Catalogue no. 12-001-X

Survey Methodology

June 2009

SURVEY
METHODOLOGY

Catalogue No. 12-001-XPB

A JOURNAL
PUBLISHED BY
STATISTICS CANADA

JUNE 2005
•
VOLUME 31
•
NUMBER 1

Statistics Canada Statistique Canada

Canada

Statistics Canada Statistique Canada

Canada

How to obtain more information

Specific inquiries about this product and related statistics or services should be directed to: Business Survey Methods Division, Statistics Canada, Ottawa, Ontario, K1A 0T6 (telephone: 1-800-263-1136).

For information about this product or the wide range of services and data available from Statistics Canada, visit our website at www.statcan.gc.ca, e-mail us at infostats@statcan.gc.ca, or telephone us, Monday to Friday from 8:30 a.m. to 4:30 p.m., at the following numbers:

Statistics Canada's National Contact Centre

Toll-free telephone (Canada and United States):

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

Local or international calls:

Inquiries line	1-613-951-8116
Fax line	1-613-951-0581

Depository Services Program

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

To access and order this product

This product, Catalogue no. 12-001-X, is available free in electronic format. To obtain a single issue, visit our website at www.statcan.gc.ca and select "Publications."

This product, Catalogue no. 12-001-X, 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:

	Single issue	**Annual subscription**
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
• 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 under "About us" > "Providing services to Canadians."

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.

∞

Statistics Canada

Business Survey Methods Division

Survey Methodology

June 2009

Published by authority of the Minister responsible for Statistics Canada

June 2009

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

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

Frequency: semi-annual

Ottawa

Cette publication est disponible en français sur demande (nº 12-001-X au catalogue).

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 cooperation and goodwill.

Survey Methodology

A Journal Published by Statistics Canada

Volume 35, Number 1, June 2009

Contents

In This Issue 1

Regular Papers

Victor M. Estevao and Carl-Erik Särndal
A new face on two-phase sampling with calibration estimators 3

Jianzhu Li and Richard Valliant
Survey weighted hat matrix and leverages 15

Jean-François Beaumont and Cynthia Bocci
A practical bootstrap method for testing hypotheses from survey data 25

Bo-Seung Choi, Jai Won Choi and Yousung Park
Bayesian methods for an incomplete two-way contingency table with application to the Ohio (Buckeye State) Polls 37

Malay Ghosh, Dalho Kim, Karabi Sinha, Tapabrata Maiti, Myron Katzoff and Van L. Parsons
Hierarchical and empirical Bayes small domain estimation of the proportion of persons without health insurance for minority subpopulations 53

Tucker McElroy and Scott Holan
A nonparametric test for residual seasonality 67

Siegfried Gabler and Partha Lahiri
On the definition and interpretation of interviewer variability for a complex sampling design 85

Barry Schouten, Fannie Cobben and Jelke Bethlehem
Indicators for the representativeness of survey response 101

Guillaume Chauvet
Stratified balanced sampling 115

In this issue

In the first paper of this issue of *Survey Methodology*, Estevao and Särndal consider the problem of calibration estimation in the context of two-phase sampling. The contributions of the paper include the choice of initial weights in the calibration procedure as well as the important problem of variance estimation. New variance estimators are proposed and results from a simulation study show that the proposed variance estimators are more efficient than the traditional ones.

Next, Li and Valliant investigate the problem of the detection of influential units in linear regression analysis of survey data. They first give an expression for the hat matrix and its associated leverages (diagonal of the hat matrix) when a weighted least squares technique is used to estimate model parameters. They then propose a decomposition of the leverages and highlight that the leverage for a given unit can be large when either its survey weight is large or its vector of explanatory variables is far from the center. They illustrate the effect of influential units on both ordinary and weighted least squares using a numerical example.

Beaumont and Bocci propose a bootstrap methodology for testing hypotheses about a vector of unknown model parameters when the sample has been drawn from a finite population. The technique uses model-based test statistics that incorporate the survey weights and can usually be obtained easily using standard software packages. Using a simulation study the authors show that the proposed method performs similarly to the Rao-Scott procedure, and better than the Wald and Bonferroni procedures when testing hypotheses about a vector of linear regression model parameters.

The paper by Park, Choi and Choi present an interesting approach to nonresponse. Studies have shown that the voting behaviour of the undecided voters can have a significant impact on the final result of an election and that by considering these undecided voters, the accuracy of election forecasting can be improved. The authors present two Bayesian models whose priors depend on information from both respondents and undecided. They analyze an incomplete two-way contingency table using four sets of data from the 1998 Ohio state polls to illustrate how to use and interpret estimation results for the elections.

Ghosh, Kim, Sinha, Maiti, Katzoff and Parsons develop hierarchical and empirical Bayes methods for estimation of proportions in small domains using unit-level models. They propose a hierarchical Bayes analogue of the generalized linear mixed model to obtain posterior means and posterior standard errors of the population small domain proportions. Using an approach based on the theory of optimal estimating functions, they also obtain emprical Bayes estimators and corresponding asymptotic mean square error estimators. The methods are illustrated using data from the National Health Interview Survey (NHIS) to obtain small domain estimates of the proportions of Asians without health insurance.

In the McElroy and Holan paper, the problem of testing for residual seasonality in seasonally adjusted data is investigated. The authors propose a statistical significance test for peaks in the spectral density of the time series under consideration that is indicative of seasonality. The theory of the proposed method developed and is illustrated and compared with existing methods through both simulation and empirical studies.

Gabler and Lahiri provide a model-assisted justification of the traditional interviewer variance formula for equal probability sampling with no spatial clustering. They then obtain, in the context of a complex sampling design, a definition of interviewer variability that appropriately accounts for unequal probabilities of selection and spatial clustering. They also propose a decomposition of total effects into effects due to weighting, spatial clustering and interviewers. Their results can help to more effectively understand and control sources of variability.

In their paper, Schouten, Cobben and Bethlehem investigate the problem of assessing the similarity between the response to a survey and the sample or population under investigation. They propose a representativeness indicator to replace response rates as a quality indicator for the impact of nonresponse bias. This indicator, called the R-indicator, is shown to be somewhat related to Cramer's V measure for the association between response and auxiliary variables. In fact, the R-indicator is better viewed as a lack of association measure since a weaker association implies that there is no evidence that nonresponse has affected the composition of the observed data. The theoretical properties of the proposed indicator are developed and it is illustrated through empirical studies.

Finally, in his article, Chauvet addresses the issue of balanced sampling when sizes in each stratum are too small for exact balancing. The author proposes an algorithm adapted to the Cube method, which guarantees balancing at the population level. A simulation study confirmed that the proposed method performed well.

Harold Mantel, Deputy Editor

Survey Methodology, June 2009
Vol. 35, No. 1, pp. 3-14
Statistics Canada, Catalogue No. 12-001-X

A new face on two-phase sampling with calibration estimators

Victor M. Estevao and Carl-Erik Särndal [1]

Abstract

This paper provides a framework for estimation by calibration in two-phase sampling designs. This work grew out of the continuing development of generalized estimation software at Statistics Canada. An important objective in this development is to provide a wide range of options for effective use of auxiliary information in different sampling designs. This objective is reflected in the general methodology for two-phase designs presented in this paper.

We consider the traditional two-phase sampling design. A phase-one sample is drawn from the finite population and then a phase-two sample is drawn as a sub-sample of the first. The study variable, whose unknown population total is to be estimated, is observed only for the units in the phase-two sample. Arbitrary sampling designs are allowed in each phase of sampling. Different types of auxiliary information are identified for the computation of the calibration weights at each phase. The auxiliary variables and the study variables can be continuous or categorical.

The paper contributes to four important areas in the general context of calibration for two-phase designs:

(1) Three broad types of auxiliary information for two-phase designs are identified and used in the estimation. The information is incorporated into the weights in two steps: a phase-one calibration and a phase-two calibration. We discuss the composition of the appropriate auxiliary vectors for each step, and use a linearization method to arrive at the residuals that determine the asymptotic variance of the calibration estimator.

(2) We examine the effect of alternative choices of starting weights for the calibration. The two "natural" choices for the starting weights generally produce slightly different estimators. However, under certain conditions, these two estimators have the same asymptotic variance.

(3) We re-examine variance estimation for the two-phase calibration estimator. A new procedure is proposed that can improve significantly on the usual technique of conditioning on the phase-one sample. A simulation in section 10 serves to validate the advantage of this new method.

(4) We compare the calibration approach with the traditional model-assisted regression technique which uses a linear regression fit at two levels. We show that the model-assisted estimator has properties similar to a two-phase calibration estimator.

Key Words: Auxiliary information; Two-phase regression estimator; Starting weights; Separate residual variance estimator; Combined residual variance estimator.

1. Introduction

The term *double sampling* refers to sampling designs whose common feature is a selection of two probability samples, denoted s_1 and s, both of them subsets of the finite population of interest, given by $U = \{1, ..., k, ..., N\}$. The sample s_1 is realized and observed prior to s. A typical study variable is denoted by y; its value y_k is obtained only for the units $k \in s$. The objective is to estimate the population y-total $Y = \sum_U y_k$ (if A is a set of units, $A \subseteq U$, then we write $\sum_A$ as a short form for $\sum_{k \in A}$ when there is no ambiguity).

Hidiroglou (2001) discusses two types of double sampling, *nested* and *non-nested*. This paper focuses on the nested type, usually referred to as two-phase sampling: The phase-two sample s is a sub-sample from the phase-one sample s_1 drawn from U, so $s \subseteq s_1 \subseteq U$.

Estimation for two-phase sampling has been examined in several earlier papers in a context where two kinds of auxiliary information are recognized and addressed by their levels: At the population level, the total $\sum_U \mathbf{x}_{1k}$ is known, where $\mathbf{x}_{1k}$ is a vector known for every $k \in s_1$; therefore, it is also known for every $k \in s$. At the level of the first sample, the vector value $\mathbf{x}_{2k}$ is observed for every $k \in s_1$, and is thereby known for every $k \in s$; the total $\sum_U \mathbf{x}_{2k}$ is unknown but can be estimated without design bias at the s_1-level. Two arguments are found in the literature for incorporating these two types of auxiliary information in estimating $Y = \sum_U y_k$: the regression fit argument and the calibration argument. Under certain conditions they can lead to identical estimators, but this is not so in general.

The regression fit argument prevails in Särndal and Swensson (1987), Särndal, Swensson and Wretman (1992), Sitter (1997), Hidiroglou and Särndal (1998), Axelson (1998) and Hidiroglou, Rao and Haziza (2006). The calibration approach in Deville and Särndal (1992) was applied to two-phase sampling by Dupont (1995). She compares the resulting calibration estimators with those obtained from the regression approach. For the same auxiliary information, the two approaches may not give

[1] Victor M. Estevao, Business Survey Methods Division, Statistics Canada, Ottawa, Ontario, Canada, K1A OT6. E-mail: victor.estevao@statcan.gc.ca; Carl-Erik Särndal, professor. E-mail: carl.sarndal@rogers.com.

identical estimators, although in practice the difference is likely to be of little consequence. Resampling for two-phase variance estimation is considered in Kott and Stukel (1997). Estevao and Särndal (2002) focus on the calibration argument and distinguish ten different ways to use all or part of the information available at the two levels. The present paper also focuses on the calibration approach. It extends earlier work by recognizing three (rather than two) types of auxiliary information, each having different characteristics.

In the regression approach, it is natural to fit two linear least squares regressions. One set of regression-predicted y-values are produced for $k \in s_1$ using both $\mathbf{x}_{1k}$ and $\mathbf{x}_{2k}$ as predictors; another set is produced for $k \in s_1$ using only the vector $\mathbf{x}_{1k}$ as predictor. Both sets of predicted y-values, as well as the known total $\sum_U \mathbf{x}_{1k}$, are used to build the regression-type estimator of Y, in the manner described in section 9.

The calibration approach is motivated by two factors: To create a set of weights that are consistent with known or estimated totals for the auxiliary variables and to reduce the variance of the estimates made for the study variable(s). We want the weights w_k in $\hat{Y}_{2P} = \sum_s w_k y_k$ to achieve consistency with the total $\sum_U \mathbf{x}_{1k}$ known at the level of the population and/or with an (approximately) unbiased estimate, made at the level of the phase-one sample, of the unknown $\sum_U \mathbf{x}_{2k}$. Since y is observed only at the ultimate level (the phase-two sample), consistency "at higher levels" on important auxiliary variables will often significantly reduce the variance of $\hat{Y}_{2P} = \sum_s w_k y_k$. We can distinguish two steps in the process leading to the weights w_k, a phase-one calibration and a phase-two calibration.

The two-phase sampling design is as follows: From the finite population of units $U = \{1, 2, \ldots k, \ldots N\}$ we select a phase-one sample s_1. The known positive inclusion probability of unit k is $\pi_{1k} = \Pr(k \in s_1)$, and the phase-one design weight is $a_{1k} = 1/\pi_{1k}$. Certain variables may be observed for the units $k \in s_1$. Then, conditionally on s_1, we select a phase-two sample s from s_1. The known and positive conditional inclusion probability of k is $\pi_{2k} = \Pr(k \in s \mid s_1)$ for $k \in s_1$, and the conditional phase-two design weight is $a_{2k} = 1/\pi_{2k}$. (to keep the notation simple, we use π_{2k} and a_{2k} rather than the more suggestive $\pi_{2k|s_1}$ and $a_{2k|s_1}$; it should be kept in mind that both π_{2k} and a_{2k} are conditional on the phase-one sample s_1). The combined or double-expansion design weight is $a_k = a_{1k}a_{2k}$ for $k \in s$. The analysis of the estimators in this article is design based. The term "(approximately) unbiased" means "(approximately) design unbiased." We assume mild conditions on the population and the two sampling designs, permitting us to discard lower order terms in the analysis of our estimators when the expected sizes of the phase-one and phase-two samples are sufficiently large.

The double-expansion estimator $\sum_s a_k y_k$ is unbiased for $Y = \sum_U y_k$. We can produce more efficient estimators by taking into account the available auxiliary information. Three types or sets of auxiliary variables (called x-variables) can be distinguished for two-phase sampling designs. These are denoted by $\mathcal{X}^{\oplus}$, $\mathcal{X}^{\dagger}$ and $\mathcal{X}^{\circ}$. Their information characteristics are specified in the following table.

Table 1.1
Sets of auxiliary variables for calibration in two-phase sampling

Set of auxiliary variables	**Auxiliary variable total over U**	**Unit variable values for $k \in s_1$**	**Unit variable values for $k \in s$**
$\mathcal{X}^{\oplus}$	known	known	known
$\mathcal{X}^{\dagger}$	known	unknown	known
$\mathcal{X}^{\circ}$	unknown	known	known

Each set may contain any number of x-variables. The three sets are mutually exclusive. The properties in the last three columns apply to every x-variable in the corresponding set. All x-variables used for calibration belong to one of these three sets.

2. Phase-one calibration

For the phase-one calibration, we use a vector $\mathbf{x}_{1k}$ of auxiliary variables selected from the set $\mathcal{X}^{\oplus}$. While it is natural to let $\mathbf{x}_{1k}$ consist of all the variables in $\mathcal{X}^{\oplus}$, the general presentation here allows us to define $\mathbf{x}_{1k}$ to include some or even none of the variables in $\mathcal{X}^{\oplus}$. The phase-one calibration weights w_{1k} are derived by modifying the phase-one starting weights a_{1k} subject to the calibration constraint $\sum_{s_1} w_{1k}\mathbf{x}_{1k} = \sum_U \mathbf{x}_{1k}$. In our formulation, the calibration weights are given for $k \in s_1$ as

$$w_{1k} = a_{1k}\left\{1 + (\mathbf{X}_1 - \hat{\mathbf{X}}_1)'\left(\sum\nolimits_{s_1} a_{1k}\mathbf{z}_{1k}\mathbf{x}'_{1k}\right)^{-1}\mathbf{z}_{1k}\right\} \qquad (2.1)$$

where $\mathbf{X}_1 = \sum_U \mathbf{x}_{1k}$, $\hat{\mathbf{X}}_1 = \sum_{s_1} a_{1k}\mathbf{x}_{1k}$ and $\mathbf{z}_{1k}$ is an instrumental vector of the same dimension as $\mathbf{x}_{1k}$. It replaces $\mathbf{x}_{1k}/\sigma^2_{1k}$ in the form of the model-assisted estimator described by Särndal, Swensson, Wretman (1992), and permits a more general specification of the calibration weights. The use of an instrumental vector is discussed in Estevao and Särndal (2000) and Deville (2002). Here and in the following, we always assume the invertibility of matrices such as the one over s_1 in (2.1) and those (over s and U) appearing later.

3. Phase-two calibration

We use a vector $\mathbf{x}_k$ of auxiliary variables to produce a set of phase-two (or final) calibration weights w_k. They are used to calculate $\hat{Y}_{2P} = \sum_s w_k y_k$ as our estimator of $Y = \sum_U y_k$. The vector $\mathbf{x}_k = (\mathbf{x}'_{k(t)}, \mathbf{x}'_{k(w)}, \mathbf{x}'_{k(a)})'$ has three components, as described below. No auxiliary variable can appear in more than one of the three vector components. These three components have different roles in the setup of the phase-two calibration equation $\sum_s w_k \mathbf{x}_k = \mathbf{X}$ and in the determination of the phase-two calibration weights.

The variables in the vector $\mathbf{x}_{k(t)}$ are selected from among those in the set $\mathcal{X}^{\oplus} \cup \mathcal{X}^{\dagger}$. This means that the total $\sum_U \mathbf{x}_{k(t)}$ is known and can be included in $\mathbf{X}$. Variables in $\mathbf{x}_{1k}$ are allowed to reoccur in $\mathbf{x}_{k(t)}$, and this is usually preferable in order to reduce the variance of the estimator. We can specify $\mathbf{x}_{k(t)} = \mathbf{x}_{1k}$, but our framework permits $\mathbf{x}_{k(t)}$ to include variables from $\mathcal{X}^{\dagger}$. This allows us to use variables with known population totals in situations where the variables are too expensive to collect for a large phase-one sample s_1 but are observable for the smaller phase-two sample s. These variables are excluded from the phase-one calibration because they are unavailable for $k \in s_1$.

The variables in $\mathbf{x}_{k(w)}$ and $\mathbf{x}_{k(a)}$ are selected from among those in the set $\mathcal{X}^{\oplus} \cup \mathcal{X}^{\dagger} \cup \mathcal{X}^{\circ}$ provided they are not already included in $\mathbf{x}_{k(t)}$. The variables in $\mathbf{x}_{k(w)}$ are those for which we want to satisfy the phase-two calibration equation $\sum_s w_k \mathbf{x}_{k(w)} = \sum_{s_1} w_{1k} \mathbf{x}_{k(w)}$, where the right-hand side is approximately unbiased for $\sum_U \mathbf{x}_{k(w)}$. The variables in $\mathbf{x}_{k(a)}$ are those for which we want to satisfy the phase-two calibration equation $\sum_s w_k \mathbf{x}_{k(a)} = \sum_{s_1} a_{1k} \mathbf{x}_{k(a)}$. Here, the right-hand side is unbiased for $\sum_U \mathbf{x}_{k(a)}$. The inclusion of both $\mathbf{x}_{k(w)}$ and $\mathbf{x}_{k(a)}$ in the definition of $\mathbf{x}_k$ allows us to calibrate on one or both of these vectors and provides a general framework for producing different estimators from the phase-two calibration.

The phase-two calibration equation is $\sum_s w_k \mathbf{x}_k = \mathbf{X}$, where $\mathbf{X}$ is the stacked auxiliary vector

$$\mathbf{X} = \begin{pmatrix} \sum_U \mathbf{x}_{k(t)} \\ \sum_{s_1} w_{1k} \mathbf{x}_{k(w)} \\ \sum_{s_1} a_{1k} \mathbf{x}_{k(a)} \end{pmatrix}. \tag{3.1}$$

A specific variable can only occur once in $\mathbf{x}_k$. Otherwise, the calibration equation may be inconsistent and admit no solution.

The starting weights for the phase-two calibration are denoted by a_k^* for $k \in s$. There is more than one reasonable choice for the a_k^*. We consider two alternatives, both of which seem natural: (1) $a_k^* = a_k = a_{1k}\, a_{2k}$, and (2) $a_k^* = w_{1k}\, a_{2k}$, where w_{1k} is the phase-one calibration weight given by (2.1).

Given the starting weights a_k^*, we determine final weights w_k subject to the calibration equation $\sum_s w_k \mathbf{x}_k = \mathbf{X}$. These final weights are given for $k \in s$ by

$$w_k = a_k^* \left\{ 1 + (\mathbf{X} - \hat{\mathbf{X}})' \left(\sum\nolimits_s a_k^* \mathbf{z}_k \mathbf{x}'_k \right)^{-1} \mathbf{z}_k \right\} \tag{3.2}$$

where $\hat{\mathbf{X}} = \sum_s a_k^* \mathbf{x}_k$ is an unbiased or approximately unbiased estimator of $\mathbf{X}$, depending on the composition of $\mathbf{x}_k$. The instrumental variable $\mathbf{z}_k$ has the same dimension as $\mathbf{x}_k$. The vectors $\mathbf{z}_{1k}$ and $\mathbf{z}_k$ are assumed to be fixed functions of $\mathbf{x}_{1k}$ and $\mathbf{x}_k$. How to choose $\mathbf{z}_{1k}$ and $\mathbf{z}_k$ is a topic we leave for others to address.

4. Comparison of two options for the starting weights

The objective in this section is to analyze how the final weights w_k in $\hat{Y}_{2P} = \sum_s w_k y_k$ depend on the specification of the starting weights a_k^* in (3.2). We consider two distinct cases based on whether or not the auxiliary variables $\mathbf{x}_k$ are used for the phase-two calibration. When we carry out the phase-two calibration, the two different choices for starting weights generally lead to different estimators. We show that these estimators are asymptotically equivalent under certain conditions, commonly found in practice. When we have no phase-two calibration, the two choices for starting weights lead to two other estimators that are usually less efficient than those obtained by performing the phase-two calibration.

4.1 Estimators with phase-two calibration ($\mathbf{x}_k \neq \phi$)

As noted previously, there are two alternatives for the starting weights a_k^* in (3.2): (1) $a_k^* = a_k = a_{1k}\, a_{2k}$, and (2) $a_k^* = w_{1k}\, a_{2k}$, where w_{1k} is the phase-one calibration weight given by (2.1). We now provide a detailed analysis of the form of the estimator under these two choices. In this subsection, we look at the more interesting case where we perform the phase-two calibration ($\mathbf{x}_k \neq \phi$). In the next subsection, we consider what happens when we do not carry out the phase-two calibration ($\mathbf{x}_k = \phi$).

Our procedure is as follows. First, we derive the linearized (asymptotic) form of $\hat{Y}_{2P}$ based on the general starting weights a_k^*. Then we substitute the two choices for a_k^* in this expression. We determine $\hat{Y}_{2P}$ based on the starting weights $a_k^* = a_k = a_{1k}\, a_{2k}$. We denote this estimator by $\hat{Y}_{2P\,a}$ and derive its linearized form, $\hat{Y}_{2P\,a\,\text{lin}}$. Similarly, we obtain $\hat{Y}_{2P}$ based on the starting weights $a_k^* = w_{1k} a_{2k}$. We refer to this estimator as $\hat{Y}_{2P\,w}$ and derive its linearized form, $\hat{Y}_{2P\,w\,\text{lin}}$. These two forms are slightly different but we prove in Result 4.2 that $\hat{Y}_{2P\,a\,\text{lin}} = \hat{Y}_{2P\,w\,\text{lin}}$ under certain conditions.

We start by inserting the weights w_k into $\hat{Y}_{2P} = \sum_s w_k\, y_k$ and writing the estimator as

$$\begin{aligned}\hat{Y}_{2P} \;=\; & \sum\nolimits_U \mathbf{x}'_{k(t)}\,\mathbf{B}_{(y;\,\mathbf{x})(t)} + \sum\nolimits_{s_1} w_{1k}\,\mathbf{x}'_{k(w)}\mathbf{B}_{(y;\,\mathbf{x})(w)} \\ & + \sum\nolimits_{s_1} a_{1k}\,\mathbf{x}'_{k(a)}\mathbf{B}_{(y;\,\mathbf{x})(a)} + \sum\nolimits_s a_k^*\, e_{(y;\,\mathbf{x})k} \\ & + (\mathbf{X} - \hat{\mathbf{X}})'(\hat{\mathbf{B}}^*_{(y;\,\mathbf{x})} - \mathbf{B}_{(y;\,\mathbf{x})}) \end{aligned} \qquad (4.1)$$

where $\hat{\mathbf{B}}^*_{(y;\,\mathbf{x})} = (\sum_s a_k^*\,\mathbf{z}_k\,\mathbf{x}'_k)^{-1}\sum_s a_k^*\,\mathbf{z}_k\,y_k$, $\mathbf{B}_{(y;\,\mathbf{x})} = (\sum_U \mathbf{z}_k\mathbf{x}'_k)^{-1}\sum_U \mathbf{z}_k\,y_k$ and $\mathbf{B}_{(y;\mathbf{x})} = (\mathbf{B}'_{(y;\mathbf{x})(t)}, \mathbf{B}'_{(y;\mathbf{x})(w)}, \mathbf{B}'_{(y;\mathbf{x})(a)})'$ is the partitioning corresponding to $\mathbf{x}_k = (\mathbf{x}'_{k(t)}, \mathbf{x}'_{k(w)}, \mathbf{x}'_{k(a)})'$. Our subscript notation of the form $(\mathbf{v}_1;\,\mathbf{v}_2)$ identifies the variables in the regression. The term $\mathbf{v}_2$ refers to the independent variables and $\mathbf{v}_1$ identifies the dependent variable or variables. For simplicity, the instrumental vectors $\mathbf{z}_{1k}$ and $\mathbf{z}_k$ are not included in the notation.

The term $e_{(y;\,\mathbf{x})k} = y_k - \mathbf{x}'_k\mathbf{B}_{(y;\,\mathbf{x})}$ is defined for $k \in U$. Note that although $e_{(y;\,\mathbf{x})k}$ looks like a regression residual, it does not arise as the result of fitting a proper regression model. We then develop the term $\sum_{s_1} w_{1k}\,\mathbf{x}'_{k(w)}\mathbf{B}_{(y;\,\mathbf{x})(w)}$ in (4.1) by inserting expression (2.1) for w_{1k} and making use of the phase-one calibration equation $\sum_{s_1} w_{1k}\mathbf{x}_{1k} = \sum_U \mathbf{x}_{1k}$. We obtain

$$\begin{aligned}\sum\nolimits_{s_1} w_{1k}\mathbf{x}'_{k(w)}\mathbf{B}_{(y;\,\mathbf{x})(w)} = \\ & \mathbf{X}'_1\,\mathbf{B}_{(\mathbf{x}\mathbf{B}_{(w)};\,\mathbf{x}_1)} + \sum\nolimits_{s_1} a_{1k}e_{(\mathbf{x}\mathbf{B}_{(w)};\,\mathbf{x}_1)k} \\ & + (\mathbf{X}_1 - \hat{\mathbf{X}}_1)'\,(\hat{\mathbf{B}}_{(\mathbf{x}\mathbf{B}_{(w)};\,\mathbf{x}_1)} - \mathbf{B}_{(\mathbf{x}\mathbf{B}_{(w)};\,\mathbf{x}_1)}) \end{aligned} \qquad (4.2)$$

where $\hat{\mathbf{B}}_{(\mathbf{x}\mathbf{B}_{(w)};\mathbf{x}_1)} = (\sum_{s_1} a_{1k}\mathbf{z}_{1k}\mathbf{x}'_{1k})^{-1}\sum_{s_1} a_{1k}\mathbf{z}_{1k}\mathbf{x}'_{k(w)}\mathbf{B}_{(y;\mathbf{x})(w)}$ converges in probability to (and is approximately unbiased for) $\mathbf{B}_{(\mathbf{x}\mathbf{B}_{(w)};\mathbf{x}_1)} = (\sum_U \mathbf{z}_{1k}\,\mathbf{x}'_{1k})^{-1}\sum_U \mathbf{z}_{1k}\,\mathbf{x}'_{k(w)}\mathbf{B}_{(y;\,\mathbf{x})(w)}$, and $e_{(\mathbf{x}\mathbf{B}_{(w)};\mathbf{x}_1)k} = \mathbf{x}'_{k(w)}\mathbf{B}_{(y;\mathbf{x})(w)} - \mathbf{x}'_{1k}\mathbf{B}_{(\mathbf{x}\mathbf{B}_{(w)};\mathbf{x}_1)}$ is defined for $k \in U$.

We can interpret $e_{(\mathbf{x}\mathbf{B}_{(w)};\,\mathbf{x}_1)k}$ as a residual arising from a population fit based on a generalized regression of $\mathbf{x}'_{k(w)}\mathbf{B}_{(y;\,\mathbf{x})(w)}$ as the dependent variable and $\mathbf{x}_{1k}$ as the predictor vector. Replacing expression (4.2) into expression (4.1) for $\hat{Y}_{2P}$ leads to

$$\begin{aligned}\hat{Y}_{2P} \;=\; & \sum\nolimits_U (\mathbf{x}'_{k(t)}\mathbf{B}_{(y;\,\mathbf{x})(t)} + \mathbf{x}'_{1k}\,\mathbf{B}_{(\mathbf{x}\mathbf{B}_{(w)};\,\mathbf{x}_1)}) \\ & + \sum\nolimits_{s_1} a_{1k}(\mathbf{x}'_{k(a)}\mathbf{B}_{(y;\,\mathbf{x})(a)} + e_{(\mathbf{x}\mathbf{B}_{(w)};\,\mathbf{x}_1)k}) \\ & + \sum\nolimits_s a_k^* e_{(y;\,\mathbf{x})k} \\ & + (\mathbf{X}_1 - \hat{\mathbf{X}}_1)'\,(\hat{\mathbf{B}}_{(\mathbf{x}\mathbf{B}_{(w)};\,\mathbf{x}_1)} - \mathbf{B}_{(\mathbf{x}\mathbf{B}_{(w)};\,\mathbf{x}_1)}) \\ & + (\mathbf{X} - \hat{\mathbf{X}})'(\hat{\mathbf{B}}^*_{(y;\,\mathbf{x})} - \mathbf{B}_{(y;\,\mathbf{x})}). \end{aligned} \qquad (4.3)$$

The following result establishes the relationship between the estimators obtained for the two choices of starting weights.

Result 4.1: The linearized forms of $\hat{Y}_{2P\,a}$ and $\hat{Y}_{2P\,w}$ are related by the equation $\hat{Y}_{2P\,w\,\mathrm{lin}} = \hat{Y}_{2P\,a\,\mathrm{lin}} + (\mathbf{X}_1 - \hat{\mathbf{X}}_1)'(\mathbf{B}_{(y;\mathbf{x}_1)} - \mathbf{B}_{(\mathbf{x};\mathbf{x}_1)}\mathbf{B}_{(y;\mathbf{x})})$.

Proof

We consider expression (4.3) under the two possible choices for a_k^*. First, with $a_k^* = a_k = a_{1k}\,a_{2k}$ we obtain $\hat{Y}_{2P\,a}$ given by

$$\begin{aligned}\hat{Y}_{2P\,a} \;=\; & \sum\nolimits_U (\mathbf{x}'_{k(t)}\mathbf{B}_{(y;\mathbf{x})(t)} + \mathbf{x}'_{1k}\,\mathbf{B}_{(\mathbf{x}\mathbf{B}_{(w)};\mathbf{x}_1)}) \\ & + \sum\nolimits_{s_1} a_{1k}(\mathbf{x}'_{k(a)}\mathbf{B}_{(y;\mathbf{x})(a)} + e_{(\mathbf{x}\mathbf{B}_{(w)};\mathbf{x}_1)k}) \\ & + \sum\nolimits_s a_k e_{(y;\mathbf{x})k} \\ & + (\mathbf{X}_1 - \hat{\mathbf{X}}_1)'\,(\hat{\mathbf{B}}_{(\mathbf{x}\mathbf{B}_{(w)};\mathbf{x}_1)} - \mathbf{B}_{(\mathbf{x}\mathbf{B}_{(w)};\mathbf{x}_1)}) \\ & + (\mathbf{X} - \hat{\mathbf{X}})'(\hat{\mathbf{B}}_{(y;\mathbf{x})} - \mathbf{B}_{(y;\mathbf{x})}). \end{aligned} \qquad (4.4)$$

The term $\hat{\mathbf{B}}_{(y;\mathbf{x})} = (\sum_s a_k\mathbf{z}_k\mathbf{x}'_k)^{-1}\sum_s a_k\mathbf{z}_k y_k$ converges in probability to (and is approximately unbiased for) $\mathbf{B}_{(y;\mathbf{x})} = (\sum_U \mathbf{z}_k\mathbf{x}'_k)^{-1}\sum_U \mathbf{z}_k y_k$. The first term is constant and does not contribute to the variance of $\hat{Y}_{2P\,a}$. The next two terms are random quantities, defined as sums over s_1 and s respectively. The last two terms are products of differences with zero or almost zero expectation. As for the product $(\mathbf{X}_1 - \hat{\mathbf{X}}_1)'\,(\hat{\mathbf{B}}_{(\mathbf{x}\mathbf{B}_{(w)};\mathbf{x}_1)} - \mathbf{B}_{(\mathbf{x}\mathbf{B}_{(w)};\mathbf{x}_1)})$, both differences are functions of the phase-one sample s_1. We know that $\hat{\mathbf{X}}_1$ is unbiased for $\mathbf{X}_1$ and $\hat{\mathbf{B}}_{(\mathbf{x}\mathbf{B}_{(w)};\mathbf{x}_1)}$ is approximately unbiased for $\mathbf{B}_{(\mathbf{x}\mathbf{B}_{(w)};\mathbf{x}_1)}$. Under fairly general conditions, $N^{-1}(\mathbf{X}_1 - \hat{\mathbf{X}}_1)'(\hat{\mathbf{B}}_{(\mathbf{x}\mathbf{B}_{(w)};\,\mathbf{x}_1)} - \mathbf{B}_{(\mathbf{x}\mathbf{B}_{(w)};\,\mathbf{x}_1)}) = O_P(n_1^{-1})$, where n_1 is the expected size of s_1, assumed sufficiently large. By a similar reasoning, $N^{-1}(\mathbf{X} - \hat{\mathbf{X}})'(\hat{\mathbf{B}}_{(y;\,\mathbf{x})} - \mathbf{B}_{(y;\,\mathbf{x})}) = O_P(n^{-1})$, where n is the expected size of s, also assumed sufficiently large. Consequently, we can drop the last two terms of (4.4), because they are of lower order than the preceding terms: $N^{-1}\sum_{s_1} a_{1k}(\mathbf{x}'_{k(a)}\,\mathbf{B}_{(y;\mathbf{x})\,(a)} + e_{(\mathbf{x}\mathbf{B}_{(w)};\mathbf{x}_1)k})$ is $O_P(n_1^{-1/2})$ and $N^{-1}\sum_s a_k e_{(y;\mathbf{x})k}$ is $O_P(n^{-1/2})$. The first three terms define the linearized form of $\hat{Y}_{2P\,a}$,

$$\begin{aligned}\hat{Y}_{2P\,a\,\mathrm{lin}} \;=\; & \sum\nolimits_U (\mathbf{x}'_{k(t)}\mathbf{B}_{(y;\,\mathbf{x})(t)} + \mathbf{x}'_{1k}\,\mathbf{B}_{(\mathbf{x}\mathbf{B}_{(w)};\,\mathbf{x}_1)}) \\ & + \sum\nolimits_{s_1} a_{1k}(\mathbf{x}'_{k(a)}\mathbf{B}_{(y;\,\mathbf{x})(a)} + e_{(\mathbf{x}\mathbf{B}_{(w)};\,\mathbf{x}_1)k}) \\ & + \sum\nolimits_s a_k e_{(y;\,\mathbf{x})k}. \end{aligned} \qquad (4.5)$$

Now let us consider expression (4.3) under the second choice, $a_k^* = w_{1k}\,a_{2k}$. This leads to $\hat{Y}_{2P\,w}$ given by

$$\begin{aligned}\hat{Y}_{2P\,w} = {} & \textstyle\sum_U (\mathbf{x}'_{k(t)}\mathbf{B}_{(y;\,\mathbf{x})(t)} + \mathbf{x}'_{1k}\,\mathbf{B}_{(\mathbf{xB}_{(w)};\,\mathbf{x}_1)}) \\ & + \textstyle\sum_{s_1} a_{1k}(\mathbf{x}'_{k(a)}\mathbf{B}_{(y;\,\mathbf{x})(a)} + e_{(\mathbf{xB}_{(w)};\,\mathbf{x}_1)k}) \\ & + \textstyle\sum_s a_k e_{(y;\,\mathbf{x})k} \\ & + (\mathbf{X}_1 - \hat{\mathbf{X}}_1)'\left(\textstyle\sum_{s_1} a_{1k}\mathbf{z}_{1k}\mathbf{x}'_{1k}\right)^{-1}\textstyle\sum_s a_k \mathbf{z}_{1k} e_{(y;\,\mathbf{x})k} \\ & + (\mathbf{X} - \hat{\mathbf{X}})'(\hat{\mathbf{B}}^w_{(y;\,\mathbf{x})} - \mathbf{B}_{(y;\,\mathbf{x})}) \qquad (4.6)\end{aligned}$$

where $\hat{\mathbf{B}}^w_{(y;\,\mathbf{x})} = (\sum_s w_{1k}a_{2k}\mathbf{z}_k\mathbf{x}'_k)^{-1}\sum_s w_{1k}a_{2k}\mathbf{z}_k y_k$ and $\hat{\mathbf{X}} = \sum_s w_{1k}a_{2k}\,\mathbf{x}_k$. The first three terms of $\hat{Y}_{2P\,w}$ are the same as those found in expression (4.4) for $\hat{Y}_{2P\,a}$. The fourth and fifth terms differ from their counterparts in (4.4). Although $\hat{\mathbf{B}}^w_{(y;\,\mathbf{x})}$ and $\hat{\mathbf{X}}$ are functions of the phase-one calibration weights w_{1k}, we do not need to replace them in $\hat{\mathbf{B}}^w_{(y;\,\mathbf{x})}$ and $\hat{\mathbf{X}}$ in the fifth term; this would simply split the lower order term $(\mathbf{X} - \hat{\mathbf{X}})'(\hat{\mathbf{B}}^w_{(y;\,\mathbf{x})} - \mathbf{B}_{(y;\,\mathbf{x})})$ into other lower order terms. Therefore, we can drop the fifth term of (4.6) when the sample sizes are sufficiently large. The fourth term can be written as follows.

$$\begin{aligned}(\mathbf{X}_1 - \hat{\mathbf{X}}_1)'\left(\textstyle\sum_{s_1} a_{1k}\mathbf{z}_{1k}\mathbf{x}'_{1k}\right)^{-1}\textstyle\sum_s a_k\mathbf{z}_{1k}e_{(y;\,\mathbf{x})k} \\ = (\mathbf{X}_1 - \hat{\mathbf{X}}_1)'(\mathbf{B}_{(y;\,\mathbf{x}_1)} - \mathbf{B}_{(\mathbf{x};\,\mathbf{x}_1)}\mathbf{B}_{(y;\,\mathbf{x})}) \\ + (\mathbf{X}_1 - \hat{\mathbf{X}}_1)'(\hat{\mathbf{B}}_{(y;\,\mathbf{x}_1)s_1} - \mathbf{B}_{(y;\,\mathbf{x}_1)}) \\ - (\mathbf{X}_1 - \hat{\mathbf{X}}_1)'(\hat{\mathbf{B}}_{(\mathbf{x};\,\mathbf{x}_1)} - \mathbf{B}_{(\mathbf{x};\,\mathbf{x}_1)})\mathbf{B}_{(y;\,\mathbf{x})} \\ + (\mathbf{X}_1 - \hat{\mathbf{X}}_1)'\left(\textstyle\sum_{s_1} a_{1k}\mathbf{z}_{1k}\mathbf{x}'_{1k}\right)^{-1} \\ \left(\textstyle\sum_s a_k\mathbf{z}_{1k}e_{(y;\,\mathbf{x})k} - \textstyle\sum_{s_1} a_{1k}\mathbf{z}_{1k}e_{(y;\,\mathbf{x})k}\right). \quad (4.7)\end{aligned}$$

The quantities in this expression are defined as follows: $\hat{\mathbf{B}}_{(\mathbf{x};\,\mathbf{x}_1)} = (\sum_{s_1} a_{1k}\,\mathbf{z}_{1k}\,\mathbf{x}'_{1k})^{-1}\sum_{s_1} a_{1k}\,\mathbf{z}_{1k}\,\mathbf{x}'_k$ and $\hat{\mathbf{B}}_{(y;\,\mathbf{x}_1)s_1} = (\sum_{s_1} a_{1k}\mathbf{z}_{1k}\mathbf{x}'_{1k})^{-1}\sum_{s_1} a_{1k}\mathbf{z}_{1k}y_k$. The statistic $\hat{\mathbf{B}}_{(y;\,\mathbf{x}_1)s_1}$ can not be computed from the phase-one sample because the values y_k are only known for $k \in s$. It is implicitly defined for the purpose of determining the linearized form. We can define such a construct in the same manner as $\hat{\mathbf{B}}_{(\mathbf{xB}_{(w)};\,\mathbf{x}_1)}$ is a function of the unknown quantity $\mathbf{B}_{(y;\,\mathbf{x})(w)}$. Now $\hat{\mathbf{B}}_{(y;\,\mathbf{x}_1)s_1}$ is approximately unbiased for its corresponding population quantity $\mathbf{B}_{(y;\,\mathbf{x}_1)} = (\sum_U \mathbf{z}_{1k}\,\mathbf{x}'_{1k})^{-1}\sum_U \mathbf{z}_{1k}\,y_k$. Similarly, $\hat{\mathbf{B}}_{(\mathbf{x};\,\mathbf{x}_1)}$ is approximately unbiased for $\mathbf{B}_{(\mathbf{x};\mathbf{x}_1)} = (\sum_U \mathbf{z}_{1k}\,\mathbf{x}'_{1k})^{-1}\sum_U \mathbf{z}_{1k}\mathbf{x}'_k$. As before, we can argue that the last three terms of (4.7) are of lower order than the first term $(\mathbf{X}_1 - \hat{\mathbf{X}}_1)'(\mathbf{B}_{(y;\,\mathbf{x}_1)} - \mathbf{B}_{(\mathbf{x};\,\mathbf{x}_1)}\mathbf{B}_{(y;\,\mathbf{x})})$, which provides the linear approximation. The substitution of this term into (4.6) leads to the linearized form of $\hat{Y}_{2P\,w}$,

$$\begin{aligned}\hat{Y}_{2P\,w\,\mathrm{lin}} = {} & \textstyle\sum_U (\mathbf{x}'_{k(t)}\mathbf{B}_{(y;\,\mathbf{x})(t)} + \mathbf{x}'_{1k}\,\mathbf{B}_{(\mathbf{xB}_{(w)};\,\mathbf{x}_1)}) \\ & + \textstyle\sum_{s_1} a_{1k}(\mathbf{x}'_{k(a)}\mathbf{B}_{(y;\,\mathbf{x})(a)} + e_{(\mathbf{xB}_{(w)};\,\mathbf{x}_1)k}) \\ & + \textstyle\sum_s a_k e_{(y;\,\mathbf{x})k} \\ & + (\mathbf{X}_1 - \hat{\mathbf{X}}_1)'(\mathbf{B}_{(y;\,\mathbf{x}_1)} - \mathbf{B}_{(\mathbf{x};\,\mathbf{x}_1)}\mathbf{B}_{(y;\,\mathbf{x})}). \qquad (4.8)\end{aligned}$$

Comparing (4.5) with (4.8), we see that $\hat{Y}_{2P\,w\,\mathrm{lin}} = \hat{Y}_{2P\,a\,\mathrm{lin}} + (\mathbf{X}_1 - \hat{\mathbf{X}}_1)'(\mathbf{B}_{(y;\,\mathbf{x}_1)} - \mathbf{B}_{(\mathbf{x};\,\mathbf{x}_1)}\mathbf{B}_{(y;\,\mathbf{x})})$ as stated in the result. This completes the proof of result 4.1.

Result 4.1 shows that in general, the linearized forms of $\hat{Y}_{2P\,w}$ and $\hat{Y}_{2P\,a}$ are not the same. However, they are the same under certain conditions. Let us consider the case of nested calibration (not to be confused with nested sampling), meaning that $\mathbf{x}_k$ includes $\mathbf{x}_{1k}$. Then $\mathbf{x}_k$ is of the form $\mathbf{x}_k = (\mathbf{x}'_{1k}, \mathbf{x}'_{+k})'$ where the vector $\mathbf{x}_{+k}$ is composed of the remaining variables. We now state and prove the following result.

Result 4.2: If $\mathbf{x}_k = (\mathbf{x}'_{1k}, \mathbf{x}'_{+k})'$ and $\mathbf{z}_k = (\mathbf{z}'_{1k}, \mathbf{z}'_{+k})'$ then $\hat{Y}_{2P\,w\,\mathrm{lin}} = \hat{Y}_{2P\,a\,\mathrm{lin}}$ and $\hat{Y}_{2P\,a}$ and $\hat{Y}_{2P\,w}$ are asymptotically equivalent.

Proof

The proof follows from result 4.1 by showing $\mathbf{B}_{(y;\,\mathbf{x}_1)} - \mathbf{B}_{(\mathbf{x};\,\mathbf{x}_1)}\mathbf{B}_{(y;\,\mathbf{x})} = \mathbf{0}$ under the specified conditions. We have

$$\mathbf{B}_{(y;\mathbf{x}_1)} - \mathbf{B}_{(\mathbf{x};\mathbf{x}_1)}\mathbf{B}_{(y;\mathbf{x})} = \left(\textstyle\sum_U \mathbf{z}_{1k}\mathbf{x}'_{1k}\right)^{-1}\left(\textstyle\sum_U \mathbf{z}_{1k}h_k\right)$$

where $h_k = y_k - \mathbf{x}'_k(\sum_U \mathbf{z}_k\,\mathbf{x}'_k)^{-1}(\sum_U \mathbf{z}_k\,y_k)$. Since $\sum_U \mathbf{z}_{1k}h_k = \mathbf{0}$ and we assume $\mathbf{z}_k = (\mathbf{z}'_{1k}, \mathbf{z}'_{+k})'$, it follows $\sum_U \mathbf{z}_{1k}h_k = \mathbf{0}$ and $\mathbf{B}_{(y;\,\mathbf{x}_1)} - \mathbf{B}_{(\mathbf{x};\,\mathbf{x}_1)}\mathbf{B}_{(y;\,\mathbf{x})} = \mathbf{0}$. Therefore from result 4.1, $\hat{Y}_{2P\,w\,\mathrm{lin}} = \hat{Y}_{2P\,a\,\mathrm{lin}}$. Since their linear forms are the same, $\hat{Y}_{2P\,a}$ and $\hat{Y}_{2P\,w}$ are asymptotically equivalent estimators.

Interestingly, Result 4.2 only requires that we include $\mathbf{x}_{1k}$ somewhere within $\mathbf{x}_k$. Obviously, it makes sense to include $\mathbf{x}_{1k}$ within the component $\mathbf{x}_{k(t)}$ of $\mathbf{x}_k$ because the $\mathbf{x}_1$-totals are known. However, we obtain the same asymptotic result as long as all variables in $\mathbf{x}_{1k}$ are included somewhere in $\mathbf{x}_k = (\mathbf{x}'_{k(t)}, \mathbf{x}'_{k(w)}, \mathbf{x}'_{k(a)})'$. In practice, we often find $\mathbf{x}_{k(t)} = \mathbf{x}_{1k}$ with $\mathbf{z}_{1k} = \mathbf{x}_{1k}$ and $\mathbf{z}_k = \mathbf{x}_k = (\mathbf{x}'_{1k}, \mathbf{x}'_{+k})'$ where $\mathbf{x}_{+k}$ is the vector for the remaining variables $\mathbf{x}_{k(w)}$ and $\mathbf{x}_{k(a)}$. This satisfies the requirements for the asymptotic equivalence of $\hat{Y}_{2P\,a}$ and $\hat{Y}_{2P\,w}$.

To study the properties of $\hat{Y}_{2P\,a}$ and $\hat{Y}_{2P\,w}$ we work with their linearized forms given respectively by (4.5) and (4.8). With appropriate definitions for the residuals e_{0k}, e_{1k} and e_{2k}, we can represent $\hat{Y}_{2P\,a\,\mathrm{lin}}$ and $\hat{Y}_{2P\,w\,\mathrm{lin}}$ as the sum of three terms: a constant term $\sum_U e_{0k}$, a phase-one expansion term $\sum_{s_1} a_{1k}e_{1k}$, and a double-expansion term $\sum_s a_k e_{2k}$,

$$\hat{Y}_{2P\,\mathrm{lin}} = \sum\nolimits_U e_{0k} + \sum\nolimits_{s_1} a_{1k}\, e_{1k} + \sum\nolimits_s a_k\, e_{2k}. \quad (4.9)$$

This makes (4.9) a suitable starting point for studying the bias and the asymptotic variance of the two estimators $\hat{Y}_{2P\,a}$ and $\hat{Y}_{2P\,w}$.

For the linearized form $\hat{Y}_{2P\,a\,\mathrm{lin}}$ given by (4.5), the three residual quantities are defined as follows for $k \in U$:

$$e_{0k} = \mathbf{x}'_{k(t)}\, \mathbf{B}_{(y;\,\mathbf{x})(t)} + \mathbf{x}'_{1k}\, \mathbf{B}_{(\mathbf{x}\mathbf{B}_{(w)};\,\mathbf{x}_1)}$$

$$e_{1k} = \mathbf{x}'_{k(a)}\mathbf{B}_{(y;\,\mathbf{x})(a)} + \mathbf{x}'_{k(w)}\mathbf{B}_{(y;\,\mathbf{x})(w)}$$

$$- \mathbf{x}'_{1k}\mathbf{B}_{(\mathbf{x}\mathbf{B}_{(w)};\,\mathbf{x}_1)}$$

$$e_{2k} = y_k - \mathbf{x}'_{k(t)}\, \mathbf{B}_{(y;\,\mathbf{x})(t)} - \mathbf{x}'_{k(w)}\mathbf{B}_{(y;\,\mathbf{x})(w)}$$

$$- \mathbf{x}'_{k(a)}\mathbf{B}_{(y;\,\mathbf{x})(a)}. \quad (4.10)$$

Note that e_{2k} is simply $e_{(y;\,\mathbf{x})k}$. Similarly, for $\hat{Y}_{2P\,w\,\mathrm{lin}}$ given by (4.8), the residuals have the following definitions for $k \in U$:

$$e_{0k} = \mathbf{x}'_{k(t)}\, \mathbf{B}_{(y;\,\mathbf{x})(t)}$$

$$+ \mathbf{x}'_{1k}(\mathbf{B}_{(\mathbf{x}\mathbf{B}_{(w)};\,\mathbf{x}_1)} + \mathbf{B}_{(y;\,\mathbf{x}_1)} - \mathbf{B}_{(\mathbf{x};\,\mathbf{x}_1)}\mathbf{B}_{(y;\,\mathbf{x})})$$

$$e_{1k} = \mathbf{x}'_{k(a)}\mathbf{B}_{(y;\,\mathbf{x})(a)} + \mathbf{x}'_{k(w)}\mathbf{B}_{(y;\,\mathbf{x})(w)}$$

$$- \mathbf{x}'_{1k}(\mathbf{B}_{(\mathbf{x}\mathbf{B}_{(w)};\,\mathbf{x}_1)} + \mathbf{B}_{(y;\,\mathbf{x}_1)} - \mathbf{B}_{(\mathbf{x};\,\mathbf{x}_1)}\mathbf{B}_{(y;\,\mathbf{x})})$$

$$e_{2k} = y_k - \mathbf{x}'_{k(t)}\, \mathbf{B}_{(y;\,\mathbf{x})(t)}$$

$$- \mathbf{x}'_{k(w)}\mathbf{B}_{(y;\,\mathbf{x})(w)} - \mathbf{x}'_{k(a)}\mathbf{B}_{(y;\,\mathbf{x})(a)}. \quad (4.11)$$

Note that in both cases, $e_{0k} + e_{1k} + e_{2k} = y_k$ for every k, and hence $\sum_U (e_{0k} + e_{1k} + e_{2k}) = \sum_U y_k = Y$. This additivity allows us to prove in section 5 that $\hat{Y}_{2P\,a}$ and $\hat{Y}_{2P\,w}$ are approximately unbiased. To save space, we concentrate on the properties of $\hat{Y}_{2P\,a}$ in the remaining sections. However, the analysis is similar for $\hat{Y}_{2P\,w}$ and the method for variance estimation proposed in section 7 can also be used for this estimator.

4.2 Estimators without the phase-two calibration ($\mathbf{x}_k = \phi$)

If there is no phase-two calibration ($\mathbf{x}_k = \phi$), then $w_k = a_k^*$. Accordingly, the final weights are either $w_k = a_k = a_{1k}\, a_{2k}$ or $w_k = w_{1k}\, a_{2k}$. The first alternative gives the double-expansion estimator $\sum_s a_k\, y_k$. The second produces a different estimator that is usually more efficient. However, both of these are generally inefficient compared to the estimators obtained by carrying out the phase-two calibration. The linearized form of the two-phase estimator with $w_k = w_{1k}\, a_{2k}$ is obtained by writing it as follows.

$$\hat{Y}_{2P} = \mathbf{X}'_1\, \mathbf{B}_{(y;\,\mathbf{x}_1)} - \hat{\mathbf{X}}'_1\mathbf{B}_{(y;\,\mathbf{x}_1)} + \sum\nolimits_s a_k y_k$$

$$+ (\mathbf{X}_1 - \hat{\mathbf{X}}_1)'\, (\hat{\mathbf{B}}_{(y;\,\mathbf{x}_1)s_1} - \mathbf{B}_{(y;\,\mathbf{x}_1)})$$

$$+ (\mathbf{X}_1 - \hat{\mathbf{X}}_1)'\Big(\sum\nolimits_{s_1} a_{1k}\mathbf{z}_{1k}\mathbf{x}'_{1k}\Big)^{-1}$$

$$\Big(\sum\nolimits_s a_k\mathbf{z}_{1k}y_k - \sum\nolimits_{s_1} a_{1k}\mathbf{z}_{1k}y_k\Big). \quad (4.12)$$

The terms $\hat{\mathbf{B}}_{(y;\,\mathbf{x}_1)s_1}$ and $\mathbf{B}_{(y;\,\mathbf{x}_1)}$ were defined in the previous section. When the samples are sufficiently large, $(\mathbf{X}_1 - \hat{\mathbf{X}}_1)'(\sum_{s_1} a_{1k}\mathbf{z}_{1k}\mathbf{x}'_{1k})^{-1}(\sum_s a_k\mathbf{z}_{1k}y_k - \sum_{s_1} a_{1k}\mathbf{z}_{1k}y_k)$ and $(\mathbf{X}_1 - \hat{\mathbf{X}}_1)'\, (\hat{\mathbf{B}}_{(y;\,\mathbf{x}_1)s_1} - \mathbf{B}_{(y;\,\mathbf{x}_1)})$ are of lower order and can be ignored. This leads to the linearized form of this estimator.

$$\hat{Y}_{2P\,\mathrm{lin}} = \mathbf{X}'_1\, \mathbf{B}_{(y;\,\mathbf{x}_1)} - \hat{\mathbf{X}}_1\mathbf{B}_{(y;\,\mathbf{x}_1)} + \sum\nolimits_s a_k\, y_k. \quad (4.13)$$

We can also write this linearized form as a sum (4.9) of three residual terms, with the residuals e_{0k}, e_{1k} and e_{2k} having following definitions for $k \in U$.

$$e_{0k} = \mathbf{x}'_{1k}\, \mathbf{B}_{(y;\,\mathbf{x}_1)}$$

$$e_{1k} = -\mathbf{x}'_{1k}\mathbf{B}_{(y;\,\mathbf{x}_1)}$$

$$e_{2k} = y_k. \quad (4.14)$$

These residuals show a resemblance to those given by (4.10) if we set $\mathbf{x}_k = \phi$ and remove $\mathbf{B}_{(y;\,\mathbf{x})}$. Note how $\mathbf{B}_{(y;\,\mathbf{x}_1)}$ has the same role as $\mathbf{B}_{(\mathbf{x}\mathbf{B}_{(w)};\,\mathbf{x}_1)}$ in (4.10). As before, $e_{0k} + e_{1k} + e_{2k} = y_k$ for every k, and hence $\sum_U (e_{0k} + e_{1k} + e_{2k}) = \sum_U y_k = Y$.

The double-expansion estimator is a special case of this estimator when we also have $\mathbf{x}_{1k} = \phi$. This means that $\mathbf{B}_{(y;\,\mathbf{x}_1)}$ is not defined. The corresponding definitions for e_{0k}, e_{1k} and e_{2k} are simply $e_{0k} = 0$, $e_{1k} = 0$ and $e_{2k} = y_k$ for $k \in U$.

In the following sections, we examine the bias and variance of the two-phase calibration estimator $\hat{Y}_{2P\,a}$ and we propose a new method for estimation of variance. We can derive corresponding results when there is no phase-two calibration because the residuals for these two groups of estimators have similar properties and linearized form. The only difference occurs in the estimation of variance. We use the same variance estimator (as described in section 7) but the residuals are estimated by using $\hat{e}_{1k} = -\mathbf{x}'_{1k}\hat{\mathbf{B}}_{(y;\,\mathbf{x}_1)s}$ where $\hat{\mathbf{B}}_{(y;\,\mathbf{x}_1)s} = (\sum_s a_k\, \mathbf{z}_{1k}\, \mathbf{x}'_{1k})^{-1} \sum_s a_k\, \mathbf{z}_{1k}\, y_k$, and $\hat{e}_{2k} = y_k$.

5. Bias and variance of the two-phase calibration estimator $\hat{Y}_{2Pa}$

The two-phase calibration estimator $\hat{Y}_{2Pa} = \sum_s w_k y_k$ is approximately unbiased for $Y = \sum_U y_k$. To show this, we derive the expectation of the linearized form given by (4.9) via the usual method of conditioning on the phase-one sample s_1. We have $E(\sum_s a_k e_{2k}) = E_{s_1} E_{s|s_1}(\sum_s a_k e_{2k}) = E_{s_1}(\sum_{s_1} a_{1k} e_{2k}) = \sum_U e_{2k}$, $E_{s_1}(\sum_{s_1} a_{1k} e_{1k}) = \sum_U e_{1k}$, and $\sum_U e_{0k}$ is a constant term, so

$$E(\hat{Y}_{2Pa\,\mathrm{lin}}) = \sum\nolimits_U (e_{0k} + e_{1k} + e_{2k}) = \sum\nolimits_U y_k = Y.$$

This shows that $\hat{Y}_{2Pa\,\mathrm{lin}}$ is unbiased for Y. By (4.4), $\hat{Y}_{2Pa} = \hat{Y}_{2Pa\,\mathrm{lin}} + R$, so the bias of $\hat{Y}_{2Pa}$ equals the expectation of R, which is the sum of the two lower order terms $(\mathbf{X}_1 - \hat{\mathbf{X}}_1)'(\hat{\mathbf{B}}_{(\mathbf{x}\mathbf{B}_{(w)};\,\mathbf{x}_1)} - \mathbf{B}_{(\mathbf{x}\mathbf{B}_{(w)};\,\mathbf{x}_1)})$ and $(\mathbf{X} - \hat{\mathbf{X}})'(\hat{\mathbf{B}}_{(y;\,\mathbf{x})} - \mathbf{B}_{(y;\,\mathbf{x})})$. As pointed out in section 4, each of these terms has expectation close to zero. It follows that $\hat{Y}_{2Pa}$ is approximately unbiased for Y.

The variance of $\hat{Y}_{2Pa} = \sum_s w_k y_k$ is closely approximated by the variance of the linearized form $\hat{Y}_{2Pa\,\mathrm{lin}}$ given by (4.9) with residuals defined by (4.10). Its first term, $\sum_U e_{0k}$, is constant and does not contribute to the variance. Therefore,

$$V(\hat{Y}_{2Pa\,\mathrm{lin}}) = V\left(\sum\nolimits_{s_1} a_{1k} e_{1k} + \sum\nolimits_s a_k e_{2k}\right). \tag{5.1}$$

We use (5.1) as the starting point for deriving a variance estimator for $\hat{Y}_{2Pa\,\mathrm{lin}}$. Two different approaches can be used and it is of interest to compare them. The one in section 7 is new and more interesting because it produces a more efficient variance estimator than the one in section 8, derived by the traditional technique of conditioning on the phase-one sample s_1. The residuals e_{1k} and e_{2k} given by (4.10) play an important role in both derivations.

6. Preliminaries for variance estimation

Our objective is to estimate the variance $V(\hat{Y}_{2Pa\,\mathrm{lin}})$ given by (5.1). This is done in sections 7 and 8 by two different arguments. The residuals e_{1k} and e_{2k} are defined for all $k \in U$ but they can not be computed. They must be replaced by estimates $\hat{e}_{1k}$ and $\hat{e}_{2k}$. These estimates, formed in the image of (4.10) are

$$\begin{aligned}\hat{e}_{1k} &= \mathbf{x}'_{k(a)}\hat{\mathbf{B}}_{(y;\,\mathbf{x})(a)} + \mathbf{x}'_{k(w)}\hat{\mathbf{B}}_{(y;\,\mathbf{x})(w)} \\ &\quad - \mathbf{x}'_{1k}\hat{\mathbf{B}}_{(\mathbf{x}\hat{\mathbf{B}}_{(w)};\,\mathbf{x}_1)} \quad \text{for } k \in s_1 \\ \hat{e}_{2k} &= y_k - \mathbf{x}'_k\,\hat{\mathbf{B}}_{(y;\,\mathbf{x})} \\ &= y_k - \mathbf{x}'_{k(t)}\,\hat{\mathbf{B}}_{(y;\,\mathbf{x})(t)} - \mathbf{x}'_{k(w)}\hat{\mathbf{B}}_{(y;\,\mathbf{x})(w)} \\ &\quad - \mathbf{x}'_{k(a)}\hat{\mathbf{B}}_{(y;\,\mathbf{x})(a)} \quad \text{for } k \in s \end{aligned} \tag{6.1}$$

where

$$\begin{aligned}\hat{\mathbf{B}}_{(y;\,\mathbf{x})} &= \left(\sum\nolimits_s a_k \mathbf{z}_k \mathbf{x}'_k\right)^{-1} \sum\nolimits_s a_k \mathbf{z}_k y_k \\ &= (\hat{\mathbf{B}}'_{(y;\,\mathbf{x})(t)}, \hat{\mathbf{B}}'_{(y;\,\mathbf{x})(w)}, \hat{\mathbf{B}}'_{(y;\,\mathbf{x})(a)})'\end{aligned}$$

and (6.2)

$$\hat{\mathbf{B}}_{(\mathbf{x}\hat{\mathbf{B}}_{(w)};\,\mathbf{x}_1)} =$$

$$\left(\sum\nolimits_{s_1} a_{1k} \mathbf{z}_{1k} \mathbf{x}'_{1k}\right)^{-1} \sum\nolimits_{s_1} a_{1k} \mathbf{z}_{1k} \mathbf{x}'_{k(w)}\hat{\mathbf{B}}_{(y;\,\mathbf{x})(w)}.$$

The term $\hat{\mathbf{B}}_{(\mathbf{x}\hat{\mathbf{B}}_{(w)};\,\mathbf{x}_1)}$ in the definition of $\hat{e}_{1k}$ is the estimate of $\mathbf{B}_{(\mathbf{x}\mathbf{B}_{(w)};\,\mathbf{x}_1)} = (\sum_U \mathbf{z}_{1k}\mathbf{x}'_{1k})^{-1}\sum_U \mathbf{z}_{1k}\mathbf{x}'_{k(w)}\mathbf{B}_{(y;\,\mathbf{x})(w)}$ in (4.10). Two replacements are required in $\mathbf{B}_{(\mathbf{x}\mathbf{B}_{(w)};\,\mathbf{x}_1)}$ to arrive at $\hat{\mathbf{B}}_{(\mathbf{x}\hat{\mathbf{B}}_{(w)};\,\mathbf{x}_1)}$: First, sums over U are replaced by appropriately weighted sums over s_1, giving $\hat{\mathbf{B}}_{(\mathbf{x}\mathbf{B}_{(w)};\,\mathbf{x}_1)} = (\sum_{s_1} a_{1k}\mathbf{z}_{1k}\mathbf{x}'_{1k})^{-1}\sum_{s_1} a_{1k}\mathbf{z}_{1k}\mathbf{x}'_{k(w)}\mathbf{B}_{(y;\,\mathbf{x})(w)}$. In this expression, $\mathbf{B}_{(y;\,\mathbf{x})(w)}$ is still unknown, so we replace it by its estimate $\hat{\mathbf{B}}_{(y;\,\mathbf{x})(w)}$ to arrive at $\hat{\mathbf{B}}_{(\mathbf{x}\hat{\mathbf{B}}_{(w)};\,\mathbf{x}_1)}$.

