - 2012)

# 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 28, No. 2, 2012

Collecting Survey Data During Armed Conflict William G. Axinn, Dirgha Ghimire, Nathalie E. Williams .....	153
Sources of Error in Survey and Administrative Data: The Importance of Reporting Procedures Jeffrey A. Groen.....	173
Management Challenges of the 2010 U.S. Census Daniel H. Weinberg.....	199
Response Rates in Business Surveys: Going Beyond the Usual Performance Measure Katherine Jenny Thompson, Broderick E. Oliver.....	221
Calibration Inspired by Semiparametric Regression as a Treatment for Nonresponse Giorgio E. Montanari, M. Giovanna Ranalli .....	239
Strategy for Modelling Nonrandom Missing Data Mechanisms in Observational Studies Using Bayesian Methods Alexina Mason, Sylvia Richardson, Ian Plewis, Nicky Best.....	279
Book Reviews.....	303

All inquires about submissions and subscriptions should be directed to [jos@scb.se](mailto:jos@scb.se)

# 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 28, No. 3, 2012

Calibrated Bayes, an Alternative Inferential Paradigm for Official Statistics Roderick J. Little.....	309
Discussion	
Jean-François Beaumont.....	335
Philippe Brion .....	341
Alan H. Dorfman .....	349
Risto Lehtonen .....	353
Paul A. Smith .....	359
Michael P. Cohen.....	363
Rejoinder	
Roderick J. Little.....	367
Improving RDD Cell Phone Samples. Evaluation of Different Pre-call Validation Methods Tanja Kunz, Marek Fuchs .....	373
Mutual Information as a Measure of Intercoder Agreement Ben Klemens.....	395
The Organization of Information in a Statistical Office Tjalling Gelsema.....	413
Unit Root Properties of Seasonal Adjustment and Related Filters William R. Bell .....	441
Book Review .....	463
In Other Journals .....	469

All inquiries about submissions and subscriptions should be directed to [jos@scb.se](mailto:jos@scb.se)



**Volume 40, No. 2, June/juin 2012**

Hui Song, Yingwei Peng and Dongsheng Tu A new approach for joint modelling of longitudinal measurements and survival times with a cure fraction .....	207
Georgios Papageorgiou Restricted maximum likelihood estimation of joint mean-covariance models.....	225
Karelyn A. Davis, Chul G. Park and Sanjoy K. Sinha Testing for generalized linear mixed models with cluster correlated data under linear inequality constraints.....	243
David Haziza and Frédéric Picard Doubly robust point and variance estimation in the presence of imputed survey data.....	259
Jieli Ding, Yanyan Liu, David B. Peden, Steven R. Kleeberger and Haibo Zhou Regression analysis for a summed missing data problem under an outcome-dependent sampling scheme .....	282
Hannes Kazianka and Jürgen Pilz Objective Bayesian analysis of spatial data with uncertain nugget and range parameters .....	304
Tingting Zhang and Jun S. Liu Nonparametric hierarchical Bayes analysis of binomial data via Bernstein polynomial priors .....	328
Kei Hirose and Sadanori Konishi Variable selection via the weighted group lasso for factor analysis models .....	345
Zhibiao Zhao and Weixin Yao Sequential design for nonparametric inference .....	362
José R. Berrendero, Antonio Cuevas and Beatriz Pateiro-López Testing uniformity for the case of a planar unknown support.....	378
Acknowledgement of referees' services/Remerciements aux membres des jurys.....	396

**Volume 40, No. 3, September/septembre 2012**

Yulia R. Gel and Bei Chen Robust Lagrange multiplier test for detecting ARCH/GARCH effect using permutation and bootstrap.....	405
Florian Ketterer and Hajo Holzmann Testing for intercept-scale switch in linear autoregression .....	427
Pierre Duchesne, Kilani Ghoudi and Bruno Rémillard On testing for independence between the innovations of several time series .....	447
Ivan Kojadinovic and Jun Yan Goodness-of-fit testing based on a weighted bootstrap: A fast large-sample alternative to the parametric bootstrap .....	480
Ramon Oller and Guadalupe Gómez A generalized Fleming and Harrington's class of tests for interval-censored data .....	501
Carlotta Ching Ting Fok, James O. Ramsay, Michal Abrahamowicz and Paul Fortin A functional marked point process model for lupus data .....	517
Grace Y. Yi and Jerald F. Lawless Likelihood-based and marginal inference methods for recurrent event data with covariate measurement error .....	530
Hongjian Zhu and Feifang Hu Interim analysis of clinical trials based on urn models .....	550
Zhong Guan, Jing Qin and Biao Zhang Information borrowing methods for covariate-adjusted ROC curve .....	569
Jiming Jiang and Thuan Nguyen Small area estimation via heteroscedastic nested-error regression.....	588
Jae Kwang Kim and Minki Hong Imputation for statistical inference with coarse data .....	604

# GUIDELINES FOR MANUSCRIPTS

Before finalizing your text for submission, please examine a recent issue of *Survey Methodology* (Vol. 32, No. 2 and onward) as a guide and note particularly the points below. Articles must be submitted in machine-readable form, preferably in Word. A pdf or paper copy may be required for formulas and figures.

## 1. Layout

- 1.1 Documents should be typed entirely double spaced with margins of at least 1½ inches on all sides.
- 1.2 The documents should be divided into numbered sections with suitable verbal titles.
- 1.3 The name (fully spelled out) 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, ω; o, O, 0; l, 1).
- 3.6 Italics are used for emphasis.

## 4. Figures and Tables

- 4.1 All figures and tables should be numbered consecutively with arabic numerals, with titles that are as self explanatory as possible, at the bottom for figures and at the top for tables.

## 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, page 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.

## 6. Short Notes

- 6.1 Documents submitted for the short notes section must have a maximum of 3,000 words.