A key point to note is that estimates $\hat{e}_{1k}$ can be obtained for $k \in s_1$, because $\mathbf{x}_{k(a)}$, $\mathbf{x}_{k(w)}$ and $\mathbf{x}_{1k}$ are all known for $k \in s_1$, but estimates $\hat{e}_{2k}$ can only be made for $k \in s$, because y_k is available only for $k \in s$. The fact that the estimates $\hat{e}_{1k}$ are available for $k \in s_1$ rather than $k \in s$ allows us to construct (in section 7) a more efficient estimator of $V(\hat{Y}_{2Pa\,\mathrm{lin}})$ than the traditional approach to variance estimation (in section 8) where all estimated residuals are calculated only for $k \in s$.

The design weights $a_{1k} = 1/\pi_{1k}$, $a_{2k} = 1/\pi_{2k}$ and $a_k = a_{1k}a_{2k}$ were defined in section 1. In the following sections, we also need the quantities given below, defined as functions of the second-order inclusion probabilities $\pi_{1k\ell} = \Pr(k \;\&\; \ell \in s_1)$ and $\pi_{2k\ell} = \Pr(k \& \ell \in s \mid s_1)$:

$$a_{1k\ell} = 1/\pi_{1k\ell},\; a_{2k\ell} = 1/\pi_{2k\ell},\; a_{k\ell} = a_{1k\ell}a_{2k\ell}$$

$$D_{1k\ell} = a_{1k}a_{1\ell} - a_{1k\ell},\; D_{2k\ell} = a_{2k}a_{2\ell} - a_{2k\ell},$$

$$D_{k\ell} = a_k a_\ell - a_{k\ell}.$$

Here, $\pi_{2k\ell}$ and $a_{2k\ell}$ are conditional on the sample s_1. All first-order and second-order inclusion probabilities are assumed positive. Using this notation and the above results, we now develop two different variance estimators in the next two sections.

7. The separate residual variance estimator

The variance of $\hat{Y}_{2Pa\,\mathrm{lin}}$ is given by (5.1), where e_{1k} and e_{2k} are defined by (4.10). It can be expanded as

$$V(\hat{Y}_{2P\,a\,\mathrm{lin}}) = V\left(\sum\nolimits_{s_1} a_{1k}\, e_{1k}\right) + V\left(\sum\nolimits_{s} a_k\, e_{2k}\right)$$

$$+ 2\ \mathrm{Cov}\left(\sum\nolimits_{s_1} a_{1k}\, e_{1k}, \sum\nolimits_{s} a_k\, e_{2k}\right). \qquad (7.1)$$

If we knew the residuals e_{1k} and e_{2k}, unbiased estimates for these three components would be given respectively by

$$\sum\nolimits_{k\in s_1}\sum\nolimits_{\ell\in s_1} D_{1k\ell}\, e_{1k}\, e_{1\ell},$$

$$\sum\nolimits_{k\in s}\sum\nolimits_{\ell\in s} D_{k\ell}\, e_{2k}\, e_{2\ell},$$

$$2\sum\nolimits_{k\in s_1}\sum\nolimits_{\ell\in s} D_{1k\ell}\, a_{2\ell}\, e_{1k}\, e_{2\ell}. \qquad (7.2)$$

The proof of unbiasedness is similar for all three components. For example, for the second one, we have

$$E_{s_1} E_{s|s_1}\left(\sum\nolimits_{k\in s}\sum\nolimits_{\ell\in s} D_{k\ell}\, e_{2k} e_{2\ell}\right)$$

$$= E_{s_1}\left(\sum\nolimits_{k\in s_1}\sum\nolimits_{\ell\in s_1} (D_{k\ell}/a_{2k\ell}) e_{2k} e_{2\ell}\right)$$

$$= \sum\nolimits_{k\in U}\sum\nolimits_{\ell\in U} (D_{k\ell}/a_{k\ell}) e_{2k} e_{2\ell}$$

$$= \sum\nolimits_{k\in U}\sum\nolimits_{\ell\in U} (a_k a_\ell/a_{k\ell}) e_{2k} e_{2\ell} - \left(\sum\nolimits_{U} e_{2k}\right)^2$$

$$= E\left[\left(\sum\nolimits_{s} a_k e_{2k}\right)^2\right] - \left[E\left(\sum\nolimits_{s} a_k e_{2k}\right)\right]^2$$

$$= V\left(\sum\nolimits_{s} a_k e_{2k}\right).$$

We now replace the unknown residuals in (7.2) by the respective estimates given by (6.1); that is, e_{1k} by $\hat{e}_{1k}$ for $k \in s_1$ and e_{2k} by $\hat{e}_{2k}$ for $k \in s$. Then, the resulting three components are added to arrive at the "separate residual" variance estimator

$$\hat{V}_{sr}(\hat{Y}_{2P\,a\,\mathrm{lin}}) = \sum\nolimits_{k\in s_1}\sum\nolimits_{\ell\in s_1} D_{1k\ell}\, \hat{e}_{1k}\, \hat{e}_{1\ell}$$

$$+ \sum\nolimits_{k\in s}\sum\nolimits_{\ell\in s} D_{k\ell}\, \hat{e}_{2k}\, \hat{e}_{2\ell}$$

$$+ 2\sum\nolimits_{k\in s_1}\sum\nolimits_{\ell\in s} D_{1k\ell} a_{2\ell}\, \hat{e}_{1k} \hat{e}_{2\ell}. \qquad (7.3)$$

The term "separate residual" and the corresponding subscript *sr* reflect the fact that (7.3) keeps the residuals separate, where $\hat{e}_{1k}$ is defined over the larger sample s_1 and $\hat{e}_{2k}$ over the smaller sample s. The fact that residuals computed for the larger sample s_1 can be advantageous for variance estimation was recognized by Axelson (1998). However, his derivation differs from our calibration approach based on $\mathbf{x}_{1k}$ and $\mathbf{x}_k$. The technique for variance estimation of the two-phase regression estimator in Hidiroglou, Rao and Haziza (2006) has certain traits in common with our approach, but there are also considerable differences.

8. The combined residual variance estimator

We arrived at (7.3) by recognizing that the estimates $\hat{e}_{1k}$ are obtainable for $k \in s_1$. The traditional approach, reviewed in this section, is to derive a variance estimator by conditioning on the phase-one sample s_1. This produces a variance estimator where all required residuals are defined for $k \in s$. Later, we compare it with the more efficient (7.3). From (5.1), we condition on the phase-one sample s_1 to obtain

$$V(\hat{Y}_{2P\,a\,\mathrm{lin}}) = V_{s_1} E_{s|s_1}\left(\sum\nolimits_{s_1} a_{1k} e_{1k} + \sum\nolimits_{s} a_k e_{2k}\right)$$

$$+ E_{s_1} V_{s|s_1}\left(\sum\nolimits_{s_1} a_{1k} e_{1k} + \sum\nolimits_{s} a_k e_{2k}\right)$$

$$= V_{s_1}\left(\sum\nolimits_{s_1} a_{1k} e_{1k} + \sum\nolimits_{s_1} a_{1k} e_{2k}\right)$$

$$+ E_{s_1} V_{s|s_1}\left(\sum\nolimits_{s} a_k e_{2k}\right)$$

$$= V_{s_1}\left(\sum\nolimits_{s_1} a_{1k} e_{12k}\right) + E_{s_1} V_{s|s_1}\left(\sum\nolimits_{s} a_k e_{2k}\right) \qquad (8.1)$$

where $e_{12k} = e_{1k} + e_{2k}$ is called the combined residual. From (4.10), we obtain the following.

$$e_{12k} = y_k - \mathbf{x}'_{k(t)}\,\mathbf{B}_{(y;\mathbf{x})(t)} - \mathbf{x}'_{1k}\mathbf{B}_{(\mathbf{x}\mathbf{B}_{(w)};\mathbf{x}_1)}$$

$$e_{2k} = y_k - \mathbf{x}'_{k(t)}\,\mathbf{B}_{(y;\mathbf{x})(t)} - \mathbf{x}'_{k(w)}\mathbf{B}_{(y;\mathbf{x})(w)}$$

$$- \mathbf{x}'_{k(a)}\mathbf{B}_{(y;\mathbf{x})(a)}. \qquad (8.2)$$

It is straightforward to define estimators of the two components $V_{s_1}(\sum_{s_1} a_{1k}\, e_{12k})$ and $E_{s_1} V_{s|s_1}(\sum_{s} a_k\, e_{2k})$. Each of these has the form of a double sum over *s* because e_{12k} and e_{2k} contain y_k which is only available for $k \in s$. The first component uses $\hat{e}_{12k} = \hat{e}_{1k} + \hat{e}_{2k} = y_k - \mathbf{x}'_{k(t)}\hat{\mathbf{B}}_{(y;\,\mathbf{x})(t)} - \mathbf{x}'_{1k}\hat{\mathbf{B}}_{(\mathbf{x}\hat{\mathbf{B}}_{(w)};\,\mathbf{x}_1)}$ for $k \in s$. We then have $\sum_{k\in s}\sum_{\ell\in s} D_{1k\ell}\, a_{2k\ell}\, \hat{e}_{12k}\, \hat{e}_{12\ell}$ as an estimator of $V_{s_1}(\sum_{s_1} a_{1k}\, e_{12k})$.

For the second component, we use the residual estimates $\hat{e}_{2k} = y_k - \mathbf{x}'_k\hat{\mathbf{B}}_{(y;\,\mathbf{x})}$ given by (6.1) for $k \in s$, and obtain $\sum_{k\in s}\sum_{\ell\in s} D_{2k\ell}\, a_{1k}\, a_{1\ell}\, \hat{e}_{2k}\, \hat{e}_{2\ell}$ as an estimator of $E_{s_1} V_{s|s_1}(\sum_{s} a_k e_{2k})$. Summing the two estimated terms we have the following variance estimator, where the subscript *cr* indicates "combined residual",

$$\hat{V}_{cr}(\hat{Y}_{2P\,a\,\mathrm{lin}}) = \sum\nolimits_{k\in s}\sum\nolimits_{\ell\in s} D_{1k\ell}\, a_{2k\ell}\, \hat{e}_{12k}\, \hat{e}_{12\ell}$$

$$+ \sum\nolimits_{k\in s}\sum\nolimits_{\ell\in s} D_{2k\ell}\, a_{1k}\, a_{1\ell}\, \hat{e}_{2k}\, \hat{e}_{2\ell}. \qquad (8.3)$$

Let us review how (7.3) and (8.3) differ. The separate residual variance estimator (7.3) starts with the expansion $V(\hat{Y}_{2Pa\,\text{lin}}) = V(\sum_{s_1} a_{1k} e_{1k}) + V(\sum_s a_k e_{2k}) + 2\,\text{Cov}(\sum_{s_1} a_{1k} e_{1k}, \sum_s a_k e_{2k})$. We estimate these three components separately as functions of the residuals e_{1k} and e_{2k}. The resulting variance expression has three terms: a double sum over s_1 in terms of e_{1k} and $e_{1\ell}$, a double sum over s in terms of e_{2k} and $e_{2\ell}$, and a cross-sum over s_1 and s in terms of $e_{1k} \in s_1$ and $e_{2\ell} \in s.$ Finally, we arrive at (7.3) by estimating e_{1k} by $\hat{e}_{1k}$ for $k \in s_1$ and e_{2k} by $\hat{e}_{2k}$ for $k \in s.$

The combined residual variance estimator (8.3) arises from the traditional conditioning on the phase-one sample s_1 as $V(\hat{Y}_{2P\,a\,\text{lin}}) = V_{s_1} E_{s|s_1}(\hat{Y}_{2P\,a\,\text{lin}}) + E_{s_1} V_{s|s_1}(\hat{Y}_{2P\,a\,\text{lin}}).$ This leads us to combine e_{1k} and e_{2k} as $e_{12k} = e_{1k} + e_{2k}$ in the first term. The second term, $E_{s_1} V_{s|s_1}(\hat{Y}_{2P\,a\,\text{lin}})$, is a function of e_{2k}. Since e_{12k} and e_{2k} can only be estimated over s, the resulting variance estimator becomes a sum of two terms, each of them expressed as a double sum over $s.$

The separate residual estimator (7.3) is more efficient than the combined residual alternative (8.3), because it is based on residuals $\hat{e}_{1k}$ obtained for the typically larger sample s_1. The advantage of (7.3) over (8.3) is illustrated by the simulation in section 10. The approach behind the separate residual variance estimator (7.3) can be extended to three-phase sampling and other complex designs. In those extensions of the technique, we proceed in a similar manner, starting by a derivation of the linearized form through an expansion of the variance components and the determination of the appropriate residuals.

9. A comparison with the two-phase regression estimator

Särndal, Swensson and Wretman (1992) developed a two-phase regression estimator for $Y = \sum_U y_k$, based on an earlier paper by Särndal and Swensson (1989). It is useful to see how this estimator, denoted here by $\hat{Y}_{\text{reg}}$, compares with the calibration estimator $\hat{Y}_{2P}$ considered in the preceding sections of this paper. When based on the same auxiliary information, the two estimators are "close" but not identical. This is because the estimator $\hat{Y}_{2P}$ is derived by calibration in each of the two phases, whereas the two-phase regression estimator $\hat{Y}_{\text{reg}}$ is derived by model-assisted reasoning.

We now describe the two-phase regression estimator of Särndal, Swensson and Wretman (1992). Their derivation involves the fit of two linear regression models with the use of the available auxiliary data; one at "the top level" and the other at "the bottom level". These authors develop a corresponding estimator of variance, via the traditional conditioning argument. We compare their variance estimator with the combined residual variance estimator (8.3), also developed by the conditioning argument. The two variance estimators do not agree exactly, because the point estimators are slightly different, but they are numerically close, as shown in this section.

Let $\mathbf{x}_{1k}$ be a vector of auxiliary variables with known population totals, and let $\mathbf{x}_k = (\mathbf{x}'_{1k}, \mathbf{x}'_{2k})'$, where both $\mathbf{x}_{1k}$ and $\mathbf{x}_{2k}$ are known vector values for $k \in s_1$. The total $\sum_U \mathbf{x}_{1k}$ is assumed known whereas the total $\sum_U \mathbf{x}_{2k}$ is unknown. The predicted values produced for $k \in s_1$ by the two regressions fitted at the "top level" and "bottom level" are given respectively by

$$\hat{y}_{1k} = \mathbf{x}'_{1k}\hat{\mathbf{B}}_{1s}$$

with (9.1)

$$\hat{\mathbf{B}}_{1s} = \left(\sum\nolimits_s a_k \mathbf{x}_{1k}\mathbf{x}'_{1k}/\sigma^2_{1k}\right)^{-1}\left(\sum\nolimits_s a_k \mathbf{x}_{1k} y_k/\sigma^2_{1k}\right)$$

and

$$\hat{y}_k = \mathbf{x}'_k\hat{\mathbf{B}}_s$$

with (9.2)

$$\hat{\mathbf{B}}_s = \left(\sum\nolimits_s a_k \mathbf{x}_k\mathbf{x}'_k/\sigma^2_k\right)^{-1}\sum\nolimits_s a_k \mathbf{x}_k y_k/\sigma^2_k.$$

The resulting two-phase regression estimator $\hat{Y}_{\text{reg}}$ of $Y = \sum_U y_k$ is

$$\hat{Y}_{\text{reg}} = \left(\sum\nolimits_U \mathbf{x}_{1k}\right)'\hat{\mathbf{B}}_{1s} + \sum\nolimits_{s_1} a_{1k}(\hat{y}_k - \hat{y}_{1k}) + \sum\nolimits_s a_k(y_k - \hat{y}_k). \tag{9.3}$$

Can $\hat{Y}_{\text{reg}}$ be interpreted as a calibration estimator? To answer this question, let us determine the implicit weights in (9.3). We can write $\hat{Y}_{\text{reg}} = \sum_s w_k y_k$, with weights w_k identified by substituting (9.1) and (9.2) into (9.3) and simplifying. We find $w_k = a_k g_k = a_{1k} a_{2k} g_k$, where the calibration factor g_k is given for $k \in s$ by

$$g_k = 1 + \left(\sum\nolimits_U \mathbf{x}_{1k} - \sum\nolimits_{s_1} a_{1k}\mathbf{x}_{1k}\right)' \left(\sum\nolimits_s a_k \mathbf{x}_{1k}\mathbf{x}'_{1k}/\sigma^2_{1k}\right)^{-1}\mathbf{x}_{1k}/\sigma^2_{1k} + \left(\sum\nolimits_{s_1} a_{1k}\mathbf{x}_k - \sum\nolimits_s a_k\mathbf{x}_k\right)' \left(\sum\nolimits_s a_k \mathbf{x}_k\mathbf{x}'_k/\sigma^2_k\right)^{-1}\mathbf{x}_k/\sigma^2_k. \tag{9.4}$$

The weights w_k are not explicitly stated in Särndal, Swensson and Wretman (1992). In what sense, if any, can w_k be considered a calibration weight? To examine this, we first replace y_k in (9.3) with $\mathbf{x}'_{1k}$. Using (9.1) and (9.2) with $y_k = \mathbf{x}'_{1k}$ gives $\sum_U \mathbf{x}'_{1k}$ as the right-hand side of (9.3). Thus, the weights $w_k = a_k g_k$ satisfy $\sum_s w_k \mathbf{x}_{1k} = \sum_U \mathbf{x}_{1k}$. Next we replace y_k in (9.3) with $\mathbf{x}'_{2k}$, again using (9.1) and (9.2) to obtain

$$\textstyle\sum_{s_1} a_{1k}\mathbf{x}'_{2k} + \left(\sum_U \mathbf{x}_{1k} - \sum_{s_1} a_{1k}\mathbf{x}_{1k}\right)' \left(\sum_s a_k \mathbf{x}_{1k}\mathbf{x}'_{1k}/\sigma^2_{1k}\right)^{-1} \left(\sum_s a_k \mathbf{x}_{1k}\mathbf{x}'_{2k}/\sigma^2_{1k}\right). \quad (9.5)$$

Although (9.5) is an approximately unbiased estimate of the unknown $\mathbf{x}_{2k}$ -total $\sum_U \mathbf{x}'_{2k}$, it does not have the usual form of the right-hand side of a phase-two calibration equation, such as $\sum_{s_1} a_{1k}\mathbf{x}'_{2k}$ or $\sum_{s_1} w_{1k}\mathbf{x}'_{2k}$. However, it is close. If we replace the two sums over s with appropriately weighted sums over s_1, then (9.5) becomes $\sum_{s_1} w_{1k}\mathbf{x}'_{2k}$ where w_{1k} is given by (2.1) with $\mathbf{z}_{1k} = \mathbf{x}_{1k}/\sigma^2_{1k}$. Thus, the implicit weights w_k in $\hat{Y}_{\text{reg}}$ calibrate exactly on the known population $\mathbf{x}_{1k}$ -total, and they come close to calibrating on the estimated $\mathbf{x}_{2k}$ -total $\sum_{s_1} w_{1k}\mathbf{x}'_{2k}$. This suggests that $\hat{Y}_{\text{reg}}$ should have properties similar to an estimator $\hat{Y}_{2P}$ obtained by defining $\mathbf{x}_k$ in $\hat{Y}_{2P}$ as $\mathbf{x}_k = (\mathbf{x}'_{1k}, \mathbf{x}'_{2k})'$ with $\mathbf{x}_{k(t)} = \mathbf{x}_{1k}$, $\mathbf{x}_{k(w)} = \mathbf{x}_{2k}$ and $\mathbf{x}_{k(a)} = \phi$. In addition, the form of the model-assisted estimator implies $\mathbf{z}_{1k} = \mathbf{x}_{1k}/\sigma^2_{1k}$ and $\mathbf{z}_k = \mathbf{x}_k/\sigma^2_k$. Since $\mathbf{x}_k$ includes $\mathbf{x}_{1k}$ it is reasonable to define $\mathbf{z}_k = \mathbf{x}_k/\sigma^2_k$ as $\mathbf{z}_k = (\mathbf{x}'_{1k}/\sigma^2_{1k}, \mathbf{x}'_{2k}/\sigma^2_{2k})'$. These specifications meet the requirements for asymptotic equivalence of $\hat{Y}_{2Pa}$ and $\hat{Y}_{2Pw}$ so we do not need to worry about the choice of starting weights in $\hat{Y}_{2P}$. We can simply work with $\hat{Y}_{2Pa}$ as the estimator comparable to $\hat{Y}_{\text{reg}}$. Now, let us look at variance estimation for $\hat{Y}_{\text{reg}}$ and the estimator $\hat{Y}_{2Pa}$ under these specifications.

The variance estimator of Särndal, Swensson and Wretman (1992) contains calibration factors denoted g_{ks} and g_{1ks_1}. They are not to be confused with g_k given by (9.4). If we disregard g_{ks} and g_{1ks_1}, both of which are near one and of limited numerical impact, their variance estimator is

$$\hat{V}(\hat{Y}_{\text{reg}}) = \sum_{k\in s}\sum_{\ell\in s} D_{1k\ell}\, a_{2k\ell}\, \hat{e}_{1ks}\, \hat{e}_{1\ell s} + \sum_{k\in s}\sum_{\ell\in s} D_{2k\ell}\, a_{1k}\, a_{1\ell}\, \hat{e}_{ks}\, \hat{e}_{\ell s} \quad (9.6)$$

where, for $k \in s$,

$$\hat{e}_{1ks} = y_k - \mathbf{x}'_{1k}\hat{\mathbf{B}}_{1s} \text{ and } \hat{e}_{ks} = y_k - \mathbf{x}'_k\hat{\mathbf{B}}_s. \quad (9.7)$$

Both components of (9.6) are double sums over s, reflecting the fact that both $\hat{e}_{1ks}$ and $\hat{e}_{ks}$ can only be obtained for $k \in s$. Formula (9.6) looks similar to formula (8.3) for the combined residual estimator but how different are the residuals in the two formulas? Let us look at the residuals for the comparable point estimator. As noted above, this estimator $\hat{Y}_{2P}$ has $\mathbf{x}'_k = (\mathbf{x}'_{1k}, \mathbf{x}'_{2k})$ with $\mathbf{x}_{k(t)} = \mathbf{x}_{1k}$, $\mathbf{x}_{k(w)} = \mathbf{x}_{2k}$, $\mathbf{x}_{k(a)} = \phi$, $\mathbf{z}_{1k} = \mathbf{x}_{1k}/\sigma^2_{1k}$ and $\mathbf{z}_k = \mathbf{x}_k/\sigma^2_k = (\mathbf{x}'_{1k}/\sigma^2_{1k}, \mathbf{x}'_{2k}/\sigma^2_{2k})'$. Under these specifications, the residuals $\hat{e}_{1k}$ and $\hat{e}_{2k}$ in (6.1) are given by

$$\hat{e}_{1k} = \mathbf{x}'_{2k}\hat{\mathbf{B}}_{(y;\,\mathbf{x})(2)} - \mathbf{x}'_{1k}\hat{\mathbf{B}}_{(\mathbf{x}\hat{\mathbf{B}}_{(2)};\,\mathbf{x}_1)} \quad \text{for } k \in s_1$$

$$\hat{e}_{2k} = y_k - \mathbf{x}'_k\hat{\mathbf{B}}_{(y;\,\mathbf{x})} = y_k - \mathbf{x}'_{1k}\hat{\mathbf{B}}_{(y;\,\mathbf{x})(1)} - \mathbf{x}'_{2k}\hat{\mathbf{B}}_{(y;\,\mathbf{x})(2)} \quad \text{for } k \in s \quad (9.8)$$

where $\hat{\mathbf{B}}_{(y;\,\mathbf{x})} = (\hat{\mathbf{B}}'_{(y;\,\mathbf{x})(1)}, \hat{\mathbf{B}}'_{(y;\,\mathbf{x})(2)})'$ corresponds to the partitioning of $\mathbf{x}_k = (\mathbf{x}'_{1k}, \mathbf{x}'_{2k})'$ and from (6.2)

$$\hat{\mathbf{B}}_{(y;\mathbf{x})} = \left(\sum_s a_k\mathbf{z}_k\mathbf{x}'_k\right)^{-1}\left(\sum_s a_k\mathbf{z}_k y_k\right)$$

$$\hat{\mathbf{B}}_{(\mathbf{x}\hat{\mathbf{B}}_{(2)};\mathbf{x}_1)} = \left(\sum_{s_1} a_{1k}\mathbf{x}_{1k}\mathbf{x}'_{1k}/\sigma^2_{1k}\right)^{-1} \left(\sum_{s_1} a_{1k}\mathbf{x}_{1k}\mathbf{x}'_{2k}\hat{\mathbf{B}}_{(y;\mathbf{x})(2)}/\sigma^2_{1k}\right). \quad (9.9)$$

The residuals $\hat{e}_{2k}$ in (9.8) are the same as $\hat{e}_{ks}$ in (9.7). But how do the residuals $\hat{e}_{12k} = \hat{e}_{1k} + \hat{e}_{2k}$, obtained by adding in (9.8), relate to their counterparts $\hat{e}_{1ks}$ in (9.7)? To find this link, we first show that $\hat{\mathbf{B}}_{1s} = (\sum_s a_k\mathbf{x}_{1k}\mathbf{x}'_{1k}/\sigma^2_{1k})^{-1} \sum_s a_k\mathbf{x}_{1k}y_k/\sigma^2_{1k}$ can be written as

$$\hat{\mathbf{B}}_{1s} = \hat{\mathbf{B}}_{(y;\,\mathbf{x})(1)} + \left(\sum_s a_k\mathbf{x}_{1k}\mathbf{x}'_{1k}/\sigma^2_{1k}\right)^{-1}\left(\sum_s a_k\mathbf{x}_{1k}\mathbf{x}'_{2k}\hat{\mathbf{B}}_{(y;\,\mathbf{x})(2)}/\sigma^2_{1k}\right). \quad (9.10)$$

To see this, we start with $\hat{\mathbf{B}}_{(y;\,\mathbf{x})}$, which by definition satisfies $\sum_s a_k\mathbf{z}_k y_k = (\sum_s a_k\mathbf{z}_k\mathbf{x}'_k)\hat{\mathbf{B}}_{(y;\,\mathbf{x})}$. This equality can also be written as $\sum_s a_k\mathbf{z}_k y_k = \sum_s a_k\mathbf{z}_k(\mathbf{x}'_{1k}\hat{\mathbf{B}}_{(y;\,\mathbf{x})(1)} + \mathbf{x}'_{2k}\hat{\mathbf{B}}_{(y;\,\mathbf{x})(2)})$. Since $\mathbf{z}_k = (\mathbf{x}'_{1k}/\sigma^2_{1k}, \mathbf{x}'_{2k}/\sigma^2_{2k})'$, the component of this equation corresponding to $\mathbf{x}_{1k}$ is $\sum_s a_k\mathbf{x}_{1k}y_k/\sigma^2_{1k} = \sum_s a_k\mathbf{x}_{1k}\mathbf{x}'_{1k}\hat{\mathbf{B}}_{(y;\,\mathbf{x})(1)}/\sigma^2_{1k} + \sum_s a_k\mathbf{x}_{1k}\mathbf{x}'_{2k}\hat{\mathbf{B}}_{(y;\,\mathbf{x})(2)}/\sigma^2_{1k}$. Premultiplying both sides by $(\sum_s a_k\mathbf{x}_{1k}\mathbf{x}'_{1k}/\sigma^2_{1k})^{-1}$, we obtain (9.10).

Then, starting with (9.8) and using the definition of $\hat{\mathbf{B}}_{(\mathbf{x}\hat{\mathbf{B}}_{(2)};\,\mathbf{x}_1)}$ given by (9.9), we have

$$\begin{aligned}\hat{e}_{12k} &= \hat{e}_{1k} + \hat{e}_{2k}\\ &= y_k - \mathbf{x}'_{1k}\hat{\mathbf{B}}_{(y;\,\mathbf{x})(1)} - \mathbf{x}'_{1k}\hat{\mathbf{B}}_{(\mathbf{x}\hat{\mathbf{B}}_{(2)};\,\mathbf{x}_1)}\\ &= y_k - \mathbf{x}'_{1k}\left\{\hat{\mathbf{B}}_{(y;\,\mathbf{x})(1)} + \left(\sum_{s_1} a_{1k}\mathbf{x}_{1k}\mathbf{x}'_{1k}/\sigma^2_{1k}\right)^{-1} \left(\sum_{s_1} a_{1k}\mathbf{x}_{1k}\mathbf{x}'_{2k}\hat{\mathbf{B}}_{(y;\,\mathbf{x})(2)}/\sigma^2_{1k}\right)\right\}.\end{aligned}$$

In the expression within curly brackets, let us replace the two a_{1k} -weighted sums over s_1 with the corresponding a_k -weighted sums over s; the result is equal to $\hat{\mathbf{B}}_{1s}$ as given by (9.10). This means $\hat{e}_{12k} \cong y_k - \mathbf{x}'_{1k}\hat{\mathbf{B}}_{1s} = \hat{e}_{1ks}$. In summary, $\hat{e}_{12k} \cong \hat{e}_{1ks}$ for $k \in s$ and $\hat{e}_{2k} = \hat{e}_{ks}$ for $k \in s$. Hence, the variance estimator (9.6) for the two-phase regression estimator $\hat{Y}_{\text{reg}}$ should be numerically close to the combined residual variance estimator (8.3) for the calibration estimator $\hat{Y}_{2P}$ defined in this section. We present empirical support for this through the simulation in next section.

10. Simulation

In this section we present a small simulation to validate the claim that the separate residual variance estimator $\hat{V}_{sr}(\hat{Y}_{2P\,a\,\mathrm{lin}})$ given by (7.3) can be considerably more efficient than the combined residual variance estimator $\hat{V}_{cr}(\hat{Y}_{2P\,a\,\mathrm{lin}})$ given by (8.3), and that the behaviour of the latter is very similar to that of the two-phase regression estimator $\hat{V}(\hat{Y}_{\mathrm{reg}})$ given by (9.6). We created a population of $N = 5{,}000$ units in two steps as follows: First, the values (u_{1k}, u_{2k}) for $k = 1, 2, \ldots, 5{,}000$ were generated by 5,000 realizations of the independent random variables $u_{1k} \sim 2\,\mathrm{Gamma}(4)$ and $u_{2k} \sim 3\,\mathrm{Gamma}(6)$, where the $\mathrm{Gamma}(a)$ distribution has density $f(x) = [\Gamma(a)]^{-1}$ $x^{a-1}e^{-x}$ for $x > 0$. Secondly, the values of the variable of interest were created as $y_k = 10 + u_{1k} + 3u_{2k} + \varepsilon_k$, $k = 1, 2, \ldots 5{,}000$, with $\varepsilon_k \sim 5\,\mathrm{Normal}(0)$, where $\mathrm{Normal}(0)$ is the standard Normal distribution with mean 0 and variance 1. The target of estimation in the experiment is the population y-total $Y = \sum_U y_k = 358{,}205$. For the phase-one calibration, we used the auxiliary vector $\mathbf{x}_{1k} = (1, u_{1k})'$ and $\mathbf{z}_{1k} = \mathbf{x}_{1k}$. That is, the weights w_{1k} for $k \in s_1$ were determined by calibration to the known total $(N, \sum_U u_{1k}) = (5{,}000,\ 39{,}611.8)$. For the phase-two calibration we used $\mathbf{x}_k = (\mathbf{x}'_{k(t)}, \mathbf{x}'_{k(w)}, \mathbf{x}'_{k(a)})'$ with $\mathbf{x}_{k(t)} = (1, u_{1k})'$, $\mathbf{x}_{k(w)} = u_{2k}$, $\mathbf{x}_{k(a)} = \phi$ and $\mathbf{z}_k = \mathbf{x}_k$. These specifications satisfy the conditions for asymptotic equivalence between $\hat{Y}_{2P\,a}$ and $\hat{Y}_{2P\,w}$. Therefore, for this simulation, we can work with $\hat{Y}_{2P\,a}$ and its linearized form $\hat{Y}_{2P\,a\,\mathrm{lin}}$.

For each phase-one sample s_1, the final weights w_k for the estimator $\hat{Y}_{2P\,a} = \sum_s w_k y_k$ were determined by calibrating to the known totals given by the vector $(N, \sum_U u_{1k}, \sum_{s_1} w_{1k}\,u_{2k}) = (5{,}000, 39{,}611.8, \sum_{s_1} w_{1k}u_{2k})$. It is important to note that it was not necessary to have $\mathbf{x}_{k(a)} = \phi$ in order to run a simulation to compare $\hat{V}_{sr}(\hat{Y}_{2P\,a\,\mathrm{lin}})$ and $\hat{V}_{cr}(\hat{Y}_{2P\,a\,\mathrm{lin}})$. However, we can not compare $\hat{V}_{cr}(\hat{Y}_{2P\,a\,\mathrm{lin}})$ and $\hat{V}(\hat{Y}_{\mathrm{reg}})$ unless we define an estimator $\hat{Y}_{2P\,a}$ comparable to $\hat{Y}_{\mathrm{reg}}$, and to achieve this we need $\mathbf{x}_{k(a)} = \phi$, as noted in section 9.

We drew repeated sample pairs (s_1, s), where s_1 is an SRS of n_1 units from U, and s is an SRS of n units from s_1. Here SRS stands for simple random sampling without replacement. We worked with different size combinations (n_1, n): (4000, 3000), (4000, 2000), (4000, 1000), (3000, 2000), (3000, 1000) and (2000, 1000). If $n = n_1$, two-phase sampling is equivalent to one-phase sampling, and $\hat{V}_{sr}(\hat{Y}_{2P\,a\,\mathrm{lin}})$ and $\hat{V}_{cr}(\hat{Y}_{2P\,a\,\mathrm{lin}})$ are identical.

For each combination (n_1, n), we realized 100,000 sample pairs (s_1, s). Based on the data for each of these outcomes, we computed the separate residual variance estimator $\hat{V}_{sr}(\hat{Y}_{2P\,a\,\mathrm{lin}})$, the combined residual variance estimator $\hat{V}_{cr}(\hat{Y}_{2P\,a\,\mathrm{lin}})$ and the variance estimator $\hat{V}(\hat{Y}_{\mathrm{reg}})$. For this purpose, we used the respective expressions that follow from (7.3), (8.3) and (9.6) when SRS is specified at each phase. To save space, these expressions are not shown here. We obtained 100,000 realized values for each of the three variance estimators. Figure 10.1 shows the distributions of the 100,000 $\hat{V}$-values for $n_1 = 4{,}000$ and $n = 2{,}000$.

The figure shows strikingly different distributions for $\hat{V}_{sr}(\hat{Y}_{2P\,a\,\mathrm{lin}})$ and $\hat{V}_{cr}(\hat{Y}_{2P\,a\,\mathrm{lin}})$. The distribution of the separate residual estimator $\hat{V}_{sr}(\hat{Y}_{2P\,a\,\mathrm{lin}})$ is much more concentrated. Thus $\hat{V}_{sr}(\hat{Y}_{2P\,a\,\mathrm{lin}})$ is more efficient than $\hat{V}_{cr}(\hat{Y}_{2P\,a\,\mathrm{lin}})$ and on average, it produces considerably shorter confidence intervals. We also note that the distribution of $\hat{V}(\hat{Y}_{\mathrm{reg}})$ is very similar to that of $\hat{V}_{cr}(\hat{Y}_{2P\,a\,\mathrm{lin}})$. This supports our analysis in section 9. Similar results were obtained for the other sample sizes in the simulation.

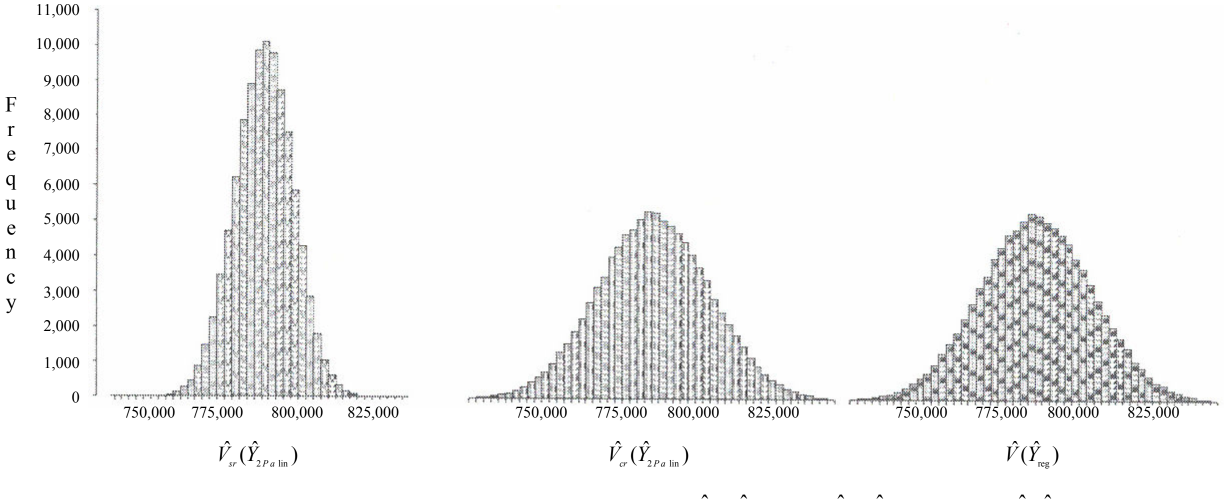


Figure 10.1 Distribution of 100,000 realized values for $\hat{V}_{sr}(\hat{Y}_{2P\,a\,\mathrm{lin}})$, $\hat{V}_{cr}(\hat{Y}_{2P\,a\,\mathrm{lin}})$ and $\hat{V}(\hat{Y}_{\mathrm{reg}})$

To obtain a measure of the efficiency of the three variance estimators, we computed the simulation variance of the 100,000 $\hat{V}$-values. These simulation variances are shown in Table 10.1, Table 10.2 and Table 10.3. The numbers are dramatically lower for $\hat{V}_{sr}(\hat{Y}_{2P\,a\,\mathrm{lin}})$ than for the other two. Table 10.4 shows the relative advantage of $\hat{V}_{sr}(\hat{Y}_{2P\,a\,\mathrm{lin}})$ over $\hat{V}_{cr}(\hat{Y}_{2P\,a\,\mathrm{lin}})$. For this population, the simulation variance of $\hat{V}_{sr}(\hat{Y}_{2P\,a\,\mathrm{lin}})$ is less than half the simulation variance of $\hat{V}_{cr}(\hat{Y}_{2P\,a\,\mathrm{lin}})$.

Table 10.1
Simulation variance for the separate residual variance estimator $\hat{V}_{sr}(\hat{Y}_{2P\,a\,\mathrm{lin}})$

		n	
n_1	3,000	2,000	1,000
4,000	64.82	95.91	484.92
3,000		1,179.62	1,806.79
2,000			13,995.94

Note: Actual values are the displayed values times 10^6.

Table 10.2
Simulation variance for the combined residual variance estimator $\hat{V}_{cr}(\hat{Y}_{2P\,a\,\mathrm{lin}})$

		n	
n_1	3,000	2,000	1,000
4,000	153.22	364.08	1,290.41
3,000		2,449.05	6,855.69
2,000			33,220.88

Note: Actual values are the displayed values times 10^6.

Table 10.3
Simulation variance for the variance estimator $\hat{V}(\hat{Y}_{\mathrm{reg}})$

		n	
n_1	3,000	2,000	1,000
4,000	153.25	364.14	1,289.79
3,000		2,449.36	6,854.52
2,000			33,210.31

Note: Actual values are the displayed values times 10^6.

Table 10.4
Ratio of entries in Table 10.1 to corresponding entries in Table 10.2

		n	
n_1	3,000	2,000	1,000
4,000	0.42	0.26	0.38
3,000		0.48	0.26
2,000			0.42

11. Discussion

In a design-based perspective on estimation for two-phase sampling designs, one can follow a regression estimation approach or a calibration estimation approach. We concentrate on the calibration approach to create approximately design-unbiased estimators. The extent of the information available for the calibration holds the key to the efficiency of the estimates. We recognize in this paper that there are three different types of auxiliary variables associated with two-phase designs. They have different information characteristics. From these we define four different auxiliary vectors; one for the phase-one calibration and the other three for the phase-two calibration. The calibration approach is suitable for analyzing the resulting estimators in a systematic manner. As the paper shows, this approach also leads to a more efficient variance estimator than the traditional method for variance estimation in two-phase designs.

References

Axelson, M. (1998). Variance estimation for the generalised regression estimator under two-phase sampling - a modified approach. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 85-89.

Deville, J.-C., and Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the Ame,rican Statistical Association*, 87, 376-382.

Deville, J.-C. (2002). La correction de la nonréponse par calage généralisé. *Actes des Journeés de Méthodologie*, I.N.S.E.E., Paris.

Dupont, F. (1995). Alternative adjustments when there are several levels of auxiliary information. *Survey Methodology*, 21, 125-136.

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

Hidiroglou, M.A. (2001). Double sampling. *Survey Methodology*, 27, 143-154.

Hidiroglou, M.A., and Särndal, C.-E. (1998). Use of auxiliary information for two-phase sampling. *Survey Methodology*, 24, 11-20.

Hidiroglou, M.A., Rao, J.N.K. and Haziza, D. (2006). Variance estimation in two phase sampling. (Accepted paper to appear in) *Australian and New Zealand Journal of Statistics*.

Kott, P.S., and Stukel, D.M. (1997). Can the jackknife be used with a two-phase sample? *Survey Methodology*, 23, 81-90.

Särndal, C.-E., and Swensson, B. (1987). A general view of estimation for two phases of selection with applications to two-phase sampling and nonresponse. *International Statistical Review*, 55, 279-294.

Särndal, C.-E., Swensson, B. and Wretman, J. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.

Sitter, R.R. (1997). Variance estimation for the regression estimator in two-phase sampling. *Journal of the American Statistical Association*, 92, 780-787.

Survey Methodology, June 2009
Vol. 35, No. 1, pp. 15-24
Statistics Canada, Catalogue No. 12-001-X

Survey weighted hat matrix and leverages

Jianzhu Li and Richard Valliant [1]

Abstract

Regression diagnostics are geared toward identifying individual points or groups of points that have an important influence on a fitted model. When fitting a model with survey data, the sources of influence are the response variable **Y**, the predictor variables **X**, and the survey weights, **W**. This article discusses the use of the hat matrix and leverages to identify points that may be influential in fitting linear models due to large weights or values of predictors. We also contrast findings that an analyst will obtain if ordinary least squares is used rather than survey weighted least squares to determine which points are influential.

Key Words: Influence; Linear regression; Survey data; Weighted least squares.

1. Introduction

In some conventional linear regression diagnostics, it is often useful to measure the influence each data point can have in determining the values of parameter estimates and, in turn, fitted values. The hat matrix and its diagonal elements, referred to as leverages, are popular techniques that are used to identify the cases that have outlying values for predictor variables, and, therefore, may be influential in model fitting if they are also associated with unusual residuals. When there is more than one predictor variable in the regression, analysts can compute leverages to summarize the collective influence of the **X** values for each observation.

In finite population estimation, a superpopulation assumption is usually used to build models. Suppose that some model fits reasonably well for the bulk of the population. For convenience, we will refer to this as the “true” model. However, the goal is usually to find a model that has some descriptive or predictive power, bearing in mind that no model is really “true”. The influence diagnostics should allow analysts to identify points that make estimated parameters deviate from that true model. Parameter estimates in linear regression using complex survey data are often derived from the pseudo maximum likelihood approach, outlined by Skinner, Holt and Smith (1989, Chapter 3), following ideas of Binder (1983). In this paper, we assume that the analyst has decided that an estimator involving sample weights is appropriate for his or her problem. As shown in later sections, the survey weighted hat matrix and leverages are useful for detecting potentially influential observations caused by not only extreme **X** values, but also by large sample weights.

Previous survey literature has discussed the effect of outliers on some survey estimates, but does not give much attention to diagnostics for linear regression models. Deville and Särndal (1992), and Potter (1990, 1993) discuss some possibilities for locating or trimming extreme survey weights when the goal is to estimate population totals and other simple descriptive statistics. Hulliger (1995) and Moreno-Rebollo, Muñoz-Reyes and Muñoz-Pichardo (1999) address the effect of outliers on the Horvitz-Thompson estimator of a population total. Smith (1987) demonstrates diagnostics based on case deletion and a form of the influence function. Zaslavsky, Schenker and Belin (2001), and Beaumont and Alavi (2004) use M-estimation based strategies to downweight the influential clusters or units. Chambers (1986), Gwet and Rivest (1992), Welsh and Ronchetti (1998), and Duchesne (1999) conduct research on outlier robust estimation techniques for totals.

A perennial question among analysts of survey data is whether to use the survey weights or not when fitting models. The collections edited by Skinner *et al.* (1989) and Chambers and Skinner (2003) discuss this issue at length. Binder and Roberts (2003, Chapter 3), Chambers, Dorfman and Sverchkov (2003, Sections 11.2.3, 11.6), Chambers and Skinner (2003, Chapter 1), Korn and Graubard (1999, Sections 4.3, 4.4), Pfeffermann (1996), and Smith (1989, Chapter 6) describe the arguments pro and con. The details can be quite mathematical and abstract but are summarized succinctly by Skinner (2003, Section 6.2.3).

We paraphrase Skinner (2003, Section 6.2.3) here in the context of fitting a linear model to predict some response Y based on a set of explanatory variables **X**. If the linear model is specified correctly and the sampling depends only on the explanatory variables in the model, then unweighted regression parameter estimates will be unbiased in a model-based sense. In particular, the assumed conditions require that the survey weights are unrelated to Y conditional on the values of the **X** predictors. However, if sampling depends on factors that may be related to Y, even after conditioning on the values of the predictors, the unweighted parameter

[1] Jianzhu Li, Westat, 1650 Research Boulevard, Rockville MD 20850; Richard Valliant, Survey Research Center, University of Michigan, and Joint Program in Survey Methodology, University of Maryland, 1218 LeFrak Hall, College Park, MD 20742.

estimators will be biased both with respect to the true model and in the design-based, repeated sampling sense. This situation is known as having an *informative* sample design in which the distribution of the sample values of Y is different from the population distribution. An example of this is given by Chambers, Dorfman and Sverchkov (2003, Section 11.2.3). If sample units are selected with probabilities proportional to some measure x of their size and Y is related to x, the sample distribution of Y will be skewed to the right of its population distribution. The situation in this example is similar to the one in our empirical study in section 5.

Using the survey weights guards against the bias that may result from not accounting for an informative sample. Also, if the model is not correctly specified, the survey-weighted regression still estimates a census parameter. That is, the weighted estimates are approximately unbiased for the best-fitting linear model that would be obtained if the entire finite population were in hand. In this paper, we assume that an analyst has made the decision to use weights in fitting a model, possibly for the reasons above, and provide one type of diagnostic for assessing the effects of certain data points.

The hat matrix and leverages we present are the same ones that are produced by standard software packages when a weighted least squares regression is done. However, the literature is missing any discussion of their use and interpretation in the context of survey-weighted regression. Korn and Graubard (1999) is one of the few references that addresses any kind of diagnostics for models fitted from survey data. Leverages are among a series of diagnostic tools and will be more effective when evaluated with residuals. Many diagnostic statistics, such as the famous Cook's distance (Cook 1977) turn out to have both leverages and residuals as components.

The literature gives somewhat ambiguous guidance on how to deal with the influential observations once they are identified. An obvious, and perhaps naïve, solution is to remove the outliers and refit the model, which makes sense when the outliers result from improperly recorded data. A natural extension of this would be to devise an automatic approach where certain rules would be used to identify influential points, delete them, and refit the model. Our presumption in this article is that, after identification of influential points and careful consideration of the reasons for the influence, an analyst will determine whether the points should be excluded from fitting. This is in contrast to setting up some procedure that would automatically exclude points based on some cutoff values.

The remainder of the paper is organized as follows. Section 2 describes the ordinary least squares hat matrix, leverages, and some of their properties. Sections 3 and 4 cover the survey-weighted hat matrix and leverages plus a decomposition that shows how points can have large leverages. The extensions to survey data apply to both single- and multi-stage designs. Section 5 gives a numerical example using a single-stage sample of mental health organizations. The last section summarizes our findings and gives some directions for additional research.

2. OLS hat matrix

A *working* model is one that is being provisionally considered by an analyst for the structure that best describes a conceptual superpopulation. It may be revised after further assessment by adding predictors, dropping predictors, or making other changes to the form of the model. Suppose that the working linear model is

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \qquad V(\boldsymbol{\varepsilon}) = \sigma^2 \mathbf{I} \tag{1}$$

where $\mathbf{Y} = (Y_1, ..., Y_n)^T$, $\mathbf{X}^T = (\mathbf{x}_1, ..., \mathbf{x}_n)$ with $\mathbf{x}_i = (x_{i1}, ..., x_{ip})^T$, $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)^T$, and $\boldsymbol{\varepsilon} = (\varepsilon_1, ..., \varepsilon_n)^T$. Assuming the $\mathbf{X}$ matrix is of full rank, the ordinary least squares (OLS) estimate of $\boldsymbol{\beta}$ is

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y} = \mathbf{A}^{-1} \mathbf{X}^T \mathbf{Y}, \tag{2}$$

where $\mathbf{A} = \mathbf{X}^T \mathbf{X}$ is a square matrix and invertible. The fitted values $\hat{\mathbf{Y}}$ corresponding to the observed values $\mathbf{Y}$ are

$$\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}\mathbf{A}^{-1} \mathbf{X}^T \mathbf{Y} = \mathbf{H}\mathbf{Y},$$

where $\mathbf{H} = \mathbf{X}\mathbf{A}^{-1}\mathbf{X}^T$ is called the hat matrix. This name was first introduced by Tukey (Belsley, Kuh and Welsch 1980, Chapter 2; Hoaglin and Welsch 1978). The leverage, $h_{ii} = \mathbf{x}_i^T \mathbf{A}^{-1} \mathbf{x}_i$, is the i^{th} element on the diagonal of the hat matrix, which measures the impact of Y_i on its own fitted value since $\hat{Y}_i = \sum_j h_{ij} Y_j = h_{ii} Y_i + \sum_{j \neq i} h_{ij} Y_j$. If h_{ii} approaches 1, Y_i has a crucial role in determining the value of $\hat{Y}_i$.

The OLS hat matrix and leverages have many special and useful properties:

(i) $\mathbf{H}$ is symmetric, or $h_{ij} = h_{ji}$;

(ii) $\mathbf{H}$ is idempotent, or $\mathbf{H} = \mathbf{H}^2$, or $\mathbf{H}(\mathbf{I} - \mathbf{H}) = \mathbf{0}$;

(iii) $\mathbf{HX} = \mathbf{X}$ or $(\mathbf{I} - \mathbf{H})\mathbf{X} = \mathbf{0}$;

(iv) $0 \le h_{ii} \le 1$;

(v) $\sum_i h_{ii} = \text{rank}(\mathbf{X}) = p$, which implies that the mean leverage is $\bar{h} = p/n$;

if model (1) has an intercept, the following two properties hold:

(vi) $\sum_i h_{ij} = 1$;

(vii) $h_{ii} = 1/n + (\mathbf{x}_i - \bar{\mathbf{x}})^T \mathbf{A}^{-1}(\mathbf{x}_i - \bar{\mathbf{x}})$, where $\bar{\mathbf{x}} = \sum_n \mathbf{x}_i / n$.

In a reasonably large data set, an individual leverage value h_{ii} is usually considered extreme if it is more than twice the mean, $\bar{h} = p/n$ (Belsley *et al.* 1980, Chapter 2). The existence of a gap between most of the cases and a few unusual cases in the empirical distribution of the leverages also provides evidence of outlying units.

3. Survey weighted hat matrix

The initial step in the pseudo maximum likelihood approach is to form the set of estimating equations that would be appropriate for a model if the entire finite population were observed. This set is a type of population total which is then estimated using design-based survey methods. Suppose that the underlying structural model is a fixed-effects linear model:

$$Y_i = \mathbf{x}_i^T \boldsymbol{\beta} + \varepsilon_i, \quad \varepsilon_i \sim \text{ind}\, N(0, v_i \sigma^2) \qquad (3)$$

where ε_i is independently normally distributed with mean 0 and variance $v_i \sigma^2$, which is known except for the constant σ^2. The pseudo maximum likelihood estimator (PMLE) of $\boldsymbol{\beta}$ is the solution to the set of estimating equations $\mathbf{X}^T \mathbf{W}\mathbf{V}^{-1}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) = 0$, with $\mathbf{V} = \text{diag}(v_1, ..., v_n)$ and $\mathbf{W} = \text{diag}(w_1, ..., w_n)$. Survey weights, which in probability samples are usually inversely proportional to inclusion probabilities, are used in the PMLE to account for an informative design in which the sample distribution of the Y's is likely to differ from that of the finite population. These equations can be solved explicitly as $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T \mathbf{W}\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T \mathbf{W}\mathbf{V}^{-1}\mathbf{Y}$. If we assume $\mathbf{V} = \mathbf{I}$, model (3) reduces to (1) and the survey-weighted (SW) estimator $\hat{\boldsymbol{\beta}}$ will consequently take the form of a weighted least squares estimator, $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T \mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T \mathbf{W}\mathbf{Y}$.

When survey weights are accounted for in the regression, the predicted values become $\hat{\mathbf{Y}} = \mathbf{H}\mathbf{Y}$, where the hat matrix includes the survey weights and is defined as

$$\mathbf{H} = \mathbf{X}(\mathbf{X}^T \mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T \mathbf{W} = \mathbf{X}\mathbf{A}^{-1}\mathbf{X}^T \mathbf{W}$$

with $\mathbf{A} = \mathbf{X}^T \mathbf{W}\mathbf{X}$. The leverages on the diagonal of the hat matrix are $h_{ii} = \mathbf{x}_i^T \mathbf{A}^{-1}\mathbf{x}_i w_i$. In this formulation, it is assumed that the analyst does not incorporate a $\mathbf{V}$ matrix in the regression. However, results can be modified to incorporate $\mathbf{V}$ simply by using $\mathbf{W}^* = \mathbf{W}\mathbf{V}^{-1}$ rather than $\mathbf{W}$. Unlike the unweighted hat matrix, the SW hat matrix is no longer symmetric for sampling designs with unequal selection probabilities (or, more generally, unequal weights). Properties (ii) – (vi) in section 2 still hold (*e.g.*, see Valliant, Dorfman and Royall 2000, Chapter 5) provided the unweighted hat matrices were replaced by the weighted ones. In addition, the SW hat matrix has extra useful, and easily verified, properties as follows:

a) $\mathbf{W}\mathbf{H} = \mathbf{W}\mathbf{X}\mathbf{A}^{-1}\mathbf{X}^T \mathbf{W} = \mathbf{H}^T \mathbf{W}$;

b) $\mathbf{X}^T \mathbf{W}(\mathbf{I} - \mathbf{H}) = \mathbf{X}^T \mathbf{W} - \mathbf{X}^T \mathbf{H}^T \mathbf{W} = \mathbf{0}$;

c) $w_{i'} h_{i'i} = w_{i'} \mathbf{x}_{i'}^T \mathbf{A}^{-1}\mathbf{x}_i w_i = w_i h_{ii'}$.

The definition of the weighted leverages indicates that a large leverage may be caused by outlying $\mathbf{X}$ values, an outlying weight, or both. Note that the formulas for the survey-weighted hat matrix and leverages apply regardless of whether the sample design uses strata or is single-stage or multi-stage. This is in contrast to diagnostics, like Cook's D, that require estimated standard errors or covariance matrices that should be specialized to fit the sample design.

4. Decomposition of leverages

Leverages can be decomposed into components that separate the effect of the weight and the $\mathbf{X}$ values for a unit. Suppose the working model is (1) and that the model contains an intercept, so that

$$\mathbf{X} = \begin{pmatrix} 1 & \mathbf{x}_1^T \\ \vdots & \vdots \\ 1 & \mathbf{x}_n^T \end{pmatrix} \equiv (\mathbf{1} \;\; \mathbf{X}_1), \text{ and } \mathbf{X}_1 = \begin{pmatrix} \mathbf{x}_1^T \\ \vdots \\ \mathbf{x}_n^T \end{pmatrix},$$

where $\mathbf{x}_i^T = (x_{i1}, \ldots, x_{i,p-1})$ are $1 \times (p-1)$ vectors, $\mathbf{1}$ is a $n \times 1$ vector with all the elements equal to 1, and $\mathbf{X}_1$ is a $n \times (p-1)$ matrix. The $\mathbf{A}$ matrix is computed as

$$\mathbf{A} = \begin{pmatrix} \mathbf{1}^T \\ \mathbf{X}_1^T \end{pmatrix} \mathbf{W}(\mathbf{1}\,\mathbf{X}_1) = \begin{pmatrix} \mathbf{1}^T \mathbf{W}\mathbf{1} & \mathbf{1}^T \mathbf{W}\mathbf{X}_1 \\ \mathbf{X}_1^T \mathbf{W}\mathbf{1} & \mathbf{X}_1^T \mathbf{W}\mathbf{X}_1 \end{pmatrix} \equiv \begin{pmatrix} \hat{N} & \hat{\mathbf{t}}_X^T \\ \hat{\mathbf{t}}_X & \mathbf{A}_1 \end{pmatrix},$$

where $\hat{\mathbf{t}}_X$ is a $(p-1) \times 1$ vector with elements $\hat{\mathbf{t}}_{Xj} = \sum_{i \in s} w_i x_{ij}$ and $\mathbf{A}_1$ is a $(p-1) \times (p-1)$ matrix. Using the inverse of a partitioned matrix,

$$\mathbf{A}^{-1} = \begin{pmatrix} \frac{1}{\hat{N}} + \frac{1}{\hat{N}} \hat{\mathbf{t}}_X^T \mathbf{S}^{-1} \hat{\mathbf{t}}_X \frac{1}{\hat{N}} & -\frac{1}{\hat{N}} \hat{\mathbf{t}}_X^T \mathbf{S}^{-1} \\ -\frac{1}{\hat{N}} \mathbf{S}^{-1} \hat{\mathbf{t}}_X & \mathbf{S}^{-1} \end{pmatrix}$$

$$= \begin{pmatrix} \frac{1}{\hat{N}} + \bar{\mathbf{x}}_W^T \mathbf{S}^{-1} \bar{\mathbf{x}}_W & -\bar{\mathbf{x}}_W^T \mathbf{S}^{-1} \\ -\mathbf{S}^{-1} \bar{\mathbf{x}}_W & \mathbf{S}^{-1} \end{pmatrix}$$

$$= \begin{pmatrix} \frac{1}{\hat{N}} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} + \begin{pmatrix} -\bar{\mathbf{x}}_W^T \\ \mathbf{I} \end{pmatrix} \mathbf{S}^{-1}(-\bar{\mathbf{x}}_W \;\; \mathbf{I})$$

where $\bar{\mathbf{x}}_W = \hat{\mathbf{t}}_X / \hat{N}$ is a $(p-1)\times 1$ vector, and $\mathbf{S} = \mathbf{A}_1 - \hat{\mathbf{t}}_X \hat{\mathbf{t}}_X^T / \hat{N}$ is a $(p-1)\times(p-1)$ matrix. Simplifying the hat matrix using the above inverse matrix, we obtain

$$\begin{aligned}\mathbf{H} &= \mathbf{X}\mathbf{A}^{-1}\mathbf{X}^T\,\mathbf{W}\\ &= \left\{\frac{1}{\hat{N}}\mathbf{1}\mathbf{1}^T + (\mathbf{X}_1 - \mathbf{1}\bar{\mathbf{x}}_W^T)\,\mathbf{S}^{-1}(-\bar{\mathbf{x}}_W\mathbf{1}^T + \mathbf{X}_1^T)\right\}\mathbf{W}\\ &= \left\{\frac{1}{\hat{N}}\mathbf{1}\mathbf{1}^T + \begin{pmatrix}\mathbf{x}_1^T - \bar{\mathbf{x}}_W^T\\ \vdots\\ \mathbf{x}_n^T - \bar{\mathbf{x}}_W^T\end{pmatrix}\mathbf{S}^{-1}(\mathbf{x}_1 - \bar{\mathbf{x}}_W, \ldots, \mathbf{x}_n - \bar{\mathbf{x}}_W)\right\}\mathbf{W}.\end{aligned}$$

Then, using the fact that $\hat{N} = n\bar{w}$ with $\bar{w} = \sum_{i=1}^n w_i / n$, the leverage of i^{th} observation, or the i^{th} diagonal element of the weighted hat matrix $\mathbf{H}$, is

$$h_{ii} = \frac{1}{n}\frac{w_i}{\bar{w}}[1 + \hat{N}(\mathbf{x}_i - \bar{\mathbf{x}}_W)^T\mathbf{S}^{-1}(\mathbf{x}_i - \bar{\mathbf{x}}_W)].$$

The quadratic form, $(\mathbf{x}_i - \bar{\mathbf{x}}_W)^T\,\mathbf{S}^{-1}\,(\mathbf{x}_i - \bar{\mathbf{x}}_W)$, defines an ellipsoid centered at $\bar{\mathbf{x}}_W$ (*e.g.*, see Weisberg 2005, Chapter 8), and $\hat{N}(\mathbf{x}_i - \bar{\mathbf{x}}_W)^T\mathbf{S}^{-1}(\mathbf{x}_i - \bar{\mathbf{x}}_W)$ is the Mahalanobis distance from $\mathbf{x}_i$ to $\bar{\mathbf{x}}_W$. Consequently, a leverage can be large if (1) w_i is large, especially relative to the average weight $\bar{w}$; or (2) $\mathbf{x}_i$ is far from the weighted average, $\bar{\mathbf{x}}_W$, of the $\mathbf{X}$, in the metric determined by the matrix $\mathbf{S}$.

For example, in a simple linear model with only one auxiliary variable, $y_i = \alpha + \beta x_i + \varepsilon_i$, $\varepsilon_i \sim (0, \sigma^2)$, the leverage of the i^{th} observation is

$$h_{ii}^W = \frac{1}{n}\frac{w_i}{\bar{w}}\left[1 + \hat{N}\frac{(x_i - \bar{x}_W)^2}{\sum_{j=1}^n w_j\,(x_j - \bar{x}_W)^2}\right].$$

where $\bar{x}_W = \sum_i w_i\,x_i / \hat{N}$.

If the error terms in the model have a general variance structure $\boldsymbol{\varepsilon} \sim (0, \mathbf{V})$ and $\mathbf{V}$ is known, the hat matrix is then defined as $\mathbf{H} = \mathbf{X}\mathbf{A}^{-1}\mathbf{X}^T\mathbf{W}\mathbf{V}^{-1}$ with

$$\begin{aligned}\mathbf{A} &= \begin{pmatrix}\mathbf{1}^T\mathbf{W}\mathbf{V}^{-1}\mathbf{1} & \mathbf{1}^T\mathbf{W}\mathbf{V}^{-1}\mathbf{X}_1\\ \mathbf{X}_1^T\mathbf{V}^{-1}\mathbf{W}\mathbf{1} & \mathbf{X}_1^T\mathbf{W}\mathbf{V}^{-1}\mathbf{X}_1\end{pmatrix}\\ &= \begin{pmatrix}\sum_s w_i / v_i & \sum_s w_i\mathbf{x}_i^T / v_i\\ \sum_s w_i\mathbf{x}_i / v_i & \sum_s w_i\mathbf{x}_i\mathbf{x}_i^T / v_i\end{pmatrix}.\end{aligned}$$

A formula for $\mathbf{A}^{-1}$ like the one above applies with $\hat{\mathbf{t}}_{XV} = \sum_s w_i\,\mathbf{x}_i / v_i$, $\hat{N}_V = \sum_s w_i / v_i$, and $\mathbf{S}_V = \mathbf{X}_1^T\mathbf{W}\mathbf{V}^{-1}\mathbf{X}_1 - \hat{\mathbf{t}}_{XV}\hat{\mathbf{t}}_{XV}^T / \hat{N}_V$. If a general $\mathbf{V}$ is used, $\hat{\mathbf{t}}_{XV}$ and $\hat{N}_V$ no longer are design-based estimates of $\mathbf{T}_X$ and N but are estimates of $\mathbf{T}_{XV} = \sum_1^N \mathbf{x}_i / v_i$ and $N_V = \sum_1^N 1/v_i$. The leverage of the i^{th} observation under this general model is

$$h_{ii} = \frac{w_i}{v_i\,\hat{N}_V}\,[1 + \hat{N}_V\,(\mathbf{x}_i - \bar{\mathbf{x}}_{WV})^T\,\mathbf{S}_V^{-1}\,(\mathbf{x}_i - \bar{\mathbf{x}}_{WV})].$$

5. Numerical example

As noted in section 1, arguments can be advanced to justify ignoring sample design features, generally, and weights, in particular, when fitting models. Roughly speaking, when a model conditions on all the design variables determining the sampling scheme and the model is correct for both the population and the sample, OLS regression can be used. Analysts may object to including design variables in a model because some are not scientifically interesting as predictors. In addition, conditioning on all design variables may not be possible, especially when the "sampling scheme" includes uncontrolled nonresponse that itself may be related to the response variable. As noted in section 1, SW provides a modicum of protection against having a misspecified model when the distribution of the sample Y's is different from that of the population due to the type of sample design used. Nevertheless, some analysts will contend that the sample design and survey weights can be ignored in specific applications and that OLS is appropriate. Thus, it is interesting to see how different the OLS diagnostics are from SW diagnostics in a real application. However, given a course of action, an analyst should use diagnostics consistent with the method of fitting. If OLS is used, the standard OLS diagnostics should be examined; if SW regression is used, SW diagnostics are appropriate. It may well be that different points are influential depending on whether one uses OLS or SW regression.

In this section we examine the hat matrix and leverages in a regression example using the 1998 Survey of Mental Health Organizations (SMHO) conducted in the U.S., which collected data on specialty mental health care organizations and general hospital mental health care services. The sample for this survey was based on a stratified single-stage design with probability proportional to size (PPS) sampling (Manderscheid and Henderson 2002; Choudhry 2000). The measure of size (MOS) used in sampling was the number of "episodes", defined as the number of patients/clients of an organization at the beginning of 1998 plus the number of new patients/clients added during calendar year 1998. Many of the analysis variables in the survey are related to the MOS, and their unweighted sample distributions will be different from the population distributions since the sample tends to have larger size units. Thus, this design is potentially informative as defined in Chambers and Skinner (2003).

The varying sizes of the mental health care organizations resulted in the values of collected variables in the sample having wide ranges, which may cause some observations to have relatively large influence on the parameter estimates of a linear regression. The model of interest in this study is to regress the total expenditure of a health organization, in 1,000's of dollars, on the number of beds set up and staffed for use and the number of additions of patients or clients during the reporting year. The SW estimator, $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\mathbf{Y}$, was used. Mimicking the procedure employed by most analysts, we did not incorporate a model variance matrix $\mathbf{V}$ in the estimate of the regression parameter. A total of 875 observations was used in the regression, each of which had non-missing values on the independent and dependent variables.

Table 1 gives a summary of the quantile values of the variables involved in the regression, including the survey weights. The total expenditure has a maximum of 519,863.3, which is almost 30,000 times the minimum, 16.6. Although not as extreme as the total expenditure, the number of beds and the number of additions also have significant differences between their maxima and minima. Because the sample was selected using a PPS design, the sample weights were associated with the sizes of the mental health organizations, with a range from 1 to 158.86. The weights we use in analysis include a nonresponse adjustment which was done separately by design stratum. In some cases, units that were selected with certainty in the initial sample did not respond and some of the responding certainties had their weights adjusted to be larger than 1. A total of 157 organizations had a weight of 1 after the nonresponse adjustment.

Table 1
Quantiles of variables in SMHO regression

	Quantiles				
Variables	**0%**	**25%**	**50%**	**75%**	**100%**
Expenditure (1,000's)	16.6	2,932.5	6,240.5	11,842.6	519,863.3
# of Beds	0	6.5	36	93	2,405
# of Additions	0	558.5	1,410	2,406	79,808
Weights	1	1.42	2.48	7.76	158.86

In the regressions that follow, we have included the units with weights of 1 in standard error estimation rather than excluding them, as would be the approach for handling certainties in purely design-based estimation. Including the certainties is consistent with the idea that a superpopulation model is being estimated and that slope coefficients would still have a variance even if a census were done. A sketch of the mathematical justification for doing this is model-dependent (not design-based) and is given in the Appendix.

Figure 1 shows scatterplots of expenditures versus beds and additions for the sample of 875 facilities (omitting one extremely large facility described below). In the first row, points are highlighted whose OLS leverage is greater than $2p/n = 0.007$. The second row shows bubbleplots with the relative size of the bubbles proportional to the weight of each case. High SW leverage points are highlighted using the same cutoff of 0.007. The distributions of the predictors are quite skewed as noted in Table 1. There is also one very large facility that is not shown in Figure 1 because it distorts the scale of the plot. That facility (denoted as observation 818 here) has (expenditures in 1,000's; beds; additions) = ($519,863.3; 2,405; 79,808) and has a survey weight of 2.22. (Observation 818 was one of the cases noted earlier that was a certainty in the initial sample but received a nonresponse adjustment, and, thus, had a final weight larger than 1.) Because its data values are far out of line with those of the other organizations, this point has the potential to affect estimates.

Table 2 reports the twenty observations with the largest SW leverages. The values of the leverages range from 0.022 to 0.389, substantially greater than the level of the rough rule of thumb 0.007. This table also shows, for these twenty cases, the OLS unweighted leverages, the ratio of individual sample weight to average sample weight and the relative absolute distance between individual X values and their weighted means. We note that unit 818 has the highest weighted and unweighted leverages, mainly resulting from its extremely large number of beds and number of additions. Since this case has a less-than-average sample weight, the OLS leverage is even larger than the weighted one. There are other similar cases such as units 271, 179, 820, 157, 163, 156, and 154, which are associated with either extreme number of beds, or extreme number of additions, or both – but have small weights. Another type of outlier results from extreme sample weights, even if the values of their auxiliary variables are not very distinct from others. Units 672, 613, 711, 801, and 611 all have sample weights more than 15 times the average weight. Their weighted leverages are identified as large, whereas the unweighted leverages are not. There is also a noticeable gap between the weighted leverages for case 331 ($h_{ii} = 0.075$) and for case 271 ($h_{ii} = 0.046$).

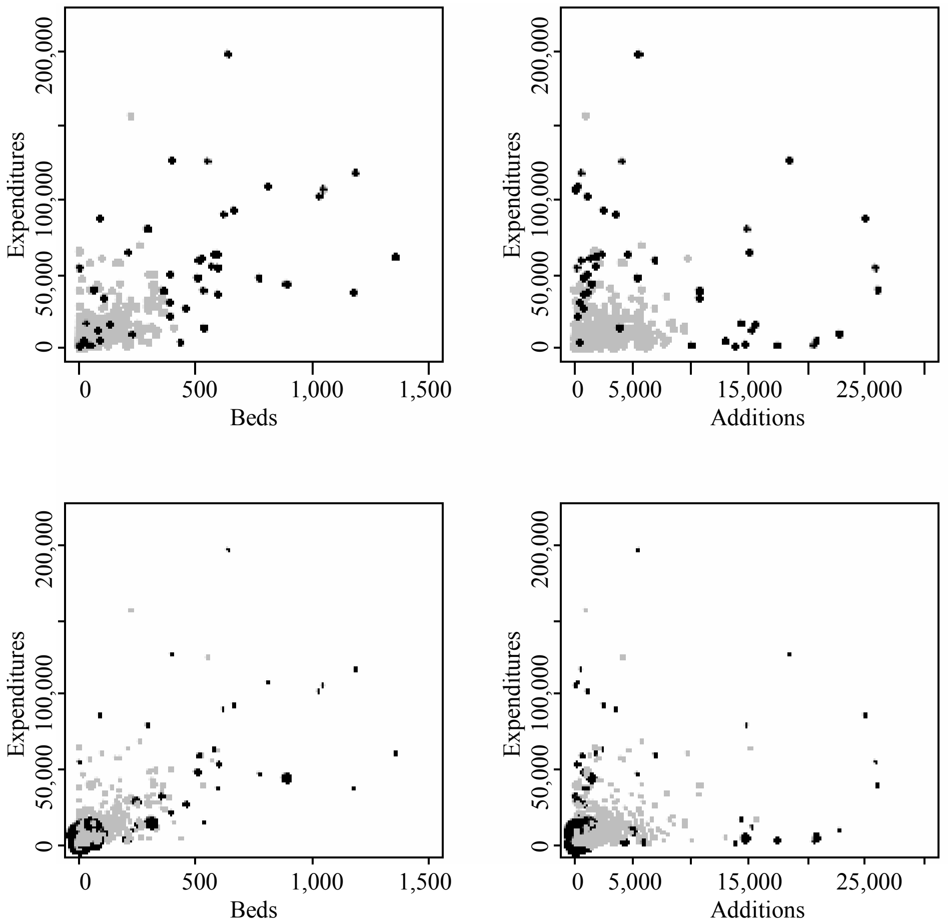


Figure 1 **Scatterplots of expenditures versus beds and additions. High leverage points based on OLS (SW) are highlighted in top (bottom) row**

Table 2
Observations with 20 largest survey weighted leverages

			Weights	**Beds**	**Additions**
Obs ID	**OLS h_{ii}**	**Weighted h_{ii}**	$w_i / \bar{w}$	$\lvert x_{1i} - \bar{x}_1^W \rvert / \bar{x}_1^W$	$\lvert x_{2i} - \bar{x}_2^W \rvert / \bar{x}_2^W$
818	0.513	0.389	0.3	49.3	64.7
189	0.037	0.245	3.4	17.7	0.3
346	0.035	0.157	2.2	0.6	16.1
366	0.017	0.105	3.0	0.7	11.1
331	0.024	0.075	1.5	0.1	13.4
271	0.068	0.046	0.4	23.7	0.0
830	0.004	0.045	5.8	5.4	0.1
628	0.056	0.045	0.4	1.0	20.3
179	0.089	0.038	0.2	27.4	0.5
672	0.002	0.034	24.2	1.0	0.8
820	0.048	0.034	0.3	0.8	19.6
207	0.012	0.030	1.3	9.5	0.3
157	0.069	0.030	0.2	23.8	0.5
163	0.017	0.027	0.8	11.4	0.8
613	0.002	0.026	18.5	1.0	0.7
711	0.002	0.024	16.8	1.0	0.9
801	0.002	0.024	17.5	0.6	0.9
156	0.055	0.023	0.2	20.9	0.9
611	0.002	0.023	15.9	1.0	0.8
154	0.051	0.022	0.2	20.5	0.1
			$\bar{w} = 6.57$	$\bar{x}_1^W = 47.83$	$\bar{x}_2^W = 1{,}214.13$

Note: observation ID is the line number of an observation in the sample.

Sizes of the sample weights can make analysts reach different conclusions when they use weighted or unweighted leverages to identify potentially influential observations. Figure 2 shows a scatterplot of weighted leverages versus unweighted ones. The two reference lines were drawn at values of 0.007. Observation 818 is omitted since it would again distort the scale of the graph. Clearly, the high leverage points identified by the SW method only, located in area A, have significantly larger weights than the points in area B, which are identified by the OLS method only.

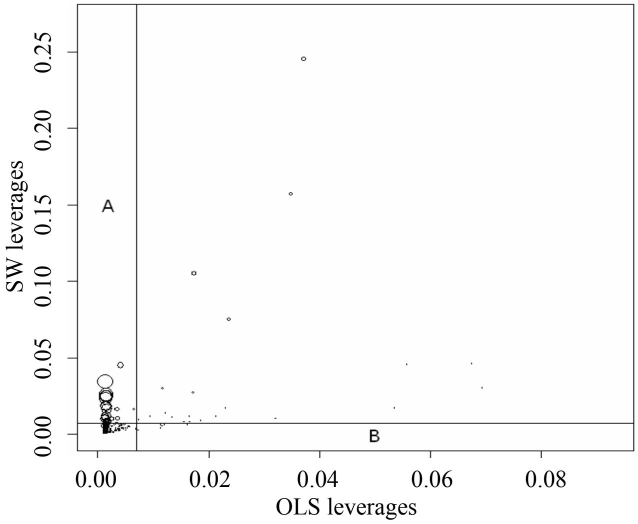


Figure 2 Plot of survey weighted leverages versus OLS unweighted leverages

Given that some potentially influential cases have been identified, the next step is to see what effect they have on parameter estimates. Table 3 shows the OLS and SW parameter estimates using all cases. Table 4 lists the OLS and SW estimates (i) omitting high leverage cases and (ii) omitting observation 818. High leverage points are those with $h_{ii} > 0.007$. However, note that different sets of points are high leverage in OLS and SW regressions. The standard errors are estimated via the usual OLS formula and the sandwich estimator (Binder 1983) for the SW estimates.

Comparing Tables 3 and 4, we see that the OLS estimates change substantially after high leverage points are deleted (section (i) of Table 4). The OLS intercept, which is significant in both tables, jumps from negative to positive. The OLS slope for beds drops by about 26% (94.16 to 69.27) when the high leverage points are dropped. The decrease is about 59% for the slope for additions. The SW estimates for beds and additions are also sensitive to the high leverage points with the slopes decreasing by 7% and 46% respectively. In all cases, the slopes are significant so that the qualitative conclusion that expenditures is related to beds and additions holds with or without the high leverage points. However, predicted values will be quite different before and after omitting these points.

The standard errors (SE's) also decrease substantially when the high leverage points are omitted. For example, the SW standard error for beds drops from 13.14 to 6.75 (a 49% reduction); the SE for additions drops from 0.76 to 0.21 (a 72% reduction). This is due to some points with extreme weights being removed in the SW regression. In contrast, the SE's for the OLS estimates actually increase when the OLS high leverage points are omitted because the sample variance of the x's decreases. This is another illustration of the considerable differences that can occur when applying the same type of diagnostic to OLS and SW regressions.

Table 3
OLS and SW parameter estimates of SMHO regression using all 875 sample cases

Independent	**OLS Estimation**			**SW Estimation**		
Variables	**Coefficient**	**SE**	***t***	**Coefficient**	**SE**	***t***
Intercept	-1,201.73	526.19	-2.28	514.08	1,157.71	0.44
# of Beds	94.16	3.03	31.08	81.23	13.14	6.18
# of Additions	2.31	0.13	18.50	1.84	0.76	2.43

Table 4
OLS and SW parameter estimates after from SMHO regression

Independent	**OLS Estimation**			**SW Estimation**		
Variables	**Coefficient**	**SE**	***t***	**Coefficient**	**SE**	***t***
(i) Deleting observations with leverages greater than 0.007						
Intercept	2,987.55	490.54	6.09	1,993.86	353.71	5.64
# of Beds	69.27	4.35	15.94	75.82	6.75	11.23
# of Additions	0.95	0.20	4.71	1.00	0.21	4.73
(ii) Deleting observation 818						
Intercept	1,979.51	537.93	3.68	2,281.17	460.35	4.96
# of Beds	81.80	2.92	27.98	68.69	8.04	8.54
# of Additions	1.19	0.14	8.41	0.79	0.29	2.75

Because point 818 is so obviously extreme, we also fitted the regression after dropping only that observation. The results are shown in section (ii) of Table 4. Omitting that single point causes noticeable changes in both OLS and SW parameter estimates. This also illustrates that a single point can affect the standard errors for estimated slopes in a survey-weighted regression, as is also the case in OLS. Observation 818 has a large residual (see Figure 3); omitting it results in the SE for Beds dropping from 13.14 in Table 3 to 8.04 in Table 4. Note that if unit 818 had a large weight, then its residual would likely be smaller since it would have more affect on the fit. If so, the SE could actually be smaller when unit 818 is included.

Another point to be gleaned from Tables 3 and 4 is that the OLS and SW estimates are much closer to each other after the high leverage points are dropped than they are before. As shown in Table 5, the OLS estimates are 16 and 26% larger than the SW estimates with all points but are 9 and 5% less than SW after dropping points.

Table 5
Ratios of OLS and SW parameter estimates before and after deleting observations with leverages greater than 0.007 from SMHO regression

	Ratio of OLS to SW estimates	
	With all points	Dropping high leverage points
Beds	1.16	0.91
Additions	1.26	0.95

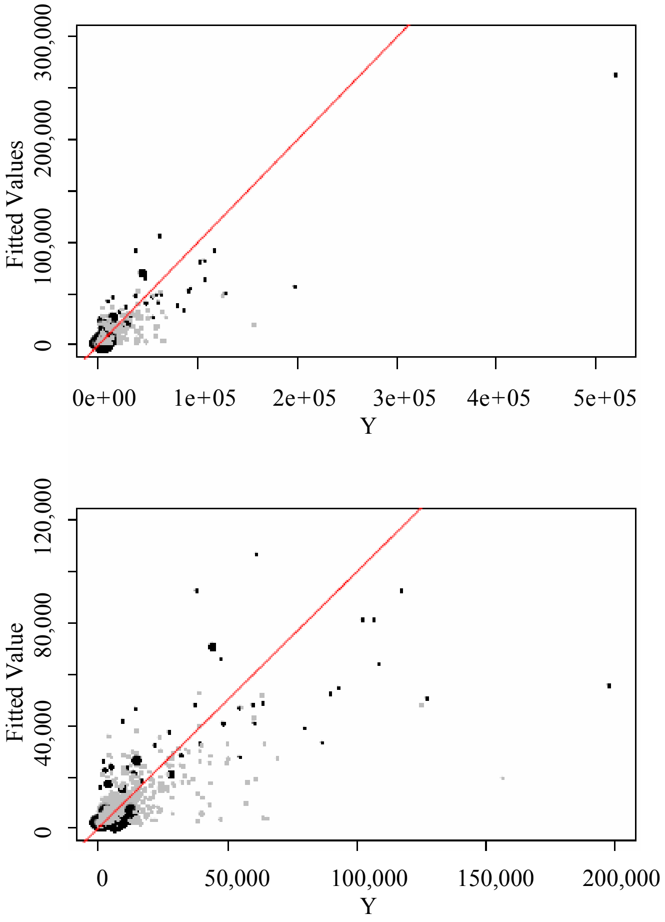


Figure 3 Plot of fitted values versus Y values. Reference line is drawn at $Y = \hat{Y}$. The upper panel includes all points. The lower panel omits the extreme observation 818. High leverage points based on SW are solid, dark circles in each panel

Leverages are usually combined with residuals to determine which points are influential in fitting the regression model because residuals can be used to detect discrepant Y values. A scatterplot of fitted values from the SW regression versus the Y values is shown in Figure 3. The high leverage points are labeled as dark solid circles. The vertical distances from the points to the 45 degree line imply the sizes of the residuals. The upper panel includes all 875 sample points; the lower panel omits observation 818 to provide better resolution for the remaining points. Note that some observations have high leverages and small residuals, while others have low leverages and large residuals. The influence of these points on the regression can be further investigated using various tools that we will not cover here. For example, Cook's distance, implicitly involving the leverage and residual, is designed to measure the effect of deleting a single observation on the overall parameter estimates. The adaptation of some basic OLS diagnostic statistics to survey data, such as DFBETAS and DFFITS, has been discussed under a single stage sampling design in Li and Valliant (2006).

6. Conclusion

Leverages and residuals are essential components of diagnostic statistics intended to identify substantial influence of a single observation or a group of observations on a fitted linear model. Survey data sets can contain influential observations whether one argues that the sample design is ignorable and ordinary least squares can be used, or that the design must be accounted for and survey weights used. The points that are influential in the two cases are not necessarily the same, as illustrated here.

Once high leverage points are identified, an important question is how to deal with them for inference. Two options are to down-weight them or drop them from model-fitting entirely. Down-weighting seems unsatisfactory in general since a point can have a high leverage not because of a large weight but rather due to having one or more unusual X's. Down-weighting may be sensible from a model-based point-of-view, assuming the model itself is correctly specified. However, the design-based idea of estimating a census parameter may then be lost. If a point has a large leverage because of extreme X's, then it may not follow the model at all and should be dropped.

However, using a mechanical procedure that automatically drops many influential observations with high leverages can lead to standard error estimates that are too small, resulting in confidence intervals that cover at less than the nominal rates and in inflated Type I error rates in hypothesis tests (Li 2007). This phenomenon is similar to well-known problems in stepwise regression (Hurvich and

Tsai 1990, Zhang 1992). Thus, a useful research topic appears to be developing inferential procedures for constructing confidence intervals and conducting hypothesis tests that account for the effects of dropping or down-weighting points.

For complex survey data, the hat matrix involves no design features except for sample weights and can be used to identify cases that have atypical weights or predictor values. Other diagnostic statistics, like Cook's D, do contain variance estimates that need to account for complex sample design features such as stratification and clustering. The adaptation and extension of additional diagnostic approaches for survey analysis will be explored in the future.

7. Acknowledgement

This material is based upon work 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. We thank the editor and referees for their thoughtful comments, which improved the paper considerably.

Appendix

Inclusion of certainties in standard error estimation

In the empirical study in section 5, we included certainty units in the standard error calculations. The justification for doing this is sketched here. Under the general model (3), the model variance of $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\mathbf{Y}$, the estimator used in the empirical study, is $\mathrm{var}_M(\hat{\boldsymbol{\beta}}) = \mathbf{A}^{-1}\mathbf{X}^T\mathbf{W}\mathbf{V}\mathbf{W}\mathbf{X}\mathbf{A}^{-1}\sigma^2$ where $\mathbf{A} = \mathbf{X}^T\mathbf{W}\mathbf{X}$ and $\mathbf{V} = \mathrm{diag}(v_i)_{i\in s}$. The sandwich variance estimator used in the study reported in section 5 is defined as

$$v(\hat{\boldsymbol{\beta}}) = \mathbf{A}^{-1}\frac{n}{n-1}\sum_{i\in s}(\mathbf{z}_i - \bar{\mathbf{z}})(\mathbf{z}_i - \bar{\mathbf{z}})^T\mathbf{A}^{-1} \qquad (4)$$

where $\mathbf{z}_i = w_i e_i \mathbf{x}_i$ with $e_i = Y_i - \mathbf{x}_i\hat{\boldsymbol{\beta}}$ and $\bar{\mathbf{z}} = \sum_{i\in s} w_i e_i \mathbf{x}_i/n$. This estimator is design consistent (see Binder 1983) in single-stage sampling if units are sampled with replacement with probabilities equal to w_i^{-1}, and there are no certainty units. If the sample contains certainties, the formula for $v(\hat{\boldsymbol{\beta}})$ would be modified to estimate the design-based variance: certainties would be excluded from the sums in (4) and $\bar{\mathbf{z}}$, and n would be changed to n_{nc}, the number of non-certainties. In the extreme case of a census, the design-based variance estimator would reduce to zero.

The estimator in (4) is approximately model-unbiased under (3) regardless of whether the sample contains certainties or not. The middle matrix in (4) can be expanded as $\sum_{i\in s}(\mathbf{z}_i - \bar{\mathbf{z}})(\mathbf{z}_i - \bar{\mathbf{z}})^T = \sum_{i\in s}\mathbf{z}_i\mathbf{z}_i^T - n\bar{\mathbf{z}}\bar{\mathbf{z}}^T$. Assuming that $e_i \approx Y_i - \mathbf{x}_i^T\boldsymbol{\beta}$, the model expectation under (3) of the first term is $E_M(\sum_{i\in s}\mathbf{z}_i\mathbf{z}_i^T) = \mathbf{X}^T\mathbf{W}\mathbf{V}\mathbf{W}\mathbf{X}\sigma^2$ while $E_M(n\bar{\mathbf{z}}\bar{\mathbf{z}}^T) = n^{-1}\mathbf{X}^T\mathbf{W}\mathbf{V}\mathbf{W}\mathbf{X}\sigma^2$. Substituting these expectations gives $E_M[v(\hat{\boldsymbol{\beta}})] = \mathrm{var}_M(\hat{\boldsymbol{\beta}})$, which holds even when some units are certainties. This also shows that $v(\hat{\boldsymbol{\beta}})$ is robust in the sense of properly reflecting the contribution of heteroscedastic variances in (3) to the model-variance of $\hat{\boldsymbol{\beta}}$ even though $\mathbf{V}$ may be unknown and not accounted for in the estimation of $\boldsymbol{\beta}$.

References

Beaumont, J.-F., and Alavi, A. (2004). Robust Generalized Regression Estimation. *Survey Methodology*, 30, 195-208.

Belsley, D.A., Kuh, E. and Welsch, R. (1980). *Regression Diagnostics*: *Identifying Influential Data and Sources of Collinearity*. New York: John Wiley & Sons, Inc.

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

Binder, D.A., and Roberts, G.R. (2003). Design-based and model-based methods for estimating model parameters, Chapter 3 in *Analysis of Survey Data*, (Eds. R. Chambers and C. Skinner). New York: John Wiley & Sons, Inc.

Chambers, R.L. (1986). Outlier robust finite population estimation. *Journal of the American Statistical Association*, 81, 1063-1069.

Chambers, R.L., Dorfman, A.H. and Sverchkov, M.Y. (2003). Nonparametric regression with complex survey data, Chapter 11 in *Analysis of Survey Data*, (Eds. R. Chambers and C. Skinner). New York: John Wiley & Sons, Inc.

Chambers, R.L., and Skinner, C.J. (2003). *Analysis of Survey Data*. New York: John Wiley & Sons, Inc.

Choudhry, G. (2000). The 1998 Survey of Mental Health Organizations Survey Design. Westat technical report prepared for Center for Mental Health Services, Substance Abuse and Mental Health Services Administration (SAMHSA), available by request to SAMHSA.

Cook, R.D. (1977). Detection of influential observation in linear regression. *Technometrics*, 19, 15-18.

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

Duchesne, P. (1999). Robust calibration estimators. *Survey Methodology*, 25, 43-56.

Gwet, J., and Rivest, L. (1992). Outlier resistant alternatives to the ratio estimator. *Journal of the American Statistical Association*, 87, 1174-1182.

Hoaglin, D.C., and Welsch, R.E. (1978). The hat matrix in regression and ANOVA (Corr: 78V32 p146). *The American Statistician*, 32, 17-22.

Hulliger, B. (1995). Outlier robust Horvitz-Thompson estimators. *Survey Methodology*, 21, 79-87.

Hurvich, C.M., and Tsai, C. (1990). The impact of model selection on inference in linear regression. *The American Statistician*, 44, 214-217.

Korn, E.L., and Graubard, B.I. (1999). *Analysis of Health Surveys*. New York: John Wiley & Sons, Inc.

Li, J. (2007). *Regression Diagnostics for Complex Survey Data*: *Identification of Influential Observations*. Unpublished doctoral dissertation, University of Maryland.

Li, J., and Valliant, R. (2006). Influence analysis in linear regression with sampling weights. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 3330-3337.

Manderscheid, R.W., and Henderson, M.J. (2002). Mental Health, United States, 2002. DHHS Publication No. SMA04-3938. Rockville MD USA: Substance Abuse and Mental Health Services Administration. available at http://mentalhealth.samhsa.gov/publications/allpubs/SMA04-3938/AppendixA.asp

Moreno-Rebollo, J.L., Muñoz-Reyes, A. and Muñoz-Pichardo, J. (1999). Influence diagnostic in survey sampling: Conditional bias. *Biometrika*, 86, 923-928.

Pfeffermann, D. (1996). The use of sampling weights for survey data analysis. *Statistical Methods in Medical Research*, 5, 239-261.

Potter, F.J. (1990). A study of procedures to identify and trim extreme sampling weights. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 225-230.

Potter, F.J. (1993). The effect of weight trimming on nonlinear survey estimates. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 758-763.

Skinner, C.J. (2003). Introduction to Part B, Chapter 6 in *Analysis of Survey Data*, (Eds. R. Chambers and C. Skinner). New York: John Wiley & Sons, Inc.

Skinner, C.J., Holt, D. and Smith, T.M.F. (eds.) (1989). *Analysis of Complex Surveys*. New York: John Wiley & Sons, Inc.

Smith, T.M.F. (1987). Influential observations in survey sampling. *Journal of Applied Statistics*, 14 , 143-152.

Smith, T.M.F. (1989). Introduction to Part B, Chapter 6 in *Analysis of Complex Surveys*, (Eds. C.J. Skinner, D. Holt and T.M.F. Smith). New York: John Wiley & Sons, Inc.

Valliant, R., Dorfman, A.H. and Royall, R.M. (2000). *Finite Population Sampling and Inference*: *A Prediction Approach*. New York: John Wiley & Sons, Inc.

Weisberg, S. (2005). *Applied Linear Regression*, Third Edition. New York: John Wiley & Sons, Inc.

Welsh, A.H., and Ronchetti, E. (1998). Bias-calibrated estimation from sample surveys containing outliers. *Journal of the Royal Statistical Society*, Series B, Methodological, 60, 413-428.

Zaslavsky, A.M., Schenker, N. and Belin, T.R. (2001). Downweighting influential clusters in surveys: Application to the 1990 post enumeration Survey. *Journal of the American Statistical Association*, 96 , 858-869.

Zhang, P. (1992). Influence after variable selection in linear regression models. *Biometrika*, 79, 741-746.

Survey Methodology, June 2009
Vol. 35, No. 1, pp. 25-35
Statistics Canada, Catalogue No. 12-001-X

A practical bootstrap method for testing hypotheses from survey data

Jean-François Beaumont and Cynthia Bocci [1]

Abstract

The bootstrap technique is becoming more and more popular in sample surveys conducted by national statistical agencies. In most of its implementations, several sets of bootstrap weights accompany the survey microdata file given to analysts. So far, the use of the technique in practice seems to have been mostly limited to variance estimation problems. In this paper, we propose a bootstrap methodology for testing hypotheses about a vector of unknown model parameters when the sample has been drawn from a finite population. The probability sampling design used to select the sample may be informative or not. Our method uses model-based test statistics that incorporate the survey weights. Such statistics are usually easily obtained using classical software packages. We approximate the distribution under the null hypothesis of these weighted model-based statistics by using bootstrap weights. An advantage of our bootstrap method over existing methods of hypothesis testing with survey data is that, once sets of bootstrap weights are provided to analysts, it is very easy to apply even when no specialized software dealing with complex surveys is available. Also, our simulation results suggest that, overall, it performs similarly to the Rao-Scott procedure and better than the Wald and Bonferroni procedures when testing hypotheses about a vector of linear regression model parameters.

Key Words: Bootstrap weights; Analysis of survey data; Hypothesis testing; Informative sampling; Linear regression; Model parameters.

1. Introduction

The bootstrap technique is becoming more and more popular in sample surveys conducted by national statistical agencies. The main reasons seem to be that it can easily deal with several situations that would be difficult to handle otherwise (*e.g.*, nonresponse weight adjustment, calibration, non-smooth statistics, *etc.*) and that it is convenient for analysts. In most of its implementations, several sets of bootstrap weights accompany the survey microdata file given to analysts; no other design information is provided. These weights are usually obtained by assuming that the first-stage sampling fractions are small enough that a without-replacement sampling design can be accurately approximated by a with-replacement sampling design. The reader is referred to Rao, Wu and Yue (1992) for a succinct but clear description of a method to construct bootstrap weights under this assumption when a stratified multistage sampling design has been used.

So far, the use of the technique in practice seems to have been mostly limited to variance estimation problems (*e.g.*, Langlet, Faucher and Lesage 2003; Yeo, Mantel, and Liu 1999; and Hughes and Brodsky 1994). On the research side, efforts have been mainly oriented towards finding an appropriate bootstrap methodology for variance estimation when the sample is drawn without replacement from a finite population (see Sitter 1992; or Shao and Tu 1995, Chapter 6, for a review of methods). Some authors have also studied the problem of determining bootstrap confidence intervals for a finite population parameter (*e.g.*, Rao and Wu 1988; Kovar, Rao and Wu 1988; Sitter 1992; and Rao *et al.* 1992). To our knowledge, there does not seem to be any literature on hypothesis testing using the bootstrap technique in survey sampling although this problem has been studied in the context of classical statistics. The reader is referred to Hall and Wilson (1991) for a discussion on bootstrap tests of hypotheses and to Efron and Tibshirani (1993) for an excellent account of the bootstrap technique in classical statistics. It is worth noting the work of Graubard, Korn and Midthune (1997) who applied the classical parametric bootstrap method to survey data in order to test the fit of a logistic regression model. Their procedure is valid when sampling is not informative.

The problem of hypothesis testing from complex survey data has been well studied in the last 30 years (*e.g.*, Rao and Scott 1981; Fay 1985; Thomas and Rao 1987; Korn and Graubard 1990; Korn and Graubard 1991; Graubard and Korn 1993; Thomas, Singh and Roberts 1996; and Rao and Thomas 2003). However, except perhaps for estimating unknown variances/covariances involved in these methods, the bootstrap technique has apparently not yet been considered for testing hypotheses. The goal of this paper is thus to propose a bootstrap methodology for testing hypotheses about a vector of unknown model parameters when the sample has been drawn from a finite population. The probability sampling design used to select the sample may be informative or not. Informally speaking, sampling is informative when the model that holds for the selected

[1] Jean-François Beaumont, Statistics Canada, Statistical Research and Innovation Division, Tunney's Pasture, R.H. Coats building, 16th floor, Ottawa, Ontario, Canada, K1A 0T6. E-mail: Jean-Francois.Beaumont@statcan.gc.ca; Cynthia Bocci, Statistics Canada, Business Survey Methods Division, Tunney's Pasture, R.H. Coats building, 11th floor, Ottawa, Ontario, Canada, K1A 0T6. E-mail: Cynthia.Bocci@statcan.gc.ca.

sample is different from the model that holds for the whole population; otherwise sampling is not informative.

Our method uses model-based test statistics that incorporate the survey weights. Such statistics are usually easily obtained using classical software packages. We approximate the distribution under the null hypothesis of these weighted model-based statistics by using bootstrap weights. An advantage of our bootstrap method over existing methods of hypothesis testing with survey data is that, once sets of bootstrap weights are provided to analysts, it is very easy to apply even when no specialized software dealing with complex surveys is available.

We introduce notation and the problem in section 2. In section 3, we describe and justify our proposed bootstrap methodology for testing hypotheses with survey data. A linear regression example is given in section 4 to illustrate the theory. We briefly describe the alternative Rao-Scott (Rao and Scott 1981), Wald and Bonferroni procedures in section 5 when testing hypotheses about a vector of linear regression model parameters. They are evaluated in section 6 and compared to our proposed bootstrap procedure through a simulation study. Finally, we conclude in the last section with a short summary and discussion.

2. Preliminaries

We assume that a finite population U of size N has been generated according to a model, specified by the analyst, that describes the conditional distribution $F(\mathbf{y}_U \mid \mathbf{X}_U;\ \boldsymbol{\beta}, \boldsymbol{\theta})$. The N-vector $\mathbf{y}_U$ contains the population values of a dependent variable y, $\mathbf{X}_U$ is an N-row matrix that contains the population values of a vector of independent variables $\mathbf{x}$, $\boldsymbol{\beta}$ is an r-vector of unknown model parameters and $\boldsymbol{\theta}$ is a potential vector of additional unknown model parameters. We are interested in testing hypotheses about $\boldsymbol{\beta}$ but not $\boldsymbol{\theta}$. We also assume that, if the entire population U could be observed, a test statistic $t(U;\ \mathbf{c})$ would be used to test the multiple linear hypothesis $\mathrm{H}_0 : \mathbf{H}\boldsymbol{\beta} = \mathbf{c}$ against the alternative hypothesis H_1: $\mathbf{H}\boldsymbol{\beta} \neq \mathbf{c}$. The $Q \times r$ matrix $\mathbf{H}$ is used to define the hypothesis to be tested and $\mathbf{c}$ is a Q-vector of constants specified by the analyst. Ideally, $t(U;\ \mathbf{c})$ is asymptotically pivotal; *i.e.*, it has an asymptotic distribution that does not depend on any unknown parameter. We consider statistics that have the following quadratic form:

$$t(U;\ \mathbf{c}) = (\mathbf{H}\hat{\boldsymbol{\beta}}_U - \mathbf{c})' \{\mathbf{A}(U)\}^{-1} (\mathbf{H}\hat{\boldsymbol{\beta}}_U - \mathbf{c}), \qquad (2.1)$$

where $\hat{\boldsymbol{\beta}}_U$ is a consistent estimator of $\boldsymbol{\beta}$ under the model and $\mathbf{A}(U)$ is some scaling matrix. Typically, $\mathbf{A}(U)$ is symmetric and positive definite.

As an illustrative example, let us assume that y_k, for all population units $k \in U$, are independently and identically distributed random variables with mean β and variance θ and that we are interested in testing the null hypothesis H_0: $\beta = c$. In this example, $Q = 1$, $r = 1$, $\mathbf{H} = 1$ and $\mathbf{X}_U = \mathbf{1}_U$, where $\mathbf{1}_U$ is a population vector of one's. A common test statistic for this problem is

$$t(U;\ \mathbf{c}) = \frac{(\hat{\beta}_U - c)^2}{\hat{\theta}_U / N}, \qquad (2.2)$$

where $\hat{\beta}_U = \sum_{k \in U} y_k / N$ and $\hat{\theta}_U = \sum_{k \in U} (y_k - \hat{\beta}_U)^2/(N-1)$. The statistic (2.2) has the same form as (2.1) if we let $A(U) = \hat{\theta}_U/N$. This statistic is usually assumed to follow the distribution χ_1^2 or $F_{1,\ N-1}$ under the null hypothesis.

As is typically the case, a random sample s of size n is selected from the finite population U according to a given probability sampling design $p(s)$. Since the dependent variable y and, possibly, the independent variables $\mathbf{x}$ are not observed for nonsample units, we may want to use the statistic $t(s; \mathbf{c})$ instead of $t(U; \mathbf{c})$. In the above example, this would lead to $t(s; \mathbf{c}) = n(\hat{\beta}_s - c)^2 / \hat{\theta}_s$, where $\hat{\beta}_s = \sum_{k \in s} y_k / n$ and $\hat{\theta}_s = \sum_{k \in s} (y_k - \hat{\beta}_s)^2/(n-1)$. However, if sampling is informative with respect to the model, it may be more appropriate and is undoubtedly more common to use a weighted test statistic of the form

$$\hat{t}(s,\ \mathbf{w}_s;\ \mathbf{c}) = (\mathbf{H}\hat{\boldsymbol{\beta}}_{ws} - \mathbf{c})' \{\hat{\mathbf{A}}(s,\ \mathbf{w}_s)\}^{-1} (\mathbf{H}\hat{\boldsymbol{\beta}}_{ws} - \mathbf{c}). \quad (2.3)$$

The n-vector $\mathbf{w}_s$ contains the survey weight of sample unit k in its k^{th} element, denoted by w_k, $\hat{\boldsymbol{\beta}}_{ws}$ is a weighted estimator for $\boldsymbol{\beta}$ and $\hat{\mathbf{A}}(s,\ \mathbf{w}_s)$ is a weighted analogue to $\mathbf{A}(s)$ in that each sample unit k is weighted by its survey weight w_k whereas there is no weighting with $\mathbf{A}(s)$. We thus have $\hat{\mathbf{A}}(s,\ \mathbf{1}_s) = \mathbf{A}(s)$, where $\mathbf{1}_s$ is a sample vector of one's. As a result, the statistic $\hat{t}(s,\ \mathbf{w}_s;\ \mathbf{c})$ is also a weighted analogue to $t(s;\ \mathbf{c})$ and we have $\hat{t}(s,\ \mathbf{1}_s;\ \mathbf{c}) = t(s;\ \mathbf{c})$. If the statistic $t(s;\ \mathbf{c})$ can be computed using some classical software package, not necessarily developed to handle survey data, the statistic $\hat{t}(s,\ \mathbf{w}_s;\ \mathbf{c})$ can also be computed using the same software package provided that it can allow each observation to be weighted by its survey weight.

Typically, the survey weight w_k, for a unit $k \in s$, is equal to the inverse of its selection probability, which may then be calibrated to account for known external information (*e.g.*, Deville and Särndal 1992). We assume that the sampling design and the survey weights are constructed so that the following two assumptions hold:

Assumption 1: $\sqrt{n}\,(\hat{\boldsymbol{\beta}}_{ws} - \boldsymbol{\beta}) \xrightarrow{mp} N(\mathbf{0}, \boldsymbol{\Sigma})$, where $\xrightarrow{mp}$ denotes convergence in distribution under the model and the sampling design, and $\boldsymbol{\Sigma}$ is the asymptotic variance-covariance matrix of $\sqrt{n}\,\hat{\boldsymbol{\beta}}_{ws}$ under the model and the sampling design. The notation "m" stands for the model while the notation "p" stands for the probability sampling design.

Assumption 2: $n\hat{\mathbf{A}}(s, \mathbf{w}_s)$ is symmetric, positive definite and *mp*-consistent for some fixed symmetric positive definite scaling matrix $\tilde{\mathbf{A}}$.

Note that assumption 2 does not require $\hat{\mathbf{A}}(s, \mathbf{w}_s)$ to be *p*-consistent for $\mathbf{A}(U)$. Indeed, $N\mathbf{A}(U)$ will be typically *m*-consistent for $\tilde{\mathbf{A}}$. Other choices could replace the weighted scaling matrix $\hat{\mathbf{A}}(s, \mathbf{w}_s)$ in (2.3). For instance, it could be replaced by an estimator of the design variance of $\mathbf{H}\hat{\boldsymbol{\beta}}_{ws}$ under simple random sampling (*e.g.*, Rao and Scott 1981). An alternative choice is the common Wald statistic. It is obtained by replacing $\hat{\mathbf{A}}(s, \mathbf{w}_s)$ in (2.3) by $\hat{\mathbf{V}}_{mp}(\mathbf{H}\hat{\boldsymbol{\beta}}_{ws})$, which is an *mp*-consistent estimator of $\mathbf{V}_{mp}(\mathbf{H}\hat{\boldsymbol{\beta}}_{ws})$; the variance of $\mathbf{H}\hat{\boldsymbol{\beta}}_{ws}$ evaluated with respect to the model and the sampling design. As pointed out in the paragraph below (2.3), an advantage of using a scaling matrix $\hat{\mathbf{A}}(s, \mathbf{w}_s)$ such that $\hat{\mathbf{A}}(s, \mathbf{1}_s) = \mathbf{A}(s)$ is that the resulting test statistic $\hat{t}(s, \mathbf{w}_s; \mathbf{c})$ can then be directly computed using classical software packages provided that they allow each observation to be weighted by its survey weight. It is thus more convenient for the users of survey data.

Continuing the above example, we may define our weighted test statistic as

$$\hat{t}(s, \mathbf{w}_s; \mathbf{c}) = \frac{(\hat{\beta}_{ws} - c)^2}{\{(\hat{N} - 1)/(n - 1)\}(\hat{\theta}_{ws}/\hat{N})}, \qquad (2.4)$$

where $\hat{N} = \sum_{k\in s} w_k$, $\hat{\beta}_{ws} = \sum_{k\in s} w_k y_k / \sum_{k\in s} w_k$ and $\hat{\theta}_{ws} = \sum_{k\in s} w_k (y_k - \hat{\beta}_{ws})^2/(\hat{N} - 1)$. In (2.4), the underlying weighted scaling matrix is $\hat{A}(s, \mathbf{w}_s) = \{(\hat{N} - 1)/(n - 1)\}(\hat{\theta}_{ws}/\hat{N})$, which does not depend on the way the weights are scaled. If they are rescaled so that $\sum_{k\in s} w_k = n$, which is typically done by analysts, then the factor $(\hat{N} - 1)/(n - 1)$ vanishes. The role of this factor, along with other regularity conditions, is to satisfy assumption 2. If the SAS® System is chosen, the test statistic (2.4) is obtained by using the WEIGHT statement in standard procedures. When the null hypothesis is true, it is well known that (2.4) unfortunately does not follow the distribution χ_1^2 or $F_{1,n-1}$ under the model and the sampling design.

To obtain a valid test procedure, we need to approximate the distribution of $\hat{t}(s, \mathbf{w}_s; \mathbf{c})$ under the null hypothesis. This can be achieved by using the following result:

Result 1: $\hat{t}(s, \mathbf{w}_s; \mathbf{H}\boldsymbol{\beta}) \xrightarrow{mp} \sum_{q=1}^{Q} \lambda_q \Omega_q$, where λ_q, for $q = 1, ..., Q$, are the eigenvalues of $\boldsymbol{\Lambda} = (\tilde{\mathbf{A}}^{-1})(\mathbf{H}\boldsymbol{\Sigma}\mathbf{H}')$ and Ω_q are independent chi-square random variables with one degree of freedom.

The proof of result 1 uses assumptions 1 and 2 and is given in the appendix. When the null hypothesis is true (*i.e.*, $\mathbf{H}\boldsymbol{\beta} = \mathbf{c}$), we thus have

$$\hat{t}(s, \mathbf{w}_s; \mathbf{c}) \xrightarrow{mp} \sum_{q=1}^{Q} \lambda_q \Omega_q. \qquad (2.5)$$

Rao and Scott (1981) used a similar result to construct their test procedures. They approximated a distribution like (2.5) by a scaled chi-square distribution that matches the estimated first two moments of the right-hand side of (2.5). Instead, we approximate the distribution of $\hat{t}(s, \mathbf{w}_s; \mathbf{c})$ under the null hypothesis by using bootstrap weights. This is described in the next section.

Before giving details of our test procedure, it is useful to note that $\hat{t}(s, \mathbf{w}_s; \mathbf{c})$ in (2.3) can be written as

$$\begin{aligned}\hat{t}(s, \mathbf{w}_s; \mathbf{c}) = {} & \hat{t}(s, \mathbf{w}_s; \mathbf{H}\boldsymbol{\beta}) \\ & + 2(\mathbf{H}\hat{\boldsymbol{\beta}}_{ws} - \mathbf{H}\boldsymbol{\beta})'\{\hat{\mathbf{A}}(s, \mathbf{w}_s)\}^{-1}(\mathbf{H}\boldsymbol{\beta} - \mathbf{c}) \\ & + (\mathbf{H}\boldsymbol{\beta} - \mathbf{c})'\{\hat{\mathbf{A}}(s, \mathbf{w}_s)\}^{-1}(\mathbf{H}\boldsymbol{\beta} - \mathbf{c}). \qquad (2.6)\end{aligned}$$

Under the null hypothesis, the last two terms on the right-hand side of (2.6) vanish and we have $\hat{t}(s, \mathbf{w}_s; \mathbf{c}) = \hat{t}(s, \mathbf{w}_s; \mathbf{H}\boldsymbol{\beta})$. When the null hypothesis is false, the third term on the right-hand side of (2.6) dominates the others as the sample size increases since the first, second and third terms are $O_p(1)$, $O_p(\sqrt{n})$ and $O_p(n)$ respectively, provided that assumptions 1 and 2 hold. Also, since $\hat{\mathbf{A}}(s, \mathbf{w}_s)$ is positive definite, the third term is always positive. Therefore, a large positive observed value of $\hat{t}(s, \mathbf{w}_s; \mathbf{c})$ compared to a large percentile of the distribution of $\hat{t}(s, \mathbf{w}_s; \mathbf{H}\boldsymbol{\beta})$ is an indication that the null hypothesis may be wrong.

3. The proposed bootstrap method

Let w_k^* denote a random bootstrap weight for unit *k*, obtained using some bootstrap procedure such as that of Rao *et al.* (1992), and let $\mathbf{w}_s^*$ be the *n*-vector that contains the random bootstrap weight w_k^* in its k^{th} element. The bootstrap estimator $\hat{\boldsymbol{\beta}}_{w^*s}$ is obtained similarly to $\hat{\boldsymbol{\beta}}_{ws}$ by replacing the survey weight w_k by its bootstrap version w_k^* for each sample unit. We also denote by $\mathbf{w}_s^{*b}$, for $b = 1, ..., B$, the B *n*-vectors containing the bootstrap weights w_k^{*b} in their k^{th} element. These B vectors are drawn independently and have the same distribution as $\mathbf{w}_s^*$; this distribution is called the bootstrap distribution and is denoted by the symbol '*'. The b^{th} bootstrap estimator $\hat{\boldsymbol{\beta}}_{w^{*b}s}$ is defined in an obvious manner.

Before describing our bootstrap test procedure, we first introduce three additional assumptions related to the construction of the bootstrap weights:

Assumption 3: $\sqrt{n}(\hat{\boldsymbol{\beta}}_{w^*s} - \hat{\boldsymbol{\beta}}_{ws}) \xrightarrow{*} N(\mathbf{0}, \hat{\boldsymbol{\Sigma}})$, where $\xrightarrow{*}$ denotes convergence in bootstrap distribution and

$\hat{\boldsymbol{\Sigma}}$ is the asymptotic bootstrap variance-covariance matrix of $\sqrt{n}\,\hat{\boldsymbol{\beta}}_{w^*s}$.

Assumption 4: $n\hat{\mathbf{A}}(s,\ \mathbf{w}_s^*)$ is *-consistent for $n\hat{\mathbf{A}}(s,\ \mathbf{w}_s)$.

Assumption 5: $\hat{\boldsymbol{\Sigma}}$ is *mp*-consistent for $\boldsymbol{\Sigma}$.

Assumptions 3 and 4 are bootstrap analogues to assumptions 1 and 2 and should be satisfied with most bootstrap methods (*e.g.*, those described in the review paper by Sitter 1992) and models (*e.g.*, linear regression model, logistic regression model, *etc*.). The reader is referred to Shao and Tu (1995, Chapter 6; in particular section 6.4.4) for greater detail.

A comment is in order about assumption 5. This assumption is equivalent to requiring that the bootstrap variance $\mathbf{V}_*(\hat{\boldsymbol{\beta}}_{w^*s})$ be *mp*-consistent for

$$\mathbf{V}_{mp}(\hat{\boldsymbol{\beta}}_{ws}) = \mathbf{E}_m\mathbf{V}_p(\hat{\boldsymbol{\beta}}_{ws}) + \mathbf{V}_m\mathbf{E}_p(\hat{\boldsymbol{\beta}}_{ws}). \qquad (3.1)$$

This means that the bootstrap distribution must reflect the variability due to both the model and the sampling design. Unfortunately, standard design-based bootstrap methods reflect only the variability due to the sampling design so that they only track the first term of the right-hand side of (3.1). Thus, these bootstrap methods do not satisfy assumption 5 in general. However, when the overall sampling fraction n/N is negligible, the second term of the right-hand side of (3.1) becomes negligible (*e.g.*, see Binder and Roberts 2003) so that the approximation $\mathbf{V}_{mp}(\hat{\boldsymbol{\beta}}_{ws}) \approx \mathbf{E}_m\mathbf{V}_p(\hat{\boldsymbol{\beta}}_{ws})$ is appropriate and design-based bootstrap methods can be used. In many household surveys, the overall sampling fraction is actually quite small. Indeed, bootstrap weights are often obtained under the assumption that the first-stage sampling fractions are small (*e.g.*, Rao *et al.* 1992). Developing bootstrap procedures that capture both terms of (3.1) is an area for future research.

Under assumptions 3 and 4, we obtain our second result:

Result 2: $\hat{t}(s,\ \mathbf{w}_s^*;\ \mathbf{H}\hat{\boldsymbol{\beta}}_{ws}) \xrightarrow{\ *\ } \sum_{q=1}^{Q}\hat{\lambda}_q\Omega_q$, where $\hat{\lambda}_q$, for $q = 1, \ldots, Q$, are the eigenvalues of $\hat{\boldsymbol{\Lambda}} = [n\hat{\mathbf{A}}(s,\ \mathbf{w}_s)]^{-1}(\mathbf{H}\hat{\boldsymbol{\Sigma}}\mathbf{H}')$ and Ω_q are again independent chi-square random variables with one degree of freedom.

The proof of result 2 is omitted as it is very similar to the proof of result 1 given in the appendix. From assumptions 2 and 5, $\hat{\boldsymbol{\Lambda}}$ is *mp*-consistent for $\boldsymbol{\Lambda}$. Thus, using results 1 and 2, the bootstrap distribution of $\hat{t}(s,\ \mathbf{w}_s^*;\ \mathbf{H}\hat{\boldsymbol{\beta}}_{ws})$ is asymptotically the same as the *mp*-distribution of $\hat{t}(s,\ \mathbf{w}_s;\ \mathbf{H}\boldsymbol{\beta})$, which is itself the same as the *mp*-distribution of $\hat{t}(s,\ \mathbf{w}_s;\ \mathbf{c})$ under the null hypothesis; the distribution that we want to approximate. This suggests the following bootstrap test procedure:

i) Obtain bootstrap weights, w_k^{*b}, for $k \in s$ and $b = 1, \ldots, B$.

ii) Compute $\hat{t}(s,\ \mathbf{w}_s^{*b};\ \mathbf{H}\hat{\boldsymbol{\beta}}_{ws})$, for $b = 1, \ldots, B$.

iii) Since a large value of $\hat{t}(s,\ \mathbf{w}_s;\ \mathbf{c})$ leads to rejecting the null hypothesis, compute the observed significance level (p-value) as

$$\frac{\#\{\hat{t}(s,\ \mathbf{w}_s^{*b};\ \mathbf{H}\hat{\boldsymbol{\beta}}_{ws}) > \hat{t}(s,\ \mathbf{w}_s;\ \mathbf{c})\}}{B}.$$

The null hypothesis is rejected if this value is lower than the significance level α (*e.g.*, 5%).

Note that the statistic to be bootstrapped is $\hat{t}(s,\ \mathbf{w}_s^{*b};\ \mathbf{H}\hat{\boldsymbol{\beta}}_{ws})$ and not $\hat{t}(s,\ \mathbf{w}_s^{*b};\ \mathbf{c})$. The use of the latter would not properly reflect the distribution under the null hypothesis and would thus violate the first guideline in Hall and Wilson (1991).

If $\hat{t}(s,\ \mathbf{w}_s;\ \mathbf{c})$ is pivotal then the second guideline of Hall and Wilson (1991) is also satisfied. The fact that $t(U;\ \mathbf{c})$ is asymptotically pivotal certainly helps in obtaining a better bootstrap test procedure. However, it does not unfortunately guarantee that $\hat{t}(s,\ \mathbf{w}_s;\ \mathbf{c})$ is also asymptotically pivotal, particularly when sampling is informative. Nevertheless, failure to use a pivotal statistic does not invalidate the above test procedure and may not reduce its power. But, it may reduce the level accuracy of the test. As pointed out by Hall and Wilson (1991), it is sometimes appropriate to disregard the second guideline. The main advantage of using the simple (possibly non-pivotal) statistic $\hat{t}(s,\ \mathbf{w}_s;\ \mathbf{c})$ in (2.3) and the bootstrap statistic $\hat{t}(s,\ \mathbf{w}_s^{*b};\ \mathbf{H}\hat{\boldsymbol{\beta}}_{ws})$ is that, once bootstrap weights have been provided on the microdata file, these statistics are easily obtained using classical software packages that ignore sampling design features. Moreover, we show in section 5, through a simulation study, that our bootstrap test procedure performs similarly to the Rao-Scott procedure and better than the Wald and Bonferroni procedures.

4. A linear regression example

To better illustrate the theory in a practical context, let us now assume that, conditional on $\mathbf{X}_U$, the random variables y_k, for $k \in U$, are independently distributed with mean $\mathrm{E}_m(y_k \mid \mathbf{X}_U) = \mathbf{x}_k'\boldsymbol{\beta}$ and variance $\mathrm{V}_m(y_k \mid \mathbf{X}_U) = \theta$, where $\mathbf{x}_k$ is an r-vector of linearly independent variables for unit k. Recall that we are interested in testing the null hypothesis H_0: $\mathbf{H}\boldsymbol{\beta} = \mathbf{c}$ against the alternative hypothesis H_1: $\mathbf{H}\boldsymbol{\beta} \neq \mathbf{c}$. If the entire population could be observed, the common statistic

$t(U;\ \mathbf{c}) =$

$$\frac{(\mathbf{H}\hat{\boldsymbol{\beta}}_U - \mathbf{c})'\left(\mathbf{H}\left(\sum_{k\in U}\mathbf{x}_k\mathbf{x}_k'\right)^{-1}\mathbf{H}'\right)^{-1}(\mathbf{H}\hat{\boldsymbol{\beta}}_U - \mathbf{c})}{Q\,\hat{\theta}_U} \qquad (4.1)$$

could be used, where

$$\hat{\boldsymbol{\beta}}_U = \left(\sum\nolimits_{k\in U} \mathbf{x}_k \mathbf{x}_k'\right)^{-1} \sum\nolimits_{k\in U} \mathbf{x}_k y_k$$

and

$$\hat{\theta}_U = \frac{\sum_{k\in U} (y_k - \mathbf{x}_k'\hat{\boldsymbol{\beta}}_U)^2}{N - r}.$$

The statistic $t\,(U;\ \mathbf{c})$ in (4.1) follows the distribution $F_{Q,\,N-r}$ under the null hypothesis. It reduces to (2.2) when $Q = r = \mathbf{H} = \mathbf{x}_k = 1$ in (4.1).

A weighted sample version of (4.1), which can be written in the form of (2.3), is

$\hat{t}\,(s,\ \mathbf{w}_s;\ \mathbf{c}) =$

$$\frac{(\mathbf{H}\hat{\boldsymbol{\beta}}_{ws} - \mathbf{c})'\left(\mathbf{H}\left(\sum\nolimits_{k\in s} w_k \mathbf{x}_k \mathbf{x}_k'\right)^{-1} \mathbf{H}'\right)^{-1} (\mathbf{H}\hat{\boldsymbol{\beta}}_{ws} - \mathbf{c})}{Q\,\hat{\theta}_{ws}\,\{(\hat{N} - r)/(n - r)\}}, \tag{4.2}$$

where

$$\hat{\boldsymbol{\beta}}_{ws} = \left(\sum\nolimits_{k\in s} w_k \mathbf{x}_k \mathbf{x}_k'\right)^{-1} \sum\nolimits_{k\in s} w_k \mathbf{x}_k y_k \tag{4.3}$$

and

$$\hat{\theta}_{ws} = \frac{\sum_{k\in s} w_k (y_k - \mathbf{x}_k'\hat{\boldsymbol{\beta}}_{ws})^2}{\hat{N} - r}. \tag{4.4}$$

For instance, the statistic $\hat{t}\,(s,\ \mathbf{w}_s;\ \mathbf{c})$ in (4.2) could be obtained by using the WEIGHT statement in the procedure REG of SAS as long as $w_k > 0$, for $k \in s$. Note that it satisfies assumption 2 and does not depend on the way the weights are scaled. Again, if the weights are rescaled so that $\sum_{k\in s} w_k = n$, the factor $(\hat{N} - r)/(n - r)$ in (4.2) vanishes. The test statistic (4.2) reduces to (2.4) when $Q = r = \mathbf{H} = \mathbf{x}_k = 1$ in (4.2), (4.3) and (4.4). The bootstrap statistic $\hat{t}\,(s,\ \mathbf{w}_s^{*b};\ \mathbf{H}\hat{\boldsymbol{\beta}}_{ws})$ as well as $\hat{\boldsymbol{\beta}}_{w^{*b}s}$ and $\hat{\theta}_{w^{*b}s}$ are obtained similarly to $\hat{t}\,(s,\ \mathbf{w}_s;\ \mathbf{c})$, $\hat{\boldsymbol{\beta}}_{ws}$ and $\hat{\theta}_{ws}$ in (4.2), (4.3) and (4.4) respectively, except that w_k is replaced by w_k^{*b} and $\mathbf{c}$ is replaced by $\mathbf{H}\hat{\boldsymbol{\beta}}_{ws}$.

Remark 1: Note that w_k^{*b} is likely to be 0 for some units $k \in s$ (see, for example, Rao *et al.* 1992). In some software packages such as SAS, the number of observations used in the analysis of the b^{th} bootstrap replicate, n^{*b}, is equal to the number of units $k \in s$ for which $w_k^{*b} > 0$. Such software packages may use $n^{*b} - r$ instead of $n - r$ when computing the bootstrap statistic $\hat{t}\,(s,\ \mathbf{w}_s^{*b};\ \mathbf{H}\hat{\boldsymbol{\beta}}_{ws})$. One must thus make sure that $n - r$ is used and, if not, that the bootstrap statistic computed from these packages is properly adjusted before applying the proposed bootstrap test procedure. One way of avoiding this problem is to add a very small positive value (*e.g.*, 1×10^{-10}) to each bootstrap weight w_k^{*b}, for $k \in s$, so that no observation is excluded from the computation of $\hat{t}\,(s,\ \mathbf{w}_s^{*b};\mathbf{H}\hat{\boldsymbol{\beta}}_{ws})$.

Remark 2: Let us define the bootstrap statistic $\hat{t}_e(s,\ \mathbf{w}_s^{*b};\ \mathbf{0})$ by replacing y_k by $e_k = y_k - \mathbf{x}_k'\hat{\boldsymbol{\beta}}_{ws}$ in $\hat{t}\,(s,\ \mathbf{w}_s^{*b};\ \mathbf{0})$, for each $k \in s$. It is not difficult to show that $\hat{t}_e(s,\ \mathbf{w}_s^{*b};\ \mathbf{0}) = \hat{t}(s,\ \mathbf{w}_s^{*b};\ \mathbf{H}\hat{\boldsymbol{\beta}}_{ws})$ so that our bootstrap procedure can be implemented using either $\hat{t}_e(s,\ \mathbf{w}_s^{*b};\ \mathbf{0})$ or $\hat{t}\,(s, \mathbf{w}_s^{*b}; \mathbf{H}\hat{\boldsymbol{\beta}}_{ws})$ when a linear regression model is used. The former may sometimes be more convenient with some software packages. This was the case in our simulation study since the use of $\hat{t}_e(s,\ \mathbf{w}_s^{*b};\ \mathbf{0})$ allowed us to get rid of manually typing the values of $\mathbf{H}\hat{\boldsymbol{\beta}}_{ws}$ for each selected sample. An informal explanation for the equality $\hat{t}_e(s,\ \mathbf{w}_s^{*b};\ \mathbf{0}) = \hat{t}\,(s,\ \mathbf{w}_s^{*b};\ \mathbf{H}\hat{\boldsymbol{\beta}}_{ws})$ can be obtained by treating $\hat{\boldsymbol{\beta}}_{ws}$ as a fixed quantity, which is actually the case under the bootstrap distribution. The bootstrap statistic $\hat{t}\,(s,\ \mathbf{w}_s^{*b};\ \mathbf{H}\hat{\boldsymbol{\beta}}_{ws})$ can thus be interpreted as a statistic aiming at testing the null hypothesis H_0^*: $\mathbf{H}\boldsymbol{\beta} = \mathbf{H}\hat{\boldsymbol{\beta}}_{ws}$ or, alternatively, H_0^*: $\mathbf{H}\boldsymbol{\gamma} = \mathbf{0}$, where $\boldsymbol{\gamma} = \boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_{ws}$. Still assuming that $\hat{\boldsymbol{\beta}}_{ws}$ is fixed, we can rewrite our linear model $\mathrm{E}_m(y_k \,|\, \mathbf{X}_U) = \mathbf{x}_k'\boldsymbol{\beta}$ as $\mathrm{E}_m(e_k \,|\, \mathbf{X}_U) = \mathbf{x}_k'\boldsymbol{\gamma}$. These observations seem to imply that using the bootstrap statistic $\hat{t}_e(s,\ \mathbf{w}_s^{*b};\ \mathbf{0})$ is equivalent to using $\hat{t}\,(s,\ \mathbf{w}_s^{*b};\ \mathbf{H}\hat{\boldsymbol{\beta}}_{ws})$, which is indeed true.

Remark 3: We have already mentioned that the WEIGHT statement is necessary to obtain a weighted statistic if the proposed bootstrap test procedure is implemented using the procedure REG of SAS. Also, the TEST statement is necessary to request the desired statistics to be produced and the "ODS OUTPUT TESTANOVA =" statement to save these requested statistics in a SAS dataset specified by the user.

5. Some alternative procedures for linear regression

In this section, we briefly describe some test procedures in the context of linear regression exposed in section 4; namely, two naïve procedures that are sometimes used in practice as well as specific implementations of the Rao-Scott, Wald and Bonferroni procedures. They will all be evaluated in the simulation study in section 6.

The Bonferroni, Wald and Rao-Scott procedures, described in sections 5.2, 5.3 and 5.4 respectively, all need an *mp*-consistent estimator $\hat{\mathbf{V}}_{mp}(\hat{\boldsymbol{\beta}}_{ws})$ of $\mathbf{V}_{mp}(\hat{\boldsymbol{\beta}}_{ws})$. In the simulation study in section 6, we have used the bootstrap variance estimator

$$\hat{\mathbf{V}}_{mp}(\hat{\boldsymbol{\beta}}_{ws}) = \frac{\sum_{b=1}^{B} (\hat{\boldsymbol{\beta}}_{w^{*b}s} - \hat{\boldsymbol{\beta}}_{ws})(\hat{\boldsymbol{\beta}}_{w^{*b}s} - \hat{\boldsymbol{\beta}}_{ws})'}{B}. \tag{5.1}$$

It is worth noting that the validity of assumption 5 is thus not only required for our proposed bootstrap method but also for the Bonferroni, Wald and Rao-Scott methods.

5.1 Two naïve procedures

The weighted version of the naïve procedure consists of using the statistic $\hat{t}\,(s,\ \mathbf{w}_s;\ \mathbf{c})$ in (4.2), which is compared to the upper tail of the distribution $F_{Q,\,n-r}$. The unweighted version uses the statistic $\hat{t}\,(s,\ \mathbf{1}_s;\ \mathbf{c})$, which is again compared to the upper tail of the distribution $F_{Q,\,n-r}$. Both procedures are not expected to work well under informative sampling but are still often used in practice, especially the weighted version. Note that if sampling is not informative, the unweighted version, that ignores the sampling design, leads to a simple, valid and reasonably powerful test.

5.2 The Bonferroni procedure

The Bonferroni procedure was studied by Korn and Graubard (1990). It is simple to use and was shown to work well in their empirical study. To describe this procedure, let $\mathbf{H}'_q$ represent the q^{th} row of $\mathbf{H}$ and c_q the q^{th} element of $\mathbf{c}$. Then, compute the Q weighted statistics

$$\hat{t}_q^{BON}(s;c_q) = \frac{(\mathbf{H}'_q\hat{\boldsymbol{\beta}}_{ws} - c_q)^2}{\mathbf{H}'_q\hat{\mathbf{V}}_{mp}(\hat{\boldsymbol{\beta}}_{ws})\mathbf{H}_q}. \qquad (5.2)$$

The largest statistic $\hat{t}_q^{BON}(s;c_q)$, for $q = 1, \ldots, Q,$ is compared to the upper tail of the distribution $F_{1,\,d}$ with a revised significance level α / Q instead of α. The number of degrees of freedom d is equal to the number of sampled primary sampling units minus the number of strata. Note that this procedure depends in general on the model parametrization used.

5.3 WALD F-procedure

An F-version of the standard Wald chi-square statistic, with adjusted denominator degrees of freedom as proposed by Fellegi (1980), can be defined as

$$\hat{t}^{W}(s;\mathbf{c}) = \frac{d - Q + 1}{Qd}(\mathbf{H}\hat{\boldsymbol{\beta}}_{ws} - \mathbf{c})'(\mathbf{H}\hat{\mathbf{V}}_{mp}(\hat{\boldsymbol{\beta}}_{ws})\,\mathbf{H}')^{-1}(\mathbf{H}\hat{\boldsymbol{\beta}}_{ws} - \mathbf{c}). \qquad (5.3)$$

The statistic $\hat{t}^{W}(s;\mathbf{c})$ is compared to the upper tail of the distribution $F_{Q,\,d-Q+1}$. This procedure is implemented in the software package SUDAAN (Research Triangle Institute 2004).

5.4 Rao-Scott F-procedure

Another procedure consists of using an F-version (see Rao and Thomas 2003) of the second-order adjusted chi-square statistic of Rao and Scott (1981), which is based on Satterthwaite's correction for the number of degrees of freedom. We use an adaptation of these authors' method for linear regression, as implemented in the software package SUDAAN (Research Triangle Institute 2004). The statistic is defined as

$$\hat{t}^{RS}(s;\mathbf{c}) = \frac{1}{\bar{\lambda}\,(1 + a^2)\,Q^*}(\mathbf{H}\hat{\boldsymbol{\beta}}_{ws} - \mathbf{c})'(\mathbf{H}\hat{\mathbf{V}}_{\text{SRS}}(\hat{\boldsymbol{\beta}}_{ws})\,\mathbf{H}')^{-1}(\mathbf{H}\hat{\boldsymbol{\beta}}_{ws} - \mathbf{c}), \qquad (5.4)$$

where $\hat{\mathbf{V}}_{\text{SRS}}(\hat{\boldsymbol{\beta}}_{ws})$ is an estimator of the variance-covariance matrix of $\hat{\boldsymbol{\beta}}_{ws}$ under a simple random sampling design, $\bar{\lambda}$ is the average of the eigenvalues of the generalized design effect matrix $[\hat{\mathbf{V}}_{\text{SRS}}(\hat{\boldsymbol{\beta}}_{ws})]^{-1}\hat{\mathbf{V}}_{mp}(\hat{\boldsymbol{\beta}}_{ws})$, a is the coefficient of variation of these eigenvalues and $Q^* = Q/(1 + a^2)$. The Rao-Scott F-statistic $\hat{t}^{RS}(s;\mathbf{c})$ is compared to the upper tail of the distribution $F_{Q^*,\,d}$.

6. Simulation study

We performed a simulation study to investigate the level and power of the above test procedures in the case of informative and non-informative sampling. In sections 6.1 and 6.2, we describe the population and sample creation respectively. We then define the null hypotheses to be tested in section 6.3, describe the methods evaluated in section 6.4 and present simulation results in section 6.5.

6.1 Generation of the populations

We generated four populations of $N = 10{,}000$ units. First, a categorical variable v_k was generated independently for each population unit k so that $v_k = i,$ for $i = 1, \ldots, I,$ with probability $P(v_k = i) = 1 / I,$ where I is the number of categories of v_k, which was set equal to 5. The dependent variable y was generated as

$$y_k = \alpha_o + \alpha_1\left(v_k - \frac{(I + 1)}{2}\right) + \sigma\varphi_k, \qquad (6.1)$$

where $\varphi_k \sim N(0,\,1)$, $\alpha_o = 10$ and $\sigma = 3.$ The four populations that we generated only differ in the choice of α_1, which controls the correlation between y and v. We considered $\alpha_1 = 0$, 0.25, 0.50 and 0.75.

6.2 Generation of samples and bootstrap weights

From each of the above four populations, 5,000 stratified simple random samples of size 100 were selected without replacement under two different stratification scenarios aimed at simulating both informative and non-informative sampling. In the case of non-informative sampling, the strata correspond exactly to the five categories of variable v defined above. In the case of informative sampling, the strata are defined by the cross-classification of variable v and another categorical variable z that depends on the random error term $\sigma\varphi_k$ in (6.1). For each population unit k, variable z was created as follows: $z_k = 1$, if $\sigma\varphi_k > 0$, and $z_k = 2$, otherwise. This leads to 10 strata in the informative case that are constructed by crossing the five categories of v with the two categories of z. Each of the 10 informative strata contains about 1,000 population units while each of the 5 non-informative strata contains about 2,000 population units.

Furthermore, two different stratum allocation schemes were used. The scheme, SCHEME_UNEQUAL, allocates the 100 sample units among the strata in the following way:

Table 1
Sample sizes for SCHEME_UNEQUAL

Informative	z \ v	1	2	3	4	5
	1	4	4	16	4	28
	2	4	8	4	24	4

			v		
Non-informative	1	2	3	4	5
	8	12	20	28	32

The second scheme, denoted SCHEME_EQUAL, assigns the same number of units in each stratum as follows:

Table 2
Sample sizes for SCHEME_EQUAL

Informative	z \ v	1	2	3	4	5
	1	10	10	10	10	10
	2	10	10	10	10	10

			v		
Non-informative	1	2	3	4	5
	20	20	20	20	20

The two different schemes lead to very different sets of survey weights. The weights resulting from the SCHEME_UNEQUAL allocation are much more variable than those from SCHEME_EQUAL. Note that we simply defined the survey weight w_k as the inverse of the selection probability of unit k.

Finally, for each selected sample, 500 design-based bootstrap weights were calculated for each sampled unit, as described in Rao *et al.* (1992), among others. In our implementation of this methodology, each bootstrap sample was selected with replacement by stratified simple random sampling with $n_h - 1$ draws from the n_h sample units in stratum h. This methodology takes the sampling design variability into account (with a slight overestimation of the design variance due to assuming with-replacement sampling) but ignores the model variability. This is acceptable since the overall sampling fraction (1/100) is small.

6.3 Null hypotheses

For each selected sample, we modeled y_k as a function of v_k using an analysis of variance model. More specifically, we defined indicator variables

$$x_{ik} = \begin{cases} 1, & \text{if } v_k = i, \\ 0, & \text{otherwise,} \end{cases}$$

for $i = 1, \ldots, I$, and fitted the linear model $y_k = \beta_0 + \sum_{i=1}^{I-1} \beta_i \, x_{ik} + \varepsilon_k$ using the weighted least-squares technique, where ε_k is a random error term with mean 0 and constant variance. We considered testing the following two null hypotheses:

$$\text{TEST1: } H_0\text{: } \beta_1 = 0$$
$$\text{TEST2: } H_0\text{: } \beta_1 = \beta_2 = \ldots = \beta_{I-1} = 0.$$

Note that both null hypotheses are true for the population with $\alpha_1 = 0$ while they are false for the other populations. The latter three populations are used to assess the power of the different test procedures under study.

6.4 Test methods

For each selected sample, we tested the above two null hypotheses using five different methods: the proposed bootstrap method, the naïve method (both unweighted and weighted versions) described in section 5.1, the Bonferroni method described in section 5.2, the Wald F method described in section 5.3 and the Rao-Scott F method described in section 5.4. Results for the naïve method are standard output in the software SAS whereas the Wald and Rao-Scott F-statistics are standard output in the SUDAAN statistical software, version 9. The Bonferroni statistics (5.2) are also obtained through SUDAAN. The proposed method is programmed in the statistical software SAS, version 8.

In addition, we also performed the simulation study using a linearized variance estimator in the Wald, Rao-Scott and Bonferroni methods instead of the bootstrap variance estimator (5.1). Rejection rates obtained using the linearized variance estimator were slightly lower but quite similar to those obtained using (5.1). Given this observation and that our focus is on bootstrap methods, we neither show nor discuss these additional results in the next section.

6.5 Simulation results

For each population, stratification scenario, allocation scheme, null hypothesis and method, we calculated the rejection rate in percentage over the 5,000 selected samples (using a 5% significance level). Results are given below in tables 3A, 3B, 4A and 4B. The results are more striking and more interesting for the null hypothesis TEST2 than the null hypothesis TEST1. We will thus focus our discussion of the results on the former.

Tables 3A and 3B contain the results in the case of informative sampling, which is of more interest to us. Let us discuss first results in table 3A for SCHEME_UNEQUAL. Both naïve methods perform poorly as they do not properly exploit sampling design information. On the one hand, the unweighted version is definitely too liberal as its rejection rate is far above 5% under the null hypothesis. On the other hand, the weighted version is too conservative and significantly lacks power when compared to other methods. The Wald method is too liberal with a rejection rate of 15.8% when H_0 is true. The simple Bonferroni method improves the situation although it is still too liberal with a rejection rate of 11.4% when H_0 is true. This result is somewhat surprising as the Bonferroni method is known to be (asymptotically) conservative. A referee suggested that we consider an improved Bonferroni method such as that developed by Benjamini and Hochberg (1995). In this simulation study, such a method would not help as it always rejects more often than the standard Bonferroni method. The Rao-Scott method significantly outperforms the Wald and Bonferroni methods under the null hypothesis with a rejection rate of 6.8%. The proposed bootstrap method is comparable to the proven but more complicated Rao-Scott method with perhaps even a slight improvement in the level with a rejection rate of 6.2% when H_0 is true. However, the Rao-Scott method is slightly more powerful than the proposed bootstrap method.

Table 3B contains results under SCHEME_EQUAL in the informative sampling scenario. Here, the weighted and unweighted versions of the naïve method yield similar results since the variability of the survey weights is quite small. Even in this case, the naïve method is definitely too conservative, which results in an extremely low power. All other methods are comparable both in terms of level (H_0 true) and power (H_0 false) although the Wald method is still slightly too liberal compared to the Bonferroni, Rao-Scott and proposed bootstrap methods with a rejection rate of 7.9% when H_0 is true.

Tables 4A and 4B contain the results in the case of non-informative sampling. Again, let us discuss first results in table 4A for SCHEME_UNEQUAL. As expected, the naïve unweighted method performs well here while the naïve weighted method becomes too liberal with a rejection rate of 12.8% when H_0 is true. In terms of the level, the proposed method is competitive to the naïve unweighted method and even slightly conservative. It outperforms the Wald method and is slightly better than the Bonferroni and Rao-Scott methods. Its power is however slightly less than these latter two competitors but still acceptable.

Table 3A
Rejection rates at the 5% significance level under SCHEME_UNEQUAL and informative sampling

SCHEME_UNEQUAL	**Informative Sampling**							
	Ho TRUE		**Ho FALSE**		**Ho FALSE**		**Ho FALSE**	
	$\alpha_1 = 0$		**$\alpha_1 = 0.25$**		**$\alpha_1 = 0.50$**		**$\alpha_1 = 0.75$**	
Method	**Test1**	**Test2**	**Test1**	**Test2**	**Test1**	**Test2**	**Test1**	**Test2**
Naïve Unweighted	37.5	100.0	85.3	100.0	98.8	100.0	100.0	100.0
Naïve Weighted	1.7	0.4	14.5	4.6	58.0	33.6	90.3	78.6
Wald	8.0	15.8	30.9	37.1	71.8	73.9	93.1	95.4
Rao-Scott	8.0	6.8	30.9	21.1	71.8	61.7	93.1	91.8
Bonferroni	8.0	11.4	30.9	32.6	71.8	68.8	93.1	91.9
Proposed Bootstrap	7.4	6.2	29.4	19.7	70.2	59.7	92.8	91.0

Table 3B
Rejection rates at the 5% significance level under SCHEME_EQUAL and informative sampling

SCHEME_EQUAL	**Informative Sampling**							
	Ho TRUE		**Ho FALSE**		**Ho FALSE**		**Ho FALSE**	
	$\alpha_1 = 0$		**$\alpha_1 = 0.25$**		**$\alpha_1 = 0.50$**		**$\alpha_1 = 0.75$**	
Method	**Test1**	**Test2**	**Test1**	**Test2**	**Test1**	**Test2**	**Test1**	**Test2**
Naïve Unweighted	0.1	0.0	6.7	0.3	58.1	16.5	97.2	79.7
Naïve Weighted	0.1	0.0	6.3	0.3	56.8	18.2	97.0	81.4
Wald	5.8	7.9	43.6	37.5	93.7	92.3	99.9	100.0
Rao-Scott	5.8	5.5	43.6	32.1	93.7	90.4	99.9	99.9
Bonferroni	5.8	6.2	43.6	33.6	93.7	88.6	99.9	99.8
Proposed Bootstrap	2.3	5.1	42.3	31.0	93.6	89.6	99.9	99.9

Table 4A
Rejection rates at the 5% significance level under SCHEME_UNEQUAL and non-informative sampling

SCHEME_UNEQUAL	Non-Informative Sampling							
	Ho TRUE		Ho FALSE		Ho FALSE		Ho FALSE	
	α_1 = 0		α_1 = 0.25		α_1 = 0.50		α_1 = 0.75	
Method	**Test1**	**Test2**	**Test1**	**Test2**	**Test1**	**Test2**	**Test1**	**Test2**
Naïve Unweighted	4.2	4.7	13.5	11.2	39.9	34.6	71.8	70.5
Naïve Weighted	11.4	12.8	24.6	23.0	56.8	50.2	83.8	81.2
Wald	7.6	8.6	16.8	17.8	42.9	42.6	72.5	76.2
Rao-Scott	7.6	6.4	16.8	12.3	42.9	32.1	72.5	72.5
Bonferroni	7.6	7.1	16.8	16.5	42.9	42.1	72.5	75.0
Proposed Bootstrap	6.3	4.5	14.4	9.2	38.5	26.4	68.2	56.4

Table 4B
Rejection rates at the 5% significance level under SCHEME_EQUAL and non-informative sampling

SCHEME_EQUAL	Non-Informative Sampling							
	Ho TRUE		Ho FALSE		Ho FALSE		Ho FALSE	
	α_1 = 0		α_1 = 0.25		α_1 = 0.50		α_1 = 0.75	
Method	**Test1**	**Test2**	**Test1**	**Test2**	**Test1**	**Test2**	**Test1**	**Test2**
Naïve Unweighted	4.9	4.5	17.2	12.4	54.3	42.2	88.2	81.7
Naïve Weighted	5.0	4.5	17.4	12.5	54.7	42.7	88.3	81.9
Wald	5.7	6.9	18.8	16.3	56.6	48.9	88.9	85.0
Rao-Scott	5.7	5.0	18.8	13.1	56.6	49.2	88.9	82.6
Bonferroni	5.7	5.4	18.8	13.7	56.6	45.1	88.9	81.8
Proposed Bootstrap	5.0	3.3	16.4	10.0	53.2	36.5	86.8	77.6

Table 4B contains results under SCHEME_EQUAL in the non-informative sampling scenario. In this table, the methods do not appear to differ drastically. As expected, the naïve method (both weighted and unweighted versions) performs well although it did not outperform the Rao-Scott and Bonferroni methods in this simulation study. The proposed method is still slightly conservative in this non-informative scenario and has slightly less power than the other methods.

To investigate the effect of large samples on the test procedures, we also performed some simulations with sample sizes that are ten times larger than in the original setup, as suggested by one reviewer. That is, we considered a population size of 100,000 and selected 1,000 samples of size 1,000 thus deliberately keeping the same small sampling fraction. From this setup, we obtained results when H_0 is true, shown in table 5, for both informative and non-informative sampling under unequal stratum allocation. As expected, all the methods other than the naïve ones have similar rejection rates that are indeed slightly lower than 5%. This illustrates that the differences between the methods become less important as the sample size increases.

Table 5
Rejection rates at the 5% significance level under SCHEME_UNEQUAL

SCHEME_UNEQUAL	Informative		Non-informative	
	Ho TRUE		Ho TRUE	
	α_1 = 0		α_1 = 0	
Method	**Test1**	**Test2**	**Test1**	**Test2**
Naïve Unweighted	100.0	100.0	3.7	3.8
Naïve Weighted	1.3	0.7	9.3	10.5
Wald	4.6	4.5	3.2	4.1
Rao-Scott	4.6	3.8	3.2	3.8
Bonferroni	4.6	4.5	3.2	3.6
Proposed Bootstrap	4.4	3.6	2.9	3.8

Overall, our proposed bootstrap method was the best in terms of the level, followed closely by the Rao-Scott method. It gave somewhat conservative results in the non-informative sampling scenarios. This was accompanied by a slight loss of power. The Rao-Scott method is a good alternative if users have access to an appropriate software package. The Bonferroni method is simple to use but may be too liberal and the Wald method is even worse. The naïve methods may have serious deficiencies, either in the level or in the power, although the naïve unweighted method is viable if one is reasonably sure that sampling is not informative.

7. Summary and discussion

We have proposed a general and simple bootstrap procedure for testing hypotheses from survey data, which could also be applied outside the survey sampling field. Our procedure uses classical model-based test statistics and is thus easy to implement for analysts using classical software packages. We have shown in a simulation study that it performed well in the context of a linear regression model. These good results are encouraging and may suggest that our proposed bootstrap procedure could be useful with other more complicated models and other statistics. The idea could also be easily adapted for the construction of bootstrap confidence intervals.

One could also consider bootstrapping an asymptotically pivotal statistic such as the Rao-Scott statistic (5.4). This would however involve double bootstrapping if $\hat{\mathbf{V}}_{mp}(\hat{\boldsymbol{\beta}}_{ws})$ is estimated using the bootstrap technique as in (5.1). Double bootstrapping requires generating another set of bootstrap replicates for each initial bootstrap replicate. Although better test procedures could potentially be obtained, double bootstrapping may not be convenient for analysts. By focusing on simpler statistics that do not involve the bootstrap technique, our test procedure avoids double bootstrapping and remains simple.

The properties of our method depend not only on the choice of the test statistic but also on the construction of the bootstrap weights. Typically, bootstrap weights capture the first two design moments of the sampling error, which should be sufficient in most cases to satisfy our bootstrap assumptions 3, 4 and 5. Bootstrap weights that also capture the third design moment could perhaps be useful for improving the level accuracy of the bootstrap test. This needs further investigation. Finally, as already pointed out in section 3, standard design-based bootstrap weights satisfy assumption 5 only when the overall sampling fraction is negligible so that the model portion of the total variance (3.1) is negligible. Research is needed to develop proper bootstrap weights, when a non-negligible sampling fraction is used, that capture both the model and the design portions of the total variance.

Acknowledgements

We sincerely thank the Associate Editor and two referees for their comments. We also thank J.N.K. Rao from Carleton University as well as David Binder and Yves Lafortune from Statistics Canada for their comments and stimulating discussions on this topic. All these comments and discussions were useful to improve the general quality of the paper and its clarity.

Appendix

Proof of result 1

From assumption 1, we can easily see that

$$\sqrt{n}\,(\mathbf{H}\hat{\boldsymbol{\beta}}_{ws} - \mathbf{H}\boldsymbol{\beta}) \xrightarrow{mp} N(\mathbf{0}, \mathbf{H}\boldsymbol{\Sigma}\mathbf{H}'). \qquad (\text{A}.1)$$

Using a standard result on quadratic forms (*e.g.*, Seber 1984, page 540) and equation (A.1), we obtain

$$n\,(\mathbf{H}\hat{\boldsymbol{\beta}}_{ws} - \mathbf{H}\boldsymbol{\beta})'\tilde{\mathbf{A}}^{-1}(\mathbf{H}\hat{\boldsymbol{\beta}}_{ws} - \mathbf{H}\boldsymbol{\beta}) \xrightarrow{mp} \sum_{q=1}^{Q}\lambda_q\Omega_q, \quad (\text{A}.2)$$

where λ_q, for $q = 1, \ldots, Q$, are the eigenvalues of $\boldsymbol{\Lambda} = (\tilde{\mathbf{A}}^{-1})(\mathbf{H}\boldsymbol{\Sigma}\mathbf{H}')$ and Ω_q are independent chi-square random variables with one degree of freedom. Therefore, from (A.2) and assumption 2, we have

$$\hat{t}\,(s, \mathbf{w}_s;\ \mathbf{H}\boldsymbol{\beta}) =$$

$$(\mathbf{H}\hat{\boldsymbol{\beta}}_{ws} - \mathbf{H}\boldsymbol{\beta})'\{\hat{\mathbf{A}}(s, \mathbf{w}_s)\}^{-1}(\mathbf{H}\hat{\boldsymbol{\beta}}_{ws} - \mathbf{H}\boldsymbol{\beta}) \xrightarrow{mp} \sum_{q=1}^{Q}\lambda_q\Omega_q.$$

References

Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society*, Series B, 57, 289-300.

Binder, D.A., and Roberts, G.R. (2003). Design-based and model-based methods for estimating model parameters. *Analysis of survey data*, (Eds. R.L. Chambers and C.J. Skinner). New-York: John Wiley & Sons, Inc.

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

Efron, B., and Tibshirani, R.J. (1993). *An introduction to the bootstrap*. New York: Chapman & Hall.

Fay, R.E. (1985). A jackknifed chi-squared test for complex samples. *Journal of the American Statistical Association*, 80, 148-157.

Fellegi, I.P. (1980). Approximate tests of independence and goodness of fit based on stratified multistage samples. *Journal of the American Statistical Association*, 75, 261-268.

Graubard, B.I., and Korn, E.L. (1993). Hypothesis testing with complex survey data: The use of classical quadratic test statistics with particular reference to regression problems. *Journal of the American Statistical Association*, 88, 629-641.

Graubard, B.I., Korn, E.L. and Midthune, D. (1997). Testing goodness-of-fit for logistic regression with survey data. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 170-174.

Hall, P., and Wilson, S.R. (1991). Two guidelines for bootstrap hypothesis testing. *Biometrics*, 47, 757-762.

Hughes, A.L., and Brodsky, M.D. (1994). Variance estimation of drug abuse episodes using the bootstrap. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 212-217.

Korn, E.L., and Graubard, B.I. (1990). Simultaneous testing of regression coefficients with complex survey data: Use of Bonferroni *t* statistics. *The American Statistician*, 44, 270-276.

Korn, E.L., and Graubard, B.I. (1991). A note on the large sample properties of linearization, jackknife and balanced repeated replication methods for stratified samples. *The Annals of Statistics*, 19, 2275-2279.

Kovar, J.G., Rao, J.N.K. and Wu, C.F.J. (1988). Bootstrap and other methods to measure errors in survey estimates. *The Canadian Journal of Statistics*, 16, Supplement, 25-45.

Langlet, É.R., Faucher, D. and Lesage, É. (2003). An application of the bootstrap variance estimation method to the Canadian Participation and Activity Limitation Survey. *Proceedings of the Joint Statistical Meetings*, American Statistical Association, 2299-2306.

Rao, J.N.K., and Scott, A.J. (1981). The analysis of categorical data from complex sample surveys: Chi-squared tests for goodness of fit and independence in two-way tables. *Journal of the American Statistical Association*, 76, 221-230.

Rao, J.N.K., and Wu, C.F.J. (1988). Resampling inference with complex survey data. *Journal of the American Statistical Association*, 83, 231-241.

Rao, J.N.K., Wu, C.F.J. and Yue, K. (1992). Some recent work on resampling methods for complex surveys. *Survey Methodology*, 18, 209-217.

Rao, J.N.K., and Thomas, D.R. (2003). Analysis of categorical response data from complex surveys: An appraisal and update. *Analysis of survey data*, (Eds. R.L. Chambers and C.J. Skinner). New-York: John Wiley & Sons, Inc.

Research Triangle Institute (2004). *SUDAAN language manual*, *release* 9.0. Research Triangle Park, NC: Research Triangle Institute.

Seber, G.A.F. (1984). *Multivariate Observations*. New-York: John Wiley & Sons, Inc.

Shao, J., and Tu, D. (1995). *The jackknife and the bootstrap*. New-York: Springer-Verlag.

Sitter, R.R. (1992). Comparing three bootstrap methods for survey data. *The Canadian Journal of Statistics*, 20, 135-154.

Thomas, D.R., and Rao, J.N.K. (1987). Small-sample comparisons of level and power for simple goodness-of-fit statistics under cluster sampling. *Journal of the American Statistical Association*, 82, 630-636.

Thomas, D.R., Singh, A.C. and Roberts, G.R. (1996). Tests of independence on two-way tables under cluster sampling: An evaluation. *International Statistical Review*, 64, 295-311.

Yeo, D., Mantel, H. and Liu, T.-P. (1999). Bootstrap variance estimation for the National Population Health Survey. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 778-783.

Survey Methodology, June 2009
Vol. 35, No. 1, pp. 37-51
Statistics Canada, Catalogue No. 12-001-X

Bayesian methods for an incomplete two-way contingency table with application to the Ohio (Buckeye State) Polls

Bo-Seung Choi, Jai Won Choi and Yousung Park [1]

Abstract

We use a Bayesian method to resolve the boundary solution problem of the maximum likelihood (ML) estimate in an incomplete two-way contingency table, using a loglinear model and Dirichlet priors. We compare five Dirichlet priors in estimating multinomial cell probabilities under nonignorable nonresponse. Three priors among them have been used for an incomplete one-way table, while the remaining two new priors are newly proposed to reflect the difference in the response patterns between respondents and the undecided. The Bayesian estimates with the previous three priors do not always perform better than ML estimates unlike previous studies, whereas the two new priors perform better than both the previous three priors and the ML estimates whenever a boundary solution occurs. We use four sets of data from the 1998 Ohio state polls to illustrate how to use and interpret estimation results for the elections. We use simulation studies to compare performance of the five Bayesian estimates under nonignorable nonresponse.

Key Words: Bayesian analysis; Nonignorable nonresponse; Contingency table; Boundary solution; EM algorithm.

1. Introduction

The problem of nonresponse is common in most surveys becoming a serious issue as the nonresponse rate increases (De Heer 1999; Groves and Couper 1998). When survey data is summarized in a two-way contingency table, the table includes fully classified counts, partially classified counts (*i.e.*, item nonresponse), and unclassified counts (*i.e.*, unit nonresponse). For example, in the Ohio (Buckeye State) Poll (BSP) (Chen and Stasny 2003), one category involves the voting preference (candidates A,B,C, or undecided) and the other category is the likelihood of voting (likely to vote, not likely to vote, and undecided). First supplemental margin contains data only on the voting preference, second contains data only on the likelihood of voting, and third is only the number of unit nonresponses (both responses unknown). Our interest is to incorporate these missing observations into estimating the true support for each candidate and to present Bayesian models to predict the winner.

In some surveys, the undecided answers are treated as a valid response category when the respondents do not have strong preference for a candidate and voting intention (Smith 1984; Rubin, Stern and Vehovar 1995). Many studies, however, have shown that the voting behavior of the undecided voters can have a significant impact on the final result and that by considering these undecided voters, the accuracy of election forecasting can be improved (Perry 1979; Fenwick, Wiseman, Becker and Heiman 1982; Myers and O'Connor 1983; Kim 1995; Chen and Stasny 2003; Martin, Traugott and Kennedy 2005). Perry (1979), among them, showed that the undecided percentage in a poll is likely to be greater than the true percentage by presenting an empirical evidence using a secret ballot approach. Kim (1995) also indicated that these undecided voters are critical, especially in cases where the number of undecided voters is greater than the gap between the two leading runners in an election race. Three of our empirical studies in Section 3 belong to this critical case. Fenwick *et al.* (1982) and Kim (1995) applied a discriminant analysis to the October 1980 poll data in Massachusetts and the 1992 USA presidential election, from which they allocated the undecided voters to candidates to show that undecided voters generally do not vote in the same proportions as their decided counterparts. When the focus is on the candidate the undecided voter may vote for, undecided responses are better treated as missing data (Myers and O'Connor 1983). As indicated in Flannelly, Flannelly and McLeod (2000) and Lau (1994), the forecasting error for the actual election results increases as the rate of undecided voters increases. To overcome this problem, Monterola, Lim, Garcia and Saloma (2001) applied a neural network approach to classify undecided voters in a public opinion survey. Smith, Skinner and Clarke (1999) and Molenberghs, Kenward and Goetghebeur (2001) utilized model based imputation methods for the 1992 British General Election Panel Survey and the 1991 Slovenian plebiscite public opinion survey. Because our main goal is to obtain more accurate forecasts by allocating undecided voters to proper cell, we treat undecided voters as missing observations in the same way as these researchers handled them.

[1] Bo-Seung Choi, Research Professor, Institute of Economics, Korea University, Seoul 136-701, Korea; Jai Won Choi, Professor, Department of Biostatistics, Medical College of Georgia, Augusta, GA 30912; Yousung Park, Professor, Department of Statistics, Korea University, Seoul 136-701, Korea. E-mail: yspark@korea.ac.kr.

Nonresponse (or undecided, equivalently) can be distinguished by three types of nonresponses (Little and Rubin 2002, page 11): missing completely at random (MCAR) means that the probability of a nonresponse on a variable of interest is independent of all survey variables including itself; missing at random (MAR) means that the probability of a nonresponse depends only on the observed data; missing not at random (MNAR) means that the probability of nonresponse depends on the unobserved values. Models for MCAR or MAR are called ignorable nonresponse models while models for MNAR are called nonignorable. For example, in a pre-election survey, if the respondents do not answer with their preference of a candidate, although they support a particular candidate, the pattern for candidate preference can be different between the respondents and nonrespondents. Then, the nonresponse mechanism is nonignorable. When data is assumed to be MCAR, the effect of nonresponse can be removed in likelihood inference (Little and Rubin 2002, page 11). However, when the nonrespondents follow a response pattern different from that of the respondents, discarding nonresponses or misspecifying the nonresponse mechanism leads to larger variances and biases in estimation (Chen 1972; Park and Brown 1994).

When nonresponse is nonignorable in contingency tables, ML estimation often yields boundary solutions where the probability of nonresponse is estimated to be zero in some cells. These boundary solutions often provide a local maximum of the likelihood function. In this case, the maximum likelihood (ML) estimates of the loglinear model parameters cannot have a unique solution and usually have large standard deviations (see Section 4 or Baker, Rosenberger and Dersimonian (1992) and Park and Brown (1994) for more detailed discussions).

The conditions where the ML estimate falls on the boundary solution have been proposed in a one-way contingency table (Baker and Laird 1988; Michiels and Molenbergs 1997). The geometric explanation for the boundary solution of the ML estimate was presented (Smith *et al.* 1999; Clark 2002). Baker *et al.* (1992) presented a sufficient and necessary condition under which the ML estimate can have a boundary solution in a two-way contingency table.

To overcome such a boundary problem in the ML estimate under the existence of nonignorable nonresponses, Park and Brown (1994) and Park (1998) proposed Bayesian approach using empirical priors based only on respondent information. Clogg, Rubin, Schenker and Schultz (1991) used constant prior for an incomplete one-way contingency table. Although they showed that, under nonignorable nonresponse, Bayesian methods provided smaller mean squared errors (MSE) than ML estimate in estimating cell expectations, our simulation study shows that this is generally not true in an incomplete two-way contingency table. Thus, we present two Bayesian models whose priors depend on information from both respondents and undecided. We, then, apply each to analyze incomplete two-way contingency table. An extension to a multi-way table is straightforward. We can easily apply this extension to weighted data from stratified or cluster sampling using appropriate covariates (see Section 2.2).

The remainder of this paper is divided into four sections. In Section 2, we consider Bayesian models with five different priors and present a generalized expectation maximization (EM) algorithm to estimate cell probabilities. In Section 3, we apply the Bayesian models to four empirical data sets from the Buckeye State Poll (BSP) and compare the Bayesian estimates with the ML estimate and the actual election results. In Section 4, we use simulation studies to compare MSEs and biases of the Bayesian estimates from different missing percentages and response patterns of the respondents and nonrespondents. In this section, we also calculate the coverage probability to examine the performance of the Bayesian estimates. Section 5 includes some concluding remarks.

2. Bayesian models

We discuss five Bayesian estimates to accommodate nonignorable nonresponse in an incomplete two-way contingency table. We present an EM algorithm to tackle the nonresponse problem in a two-way contingency table in Section 2.1. Then, in Section 2.2, we specify five priors and extend our approach to a multi-way contingency table.

Let X_1 and X_2 be response variables indexed by I and J categories, respectively, in a two-way contingency table. We also let $R_1 = 1$ when X_1 is observed and $R_1 = 2$ when X_1 is missing. Similarly, $R_2 = 1$ when X_2 is observed and $R_2 = 2$ when X_2 is missing. Then the full array of X_1, X_2, R_1, and R_2 constructs a $I \times J \times 2 \times 2$ contingency table in which we have completely classified counts, partially classified counts, and unclassified counts. To distinguish these three types of observations, let y_{ijkl} be the count belonging to the i^{th} category of X_1, the j^{th} category of X_2, the k^{th} value of R_1, and the l^{th} value of R_2. Thus, y_{ij11} is used for the completely classified counts, y_{i+12} and y_{+j21} for the respective column and row supplemental margins, and y_{++22} for the unclassified counts. We assume a multinomial distribution for these three types of observations to have the following log likelihood:

$$l = \sum_i \sum_j y_{ij11} \cdot \log(\pi_{ij11}) + \sum_i y_{i+12} \cdot \log(\pi_{i+12}) + \sum_j y_{+j21} \cdot \log(\pi_{+j21}) + y_{++22} \cdot \log(\pi_{++22}) \qquad (1)$$

where $\pi_{ijkl} = \Pr[X_1 = i,\ X_2 = j,\ R_1 = k,\ R_2 = l]$ and $N = \sum_{i,j,k,l} y_{ijkl}$ is fixed.

Since this likelihood function involves more parameters than degrees of freedom available for estimation, we link π_{ijkl} to relevant covariates using a loglinear function. Since no explanatory variable is available, we do not use any explanatory variables. However, the loglinear model can easily incorporate explanatory variables in the same way as it incorporates the categorical variables (see Baker and Laird 1988 and Park and Brown 1994 for details).

A nonignorable nonresponse model for all of the variables X_1, X_2, R_1, and R_2 is defined by

$$\begin{aligned}\log(m_{ijkl}) &= \beta_0 + \beta^i_{X_1} + \beta^j_{X_2} + \beta^k_{R_1} + \beta^l_{R_2} \\ &\quad + \beta^{ik}_{X_1R_1} + \beta^{jl}_{X_2R_2} + \beta^{ij}_{X_1X_2} + \beta^{kl}_{R_1R_2}\end{aligned}$$
$$\text{for } i = 1, \ldots, I,\ j = 1, \ldots, J,\ k = 1, 2, \text{ and } l = 1, 2 \quad (2)$$

where $m_{ijkl} = N \cdot \pi_{ijkl}$ is the expected cell count for the $(i, j, k, l)^{\text{th}}$ category and the sum of each β-term over any of its respective super script(s) is zero.

This loglinear model is saturated since the number of parameters is exactly the same as the number of cells observed from the incomplete two-way contingency table. This model is also a nonignorable nonresponse model because of the interaction terms between X_1 and R_1 and between X_2 and R_2, implying that the nonresponse of each response variable depends on its own status. The loglinear model is a tool frequently used for analyzing incomplete contingency tables with nonignorable nonresponses. Let p be the number of parameters (*i.e.*, β) to be estimated. We introduce the $p \times 1$ design vector $\mathbf{z}_{ijkl}$ to indicate the affiliation of the observation belonging to the $(i, j, k, l)^{\text{th}}$ category. Then the loglinear model given in (2) can be rewritten as

$$\log \mathbf{m} = Z\boldsymbol{\beta} \quad (3)$$

where the $I \times J \times 2 \times 2$ vector $\mathbf{m}$ is the cell expectation and $\boldsymbol{\beta}$ is the vector representation of the β s. To avoid a boundary solution of the ML estimate in model (2), we impose Dirichlet priors to the cell probabilities $(\pi_{ij11}, \pi_{ij12}, \pi_{ij21}, \pi_{ij22})$ as given by

$$\prod_i \prod_j \pi_{ij11}^{\delta_{ij11}} \cdot \pi_{ij12}^{\delta_{ij12}} \cdot \pi_{ij21}^{\delta_{ij21}} \cdot \pi_{ij22}^{\delta_{ij22}} \quad (4)$$

where the hyper parameters, the δ_{ijkl} s are specified in Section 2.2. These Dirichlet priors produce an explicit and convenient form of a posterior distribution because they are conjugated to a multinomial distribution (Clogg *et al.* 1991; Park and Brown 1994; Forster and Smith 1998). Together with (3), the multinomial distribution of (1) for observations, and the prior distribution (4), we have the following log posterior distribution:

$$\begin{aligned} l_{pos} &= \sum_i \sum_j y_{ij11} \cdot (\mathbf{z}_{ij11} \cdot \boldsymbol{\beta}) \\ &\quad - \sum_i \sum_j y_{ij11} \cdot \log\Big(\sum_{i,j,k,l} \exp(\mathbf{z}_{ijkl} \cdot \boldsymbol{\beta}) \Big) \\ &\quad + \sum_i y_{i+12} \cdot \log\Big(\sum_j \exp(\mathbf{z}_{ij12} \cdot \boldsymbol{\beta}) \Big) \\ &\quad - \sum_i y_{i+12} \cdot \log\Big(\sum_{i,j,k,l} \exp(\mathbf{z}_{ijkl} \cdot \boldsymbol{\beta}) \Big) \\ &\quad + \sum_j y_{+j21} \cdot \log\Big(\sum_i \exp(\mathbf{z}_{ij21} \cdot \boldsymbol{\beta}) \Big) \\ &\quad - \sum_j y_{+j21} \cdot \log\Big(\sum_{i,j,k,l} \exp(\mathbf{z}_{ijkl} \cdot \boldsymbol{\beta}) \Big) \\ &\quad + y_{++22} \cdot \log\Big(\sum_i \sum_j \exp(\mathbf{z}_{ij22} \cdot \boldsymbol{\beta}) \Big) \\ &\quad - y_{++22} \cdot \log\Big(\sum_{i,j,k,l} \exp(\mathbf{z}_{ijkl} \cdot \boldsymbol{\beta}) \Big) \\ &\quad + \sum_{i,j,k,l} \delta_{ijkl} \cdot (\mathbf{z}_{ijkl} \cdot \boldsymbol{\beta}) \\ &\quad - \sum_{i,j,k,l} \delta_{ijkl} \cdot \log\Big(\sum_{i,j,k,l} \exp(\mathbf{z}_{ijkl} \cdot \boldsymbol{\beta}) \Big). \end{aligned} \quad (5)$$

Equation (5) is rather complex and thus we use the EM algorithm to estimate the parameters (*i.e.*, $\boldsymbol{\beta}$).

2.1 The EM algorithm

We maximize the posterior distribution given in (5) over the parameter $\boldsymbol{\beta}$ using the generalized expectation maximization (GEM) algorithm (Dempster, Laird and Rubin 1977) with the following E and M steps.

E-step: Using augmented y_{ij12}, y_{ij21}, and y_{ij22} for $i = 1, \ldots, I$ and $j = 1, \ldots, J$, the posterior (5) can be written as

$$\begin{aligned} l_{a.\text{pos}} &= \sum_i \sum_j (y_{ij11} + \delta_{ij11}) \log(\pi_{ij11}) \\ &\quad + \sum_i \sum_j (y_{ij12} + \delta_{ij12}) \log(\pi_{ij12}) \\ &\quad + \sum_i \sum_j (y_{ij21} + \delta_{ij21}) \log(\pi_{ij21}) \\ &\quad + \sum_i \sum_j (y_{ij22} + \delta_{ij22}) \log(\pi_{ij22}). \end{aligned} \quad (6)$$

To determine the expected augmented log posterior in (6), we average over the missing counts y_{ij12}, y_{ij21}, and y_{ij22} conditioning on the current parameter estimates, π^{old}_{ijkl}, and the marginal sums y_{i+12}, y_{+j21}, and y_{++22}:

$$
\begin{aligned}
E_{\text{old}}[l_{a.\text{pos}}] &= \sum_i \sum_j (y_{ij11} + \delta_{ij11}) \cdot \log(\pi_{ij11}) \\
&+ \sum_i \sum_j (E_{\text{old}}[y_{ij12} \mid \pi_{ijkl}^{\text{old}}, y_{i+12}] + \delta_{ij12}) \cdot \log(\pi_{ij12}) \\
&+ \sum_i \sum_j (E_{\text{old}}[y_{ij21} \mid \pi_{ijkl}^{\text{old}}, y_{+j21}] + \delta_{ij21}) \cdot \log(\pi_{ij21}) \\
&+ \sum_i \sum_j (E_{\text{old}}[y_{ij22} \mid \pi_{ijkl}^{\text{old}}, y_{++22}] + \delta_{ij22}) \cdot \log(\pi_{ij22}). \quad (7)
\end{aligned}
$$

Since y_{ij12}, y_{ij21}, and y_{ij22} are multinomial random variates conditioned on the respective marginal sum y_{i+12}, y_{+j21}, and y_{++22}, the conditional expectations in the equation (7) are given by

$$
E_{\text{old}}(y_{ij12} \mid \pi_{ijkl}^{\text{old}}, y_{i+12}) = y_{i+12} \frac{m_{ij12}^{\text{old}}}{m_{i+12}^{\text{old}}},
$$

$$
E_{\text{old}}(y_{ij21} \mid \pi_{ijkl}^{\text{old}}, y_{+j21}) = y_{+j21} \frac{m_{ij21}^{\text{old}}}{m_{+j21}^{\text{old}}},
$$

and

$$
E_{\text{old}}(y_{ij22} \mid \pi_{ijkl}^{\text{old}}, y_{++22}) = y_{++22} \frac{m_{ij22}^{\text{old}}}{m_{++22}^{\text{old}}}
$$

where $m_{ijkl}^{\text{old}} = N \cdot \pi_{ijkl}^{\text{old}}$.

M-step: In this step, we maximize the expected log posterior (7) using the pseudo observations $\tilde{y}_{ij11} = y_{ij11} + \delta_{ij11}, \tilde{y}_{ij12} = y_{i+12}\, m_{ij12}^{\text{old}} / m_{i+12}^{\text{old}} + \delta_{ij12}$, $\tilde{y}_{ij21} = y_{+j21}\, m_{ij21}^{\text{old}} / m_{+j21}^{\text{old}} + \delta_{ij21}$, and $\tilde{y}_{ij22} = y_{++22}\, m_{ij22}^{\text{old}} / m_{++22}^{\text{old}} + \delta_{ij22}$. We impose the constraints on these pseudo observations so that their marginal sums are the same as the corresponding marginal sums of observations: $\tilde{y}_{++11} = y_{++11}$, $\tilde{y}_{i+12} = y_{i+12}$, $\tilde{y}_{+j21} = y_{+j21}$, and $\tilde{y}_{++22} = y_{++22}$. Under these constraints, the pseudo observations are now

$$
y_{ijkl}^* = \begin{cases}
\tilde{y}_{ij11} \dfrac{y_{++11}}{y_{++11} + \delta_{++11}} & \text{for k = 1 and l = 1} \\
\tilde{y}_{ij12} \dfrac{y_{i+12}}{y_{i+12} + \delta_{i+12}} & \text{for k = 1 and l = 2} \\
\tilde{y}_{ij21} \dfrac{y_{+j21}}{y_{+j21} + \delta_{+j21}} & \text{for k = 2 and l = 1} \\
\tilde{y}_{ij22} \dfrac{y_{++22}}{y_{++22} + \delta_{++22}} & \text{for k = 2 and l = 2.}
\end{cases}
$$

Then, the expected log posterior function (7) becomes

$$
\begin{aligned}
E_{\text{old}}[l_{a.\text{pos}}] &= \sum_i \sum_j y_{ij11}^* \cdot \log(\pi_{ij11}) \\
&+ \sum_i \sum_j y_{ij12}^* \cdot \log(\pi_{ij12}) \\
&+ \sum_i \sum_j y_{ij21}^* \cdot \log(\pi_{ij21}) \\
&+ \sum_i \sum_j y_{ij22}^* \cdot \log(\pi_{ij22}).
\end{aligned}
$$

This equation has the same form as the likelihood obtained from a four-way contingency table with fully observed cell counts y_{ijkl}^* s. Thus, using the iterative re-weighted least squares method (Agresti 2002, page 342), we obtain the maximum posterior estimator (MPE) of $\boldsymbol{\beta}$ as follows:

$$
\boldsymbol{\beta}^{(t+1)} = (Z^T \hat{V}_t^{-1} Z)^{-1} Z^T \hat{V}_t^{-1} \gamma^{(t)},
$$

where $\gamma^{(t)}$ has element $\gamma_{ijkl}^{(t)} = \log m_{ijkl}^{(t)} + (y_{ijkl} - m_{ijkl}^{(t)}) / m_{ijkl}^{(t)}$ and $\hat{V}_t = [\text{diag}(\mathbf{m}^{(t)})]^{-1}$. We finally iterate these E and M-steps until a convergence criterion is achieved. The convergence criterion we use is $\varepsilon \le 10^{-6}$, where ε is the difference between two consecutive log posterior functions.

Let $Y_{\text{obs}} = (y_{ij11}, y_{i+12}, y_{+j21}, y_{++22})$ and $Y_{\text{mis}} = (y_{ij12}, y_{ij21}, y_{ij22})$ for $i = 1, \ldots, I$ and $j = 1, \ldots, J$ be the observed count vector and the missing count vector, respectively. Then the log posterior distribution (5) can be written as

$$
\begin{aligned}
l_{\text{pos}} = l(\boldsymbol{\beta} \mid Y_{\text{obs}}) &= l(\boldsymbol{\beta} \mid Y_{\text{obs}}, Y_{\text{mis}}) \\
&- \log f(Y_{\text{mis}} \mid Y_{\text{obs}}, \boldsymbol{\beta}). \quad (8)
\end{aligned}
$$

By taking differentiation twice with respect to $\boldsymbol{\beta}$, (8) yields

$$
\begin{aligned}
\frac{\partial^2 l(\boldsymbol{\beta} \mid Y_{\text{obs}})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} &= \frac{\partial^2 l(\boldsymbol{\beta} \mid Y_{\text{obs}}, Y_{\text{mis}})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} \\
&- \frac{\partial^2 \log f(Y_{\text{mis}} \mid Y_{\text{obs}}, \boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} \\
&= - \mathbf{Z}^T [\text{diag}(\mathbf{m}) - \mathbf{m}\mathbf{m}^T / N] \, \mathbf{Z} \\
&+ \mathbf{Z}^T [\text{diag}(\boldsymbol{\pi}) - \boldsymbol{\pi}\boldsymbol{\pi}^T] \, AB\mathbf{Z}, \quad (9)
\end{aligned}
$$

where $\boldsymbol{\pi}$ is vector expression of cell probabilities π_{ijkl} and A, B are given by

$$A = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & \operatorname{diag}\left(\dfrac{y_{i+12}^2}{y_{i+12} + \delta_{i+12}} \dfrac{m_{ij12}}{m_{i+12}}\right) & 0 & 0 \\ 0 & 0 & \operatorname{diag}\left(\dfrac{y_{+j21}^2}{y_{+j21} + \delta_{+j21}} \dfrac{m_{ij21}}{m_{+j21}}\right) & 0 \\ 0 & 0 & 0 & \operatorname{diag}\left(\dfrac{y_{++22}^2}{y_{++22} + \delta_{++22}} \dfrac{m_{ij22}}{m_{++22}}\right) \end{pmatrix}$$

and

$$B = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & I_{IJ} - B^{12} & 0 & 0 \\ 0 & 0 & I_{IJ} - B^{21} & 0 \\ 0 & 0 & 0 & I_{IJ} - B^{22} \end{pmatrix}.$$

Here, to save the space and since there is no difficulty to extend for general i and j, B^{12}, B^{21}, and B^{22} are illustrated only for $I = 2$ and $J = 3$:

$$B^{12} = \begin{pmatrix} \frac{m_{1112}}{m_{1+12}} & \frac{m_{1212}}{m_{1+12}} & \frac{m_{1312}}{m_{1+12}} & 0 & 0 & 0 \\ \frac{m_{1112}}{m_{1+12}} & \frac{m_{1212}}{m_{1+12}} & \frac{m_{1312}}{m_{1+12}} & 0 & 0 & 0 \\ \frac{m_{1112}}{m_{1+12}} & \frac{m_{1212}}{m_{1+12}} & \frac{m_{1312}}{m_{1+12}} & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{m_{2112}}{m_{2+12}} & \frac{m_{2112}}{m_{2+12}} & \frac{m_{2112}}{m_{2+12}} \\ 0 & 0 & 0 & \frac{m_{2212}}{m_{2+12}} & \frac{m_{2212}}{m_{2+12}} & \frac{m_{2212}}{m_{2+12}} \\ 0 & 0 & 0 & \frac{m_{2312}}{m_{2+12}} & \frac{m_{2312}}{m_{2+12}} & \frac{m_{2312}}{m_{2+12}} \end{pmatrix},$$

$$B^{21} = \begin{pmatrix} \frac{m_{1121}}{m_{+121}} & 0 & 0 & \frac{m_{2121}}{m_{+121}} & 0 & 0 \\ 0 & \frac{m_{1221}}{m_{+221}} & 0 & 0 & \frac{m_{2221}}{m_{+221}} & 0 \\ 0 & 0 & \frac{m_{1321}}{m_{+321}} & 0 & 0 & \frac{m_{2321}}{m_{+321}} \\ \frac{m_{1121}}{m_{+121}} & 0 & 0 & \frac{m_{2121}}{m_{+121}} & 0 & 0 \\ 0 & \frac{m_{1221}}{m_{+221}} & 0 & 0 & \frac{m_{2221}}{m_{+221}} & 0 \\ 0 & 0 & \frac{m_{1321}}{m_{+321}} & 0 & 0 & \frac{m_{2321}}{m_{+321}} \end{pmatrix},$$

and

$$B^{22} = \begin{pmatrix} \frac{m_{1122}}{m_{++22}} & \frac{m_{1222}}{m_{++22}} & \frac{m_{1322}}{m_{++22}} & \frac{m_{2122}}{m_{++22}} & \frac{m_{2222}}{m_{++22}} & \frac{m_{2322}}{m_{++22}} \\ \frac{m_{1122}}{m_{++22}} & \frac{m_{1222}}{m_{++22}} & \frac{m_{1322}}{m_{++22}} & \frac{m_{2122}}{m_{++22}} & \frac{m_{2222}}{m_{++22}} & \frac{m_{2322}}{m_{++22}} \\ \frac{m_{1122}}{m_{++22}} & \frac{m_{1222}}{m_{++22}} & \frac{m_{1322}}{m_{++22}} & \frac{m_{2122}}{m_{++22}} & \frac{m_{2222}}{m_{++22}} & \frac{m_{2322}}{m_{++22}} \\ \frac{m_{1122}}{m_{++22}} & \frac{m_{1222}}{m_{++22}} & \frac{m_{1322}}{m_{++22}} & \frac{m_{2122}}{m_{++22}} & \frac{m_{2222}}{m_{++22}} & \frac{m_{2322}}{m_{++22}} \\ \frac{m_{1122}}{m_{++22}} & \frac{m_{1222}}{m_{++22}} & \frac{m_{1322}}{m_{++22}} & \frac{m_{2122}}{m_{++22}} & \frac{m_{2222}}{m_{++22}} & \frac{m_{2322}}{m_{++22}} \\ \frac{m_{1122}}{m_{++22}} & \frac{m_{1222}}{m_{++22}} & \frac{m_{1322}}{m_{++22}} & \frac{m_{2122}}{m_{++22}} & \frac{m_{2222}}{m_{++22}} & \frac{m_{2322}}{m_{++22}} \end{pmatrix}.$$

We observe that the observed data information $\partial^2 l(\boldsymbol{\beta} \mid Y_{\text{obs}}) / \partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T$ is equal to the augmented data information minus the missing data information. As shown in Gelman, Carlin, Stern and Rubin (2004, page 103), the inverse of the observed data information evaluated at the MPE of $\boldsymbol{\beta}$ is the variance of the MPE of $\boldsymbol{\beta}$.

2.2 Specification of priors

To complete the EM algorithm, we need to determine the hyper-parameters, δ_{ijkl} s. We set the sum of priors $\sum_{i,j,k,l} \delta_{ijkl}$ equal to the number of parameters involved in the loglinear model, p, as suggested by Clogg *et al.* (1991). Under this constraint, we propose five types of priors as follows. We first allocate δ_{ijkl} for the MPE of m_{ijkl} to shrink toward the MLE obtained under ignorable non-response. That is, we determine δ_{ijkl} depending only on the known response counts y_{ij11} and call them respondent-driven priors.

The first type of respondent-driven prior is, for all $i = 1, \ldots, I$ and $j = 1, \ldots, J$,

$$\delta_{ij11} = \nabla_{11} \frac{y_{ij11}}{y_{++11}}, \delta_{ij12} = \nabla_{12} \frac{y_{ij11}}{y_{++11}}, \delta_{ij21} = \nabla_{21} \frac{y_{ij11}}{y_{++11}}, \text{ and } \delta_{ij22} = \nabla_{22} \frac{y_{ij11}}{y_{++11}} \quad (10)$$

where $\nabla_{kl} = p \cdot y_{++kl} / y_{++++}$ for $k = 1, 2$ and $l = 1, 2.$ The second type of respondent-driven prior gives no prior (*i.e*, no need of prior as described below) on π_{ij11} in the first type of priors. That is, the second type is the same as the first type except $\delta_{ij11} = 0$ for all i and $j.$ In the case of a one-way contingency table (*i.e.*, either X_1 or X_2 is fully observed without missing information) and $y_{++22} = 0,$ the first type is reduced to the priors used in Park (1998), whereas the second type is reduced to the priors used in Park and Brown (1994). These two types of respondent-driven priors may be too simplistic because the non-respondents are usually assumed to have a different response pattern from the respondents under a nonignorable nonresponse model. For example, the candidate preference of nonrespondents could be different from that of respondents in a pre-election survey.

In order to define the third type of prior, denote $\hat{m}_{ijkl}$ as the MLE for m_{ijkl}. The closed form of $\hat{m}_{ijkl}$ can be obtained from Baker *et al.* (1992) where some $\hat{m}_{ijkl}$ could be zero because of boundary solutions. For example, when a supplemental column margin has a boundary solution in an incomplete 2×2 table, the MLEs are $\hat{m}_{1j11} = y_{1j11},$

$$\hat{m}_{2j11} = \frac{y_{2+11}(y_{2j11} + y_{+j21})}{y_{2+11} + y_{++21}},\ \hat{m}_{ij12} = \hat{m}_{ij11} b_j$$

where b_j is the solution of $\sum_{j=1}^{2} y_{ij11} b_j = y_{i+12},\ \hat{m}_{1j21} = 0,$

$$\hat{m}_{2j21} = \hat{m}_{2j11} \frac{y_{++21}}{y_{2+11}},\ \hat{m}_{1j22} = 0,$$

and $\hat{m}_{2j22} = \hat{m}_{2j12}\, y_{++22} / y_{2+12}.$ Therefore, these ML estimates accommodate both the information of respondents and nonrespondents, as well. The ML estimates can also be obtained from our EM algorithm in Section 2.1 by setting $\delta_{ijkl} = 0$ for all i, j, k and $l.$ Using these ML estimates, we define the third type of prior as

$$\delta_{ij11} = \nabla_{11} \cdot \left(\frac{\hat{m}_{ij11}}{\hat{m}_{++11}} \right), \delta_{ij12}$$
$$= \nabla_{12} \cdot \left(\frac{\hat{m}_{ij12}}{\hat{m}_{++12}} + \frac{1}{I \cdot J} \right) \cdot \frac{1}{2},$$
$$\delta_{ij21} = \nabla_{21} \cdot \left(\frac{\hat{m}_{ij21}}{\hat{m}_{++21}} + \frac{1}{I \cdot J} \right) \cdot \frac{1}{2},$$

and (11)

$$\delta_{ij22} = \nabla_{22} \cdot \left(\frac{\hat{m}_{ij22}}{\hat{m}_{++22}} + \frac{1}{I \cdot J} \right) \cdot \frac{1}{2}$$

where $\nabla_{kl} = p \cdot \hat{m}_{++kl} / \hat{m}_{++++}$ for $k, l = 1, 2,$ and the term $1/IJ$ is the constant prior of Clogg *et al.* (1991) to prevent possible boundary solutions for m_{ij12}, m_{ij21}, and m_{ij22} (also see the fifth prior below). Thus, we allocate the third prior of δ_{ijkl} for the MPE of m_{ijkl} to shrink toward the ML obtained under the nonignorable nonresponse, whereas the first prior is obtained under an ignorable nonresponse model.

The fourth type of prior is defined by letting $\delta_{ij11} = 0$ in (11) as we did in obtaining the second type of prior from the first type. The last type of prior is from Clogg *et al.* (1991) defined as

$$\delta_{ij11} = 0, \delta_{ij12} = \frac{p}{3} \cdot \left(\frac{1}{I \cdot J} \right), \delta_{ij21} = \frac{p}{3} \cdot \left(\frac{1}{I \cdot J} \right),$$

and (12)

$$\delta_{ij22} = \frac{p}{3} \cdot \left(\frac{1}{I \cdot J} \right).$$

These five types of priors are summarized in Table 1 and are compared in the next section using empirical data and simulation studies.

Table 1
Five types of priors δ_{ijkl} ($\hat{m}_{ijkl}$ is MLE, *I* and *J* are the numbers of row and columns in a two-way table, and *p* is the number of parameters)

	δ_{ij11}	δ_{ij12}	δ_{ij21}	δ_{ij22}	
Type I	$\nabla_{11} \frac{y_{ij11}}{y_{++11}}$	$\nabla_{12} \frac{y_{ij11}}{y_{++11}}$	$\nabla_{21} \frac{y_{ij11}}{y_{++11}}$	$\nabla_{22} \frac{y_{ij11}}{y_{++11}},$	$\nabla_{kl} = p \cdot \frac{y_{++kl}}{y_{++++}}$
Type II	0	$\nabla_{12} \frac{y_{ij11}}{y_{++11}}$	$\nabla_{21} \frac{y_{ij11}}{y_{++11}}$	$\nabla_{22} \frac{y_{ij11}}{y_{++11}},$	$\nabla_{kl} = p \cdot \frac{y_{++kl}}{y^{*}_{++++}}$
Type III	$\nabla_{11} \cdot \left(\frac{\hat{m}_{ij11}}{\hat{m}_{++11}} \right)$	$\frac{\nabla_{12}}{2} \left(\frac{\hat{m}_{ij12}}{\hat{m}_{++12}} + \frac{1}{IJ} \right)$	$\frac{\nabla_{21}}{2} \left(\frac{\hat{m}_{ij21}}{\hat{m}_{++21}} + \frac{1}{IJ} \right)$	$\frac{\nabla_{22}}{2} \left(\frac{\hat{m}_{ij22}}{\hat{m}_{++22}} + \frac{1}{IJ} \right),$	$\nabla_{kl} = p \cdot \frac{\hat{m}_{++kl}}{\hat{m}_{++++}}$
Type IV	0	$\frac{\nabla_{12}}{2} \left(\frac{\hat{m}_{ij12}}{\hat{m}_{++12}} + \frac{1}{IJ} \right)$	$\frac{\nabla_{21}}{2} \left(\frac{\hat{m}_{ij21}}{\hat{m}_{++21}} + \frac{1}{IJ} \right)$	$\frac{\nabla_{22}}{2} \left(\frac{\hat{m}_{ij22}}{\hat{m}_{++22}} + \frac{1}{IJ} \right),$	$\nabla_{kl} = p \cdot \frac{\hat{m}_{++kl}}{\hat{m}^{*}_{++++}}$
Type V	0	$\nabla_{12} \left(\frac{1}{I \cdot J} \right)$	$\nabla_{21} \left(\frac{1}{I \cdot J} \right)$	$\nabla_{22} \left(\frac{1}{I \cdot J} \right),$	$\nabla_{kl} = \frac{p}{3}$

$y^{*}_{++++} = y_{++++} - y_{++11}$ and $\hat{m}^{*}_{++++} = \hat{m}_{++++} - \hat{m}_{++11}$

Up to this point, we have presented methods for a two-way table, and y_{ijkl} is defined for the count of the (i, j) cell of the i^{th} row and j^{th} column (*i.e.*, $X_1 = i, X_2 = j$), and indicator R_1 for a missing row and R_2 for a missing column (*i.e.*, $R_1 = k, R_2 = l$). This can be easily extended to the 3-way table. Denote y_{ijklmn} to be the $(i, j, k)^{\text{th}}$ cell count for the three response variables (*i.e.*, $X_1 = i$, $X_2 = j$, and $X_3 = k$) and respective missing rows and columns (*i.e.*, $R_1 = l$, $R_2 = m$, and $R_3 = n$ for $l, m, n = 1, 2$). Thus, $lmn = 111$ implies that all of the three variables are observed, $lmn = 112$ implies that X_1 and X_2 are observed but X_3 is missing; similarly for $lmn = 121, 122, 211, 212, 221, 222$; 1 is for observed and 2 designates missing. Accordingly, the EM algorithm and priors for an incomplete three-way contingency table can be defined. The conditional expectation in the E-step for the $(i, j, k)^{\text{th}}$ cell with unknown information of k margin is

$$E_{\text{old}}(y_{ijk112} \mid \pi^{\text{old}}_{ijklmn}, y_{ij+112}) = y_{ij+112}\frac{m^{\text{old}}_{ijk112}}{m^{\text{old}}_{ij+112}}.$$

Similarly,

$$E_{\text{old}}(y_{ijk122} \mid \pi^{\text{old}}_{ijklmn}, y_{i++122}) = y_{i++122}\frac{m^{\text{old}}_{ijk122}}{m^{\text{old}}_{i++122}}.$$

and

$$E_{\text{old}}(y_{ijk222} \mid \pi^{\text{old}}_{ijklmn}, y_{+++222}) = y_{+++222}\frac{m^{\text{old}}_{ijk222}}{m^{\text{old}}_{+++222}}.$$

Other expectations and five types of priors can be similarly defined.

The Buckeye state poll is a Random Digit Dialing (RDD). No modification is necessary for the Bayesian procedures if the RDD is strictly a self-weighting survey (Lavrakas 1993; Potthoff 1994). However, RDD is not always done by a self-weighting design. For example, a telephone sample comprises a sample of households, not persons. If one person is interviewed in a household, a weight should be superimposed on the response by the number of persons in the household. A weight is also needed for the households with more than one telephone number. If an accurate estimate of the total number of households is available, stratification by region or state is possible and weighting must be considered in a comprehensive analysis. RDD was used in the 1998 Ohio election surveys. In this study, our method and models do not include weighting from stratification, clustering, and other factors leading to different probabilities of selection in a telephone survey.

However, further extension can be made for such weighting. A simple extension below shows how to accommodate a typical stratification. In a three-way table, let X_3 be the third response variable indexed by h $(h = 1, \ldots, H)$ that is assumed to be always observed. The H categories can be strata in a stratified sampling. Since X_3 is always observed, the corresponding missingness variable R_3 is equal to 1 and its observation can be denoted by y_{ijhlm1}. Then, we can write the following log likelihood for each stratum h:

$$l_h = \sum_{i=1}^{I}\sum_{j=1}^{J} y_{ijh111}\log(\pi_{ijh11}) + \sum_{i=1}^{I} y_{i+h121}\log(\pi_{i+h12}) + \sum_{j=1}^{J} y_{+jh211}\log(\pi_{+jhk21}) + y_{++h221}\log(\pi_{++h22})$$

where $\pi_{ijhlm} = P[X_1 = i, X_2 = j, R_1 = l, R_2 = m \mid X_3 = h]$. Thus, the terminology X_3 used for a three-way table acts as an indicator for strata. For each stratum h, the likelihood of (13) is exactly the same as that of a 2-way table.

Then, a log linear model for the cell expectation $m_{ijhlm} = N_h \cdot \pi_{ijhlm}$ can be defined in a similar way as in (2) where $N_h = \sum_{i,j,l,m} y_{ijhlm}$ for each $h = 1, 2, \ldots, H$. A nonignorable nonresponse model is given by

$$\log(m_{ijhlm}) = \beta_{0h} + \beta^{i}_{X_1h} + \beta^{j}_{X_2h} + \beta^{l}_{R_1h} + \beta^{m}_{R_2h} + \beta^{ij}_{X_1X_2h} + \beta^{il}_{X_1R_1h} + \beta^{jm}_{X_2R_2h}. \quad (13)$$

To avoid a boundary solution problem as in Section 2, we use the Dirichlet priors for π_{ijhlm}

$$\prod_i\prod_j \pi_{ijh11}^{\delta_{ijh11}} \cdot \pi_{ijh12}^{\delta_{ijh12}} \cdot \pi_{ijh21}^{\delta_{ijh21}} \cdot \pi_{ijh22}^{\delta_{ijh22}}.$$

Then, we follow exactly the same procedures as shown in Section 2 to estimate the cell expectations m_{ijhlm} for each $h = 1, 2\ldots, H$. The estimate of the $(i, j)^{\text{th}}$ cell expectation is

$$\hat{E}(y_{ij}) = \sum_{h=1}^{H} w_h \sum_{l,m} \hat{m}_{ijhlm}$$

where w_h is the known weight for the h^{th} stratum and $\hat{m}_{ijhlm}$ is the m_{ijhlm} evaluated at the MPE of $\boldsymbol{\beta}$. For example, $w_h = N_h / \sum_h N_h$ is for a stratified sample where N_h is the population size of the h^{th} stratum.

The variance-covariance matrix of an approximation to the distribution of $\hat{\mathbf{m}}$ is

$$\frac{\partial \hat{\mathbf{m}}^T}{\partial \boldsymbol{\beta}} \text{Var}(\hat{\boldsymbol{\beta}}_{\mathbf{MPE}}) \frac{\partial \hat{\mathbf{m}}}{\partial \boldsymbol{\beta}} \quad (14)$$

where $\hat{\mathbf{m}}$ is a vector expression of the cell estimates $\hat{m}_{ijhlm}$, $\hat{\boldsymbol{\beta}}_{\mathbf{MPE}}$ is the MPE of $\boldsymbol{\beta}$ and its variance $\text{Var}(\hat{\boldsymbol{\beta}}_{\mathbf{MPE}})$ is given by the inverse of (9), and $\partial \mathbf{m} / \partial \boldsymbol{\beta} = N_h \times [\text{diag}(\hat{\boldsymbol{\pi}}) - \hat{\boldsymbol{\pi}}\hat{\boldsymbol{\pi}}^T]\mathbf{Z}$ where $\hat{\boldsymbol{\pi}}$ has

$$\hat{\pi}_{ijhlm} = \pi_{ijhlm}(\hat{\boldsymbol{\beta}}_{\mathbf{MPE}}) = \frac{\exp(\mathbf{z}_{ijhlm}\hat{\boldsymbol{\beta}}_{\mathbf{MPE}})}{\sum_{k\in(i,j,h,l,m)}\exp(\mathbf{z}_{k}\hat{\boldsymbol{\beta}}_{\mathbf{MPE}})}$$

as its typical element.

3. An application to a Buckeye State Poll

In forecasting the winner in a poll, the accuracy of the poll often depends on how to handle undecided voters who are likely to vote but who have not yet decided their preference for a candidate. We compare the Bayesian estimates based on the five types of priors with the ML estimate using the Buckeye State Poll (BSP) conducted in 1998 by the Center for Survey Research at Ohio State University. The BSP surveys produced incomplete two-way contingency tables with one category being candidate preference and the other category being the likelihood of voting in the November 1998 races for Ohio Governor, Attorney-General, Mayor of Columbus, and Treasurer. Table 2 summarizes these four polls and shows a substantial number of undecided voters.

For comparison, we consider the following ignorable Model 1 and the two nonignorable nonresponse Model 2 and Model 3.

$$\text{Model 1:}\quad \log(m_{ijkl}) = \beta_0 + \beta^{i}_{X_1} + \beta^{j}_{X_2} + \beta^{k}_{R_1} + \beta^{l}_{R_2} + \beta^{ij}_{X_1X_2} + \beta^{kl}_{R_1R_2},$$

$$\text{Model 2:}\quad \log(m_{ijkl}) = \beta_0 + \beta^{i}_{X_1} + \beta^{j}_{X_2} + \beta^{k}_{R_1} + \beta^{l}_{R_2} + \beta^{ik}_{X_1R_1} + \beta^{jl}_{X_2R_2} + \beta^{ij}_{X_1X_2} + \beta^{kl}_{R_1R_2},$$

$$\text{Model 3:}\quad \log(m_{ijkl}) = \beta_0 + \beta^{i}_{X_1} + \beta^{j}_{X_2} + \beta^{k}_{R_1} + \beta^{l}_{R_2} + \beta^{il}_{X_1R_2} + \beta^{jk}_{X_2R_1} + \beta^{ij}_{X_1X_2} + \beta^{kl}_{R_1R_2}.$$

Model 1 is missing completely at random, and cases with missing data can be ignorable in likelihood inferences. Model 2 and Model 3 are nonignorable where the probability of missing a variable depends on itself in Model 2 while the probability in Model 3 depends on the other variable. Note that the ML estimates in Model 1 and Model 3 are not on the boundary of the parameter space as shown by Baker *et al.* (1992). Moreover, since we found that, under Model 1 and Model 3, all of the five Bayesian estimates for the expected cell counts are not only fairly close to the ML estimate and their standard deviations are almost the same, we only present the ML estimates for Model 1 and Model 3.

We denote the ML estimates under ignorable Model 1, nonignorable Model 2, and nonignorable Model 3 by $IG1_{ML}$, $NON2_{ML}$, and $NON3_{ML}$, respectively. IG and NON stand for ignorable and nonignorable, respectively. We also let $NON2^{BE}_{i}$ be the Bayesian estimator using the i^{th} type of priors under Model 2. That is, $NON2^{BE}_{1}$ uses the respondent-driven priors of (10) and $NON2^{BE}_{2}$ is the same priors as $NON2^{BE}_{1}$ except for $\delta_{ij11} = 0$. Similarly, $NON2^{BE}_{3}$ is given by (11) and $NON2^{BE}_{4}$ is the same priors except for $\delta_{ij11} = 0$. $NON2^{BE}_{5}$ is the Bayesian estimate using the constant priors of (12). In addition, we can use the Stasny method (1986, 1988) to estimate the expected cell counts under Model 1 and Model 3 that she implicitly assumed. However, her estimates appear to be exactly the same as $IG1_{ML}$.

Table 2
Observed data for BSP pre-election surveys

	Governor race				Attorney-general race		
	Fisher	**Taft**	**Others**	**Undecided**	**Montgomery**	**Cordray**	**Undecided**
Likely to vote	112	140	23	61	197	82	57
Unlikely to vote	96	108	21	73	161	65	75
Undecided	7	11	1	4	15	4	0
	Mayor race				**Treasurer race**		
	Coleman	**Teater**	**Espy**	**Undecided**	**Deters**	**Donofrio**	**Undecided**
Likely to vote	40	32	25	30	127	119	90
Unlikely to vote	37	47	41	56	127	90	84
Undecided	0	2	1	0	10	7	0

The top table in Table 3 shows predicted values of elections using only “likely to vote” for the four races and their standard deviations in parentheses. The standard deviations are close to each other and show significant differences between the first and second leading candidates, except in the race for Mayor. This table also includes the actual election results and shows whether or not the ML estimates fall into the boundary solutions.

The bottom table shows the predictions of elections using both “likely to vote” and “unlikely to vote” to see what happens if those who responded to “unlikely to vote” actually voted. Comparing the two tables, we may conclude that the winners for Governor, Attorney-General, and the Treasurer’s elections remained unchanged regardless of the likelihood of voting, whereas the winner could have changed in the Mayor’s election if most of those who were “unlikely to vote” actually voted.

Based on Table 3, we can classify the 7 estimates, except $NON2_{ML}$, into two groups: $NON2_3^{BE}$, $NON2_4^{BE}$, and $NON2_5^{BE}$ to the first group and the remaining four estimates, $NON2_1^{BE}$, $NON2_2^{BE}$, $IG1_{ML}$, and $NON3_{ML}$ to the second group. As expected, since the priors δ_{ijkl} for $NON2_1^{BE}$ and $NON2_2^{BE}$ are so defined that the estimate of m_{ijkl} shrinks toward the ML under an ignorable nonresponse model, these two Bayesian estimates are very close to $IG1_{ML}$ and hence have little advantage over the $IG1_{ML}$. It is also interesting to note that $NON3_{ML}$ is almost the same as $IG1_{ML}$ although their loglinear models are differently specified.

There is no general criterion to evaluate whether an ignorable nonresponse model or a nonignorable nonresponse model is appropriate. However, as stated in Chen and Stasny (2003), the assumption of nonignorability for a nonresponse may be a reasonable assumption in the Buckeye State Poll study because people might be reluctant to express their preference for an unpopular candidate, or their current preferences are not firm or accurate at the time of the interview. In this regard, the $NON2_1^{BE}$, $NON2_2^{BE}$, and $NON3_{ML}$ may not be appropriate in these particular case studies because they are almost the same as the $IG1_{ML}$ of Model 1.

Table 3
Prediction of elections based on the October 98 and April 98 Buckeye State Polls (the unit is % and the numbers in parentheses are standard deviations)

	Governor			Mayor			Attorney-General		Treasurer	
	Fisher	Taft	Others	Coleman	Teater	Espy	Mongomery	Cordray	Deters	Donofrio
Likely to vote only used										
$NON2_{ML}$	33.2(2.75)	42.1(3.00)	24.8	31.5(4.65)	25.3(4.23)	43.2	75.6(3.71)	24.4	57.0(3.48)	43.0
$NON2_1^{BE}$	40.6(3.04)	48.5(3.27)	10.9	38.1(5.14)	34.2(4.78)	27.7	72.1(3.61)	27.9	52.7(3.36)	47.3
$NON2_2^{BE}$	40.9(3.01)	50.7(3.20)	8.40	39.9(5.04)	33.6(4.83)	26.5	71.0(3.59)	29.0	52.1(3.34)	47.9
$NON2_3^{BE}$	35.8(2.85)	44.5(3.08)	19.7	35.6(4.87)	29.3(4.51)	35.1	63.0(3.67)	37.0	54.3(3.41)	45.7
$NON2_4^{BE}$	36.3(2.87)	45.2(3.11)	18.6	35.9(4.91)	29.4(4.52)	34.6	63.0(3.64)	37.0	53.9(3.40)	46.1
$NON2_5^{BE}$	38.9(2.99)	47.4(3.20)	13.7	37.7(4.99)	33.6(4.77)	28.7	66.0(3.54)	34.0	51.5(3.32)	48.5
$IG1_{ML}$	40.6(3.03)	51.2(3.28)	8.20	40.8(5.16)	33.4(4.76)	25.8	70.9(3.59)	29.1	51.8(3.32)	48.2
$NON3_{ML}$	40.6(3.03)	51.2(3.28)	8.20	40.9(5.16)	33.3(4.75)	25.8	70.9(3.58)	29.1	51.7(3.32)	48.3
Actual result	45	50	5	39	37	24	63	37	57	43
Boundary		yes			yes		yes		no	
Likely to vote + Unlikely to vote										
$NON2_{ML}$	32.7(1.83)	39.4(1.91)	27.8	24.8(2.45)	26.2(2.49)	49.0	77.0(1.64)	23.0	60.2(1.93)	39.8
$NON2_1^{BE}$	41.3(1.93)	46.4(1.96)	12.3	30.7(2.68)	37.1(2.75)	32.2	72.8(1.74)	27.2	56.0(1.96)	44.0
$NON2_2^{BE}$	41.9(1.93)	49.2(1.95)	8.90	32.7(2.63)	36.5(2.76)	30.8	71.4(1.77)	28.6	55.3(1.96)	44.7
$NON2_3^{BE}$	35.4(1.87)	41.8(1.93)	22.7	27.8(2.55)	30.5(2.62)	41.7	61.0(1.72)	39.0	57.6(1.95)	42.4
$NON2_4^{BE}$	36.0(1.88)	42.6(1.93)	21.4	28.7(2.57)	30.6(2.62)	40.7	60.9(1.75)	39.1	57.2(1.95)	42.8
$NON2_5^{BE}$	39.1(1.91)	45.1(1.95)	15.8	30.7(2.63)	35.8(2.74)	33.5	64.8(1.88)	35.2	54.8(1.96)	45.2
$IG1_{ML}$	41.5(1.96)	49.8(1.96)	8.70	33.9(2.70)	36.1(2.74)	29.9	71.2(1.78)	28.8	55.0(1.96)	45.0
$NON3_{ML}$	41.5(1.96)	49.8(1.96)	8.70	34.1(2.71)	36.0(2.74)	29.9	71.1(1.78)	28.9	55.0(1.96)	45.0

Compared to actual election results, $NON2_{ML}$ gives the worst prediction for Governor, Mayor, and Attorney-General because the $NON2_{ML}$ lies on a boundary solution; whereas it provides the best prediction for Treasurer because it does not lie on a boundary solution. In the Attorney-General's election, $NON2_3^{BE}$ and $NON2_4^{BE}$ not only predicted the exact actual result but also are quite different from the other estimates. Since $NON2_3^{BE}$ and $NON2_4^{BE}$ have the priors to reflect different response patterns between respondents and the undecided, we can infer that the undecided voters in the Attorney-General race have quite different preference for the candidate from the respondents (*i.e.*, $NON2_3^{BE}$ and $NON2_4^{BE}$ allocate 19.4 % of the undecided voters who are likely to vote for Montgomery and 80.6% for Cordray, whereas the data in Table 2 indicates the percentage of Montgomery *vs* Cordray is 29.4% *vs* 70.6% among respondents who are likely to vote).

To see this difference between the respondents and undecided voters in terms of parameter estimates and to examine the effect of occurrence of the boundary solution on the estimates under the nonignorable Model 2, we present the ML estimates and $NON2_3^{BE}$ estimates and their corresponding standard deviations for the Attorney-General race in Table 4. Because of a boundary solution, all of the ML estimates have too large standard deviations as expected. On the other hand, $NON2_3^{BE}$ is very stable. Since $\beta^{11}_{X_1X_2} = 0.0472$ is the smallest and its standard deviation is relatively large, we neglect $\beta^{11}_{X_1X_2}$ to avoid complexity of interpretation. Under $\beta^{11}_{X_1X_2} = 0,$ it is not difficult to show that, using the estimates of $NON2_3^{BE}$ in Table 4,

$$\log\frac{m_{1j1l}}{m_{2j1l}} = 2(\beta^1_{X_1} + \beta^{11}_{X_1R_1}) = 0.09$$

and

$$\log\frac{m_{1j2l}}{m_{2j2l}} = 2(\beta^1_{X_1} - \beta^{11}_{X_1R_1}) = 1.3916$$

for each fixed j and $l,$ and

$$\log\frac{m_{i1k1}}{m_{i2k1}} = 2(\beta^1_{X_2} + \beta^{11}_{X_2R_2}) = 0.8982$$

and

$$\log\frac{m_{i1k2}}{m_{i2k2}} = 2(\beta^1_{X_2} - \beta^{11}_{X_2R_2}) = -1.4942$$

for each fixed i and $k.$ Thus, by

$$\log\frac{m_{1j1l}}{m_{2j1l}} = 2(\beta^1_{X_1} + \beta^{11}_{X_1R_1}) = 0.09,$$

those who are likely to vote (*i.e.*, $i = 1$) are 1.09 times (*i.e.*, $e^{0.09}$) more than those who are unlikely to vote (*i.e.*, $i = 2$) among respondents ($k = 1$), whereas, by

$$\log\frac{m_{1j2l}}{m_{2j2l}} = 2(\beta^1_{X_1} - \beta^{11}_{X_1R_1}) = 1.3916,$$

likely voters of $i = 1$ are 4.02 times (*i.e.*, $e^{1.3916}$) more than unlikely voters of $i = 2$ among undecided ($k = 2$); by

$$\log\frac{m_{i1k1}}{m_{i2k1}} = 2(\beta^1_{X_2} + \beta^{11}_{X_2R_2}) = 0.8982,$$

those who vote for Montgomery are 2.46 times more than those who vote for Cordray among respondents; whereas, by

$$\log\frac{m_{i1k2}}{m_{i2k2}} = 2(\beta^1_{X_2} - \beta^{11}_{X_2R_2}) = -1.4942,$$

unlikely voters are 4.46 times more than likely voters among the undecided. This implies that the response pattern is much different between respondents and the undecided.

Table 4
ML and the third type Bayesian Estimates under nonignorable Model 2 for Attorney-General (the standard deviations are in parentheses)

	β_0	$\beta^1_{X_1}$	$\beta^1_{X_2}$	$\beta^1_{R_1}$	$\beta^1_{R_2}$	$\beta^{11}_{X_1R_1}$	$\beta^{11}_{X_2R_2}$	$\beta^{11}_{X_1X_2}$	$\beta^{11}_{R_1R_2}$
$NON2_{ML}$	-3.3735	-1.9487	3.2134	4.8496	4.8186	2.0283	-2.7594	-0.0452	-1.5588
		(3.120)	(8.515)	(3.996)	(8.871)	(3.120)	(8.512)	(0.045)	(2.501)
$NON2_3^{BE}$	0.6860	0.3704	-0.1490	3.3024	2.2942	-0.3254	0.5981	0.0472	-1.5450
		(0.118)	(0.052)	(2.501)	(2.501)	(0.117)	(0.052)	(0.041)	(2.501)

The extent of this difference can be measured by the most important terms, $\beta^{11}_{X_1R_1}$ and $\beta^{11}_{X_2R_2}$, in the nonignorable nonresponse model, Model 2. Since

$$\beta^{11}_{X_1R_1} = \frac{1}{4}\log\frac{m_{1111}/m_{2111}}{m_{1121}/m_{2121}} = -0.3254$$

and

$$\beta^{11}_{X_2R_2} = \frac{1}{4}\log\frac{m_{1111}/m_{1211}}{m_{1112}/m_{1212}} = 0.5981,\ \beta^{11}_{X_1R_1}$$

is the log-odds ratio that shows the log difference between the ratio of the number of those "likely to vote" to that of those "unlikely to vote" among the decided voters for Montgomery and the same ratio among the undecided voters who prefer Montgomery but who do not express their likelihood of voting. Whereas, $\beta^{11}_{X_2R_2}$ is the log-odds ratio that shows the log difference between the ratio of the number of voters for Montgomery to the voters for Cordray among the decided who are likely to vote and the same ratio among the undecided voters who are likely to vote but who do not express their candidate preference. Thus, among voters for Montgomery, the possibility for the undecided voters to vote relative to not voting is about 3.67 times

$$\left(i.e.,\ \frac{m_{1111}/m_{2111}}{m_{1121}/m_{2121}} = e^{4\times -0.3254} = 3.67^{-1}\right)$$

larger than the possibility of the decided, implying that Montgomery needs a strategy to raise the turnout of voters. On the other hand, among those likely to vote, the supporting rate of the decided for Montgomery is about 10.94 times

$$\left(i.e.,\ \frac{m_{1111}/m_{1211}}{m_{1112}/m_{1212}} = e^{4\times 0.5981} = 10.94\right)$$

larger than the undecided voters for Montgomery, implying that most of the undecided voters not exposing their preference of candidate are likely to vote for Cordray as the Attorney-General. This also confirms the popular account that voters are inclined to remain "undecided" in a poll if they support the candidate who is seen as inferior in a race and that the voters are inclined to abstain from voting if they support the candidate who certainly dominates the race.

4. Simulation study

We consider a 2×2 contingency table with supplemental margins to compare the performance of the five Bayesian estimates described in Section 2 for different missing percentages and different response patterns under the following nonignorable nonresponse model (*i.e.*, Model 2):

$$\log(m_{ijkl}) = \beta_0 + \beta^i_{X_1} + \beta^j_{X_2} + \beta^k_{R_1} + \beta^l_{R_2} + \beta^{ik}_{X_1R_1} + \beta^{jl}_{X_2R_2} + \beta^{ij}_{X_1X_2} + \beta^{kl}_{R_1R_2}.$$

Thus, we only compare $NON2_{ML}$ and $NON2^{BE}_i$ for $i = 1, ..., 5$ in this simulation study.

Since there are two levels in all of X_1, X_2, R_1, and R_2, there are 8 parameters to be determined for the simulation study. From the equations of

$$4\beta^{11}_{X_1R_1} = \log\frac{m_{1111}/m_{2111}}{m_{1121}/m_{2121}} \text{ and } 4\beta^{11}_{X_2R_2} = \log\frac{m_{1111}/m_{1211}}{m_{1112}/m_{1212}},$$

$$\beta^{11}_{X_1R_1} = \beta^{11}_{X_2R_2} = 0$$

means that there is no difference in the response pattern between respondents and undecided. The bigger $\beta^{11}_{X_1R_1}$ and $\beta^{11}_{X_2R_2}$ are, the more different the response pattern between respondents and undecided voters is. We vary these two parameters from 0.2 to 0.8 with an increment of 0.2. We set the missing percentage to 20% and 30% by adjusting $\beta^1_{X_1}$ and $\beta^1_{R_1}$ and fixing

$$\frac{m_{1111}/m_{1211}}{m_{2111}/m_{2211}} = 5,\ \frac{m_{1111}/m_{1112}}{m_{1112}/m_{1122}} = 2,$$

and

$$N = \textstyle\sum_{ijkl} m_{ijkl} = 1{,}000.$$

This implies that the size and missing percentage for the cell of $X_1 = 1$ and $X_2 = 1$ are approximately 5 times and 2 times the size of the other three cells, respectively.

We generate a large number of samples $\{y_{ijkl}, i, j, k, l = 1, 2\}$ from the above setting until we have 1,000 random samples with boundary solutions and the other 1,000 with no boundary solutions. The occurrence of a boundary solution is determined by the criterion given in Michiels and Molenberghs (1997) (also see Clarke (2002), Smith *et al.* (1999) for more details). Using $\{y_{ij11}, y_{i+12}, y_{+j21}, y_{++22}, i, j, = 1, 2\}$ obtained from the generated data, the expected cell counts m_{ijkl}'s are estimated by each of the five Bayesian estimates and the ML estimate described in Section 2.

We calculate mean squared errors (MSEs) and absolute biases of $NON2_{ML}$, $NON2^{BE}_1$, ..., $NON2^{BE}_5$ for $\{\sum_{kl} m_{ijkl}, i, j = 1, 2\}$. Then we take the mean over the four MSEs and the four absolute biases, which we obtain from each estimate to see the overall performance of the estimate. Similarly, we calculate mean MSEs and mean absolute biases for $\{m_{ij12} + m_{ij21} + m_{ij22}, i, j = 1, 2\}$ to see the performance of each estimate in imputing the nonresponses.

Table 5 shows the ratios of the mean MSEs and mean absolute biases of the five Bayesian estimates (*i.e.*, $NON2_1^{BE}$, ..., $NON2_5^{BE}$), relative to the ML estimate (*i.e.*, $NON2_{ML}$) when the boundary solutions occur; whereas Table 5 shows the ratios when no boundary occurs. Thus, values less than 1 imply that the corresponding Bayesian estimate has a smaller mean MSE or a smaller mean absolute bias than the ML estimate. Both tables only show the cases for $\beta^{11}_{X_1R_1} < \beta^{11}_{X_2R_2}$ and for 20% of the missing percentage because the MSEs and biases are almost symmetric about the coordinate of $(\beta^{11}_{X_1R_1}, \beta^{11}_{X_2R_2})$. They increase as we increase the missing percentage to 30% while keeping the same patterns of the MSEs and biases as those of the missing 20%.

Table 5, where a boundary solution occurs, shows that $NON2_1^{BE}$, $NON2_3^{BE}$, $NON2_4^{BE}$ have smaller MSEs than the ML estimate (*i.e.*, $NON2_{ML}$) for all values of $\beta^{11}_{X_1R_1}$ and $\beta^{11}_{X_2R_2}$, except $(\beta^{11}_{X_1R_1}, \beta^{11}_{X_2R_2}) = (0.8, 0.8)$. Here, $NON2_3^{BE}$ has a smaller MSE than the ML estimate. This is true for the absolute biases. On the other hand, Table 6, where no boundary solution occurs, shows that only $NON2_3^{BE}$ is comparable to the ML estimate in the MSE although it is slightly biased. In particular, $NON2_3^{BE}$ has a smaller MSE than the ML estimate as long as $\beta^{11}_{X_1R_1} \neq 0.8$ or $\beta^{11}_{X_2R_2} \neq 0.8$ (*i.e.*, The response pattern between respondents and nonrespondents is not very different.).

Table 5
Ratios of mean MSEs and mean absolute biases of Bayesian estimates relative to the ML estimate when boundary solutions occur under a 20% missing percentage (the ratios for absolute biases are in parentheses)

	$(\beta^{11}_{X_1R_1}, \beta^{11}_{X_2R_2})$	$NON2_1^{BE}$	$NON2_2^{BE}$	$NON2_3^{BE}$	$NON2_4^{BE}$	$NON2_5^{BE}$
For $\{m_{ij11} + m_{ij12} + m_{ij21} + m_{ij22},\ i,\ j = 1,\ 2\}$	(0.2, 0.2)	0.68(0.66)	0.47(0.22)	0.76(0.76)	0.65(0.48)	0.42(0.05)
	(0.2, 0.4)	0.68(0.48)	0.57(0.20)	0.77(0.68)	0.60(0.29)	0.56(0.30)
	(0.2, 0.6)	0.67(0.23)	0.73(0.66)	0.77(0.57)	0.64(0.10)	0.69(0.64)
	(0.2, 0.8)	0.77(0.26)	1.08(1.55)	0.83(0.43)	0.76(0.28)	0.95(1.34)
	(0.4, 0.4)	0.65(0.32)	0.69(0.57)	0.76(0.63)	0.61(0.17)	0.65(0.52)
	(0.4, 0.6)	0.58(0.14)	0.83(0.90)	0.71(0.56)	0.56(0.06)	0.69(0.71)
	(0.4, 0.8)	0.75(0.36)	1.46(2.07)	0.78(0.36)	0.74(0.42)	1.12(1.61)
	(0.6, 0.6)	0.66(0.22)	1.35(1.73)	0.73(0.43)	0.66(0.16)	1.01(1.29)
	(0.6, 0.8)	0.85(0.87)	2.27(3.19)	0.76(0.17)	0.83(0.81)	1.52(2.35)
	(0.8, 0.8)	1.12(1.93)	3.58(5.49)	0.83(0.24)	1.04(1.67)	2.18(3.95)
For $\{m_{ij12} + m_{ij21} + m_{ij22},\ i,\ j = 1,\ 2\}$	(0.2, 0.2)	0.57(0.63)	0.27(0.13)	0.69(0.74)	0.41(0.40)	0.28(0.31)
	(0.2, 0.4)	0.54(0.46)	0.37(0.34)	0.68(0.68)	0.42(0.24)	0.44(0.57)
	(0.2, 0.6)	0.51(0.19)	0.69(0.94)	0.65(0.55)	0.47(0.10)	0.69(0.88)
	(0.2, 0.8)	0.63(0.35)	1.39(2.08)	0.71(0.34)	0.62(0.47)	1.11(1.52)
	(0.4, 0.4)	0.49(0.35)	0.54(0.64)	0.65(0.64)	0.42(0.17)	0.57(0.76)
	(0.4, 0.6)	0.48(0.17)	0.98(1.24)	0.62(0.51)	0.45(0.17)	0.85(1.04)
	(0.4, 0.8)	0.62(0.44)	1.81(2.33)	0.67(0.35)	0.61(0.55)	1.35(1.81)
	(0.6, 0.6)	0.55(0.42)	1.70(1.90)	0.63(0.41)	0.54(0.40)	1.28(1.51)
	(0.6, 0.8)	0.78(0.92)	2.91(3.43)	0.69(0.14)	0.75(0.92)	1.96(2.64)
	(0.8, 0.8)	1.13(1.96)	4.63(5.72)	0.75(0.33)	1.02(1.77)	2.86(4.24)

Table 6
Ratios of mean MSEs and mean absolute biases of Bayesian estimates relative to the ML estimate when no boundary solution occurs under a 20% missing percentage (the ratios for absolute biases are in parentheses)

	$(\beta^{11}_{X_1R_1}, \beta^{11}_{X_2R_2})$	$NON2^{BE}_1$	$NON2^{BE}_2$	$NON2^{BE}_3$	$NON2^{BE}_4$	$NON2^{BE}_5$
For $\{m_{ij11} + m_{ij12} + m_{ij21} + m_{ij22},\ i,\ j = 1,\ 2\}$	(0.2, 0.2)	0.99(3.37)	1.05(7.00)	0.94(2.51)	0.93(4.89)	1.06(8.96)
	(0.2, 0.4)	0.98(2.57)	1.21(5.13)	0.97(1.89)	1.00(3.26)	1.24(5.56)
	(0.2, 0.6)	1.04(2.18)	1.52(3.84)	0.95(1.67)	1.06(2.38)	1.43(3.71)
	(0.2, 0.8)	1.12(2.04)	1.75(3.53)	1.00(1.48)	1.13(2.14)	1.52(3.21)
	(0.4, 0.4)	1.03(2.40)	1.49(4.66)	0.97(1.69)	1.05(2.74)	1.39(4.46)
	(0.4, 0.6)	1.20(2.17)	2.11(3.85)	1.00(1.52)	1.22(2.24)	1.78(3.42)
	(0.4, 0.8)	1.28(2.09)	2.36(3.67)	1.05(1.45)	1.26(2.09)	1.86(3.12)
	(0.6, 0.6)	1.22(2.16)	2.49(3.90)	0.96(1.48)	1.21(2.15)	1.90(3.32)
	(0.6, 0.8)	1.52(1.99)	3.19(3.39)	1.11(1.38)	1.45(1.91)	2.29(2.77)
	(0.8, 0.8)	1.66(1.96)	3.64(3.27)	1.14(1.36)	1.52(1.83)	2.43(2.59)
For $\{m_{ij12} + m_{ij21} + m_{ij22},\ i,\ j = 1,\ 2\}$	(0.2, 0.2)	0.88(2.59)	0.89(5.66)	0.87(2.26)	0.89(4.55)	1.21(8.69)
	(0.2, 0.4)	0.93(2.40)	1.27(4.86)	0.93(1.78)	1.00(3.08)	1.50(5.29)
	(0.2, 0.6)	1.09(2.11)	1.93(3.97)	0.98(1.40)	1.15(2.29)	1.85(3.61)
	(0.2, 0.8)	1.24(2.13)	2.36(3.90)	1.02(1.48)	1.27(2.18)	2.06(3.19)
	(0.4, 0.4)	1.03(2.18)	1.81(4.30)	0.96(1.60)	1.12(2.62)	1.85(4.39)
	(0.4, 0.6)	1.23(2.28)	2.62(4.28)	0.99(1.48)	1.29(2.42)	2.28(3.80)
	(0.4, 0.8)	1.42(2.05)	3.26(3.70)	1.07(1.42)	1.44(2.07)	2.53(3.09)
	(0.6, 0.6)	1.33(2.07)	3.22(3.95)	0.99(1.36)	1.36(2.14)	2.54(3.43)
	(0.6, 0.8)	1.65(2.09)	4.14(3.74)	1.13(1.43)	1.61(2.07)	2.98(3.13)
	(0.8, 0.8)	1.91(2.02)	4.48(3.50)	1.16(1.39)	1.66(1.93)	3.03(2.83)

Park and Brown (1994) used $NON2^{BE}_2$ to estimate expected cell counts in an incomplete one-way table under a nonignorable nonresponse mechanism. They showed by simulation studies that $NON2^{BE}_2$ has a smaller MSE than the ML estimate although it is biased more than the ML. However, larger values than 1 for $NON2^{BE}_2$ in Table 5 and Table 6 indicate that this is not true in an incomplete two-way table regardless of the boundary solution and that Bayesian methods are not always better than the ML even when a boundary solution occurs. A reason that our simulation results differ from those of Park and Brown (1994) when a boundary solution occurs is attributed to the choice of $(\beta^{11}_{X_1R_1}, \beta^{11}_{X_2R_2})$ where Park and Brown performed their simulation only for $\beta^{11}_{X_1R_1} = \beta^{11}_{X_2R_2} = 0.34$. As shown in Table 5, $NON2^{BE}_2$ is better than the ML when $\beta^{11}_{X_1R_1} \leq 0.4$ and $\beta^{11}_{X_2R_2} \leq 0.4$, whereas $NON2^{BE}_2$ is worse than the ML when the response pattern between respondents and nonrespondents is much different (*i.e.*, $\beta^{11}_{X_1R_1} \geq 0.6$ or $\beta^{11}_{X_2R_2} \geq 0.6$).

Table 7 provides the mean of the standard deviations and the 95% coverage probabilities for $\beta^{11}_{X_1R_1}$. Here, we used the variance formula given in (9) to calculate the standard deviations and the 95% coverage probabilities are the coverage rates for nominal 95% confidence intervals. When a boundary solution occurs, although the coverage probability of the ML estimate is closest to the 95% nominal coverage level, the ML estimate has too large a standard deviation to use in practice. Such large standard deviations are due to the boundary problem of the ML estimate. The coverage probabilities of $NON2^{BE}_3$ are the closest to the 95% nominal coverage level among the Bayesian estimates, while those of the other Bayesian estimates are generally smaller than the 95% nominal coverage level. This implies that the Bayesian estimates other than $NON2^{BE}_3$ underestimate their standard deviations.

When no boundary solution occurs (the second table in Table 7), the standard deviations of the ML estimate are much more stable, compared to those for the boundary solution case. The coverage probability decreases as $\beta^{11}_{X_1R_1}$ and $\beta^{11}_{X_2R_2}$ increase. In particular, the coverage probabilities of NON^{BE}_1, NON^{BE}_2, and NON^{BE}_5 are seriously smaller than the 95% nominal coverage level when the response pattern between the respondents and undecided voters is much different (*i.e.*, $\beta^{11}_{X_1R_1} \geq 0.6$ and $\beta^{11}_{X_2R_2} \geq 0.6$).

Table 7
Mean of standard deviations and 95% coverage probabilities (in parentheses) for $\beta^{11}_{X_1R_1}$

	$(\beta^{11}_{X_1R_1}, \beta^{11}_{X_2R_2})$	$NON2_{ML}$	$NON2^{BE}_1$	$NON2^{BE}_2$	$NON2^{BE}_3$	$NON2^{BE}_4$	$NON2^{BE}_5$
boundary	(0.2, 0.2)	89.5(0.974)	0.082(0.978)	0.064(0.978)	0.093(0.973)	0.071(0.972)	0.060(0.957)
	(0.2, 0.4)	158.3(0.959)	0.072(0.963)	0.096(0.963)	0.079(0.958)	0.066(0.958)	0.058(0.940)
	(0.2, 0.6)	135.3(0.940)	0.065(0.941)	0.057(0.941)	0.071(0.941)	0.062(0.939)	0.056(0.922)
	(0.2, 0.8)	57.4(0.930)	0.070(0.938)	0.061(0.935)	0.076(0.928)	0.066(0.928)	0.060(0.908)
	(0.4, 0.4)	153.4(0.961)	0.079(0.920)	0.061(0.913)	0.096(0.956)	0.072(0.949)	0.060(0.911)
	(0.4, 0.6)	82.2(0.955)	0.072(0.893)	0.059(0.883)	0.086(0.951)	0.069(0.940)	0.058(0.874)
	(0.4, 0.8)	51.2(0.933)	0.071(0.862)	0.059(0.849)	0.084(0.926)	0.068(0.917)	0.059(0.846)
	(0.6, 0.6)	175.5(0.946)	0.077(0.820)	0.060(0.781)	0.101(0.943)	0.074(0.921)	0.061(0.823)
	(0.6, 0.8)	159.6(0.924)	0.071(0.728)	0.057(0.657)	0.089(0.913)	0.069(0.880)	0.058(0.737)
	(0.8, 0.8)	72.8(0.920)	0.070(0.572)	0.056(0.330)	0.093(0.900)	0.070(0.842)	0.058(0.607)
no-boundary	(0.2, 0.2)	0.068(0.949)	0.060(0.959)	0.056(0.959)	0.062(0.937)	0.058(0.935)	0.055(0.922)
	(0.2, 0.4)	0.066(0.960)	0.060(0.970)	0.056(0.970)	0.061(0.935)	0.058(0.931)	0.055(0.951)
	(0.2, 0.6)	0.064(0.940)	0.058(0.945)	0.055(0.945)	0.059(0.959)	0.057(0.919)	0.054(0.909)
	(0.2, 0.8)	0.069(0.933)	0.063(0.944)	0.059(0.941)	0.065(0.926)	0.062(0.925)	0.058(0.920)
	(0.4, 0.4)	0.074(0.910)	0.061(0.836)	0.055(0.828)	0.064(0.899)	0.059(0.884)	0.055(0.824)
	(0.4, 0.6)	0.074(0.915)	0.060(0.815)	0.055(0.806)	0.064(0.922)	0.059(0.879)	0.055(0.792)
	(0.4, 0.8)	0.073(0.891)	0.061(0.786)	0.056(0.771)	0.064(0.873)	0.060(0.852)	0.056(0.763)
	(0.6, 0.6)	0.078(0.859)	0.061(0.567)	0.055(0.470)	0.067(0.853)	0.061(0.795)	0.056(0.572)
	(0.6, 0.8)	0.076(0.843)	0.060(0.515)	0.054(0.402)	0.065(0.817)	0.060(0.767)	0.055(0.556)
	(0.8, 0.8)	0.080(0.755)	0.059(0.110)	0.053(0.017)	0.065(0.728)	0.059(0.607)	0.055(0.158)

5. Concluding remarks

We investigated the Bayesian analysis for incomplete two-way contingency tables with nonignorable non-response. In this situation, the ML estimates often fall on the boundary solution. These boundary solutions can yield $G^2 > 0$ even for a saturated model (Baker *et al.* 1992; Park and Brown 1994). This means that the G^2 may not be appropriate as a statistic for model specification. To avoid the boundary solution problem and to obtain a statistic such as a Bayes factor for model specification regardless of a boundary solution, we proposed Bayesian estimation methods using five different priors. Two of them are new and the remaining three have been previously used for analyzing an incomplete one-way table. These two new priors accommodate different response patterns between respondents and nonrespondents.

Data analysis shows that these new two priors are more reasonable in the sense that they accommodate the nonignorable nonresponse mechanism better and produce estimates close to the actual results. Moreover, with the previous three priors, our simulation study shows that the Bayesian estimates can have larger MSEs than those of the ML estimates for a contingency table with no boundary solution and a boundary solution as well, contrary to the previous studies. However, when a boundary solution occurs, the two new priors perform better than the previous three priors and the ML estimates in the sense that they have generally smaller MSEs, smaller biases, and coverage probabilities closer to the nominal coverage level.

We have briefly discussed the weighting issues at Section 2.2. However, these issues need much more rigorous discussion than we did in that section. Our discussion can be further extended to include not only different weights but also response biases and other sources of biases and variations. These problems can be carefully developed on an extended paper at a later time.

Acknowledgements

This research was supported by a Korea University Grant (K0822301).

References

Agresti, A. (2002). *Categorical Data Analysis.* 2nd Edition. New York: John Wiley & Sons, Inc.

Baker, S.G., and Laird, N.M. (1988). Regression analysis for categorical variables with outcome subject to nonignorable nonresponse. *Journal of the American Statistical Association*, 83, 62-69.

Baker, S.G., Rosenberger, W.F. and Dersimonian, R. (1992). Closed-form estimates for missing counts in two-way contingency tables. *Statistics in Medicine*, 11, 643-657.

Chen, T. (1972). Mixed-up frequencies and missing data in contingency tables. Unpublished Ph.D. dissertation, University of Chicago, Dept. of Statistics.

Chen, Q.L., and Stasny, E.A. (2003). Handling undecided voters: Using missing data methods in election forecasting. *Technical Report*, Department of Statistics, The Ohio State University.

Clarke, P.S. (2002). On boundary solutions and identifiability in categorical regression with non-ignorable non-response. *Biometrical Journal*, 44, 701-717.

Clogg, C.C., Rubin, D.B., Schenker, N. and Schultz, B. (1991). Multiple imputation of industry and occupation codes in Census Public use-samples using Bayesian logistic regression. *Journal of the American Statistical Association*, 86, 68-78.

De Heer W. (1999). International response trends of an international survey. *Journal of Official Statistics*, 15, 129-142.

Dempster, A.P., Laird, N.M. and Rubin, D.M. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, Series B, 39, 1-38.

Flannelly, K.J., Flannelly, L.T. and McLeod, M.S. Jr. (2000). Reducing undecided voters and other sources of error in election surveys. *International Journal of Market Research*, 42, 231-237.

Fenwick, I, Wiseman, F, Becker, J.F. and Heiman, J.R. (1982). Classifying undecided voters in pre-election polls. *Public Opinion Quarterly*, 46, 383-391.

Forster, J.J., and Smith, R.W.F. (1998). Model-based inference for categorical survey data subject to non-ignorable non-response. *Journal of the Royal Statistical Society B*, 60, 57-70.

Gelman, A., Carlin, J.P., Stern, H.S. and Rubin, D.B. (2004). *Bayesian Data Analysis*. 2nd Edition. New York: Chapman and Hall/CRC.

Groves, R.M., and Couper, M.P. (1998). *Nonresponse in Household Interview Surveys*. New York: John Wiley & Sons, Inc.

Kim, T. (1995). Discriminant analysis as a prediction tool for uncommitted voters in pre-election polls. *International Journal of Public Opinion Research*, 7, 110-127.

Lau, R.R. (1994). An analysis of the accuracy of “trial heat” polls during the 1992 presidential elections. *Public Opinion Quarterly*, 59, 589-605.

Lavrakas, P.J. (1993). *Telephone Survey Method: Sampling, selection, and supervision*. 2nd Edition. Newbury Park, Calif.: Sage.

Little, J.A., and Rubin, D.B. (2002). *Statistical analysis with missing data*. 2nd Edition. New York: John Wiley & Sons, Inc.

Martin, E.A., Traugott, M.W. and Kennedy, C. (2005). A review and proposal for a new measure of poll accuracy. *The Public Opinion Quarterly*, 69, 342-369.

Michiels, B., and Molenberghs, G. (1997). Protective estimation of longitudinal categorical data with nonrandom drop-out. *Communications in Statistics: Theory and Methods*, 26, 65-94.

Molenberghs, G., Kenward, M.G. and Goetghebeur, E. (2001). Sensitivity analysis for incomplete contingency tables: The Slovenian plebiscite case. *Applied Statistics*, 50, 15-29.

Monterola, C., Lim, M., Garcia, F. and Saloma, C. (2001). Feasibility of a neural network as classifier of undecided respondents in a public opinion survey. *International Journal of Public Opinion Research*, 14, 222-299.

Myers, D.J., and O’Connor, R.E. (1983). The undecided respondents in mandatory voting settings: A Venezuelan exploration. *The Western Political Quarterly*, 36, 420-433.

Park, T., and Brown, M.B. (1994). Models for categorical data with nonignorable nonresponse. *Journal of the American Statistical Association*, 89, 44-52.

Park, T. (1998). An approach to categorical data with nonignorable nonresponse. *Biometrics*, 54, 1579-1690.

Perry, P. (1979). Certain problem in election survey methodology. *Public Opinion Quarterly*, 43, 312-325.

Potthoff, R.F. (1994). Telephone sampling in epidemiologic research: To reap the benefits, avoid the pitfalls. *American Journal of Epidemiology*, 139, 967-978.

Rubin, D.B., Stern, H.S. and Vehovar, V. (1995). Handling “Don’t Know” survey responses: the case of the Slovenian plebiscite. *Journal of the American Statistical Association*, 90, 822-828.

Smith, P.W.F., Skinner, C.J. and Clarke, P.S. (1999). Allowing for non-ignorable nonresponse in the analysis of voting intention data. *Applied Statistics*, 48, 563-577.

Smith, T.W. (1984). Non attitudes: A review and evaluation. In *Surveying Subjective Phenomena*, (Eds. C.F. Turner and E. Martin), New York: Russell Sage Foundation, 2, 215-255.

Stasny, E.A. (1986). Estimating gross flow using panel data with nonresponse: An example from the Canadian Labor Force survey. *Journal of the American Statistical Association*, 81, 42-47.

Stasny, E.A. (1988). Modelling nonignorable nonresponse in categorical panel data with an example in estimating Gross Labor-Force flows. *Journal of Business and Economic Statistics*, 6, 207-219.

Survey Methodology, June 2009
Vol. 35, No. 1, pp. 53-66
Statistics Canada, Catalogue No. 12-001-X

Hierarchical and empirical Bayes small domain estimation of the proportion of persons without health insurance for minority subpopulations

Malay Ghosh, Dalho Kim, Karabi Sinha, Tapabrata Maiti, Myron Katzoff and Van L. Parsons [1]

Abstract

The paper considers small domain estimation of the proportion of persons without health insurance for different minority groups. The small domains are cross-classified by age, sex and other demographic characteristics. Both hierarchical and empirical Bayes estimation methods are used. Also, second order accurate approximations of the mean squared errors of the empirical Bayes estimators and bias-corrected estimators of these mean squared errors are provided. The general methodology is illustrated with estimates of the proportion of uninsured persons for several cross-sections of the Asian subpopulation.

Key Words: Asian; Bias-corrected; Mean squared error; Second order accurate.

1. Introduction

The main motivation behind this work was small domain estimation of the proportion of individuals without health insurance for different minority subpopulations. The small domains were constructed on the basis of age, sex, race and the region where the person belongs. The National Health Interview Survey (NHIS) data provide the individual level binary response (that is whether or not a person has health insurance) along with individual level covariates. The data can be obtained at http://www.cdc.gov/nchs/nhis.htm. The design of NHIS is discussed in Botman, Moore, Moriarity and Parsons (2000).

In a typical year the NHIS samples dwelling units, the collective members of each unit being referred to as a household, and members with a “strong” relationship being referred to as a family. (Structural units are more explicitly defined in Chapter 5.2 in the Census document at www.census.gov/prod/2002pubs/tp63rv.pdf). Each year the NHIS data contain about 40,000 households, of which over 98% are one-family households, and contain about 100,000 persons. For “family-type” questions, *e.g.*, on insurance coverage, all adults at home are invited to participate in the interview, but proxy adult response is also allowed. Children require an adult proxy.

The original survey for any given year contains data on more than 100,000 individuals and on over 800 variables. Of these individuals, we have information on the primary response variable, namely whether a person has health insurance or not. In addition, there is information on demographic characteristics such as age, sex, race, region, education, income status, medical condition, disability conditions (if any) and many other socio-economic factors.

For the entire US population, the direct estimates for these domains, namely the weighted sample proportions, are fairly reliable, since the sample size for each domain is reasonably large. This need not be the case though when our analysis is targeted towards specific subpopulations, such as Hispanics, Asians and similar minority sectors of the community.

For a targeted minority subpopulation, the sample size in a domain is not always very large. Hence, the direct estimates may not be very reliable, being accompanied with large standard errors and coefficients of variation. This calls for the use of small domain estimation techniques, where indirect estimates are obtained for these domains based on implicit or explicit models. These models help building a link between these domains, and thus produce typically estimates of greater precision by borrowing strength.

We employ both hierarchical Bayes (HB) and empirical Bayes (EB) methodology to obtain small domain estimates and find also the associated measures of precision. The analysis is based on a HB analogue of the generalized linear mixed model (GLMM) to obtain posterior means and posterior standard errors of the population small domain proportions. The method was proposed in Ghosh, Natarajan, Stroud and Carlin (1998). The EB approach is based on the theory of optimal estimating functions. We obtain EB estimators and the corresponding approximate mean squared error estimators by an asymptotic method analogous to that of Prasad and Rao (1990) and Ghosh and Maiti (2004). While the procedure of Ghosh and Maiti (2004) is based on area-level data, the present approach uses unit level data. Hence, by necessity, one needs some modification of the procedure proposed in Ghosh and Maiti (2004) in developing the estimators. Also, the general methodology,

[1] Malay Ghosh, University of Florida; Dalho Kim, Kyungpook National University; Karabi Sinha, University of California at Los Angeles; Tapabrata Maiti, Michigan State University; Myron Katzoff, National Center for Health Statistics; Van L. Parsons, National Center for Health Statistics.

like that of Ghosh and Maiti is not restricted only to binary data. The methodology is applicable to the natural exponential family with quadratic variance functions. (Morris 1982, 1983). The development of mean squared errors of the estimates under the proposed model is somewhat simpler than that of Ghosh and Maiti (2004) for the binary case. Moreover, like Ghosh and Maiti (2004), our analysis utilizes the survey weights along with the model to derive the small domain estimates. Thus, our method, in some sense, can be regarded as design-assisted model-based estimation.

Survey weights attached to individual sampling units are usually proportional to inverses of their selection probabilities. They are often used to produce design-unbiased estimators. The classic example is the celebrated Horvitz-Thompson estimator. However, while such estimators guard against model failure, they may result in loss of efficiency if the assumed model is true. For example, in a simple Bayesian set up, if $y_i \mid \theta_i$ are independently distributed $\mathrm{N}(\theta_i, 1)$, while θ_i are independently and identically distributed $\mathrm{N}(\mu, A), (i = 1, \ldots, n)$, then the Bayes estimator (posterior mean) of $\overline{\theta} = n^{-1} \sum_{i=1}^{n} \theta_i$ is $n^{-1} \sum_{i=1}^{n} [(1 - B) y_i + B\mu] = (1 - B)\overline{y} + B\mu$, where $B = (1 + A)^{-1}$. This estimator has Bayes risk $n^{-1}(1 - B)$ under the assumed model. On the other hand, the estimator $\sum_{i=1}^{n} w_i y_i$ of $\overline{\theta}$, with $\sum_{i=1}^{n} w_i = 1$ has Bayes risk $n^{-1}(1 - B) + E[(1 - B)\overline{y} + B\mu - \sum_{i=1}^{n} w_i y_i]^2$. If, however, the assumed model is not true, for example, θ_i are independently and identically distributed $\mathrm{N}(\mu, A), (i = 1, \ldots, n)$, where A departs widely from A_0, then the Bayes risk of the estimator $(1 - B)\overline{y} + B\mu$ of $\overline{\theta}$ has Bayes risk $n^{-1}(1 - B_0) + (B - B_0)^2 (\overline{y} - \mu)^2, B_0 = (1 + A_0)^{-1}$, which can be quite larger than the corresponding Bayes risk of $\sum_{i=1}^{n} w_i y_i$ depending of course on B_0, μ and the w_i.

The present paper produces small domain estimates of the proportion of uninsured persons for the Asian population. The estimates and measures of precision are based both on the hierarchical Bayesian model as well as the EB model. The analysis was done for all the individual years 1997-2000. For brevity, the results are reported only for the year 2000. We carried out a similar analysis for the Hispanic population also. In this case, the number of small domains was 336. Since the methodology was the same as that for the Asians, to save space, we have not included in this paper that analysis as well.

The Asian group is formally composed of the (1) Chinese, (2) Filipino, (3) Asian Indian, and (4) others such as Koreans, Vietnamese, Japanese, Hawaiian, Samoan, Guamanian *etc*. These individuals are assigned to specific domains depending on their age, race, gender and the region they come from. There are 3 age-groups (0-17, 18-64 and 65+), 2 Genders, 4 Races and 4 Regions depending on the size of the Metropolitan Statistical Area ($< 499{,}999$; 500,000-999,999; 1,000,000-2,499,999 $> 2{,}500{,}000$). Thus, the total number of domains equals $3 \times 2 \times 4 \times 4 = 96$. When the individuals are distributed to their respective domains, it turns out that many of the domains contain only a few observations. Indeed, there are several domains with a sample of size 1, while one domain has sample size zero.

The outline of the remaining sections is as follows. Section 2 addresses the selection of covariates for the Asians. Section 3 discusses the general HB methodology needed for obtaining the small domain estimates and the associated measures of precision. Section 4 discusses the adequacy of the proposed HB model. Section 5 discusses an alternative EB methodology, finds second order correct (to be made precise later) mean squared errors (MSE's) of the proposed EB estimators, and also second order correct approximation of these MSE's. Section 6 finds the small domain estimates and the corresponding measures of precision for the Asian subpopulation in 2000 using both the HB and the EB methodology, and these estimates are compared with the direct estimates. Some concluding remarks are made in Section 7.

2. Selection of covariates

As mentioned in the introduction, the number of covariates exceeds 800. Inclusion of all of them in the initial model is impractical and unnecessary. We started with what we deemed to be a meaningful set of 6 covariates and used a fully stepwise selection process (with a significance level of 0.05) to finally come up with the best model.

The six covariates that we considered were: (1) legal marital status, (2) family size, (3) education level, (4) total earnings from the previous year, (5) total family income, and (6) full time working status.

After the stepwise procedure, our final model included, along with the intercept term, the covariates family size, education level, and total family income.

Since the SURVEYREG procedure in SAS Version 8 fits linear regression models and produces hypothesis tests and estimates for survey data, we used this procedure for our covariate selection. Logistic regression for covariate selection was not available at the time when this research was done. It may be noted though that SURVEYREG acounts for clustering and unequal weighting, and produces standard errors that correctly account for complex survey designs.

3. Hierarchical Bayesian analysis

A general one-parameter exponential family model is given by

$$f(y_{ij} \mid \theta_{ij}) = \exp[\xi_{ij}\{y_{ij}\theta_{ij} - \psi(\theta_{ij})\}]h(y_{ij}; \xi_{ij}), \qquad (3.1)$$

$j = 1, \ldots, n_i, i = 1, \ldots, k.$ Here y_{ij} is the response of the j^{th} unit in the i^{th} small domain, while ξ_{ij}, the "so-called" overdispersion parameters are assumed to be known, and are taken as 1 without loss of generality. This is because one can otherwise work with the transformed parameters. $\zeta_{ij} = \xi_{ij}\theta_{ij}$. The function h is a positive function which depends on the y_{ij}, but not on the θ_{ij}. If y_{ij} is binary with success probability p_{ij}, then $\theta_{ij} = \text{logit}(p_{ij})$. In our example, y_{ij}, the response of the j^{th} individual in the i^{th} small domain, is 1 or 0 depending on whether the person does not or does have health insurance. We are interested in estimation of $\bar{\mu}_{iw} = \sum_{j=1}^{n_i} w_{ij} p_{ij}$, the domain specific weighted averages of the population proportions. In this case, the direct estimator of $\bar{\mu}_{iw}$ is $\sum_{j=1}^{n_i} w_{ij} y_{ij}$. These direct estimators are usually subject to large standard errors and coefficients of variation. The survey weights w_{ij} are assumed to be known, and are normalized so that $\sum_{j=1}^{n_i} w_{ij} = 1$ for all $i = 1, \ldots, k$. It must be admitted though that often in practice, the w_{ij} are only estimates, for example taking into account post-stratification and non-response. However, the actual mechanism used to generate these weights are unavailable to secondary users of the data, and we need to assume the weights to be known. Another important example, not specifically considered in this paper, is $y_{ij} \sim \text{Poisson}(\lambda_{ij})$, so that $\theta_{ij} = \log(\lambda_{ij})$. One can use a Poisson model here based on the domain level counts of uninsured people The difficulty lies in the fact that in the present example, we have individual level and *not* domain level covariates. Modelling the counts via domain-level covariates is not possible in this situation.

In this section, we discuss how to carry out the analysis for the general hierarchical Bayesian model when we are interested in estimating $\mu_{ij} = E(y_{ij} \mid \theta_{ij}) = \psi'(\theta_{ij})$. Since $\psi''(\theta_{ij}) = \text{var}(y_{ij} \mid \theta_{ij})$, μ_{ij} is a one-to-one function of θ_{ij}. In particular, $\mu_{ij} = p_{ij}$ in the binary case. Specific applications will be considered in Section 5.

The next stage of the model is

$$\theta_{ij} = \boldsymbol{x}_{ij}^T \boldsymbol{b} + u_i; \; j = 1, \ldots, n_i, i = 1, \ldots, k, \qquad (3.2)$$

where $\boldsymbol{x}_{ij}$ are the design vectors, or equivalently the predictor vectors, $\boldsymbol{b}$ is the vector of regression parameters, and u_i are the random effects. It is assumed that u_i are iid $\text{N}(0, \sigma_u^2)$. Also, let $\boldsymbol{X}^T = (\boldsymbol{x}_{11}, \ldots, \boldsymbol{x}_{1n_1}, \ldots, \boldsymbol{x}_{k1}, \ldots, x_{kn_k})$, and assume that $\boldsymbol{X}$ is a full rank matrix.

Finally, it is assumed that $\boldsymbol{b}$ and σ_u^2 are mutually independent, where $\boldsymbol{b}$ has the improper uniform prior on, R^P, and σ_u^2 has an inverse gamma distribution with parameters $c/2.d/2.i.e.$, $\pi(\sigma_u^2) \propto \exp(-c/2\sigma_u^2)(\sigma_u^2)^{-d/2-1}$, $c > 0$.

Let $\boldsymbol{y} = (y_{11}, \ldots, y_{1n_1}, \ldots, y_{k1}, \ldots, y_{kn_k})^T$, and $\boldsymbol{\theta} = (\theta_{11}, \ldots, \theta_{1n_1}, \ldots, \theta_{k1}, \ldots, \theta_{kn_k})^T$. Then the joint posterior is given by

$$\pi(\boldsymbol{\theta}, \boldsymbol{b}, \sigma_u^2 \mid \boldsymbol{y}) \propto \prod_{i=1}^{k} \prod_{j=1}^{n_i} f(y_{ij} | \theta_{ij})$$

$$\times (\sigma_u^2)^{-k/2} \exp\left[-\frac{1}{2\sigma_u^2} \sum_{i=1}^{m} \sum_{j=1}^{n_i} (\theta_{ij} - \boldsymbol{x}_{ij}^T \boldsymbol{b})^2\right]$$

$$\times (\sigma_u^2)^{-d/2-1} \exp\left(-\frac{c}{2\sigma_u^2}\right). \qquad (3.3)$$

This is a nonconjugate Bayesian analysis, and is not implementable analytically. Instead, we use the Markov chain Monte Carlo (MCMC) numerical integration technique. In particular, we employ the Gibbs sampler. The general MCMC technique is discussed in many places. A convenient reference is Tanner (1996, Chapter 6).

In order to implement the Gibbs sampler, we need to find the full conditionals of θ_{ij}, $\boldsymbol{b}$ and σ_u^2. The full conditionals are given by

$$\sigma_u^2 \mid \boldsymbol{\theta}, \boldsymbol{b}, \boldsymbol{y} \sim \text{IG}\left(\frac{\sum_{i=1}^{k} \sum_{j=1}^{n_i} (\theta_{ij} - \boldsymbol{x}_{ij}^T \boldsymbol{b})^2 + c}{2}, \frac{k+d}{2}\right);$$

$$\boldsymbol{b} \mid \boldsymbol{\theta}, \sigma_u^2, \boldsymbol{y} \sim N((\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}, \sigma_u^2(\boldsymbol{X}^T\boldsymbol{X})^{-1});$$

$$\theta_{ij} \mid \boldsymbol{b}, \sigma_u^2, \boldsymbol{y} \sim f(y_{ij} \mid \theta_{ij}) \exp\left[-\frac{1}{2\sigma_u^2}(\theta_{ij} - \boldsymbol{x}_{ij}^T \boldsymbol{b})^2\right].$$

Our data analysis is based on generating samples from the above conditionals specialized to the binary case. Generation of samples from the conditionals of σ_u^2 and $\boldsymbol{b}$ is standard. This is not so for the θ_{ij}, and requires the Metropolis-Hastings algorithm. For a discussion of this algorithm, we refer once again to Tanner (1996).

If $\hat{\mu}_{ij}^{(r)}$ denotes the sampled value of μ_{ij} generated from the r^{th} draw, and the number of draws is R, then the Monte Carlo estimate of $E(\mu_{ij} \mid \boldsymbol{y})$ is $R^{-1}\sum_{r=1}^{R} \hat{\mu}_{ij}^{(r)}$. Similarly, the Monte-Carlo estimate of $\text{var}(\mu_{ij} \mid \boldsymbol{y})$ is $R^{-1}\sum_{r=1}^{R} (\hat{\mu}_{ij}^{(r)})^2 - (R^{-1}\sum_{r=1}^{R} \hat{\mu}_{ij}^{(r)})^2$. Finally, Monte-Carlo estimate of $\text{cov}(\mu_{ij}, \mu_{i'j'} \mid \boldsymbol{y})$ is given by $R^{-1}\sum_{r=1}^{R} (\hat{\mu}_{ij}^{(r)} \hat{\mu}_{i'j'}^{(r)}) - (R^{-1}\sum_{r=1}^{R} \hat{\mu}_{ij}^{(r)}) (R^{-1}\sum_{r=1}^{R} \hat{\mu}_{i'j'}^{(r)})$. Based on these calculations, it is now immediate to find $E[\bar{\mu}_{iw} \mid \boldsymbol{y}] = \sum_{j=1}^{n_i} w_{ij} E(\mu_{ij} \mid \boldsymbol{y})$ and $V[\bar{\mu}_{iw} \mid \boldsymbol{y}] = \sum_{j=1}^{n_i} w_{ij}^2 V(\mu_{ij} \mid \boldsymbol{y}) + \sum_{1 \le j \ne j' \le n_i} w_{ij} w_{ij'} \text{ Cov}(\mu_{ij}, \mu_{ij'} \mid \boldsymbol{y})$. In contrast, the direct unbiased estimator of $\bar{\mu}_{iw}$ is given by $\bar{y}_{iw} = \sum_{j=1}^{n_i} w_{ij} y_{ij}$. However, as noted earlier, for many of these domains, the sample sizes are so small that these unbiased estimators are subject to large standard errors and coefficients of variation.

4. Empirical Bayes estimation

Once again, let y_{ij} denote the response of the j^{th} unit in the i^{th} small domain $(j=1,\ldots,n_i; i=1,\ldots,k)$. Also, we assume the exponential family model for the y_{ij} as given in (3.1), but it is assumed in addition that the y_{ij} has a probability function or a probability density function belonging to the natural exponential family quadratic variance function (NEF-QVF) class. We may recall that $\mu_{ij}=E(y_{ij}\mid\theta_{ij})=\psi'(\theta_{ij})$. With the quadratic variance function structure, $\text{Var}(y_{ij}\mid\theta_{ij})=Q(\mu_{ij})=v_0+v_1\mu_{ij}+v_2\mu_{ij}^2$, where v_0, v_1 and v_2 are not simultaneously zero. Morris (1982, 1983) has characterized distributions belonging to the NEF-QVF family. The family consists of the six basic distributions, namely, (i) Bernoulli, (ii) Poisson, (iii) normal with known variance, (iv) geometric, (v) exponential, (vi) hyperbolic secant, and their convolutions. In this way, binomial, negative binomial and gamma distributions also belong to this family. For the Bernoulli distribution, $v_0=0, v_1=1$ and $v_2=-1$. For the Poisson distribution, $v_0=v_2=0$ and $v_1=1$. For the normal distribution with known variance $\sigma^2, \xi_{ij}=\sigma^{-2}, v_0=1$ and $v_1=v_2=0$. Once again we will assume without loss of generality that $\xi_{ij}=1$.

We propose in this section EB estimators of the small domain means. To this end, we begin with the general NEF-QVF family of distributions along with a conjugate prior for the canonical parameter of the exponential model. Together they constitute an overdispersed NEF-QVF family of distributions. Specifically, we consider the conjugate prior with pdf

$$\pi(\theta_{ij})=\exp[\lambda\{m_{ij}\theta_{ij}-\psi(\theta_{ij})\}]C(\lambda, m_{ij}) \tag{4.1}$$

for θ_{ij}, where $m_{ij}=g(\boldsymbol{x}_{ij}^T\boldsymbol{b}), j=1,\ldots,n_i;\ i=1,\ldots,k$. Here $\boldsymbol{x}_{ij}$ is the design vector associated with the j^{th} unit in the i^{th} small domain, and g is the link function. Then (Morris 1983),

$$E(\mu_{ij})=m_{ij};\ \text{var}(\mu_{ij})=Q(m_{ij})/(\lambda-v_2), \tag{4.2}$$

where we assume that $\lambda>\max(0, v_2)$. Since $\text{var}(\mu_{ij})$ is strictly decreasing in λ, we may interpret the latter as the precision parameter.

We first obtain the Bayes estimator of μ_{ij}. This is given by (Morris 1983)

$$E(\mu_{ij}\mid y_{ij})=\frac{1}{\lambda+1}y_{ij}+\frac{\lambda}{\lambda+1}m_{ij}(\boldsymbol{b}).$$

The above can also be viewed as the best linear unbiased predictor (BLUP) of μ_{ij}. To see this, we calculate

$$E(y_{ij})=E(\mu_{ij})=m_{ij};\ \text{cov}(y_{ij},\mu_{ij})=\text{var}(\mu_{ij})$$

$$=Q(m_{ij})/(\lambda-v_2);\ \text{var}(y_{ij})=\frac{\lambda+1}{\lambda-v_2}Q(m_{ij}).$$

Hence, the BLUP of μ_{ij} is given by

$$m_{ij}(\boldsymbol{b})+\frac{\text{cov}(y_{ij},\mu_{ij})}{\text{var}(y_{ij})}(y_{ij}-m_{ij}(\boldsymbol{b}))$$

$$=\frac{1}{\lambda+1}y_{ij}+\frac{\lambda}{\lambda+1}m_{ij}(\boldsymbol{b}). \tag{4.3}$$

Thus the Bayes estimator of $\bar{\mu}_{iw}=\sum_{j=1}^{n_i}w_{ij}\mu_{ij}$ is given by $\sum_{j=1}^{n_i}w_{ij}E(\mu_{ij}\mid y_{ij})$.

In practice, however, $\boldsymbol{b}$ and λ are unknown, and need to be estimated from the marginals of the y_{ij}. However, except for the normal distribution, these marginals are fairly complicated, and finding MLE's from the marginal likelihoods can become quite formidable. Instead, we find estimates based on some optimal unbiased estimating equations (Godambe and Thompson 1989) which requires only evaluation of the first four moments of these marginals. To this end, we begin with the the elementary unbiased estimating functions $ig_{1ij}=y_{ij}-m_{ij}$ and $g_{2ij}=(y_{ij}-m_{ij})^2-(\lambda+1)/(\lambda-v_2)V(m_{ij})$. In order to construct the optimal estimating equations, let

$$\boldsymbol{D}_{ij}^T=\begin{bmatrix}-E\left(\dfrac{\partial g_{1ij}}{\partial\boldsymbol{b}}\right) & -E\left(\dfrac{\partial g_{2ij}}{\partial\boldsymbol{b}}\right)\\ -E\left(\dfrac{\partial g_{1ij}}{\partial\lambda}\right) & -E\left(\dfrac{\partial g_{2ij}}{\partial\lambda}\right)\end{bmatrix}.$$

Also, let

$$\boldsymbol{\Sigma}_{ij}=\begin{bmatrix}\mu_{2ij} & \mu_{3ij}\\ \mu_{3ij} & \mu_{4ij}-\mu_{2ij}^2\end{bmatrix},$$

where $\mu_{rij}=E(y_{ij}-m_{ij})^r$ is the r^{th} central moment of y_{ij} based on its marginal distribution. The optimal estimating equations are then given by $\sum_{i=1}^k\sum_{j=1}^{n_i}\boldsymbol{D}_{ij}^T\boldsymbol{\Sigma}_{ij}^{-1}\boldsymbol{g}_{ij}=\boldsymbol{0}$, where $\boldsymbol{g}_{ij}=(g_{1ij}\ g_{2ij})^T$. We obtain estimates of $\boldsymbol{b}$ and λ (if they exist) by solving these equations. The solutions of these equations are found by the Nelder-Meade algorithm.

Unfortunately, the above method fails for binary data. In this case, $v_2=-1$ so that $\text{var}(y_{ij})$ does not depend on λ. Indeed, the marginal beta-binary distributions of the y_{ij} are unidentifiable in λ. A simple way to verify this is that if $y\mid p\sim\text{Bin}(1,p)$, and $p\sim\text{Beta}(\lambda m, \lambda(1-m))$, then $E(y)=E(p)=m$, and a binary distribution is completely characterized by its mean. The problem does not occur for a Binomial(n, p) distribution with $n\geq 2$ since with the same marginal for p, the mgf of the marginal distribution of the binomial y is $E[(p\exp(t)+1-p)^n]$ which depends on λ.

For binary $y_{ij}, \partial g_{1ij} / \partial \lambda = \partial g_{2ij} / \partial \lambda = 0$ so that the second element of the vector $\sum_{i=1}^{k} \sum_{j=1}^{n_i} \boldsymbol{D}_{ij}^T \boldsymbol{\Sigma}_{ij}^{-1} \boldsymbol{g}_{ij}$ is zero. Accordingly, the proposed estimating equations approach fails to estimate λ. The basic data, to be considered in our application, is binary, and this necessitates modification of the proposed procedure.

We have thus considered the optimal estimating function (for known λ)

$$\sum_{i=1}^{k} \sum_{j=1}^{n_i} [(y_{ij} - m_{ij})/(\text{var}(y_{ij})] \frac{\partial m_{ij}}{\partial \boldsymbol{b}} = \mathbf{0},$$

since $\partial g_{1ij} / \partial \boldsymbol{b} = -\partial m_{ij} / \partial \boldsymbol{b}$. It may be noted also that in this case $\text{var}(y_{ij}) = V(m_{ij}) = m_{ij}(1 - m_{ij})$. Also, with the logistic representation, $m_{ij}(\boldsymbol{b}) = \exp(\boldsymbol{x}_{ij}^T \boldsymbol{b})/[1 + \exp(\boldsymbol{x}_{ij}^T \boldsymbol{b})]$, one gets $\partial m_{ij} / \partial \boldsymbol{b} = -m_{ij}(1 - m_{ij}) \boldsymbol{x}_{ij}$. Thus $\boldsymbol{b}$ is estimated from the estimating equations $\sum_{i=1}^{k} \sum_{j=1}^{n_i} \boldsymbol{x}_{ij} y_{ij} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} \boldsymbol{x}_{ij} m_{ij}$. Denoting this estimator by $\hat{\boldsymbol{b}}$, the EB estimator of μ_{ij} is given by

$$\hat{\mu}_{ij}^{\text{EB}} = \frac{1}{\lambda + 1} y_{ij} + \frac{\lambda}{\lambda + 1} m_{ij}(\hat{\boldsymbol{b}}). \qquad (4.4)$$

Accordingly, the EB estimator of $\bar{\mu}_{iw}$ is $\hat{\bar{\mu}}_{iw}^{\text{EB}} = \sum_{j=1}^{n_i} w_{ij} \hat{\mu}_{ij}^{\text{EB}}$.

The procedure described above assumes a known λ. One can find estimates for the μ_{ij} for different choices of λ. In this article, we have tried $\lambda = 0.1, 0.5$ and 1, and have compared the estimates with the corresponding HB estimates.

Next, in this section, we find the mean squared errors (MSE) and also the estimated MSE's of $\hat{\bar{\mu}}_{iw}^{\text{EB}}$ assuming known λ. We state two theorems in this section. Some notations are needed before stating these theorems. Let $\boldsymbol{M} = \text{Diag}(m_{11}, \ldots, m_{1n_1}, \ldots, m_{k1}, \ldots, m_{kn_k})$ and $\boldsymbol{\Sigma}(\boldsymbol{b}) = \boldsymbol{X}^T \boldsymbol{M} (\boldsymbol{I} - \boldsymbol{M}) \boldsymbol{X} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} m_{ij}(1 - m_{ij}) \boldsymbol{x}_{ij} \boldsymbol{x}_{ij}^T$. Also, let $n_T = \sum_{i=1}^{k} n_i$. It is assumed that $1 \leq n_i \leq C$ for every *i*, so that $n_T = O_e(k)$, where O_e denotes the exact order. The two theorems are now given below.

Theorem 1. Assume $\boldsymbol{\Sigma}(\boldsymbol{b}) = O_e(k)$, *i.e.*, each element of $\boldsymbol{\Sigma}(\boldsymbol{b})$ is bounded below by some constant C_1, and is bounded above by some constant C_2, where $0 < C_1 < C_2 < \infty$. Then an approximate expression for $\text{MSE}(\hat{\bar{\mu}}_{iw}^{\text{EB}})$ correct up to $O(k^{-1})$ is given by

$$\begin{aligned}\text{MSE}(\hat{\bar{\mu}}_{iw}^{\text{EB}}) \doteq{} & \frac{\lambda}{(\lambda + 1)^2} \sum_{j=1}^{n_i} w_{ij}^2 m_{ij}(b)(1 - m_{ij}(b)) \\ & + \frac{\lambda^2}{(\lambda + 1)^2} \left[\sum_{j=1}^{n_i} w_{ij} m_{ij}(b)(1 - m_{ij}(b)) \boldsymbol{x}_{ij} \right]^T \\ & \times \boldsymbol{\Sigma}^{-1}(\boldsymbol{b}) \left[\sum_{j=1}^{n_i} w_{ij} m_{ij}(b)(1 - m_{ij}(b)) \boldsymbol{x}_{ij} \right]. \qquad (4.5)\end{aligned}$$

Theorem 2. Assume $\boldsymbol{\Sigma}(\boldsymbol{b}) = O_e(k)$. Then the following approximation to $\text{MSE}(\hat{\bar{\mu}}^{\text{EB}})$ holds correct up to $O(k^{-1})$.

$$\begin{aligned} & \frac{\lambda}{(1 + \lambda)^2} \sum_{j=1}^{n_i} \Bigg[m_{ij}(\hat{\boldsymbol{b}})(1 - m_{ij}(\hat{\boldsymbol{b}})) - (1 - 2m_{ij}(\hat{\boldsymbol{b}}))m_{ij}(\hat{\boldsymbol{b}}) \\ & \qquad (1 - m_{ij}(\hat{\boldsymbol{b}})) \frac{1}{2} \boldsymbol{\Sigma}^{-1}(\hat{\boldsymbol{b}}) \begin{pmatrix} tr(\boldsymbol{\Sigma}^{-1}(\hat{\boldsymbol{b}}) \boldsymbol{K}_1(\hat{\boldsymbol{b}})) \\ \cdot \\ \cdot \\ \cdot \\ tr(\boldsymbol{\Sigma}^{-1}(\hat{\boldsymbol{b}}) \boldsymbol{K}_p(\hat{\boldsymbol{b}})) \end{pmatrix} \\ & \qquad + m_{ij}^2(\hat{\boldsymbol{b}})(1 - m_{ij}(\hat{\boldsymbol{b}}))^2 \boldsymbol{x}_{ij}^T \boldsymbol{\Sigma}^{-1}(\hat{\boldsymbol{b}}) \boldsymbol{x}_{ij} \Bigg] \\ & \qquad + \frac{\lambda^2}{(\lambda + 1)^2} \left[\sum_{j=1}^{n_i} w_{ij} m_{ij}(\hat{\boldsymbol{b}})(1 - m_{ij}(\hat{\boldsymbol{b}})) \boldsymbol{x}_{ij} \right]^T \\ & \qquad \times \boldsymbol{\Sigma}^{-1}(\hat{\boldsymbol{b}}) \left[\sum_{j=1}^{n_i} w_{ij} m_{ij}(\hat{\boldsymbol{b}})(1 - m_{ij}(\hat{\boldsymbol{b}})) \boldsymbol{x}_{ij} \right]. \qquad (4.6)\end{aligned}$$

The proofs of these theorems are deferred to the Appendix. We will apply these results in finding approximate estimates of MSE's of EB estimators in the next section. However, before that, the following point is worth noting.

If one denotes the coefficient of $\lambda/(1 + \lambda)^2$ by $B_i(\hat{b})$ and the coefficient of $\lambda^2/(1 + \lambda)^2$ by $C_i(\hat{b})$ in Theorem 2, then noting that $B_i(\hat{b}) = O(1)$ and $C_i(\hat{b}) = O(k^{-1})$, for large k, $\text{MSE}(\hat{\bar{\mu}})$ is maximized at $\hat{\lambda} = (B_i(\hat{b}))/(B_i(\hat{b}) - 2C_i(\hat{b}))$ which is typically very close to 1. The resulting prior with $\hat{\lambda}$ replacing λ is the data adaptive approximate least favorable prior. In the example to be considered, this estimated λ turns out to be 1.003 which conforms the above observation.

5. Small domain estimates for Asians

We first describe how the small domains are constructed. Consider the 4-tuple (k_1, k_2, k_3, k_4), where $k_1 = 1, 2, 3$ or 4 according as the person is Chinese, Filipino, Asian Indian or Islanders. Next $k_2 = 1$ or 2 according as the person is a male or a female. Then $k_3 = 1, 2$ or 3 according as the person belongs to the age-group 0-17, 18-64 or 65+. Finally, $k_4 = 1, 2, 3$ or 4 according as the person belongs to a Metropolitan Statistical Area (MSA) of size $\leq 499{,}999$, 500,000 - 999,999, 1,000,000 - 2,499,999 or $\geq 2{,}500{,}000$. A small domain is now numbered by the formula $24(k_1 - 1) + 12(k_2 - 1) + 4(k_3 - 1) + k_4$ corresponding to the 4-tuple (k_1, k_2, k_3, k_4). For example, the small domain consisting of Filipino females belonging to the age-group

18-64 and a MSA of size $500,000-999,999$ is numbered 42.

The basic data consist of $y_{ij}=1$ or 0 if the j^{th} individual in the i^{th} small domain does not (does) have health insurance;

$\tilde{w}_{ij} =$ the sampling weight attached to the j^{th} unit in the i^{th} small domain;

$w_{ij} =$ $\tilde{w}_{ij}/\sum_{j=1}^{n_i}\tilde{w}_{ij}$ so that $\sum_{j=1}^{n_i} w_{ij}=1$ for each *i*.

$x_{ij1} =$ the family size of the j^{th} unit in the i^{th} small domain;

$x_{ij2} =$ the education level of the j^{th} unit in the i^{th} small domain;

$x_{ij3} =$ total family income of the j^{th} unit in the i^{th} small domain;

Let $p_{ij}=E(y_{ij})$. For the HB analysis, we model

$$\theta_{ij}=\text{logit}(p_{ij})=b_0+b_1x_{ij1}+b_2x_{ij2}+b_3x_{ij3}+u_i,$$
$$j=1,\ldots,n_i,\ i=1,\ldots,96.$$

The direct domain estimates are given by $\hat{p}_{iw}=\sum_{j=1}^{n_i}w_{ij}y_{ij}$. The corresponding hierarchical Bayes estimates are given by $\hat{p}_{iw}^{\text{HB}}=\sum_{j=1}^{n_i}w_{ij}E(p_{ij}\mid \boldsymbol{y})$. We use MCMC as described in Section 2 to obtain these estimates. They are referred to in the table as HB. The associated posterior standard errors are referred to as se(HB). Our hyperprior considers: $c=0.2, 0.02, 0.002$; $d=0.2, 0.02, 0.002$. The results are very insensitive to the choice of the hyperpriors, and are reported only for $c=d=0.02$. In addition, we have EB estimators for different choices of the parameter λ. The results are reported for $\lambda=0.1, 0.5$ and 1.

Table 1 provides the various estimates of uninsured Asian people and the associated standard errors for the different small domains for the year 2000. Domain 2 is excluded due to zero sample size. Domain 2 refers to Male Filipinos in the age group 0-17 belonging to MSA's of size 500,000 - 999,999. The measures of precision (posterior s.d.'s) associated with the HB estimates are denoted by se(HB) and are given by the formula $\text{se}^2(\text{HB})=\text{var}(\sum_{j=1}^{n_i}w_{ij}p_{ij}\mid \boldsymbol{y})$. One of the advantages of the HB or EB estimates is that for domains with very small sample sizes, often the direct estimates of the proportion of uninsured is zero, whereas the former provide small but non-zero estimates. We chose not to collapse the direct estimates for domains with very small sample sizes. The unit level covariates were quite distinct, and there was no meaningful way to combine them. We note also that when $\lambda=0.5$, *i.e.*, the direct and synthetic estimates have 1: 2 weight ratio, the EB and HB estimates are very close.

Table 1
Small domain estimates of the proportions of uninsured Asians: Year 2000

Domain	n_i	Direct	'97-'99 average	HB	se (HB)	EB $\lambda=0.5$	EB $\lambda=1$	se (EB) $\lambda=0.5$	se (EB) $\lambda=1$
1	10	0.126	0.034	0.133	0.043	0.148	0.158	0.057	0.060
2	0	-	0.085	-	-	-	-	-	-
3	24	0.063	0.016	0.074	0.025	0.076	0.082	0.037	0.039
4	28	0.146	0.105	0.150	0.027	0.163	0.171	0.041	0.043
5	20	0.138	0.265	0.143	0.032	0.153	0.160	0.043	0.046
6	17	0.112	0.124	0.120	0.032	0.134	0.144	0.019	0.021
7	78	0.097	0.100	0.104	0.015	0.107	0.112	0.022	0.024
8	66	0.274	0.229	0.253	0.023	0.240	0.224	0.072	0.076
9	5	0.173	0.000	0.164	0.061	0.160	0.154	0.078	0.082
10	6	0.000	0.000	0.033	0.051	0.082	0.123	0.070	0.074
11	7	0.000	0.084	0.032	0.047	0.090	0.134	0.054	0.057
12	11	0.335	0.000	0.302	0.056	0.275	0.245	0.060	0.064
13	7	0.134	0.061	0.134	0.045	0.130	0.128	0.103	0.110
14	2	0.000	0.151	0.020	0.064	0.026	0.039	0.031	0.033
15	27	0.000	0.104	0.023	0.023	0.035	0.052	0.032	0.034
16	29	0.113	0.191	0.119	0.024	0.123	0.127	0.033	0.035
17	27	0.120	0.223	0.127	0.025	0.141	0.152	0.044	0.047
18	14	0.000	0.106	0.024	0.030	0.041	0.062	0.019	0.021
19	77	0.131	0.111	0.133	0.015	0.133	0.134	0.021	0.023
20	75	0.223	0.222	0.213	0.018	0.207	0.200	0.089	0.095
21	3	0.000	0.000	0.022	0.056	0.028	0.043	0.070	0.074
22	6	0.000	0.184	0.026	0.045	0.052	0.079	0.071	0.075
23	8	0.000	0.022	0.037	0.050	0.108	0.162	0.063	0.067
24	9	0.000	0.000	0.029	0.042	0.062	0.093	0.052	0.055
25	10	0.000	0.083	0.023	0.034	0.031	0.046	0.061	0.065
26	6	0.000	0.018	0.020	0.039	0.029	0.044	0.031	0.033
27	32	0.098	0.041	0.105	0.023	0.108	0.114	0.035	0.037
28	23	0.000	0.092	0.024	0.025	0.037	0.055	0.032	0.034
29	25	0.187	0.211	0.173	0.030	0.151	0.134	0.035	0.037
30	23	0.227	0.076	0.210	0.032	0.188	0.169	0.021	0.022

Table 1 (continued)
Small domain estimates of the proportions of uninsured Asians: Year 2000

Domain	n_i	Direct	’97-’99	HB	se (HB)	EB	EB	se (EB)	se (EB)
			average			λ = 0.5	λ = 1	λ = 0.5	λ = 1
31	71	0.118	0.059	0.123	0.016	0.125	0.128	0.024	0.026
32	50	0.109	0.156	0.113	0.019	0.112	0.113	0.113	0.120
33	2	0.000	0.000	0.024	0.071	0.037	0.055	0.115	0.122
34	2	0.000	0.000	0.026	0.073	0.047	0.070	0.058	0.061
35	8	0.108	0.000	0.113	0.042	0.112	0.114	0.067	0.071
36	7	0.000	0.000	0.030	0.045	0.065	0.098	0.051	0.054
37	9	0.062	0.197	0.069	0.035	0.062	0.063	0.036	0.038
38	17	0.000	0.019	0.019	0.024	0.023	0.034	0.037	0.040
39	24	0.117	0.022	0.124	0.028	0.134	0.142	0.040	0.043
40	20	0.000	0.070	0.028	0.029	0.052	0.078	0.025	0.027
41	50	0.163	0.145	0.160	0.020	0.156	0.153	0.027	0.029
42	38	0.141	0.114	0.139	0.021	0.133	0.130	0.020	0.022
43	76	0.104	0.090	0.112	0.016	0.120	0.128	0.020	0.022
44	73	0.142	0.149	0.142	0.016	0.139	0.137	0.119	0.127
45	2	0.000	0.000	0.027	0.076	0.051	0.076	0.090	0.095
46	3	0.000	0.052	0.021	0.056	0.023	0.035	0.052	0.055
47	10	0.000	0.072	0.024	0.034	0.044	0.066	0.068	0.072
48	7	0.000	0.172	0.029	0.045	0.068	0.102	0.051	0.054
49	10	0.087	0.364	0.095	0.037	0.099	0.105	0.078	0.083
50	5	0.000	0.000	0.027	0.050	0.053	0.080	0.032	0.034
51	23	0.038	0.092	0.053	0.023	0.056	0.066	0.037	0.039
52	21	0.243	0.195	0.223	0.037	0.198	0.176	0.030	0.032
53	31	0.114	0.184	0.120	0.022	0.121	0.124	0.040	0.042
54	18	0.202	0.169	0.195	0.031	0.188	0.182	0.019	0.020
55	74	0.094	0.115	0.102	0.015	0.102	0.106	0.019	0.020
56	83	0.204	0.296	0.192	0.017	0.178	0.165	0.133	0.141
57	2	0.000	0.124	0.029	0.082	0.062	0.092	0.146	0.154
58	1	0.000	0.000	0.019	0.087	0.023	0.035	0.000	0.000
59	2	0.000	0.196	0.020	0.063	0.021	0.032	0.103	0.194
60	8	0.112	0.116	0.120	0.044	0.132	0.143	0.059	0.063
61	16	0.202	0.140	0.187	0.036	0.169	0.152	0.040	0.043
62	3	0.301	0.163	0.276	0.086	0.252	0.227	0.100	0.107
63	33	0.055	0.093	0.069	0.020	0.073	0.082	0.028	0.030
64	28	0.105	0.275	0.112	0.024	0.115	0.120	0.032	0.034
65	33	0.126	0.133	0.129	0.021	0.126	0.126	0.029	0.031
66	13	0.393	0.290	0.350	0.054	0.323	0.288	0.048	0.051
67	70	0.080	0.136	0.089	0.015	0.088	0.093	0.019	0.021
68	75	0.179	0.233	0.171	0.017	0.159	0.149	0.019	0.021
69	1	0.000	0.000	0.851	0.248	0.705	0.558	0.163	0.173
70	2	0.361	0.000	0.331	0.098	0.299	0.268	0.119	0.126
71	4	0.000	0.091	0.023	0.050	0.032	0.048	0.077	0.082
72	2	0.000	0.182	0.045	0.101	0.157	0.236	0.155	0.165
73	45	0.271	0.144	0.256	0.026	0.256	0.249	0.028	0.030
74	10	0.000	0.044	0.024	0.034	0.034	0.051	0.051	0.055
75	83	0.149	0.097	0.150	0.016	0.160	0.166	0.020	0.021
76	59	0.113	0.205	0.120	0.018	0.128	0.136	0.023	0.024
77	68	0.338	0.224	0.313	0.025	0.302	0.284	0.023	0.024
78	39	0.098	0.138	0.103	0.020	0.102	0.104	0.026	0.028
79	122	0.110	0.163	0.117	0.013	0.125	0.133	0.016	0.017
80	125	0.308	0.314	0.281	0.020	0.262	0.239	0.016	0.017
81	7	0.000	0.000	0.029	0.043	0.066	0.099	0.065	0.069
82	12	0.000	0.045	0.025	0.032	0.047	0.070	0.048	0.051
83	13	0.049	0.017	0.068	0.035	0.088	0.108	0.050	0.053
84	4	0.000	0.061	0.028	0.056	0.060	0.091	0.088	0.093
85	32	0.189	0.113	0.193	0.027	0.217	0.231	0.035	0.037
86	10	0.136	0.056	0.137	0.036	0.127	0.123	0.051	0.054
87	52	0.192	0.098	0.185	0.021	0.184	0.180	0.024	0.026
88	65	0.153	0.120	0.155	0.018	0.162	0.166	0.022	0.024
89	71	0.285	0.210	0.265	0.022	0.256	0.242	0.022	0.023
90	57	0.086	0.146	0.095	0.017	0.102	0.110	0.022	0.024
91	153	0.149	0.167	0.150	0.011	0.156	0.160	0.014	0.015
92	138	0.308	0.285	0.283	0.020	0.266	0.244	0.015	0.017
93	10	0.000	0.000	0.030	0.041	0.073	0.110	0.059	0.063
94	16	0.067	0.015	0.081	0.029	0.090	0.101	0.042	0.044
95	18	0.108	0.018	0.123	0.032	0.145	0.163	0.046	0.049
96	14	0.111	0.087	0.125	0.039	0.160	0.185	0.050	0.053

The HB estimates of the proportion of uninsured for Asians vary in the 2%-35% range for the different small domains excluding domain 69. Admittedly, the EB and HB estimates for domain 69 are very adversely affected due to small sample size. We also report the standard errors associated with the HB estimates, and estimated approximate root mean squares accompanying the EB estimates. The proposed approach largely overcomes the valid criticism that naive EB estimates of standard errors (which ignore the $O(k^{-1})$ term) are typically underestimates. We have also provided a column giving the 3-year average of the direct estimates in 1997-1999. This is primarily to examine whether domains with zero direct estimates in 2000 also possess the same feature in other years, and also for comparison of EB and HB estimates with these estimates rather than the direct estimates. It turns out that with very few exceptions, the 1997-1999 average do not conform very much to the direct estimates. However, domain 69 still has zero direct estimate.

Table 2 provides the summary table for the proprtion of uninsured for the three age groups 0-17, 18-64 and 65+ individually for Chinese (Asian 1), Filipino (Asian 2), Asian Indian (Asian 3) and other Asians (Asian 4). It turns out that at this higher level of aggregation, both the EB and HB small domain estimates are fairly close to the corresponding direct estimates except possibly for the age-group 65+. This seems to be quite satisfactory, since at this level of aggregation, the direct estimates often serve as benchmarks for comparison purpose.

Table 2
Proportions without health insurance coverage by age group and Asian group in 2000

	Direct	HB	EB (λ = 0.5)	EB (λ = 1)
0-17 years				
Total	0.120	0.126	0.131	0.137
Asian 1	0.087	0.097	0.105	0.114
Asian 2	0.046	0.062	0.071	0.083
Asian 3	0.113	0.117	0.114	0.114
Asian 4	0.165	0.165	0.171	0.175
18-64 years				
Total	0.177	0.172	0.168	0.164
Asian 1	0.162	0.160	0.160	0.159
Asian 2	0.137	0.137	0.135	0.134
Asian 3	0.150	0.147	0.141	0.137
Asian 4	0.219	0.208	0.203	0.195
65+ years				
Total	0.063	0.080	0.103	0.123
Asian 1	0.083	0.097	0.123	0.143
Asian 2	0.021	0.043	0.064	0.085
Asian 3	0.119	0.126	0.136	0.145
Asian 4	0.055	0.075	0.100	0.123

6. Model diagnostics and implementation of the hierarchical Bayesian model

We followed Gelman and Rubin (1992) for the implementation and convergence diagnostics of the Gibbs sampler. In particular we took 5 chains each of size 1,000 with an initial burning period of 1,000 iterations. We checked the potential scale reduction factors for convergence and these appeared to be very close to unity (= 1 at convergence) for each one of the parameters. A number of other diagnostics criteria are available in the literature, and are implemented via the software CODA. A partial output is provided in the Figure 1. The left side shows the overlap of the 5 parallel chains, and the right side shows the posterior inference for each parameter and the deviance (-2 log likelihood). For details regarding the description of the software that we used, we refer to Appendix C of Gelman, Carlin, Stern and Rubin (2004).

A Bayesian way to check the fit of a model to data is to draw simulated values from the posterior predictive distribution of replicated data and compare these samples to observed data. A wide departure between the generated and the observed data indicates lack of fit of the model. Following Gelman *et al.* (2004), we calculated the Bayesian *p*-values for checking the goodness-of-fit of the proposed Bayesian models. The general rationale behind such calculations is as follows. Let y denote the vector of observed data, ξ the vector of unknown parameters, $f(y \mid \xi)$ the density of y given ξ and $\Pi(\xi \mid y)$, the posterior density of ξ given y. Suppose one has drawn samples $\xi^{(1)}, \ldots, \xi^{(R)}$ from this posterior distribution using MCMC simulation. Simulate now R hypothetical replicates of the data, say $y^{(1)}, \ldots, y^{(R)}$, where $y^{(l)}, (l = 1, \ldots, R)$ is drawn from the conditional distribution of y given the simulated $\xi^{(l)}$. If the model is reasonably accurate, these hypothetical replicates should be similar to the observed data y. This is formally done by first choosing a divergence variable, say $d(y, \xi)$ which will have an extreme value if the data y are in complete disagreement with the given model. Then a *p*-value is estimated by the proportion of cases in which the simulated divergence variable exceeds the realized value of the same. Thus the estimated *p*-value (usually referred to as the posterior predictive *p*-value) is equal to $R^{-1}\sum_{l=1}^{R} I_{[d(y^{(l)}, \xi^{(l)}) \geq d(y, \xi^{(l)})]}$, where I is the usual indicator function. One way of checking the goodness of fit of the model is by a scatter plot of realized values $d(y, \xi^{(l)})$ against the predictive values $d(y^{(l)}, \xi^{(l)})$ on the same scale. A good fit is indicated by about half the points in the scatter plot falling above the 45^0 line, and half falling below. In other words, for large samples, the estimated *p*-value will not be far away from one half. Of course, one may also carry out a graphical analysis by using different plots for different subgroups, thereby allowing visualization of possible local model failure which may otherwise be obscured in the aggregate plot.

There are several possible choices of the divergence variable *d*. We considered a particular one in the present case. Noting that $E(Y_{ij} \mid p_{ij}) = p_{ij} = \exp(\theta_{ij})/(1+\exp(\theta_{ij}))$, one can consider the squared standardized residuals $((y_{ij} - p_{ij}^{(l)})^2)/(p_{ij}^{(l)}(1-p_{ij}^{(l)}))$, where $p_{ij}^{(l)} = \exp(\theta_{ij}^{(l)})/(1+\exp(\theta_{ij}^{(l)}))$ are the generated values of the p_{ij} from the l^{th} iteration. Then the divergence variable *d* is

$$d(y, p^{(l)}) = \sum_{i=1}^{95} \sum_{j=1}^{n_i} \frac{(y_{ij} - p_{ij}^{(l)})^2}{p_{ij}^{(l)}(1-p_{ij}^{(l)})}$$

$$d(y^{(l)}, p^{(l)}) = \sum_{i=1}^{95} \sum_{j=1}^{n_i} \frac{(y_{ij}^{(l)} - p_{ij}^{(l)})^2}{p_{ij}^{(l)}(1-p_{ij}^{(l)})}.$$

Clearly, there are other possible choices of *d*. Gelfand and Ghosh (1998) proposed a number of divergence measures, and studied their properties.

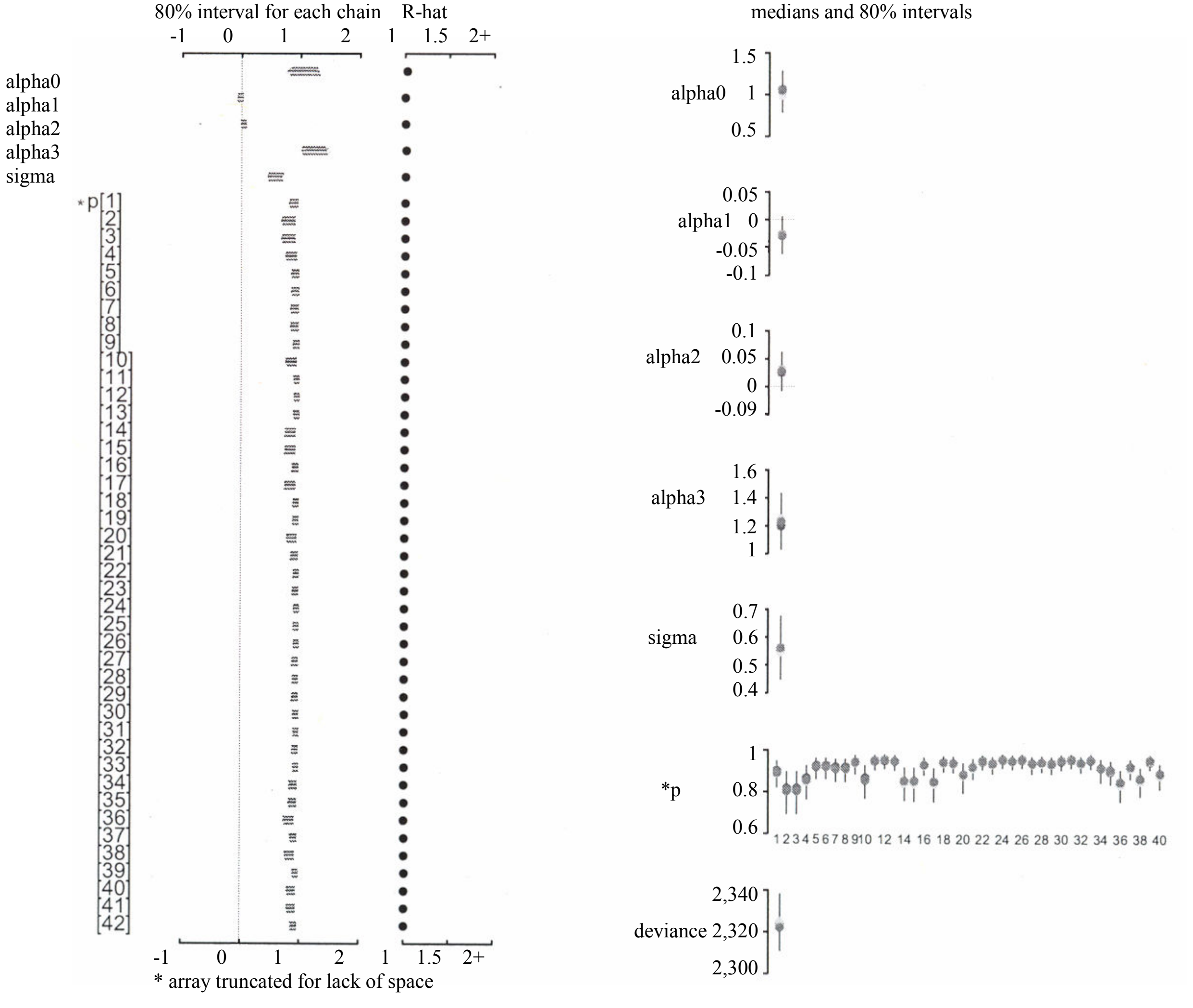


Figure 1 Bugs model at "asian_model.bug", 5 chains, each with 1,000 iterations

For the hierarchical Bayesian logistic regression model, the estimated *p*-value is 0.4216 for $(c, d) = (0.02, 0.02)$. The other choices of (c, d) produce similar values. The *p*-value bigger than 0.3 is usually treated as a good fit. Thus the proposed HB procedure seems to work well in this situation.

We have also calculated the $p_D =$ var(deviance)/2 and the *deviance information criterion* DIC or the estimated predictive deviance. The p_D can be thought of as the number of ‘unconstrained parameters in the model, where a parameter is counted as 1 if it is a part of the original model (data distribution) and is 0 if it is associated with any prior distribution. The DIC is estimated as

$$\text{DIC} = 2\hat{D}(y, \theta^{(l)}) - \hat{D}(y, \hat{\theta})$$

where $\hat{D}(y, \hat{\theta})$ is the deviance calulated at the estimated parameters and $\hat{D}(y, \theta^{(l)})$ is the esimated deviance using posterior simulation. For details, see Gelman *et al.* (2004).

For our HB analysis $p_D = 56.75$ and DIC = 2,414.41. Usually the p_D and DIC are used as criteria of model fitting and to select the model with best predictive power. Thus, we fit also the simple logistic regression model (current model without any random effects) which means that there is no data pooling, and the estimated p_D and DIC are 22.60 and 2,379.55 respectively. The corresponding *p*-value is 0.3848. Thus the proposed model seems to fit the data reasonably well

7. Summary, future work and discussion

Estimating the proportion of uninsured people, especially among the minorities, is definitely a problem of great importance, and is likely to affect the policy making of Federal and State agencies. We have just started addressing this very important issue, and have provided both empirical and hierarchical Bayesian small domain estimates for the Asian subpopulation cross-classified by age, sex and other demographic characteristics. We have also discussed the adequacy of our model fit via posterior predictive *p*-value. Much work remains to be done however. In particular, we want to extend the present findings to the analysis of bivariate and multivariate binary data.

As pointed out by a reviewer, the present analysis ignores household clustering in the likelihood, since the original survey was a household survey, and very definitely, insurance coverage is correlated within households. However, we have assumed only a conditionally independent hierarchical model given the covariates and the random effects. Once, we have assigned distributions to the random effects, and subsequently distributions to the regression coefficients and the variance components, dependence is built automatically in the final model, both at the unit and domain levels. Moreover, as mentioned earlier, adequacy of the hierarchical Bayesian model has been tested through posterior predictive *p*-values.

As a final comment, the research presented here is for illustrative purposes only. Implementation of this method for policy related matters would require further consideratuions of the methods and adherence to institutional standards for official policy release.

Acknowledgements

We thank the AE and two reviewers for their constructive comments on earlier drafts of the paper. The first author’s research was partially funded by NSF Grants SES-0317589 and SES-0631426. The fourth author’s research was partially supported by an NSF Grant SES-0318184. The research was also partially supported by NCHS/CDC under Project Number 282286285 entitled “Model-Based Subdomain Estimates”. The findings and conclusions in this paper are those of the authors and do not necessarily represent the views of the National Center for Health Statistics, Centers for Disease Control and Prevention.

Appendix

Proof of Theorem 1.

$$\text{MSE}(\hat{\mu}_{iw}^{\text{EB}})$$

$$= E(\hat{\mu}_{iw}^{\text{EB}} - \hat{\mu}_{iw})^2 = E\left(\sum_{j=1}^{n_i} w_{ij}(\hat{p}_{ij}^{\text{EB}} - p_{ij})\right)^2$$

$$= E\left(\sum_{j=1}^{n_i} w_{ij}(\hat{p}_{ij}^{\text{EB}} - \hat{p}_{ij}^{B} + \hat{p}_{ij}^{B} - p_{ij})\right)^2$$

$$= E\left(\sum_{j=1}^{n_i} w_{ij}(\hat{p}_{ij}^{\text{EB}} - \hat{p}_{ij}^{B})\right)^2 + E\left(\sum_{j=1}^{n_i} w_{ij}(\hat{p}_{ij}^{B} - p_{ij})\right)^2$$

$$+ 2E\left[\left(\sum_{j=1}^{n_i} w_{ij}(\hat{p}_{ij}^{\text{EB}} - \hat{p}_{ij}^{B})\right)\left(\sum_{j=1}^{n_i} w_{ij}(\hat{p}_{ij}^{B} - p_{ij})\right)\right].$$

Noting that $E(p_{ij} \mid \boldsymbol{y}) = \hat{p}_{ij}^{B}$,

$$E\left[\left(\sum_{j=1}^{n_i} w_{ij}(\hat{p}_{ij}^{\text{EB}} - \hat{p}_{ij}^{B})\right)\left(\sum_{j=1}^{n_i} w_{ij}(\hat{p}_{ij}^{B} - p_{ij})\right)\right]$$

$$= E\left[\left(\sum_{j=1}^{n_i} w_{ij}(\hat{p}_{ij}^{\text{EB}} - \hat{p}_{ij}^{B})\right) \times E\left(\sum_{j=1}^{n_i} w_{ij}(\hat{p}_{ij}^{B} - p_{ij}) \mid \boldsymbol{y}\right)\right] = 0.$$

Hence,

$$\mathrm{MSE}(\hat{\mu}_{iw}^{\mathrm{EB}}) = E\left(\sum_{j=1}^{n_i} w_{ij}(\hat{p}_{ij}^{\mathrm{EB}} - \hat{p}_{ij}^{B})\right)^2 + E\left(\sum_{j=1}^{n_i} w_{ij}(\hat{p}_{ij}^{B} - p_{ij})\right)^2. \qquad \text{(A.1)}$$

But

$$E\left(\sum_{j=1}^{n_i} w_{ij}(\hat{p}_{ij}^{B} - p_{ij})\right)^2 = \sum_{j=1}^{n_i} w_{ij}^2 E(\hat{p}_{ij}^{B} - p_{ij})^2.$$

Next we calculate

$$\begin{aligned}
&E(\hat{p}_{ij}^{B} - p_{ij})^2 \\
&= E\left[\frac{1}{\lambda+1}y_{ij} + \frac{\lambda}{\lambda+1}m_{ij}(\boldsymbol{b}) - p_{ij}\right]^2 \\
&= E\left[\frac{1}{\lambda+1}(y_{ij} - p_{ij}) + \frac{\lambda}{\lambda+1}(m_{ij}(\boldsymbol{b}) - p_{ij})\right]^2 \\
&= \frac{1}{(\lambda+1)^2}E(y_{ij}-p_{ij})^2 + \frac{\lambda^2}{(\lambda+1)^2}E\left(m_{ij}(\boldsymbol{b}) - p_{ij}\right)^2 \\
&\quad + \frac{2\lambda}{(\lambda+1)^2}E(y_{ij}-p_{ij})(m_{ij}(\boldsymbol{b})-p_{ij}) \\
&= \frac{1}{(\lambda+1)^2}E(p_{ij}(1-p_{ij})) + \frac{\lambda^2}{(\lambda+1)^2}V(p_{ij}) + 0 \\
&= \frac{1}{(\lambda+1)^2}\left(\frac{\lambda}{(\lambda+1)}m_{ij}(\boldsymbol{b})(1-m_{ij}(\boldsymbol{b}))\right) \\
&\quad + \frac{\lambda^2}{(\lambda+1)^2}\left(\frac{m_{ij}(\boldsymbol{b})(1-m_{ij}(\boldsymbol{b}))}{\lambda+1}\right) \\
&= \frac{\lambda m_{ij}(\boldsymbol{b})(1-m_{ij}(\boldsymbol{b}))}{(\lambda+1)^2},
\end{aligned}$$

so that

$$E\left[\sum_{j=1}^{n_i} w_{ij}(\hat{p}_{ij}^{B} - p_{ij})\right]^2 = \frac{\lambda}{(\lambda+1)^2}\sum_{j=1}^{n_i} w_{ij}^2 m_{ij}(\boldsymbol{b})(1-m_{ij}(\boldsymbol{b})). \qquad \text{(A.2)}$$

Finally, we calculate,

$$\begin{aligned}
&E\left[\sum_{j=1}^{n_i} w_{ij}(\hat{p}_{ij}^{\mathrm{EB}} - \hat{p}_{ij}^{B})\right]^2 \\
&= \frac{\lambda^2}{(\lambda+1)^2}E\left[\sum_{j=1}^{n_i} w_{ij}(m_{ij}(\hat{\boldsymbol{b}}) - m_{ij}(\boldsymbol{b}))\right]^2 \\
&= \frac{\lambda^2}{(\lambda+1)^2}E\Bigg[\sum_{j=1}^{n_i} w_{ij}^2(m_{ij}(\hat{\boldsymbol{b}}) - m_{ij}(\boldsymbol{b}))^2 \\
&\qquad + \sum\sum_{1\le j\ne j'\le n_i} w_{ij}w_{ij'}(m_{ij}(\hat{\boldsymbol{b}}) - m_{ij}(\boldsymbol{b})) \\
&\qquad (m_{ij'}(\hat{\boldsymbol{b}}) - m_{ij'}(\boldsymbol{b}))\Bigg]. \qquad \text{(A.3)}
\end{aligned}$$

By two-step Taylor expansion,

$$m_{ij}(\hat{\boldsymbol{b}}) \doteq m_{ij}(\boldsymbol{b}) + \left(\frac{\partial m_{ij}(\boldsymbol{b})}{\partial \boldsymbol{b}}\right)^T (\hat{\boldsymbol{b}} - \boldsymbol{b}) + \frac{1}{2}(\hat{\boldsymbol{b}} - \boldsymbol{b})^T \frac{\partial^2 m_{ij}(\boldsymbol{b})}{\partial \boldsymbol{b}\partial \boldsymbol{b}^T}(\hat{\boldsymbol{b}} - \boldsymbol{b}).$$

Noting that $(\partial^2 m_{ij}(\boldsymbol{b}))/(\partial \boldsymbol{b}\partial \boldsymbol{b}^T) = (1-2m_{ij}(\boldsymbol{b}))m_{ij}(\boldsymbol{b})\,(1-m_{ij}(\boldsymbol{b}))\boldsymbol{x}_{ij}\boldsymbol{x}_{ij}^T$, it follows that

$$\begin{aligned}
&E[m_{ij}(\hat{\boldsymbol{b}}) - m_{ij}(\boldsymbol{b})]^2 \\
&\doteq E\Bigg[\, m_{ij}(\boldsymbol{b})(1-m_{ij}(\boldsymbol{b}))\boldsymbol{x}_{ij}^T(\hat{\boldsymbol{b}} - \boldsymbol{b}) + \frac{1}{2}(\hat{\boldsymbol{b}} - \boldsymbol{b})^T \\
&\qquad m_{ij}(\boldsymbol{b})(1-m_{ij}(\boldsymbol{b}))(1-2m_{ij}(\boldsymbol{b}))\boldsymbol{x}_{ij}\boldsymbol{x}_{ij}^T(\hat{\boldsymbol{b}} - \boldsymbol{b})\Bigg]^2 \\
&= m_{ij}^2(\boldsymbol{b})(1-m_{ij}(\boldsymbol{b}))^2 E\Bigg[\,\boldsymbol{x}_{ij}^T(\hat{\boldsymbol{b}} - \boldsymbol{b}) + \frac{1}{2}(1-2m_{ij}(\boldsymbol{b})) \\
&\qquad (\hat{\boldsymbol{b}} - \boldsymbol{b})^T \boldsymbol{x}_{ij}\boldsymbol{x}_{ij}^T(\hat{\boldsymbol{b}} - \boldsymbol{b})\Bigg]^2. \qquad \text{(A.4)}
\end{aligned}$$

The first neglected term is $O_p(\|\hat{\boldsymbol{b}} - \boldsymbol{b}\|^3)$. From Sarkar and Ghosh (1998), $\hat{\boldsymbol{b}} - \boldsymbol{b}$ is asymptotically $\mathrm{N}(0, \boldsymbol{\Sigma}^{-1}(\boldsymbol{b}))$, where $\boldsymbol{\Sigma}(\boldsymbol{b})$ is defined before Theorem 1. With the assumption that $\boldsymbol{\Sigma}(\boldsymbol{b}) = O_e(k)$, it follows that $\boldsymbol{\Sigma}^{-1}(\boldsymbol{b}) = O_e(k^{-1})$. Thus, $\|\hat{\boldsymbol{b}} - \boldsymbol{b}\| = O_p(k^{-1/2})$. Hence, the first neglected term is $O_p(k^{-3/2})$. Next, we observe that

$$\begin{aligned}
E[\boldsymbol{x}_{ij}^T(\hat{\boldsymbol{b}} - \boldsymbol{b})]^2 &= E[(\hat{\boldsymbol{b}} - \boldsymbol{b})^T \boldsymbol{x}_{ij}\boldsymbol{x}_{ij}^T(\hat{\boldsymbol{b}} - \boldsymbol{b})] \\
&= tr[\boldsymbol{x}_{ij}\boldsymbol{x}_{ij}^T E(\hat{\boldsymbol{b}} - \boldsymbol{b})(\hat{\boldsymbol{b}} - \boldsymbol{b})^T]. \qquad \text{(A.5)}
\end{aligned}$$

In order to find $E[(\hat{\boldsymbol{b}}-\boldsymbol{b})(\hat{\boldsymbol{b}}-\boldsymbol{b})^T]$, we proceed as follows: Let $\boldsymbol{T}(\boldsymbol{b})=\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-m_{ij}(\boldsymbol{b}))\boldsymbol{x}_{ij}$ so that $\boldsymbol{T}(\hat{\boldsymbol{b}})=\boldsymbol{0}$. By one-step Taylor expansion, $\boldsymbol{0}=\boldsymbol{T}(\hat{\boldsymbol{b}})=\boldsymbol{T}(\boldsymbol{b})+[\nabla\boldsymbol{T}(\boldsymbol{b})]^T$ $(\hat{\boldsymbol{b}}-\boldsymbol{b})+O_p(n_T^{-1})$, where

$$\begin{aligned}\nabla\boldsymbol{T}(\boldsymbol{b}) &= -\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(\frac{\partial m_{ij}(\boldsymbol{b})}{\partial\boldsymbol{b}}\right)\boldsymbol{x}_{ij}^T\\ &= -\sum_{i=1}^{k}\sum_{j=1}^{n_i} m_{ij}(\boldsymbol{b})(1-m_{ij}(\boldsymbol{b}))\boldsymbol{x}_{ij}\boldsymbol{x}_{ij}^T\\ &= -\boldsymbol{X}^T\boldsymbol{M}(\boldsymbol{I}-\boldsymbol{M})\boldsymbol{X}\\ &= -\boldsymbol{\Sigma}(\boldsymbol{b}). \qquad (A.6)\end{aligned}$$

Thus, $\hat{\boldsymbol{b}}-\boldsymbol{b}=\boldsymbol{\Sigma}^{-1}\boldsymbol{T}(\boldsymbol{b})+O_p(n_T^{-1})$. Since $V(y_{ij})=m_{ij}(\boldsymbol{b})$ $(1-m_{ij}(\boldsymbol{b}))$, $V(T(\boldsymbol{b}))=\boldsymbol{\Sigma}(\boldsymbol{b})$. Hence $E[(\hat{\boldsymbol{b}}-\boldsymbol{b})(\hat{\boldsymbol{b}}-\boldsymbol{b})^T]=$ $\boldsymbol{\Sigma}^{-1}(\boldsymbol{b})+O(n_T^{-3/2})$. Also, in (8.5), we have $E(\boldsymbol{x}_{ij}^T)(\hat{\boldsymbol{b}}-\boldsymbol{b})^2=$ $tr(\boldsymbol{x}_{ij}\boldsymbol{x}_{ij}^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{b}))+O_p(n_T^{-1})$. Accordingly, by (A.4) and (A.5), we have the approximation

$$E[m_{ij}(\hat{\boldsymbol{b}})-m_{ij}(\boldsymbol{b})]^2=m_{ij}^2(\boldsymbol{b})(1-m_{ij}(\boldsymbol{b}))^2\boldsymbol{x}_{ij}^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{b})\boldsymbol{x}_{ij} \qquad (A.7)$$

which is correct up to $O(n_T^{-1})$ by our assumption.
Note that the neglected term

$$E[\boldsymbol{x}_{ij}^T(\hat{\boldsymbol{b}}-\boldsymbol{b})(1-2m_{ij}(\boldsymbol{b}))(\hat{\boldsymbol{b}}-\boldsymbol{b})^T\boldsymbol{x}_{ij}\boldsymbol{x}_{ij}^T(\hat{\boldsymbol{b}}-\boldsymbol{b})]$$

$$=O(n_T^{-3/2})$$

since

$$\begin{aligned}&E[\boldsymbol{x}_{ij}^T(\hat{\boldsymbol{b}}-\boldsymbol{b})(1-2m_{ij}(\boldsymbol{b}))(\hat{\boldsymbol{b}}-\boldsymbol{b})^T\boldsymbol{x}_{ij}\boldsymbol{x}_{ij}^T(\hat{\boldsymbol{b}}-\boldsymbol{b})]\\ &=(1-2m_{ij}(\boldsymbol{b}))E[\boldsymbol{x}_{ij}^T(\hat{\boldsymbol{b}}-\boldsymbol{b})(\hat{\boldsymbol{b}}-\boldsymbol{b})^T\boldsymbol{x}_{ij}]\\ &=(1-2m_{ij}(\boldsymbol{b}))E[\boldsymbol{x}_{ij}^T(\hat{\boldsymbol{b}}-\boldsymbol{b})]^3\\ &=O(n_T^{-3/2}).\end{aligned}$$

Similarly, note that $E[\boldsymbol{x}_{ij}^T(\hat{\boldsymbol{b}}-\boldsymbol{b})]^4=O(n_T^{-2})$ and

$$\begin{aligned}&E[(m_{ij}(\hat{\boldsymbol{b}})-m_{ij}(\boldsymbol{b}))(m_{ij'}(\hat{\boldsymbol{b}})-m_{ij'}(\boldsymbol{b}))]\\ &=m_{ij}(\boldsymbol{b})(1-m_{ij}(\boldsymbol{b}))m_{ij'}(\boldsymbol{b})(1-m_{ij'}(\boldsymbol{b}))\boldsymbol{x}_{ij}^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{b})\boldsymbol{x}_{ij'}\\ &\quad+O(n_T^{-3/2}). \qquad (A.8)\end{aligned}$$

This leads to

$$\begin{aligned}&E\left[\sum_{j=1}^{n_i}w_{ij}(\hat{p}_{ij}^{\text{EB}}-\hat{p}_{ij}^{B})\right]^2\\ &=\frac{\lambda^2}{(\lambda+1)^2}\Bigg[\sum_{j=1}^{n_i}w_{ij}^2m_{ij}^2(\boldsymbol{b})(1-m_{ij}(\boldsymbol{b}))^2\boldsymbol{x}_{ij}^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{b})\boldsymbol{x}_{ij}\\ &\qquad+\sum\sum_{1\le j\ne j'\le n_i}w_{ij}w_{ij'}m_{ij}(\boldsymbol{b})(1-m_{ij}(\boldsymbol{b}))m_{ij'}(\boldsymbol{b})\\ &\qquad(1-m_{ij'}(\boldsymbol{b}))\boldsymbol{x}_{ij}^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{b})\boldsymbol{x}_{ij'}\Bigg]+O(n_T^{-3/2})\\ &=\frac{\lambda^2}{(\lambda+1)^2}\left[\sum_{j=1}^{n_i}w_{ij}m_{ij}(\boldsymbol{b})(1-m_{ij}(\boldsymbol{b}))\boldsymbol{x}_{ij}\right]^T\\ &\times\boldsymbol{\Sigma}^{-1}(\boldsymbol{b})\left[\sum_{j=1}^{n_i}w_{ij}m_{ij}(\boldsymbol{b})(1-m_{ij}(\boldsymbol{b}))\boldsymbol{x}_{ij}\right]+O(n_T^{-3/2}). \qquad (A.9)\end{aligned}$$

Since $\boldsymbol{\Sigma}^{-1}(\boldsymbol{b})=O(k^{-1})$, and $n_T=O_e(k)$ by our assumption, the theorem follows from (A.2) and (A.9).

Proof of Theorem 2. We first note that $\hat{\boldsymbol{b}}=\boldsymbol{b}+$ $O_p(n_T^{-1/2})=\boldsymbol{b}+O_p(K^{-1})$ and $\boldsymbol{\Sigma}^{-1}(\boldsymbol{b})=O(k^{-1})$. Hence, the second term in the right hand side of (8.7) is approximated by

$$\begin{aligned}&c\left[\sum_{j=1}^{n_i}w_{ij}m_{ij}(\hat{\boldsymbol{b}})(1-m_{ij}(\hat{\boldsymbol{b}}))\boldsymbol{x}_{ij}\right]^T\\ &\qquad\boldsymbol{\Sigma}^{-1}(\hat{\boldsymbol{b}})\left[\sum_{j=1}^{n_i}w_{ij}m_{ij}(\hat{\boldsymbol{b}})(1-m_{ij}(\hat{\boldsymbol{b}}))\boldsymbol{x}_{ij}\right] \qquad (A.10)\end{aligned}$$

$(c=\lambda^2/(1+\lambda)^2)$ which is correct up to $\mathrm{O}(k^{-1})$.
However, if we estimate $m_{ij}(\boldsymbol{b})(1-m_{ij}(\boldsymbol{b}))$ simply by $m_{ij}(\hat{\boldsymbol{b}})(1-m_{ij}(\hat{\boldsymbol{b}}))$, we will be ignoring the $O(k^{-1})$ term. Thus, we need a careful approximation of the bias $E(\hat{\boldsymbol{b}}-\boldsymbol{b})$ to achieve the desired approximation. To this end, we follow Cox and Snell (1968).
We begin with the identity

$$\begin{aligned}&E[m_{ij}(\hat{\boldsymbol{b}})(1-m_{ij}(\hat{\boldsymbol{b}}))]\\ &=E[(m_{ij}(\boldsymbol{b})+m_{ij}(\hat{\boldsymbol{b}})-m_{ij}(\boldsymbol{b}))(1-m_{ij}(\boldsymbol{b})+m_{ij}(\boldsymbol{b})-m_{ij}(\hat{\boldsymbol{b}}))]\\ &=m_{ij}(\boldsymbol{b})(1-m_{ij}(\boldsymbol{b}))+(1-2m_{ij}(\boldsymbol{b}))E[m_{ij}(\hat{\boldsymbol{b}})-m_{ij}(\boldsymbol{b})]\\ &-E[m_{ij}(\hat{\boldsymbol{b}})-m_{ij}(\boldsymbol{b})]^2.\end{aligned}$$

Now, again by a two-step Taylor expansion,

$$E[m_{ij}(\hat{\boldsymbol{b}}) - m_{ij}(\boldsymbol{b})] = \left[\frac{\partial m_{ij}(\boldsymbol{b})}{\partial \boldsymbol{b}}\right]^T E(\hat{\boldsymbol{b}} - \boldsymbol{b})$$

$$+\frac{1}{2}E\left[(\hat{\boldsymbol{b}}-\boldsymbol{b})^T \frac{\partial^2 m_{ij}(\boldsymbol{b})}{\partial\boldsymbol{b}\partial\boldsymbol{b}^T}(\hat{\boldsymbol{b}}-\boldsymbol{b})\right] + O(n_T^{-3/2}).$$

In order to find $E(\hat{\boldsymbol{b}}-\boldsymbol{b}),$ we proceed as follows. We begin with the second order Taylor expansion

$$0 = T_r(\hat{\boldsymbol{b}}) = T_r(\boldsymbol{b}) + \sum_{s=1}^{p}(\hat{b}_s - b_s)\frac{\partial T_r(\boldsymbol{b})}{\partial b_s}$$

$$+\frac{1}{2}\sum_{s=1}^{p}\sum_{t=1}^{p}(\hat{b}_s - b_s)(\hat{b}_t - b_t)\frac{\partial^2 T_r(\boldsymbol{b})}{\partial b_s \partial b_t} + O(n_T^{-3/2}).$$

Taking expectations and following Cox and Snell (1968),

$$0 = E(T_r(\hat{\boldsymbol{b}}))$$

$$= \sum_{s=1}^{p}\left[E(\hat{b}_s - b_s)E\left(\frac{\partial T_r(\boldsymbol{b})}{\partial b_s}\right) + \mathrm{Cov}\left(\hat{b}_s - b_s, \frac{\partial T_r(\boldsymbol{b})}{\partial b_s}\right)\right]$$

$$+\frac{1}{2}\sum_{s=1}^{p}\sum_{t=1}^{p}(E(\hat{b}_s - b_s)(\hat{b}_t - b_t))\left(\frac{\partial^2 T_r(\boldsymbol{b})}{\partial b_s \partial b_t}\right)$$

$$+\frac{1}{2}\sum_{s=1}^{p}\sum_{t=1}^{p}\mathrm{Cov}\left[(\hat{b}_s - b_s)(\hat{b}_t - b_t), \left(\frac{\partial^2 T_r(\boldsymbol{b})}{\partial b_s \partial b_t}\right)\right]$$

$$+O(n_T^{-3/2}) = -\sum_{s=1}^{p}E(\hat{b}_s - b_s)\sigma_{rs}$$

$$+\sum_{s=1}^{p}\sum_{u=1}^{p}\mathrm{Cov}\left[\sigma^{su}(\boldsymbol{b})T_u(\boldsymbol{b}), \frac{\partial T_r(\boldsymbol{b})}{\partial b_s}\right]$$

$$+\frac{1}{2}\sum_{s=1}^{p}\sum_{t=1}^{p}\sigma^{st}E\left[\frac{\partial^2 T_r(\boldsymbol{b})}{\partial b_s \partial b_t}\right] + O(n_T^{-3/2}). \qquad \text{(A.11)}$$

Note $\mathrm{Cov}[\sigma^{su}(\boldsymbol{b})T_u(\boldsymbol{b}), \partial T_r(\boldsymbol{b})/\partial b_s] = 0$ since $\partial T_r(\boldsymbol{b})/\partial b_s$ is a constant independent of the y_{ij}.
Similarly,

$$\mathrm{Cov}\left[(\hat{b}_s - b_s)(\hat{b}_t - b_t), \left(\frac{\partial^2 T_r(\boldsymbol{b})}{\partial b_s \partial b_t}\right)\right] = 0.$$

Also, let

$$K_{rst} = E\left[\frac{\partial^2 T_r(\boldsymbol{b})}{\partial b_s \partial b_t}\right]$$

$$= \frac{\partial}{\partial b_t}\sum_{i=1}^{k}\sum_{j=1}^{n_i} - m_{ij}(\boldsymbol{b})(1 - m_{ij}(\boldsymbol{b}))x_{ijr}\, x_{ijs}$$

$$= -\sum_{i=1}^{k}\sum_{j=1}^{n_i}(1 - 2m_{ij}(\boldsymbol{b}))m_{ij}(\boldsymbol{b})(1 - m_{ij}(\boldsymbol{b}))x_{ijr}\, x_{ijs}\, x_{ijt}. \qquad \text{(A.12)}$$

Thus, one has

$$\sum_{s=1}^{k}\sigma_{rs}E(\hat{b}_s - b_s) \doteq \sum_{s=1}^{k}\sum_{t=1}^{p}\sigma^{su}K_{rst},\ r = 1, ..., p.$$

In matrix notations, one gets

$$\boldsymbol{\Sigma}\, E(\hat{\boldsymbol{b}} - \boldsymbol{b}) = \frac{1}{2}\begin{pmatrix} tr(\boldsymbol{\Sigma}^{-1}\boldsymbol{K}_1) \\ . \\ . \\ . \\ tr(\boldsymbol{\Sigma}^{-1}\boldsymbol{K}_p) \end{pmatrix}$$

where $\boldsymbol{K}_r = ((K_{rst})).$

Hence,

$$E(\hat{\boldsymbol{b}} - \boldsymbol{b}) \doteq \frac{1}{2}\boldsymbol{\Sigma}^{-1}\begin{pmatrix} tr(\boldsymbol{\Sigma}^{-1}\boldsymbol{K}_1) \\ . \\ . \\ . \\ tr(\boldsymbol{\Sigma}^{-1}\boldsymbol{K}_p) \end{pmatrix} + O(n_T^{-3/2}).$$

Since $n_T = O_e(k),$ the theorem follows.

References

Botman, S.L., Moore, T.F., Moriarity, C.L. and Parsons, V.L. (2000). Design and estimation for the National Health Interview Survey, 1995-2004. *Vital and Health Statistics*, 2, 130.

Cox, D.R., and Snell, E.J. (1968). A general distribution of residuals (with discussion). *Journal of the Royal Statistical Society*, Series B, 30, 248-275.

Ghosh, M., and Maiti, T. (2004). Small-area estimation based on natural exponential family quadratic variance function models and survey weights. *Biometrika*, 91, 95-112.

Ghosh, M., Natarajan, K., Stroud, T.W.F. and Carlin, B.P. (1998). Generalized linear models for small area estimation. *Journal of the American Statistical Association*, 93, 273-282.

Gelman, A., Carlin, J., Stern, H. and Rubin, D. (2004). *Bayesian Data Analysis*. Chapman & Hall/CRC.

Gelman, A., and Rubin, D. (1992). Inference from iterative simulation using multiple sequences (with discussion). *Statistical Science*, 457-511.

Godambe, V.P., and Thompson, M.E. (1989). An extension of quasi-likelihood estimation (with discussion). *Journal of Statistical Planning and Infererence*, 22, 137-152.

Morris, C. (1982). Natural exponential families with quadratic variance functions. *Annals of Statistics*, 10, 65-80.

Morris, C. (1983). Natural exponential families with quadratic variance functions. *Annals of Statistics*, 11, 515-529.

Prasad, N.G.N., and Rao, J.N.K. (1990). The estimation of the mean squared error of small area estimators. *Journal of the American Statistical Association*, 85, 163-171.

Sarkar, S., and Ghosh, M. (1998). Empirical Bayes estimation of local area means for NEF-QVF superpopulations. *Sankhyā*, Series B, 60, 464-487.

Tanner, M.A. (1996). *Tools for Statistical Inference*. New York: Springer.

Survey Methodology, June 2009
Vol. 35, No. 1, pp. 67-83
Statistics Canada, Catalogue No. 12-001-X

A nonparametric test for residual seasonality

Tucker McElroy and Scott Holan [1]

Abstract

Peaks in the spectrum of a stationary process are indicative of the presence of stochastic periodic phenomena, such as a stochastic seasonal effect. This work proposes to measure and test for the presence of such spectral peaks via assessing their aggregate slope and convexity. Our method is developed nonparametrically, and thus may be useful during a preliminary analysis of a series. The technique is also useful for detecting the presence of residual seasonality in seasonally adjusted data. The diagnostic is investigated through simulation and an extensive case study using data from the U.S. Census Bureau and the Organization for Economic Co-operation and Development (OECD).

Key Words: Multiple testing; Nonparametric density estimation; Seasonal adjustment; Spectral density.

1. Introduction

The presence of a peak in the spectrum of a stationary process is indicative of periodic behavior, such as seasonality or a trading day effect. There is a widespread interest in the identification of such peaks in the engineering and econometrics literature, since a pronounced spectral node will exert a potent influence on the dynamics of the stochastic process. A peak indicates a range of frequencies that offer a relatively large contribution to the overall variance of the stochastic process. If the strength of the peak, assessed through its height and width relative to neighboring values, is sufficiently pronounced, any model of the dynamics that ignores the corresponding periodicities will be misspecified. In both engineering and econometrics, one may be interested in signal extraction or forecasting, both of which are sensitive to the presence of spectral peaks.

By a spectral peak, we refer to a region of the spectral density that has greater spectral mass than its immediate neighbors; a more precise definition is developed below. Due to the applications that we have in mind, our peaks have finite height, and thus correspond to stochastic periodic effects in a stationary process. Thus, we are not principally concerned with the detection of fixed (deterministic) periodic effects, nor with nonstationary periodic phenomena (though we make some extensions to this case in Section 3.4 below), as both of these correspond to a spectral peak with infinite height. The vast literature dealing with the detection of fixed effects is discussed in Priestley (1981); for our applications the periodic aspects of the data are not fixed, but instead evolve over time.

In this paper we focus on the application to seasonal adjustment. Specifically, we concentrate on so-called seasonal peaks, which may occur at the seasonal frequencies (assuming a monthly sampling interval) $\pi/6$, $2\pi/6$, $3\pi/6$, $4\pi/6$, $5\pi/6$, and $6\pi/6$. The detection of seasonality and residual seasonality presents an important practical problem in federal statistics, and the spectrum is a natural tool towards this end. The frequency domain approach to the detection and analysis of seasonality enjoys wide popularity, because it provides a very natural way to view quasi-periodic behavior. In fact, seasonality is − informally speaking − characterized by the presence of at least one seasonal peak in the spectrum (Nerlove 1964). Frequency domain methods are now employed in X-12-ARIMA (Findley, Monsell, Bell, Otto and Chen 1998) and are part of TRAMO-SEATS (Maravall and Caporello 2004), the two most widely-used seasonal adjustment programs available to the public. Note that frequency domain methods can be implemented via either a parametric (*i.e.*, model-based) or nonparametric approach. We develop a nonparametric diagnostic, which can be invoked to determine the efficacy of *any* seasonal adjustment procedure, either model-based or nonparametric. As noted in Findley, Monsell, Bell, Otto, and Chen (1998), the use of fixed periodic functions alone to model seasonality is typically inadequate for economic data (also see the discussion in Bell and Hillmer 1984).

Spectral peaks at seasonal frequencies in a seasonally adjusted series may indicate inadequacy of the seasonal filters − see Soukup and Findley (1999) for a discussion. At a minimum, seasonal adjustment filters should remove *nonstationary* seasonality and any fixed periodic effects − those phenomena in the observed series that contribute a seasonal pole to the spectrum. However, there is a consensus among seasonal adjusters that it is also desirable to remove some aspects of the *stationary* seasonality as well − hence the explosion of effort in developing model-based seasonal adjustment filters (Bell and Hillmer 1984).

[1] Tucker McElroy, Statistical Research Division, U.S. Census Bureau, 4600 Silver Hill Road, Washington, D.C. 20233-9100. E-mail: tucker.s.mcelroy@census.gov; Scott Holan, Department of Statistics, University of Missouri-Columbia, 146 Middlebush Hall, Columbia, MO, 65211-6100. E-mail: holans@missouri.edu.

The most important prior literature on this topic is Soukup and Findley (1999), which proposes using an autoregressive spectrum to find "visually significant" peaks – essentially the value of the spectrum at each seasonal frequency (or trading day frequency) is compared to its nearest neighbors, and is classified as a peak if the discrepancy is suitably large. This method is currently implemented in the X-12-ARIMA program from the U.S. Census Bureau (2002). One limitation of this approach is that it has really no statistical component: the significance is not statistical – *i.e.*, it is not associated with a hypothesis test – and the thresholds to determine "visual significance" are determined in an *ad hoc* fashion. This paper provides a statistical significance test for peak detection, and can thus be used to offer supplementary statistical evidence of the presence of a peak.

Another related paper is Newton and Pagano (1983), which develops consistent estimators for the local maximizers of the spectrum. Our approach is slightly different, in that we already know the frequencies of interest (the six seasonal frequencies) but seek to test for the presence of a statistically significant peak. Viewing the true spectral density f as a smooth function (this can be quantified through sufficiently rapid decay of the autocovariance function), a peak is a frequency λ_0 such that

$$\dot{f}(\lambda_0) = 0 \qquad \ddot{f}(\lambda_0) < 0, \tag{1}$$

where $\dot{f}$ and $\ddot{f}$ denote first and second derivatives. Clearly, the second derivative must be negative *with some significance* in order for the concept to be meaningful. Upon further reflection, it seems that examining the infinitesimal geometry of f at the single point λ_0 is naïve, since any small spike in the side of a monotonic function may satisfy (1) while being dissociated from more intuitive notions of what constitutes a peak. Therefore, we must have negative convexity in a reasonably large neighborhood of λ_0. This thinking leads to the diagnostic of this paper: aggregate measures of the slope and convexity of the spectral density, appropriately normalized. Mathematically, these will take the form of kernel-smoothed periodogram estimates, but without the bandwidth being dependent on sample size.

In Section 2 we develop the mathematical ideas of this method, illustrated through two carefully chosen choices of kernels. Section 3 shows how statistical estimators can be formulated, and how statistical peak hypotheses can be tested. The methodology is tested in Section 4; simulations provide a finite sample description of the size and power of our test. We further demonstrate the utility of our methods through an extensive case study involving 130 time series from the U.S. Census Bureau and the Organization for Economic Co-operation and Development (OECD). We use some concepts from the multiple testing literature (Hochberg 1988) to combine tests based on the individual frequencies together into one diagnostic. Section 5 concludes, and all theorems and proofs are left to the Appendix.

2. Measuring the local geometry of the spectrum

We begin by discussing the geometry of the spectral density (or spectrum) of the time series under consideration. The starting point is to consider measures of slope and convexity of the spectrum that are completely deterministic (*cf.* the approach of Newton and Pagano 1983); later in Section 3 we will consider statistical measures. In Section 2.1 we introduce the concepts of slope and convexity measures. The relevancy of these measures to peak identification is discussed in 2.2, while 2.3 provides two simple kernels as explicit examples.

Suppose that, after suitable transformations and differencing if necessary, $X_1, X_2, ..., X_n$ is a sample from a zero-mean stationary stochastic process. We will use the notation $X = (X_1, X_2, ..., X_n)'$. The spectral density $f(\lambda)$ is well-defined so long as the autocovariance function $\gamma_f(h)$ is absolutely summable, and is given by

$$f(\lambda) = \sum_{h=-\infty}^{\infty} \gamma_f(h) e^{-ih\lambda} \tag{2}$$

with $i = \sqrt{-1}$ and $\lambda \in [-\pi, \pi]$. It follows that the inverse Fourier transform yields

$$\gamma_f(h) = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(\lambda) e^{ih\lambda} d\lambda, \tag{3}$$

a relation that we will use repeatedly in the sequel. Of course this relationship between γ_g and g holds for any integrable function g, not just a spectral density. Furthermore, denoting the Toeplitz matrix associated with γ_g by $\Sigma(g)$, it follows that

$$\Sigma_{jk}(g) = \gamma_g(j - k) = \frac{1}{2\pi} \int_{-\pi}^{\pi} g(\lambda) e^{i(j-k)\lambda} d\lambda.$$

Now from (2), f is d times continuously differentiable if $\sum_{h=-\infty}^{\infty} |h|^d |\gamma_f(h)| < \infty$. We assume that f is twice continuously differentiable for the remainder of the paper (this space of functions will be abbreviated as C^2).

2.1 Measures of slope and convexity

The local geometry of a C^2 function can be described through its first and second derivatives; an aggregate measure of these derivatives is obtained by integrating over a band of frequencies. Alternatively, one may integrate against a function A that has compact support over this band, so long as A provides a suitable proxy for integration over the band. We denote this integral via the general device of a functional θ_A, where

$$\theta_A(f) = \frac{1}{2\pi}\int_{-\pi}^{\pi} A(\lambda) f(\lambda)\, d\lambda. \qquad (4)$$

The function A will be referred to as the "kernel" of this functional. Hence the aggregate slope and convexity measures are defined by $\theta_A(\dot{f})$ and $\theta_A(\ddot{f})$, where each dot denotes a single derivative. These functionals give a summary measure of slope and convexity of f over some band $[\mu - \beta/2,\ \mu + \beta/2] \subset [0,\ \pi]$, and the corresponding kernels will therefore be denoted $A_{\beta,\,\mu}$. We consider kernels with the following properties: (i) $A_{\beta,\,\mu}$ is a C^2 function on $[-\pi,\ \pi]$; (ii) $A_{\beta,\,\mu}$ is zero outside the band $[\mu - \beta/2,\ \mu + \beta/2]$; (iii) $A_{\beta,\,\mu}$ is symmetric about μ on this band; (iv) $\dot{A}_{\beta,\,\mu}(\mu \pm \beta/2) = 0$. Condition (iii) ensures that the location of the peak in f is not shifted by employing the kernel $A_{\beta,\,\mu}$. Note that we do not impose that the total integral of $A_{\beta,\,\mu}$ be unity, because later we will employ a normalization that will automatically account for the total mass of the kernel. Now by (iv) and integration by parts in (4), we obtain

$$\theta_{A_{\beta,\,\mu}}(\dot{f}) = -\theta_{\dot{A}_{\beta,\,\mu}}(f)$$
$$\theta_{A_{\beta,\,\mu}}(\ddot{f}) = \theta_{\ddot{A}_{\beta,\,\mu}}(f). \qquad (5)$$

These formulas are convenient, because they only require a knowledge of f, not its derivatives (assuming that we can compute $\dot{A}_{\beta,\,\mu}$ and $\ddot{A}_{\beta,\,\mu}$). Following the extensive literature on kernels in nonparametric regression and spectral density estimation, we can start with an even kernel A defined on the band $[-\pi,\ \pi]$ that satisfies (i) and $\dot{A}(\pm\pi) = 0$. Then $A_{\beta,\,\mu}$ is defined via

$$A_{\beta,\,\mu}(\lambda) = \frac{2\pi}{\beta} A\left(\frac{2\pi}{\beta}(\lambda - \mu)\right),$$

and is zero outside the band of frequencies $[\mu - \beta/2,\ \mu + \beta/2]$. Clearly we must impose $\beta \le 2\mu$ and $\beta \le 2(\pi - \mu)$, so that $[\mu - \beta/2,\ \mu + \beta/2] \subset [0,\ \pi]$; and the kernel $A_{\beta,\,\mu}$ satisfies conditions (i)-(iv). Note that we cannot construct these types of measures for μ equal to 0 or π. Using a change of variables, we see that

$$\gamma_{A_{\beta,\,\mu}}(h) = \exp\{ih\mu\}\gamma_A(h\beta/2\pi), \qquad (6)$$

so that the effect of β and μ are in some sense separable. Note that we typically evaluate γ_A at non-integer values, so these relations are obtained by extending (3) to non-integer arguments. The fact that $\gamma_{A_{\beta,\,\mu}}$ is complex-valued may seem troubling, but actually only its real portion will enter into our statistical estimators. Of course, we are ultimately interested in $\dot{A}_{\beta,\,\mu}$ and $\ddot{A}_{\beta,\,\mu}$, which are given by

$$\dot{A}_{\beta,\,\mu}(\lambda) = \frac{4\pi^2}{\beta^2}\dot{A}\left(\frac{2\pi}{\beta}(\lambda - \mu)\right),$$

and

$$\ddot{A}_{\beta,\,\mu}(\lambda) = \frac{8\pi^3}{\beta^3}\ddot{A}\left(\frac{2\pi}{\beta}(\lambda - \mu)\right).$$

Later, we will consider the squares of such kernels, and their corresponding inverse Fourier transforms. Hence assuming that $[\mu - \beta/2,\ \mu + \beta/2] \subset [0,\ \pi]$, the squares are given by

$$\dot{A}^2_{\beta,\,\mu}(\lambda) = \frac{16\pi^4}{\beta^4}\dot{A}^2\left(\frac{2\pi}{\beta}(\lambda - \mu)\right)$$

and

$$\ddot{A}^2_{\beta,\,\mu}(\lambda) = \frac{64\pi^6}{\beta^6}\ddot{A}^2\left(\frac{2\pi}{\beta}(\lambda - \mu)\right).$$

Finally, we notice from (4) that we can rewrite $\theta_A(f)$ as

$$\theta_A(f) = \sum_{h=-\infty}^{\infty}\gamma_A(h)\gamma_f(h). \qquad (7)$$

Thus it may be advantageous to determine the $\gamma_A(h)$ sequence from the kernel A. Taking the inverse Fourier Transform of the above slope and convexity kernels, we can construct $\Sigma(\dot{A}_{\beta,\,\mu})$, $\Sigma(\ddot{A}_{\beta,\,\mu})$, $\Sigma(\dot{A}^2_{\beta,\,\mu})$ and $\Sigma(\ddot{A}^2_{\beta,\,\mu})$, as follows:

$$\gamma_{\dot{A}_{\beta,\,\mu}}(h) = \frac{2\pi}{\beta}\exp\{ih\mu\}\gamma_{\dot{A}}(h\beta/2\pi),$$
$$\gamma_{\ddot{A}_{\beta,\,\mu}}(h) = \frac{4\pi^2}{\beta^2}\exp\{ih\mu\}\gamma_{\ddot{A}}(h\beta/2\pi),$$
$$\gamma_{\dot{A}^2_{\beta,\,\mu}}(h) = \frac{8\pi^3}{\beta^3}\exp\{ih\mu\}\gamma_{\dot{A}^2}(h\beta/2\pi),$$
$$\gamma_{\ddot{A}^2_{\beta,\,\mu}}(h) = \frac{32\pi^5}{\beta^5}\exp\{ih\mu\}\gamma_{\ddot{A}^2}(h\beta/2\pi). \qquad (8)$$

Thus, if we have the time-domain information $\gamma_f(h)$ for the process $\{X_t\}$, we can compute slope and convexity measures using (7) given the inverse Fourier transform sequence of the appropriate kernels. Since $\gamma_f(h)$ is a symmetric sequence, we only need to consider the real portion of $\gamma_A(h)$ if it happens to be complex.

2.2 Troughs and peaks

The aggregate measures of spectral slope and convexity previously described provide the building blocks for determinants of the local spectral geometry. Our overall interest is in determining whether a given interval of the spectrum is a peak or a trough (or is monotonic). In the second order geometry of calculus, a local maximum has the defining property that the first derivative is zero and the second derivative is strictly negative. Obviously this requires looking, sequentially, at a slope measure and a convexity measure, defined over the same band of frequencies.

In order to test for the presence of a peak, the sequential approach can be seen as making inferential statements about $\theta_{A_{\beta,\,\mu}}(\dot{f})$ and $\theta_{A_{\beta,\,\mu}}(\ddot{f})$. Note, in making these inferential statements we choose μ ahead of time, according to where in the spectrum we wish to detect a peak (or trough); β is chosen according to which frequencies we wish to exclude, a decision based on how local we wish our viewpoint of the spectrum to be. Then we say that μ is a β-aggregate peak (with respect to A) of the spectrum if

$$\theta_A(\dot{f}) = 0 \qquad \text{and} \qquad \theta_A(\ddot{f}) < 0.$$

The sequential aspect comes from the idea that we generally determine whether $\theta_A(\dot{f}) = 0$ first, and then determine the convexity; this will become more apparent when we consider statistical testing in Section 3.2. In a similar manner we define a β-aggregate trough when $\theta_A(\ddot{f}) > 0$. In terms of hypothesis testing for a peak, we have

$$H_0^{(1)}\colon \theta_A(\dot{f}) = 0 \quad vs. \quad H_a^{(1)}\colon \theta_A(\dot{f}) \neq 0$$
$$H_0^{(2)}\colon \theta_A(\ddot{f}) = 0 \quad vs. \quad H_a^{(2)}\colon \theta_A(\ddot{f}) < 0.$$

The unusual aspect of this hypothesis test is that we wish to fail to reject $H_0^{(1)}$ first, and then conditional on this test we want to reject $H_0^{(2)}$ in favor of the alternative $H_a^{(2)}$.

2.3 Examples of kernels

There are a host of kernels that satisfy conditions (i) through (iv); we can simply borrow from the literature on nonparametric density estimation. For example, the Parzen and Tukey-Hanning (TH) kernels (discussed in Priestley 1981) are suitable, whereas the Bartlett and Daniell kernels are inappropriate, since (iv) does not hold. In general, one only needs to use (8) to determine the inverse Fourier transforms. In this section, we consider two examples: Quartic and TH. The advantage of these kernels is that they have easily computable first and second derivatives, and their inverse Fourier transforms can be obtained explicitly.

Example 1: *Quartic Kernel*

We begin by considering a polynomial kernel of degree four, namely a quartic. Imposing all of the constraints (i) through (iv) yields the following form:

$$A(\lambda) = \frac{15}{8\pi^4}(\lambda^4 - 2\pi^2\lambda^2 + \pi^4),$$

$$\dot{A}(\lambda) = \frac{15}{8\pi^4}(4\pi^3 - 4\pi^2\lambda), \text{ and}$$

$$\ddot{A}(\lambda) = \frac{15}{8\pi^4}(12\pi^2 - 4\pi^2).$$

Taking the inverse Fourier transform of the slope and convexity kernels (and their squares) yields

$$\gamma_{\dot{A}}(h) = \frac{15i}{\pi^5}\left(\frac{\pi^2\sin\pi h}{h^2} + \frac{3\pi\cos\pi h}{h^3} - \frac{3\sin\pi h}{h^4}\right),$$

$$\gamma_{\ddot{A}}(h) = \frac{15}{\pi^5}\left(\frac{\pi^2\sin\pi h}{h} + \frac{3\pi\cos\pi h}{h^2} - \frac{3\sin\pi h}{h^3}\right),$$

$$\gamma_{\dot{A}^2}(h) = \frac{225}{\pi^9}\left(-\frac{2\pi^4\sin\pi h}{h^3} - \frac{18\pi^3\cos\pi h}{h^4}\right.$$
$$\left. + \frac{78\pi^2\sin\pi h}{h^5} + \frac{180\pi\cos\pi h}{h^6} - \frac{18\sin\pi h}{h^7}\right),$$

$$\gamma_{\ddot{A}^2}(h) = \frac{225}{4\pi^{11}}\left(\frac{\pi^4\sin\pi h}{h} + \frac{6\pi^3\cos\pi h}{h^2}\right.$$
$$\left. - \frac{24\pi^2\sin\pi h}{h^3} - \frac{54\pi\cos\pi h}{h^4} + \frac{54\sin\pi h}{h^5}\right),$$

to which we apply (8) and obtain

$$\gamma_{\dot{A}_{\beta,\,\mu}}(h) = \frac{30i}{\beta}\exp\{ih\mu\}\left(\frac{\sin k}{k^2} + \frac{3\cos k}{k^3} - \frac{3\sin k}{k^4}\right),$$

$$\gamma_{\ddot{A}_{\beta,\,\mu}}(h) = \frac{30}{\beta^2\pi}\exp\{ih\mu\}\left(\frac{\sin k}{k} + \frac{3\cos k}{k^2} - \frac{3\sin k}{k^3}\right),$$

$$\gamma_{\dot{A}^2_{\beta,\,\mu}}(h) = \frac{1{,}800\pi}{\beta^3}\exp\{ih\mu\}\left(\frac{2\sin k}{k^3} + \frac{18\cos k}{k^4}\right.$$
$$\left. - \frac{78\sin k}{k^5} - \frac{180\cos k}{k^6} + \frac{180\sin k}{k^7}\right), \text{ and}$$

$$\gamma_{\ddot{A}^2_{\beta,\,\mu}}(h) = \frac{1{,}800}{\beta^5\pi}\exp\{ih\mu\}\left(\frac{\sin k}{k} + \frac{6\cos k}{k^2}\right.$$
$$\left. - \frac{24\sin k}{k^3} - \frac{54\cos k}{k^4} + \frac{54\sin k}{k^5}\right),$$

where $k = h\beta/2$. Note that $\gamma_{\dot{A}^2_{\beta,\,\mu}}(0) = 240\pi/(7\beta^3)$ and $\gamma_{\ddot{A}^2_{\beta,\,\mu}}(0) = 360/(\beta^5\pi)$ follow by application of L'Hopital's rule. These formulas allow us to construct the appropriate Toeplitz matrices for the diagnostic (as discussed in Section 3.1 below, it suffices to consider the real part of these sequences).

Example 2: *TH Kernel*

A similar shape to the quartic can be obtained through the use of a cosine function. The following choice satisfies all the stated conditions on a kernel:

$$A(\lambda) = \frac{1}{2\pi}(1 + \cos\lambda),$$

$$\dot{A}(\lambda) = \frac{1}{2\pi}(-\sin\lambda), \text{ and}$$

$$\ddot{A}(\lambda) = \frac{1}{2\pi}(-\cos\lambda).$$

This function is identical to the Tukey-Hanning lag window, though here we apply it as a spectral window (see Priestley 1981). Hereafter it will be referred to as the TH kernel. Taking the inverse Fourier transform of the slope and convexity kernels (and their squares) yields

$$\gamma_{\dot{A}}(h)=\frac{i}{4\pi^2}\left(\frac{\sin\pi(h+1)}{h+1}-\frac{\sin\pi(h-1)}{h-1}\right),$$

$$\gamma_{\ddot{A}}(h)=-\frac{1}{4\pi^2}\left(\frac{\sin\pi(h+1)}{h+1}+\frac{\sin\pi(h-1)}{h-1}\right),$$

$$\gamma_{\dot{A}^2}(h)=\frac{1}{16\pi^3}\left(\frac{2\sin\pi h}{h}-\frac{\sin\pi(h+2)}{h+2}-\frac{\sin\pi(h-2)}{h-2}\right), \text{ and}$$

$$\gamma_{\ddot{A}^2}(h)=\frac{1}{16\pi^3}\left(\frac{2\sin\pi h}{h}+\frac{\sin\pi(h+2)}{h+2}+\frac{\sin\pi(h-2)}{h-2}\right).$$

Now applying (8) yields

$$\gamma_{\dot{A}_{\beta,\,\mu}}(h) = \frac{i}{2\beta}\exp\{ih\mu\}\left(\frac{\sin(k+\pi)}{k+\pi}-\frac{\sin(k-\pi)}{k-\pi}\right),$$

$$\gamma_{\ddot{A}_{\beta,\,\mu}}(h) = -\frac{\pi}{\beta^2}\exp\{ih\mu\}\left(\frac{\sin(k+\pi)}{k+\pi}+\frac{\sin(k-\pi)}{k-\pi}\right),$$

$$\gamma_{\dot{A}^2_{\beta,\,\mu}}(h) = \frac{\pi}{2\beta^3}\exp\{ih\mu\}\left(\frac{2\sin k}{k}-\frac{\sin(k+2\pi)}{k+2\pi}-\frac{\sin(k-2\pi)}{k-2\pi}\right), \text{ and}$$

$$\gamma_{\ddot{A}^2_{\beta,\,\mu}}(h) = \frac{2\pi^3}{\beta^5}\exp\{ih\mu\}\left(\frac{2\sin k}{k}-\frac{\sin(k+2\pi)}{k+2\pi}+\frac{\sin(k-2\pi)}{k-2\pi}\right),$$

where $k = h\beta/2$. Note that $\gamma_{\dot{A}^2_{\beta,\,\mu}}(0) = \pi/\beta^3$ and $\gamma_{\ddot{A}^2_{\beta,\,\mu}}(0) = 4\pi^3/\beta^5$ follow by application of L'Hopital's rule (using the convention that $\sin(0)/0 = 1$). These formulas allow us to construct the appropriate Toeplitz matrices for the diagnostic (again, as discussed in Section 3.1 below, it suffices to consider the real part of these sequences).

3. Statistical methodology

Of course we do not typically have knowledge of the spectrum f, and thus it is usually necessary to form estimates from the data. In this section we describe statistical estimates of slope and convexity measures that are consistent and simple to compute in the time-domain. Under some mild additional assumptions, these estimates are asymptotically normal, which will be advantageous when performing hypothesis tests. In Section 3.1 the statistical estimates are defined, and their asymptotic properties are discussed. Section 3.2 discusses the application to peak testing, and 3.3 gives an extension to joint peak testing, which facilitates an important application in seasonal adjustment. Section 3.4 discusses extensions to trend nonstationary data.

3.1 Estimators of slope and convexity

We begin by noting that the quadratic form (for any integrable function g)

$$\frac{1}{n}X'\Sigma(g)X = \frac{1}{2\pi}\int_{-\pi}^{\pi} g(\lambda)\, I(\lambda)\, d\lambda,$$

where I denotes the periodogram. Although the periodogram is typically defined at the Fourier frequencies $(2\pi j/n;\ j = 1, \ldots, \lfloor n/2 \rfloor)$ we define it at a continuous band of frequencies as follows

$$I(\lambda) = \frac{1}{n}\left|\sum_{t=1}^{n} X_t\, e^{-it\lambda}\right|^2$$

$$= \sum_{h=1-n}^{n-1} R(h)\, e^{-it\lambda},\ \lambda \in [-\pi, \pi] \qquad (9)$$

with $R(h)$ equal to the sample (uncentered) autocovariance function. This gives an elegant way of passing from the time-domain to the frequency-domain, and is well-known in the time series literature (see Taniguchi and Kakizawa 2000). Moreover, such integrals of the periodogram are generally consistent, *i.e.*, $\theta_g(I) \xrightarrow{a.s.} \theta_g(f)$ as $n \to \infty$, under mild conditions discussed below (note that the inconsistency of the periodogram is resolved by the spectral aggregation against the function g, as shown in the Appendix). Therefore, we obtain statistical estimates of the slope and convexity measures f by using a "plug in" approach, *i.e.*, we simply replace f by I in $\theta_{A_{\beta,\,\mu}}$. In particular,

$$\hat{\theta}_{A_{\beta,\,\mu}}(\dot{f}) = -\theta_{\dot{A}_{\beta,\,\mu}}(I) = -\frac{1}{n}X'\Sigma(\dot{A}_{\beta,\,\mu})X, \text{ and}$$

$$\hat{\theta}_{A_{\beta,\,\mu}}(\ddot{f}) = \theta_{\ddot{A}_{\beta,\,\mu}}(I) = \frac{1}{n}X'\Sigma(\ddot{A}_{\beta,\,\mu})X. \qquad (10)$$

This definition makes use of (5), which accounts for the minus sign in the slope measure. In order to compute the estimate, we utilize the time-domain representation (expressed as a quadratic form). This representation is convenient in that we only need determine a suitable length of the sequences $\gamma_{\dot{A}_{\beta,\,\mu}}(h)$ and $\gamma_{\ddot{A}_{\beta,\,\mu}}(h)$, form the Toeplitz matrices $\Sigma(\dot{A}_{\beta,\,\mu})$ and $\Sigma(\ddot{A}_{\beta,\,\mu})$, and then compute the quadratic forms. Note that the inverse Fourier transforms of $\dot{A}_{\beta,\,\mu}$ and $\ddot{A}_{\beta,\,\mu}$ need only be determined once (see Section 2.3 for some explicit examples) and can be done ahead of time, and then applied repeatedly to many different time series.

In order to compute the time domain representation of the slope and convexity measures in (10), we utilize (8), *e.g.*, see the formulas in Examples 1 and 2. Of course, this will in general result in $\Sigma(\dot{A}_{\beta,\mu})$ and $\Sigma(\ddot{A}_{\beta,\mu})$ being complex. However, even if $\Sigma(g)$ (where g can be $A_{\beta,\mu}$, $\dot{A}_{\beta,\mu}$, or $\ddot{A}_{\beta,\mu}$) is a complex Toeplitz matrix, $X'\Sigma(g)X$ will always be real. From (8), it is easy to see that $\Sigma(g) = M + iN$ where M is real, symmetric, and Toeplitz, and N is real, skew-symmetric, and Toeplitz. Hence $X'NX = 0$ for any vector X, so that $X'\Sigma(g)X = X'MX$. Therefore, for the purposes of computing the statistical slope and convexity measures, we may take the real part of $\gamma_A(h)$ in (8).

Not only are these statistical estimates consistent, they are also asymptotically normal under some additional conditions (discussed in the Appendix). However, in order to construct a suitable normalization it will be necessary to estimate their variation. The asymptotic variance of $\theta_g(I)$ is $\theta_{g^2}(f^2)$ (if g is supported on $[0, \pi]$), which can be consistently estimated via $\theta_{g^2}(I^2)/2$. (The factor of 2 is required, since the integral of I^2 tends to the corresponding integral of $2f^2$ – see Chiu (1988)). This can be given a time-domain representation as follows. Let $R = \{R(1-n), ..., R(0), ..., R(n-1)\}'$ be a $2n-1$ vector of sample autocovariances, and let $\Sigma(g^2)$ be $2n-1$ dimensional in the following formula: $R'\Sigma(g^2)R/2 = \theta_{g^2}(I^2)/2$. This relationship can be easily verified using (9). Thus we will normalize $\theta_g(I)$ by the square root of $\theta_{g^2}(I^2)/2$. Hence our normalized statistical measures of slope and convexity are given by

$$-\psi_{\dot{A}_{\beta,\mu}}(I) = -\frac{\theta_{\dot{A}_{\beta,\mu}}(I)}{\sqrt{\theta_{\dot{A}^2_{\beta,\mu}}(I^2)/2}} = -\frac{1}{n}\frac{X'\Sigma(\dot{A}_{\beta,\mu})X}{\sqrt{R'\Sigma(\dot{A}^2_{\beta,\mu})R/2}}$$

and

$$\psi_{\ddot{A}_{\beta,\mu}}(I) = -\frac{\theta_{\ddot{A}_{\beta,\mu}}(I)}{\sqrt{\theta_{\ddot{A}^2_{\beta,\mu}}(I^2)/2}} = \frac{1}{n}\frac{X'\Sigma(\ddot{A}_{\beta,\mu})X}{\sqrt{R'\Sigma(\ddot{A}^2_{\beta,\mu})R/2}},$$

where the dimensions of the Σ matrices are either n or $2n-1$ as appropriate. The asymptotic properties of $\psi_{\dot{A}_{\beta,\mu}}(I)$ and $\psi_{\ddot{A}_{\beta,\mu}}(I)$ are discussed in the Appendix. In summary, both $-\sqrt{n}\psi_{\dot{A}_{\beta,\mu}}(I)$ and $\sqrt{n}\psi_{\ddot{A}_{\beta,\mu}}(I)$ are marginally asymptotically $N(0, 1)$ under $\mathrm{H}_0^{(1)}$ and $\mathrm{H}_0^{(2)}$ respectively and the assumptions discussed in the Appendix. Simulations indicate that the variance normalization is slow to converge, and its correlation with numerator causes a degree of non-normality in smaller samples. Based on the histogram of the distribution simulated under a Gaussian white noise Null hypothesis with $n = 360$ and 10,000 replications (Figure 1) there is close agreement to the normal distribution, except at the extremes in the tails. Section 4 explores this behavior further through simulation studies.

3.2 Applications to single peak testing

We now consider the application to peak testing. Recall that we have an initial Null Hypothesis $H_0^{(1)}$ that we must fail to reject in order to proceed. This can be interpreted as saying there is insufficient evidence to conclude that the first derivative (slope) of the spectral density is significantly different from zero. Now we know that $-\sqrt{n}\psi_{\dot{A}_{\beta,\mu}}(I)$ is asymptotically $N(0, 1)$ under $\mathrm{H}_0^{(1)}$ and the assumptions discussed in the Appendix. If we further suppose that a sufficiently small value x is obtained for the test statistic, we will not be able to reject $H_0^{(1)}$ with any confidence. In that case, we can consider the hypothesis $H_0^{(2)}$, which we seek to reject; this is tested via $\sqrt{n}\psi_{\ddot{A}_{\beta,\mu}}(I)$. Although $-\sqrt{n}\psi_{\dot{A}_{\beta,\mu}}(I)$ and $\sqrt{n}\psi_{\ddot{A}_{\beta,\mu}}(I)$ are asymptotically correlated (see Theorem 1 of the Appendix) we will consider the slope and convexity tests as if they were done separately (this correlation can be estimated, and used to determine the distribution of the convexity diagnostic conditional on the slope diagnostic; however, the interpretation of p-values becomes muddled. For simplicity, we treat the tests separately, one at a time, and do not explicitly account for the correlation). Our testing procedure is then conducted as follows:

1. Perform the 2-sided test of $H_0^{(1)}$ using $-\sqrt{n}\psi_{\dot{A}_{\beta,\mu}}(I)$.
2. Let p be the p-value associated with the first test statistic's value $x = -\sqrt{n}\psi_{\dot{A}_{\beta,\mu}}(I)$, with x and p related by $p = 2\Phi(-|x|)$.
3. If $p > 0.05$ (or some other pre-determined tolerance level) proceed; else conclude that there is no peak present.
4. Perform the lower 1-sided test of $H_0^{(2)}$ using $\sqrt{n}\psi_{\ddot{A}_{\beta,\mu}}(I)$.
5. Reject $H_0^{(2)}$ and conclude that there is a peak if $\sqrt{n}\psi_{\ddot{A}_{\beta,\mu}}(I) < \Phi^{-1}(\alpha)$, where α is the level of the convexity test.

3.3 Joint peak testing: Application to seasonal adjustment

We now consider the situation where we wish to test for several spectral peaks simultaneously. Clearly we could design a kernel with several nodes, one at each peak, but this would merely be the sum of several individual spectral peak diagnostics. It would have the disadvantage that a significant spectral peak in one place could cancel a significant spectral trough elsewhere. Therefore, we would prefer a test that examines a set of spectral diagnostics within a multiple testing paradigm.

For example, consider the context of testing for spectral peaks in seasonally adjusted data. There are six seasonal peaks of interest, but we must restrict attention to five due to aliasing problems (the peak at frequency π cannot be identified). If one or more of the spectral peaks is significant, we must reject our seasonal adjustment procedure (since it has failed to remove all of the peaks); therefore, we are in a multiple testing situation, and will utilize a method that controls the familywise error rate (FWER) proposed by Hochberg (1988) and described in Benjamini and Hochberg (page 294, 1995). Restricting attention to the issue of convexity, we have Null Hypotheses $H_0^{(2)}$ for each of the five seasonal frequencies. In our setting, the procedure of Hochberg (1988) is to compute p-values for the convexity test at each of the five seasonal frequencies, and order them as $p_{(1)} \leq p_{(2)} \leq p_{(3)} \leq p_{(4)} \leq p_{(5)}$, with corresponding Null Hypotheses denoted by $H_{(i)}$. For a specified FWER of level α (*e.g.*, $\alpha = 0.05$), let k be the largest i for which $p_{(i)} \leq i/(6 - i)\alpha$; then reject all $H_{(i)}$ for $i \leq k$.

When using such a procedure, we should make Type I errors – *i.e.*, identifying at least one seasonal frequency as having negative convexity when none is present – roughly α proportion of the time (if we were to restrict attention to $H_0^{(2)}$, the convexity hypothesis). The advantage of the Hochberg familywise error rate approach (H-FWER) is that it dramatically improves the statistical power compared to other methods. The validity of this method requires independence of the test statistics under consideration, and so for this reason we take five kernels $A_1, ..., A_5$ – centered at the seasonal frequencies $\pi/6, ..., 5\pi/6$ respectively – that have disjoint support. Then Theorem 1 can be generalized to obtain asymptotic independence of the five convexity test statistics (see the discussion after Theorem 1 in the Appendix). Of course, we also conduct five separate tests of the slope at each seasonal frequency, where we must fail to reject in each case in order to proceed.

As a final remark, we note that in practice a seasonal adjustment is rarely rejected on the basis of significant spectral mass at the fifth seasonal frequency of $5\pi/6$ (Findley 2006). This is partly due to the difficulty in assigning an interpretation to this frequency. Therefore, practitioners may be more interested in a "four-peak test" that focuses on the first four seasonal frequencies; one obtains this test by an obvious modification of the H-FWER procedure described above.

3.4 Extending to nonstationary data

The methodology given above assumes that the data are a sample from a stationary process. However, in the context of seasonal adjustment, it is usually the case that the seasonally adjusted data are once or twice integrated. In this case one would difference the seasonally adjusted data once or twice, and then apply the diagnostics. Now application of the differencing operators $1 - B$ and $(1 - B)^2$ are essentially high-pass filters, which can be expected to attenuate residual spectral peaks close to frequency zero (in particular, the first seasonal frequency at $\pi/6$). Thus it may be desirable to apply the diagnostic to the pseudo-spectrum instead; this can be done if the support of the kernel is bounded away from the poles in the spectral density.

Suppose that the observed data are now $Y_{1-d}, ..., Y_n$ for d the order of trend differencing (so usually $d = 1$ or 2). When the observed data are differenced, we obtain the sample X, which is strictly stationary. The pseudo-spectral density of the $\{Y_t\}$ process is $g(\lambda) = f(\lambda)\,|1 - e^{-i\lambda}|^{-2d}$, where f is the spectrum of $\{X_t\}$. This pseudo-spectrum could be estimated via $\hat{g}(\lambda) = I(\lambda)\,|1 - e^{-i\lambda}|^{-2d}$, where I is the periodogram of X as before; this is the re-coloring approach of Nerlove (1964). Then $\theta_A(g)$ is well-defined so long as $A(\lambda)\,|1 - e^{-i\lambda}|^{-2d}$ is an integrable function; essentially we must ensure that frequency zero is excluded from the support of the kernel A. Since A is centered around seasonal frequencies in practice, we can easily contrive this condition. The corresponding estimator is then

$$\hat{\theta}_A(\ddot{g}) = \theta_{\ddot{A}b}(I),$$

where $b(\lambda) = |1 - e^{-i\lambda}|^{-2d}$. The estimator is well-defined if $\ddot{A}b$ is integrable; moreover the asymptotic properties discussed in the Appendix for the stationary case extend to this case as well, so long as $\ddot{A}b$ is bounded.

This extension may be more appealing to some researchers. However, the cost is that the inverse Fourier transform of $\ddot{A}b$ must be determined, which requires some additional mathematical work. In the simulation studies and data illustrations in Section 4 we trend difference the seasonally adjusted data, but do not implement the correction factor b in the kernel.

4. Empirical studies

Having developed the theoretical aspects of the spectral diagnostic, we now turn to its performance in practice. We first present some results obtained from simulation, which provide insight into the size and power properties of the test statistic in finite samples. Then we investigate the size and power empirically, by applying the spectral diagnostics to a suite of 130 time series (65 U.S. Census Bureau series and 65 OECD series); we consider both the original and the seasonally adjusted series, and make comparisons to the Visual Significance, M7 and M8 quality control diagnostics of X-12-ARIMA (U.S. Census Bureau 2002). Additional empirical studies can be found in Evans, Holan and McElroy (2006).

4.1 Simulation study

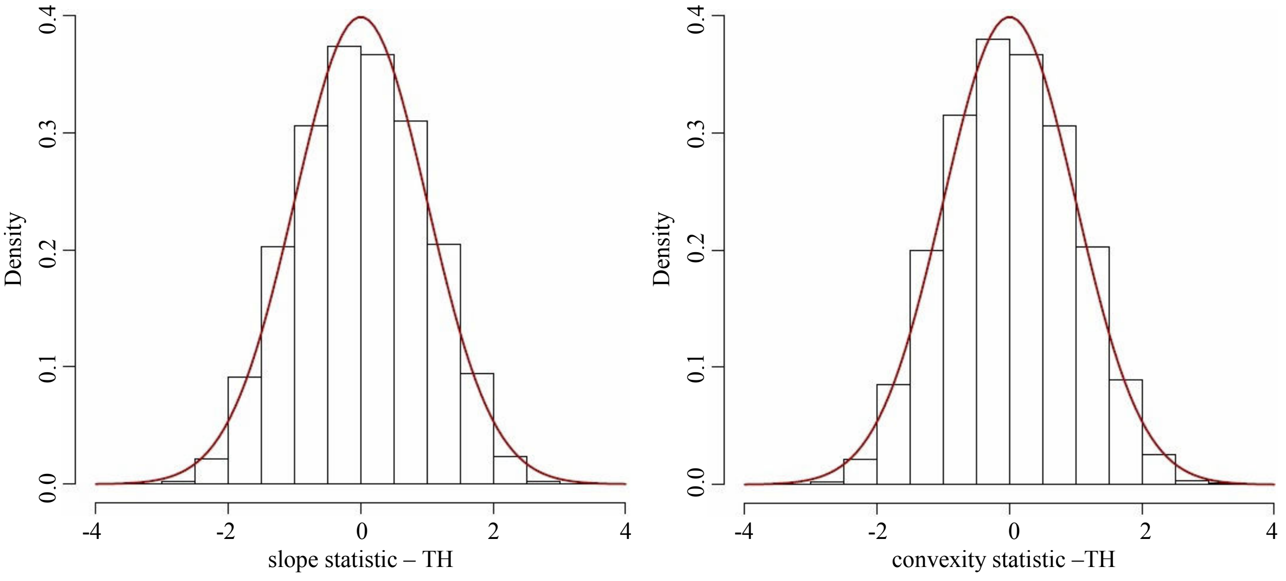
To evaluate the performance of our diagnostics we conducted several simulations. The first set of simulations examines size (level) for the single peak diagnostic. For this simulation we considered the slope and convexity diagnostics separately. Although in practice, when considering the slope diagnostic, we wish to fail to reject the Null hypothesis $\mathrm{H}_0^{(1)}$, here we are interested in empirically investigating the distributional properties, and so we impose the usual definition of size for this study. So we simulated Gaussian white noise which satisfies the assumptions of Theorem 1 as well as satisfying $\psi_{\dot{A}}(I) = \psi_{\ddot{A}}(I) = 0$, so that $\mathrm{H}_0^{(1)}$ and $\mathrm{H}_0^{(2)}$ are true. Of course, there are many processes for which both $\mathrm{H}_0^{(1)}$ and $\mathrm{H}_0^{(2)}$ are true simultaneously – for example, any process with locally flat spectral density; however, due to asymptotic considerations it suffices to consider white noise. For a (large) sample size of $n = 360$, using the TH kernel with $\mu = \pi/6$ and $\beta = \pi/6$ (this corresponds to a kernel centered on the interval $[0, \pi/6]$), 10,000 repetitions yields an empirical distribution of the normalized diagnostics, $\psi_{\dot{A}}(I)$ and $\psi_{\ddot{A}}(I)$, whose histograms are displayed in Figure 1. Henceforth, let δ and α denote the levels associated with the slope and convexity tests respectively. Note that in this case we define level to mean the probability of rejecting $\mathrm{H}_0^{(j)}$ $(j = 1, 2)$ when $\mathrm{H}_0^{(j)}$ is true. Although, in practice, in the case of the slope hypothesis we wish to fail to reject we follow the strict definition of level and assume (for the purposes of this simulation) that the null hypothesis $\mathrm{H}_0^{(1)}$ for the slope holds true. Similarly the null hypothesis for the convexity is $\mathrm{H}_0^{(2)}$. Both the slope and convexity hypotheses are evaluated independently. Table 1 summarizes the results using both kernels from Section 2, for various sample sizes; the indicated δ, α-levels are for the nominal 5 percent level. Additionally, other choices of μ and β (not shown here) yielded similar results. As depicted in this study, in smaller samples, we observed skewness in the distribution which seems to be due to correlation between $\theta_{A}(I)$ and $\theta_{A^2}(I^2)$. Also, note that the size for the convexity test is larger for the quartic kernel than for the TH kernel.

Next we consider the empirical power for our single peak diagnostic. In this setting we evaluate the power based on a joint test of the slope and convexity. Specifically, we wish to fail to reject $\mathrm{H}_0^{(1)}$ while simultaneously rejecting $\mathrm{H}_0^{(2)}$, at $\delta = \alpha = 0.05$, and thus correctly identify spectral peaks. Since our composite Null hypothesis is that there is no peak, the Alternative hypothesis includes processes such as the $AR(2)$ given by

$$(1 - 2\rho\cos\omega B + \rho^2 B^2) X_t = \varepsilon_t \qquad (11)$$

with white noise variance σ^2, associated with some fixed frequency $\omega \in [0, \pi]$. The spectrum associated with the process in (11) is given by $f(\lambda) = \sigma^2 |1 - 2\rho\cos\omega e^{-i\lambda} + \rho^2 e^{-2i\lambda}|^{-2}$, which is maximized at $\lambda_0 = \cos^{-1}(\cos\omega(1 + \rho^2)/2\rho)$. Therefore one can explore the power of a peak-testing procedure by simulating from (11) with various choices of ρ, ω, and σ. Table 2 presents the result of 10,000 simulations, of various sample sizes, from the $AR(2)$ cycle model given in (11) with peak at $\mu = \pi/6$ and bandwidth set at $\beta = \pi/6$. The peak strength is parametrized through ρ, which we vary from 0.85 to 0.95; clearly $\mathrm{H}_0^{(1)}$ and $\mathrm{H}_a^{(2)}$ are both true for this model. In other words, there are spectral peaks, of different heights, at $\lambda = \pi/6$. This AR cycle model was chosen because it provides a convenient parametrization of spectral peak location and shape. Additionally, this choice of β is compatible with the seasonal adjustment setting, as this provides the maximum window width while avoiding overlapping spectral peaks. As expected, the power of our diagnostic increases with sample size and peakedness, ranging from 0.227 (quartic kernel) in small samples having weak spectral peak to ≈ 0.95 (TH kernel) in larger samples having a more pronounced spectral peak (see Table 2). Note that in this procedure the innovation variance is set equal to one, but it is immaterial due to the normalization of the diagnostic. In summary, both the quartic and TH kernels possess decent size and power properties. Generally, the quartic kernel seems to have superior size and power, so it would be preferable for spectra of this form (note that the lower power of the TH kernel is in part due to its being undersized). Additionally, it seems that smaller values of β (results not shown) require a greater sample size; a smaller β corresponds to a more refined "viewing" of the spectral peak, which would require more data to handle the resolution.

Although the individual peak testing scenario provides the foundation for our joint testing framework, as noted, the joint testing framework provides important methodology for applications in federal statistics. The application of importance is the evaluation of effective seasonal adjustment through the exploration of residual seasonality. Thus, it is of particular interest to know how our multiple testing approach performs in simulation. Therefore, in order to investigate the size and power associated with our joint test, we simulated 10,000 repetitions from a Gaussian white noise process and from an $AR(25)$ model obtained as a fit to the Current Employment Series (Employed Males, aged 16 to 19). Our goal in the power study was to construct an $AR(p)$ process (because of its ease in simulation and desirable theoretical properties as a parametric spectral estimator – see Parzen 1983) with (stationary) spectral peaks that are realistic, or close to what might be found in practice. Thus we obtain our $AR(25)$ model – fitted via maximum likelihood using AIC – which has similar seasonal dynamics (local spectral behavior) to the Current Employment Series (CES).



(a)

(b)

Figure 1 Histogram of distribution of $-\sqrt{n}\,\psi_{\dot{A}_{\beta,\mu}}(I)$ (a) and $\sqrt{n}\,\psi_{\ddot{A}_{\beta,\mu}}(I)$ (b) under a Gaussian white noise Null hypothesis using the TH kernel. The sample size is $n = 360$ with 10,000 replications

Table 1
Results of size simulation for the single peak diagnostic. Here $\mu = \beta = \pi/6$ and 10,000 repetitions were used. The slope and convexity diagnostics were investigated separately for both the quartic and TH kernels

	Size for Single Peak $\mu = \beta = \pi/6$											
	Slope						Convexity					
	Quartic Kernel			TH Kernel			Quartic Kernel			TH Kernel		
n	Mean	Stdev	δ -level	Mean	Stdev	δ -level	Mean	Stdev	α -level	Mean	Stdev	α -level
120	0.003	0.903	0.007	-0.011	0.903	0.008	-0.065	0.852	0.032	0.025	0.888	0.018
144	-0.004	0.920	0.014	-0.011	0.927	0.015	-0.077	0.882	0.042	0.006	0.892	0.025
180	-0.003	0.942	0.022	0.002	0.920	0.017	-0.071	0.892	0.043	0.005	0.902	0.028
288	0.003	0.954	0.027	-0.002	0.950	0.025	-0.072	0.921	0.051	-0.006	0.926	0.033
360	0.003	0.962	0.032	-0.009	0.954	0.031	-0.056	0.922	0.051	0.006	0.951	0.040

Table 2
Results of power simulation for the single peak diagnostic. Here $\mu = \beta = \pi/6$ and 10,000 repetitions were used. The alternative hypothesis is given by the $AR(2)$ model defined by (11). The slope and convexity diagnostics were investigated simultaneously for both the quartic and TH kernels using $\delta = \alpha = 0.05$ for both tests (see Section 4.1)

	Power for Single Peak $\mu = \beta = \pi/6$ - $(\delta, \alpha) = (0.05, 0.05)$					
	Quartic Kernel			TH Kernel		
n	$\rho = 0.85$	$\rho = 0.90$	$\rho = 0.95$	$\rho = 0.85$	$\rho = 0.90$	$\rho = 0.95$
120	0.227	0.438	0.758	0.147	0.335	0.670
144	0.287	0.532	0.856	0.208	0.431	0.799
180	0.354	0.643	0.923	0.272	0.567	0.901
288	0.447	0.755	0.949	0.372	0.706	0.950
360	0.601	0.872	0.937	0.537	0.859	0.948

To evaluate size we considered both a test based on convexity alone and a test based on the slope and convexity simultaneously. The tests based on convexity alone (C) were performed at the nominal α-levels of 0.05 and 0.10, using the H-FWER method to control the FWER. The tests based on both the slope and convexity simultaneously (S, C) were performed as follows:

1. Perform multiple tests of convexity, $H_0^{(2)}$, using the H-FWER method to control the FWER at level α (which is either 0.05 or 0.10).
2. For any peaks found significant in Step 1 perform individual slope test, $H_0^{(1)}$, at level δ (which is either 0.10 or 0.25). Note here we wish to fail to reject $H_0^{(1)}$ in order to declare any "peaks" as statistically significant.
3. Declare there is a statistically significant peak if Step 1 finds any seasonal frequency with significant aggregate convexity in the spectrum, and if Step 2 simultaneously fails to find any significant aggregate slope for the corresponding seasonal frequency.

The results of this simulation are summarized in Table 3. One aspect of this procedure that deserves further explanation is Step 2 where δ (the level for the slope test) is taken equal to 0.10 and 0.25. Although the slope testing aspect of the procedure is conducted on an individual peak basis, it seems reasonable to try and be conservative. The issue here is that even if some of the individual slope tests are rejected we may still proceed in other cases. Thus the situation encountered here differs from the classical "no peaks" hypothesis which can be rejected if a single peak is found. Of course, since we are conducting each slope hypothesis test on an individual peak basis, any δ-level greater than 0.05 would be considered more conservative.

While we cannot expect the combined procedure (S, C) to have size approaching the nominal (because using the slope test throws off the Type I error rate), neither is the size highly accurate in the case of using the convexity test alone (C), as can be seen by examining the $\alpha = 0.10$ case with $n = 288$, 360. Here the convexity for the quartic kernel is over-sized, whereas in the single peak case the convexity test has accurate size (Table 1) for these sample sizes. Note that H-FWER only produces an approximately correctly sized procedure; another factor is that the five peak tests are only asymptotically independent. It is for these reasons that the empirical size found in Table 3 differ somewhat from the nominal levels.

To investigate power we considered the same 3 step procedure as outlined above. However, for this simulation we only considered joint slope - convexity testing and examined four pairs of (δ, α) levels; (δ, α) = (0.10, 0.05), (0.25, 0.05), (0.10, 0.10) and (0.25, 0.10). The results of this simulation (Table 4) indicate tremendous power even at sample sizes as small as $n = 120$. This is extremely important as $n = 120$ is representative of the size samples encountered in practice when conducting seasonal adjustment (*i.e.*, 10 years of monthly data). For samples sizes $n = 144$, greater than 90% power was achieved.

Table 3
Results of size simulation for the multiple peak diagnostic. Here 10,000 repetitions were used. The convexity test was investigated separately with the FWER controlled at $\alpha = 0.05$ and $\alpha = 0.10$ using the H-FWER method. Additionally, the slope and convexity diagnostics were investigated simultaneously using the H-FWER method for the convexity controlling the FWER at $\alpha = 0.05$ and $\alpha = 0.10$, while the slope was evaluated at $\delta = 0.25$ and $\delta = 0.10$. Both the quartic and TH kernels were used for both tests (see Section 4.1)

	Size for Multiple Peaks H-FWER											
	C = 0.05		(S, C) = (0.10, 0.05)		(S, C) = (0.25, 0.05)		C = 0.10		(S, C) = (0.10, 0.10)		(S, C) = (0.25, 0.10)	
n	Quartic	TH	Quartic	TH	Quartic	TH	Quartic	TH	Quartic	TH	Quartic	TH
120	0.006	0.002	0.006	0.002	0.005	0.002	0.076	0.047	0.070	0.044	0.076	0.046
144	0.009	0.002	0.011	0.002	0.008	0.003	0.087	0.053	0.090	0.051	0.086	0.050
180	0.019	0.005	0.020	0.006	0.017	0.005	0.107	0.062	0.097	0.059	0.093	0.057
288	0.031	0.009	0.026	0.008	0.025	0.008	0.117	0.069	0.116	0.069	0.112	0.068
360	0.042	0.012	0.045	0.019	0.035	0.015	0.140	0.087	0.133	0.084	0.129	0.087

Table 4
Results of power simulation for the multiple peak diagnostic. Here 10,000 repetitions were used. The slope and convexity diagnostics were investigated simultaneously using the H-FWER method for the convexity controlling the FWER at $\alpha = 0.05$ and $\alpha = 0.10$. For the slope $\delta = 0.25$ and $\delta = 0.10$. Both the quartic and TH kernels were used for both tests (see Section 4.1)

	Power for Multiple Peaks H-FWER							
	(0.10, 0.05)		(0.25, 0.05)		(0.10, 0.10)		(0.25, 0.10)	
n	Quartic	TH	Quartic	TH	Quartic	TH	Quartic	TH
120	0.877	0.897	0.860	0.881	0.997	0.997	0.996	0.998
144	0.943	0.952	0.942	0.950	0.999	1.00	0.999	1.00
180	0.989	0.992	0.989	0.992	1.00	1.00	0.999	1.00
288	1.00	1.00	0.999	0.999	1.00	1.00	0.999	1.00
360	1.00	1.00	0.998	0.999	1.00	1.00	0.998	0.999

4.2 Case studies

We also considered 130 time series, 65 from the U.S. Census Bureau and 65 from OECD. These series consist of 35 U.S. Manufacturing series, 10 U.S. Housing series, 10 U.S. Import/Export series, and 10 U.S. Retail series; there are also 22 German series, 15 Euro-area series, 11 French series, and 17 Great Britain series from OECD, covering the sectors of manufacturing, retail, wholesale, foreign trade, unemployment, and industry. For every series, we computed the seasonal peak tests for both the raw data (logged) and the seasonally adjusted data (logged) – using the x11 specification of X-12-ARIMA – with both the quartic and TH kernels. We employed the H-FWER procedure controlling the FWER at $\alpha = 0.05$, 0.10, and where the threshold for the slope tests was $\delta = 0.25$ and $\delta = 0.10$ at each peak (see Section 4.1). Note that $\delta = 0.25$ and $\delta = 0.10$ produced similar results and thus, for the sake of brevity, only results for $\delta = 0.10$ are presented here. The results for $\delta = 0.25$ are available upon request from the first author. For both the raw and seasonally adjusted data, a single trend difference was used (as is the case for the Visual Significance diagnostic, described below) before applying the seasonal peaks test.

In addition, we present the M7 and M8 statistics as well as the results of the Visual Significance (VS) diagnostic both before and after adjustment. The M7 quality control statistic measures the amount of stable seasonality relative to the moving seasonality in the original series, with values greater than 1 indicating that the seasonality in the series is not identifiable (Lothian and Morry 1978); similarly, the M8 statistic measures the size of the fluctuations in the seasonal component, with a similar interpretation. We also considered the robust nonparametric Kruskal-Wallis test (U.S. Census Bureau 2002) for the presence of seasonality assuming stability. VS is based on an $AR(30)$ spectrum estimate of the raw and seasonally adjusted series, and is described in Soukup and Findley (1999). For Tables 5-10, each cell entry lists which seasonal frequencies were found to have a significant peak, with j corresponding to $\pi j / 6$ for j = 1, 2, 3, 4, 5; an entry of $\emptyset$ indicates that no peaks were detected. For the M7 and M8 diagnostics, only the value is reported since there is no associated p-value (and they are only pertinent to the raw series).

The results of this empirical study can be found in Tables 5-10. All of the Kruskal-Wallis statistics were significant with $p = 0.000$, so these are not reported in the tables. The set of columns corresponding to the "Original Data" heading can be seen as giving empirical power (for each subset of series), assuming that each series is indeed seasonal and has seasonal spectral peaks. That is, the Total " correct " number gives the proportion of times each method correctly identified seasonality, and hence this proportion is a crude proxy for empirical power. We also report the average number of peaks that were identified, which is an empirical measure of the efficacy of the methods (the more peaks correctly identified, the better). The set of columns for "SA Data" gives an empirical size (for each subset of series), assuming that seasonal adjustment has indeed removed the spectral peaks. These are rough considerations, since we do not really know a priori whether the SA Data has been adequately adjusted.

The VS identifies all of the raw series as seasonal and most of the SA series as having no spectral seasonal peaks; the M7 and M8 diagnostics perform similarly, though of course they do not indicate which seasonal peaks are present in the raw data. Our procedure indicates a few cases (when $\alpha = 0.10$ for the convexity tests) where the adjustment may be inadequate, but these are within the scope of the expected proportion of Type I errors. For the raw series, the empirical power (*i.e.*, total proportion correct) for our method ranges between 0.66 and 0.89, with higher power for the $\alpha = 0.10$ level, as expected. In many cases the indicated peaks are the same as VS, but sometimes are quite different. Note that the average number of peaks detected for raw series was typically much higher for our procedure over the VS method, which often had an average around 3.2. When the α level was increased from 0.05 to 0.10, our method naturally increased in the average number of peaks detected; VS cannot be tuned in this way. Conversely, for SA data the average number of peaks detected tended to be less than one for our method (with the exception of the German series).

The results are fairly similar for the quartic and TH kernels. Although the M7, M8, and VS diagnostics have slightly better performance than our spectral peak procedure with $\alpha = 0.10$, it is important to note that our method provides a level of detail that M7 and M8 cannot replicate, while the VS diagnostic does not provide a p-value for any of the peaks (neither do M7 or M8). Overall, we find the results to be very encouraging and informative.

5. Conclusion

This paper presents an innovative approach to the statistical identification of spectral peaks. The convexity diagnostic computes an average of the periodogram weighted by the second derivative of a typical kernel, such as the Tukey-Hanning lag window. Implicitly this type of statistic involves a comparison of an average of the periodogram near a given frequency to its average somewhat further out; this follows from the general shape of $\dot{A}_{\beta,\,\mu}$. The slope diagnostic helps to screen out cases where there is negative convexity but also a large increase/decrease in the spectrum. That the method actually works as intended is borne out by the simulations and analysis results reported in Tables 1-10.

Table 5
Data analyses for 35 Manufacturing Series (U.S. Census Bureau) comparing our multiple peak diagnostic with VS, M7, and M8 diagnostics. Our multiple peak diagnostic uses the H-FWER method to control the FWER at $\alpha = 0.05$ and $\alpha = 0.10$ and examines the slope at $\delta = 0.10$ (see Section 4.2)

	Data Analyzes – Manufacturing Series											
	Original Data							SA Data				
	H-FWER 0.10/0.05		H-FWER 0.10/0.10		VS	M7	M8	H-FWER 0.10/0.05		H-FWER 0.10/0.10		VS
series	quartic	TH	quartic	TH				quartic	TH	quartic	TH	
M_1	2345	1235	2345	1235	12	0.24	0.39	∅	∅	∅	∅	∅
M_2	12345	12345	12345	12345	1235	0.20	0.32	∅	∅	∅	∅	∅
M_3	12345	12345	12345	12345	1235	0.28	0.46	∅	∅	∅	∅	∅
M_4	∅	∅	12345	∅	12	0.28	0.44	∅	∅	∅	∅	∅
M_5	12345	12345	12345	12345	12345	0.27	0.47	∅	∅	12345	12345	∅
M_6	∅	∅	123	123	12	0.28	0.49	∅	∅	∅	∅	∅
M_7	∅	∅	123	123	24	0.50	0.79	∅	∅	∅	∅	∅
M_8	12345	12345	12345	12345	12345	0.18	0.37	∅	∅	∅	∅	∅
M_9	12345	12345	12345	12345	124	0.42	0.73	∅	∅	∅	∅	∅
M_{10}	∅	∅	∅	∅	1	0.38	0.72	∅	∅	∅	∅	∅
M_{11}	∅	∅	12345	1234	123	0.15	0.27	∅	∅	∅	∅	∅
M_{12}	1234	1234	12345	12345	1234	0.30	0.54	∅	∅	∅	∅	∅
M_{14}	∅	∅	1234	1234	1234	0.24	0.39	∅	∅	∅	∅	∅
M_{15}	12345	12345	12345	12345	12345	0.23	0.43	∅	∅	∅	∅	∅
M_{16}	1234	1234	1234	1234	1234	0.23	0.40	∅	∅	∅	∅	∅
M_{17}	∅	∅	1234	12345	12	0.64	0.66	∅	∅	∅	∅	∅
M_{18}	12345	12345	12345	12345	245	0.20	0.37	∅	∅	∅	∅	∅
M_{19}	∅	∅	∅	∅	4	0.86	1.00	∅	∅	∅	∅	∅
M_{20}	∅	∅	∅	12345	4	0.56	0.84	∅	∅	∅	∅	∅
M_{21}	12345	12345	12345	12345	1234	0.37	0.58	∅	∅	∅	∅	∅
M_{22}	12345	12345	12345	12345	1234	0.26	0.45	∅	∅	∅	∅	∅
M_{23}	12345	12345	12345	12345	1234	0.20	0.47	∅	∅	∅	∅	∅
M_{24}	12345	12345	12345	12345	2345	0.26	0.43	∅	∅	∅	∅	∅
M_{25}	12345	12345	12345	12345	12345	0.27	0.42	∅	∅	∅	∅	∅
M_{26}	12345	12345	12345	12345	1235	0.37	0.62	∅	∅	∅	∅	∅
M_{27}	1345	1234	1345	1234	2345	0.25	0.22	∅	∅	∅	∅	∅
M_{28}	∅	∅	∅	∅	24	0.57	0.44	∅	∅	∅	∅	∅
M_{29}	∅	12345	12345	12345	24	0.78	1.13	∅	∅	∅	∅	∅
M_{30}	123	1234	12345	12345	245	0.45	0.65	∅	∅	∅	∅	∅
M_{31}	∅	∅	123	123	4	0.64	0.46	∅	∅	1234	1234	∅
M_{32}	1235	12345	1235	12345	12345	0.21	0.37	∅	∅	∅	∅	∅
M_{33}	12345	12345	12345	1234	1234	0.24	0.38	∅	∅	∅	∅	∅
M_{34}	12345	12345	12345	12345	234	0.46	0.85	∅	∅	∅	∅	∅
M_{35}	12345	12345	12345	12345	2345	0.25	0.66	∅	∅	∅	∅	∅
M_{36}	12345	12345	12345	12345	123	1.32	1.56	∅	∅	∅	∅	∅
Total "correct"	23/35	24/35	31/35	31/35	35/35	34/35	33/35	35/35	35/35	33/35	33/35	35/35
Average Number	3.09	3.29	4.09	4.09	3.23			0	0	0.26	0.26	0

Table 6
Data analyses for 30 U.S Census Bureau Series (10 Housing, 10 Import/Export and 10 Retail Sales) comparing our multiple peak diagnostic with VS, M7, and M8 diagnostics. Our multiple peak diagnostic uses the H-FWER method to control the FWER at $\alpha = 0.05$ and $\alpha = 0.10$ and examines the slope at $\delta = 0.10$ (see Section 4.2)

Data Analyzes – Manufacturing Series												
	Original Data							**SA Data**				
	H-FWER 0.10/0.05		**H-FWER 0.10/0.10**		**VS**	**M7**	**M8**	**H-FWER 0.10/0.05**		**H-FWER 0.10/0.10**		**VS**
series	**quartic**	**TH**	**quartic**	**TH**				**quartic**	**TH**	**quartic**	**TH**	
MW1Fam	12345	12345	12345	12345	12	0.13	0.25	∅	∅	∅	∅	∅
NWTot	12345	12345	12345	12345	12	0.18	0.31	∅	∅	∅	∅	∅
NE1Fam	12345	12345	12345	12345	12	0.16	0.33	∅	∅	∅	∅	∅
NETot	12345	12345	12345	12345	123	0.25	0.27	∅	∅	∅	∅	∅
S1Fam	12345	12345	12345	12345	125	0.22	0.47	∅	∅	∅	∅	∅
STot	124	124	1245	1245	125	0.29	0.57	∅	∅	∅	∅	∅
US1Fam	12345	12345	12345	12345	125	0.17	0.39	∅	∅	∅	∅	∅
USTot	12345	12345	12345	12345	125	0.20	0.42	∅	∅	∅	∅	∅
W1Fam	1234	1234	12345	12345	125	0.21	0.44	∅	∅	∅	∅	∅
WTot	1234	1234	12345	1234	12	0.27	0.56	∅	∅	∅	∅	∅
Total "correct"	10/10	10/10	10/10	10/10	10/10	10/10	10/10	10/10	10/10	10/10	10/10	10/10
Import/Export Series												
	Original Data							**SA Data**				
	H-FWER 0.10/0.05		**H-FWER 0.10/0.10**		**VS**	**M7**	**M8**	**H-FWER 0.10/0.05**		**H-FWER 0.10/0.10**		**VS**
series	**quartic**	**TH**	**quartic**	**TH**				**quartic**	**TH**	**quartic**	**TH**	
M00120	12345	∅	12345	12345	125	0.23	0.48	∅	∅	∅	∅	∅
M00190	12345	12345	12345	12345	1235	0.38	0.59	∅	∅	∅	∅	∅
M3000	12345	12345	12345	12345	234	0.48	0.95	∅	∅	∅	∅	∅
M3010	1234	1234	1234	12345	2345	0.52	0.88	∅	∅	∅	∅	∅
M12060	12345	12345	12345	12345	123	0.53	0.77	∅	∅	∅	∅	∅
X3	12345	12345	12345	12345	2345	0.57	0.94	∅	∅	∅	∅	∅
X00300	134	134	134	134	2	0.56	0.97	∅	∅	∅	∅	∅
X3020	12345	12345	12345	12345	12345	0.39	0.70	∅	∅	∅	∅	∅
X3022	12345	12345	12345	12345	23	0.69	1.04	∅	∅	∅	∅	∅
X10140	1234	1234	1234	1234	15	0.29	0.47	∅	∅	∅	∅	∅
Total "correct"	10/10	9/10	10/10	10/10	10/10	10/10	9/10	10/10	10/10	10/10	10/10	10/10
Retail Series												
	Original Data							**SA Data**				
	H-FWER 0.10/0.05		**H-FWER 0.10/0.10**		**VS**	**M7**	**M8**	**H-FWER 0.10/0.05**		**H-FWER 0.10/0.10**		**VS**
series	**quartic**	**TH**	**quartic**	**TH**				**quartic**	**TH**	**quartic**	**TH**	
s0b441x0	12345	12345	12345	12345	135	0.22	0.41	∅	∅	∅	∅	∅
s0b 44000	12345	12345	12345	12345	2345	0.12	0.26	∅	∅	∅	∅	∅
s0b 44100	12345	12345	12345	12345	135	0.21	0.40	∅	∅	∅	∅	∅
s0b 44130	12345	12345	12345	12345	1235	0.21	0.42	∅	∅	∅	∅	∅
s0b 44200	12345	12345	12345	12345	12345	0.13	0.27	∅	∅	∅	∅	∅
s0b 44300	1234	12345	1234	12345	12345	0.12	0.18	∅	∅	∅	∅	∅
s0b 44312	1234	1234	1234	1234	12345	0.31	0.48	∅	∅	∅	∅	∅
s0b 44400	12345	12345	12345	12345	1235	0.16	0.32	∅	∅	∅	∅	∅
s0b 44410	12345	12345	12345	12345	1235	0.14	0.32	∅	∅	∅	∅	∅
s0b 44500	12345	12345	12345	12345	235	0.14	0.23	∅	∅	∅	∅	∅
Total "correct"	10/10	10/10	10/10	10/10	10/10	10/10	10/10	10/10	10/10	10/10	10/10	10/10
Grand Total "correct"	30/30	29/30	30/30	30/30	30/30	30/30	29/30	30/30	30/30	30/30	30/30	30/30
Average Number	4.67	4.5	4.77	4.8	3.23			0	0	0	0	0

Table 7
Data analyses for 22 German OECD Series comparing our multiple peak diagnostic with VS, M7, and M8 diagnostics. Our multiple peak diagnostic uses the H-FWER method to control the FWER at $\alpha = 0.05$ and $\alpha = 0.10$ and examines the slope at $\delta = 0.10$ (see Section 4.2)

Data Analyzes – OECD DEU												
	Original Data							SA Data				
Series	H-FWER 0.10/0.05		H-FWER 0.10/0.10		VS	M7	M8	H-FWER 0.10/0.05		H-FWER 0.10/0.10		VS
DEU	quartic	TH	quartic	TH				quartic	TH	quartic	TH	
PRMNCG03	12345	12345	12345	12345	12345	0.17	0.33	∅	∅	∅	∅	∅
PRMNCS01	1234	1234	1234	1234	235	0.25	0.79	∅	∅	∅	∅	∅
PRMNIG01	12345	12345	12345	12345	12345	0.28	0.48	∅	∅	∅	∅	∅
PRMNTO01	134	134	134	134	12345	0.26	0.46	∅	∅	∅	∅	∅
PRMNVG01	1235	1235	1235	1235	2345	0.29	0.46	∅	∅	∅	∅	∅
SLMNCD01	1245	1235	1245	1235	23	0.22	0.40	∅	∅	∅	∅	∅
SLMNCN01	1345	2345	1345	2345	123	0.37	0.72	∅	∅	∅	∅	∅
SLMNDM01	2345	2345	2345	2345	123	0.32	0.63	∅	∅	∅	∅	∅
SLMNEX01	12345	12345	12345	12345	12	0.32	0.51	1:5	∅	12345	12345	∅
SLMNIG01	2345	2345	2345	2345	123	0.21	0.65	∅	∅	∅	∅	∅
SLMNTO01	245	345	245	345	23	0.20	0.66	∅	∅	234	∅	∅
SLRTCR01	1345	1345	1345	1345	1234	0.19	0.51	∅	∅	∅	∅	∅
SLRTTO01	12345	12345	12345	12345	12345	0.12	0.25	∅	∅	∅	∅	∅
SLRTTO02	12345	12345	12345	12345	12345	0.13	0.29	∅	∅	∅	∅	∅
SLWHTO01	2345	2345	2345	2345	123	0.20	0.62	∅	∅	134	123	∅
SLWHTO02	2345	2345	2345	2345	123	0.20	0.62	∅	∅	134	123	∅
UNLVRG01	23	23	23	23	124	0.23	0.48	∅	∅	1	∅	5
UNLVSUMA	345	345	345	345	12	0.30	0.53	12	∅	1245	12345	∅
UNLVSUTT	234	234	234	234	12	0.24	0.53	∅	∅	45	45	25
UNRTRG01	235	1245	235	1245	123	0.19	0.59	∅	∅	2345	2345	∅
XTEXVA01	1234	1234	1234	1234	23	0.28	0.77	∅	∅	∅	∅	∅
XTIMVA01	234	1234	2345	12345	23	0.31	0.95	∅	∅	∅	∅	∅
Total “correct”	22/22	22/22	22/22	22/22	22/22	22/22	22/22	20/22	22/22	14/22	16/22	20/22
Average Number	3.86	3.95	3.91	4.00	3.23			0	0	1.14	1.00	0.14

Table 8
Data analyses for 15 Euro-area OECD Series comparing our multiple peak diagnostic with VS, M7, and M8 diagnostics. Our multiple peak diagnostic uses the H-FWER method to control the FWER at $\alpha = 0.05$ and $\alpha = 0.10$ and examines the slope at $\delta = 0.10$ (see Section 4.2)

Data Analyzes – OECD EMU												
	Original Data							SA Data				
series	H-FWER 0.10/0.05		H-FWER 0.10/0.10		VS	M7	M8	H-FWER 0.10/0.05		H-FWER 0.10/0.10		VS
EMU	quartic	TH	quartic	TH				quartic	TH	quartic	TH	
PRCNTO01	345	1345	345	1345	12345	0.14	0.44	∅	∅	∅	∅	∅
PRINTO01	12345	12345	12345	12345	12345	0.10	0.23	∅	∅	∅	∅	∅
PRMNCG03	1245	1245	1245	1245	12345	0.12	0.32	1	1	1234	1234	∅
PRMNCS01	1234	12345	1234	12345	1234	0.22	0.47	∅	∅	∅	∅	∅
PRMNIG01	12345	12345	12345	12345	2345	0.15	0.27	∅	∅	∅	∅	∅
PRMNTO01	2345	2345	2345	2345	2345	0.14	0.23	∅	∅	1	∅	∅
PRMNVG01	1234	12345	1234	12345	2345	0.13	0.23	∅	∅	∅	∅	∅
SLMNCN02	12345	12345	12345	12345	123	0.31	0.57	∅	∅	∅	∅	∅
SLMNIG02	12345	12345	12345	12345	23	0.21	0.45	∅	∅	∅	∅	∅
SLMNTO02	1345	2345	1345	2345	23	0.20	0.41	24	∅	24	34	∅
SLMNVG02	12345	12345	12345	12345	2345	0.17	0.30	1	1	1245	1345	∅
SLRTTO01	12345	12345	12345	12345	12345	0.05	0.12	∅	∅	∅	∅	∅
SLRTTO02	12345	12345	12345	12345	12345	0.05	0.11	∅	∅	∅	∅	∅
XTEXVA01	1345	2345	1345	2345	23	0.31	0.57	∅	∅	12	12	∅
XTIMVA01	2345	2345	2345	2345	23	0.40	0.72	∅	∅	∅	∅	∅
Total “correct”	15/15	15/15	15/15	15/15	15/15	15/15	15/15	12/15	13/15	10/15	11/15	15/15
Average Number	4.40	4.60	4.40	4.60	3.73			0.20	0.13	0.87	0.80	0

Table 9
Data analyses for 11 French OECD Series comparing our multiple peak diagnostic with VS, M7, and M8 diagnostics. Our multiple peak diagnostic uses the H-FWER method to control the FWER at $\alpha = 0.05$ and $\alpha = 0.10$ and examines the slope at $\delta = 0.10$ (see Section 4.2)

Data Analyzes – OECD FRA												
Original Data								SA Data				
series	H-FWER 0.10/0.05		H-FWER 0.10/0.10		VS	M7	M8	H-FWER 0.10/0.05		H-FWER 0.10/0.10		VS
FRA	quartic	TH	quartic	TH				quartic	TH	quartic	TH	
PRAFAG01	12345	12345	12345	12345	123	0.13	0.29	∅	∅	∅	∅	∅
PRNCTO01	2345	2345	2345	2345	235	0.14	0.44	∅	∅	∅	∅	∅
PRMNCG01	12345	12345	12345	12345	234	0.15	0.38	1	1	1	1	∅
PRMNCS01	12345	12345	12345	12345	234	0.25	0.58	∅	∅	∅	∅	∅
PRMNIG01	12345	12345	12345	12345	12345	0.11	0.26	∅	∅	∅	∅	∅
PRMNTO01	12345	12345	12345	12345	12345	0.16	0.29	123	123	123	1234	∅
PRMNVE01	12345	12345	12345	12345	12345	0.24	0.34	∅	∅	1245	1245	∅
SLRTCR01	1345	2345	1345	2345	123	0.27	0.71	∅	∅	∅	∅	∅
SLRTTO02	12345	12345	12345	12345	12345	0.16	0.36	∅	∅	∅	∅	∅
XTEXVA01	1345	1345	1345	1345	23	0.14	0.44	∅	∅	∅	∅	∅
XTIMVA01	1245	1245	1245	1245	23	0.18	0.54	∅	∅	∅	∅	∅
Total "correct"	11/11	11/11	11/11	11/11	11/11	11/11	11/11	9/11	9/11	8/11	8/11	11/11
Average Number	4.64	4.64	4.64	4.64	3.55			0.36	0.36	0.73	0.81	0

Table 10
Data analyses for 17 Great Britain OECD Series comparing our multiple peak diagnostic with VS, M7, and M8 diagnostics. Our multiple peak diagnostic uses the H-FWER method to control the FWER at $\alpha = 0.05$ and $\alpha = 0.10$ and examines the slope at $\delta = 0.10$ (see Section 4.2)

Data Analyzes – OECD GBR												
Original Data								SA Data				
series	H-FWER 0.10/0.05		H-FWER 0.10/0.10		VS	M7	M8	H-FWER 0.10/0.05		H-FWER 0.10/0.10		VS
GBR	quartic	TH	quartic	TH				quartic	TH	quartic	TH	
PPIAMP01	1234	1234	1234	1234	24	0.56	1.58	∅	∅	∅	∅	∅
PPIAMP02	1	1	123	123	2	0.53	0.92	∅	∅	∅	∅	∅
PPIPFU01	1345	12345	1345	12345	12	0.64	0.59	∅	∅	∅	∅	∅
PRINTO01	1345	1345	1345	1345	23	0.16	0.40	∅	∅	∅	∅	∅
PRMNCG02	2345	2345	2345	2345	123	0.23	0.56	∅	∅	∅	∅	∅
PRMNCG03	12345	12345	12345	12345	123	0.20	0.49	∅	∅	∅	∅	∅
PRMNCS01	123	12	123	1234	12	0.68	1.31	∅	∅	∅	∅	∅
PRMNIG01	2345	2345	2345	2345	123	0.15	0.47	1	∅	∅	∅	∅
PRMNTO01	1345	1345	1345	1345	23	0.17	0.45	∅	∅	∅	∅	∅
PRMNVE02	12345	12345	12345	12345	12345	0.25	0.76	∅	∅	∅	∅	∅
PRMNVE03	1234	1234	1234	1234	234	0.29	0.91	∅	∅	∅	∅	∅
PRMNVG01	124	134	124	134	234	0.18	0.58	∅	∅	∅	∅	∅
SLRTCR03	1345	1345	1345	1345	124	0.42	0.74	124	123	124	123	∅
SLRTTO02	12345	12345	12345	12345	12345	0.05	0.15	1234	∅	1234	12345	∅
UNLVRG01	12345	12345	12345	12345	1245	0.63	0.61	∅	∅	∅	∅	∅
XTEXVA01	134	134	1345	1345	23	0.34	1.02	∅	∅	∅	∅	∅
XTIMVA01	23	23	23	23	23	0.31	0.90	∅	∅	∅	∅	∅
Total "correct"	17/17	17/17	17/17	17/17	17/17	17/17	14/17	14/17	16/17	13/17	15/17	17/17
Average Number	3.76	3.76	3.94	4.06	2.76			0.47	0.18	0.53	0.47	0

For the multiple peak-testing scenario, we employ known results from the multiple testing literature (*i.e.*, the applications to controlling FWER) to combine the p-values from the five seasonal frequencies in such a way to dramatically increase statistical power, as demonstrated in Table 4. Although there is some departure in the size (Table 3) for the multiple peak testing, the results are still quite usable. On a typical batch of seasonal series, the number of Type I errors are as expected, and the power is quite decent (Tables 5-10). While our method compares quite favorably to the VS, M7, and M8 diagnostics, neither of the diagnostics provide a p-value and only the former can distinguish which spectral peaks are contributing to seasonal behavior. This aspect is important to the seasonal adjuster, who wants to know not only that there may be residual seasonality, but also at what seasonal frequencies, so as to take appropriate action to alter the seasonal adjustment filters (this can be done by smoothing over additional years, which is accomplished by changing the seasonal filters in X-11-ARIMA; alternatively, one might consider shortening the series. For current research on a model-based approach to designing SA filters targeted for specific seasonal frequencies, see Aston, Findley, McElroy, Wills, and Martin (2007)).

The choice of kernel surely has some impact on the results, although we found little difference in practice between the quartic and TH kernels; the TH may be marginally more powerful. Of course, plenty of other popular kernels may also be utilized by a practitioner, and we have only chosen two that seemed intuitive and straightforward to implement. The choice of the location μ is clearly dictated by the characterization of seasonality. Since statistical power generally decreased with β, we always recommend taking the maximal β such that the kernel supports are disjoint, which guarantees the asymptotic independence property of the various diagnostics that is crucial to our multiple testing method.

Finally, the asymptotic results require that the data be differenced to stationarity. Recognizing that economic time series are typically nonstationary, it is desirable to trend-difference seasonally adjusted data before applying our diagnostic. This differencing may dampen the detection of the first seasonal peak, so practitioners may "re-color" the data as described in Section 3.4.

Acknowledgements

This paper is released to inform interested parties of ongoing research and to encourage discussion of work in progress. The views expressed are those of the authors and not necessarily those of the U.S. Census Bureau. Holan's research was supported by an ASA/NSF/BLS research fellowship.

Appendix

Here we derive asymptotic formulas for the statistical measures ψ_A of slope and convexity. These results can then be applied in the testing paradigm to get asymptotic critical values. Some mild conditions on the data are required for the asymptotic theory; we follow the material in Taniguchi and Kakizawa (2000, Section 3.1.1). Condition (B), due to Brillinger (1981), states that the process is strictly stationary and condition (B1) of Taniguchi and Kakizawa (2000, page 55) holds. Condition (HT), due to Hosoya and Taniguchi (1982), states that the process has a linear representation, and conditions (H1) through (H6) of Taniguchi and Kakizawa (2000, pages 55-56) hold. Assumption 1 (8) of Chiu (1988) is a summability condition on various higher order cumulants, which is satisfied, for example, by a Gaussian process with spectral density in C^2. None of these conditions are stringent; for example, a causal linear process with fourth moments satisfies (HT). The main result is a joint convergence of any two measures $\psi_A(I)$; *e.g.*, these can be a slope and convexity measure with the same kernel A. We present the general theorem that covers these two cases.

Theorem 1 *Suppose that the fourth order cumulants of* $\{X_t\}$ *vanish*; *that either condition* (*B*) *or* (*HT*) *holds*; *and that Assumption* 1 (8) *of Chiu* (1988) *holds*. *Let the kernels* A *and* B *satisfy conditions* (*i*) *through* (*iv*) *of Section* 2.1. *Then*

$$\left\{\sqrt{n}\frac{(\theta_A(I) - \theta_A(f))}{\sqrt{\theta_{A^2}(I^2)/2}},\right.$$

$$\left.\sqrt{n}\frac{(\theta_B(I) - \theta_B(f))}{\sqrt{\theta_{B^2}(I^2)/2}}\right\} \overset{\mathcal{L}}{\Rightarrow} N(0, V)$$

as $n \to \infty$. *Here* 0 *denotes the zero vector* $(0, 0)'$, *and* V *is a* 2×2 *matrix with entries*

$$V_{11} = V_{22} = 1 \quad V_{12} = V_{21} = \frac{\theta_{AB}(f^2)}{\sqrt{\theta_{A^2}(f^2)\theta_{B^2}(f^2)}}.$$

Proof. First we establish that $\theta_{A^2}(I^2) \xrightarrow{a.s.} 2\theta_{A^2}(f^2)$. Since the kernel A is continuous in an interval (such as $[\mu - \beta/2, \mu + \beta/2]$), this result follows directly from Corollary 1 of Chiu (1988), noting that they deal with the Riemann sums approximation to the integral functional (Chiu (1988) also defines the periodogram with a 2π factor). Of course the same results holds with B in place of A. Secondly, consider the joint convergence of $\theta_A(I)$ and

$\theta_B(I)$. We use the Cramer-Wold device, and apply Lemma 3.1.1 of Taniguchi and Kakizawa (2000), appropriately generalized to include non-even functions (*cf.* Theorem 3 of Chiu (1988)). Hence for any x, y real,

$$\sqrt{n}\left(x\,\frac{(\theta_A(I) - \theta_A(f))}{\sqrt{\theta_{A^2}(f^2)}} + y\,\frac{(\theta_B(I) - \theta_B(f))}{\sqrt{\theta_{B^2}(f^2)}} \right)$$

$$\overset{\mathcal{L}}{\Rightarrow} N\left(0,\ \frac{1}{2\pi}\int_{-\pi}^{\pi} (C(\lambda)C(-\lambda) + C^2(\lambda))\, f^2(\lambda)\, d\lambda\right)$$

using Slutsky's Theorem (Bickel and Doksum 1977), where the kernel C is defined by

$$C(\lambda) = \frac{x}{\sqrt{\theta_{A^2}(f^2)}}\,A(\lambda) + \frac{y}{\sqrt{\theta_{B^2}(f^2)}}\,B(\lambda).$$

Clearly $C(\lambda)\;C(-\lambda) = 0$ and

$$C^2(\lambda) = \frac{x^2}{\theta_{A^2}(f^2)}\,A^2(\lambda)$$

$$+\,2\frac{xy}{\sqrt{\theta_{A^2}(f^2)}\sqrt{\theta_{B^2}(f^2)}}\,A(\lambda)\,B(\lambda)$$

$$+\,\frac{y^2}{\theta_{B^2}(f^2)}\,B^2(\lambda).$$

By taking x and y to be zero and one in various combinations, we deduce the stated variance matrix V.

Next we discuss the multiple-peak testing scenario. So suppose that we have a finite collection of kernels A_i for $i = 1, 2, ..., d$, each of which satisfies the assumptions of Section 2. Then we can easily generalize Theorem 1 from two to d kernels as follows. The asymptotic covariance matrix V will have ij^{th} entry

$$\frac{\theta_{A_i A_j}(f^2)}{\sqrt{\theta_{A_i^2}(f^2)\,\theta_{A_j^2}(f^2)}}.$$

Thus, if the support for any two kernels is disjoint we obtain asymptotic independence, and can therefore invoke the H-FWER multiple testing procedure.

References

Aston, J., Findley, D., McElroy, T., Wills, K. and Martin, D. (2007). New ARIMA Models for Seasonal Time Series and Their Application to Seasonal Adjustment and Forecasting. *SRD Research Report No. RRS 2007*-14, *U.S. Census Bureau.*

Bell, W., and Hillmer, S. (1984). Issues involved with the seasonal adjustment of economic time series. *Journal of Business and Economic Statistics*, 2, 291-320.

Bickel, P., and Doksum, K. (1977). *Mathematical Statistics*: *Basic Ideas and Selected Topics*. Englewood Cliffs, New Jersey: Prentice Hall.

Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society*, Series B, 57, 289-300.

Brillinger, D. (1981). *Time Series Data Analysis and Theory*. San Francisco: Holden-Day.

Chiu, S. (1988). Weighted least squares estimators on the frequency domain for the parameters of a time series. *The Annals of Statistics*, 16, 1315-1326.

Evans, T., Holan, S. and McElroy, T. (2006). Evaluating measures for assessing spectral peaks. *2006 Proceedings American Statistical Association*, [CD-ROM]: Alexandria, VA}.

Findley, D. (2006). Personal communication.

Findley, D.F., Monsell, B.C., Bell, W.R., Otto, M.C. and Chen, B.C. (1998). New capabilities and methods of the X-12-ARIMA seasonal adjustment program. *Journal of Business and Economic Statistics*, 16, 127-177 (with discussion).

Hosoya, Y., and Taniguchi, M. (1982). A central limit theorem for stationary processes and the parameter estimation of linear processes. *The Annals of Statistics*, 10, 132-153.

Lothian, J., and Morry, M. (1978). A test for identifiable seasonality when using the X-11-ARIMA program. Working Paper, Time Series Research and Analysis Division, Statistics Canada.

Maravall, A., and Caporello, G. (2004). Program TSW: Revised Reference Manual. *Working Paper 2004*, *Research Department*, *Bank of Spain*. http://www.bde.es.

Nerlove, M. (1964). Spectral analysis of seasonal adjustment procedures. *Econometrica*, 32, 241-286.

Newton, H., and Pagano, M. (1983). A method for determining periods in time series. *Journal of the American Statistical Association*, 78, 152-157.

Parzen, E. (1983). Autoregressive spectral estimation. *Handbook of Statistics III*, (Ed. D. Brillinger and P. Krishnaiah), Amsterdam: North Holland, 221-247.

Priestley, M. (1981). *Spectral Analysis and Time Series*. London: Academic Press.

Soukup, R.J., and Findley, D.F. (1999). On the spectrum diagnostics used by X-12-ARIMA to indicate the presence of trading day effects after modeling or adjustment. Also www.census.gov/pub/ts/papers/rr9903s.pdf. *Proceedings of the Business and Economic Statistics Section*, American Statistical Association, 144-149.

Taniguchi, M., and Kakizawa, Y. (2000). *Asymptotic Theory of Statistical Inference for Time Series*. New York City, New York: Springer-Verlag.

U.S. Census Bureau (2002). X-12 ARIMA Reference Manual (Version 0.2.10), Washington, DC.

Survey Methodology, June 2009
Vol. 35, No. 1, pp. 85-99
Statistics Canada, Catalogue No. 12-001-X

On the definition and interpretation of interviewer variability for a complex sampling design

Siegfried Gabler and Partha Lahiri [1]

Abstract

Interviewer variability is a major component of variability of survey statistics. Different strategies related to question formatting, question phrasing, interviewer training, interviewer workload, interviewer experience and interviewer assignment are employed in an effort to reduce interviewer variability. The traditional formula for measuring interviewer variability, commonly referred to as the interviewer effect, is given by $ieff := deff\_int = 1 + (\overline{n}_{\text{int}} - 1)\rho_{\text{int}}$, where ρ_{int} and $\overline{n}_{\text{int}}$ are the intra-interviewer correlation and the simple average of the interviewer workloads, respectively. In this article, we provide a model-assisted justification of this well-known formula for equal probability of selection methods (epsem) with no spatial clustering in the sample and equal interviewer workload. However, spatial clustering and unequal weighting are both very common in large scale surveys. In the context of a complex sampling design, we obtain an appropriate formula for the interviewer variability that takes into consideration unequal probability of selection and spatial clustering. Our formula provides a more accurate assessment of interviewer effects and thus is helpful in allocating more reasonable amount of funds to control the interviewer variability. We also propose a decomposition of the overall effect into effects due to weighting, spatial clustering and interviewers. Such a decomposition is helpful in understanding ways to reduce total variance by different means.

Key Words: Interviewer effect; Interviewer workloads; Intra-interviewer correlation; Spatial clustering; Unequal weighting.

1. Introduction

A major source of measurement errors in surveys is due to the interviewer. This fact was recognized as early as 1929 by Rice and later by many survey researchers. Factors such as the quality of questionnaire design and the interviewer can influence the interviewer effects on survey statistics.

The interviewer can introduce homogeneity in survey data, which generally reduces the effective sample size and thereby increases the total variance of a survey estimator. The within interviewer homogeneity has been traditionally measured by the intra-interviewer correlation coefficient ρ_{int}. The magnitude of the intra-interviewer correlation was studied by many researchers, mostly in the context of telephone surveys without any spatial clustering effects (Kish 1962; Gray 1956; Hanson and Marks 1958; Tucker 1983; Groves and Magilavy 1986; Heeb and Gmel 2001, and others). Researchers have argued that the nature of the survey items may affect the value of ρ_{int}. Attitude items and complex factual items are considered more sensitive to the intra-interviewer correlation than simple factual items are (Collins and Butcher 1982; Feather 1973; Fellegi 1964; Gray 1956; Hansen, Hurwitz and Bershad 1961). According to Groves (1989), values above 0.1 are seldom observed. See Schnell and Kreuter (2005) for further discussion on this issue.

As noted by several researchers, the standard interviewer effect formula $1 + (\overline{n}_{\text{int}} - 1)\rho_{\text{int}}$ suggests that even with a small intra-interviewer correlation, the interviewer effect could be substantial simply due to a high average interviewer workload. For example, when $\rho_{\text{int}} = 0.01$ and $\overline{n}_{\text{int}} = 70$ we have $ieff = 1.69$ (Schnell and Kreuter 2005). Note that a high average interviewer workload (*e.g.*, between 60 and 70) is very common in telephone surveys (Tucker 1983; Groves and Magilavy 1986). For the European Social Survey, Philippens and Loosveldt (2004) provided box plots of the intra-interviewer correlations and the interviewer workloads for 18 participating countries.

The interviewer effect or variance is generally defined as the inflation to the total variance caused solely by the interviewers. For an epsem design with equal interviewer workload, the interviewer variance for the sample mean is simply given by $1 + (n_{\text{int}} - 1)\rho_{\text{int}}$, where n_{int} is the common interviewer workload. For complex surveys with unequal interviewer workload, survey researchers frequently use a simple modification of this formula where the common interviewer workload is replaced by the average interviewer workload, *i.e.*, the formula $1 + (\overline{n}_{\text{int}} - 1)\rho_{\text{int}}$. In Section 2, we argue that this standard formula $1 + (\overline{n}_{\text{int}} - 1)\rho_{\text{int}}$ cannot be interpreted as an inflation to the total variance caused by the interviewers even for an epsem design with unequal interviewer workload. In Sections 2-4, we observe that the interviewer variance definition depends

[1] Siegfried Gabler, GESIS, P.O. Box 12 21 55, 68072 Mannheim, Germany. E-mail: siegfried.gabler@gesis.org; Partha Lahiri, University of Maryland, College Park, U.S.A. E-mail: plahiri@survey.umd.edu.

on the nature of the complex sampling design and also on the interviewer workload assignment. In this paper, we provide appropriate definitions of the interviewer variance in different survey scenarios. A reliable definition of the interviewer variance is helpful in determining actions that need to be taken in order to reduce interviewer variability. This paper is foremost applicable to the planning of surveys rather than analyzing survey data. In other words, in this paper we have concentrated on the definitions and interpretation of the interviewer variability and not on estimating it from a given survey.

In Section 2, we consider an epsem design with no spatial clustering and provide a model-assisted interpretation of *ieff*. We show that for the equal interviewer workload *ieff* is simply the ratio of the variances of the sample mean under a correlated model that accounts for the homogeneity of the observations collected by the same interviewer and a simple uncorrelated model that fails to account for such homogeneity. Thus, multiplying the variance of the sample mean for simple random sampling by the *ieff* one can obtain the total variance of the sample mean that incorporates both the sampling and the interviewer variability. This is a very intuitive interpretation of *ieff* and complements the model-assisted justification given earlier by Kish (1962). In this section, we also show that for an epsem design *ieff* is lower than the model-assisted interviewer effect formula if the interviewer workload varies and the intra-interviewer correlation is positive. Thus, the survey designer who uses *ieff* would give less effort to control interviewer variability than is really needed. In this situation, an appropriate interviewer effect formula can be obtained from *ieff* when a weighted average interviewer workload is used in place of the usual simple average.

In Section 3, we entertain the possibility of unequal weighting but no spatial clustering. We obtain a model-assisted interpretation for *ieff* if and only if the respondents interviewed by the same interviewer share the same sampling weight and the interviewer workload is inversely proportional to the square of the common weight for the interviewer. Interestingly, unlike the epsem design, equal interviewer workload does not necessarily guarantee a model-assisted interpretation for *ieff*. When there is an equal interviewer workload and there is at least one interviewer for which the respondents do not all share the same sampling weight, we show that *ieff* is always higher than the model-assisted formula. We also point out the factors that cause the difference between these two formulae. These results have a practical relevance in terms of saving survey costs. To be specific, the survey designer who uses *ieff* is likely to allocate more funds to control interviewer variability than is really needed. We have also cited some situations where *ieff* could have an underestimation problem and thus survey designers who use *ieff* could give less emphasis to control the interviewer effects. Our formula provides a more accurate assessment of interviewer variability and thus is helpful in the allocation of more reasonable amount of funds to control the interviewer variability. Furthermore, the change in planning formulae will affect the sample size.

In many large scale sample surveys, due to various organizational and financial reasons such as the absence of a general population register or to reduce the overall survey costs, a multi-stage clustered sampling design is considered to be a cost-efficient alternative to simple random sampling. Under a multi-stage clustered sampling design, respondents who live in close spatial proximity of each other get selected. Respondents living in the same spatial cluster tend to share similar attitudes because of their similar socio-economic background and hence increase the internal homogeneity of the survey data. This spatial homogeneity violates the iid (independently identically distributed) assumption frequently used in standard statistical inferential procedures and so does the clustering within the interviewers. This fact has been recognized by many survey researchers and adjustments to various statistical procedures and the related software issues have been addressed in the literature (see Rao and Scott 1984; Skinner, Holt and Smith 1989; Biemer and Trewin 1997; Chambers and Skinner 2003; among others). In Section 4, we present a new definition of the interviewer variability in the presence of unequal weighting and spatial clustering. In the presence of spatial clustering, we argue that *ieff* generally has a tendency to overestimate the interviewer variability. Thus for complex surveys involving spatial clustering, *ieff* may unnecessarily give a false alarm regarding the magnitude of the interviewer variability.

In Section 5, we discuss the effects due to the combined effects of weighting, spatial clustering and the interviewer. The formula for overall effects offers an accurate determination of the sample size at the planning stage. We provide a nice factorization of the overall effects into the effects due to weighting, clustering and interviewer. Such a decomposition of the overall effects can be useful in understanding ways to reduce the total variance by different means. In discussing Verma, Scott and O'Muircheartaigh (1980), Hedges mentioned the need for such an overall effect formula. We generalize a formula earlier proposed by Davis and Scott (1995) to a non-epsem design and for a general correlation model valid for both discrete and continuous data. We present proofs of all the technical results in the Appendix.

2. EPSEM design with no spatial clustering

Let y_{ik} denote the observation obtained from the k^{th} respondent interviewed by the i^{th} interviewer $(i = 1, ..., I;\ k = 1, ... n_i)$. Define $n = \sum_{i=1}^{I} n_i$, the total sample size, $\bar{y} = 1/n \sum_{i=1}^{I} \sum_{k=1}^{n_i} y_{ik}$, the unweighted sample mean, and $\bar{n}_{\text{int}}(\mathbf{a}) = \sum_{i=1}^{I} a_i n_i$, a weighted average of the interviewer workload, where a_i is an arbitrary weight attached to the i^{th} interviewer workload and $\mathbf{a} = (a_1, ..., a_I)$.

We shall first provide a model-assisted justification of the traditional interviewer effect formula, *i.e.*, $ieff = 1 + (\bar{n}_{\text{int}} - 1)\rho_{\text{int}}$, where $\bar{n}_{\text{int}}$ is the unweighted average of interviewer workload. Note that $\bar{n}_{\text{int}} = \bar{n}_{\text{int}}(\mathbf{a}_0)$, with $\mathbf{a}_0 = (a_{01}, ..., a_{0I})$, $a_{0i} = 1/I$ and $ieff = ieff(\mathbf{a}_0)$. Using Result 1 given in the Appendix, we get

$$ieff(\mathbf{a}_1) = \frac{\text{Var}_{M_2}(\bar{y})}{\text{Var}_{M_1}(\bar{y})} = 1 + [\bar{n}_{\text{int}}(\mathbf{a}_1) - 1]\rho_{\text{int}},$$

where $\mathbf{a}_1 = (a_{11}, ..., a_{1I})$, with $a_{1i} = n_i/n$. In the above, $\text{Var}_{M_1}(\bar{y})$ and $\text{Var}_{M_2}(\bar{y})$ are the variances of $\bar{y}$ under the following two models, respectively,

$$M_1\text{: Cov}(y_{ik}, y_{i'k'}) = \begin{cases} \sigma^2 & \text{if } i = i',\ k = k', \\ 0 & \text{otherwise,} \end{cases}$$

$$M_2\text{: Cov}(y_{ik}, y_{i'k'}) = \begin{cases} \sigma^2 & \text{if } i = i',\ k = k', \\ \rho_{\text{int}}\sigma^2 & \text{if } i = i',\ k \neq k', \\ 0 & \text{otherwise.} \end{cases}$$

Note that unlike model M_1, model M_2 introduces homogeneity of the observations collected by the same interviewer.

Remark 2.1: It follows from the corollary to Result 1, given in the Appendix, that for $\rho_{\text{int}} > 0$, $ieff(\mathbf{a}_1) = ieff$ if and only if $n_i = n/I$ for all i, *i.e.*, if and only if each interviewer has the same workload. For the balanced case, Kish (1962) provided a model-assisted justification of $ieff$ using a linear mixed model, which is a special case of M_2. For the unbalanced case, it is interesting to note the similarity between the interviewer variability formula $ieff(\mathbf{a}_1)$ and the design effects formula given in (A3) of Holt in discussing Verma *et al*. (1980).

Remark 2.2: It follows from the corollary to Result 1 that if $\rho_{\text{int}} > 0$ and n_i's are not equal then $ieff(\mathbf{a}_1) > ieff$.

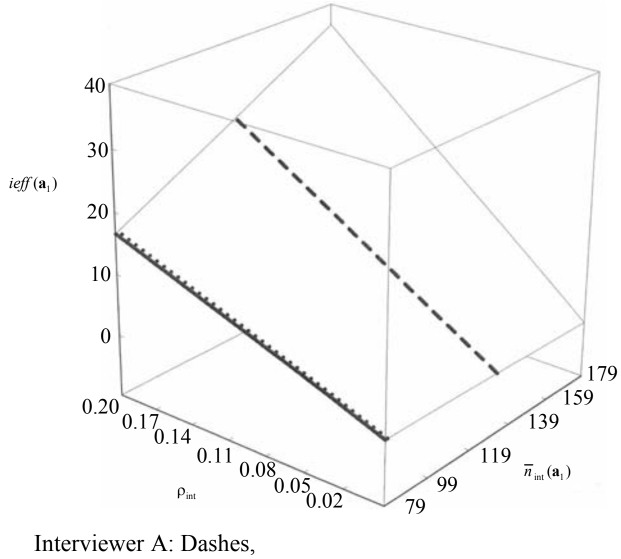
In the following example, we demonstrate the extent to which $ieff(\mathbf{a}_1)$ and $ieff$ could differ for different interviewer workload patterns.

Example 1: In Table 1, we consider three different workload assignments for ten interviewers, each with $n = 790$. Case A) represents the most variable workload assignment with a standard deviation = 68.3; Case B) is nearly balanced with a standard deviation = 9.5; Case C) corresponds to the equal interviewer assignment.

Table 1
Three different interviewer workload assignments (Example 1)

	Interviewer workload pattern		
Interviewer	**A)**	**B)**	**C)**
1	4	70	79
2	10	70	79
3	20	70	79
4	34	70	79
5	52	70	79
6	74	88	79
7	100	88	79
8	130	88	79
9	164	88	79
10	202	88	79
n	790	790	790
$\bar{n}_{\text{int}}(\mathbf{a}_1)$	132	80	79

Let $ieff(\mathbf{a}_{1;A})$, $ieff(\mathbf{a}_{1;B})$, and $ieff(\mathbf{a}_{1;C}) = ieff$ denote $ieff(\mathbf{a}_1)$, the model-assisted interviewer variance formula corresponding to the cases A, B and C, respectively. For $\rho_{\text{int}} > 0$ the function $ieff(\mathbf{a}_1)$ is Schur-convex, which explains the fact $ieff(\mathbf{a}_{1;A}) \geq ieff(\mathbf{a}_{1;B}) \geq ieff(\mathbf{a}_{1;C}) = ieff$. Figure 1 provides the values of the interviewer variance obtained from the standard formula (*i.e.*, $ieff$) and our model-assisted interview variance formula for all combinations of the two influencing factors, *i.e.*, weighted average of interviewer workload and the intra-interviewer correlation. From Figure 1, it is interesting to note that $ieff$ could underreport by about 100%.



Interviewer A: Dashes,
Interviewer B: Spaced dots and
Interviewer C: Solid line

Figure 1 A graph of $ieff(\mathbf{a}_1)$ *vs.* ρ_{int} for different $\bar{n}_{\text{int}}(\mathbf{a}_1)$

3. Unequal weighting with no spatial clustering

In this section, we consider the situation when we have unequal weights. Let w_{ik} be the survey weight attached to the k^{th} respondent interviewed by the i^{th} interviewer. In this situation, a weighted mean $\bar{y}_w = \sum_i \sum_k w_{ik} y_{ik} / \sum_i \sum_k w_{ik}$ is a popular estimator of the finite population mean (See Brewer 1963; Hájek 1971) and the model-assisted interviewer variance formula is given by

$$ieff_w = \frac{\text{Var}_{M_2}(\bar{y}_w)}{\text{Var}_{M_1}(\bar{y}_w)} = 1 + \rho_{\text{int}} \left(\frac{\sum_i \left(\sum_k w_{ik} \right)^2}{\sum_i \sum_k w_{ik}^2} - 1 \right).$$

See Result 1 given in the Appendix.

Define $\bar{w}_i = 1/n_i \sum_{k=1}^{n_i} w_{ik}$, the average survey weight for the i^{th} interviewer and $\sigma_i^2 = 1/n_i \sum_k w_{ik}^2 - \bar{w}_i^2$, the variance of the survey weights for the i^{th} interviewer. It can be shown that

$$ieff_w = 1 + \rho_{\text{int}} (\bar{n}_w - 1),$$

where

$$\bar{n}_w = \frac{\sum_i n_i^2 \bar{w}_i^2}{\sum_i n_i \bar{w}_i^2 + \sum_i n_i \sigma_i^2}.$$

Note that, in general, $ieff_w$ cannot be written in the form $ieff_w = 1 + \rho_{\text{int}} (\bar{n}_{\text{int}}(\mathbf{a}) - 1)$ with $\sum_i a_i = 1$.

Remark 3.1: From Result 2 in the Appendix, we have

$$ieff_w \leq ieff(\mathbf{a}_2),$$

where

$$\mathbf{a}_2 = (a_{21}, \ldots, a_{2I}), \text{ with } a_{2i} = \frac{\sum_k w_{ik}^2}{\sum_i \sum_k w_{ik}^2}.$$

In the above, for $\rho_{\text{int}} > 0$, $ieff_w = ieff(\mathbf{a}_2)$ if and only if all σ_i^2 are zero. Thus, $ieff(\mathbf{a}_2)$ can be interpreted as a conservative interviewer variance.

Equality holds if and only if $w_{ik} = \bar{w}_i$ for all i and k in which case

$$ieff_w = ieff(\mathbf{a}_2^*),$$

where

$$\mathbf{a}_2^* = (a_{21}^*, \ldots, a_{2I}^*), \text{ with } a_{2i}^* = \frac{n_i \bar{w}_i^2}{\sum_i n_i \bar{w}_i^2}.$$

Thus, the formulae $ieff_w$ and $ieff(\mathbf{a}_2^*)$ are equivalent if and only if the survey weights are all the same for a given interviewer. One example of such a design is an epsem design for which we have

$$a_{2i}^* = \frac{n_i}{n}$$

and

$$ieff_w = ieff(\mathbf{a}_2^*) = ieff(\mathbf{a}_1).$$

Now we shall try to understand the factors that explain the difference between $ieff_w$ and $ieff$. To this end, define

$\bar{w} = 1/n \sum_{i=1}^{I} \sum_{k=1}^{n_i} w_{ik} = \sum_{i=1}^{I} n_i / n \, \bar{w}_i$, the average survey weight for all interviewers,

$SSB = \sum_{i=1}^{I} n_i (\bar{w}_i - \bar{w})^2$, the between interviewer sum of squares of the survey weights,

$SSW = \sum_{i=1}^{I} \sum_{k=1}^{n_i} (w_{ik} - \bar{w}_i)^2 = \sum_{i=1}^{I} n_i \sigma_i^2$, the within interviewer sum of squares of the survey weights,

$SST = SSB + SSW$, the total sum of squares of the survey weights,

$\tau_w = SSW/SST$, an indicator of the relative contribution of the within interviewer variability of survey weights to the total variability,

$CV_w = \sqrt{SST/n} \, / \, \bar{w}$, the coefficient of variation of the survey weights in the entire sample.

It can be shown that (see Result 4)

$ieff_w - ieff$

$$= \frac{\bar{n}_{\text{int}}}{SST + n\bar{w}^2} \left[\sum_{i=1}^{I} \left(\frac{n_i}{\bar{n}_{\text{int}}} - 1 \right) n_i \bar{w}_i^2 - SSW \right] \rho_{\text{int}} \tag{1}$$

$$= \frac{\bar{n}_{\text{int}}}{(1 + CV_w^{-2}) SST} \left[\sum_{i=1}^{I} \left(\frac{n_i}{\bar{n}_{\text{int}}} - 1 \right) n_i \bar{w}_i^2 - SSW \right] \rho_{\text{int}} \tag{2}$$

$$= \frac{\bar{n}_{\text{int}} \tau_w}{1 + CV_w^{-2}} \left(\frac{\sum_{i=1}^{I} \left(\frac{n_i}{\bar{n}_{\text{int}}} - 1 \right) n_i \bar{w}_i^2}{SSW} - 1 \right) \rho_{\text{int}}. \tag{3}$$

Remark 3.2: We can use formula (1) in any situation. For epsem designs, we have

$$ieff_w - ieff = \rho_{\text{int}} \frac{\bar{n}_{\text{int}}}{n} \sum_{i=1}^{I} \left(\frac{n_i}{\bar{n}_{\text{int}}} - 1 \right) n_i.$$

Note that an application of the Cauchy-Schwarz inequality suggests $ieff_w - ieff \geq 0$ with equality if and only if $n_i = n/I$ for all i.

Remark 3.3: We can use (2) if $SST \neq 0$, *i.e.*, if the design is not epsem. If $\rho_{\text{int}} > 0$, (2) implies

$$ieff_w - ieff \le 0 \quad \text{if and only if} \; \sum_{i=1}^{I}\left(\frac{n_i}{\bar{n}_{\text{int}}}-1\right) n_i \bar{w}_i^2 \le SSW.$$

If high interviewer workload tends to be associated with small average survey weights and vice versa and $SSW \neq 0$, we can expect $ieff$ to be a conservative value of the actual interviewer variance $ieff_w$. In Example 2, c) and d), we have such a situation.

Now, we have $ieff_w = ieff$ if and only if $w_{ik} = \bar{w}_i$ (or, equivalently, $SSW = 0$) and $n_i \bar{w}_i^2 / \sum_i n_i \bar{w}_i^2 = 1/I$ for all i and k, *i.e.*, $ieff_w = ieff$ if and only if $w_{ik} = \bar{w}_i$ and $\bar{w}_i \propto 1/\sqrt{n_i}$ for all i and k.

Thus, for a non-epsem design, equal interviewer workload does not necessarily provide us a model-assisted interpretation for $ieff$. For example, if the survey weights vary within at least one interviewer, we will not have a model-assisted interpretation of $ieff$. Obviously, for an epsem design the two formulae are equivalent if and only if we have equal interviewer workload.

Remark 3.4: If the interviewer workload is the same for all interviewers, we have

$$ieff_w - ieff = -\frac{\bar{n}_{\text{int}}\tau_w}{1 + CV_w^{-2}}\rho_{\text{int}}$$

(assume $SST \neq 0$). Thus, $ieff$ is a conservative value of the actual interviewer effect $ieff_w$. Furthermore, $|ieff_w - ieff|$ is an increasing function of the common interviewer workload $\bar{n}_{\text{int}}$ and $\tau_w/(1 + CV_w^{-2})$ (for fixed CV_w^{-2}, the latter is an increasing function of τ_w). The same interviewer workload is given in Example 2 a).

Remark 3.5: We can use formula (3) if $SSW > 0$, *i.e.*, if there is at least one interviewer for which weights are not all equal.

Example 2.

Table 2 presents eight different combinations of $(n_i, \bar{w}_i, \sigma_i^2)$. The first combination assumes equal n_i values but unequal weights. The second combination assumes $\bar{w}_i^2 \propto \sigma_i^2$. The other six combinations show all possible ordering of $\bar{n}_{\text{int}}$, $\bar{n}_{\text{int}}(\mathbf{a}_1)$, $\bar{n}_w$, $\bar{n}_{\text{int}}(\mathbf{a}_2)$ and, therefore, $ieff$, $ieff(\mathbf{a}_1)$, $ieff_w$, $ieff(\mathbf{a}_2)$ taking into consideration that $ieff \le ieff(\mathbf{a}_1)$ and $ieff_w \le ieff(\mathbf{a}_2)$.

Table 2
Ordering of interviewer effects formulae for several parameter combinations (Example 2); in the last column $\rho_{\text{int}} = 0.01$

	n_i	$\bar{w}_i$	σ_i^2	$\bar{n}_{\text{int}}$	$\bar{n}_{\text{int}}(\mathbf{a}_1)$	$\bar{n}_w$	$\bar{n}_{\text{int}}(\mathbf{a}_2)$	Interviewer effects	$ieff / ieff_w$
a)	25	1.022	0.299	25	25	19.20	25	$ieff = ieff(\mathbf{a}_1) = ieff(\mathbf{a}_2) > ieff_w$	1.003
	25	1.036	0.375						
	25	0.998	0.276						
	25	0.945	0.260						
b)	10	1	1	25	30	15	30	$ieff_w < ieff < ieff(\mathbf{a}_1) = ieff(\mathbf{a}_2)$	1.007
	20	1	1						
	30	1	1						
	40	1	1						
c)	10	1	1	25	30	7.5	32.5	$ieff_w < ieff < ieff(\mathbf{a}_1) < ieff(\mathbf{a}_2)$	1.023
	20	1	2						
	30	1	3						
	40	1	4						
d)	10	1	4	25	30	10	26.7	$ieff_w < ieff < ieff(\mathbf{a}_2) < ieff(\mathbf{a}_1)$	1.015
	20	1	3						
	30	1	2						
	40	1	1						
e)	10	4	144	25	30	1.80	11.71	$ieff_w < ieff(\mathbf{a}_2) < ieff < ieff(\mathbf{a}_1)$	0.998
	20	2	9						
	30	0.333	0.555						
	40	0.250	0.125						
f)	10	0.333	0.025	25	30	31.82	35.26	$ieff < ieff(\mathbf{a}_1) < ieff_w < ieff(\mathbf{a}_2)$	1.015
	20	0.666	0.075						
	30	1	0.125						
	40	1.333	0.175						
g)	10	1	0.010	25	30	29.13	30.10	$ieff < ieff_w < ieff(\mathbf{a}_1) < ieff(\mathbf{a}_2)$	0.999
	20	1	0.020						
	30	1	0.030						
	40	1	0.040						
h)	10	1	0.004	25	30	29.94	29.99	$ieff < ieff_w < ieff(\mathbf{a}_2) < ieff(\mathbf{a}_1)$	0.998
	20	1	0.003						
	30	1	0.002						
	40	1	0.001						

In the example, $\sum_i n_i \bar{w}_i = n$. We now explain the eight different patterns.

a) Since all n_i are equal, $ieff = ieff(\mathbf{a}_1) = ieff(\mathbf{a}_2)$. Moreover, $ieff_w$ is smaller than the rest because of the fact that $\sigma_i^2 > 0$.

b) Since σ_i^2 are relatively large, $ieff_w < ieff$. Also, $\sigma_i^2 = c \cdot \bar{w}_i^2$ implies $ieff(\mathbf{a}_1) = ieff(\mathbf{a}_2)$.

c) Since σ_i^2 are relatively large, $ieff_w < ieff$. Moreover, since $\bar{w}_i^2 + \sigma_i^2$ and n_i are both increasing, we have $ieff(\mathbf{a}_1) < ieff(\mathbf{a}_2)$.

d) Since σ_i^2 are relatively large, $ieff_w < ieff$. Since $\bar{w}_i^2 + \sigma_i^2$ is decreasing and n_i is increasing, we have $ieff(\mathbf{a}_2) < ieff(\mathbf{a}_1)$.

e) Since σ_i^2 are relatively large, $ieff_w < ieff$. Also, $\bar{w}_i^2$ and σ_i^2 are decreasing and n_i is increasing implying $ieff(\mathbf{a}_2) < ieff(\mathbf{a}_1)$.

f) The fact that $\bar{w}_i^2$ and n_i are increasing implies that $ieff_w > ieff$; since σ_i^2 and n_i are both increasing, we have $ieff(\mathbf{a}_1) < ieff(\mathbf{a}_2)$.

g) Since $\bar{w}_i^2$ and n_i are increasing, we have $ieff_w > ieff$ and since σ_i^2 is increasing, we have $ieff(\mathbf{a}_1) < ieff(\mathbf{a}_2)$. Moreover, $ieff_w < ieff(\mathbf{a}_1)$ since σ_i^2 is smaller than that in f).

h) Since $\bar{w}_i^2$ and n_i are increasing, we have $ieff_w > ieff$ and since σ_i^2 is decreasing, we have $ieff(\mathbf{a}_2) < ieff(\mathbf{a}_1)$.

4. Unequal weighting and spatial clustering

In this section, we obtain an appropriate interviewer variance formula in the presence of spatial clustering and unequal probability of selection. Consider the situation when more than one interviewer work independently in the same psu and the respondents in each psu are randomly assigned to the interviewers. We shall assume that no interviewer works in more than one psu. Such a design was considered in Biemer and Stokes (1985). Now we shall separate the interviewer effect from psu effect (*i.e.*, spatial clustering) and unequal weighting. Let y_{pik} and w_{pik} be the observation and the associated survey weight for the k^{th} respondent in the p^{th} psu interviewed by the i^{th} interviewer ($p = 1, ..., P$; $i = 1, ... I_p$; $k = 1, ..., n_{pi}$). Let $n_p = \sum_{i=1}^{I_p} n_{pi}$ be the number of sampling units in psu p.

In this case, we use the following weighted average to estimate the finite population mean:

$$\bar{y}_w = \frac{\sum_{p=1}^{P}\sum_{i=1}^{I_P}\sum_{k=1}^{n_{pi}} w_{pik}\, y_{pik}}{\sum_{p=1}^{P}\sum_{i=1}^{I_P}\sum_{k=1}^{n_{pi}} w_{pik}}.$$

Define

$$ieff_{s,\,w} = \frac{\text{Var}_{M_4}(\bar{y}_w)}{\text{Var}_{M_3}(\bar{y}_w)},$$

where the suffixes s and w signify the presence of spatial clustering and unequal weighting. In the above, $\text{Var}_{M_3}(\bar{y}_w)$ and $\text{Var}_{M_4}(\bar{y}_w)$ are the variances of $\bar{y}_w$ under the following two models respectively

$$M_3\text{: } \text{Cov}(y_{pik},\, y_{p'i'k'}) = \begin{cases} \sigma^2 & \text{if } p = p',\ i = i',\ k = k' \\ \rho_C \sigma^2 & \text{if } p = p',\ k \neq k' \\ 0 & \text{otherwise} \end{cases}$$

$$M_4\text{: } \text{Cov}(y_{pik}, y_{p'i'k'}) = \begin{cases} \sigma^2 & \text{if } p = p',\ i = i',\ k = k' \\ \rho_C \sigma^2 & \text{if } p = p',\ i \neq i' \\ \rho \sigma^2 & \text{if } p = p',\ i = i',\ k \neq k' \\ 0 & \text{if } p \neq p' \end{cases}$$

In the above, ρ_C is the intra-psu correlation and ρ is the combined interviewer and psu intra-class correlation. Define $\rho_{\text{int}} = \rho - \rho_C$, intra-interviewer correlation. Usually, $\rho_{\text{int}} > 0$.

From Result 5, we have

$$ieff_{s,\,w} = 1 + \rho_{\text{int}} \frac{\bar{n}_{\text{int}}(\mathbf{A}_w) - 1}{1 + \rho_C(\bar{n}_{\text{psu}}(\mathbf{b}_w) - 1)},$$

where

$$\mathbf{A}_w = ((a_{wpi}))_{\substack{i=1,\,...,\,I_p \\ p=1,\,...,\,P}} \text{ and } a_{wpi} = \frac{n_{pi}\bar{w}_{pi}^2}{\sum_{p=1}^{P}\sum_{i=1}^{I_p}\sum_{k=1}^{n_{pi}} w_{pik}^2}$$

with

$$\bar{w}_{pi} = \frac{1}{n_{pi}}\sum_{k}^{n_{pi}} w_{pik},$$

$$\bar{n}_{\text{int}}(\mathbf{A}_w) = \sum_{p=1}^{P}\sum_{i}^{I_P} a_{wpi} n_{pi} = \frac{\sum_{p=1}^{P}\sum_{i}^{I_P}\left(\sum_{k}^{n_{pi}} w_{pik}\right)^2}{\sum_{p=1}^{P}\sum_{i=1}^{I_p}\sum_{k=1}^{n_{pi}} w_{pik}^2},$$

and

$$\mathbf{b}_w = (b_{wp})_{p=1,\ldots,P} \quad \text{and} \quad b_{wp} = \frac{n_p \bar{w}_p^2}{\sum_{p=1}^{P}\sum_{i=1}^{I_p}\sum_{k=1}^{n_{pi}} w_{pik}^2},$$

with

$$\bar{w}_p = \frac{1}{n_p}\sum_{i=1}^{I_p}\sum_{k}^{n_{pi}} w_{pik} = \frac{1}{n_p}\sum_{i=1}^{I_p} n_{pi}\bar{w}_{pi},$$

and

$$\bar{n}_{\text{psu}}(\mathbf{b}_w) = \sum_{p=1}^{P} b_{wp} n_p = \frac{\sum_{p=1}^{P}\left(\sum_{i}^{I_P}\sum_{k}^{n_{pi}} w_{pik}\right)^2}{\sum_{p=1}^{P}\sum_{i=1}^{I_p}\sum_{k=1}^{n_{pi}} w_{pik}^2}.$$

Note that $\bar{n}_{\text{int}}(\mathbf{A}_w) \le \bar{n}_{\text{psu}}(\mathbf{b}_w)$ with equality if and only if $I_P = 1$. Also note that $\bar{n}_{\text{int}}(\mathbf{A}_w)$ is invariant of the allocation of the interviewers to the psu's while $\bar{n}_{\text{psu}}(\mathbf{b}_w)$ is not.

Remark 4.1: If $\rho_C = 0$ we get

$$ieff_{s,w} = 1 + \rho_{\text{int}}(\bar{n}_{\text{int}}(\mathbf{A}_w) - 1).$$

This formula is similar to $ieff_w$ given in Section 2. Thus, all the comments given in Remark 2.1 apply here. Note that $\bar{n}_{\text{int}}(\mathbf{A}_w)$, just like $\bar{n}_w$, cannot be generally written in the form $\bar{n}_{\text{int}}(\mathbf{A}_w) = \sum_{p=1}^{P}\sum_{i=1}^{I_p} a_{wpi}n_{pi}$ with $\sum_{p=1}^{P}\sum_{i=1}^{I_p} a_{wpi} = 1$; the same comment applies to $\bar{n}_{\text{psu}}(\mathbf{b}_w)$.

Remark 4.2: Define

$$\bar{n}_{\text{int}}(\mathbf{A}) = \sum_{p=1}^{P}\sum_{i}^{I_P} a_{pi}n_{pi}, \text{ where } \mathbf{A} = ((a_{pi})), \text{ with } a_{pi} = \frac{n_{pi}}{n},$$

and

$$\bar{n}_{\text{psu}}(\mathbf{b}) = \sum_{p=1}^{P} b_p n_p, \text{ where } \mathbf{b} = (b_1, \ldots, b_P) \text{ with } b_p = \frac{n_p}{n}.$$

If $\rho_C \neq 0$ but we have an epsem design, then we drop the suffix w in $ieff_{s,w}$. Note that

$$\begin{aligned} ieff_s &= 1 + \rho_{\text{int}}\frac{\bar{n}_{\text{int}}(\mathbf{A}) - 1}{1 + \rho_C[\bar{n}_{\text{psu}}(\mathbf{b}) - 1]} \\ &= 1 + \frac{\rho_{\text{int}}}{\rho_C}\cdot\frac{\bar{n}_{\text{int}}(\mathbf{A}) - 1}{\bar{n}_{\text{psu}}(\mathbf{b}) - 1}\cdot\frac{\rho_C(\bar{n}_{\text{psu}}(\mathbf{b}) - 1)}{1 + \rho_C(\bar{n}_{\text{psu}}(\mathbf{b}) - 1)} \end{aligned}$$

so that

$$ieff_s < 1 + \frac{\rho_{\text{int}}}{\rho_C}\cdot\frac{\bar{n}_{\text{int}}(\mathbf{A}) - 1}{\bar{n}_{\text{psu}}(\mathbf{b}) - 1} < 1 + \frac{\rho_{\text{int}}}{\rho_C}\cdot\frac{\bar{n}_{\text{int}}(\mathbf{A})}{\bar{n}_{\text{psu}}(\mathbf{b})}.$$

It can be readily seen that the right side of the inequality increases with the ratios ρ_{int}/ρ_C and

$$\frac{\bar{n}_{\text{int}}(\mathbf{A}) - 1}{\bar{n}_{\text{psu}}(\mathbf{b}) - 1}.$$

We have

$$ieff_s - ieff = \rho_{\text{int}}\frac{\dfrac{\bar{n}_{\text{int}}(\mathbf{A}) - 1}{\bar{n}_{\text{int}} - 1} - [1 + \rho_C(\bar{n}_{\text{psu}}(\mathbf{b}) - 1)]}{1 + \rho_C(\bar{n}_{\text{psu}}(\mathbf{b}) - 1)}(\bar{n}_{\text{int}} - 1).$$

Thus, for $\rho_{\text{int}} > 0$,

$ieff_s < ieff$ if and only if

$$Deff_s := 1 + \rho_C(\bar{n}_{\text{psu}}(\mathbf{b}) - 1) > \frac{\bar{n}_{\text{int}}(\mathbf{A}) - 1}{\bar{n}_{\text{int}} - 1},$$

i.e., if and only if the design effect due to the spatial clustering is larger than the ratio of the weighted average of the interviewer workload –1 and the average interviewer workload –1. If the interviewer workload is the same for all the interviewers, the right hand side of the inequality is 1 and so the inequality is always valid. It is interesting to note that $ieff \approx 4 \cdot ieff_s$ if $\rho_{\text{int}} = 0.1$, $\rho_C = 0.05$, $\bar{n}_{\text{psu}}(b) = 140$, and $\bar{n}_{\text{int}} = 70$.

Remark 4.3: In the general case, we have

$$ieff_{s,w} - ieff = \rho_{\text{int}}\left(\frac{\bar{n}_{\text{int}}(\mathbf{A}_w) - 1}{1 + \rho_C(\bar{n}_{\text{psu}}(\mathbf{b}_w) - 1)} - (\bar{n}_{\text{int}} - 1)\right).$$

Thus, for $\rho_{\text{int}} > 0$,

$ieff_{s,w} < ieff$ if and only if

$$Deff_{s,w} := 1 + \rho_C(\bar{n}_{\text{psu}}(\mathbf{b}_w) - 1) > \frac{\bar{n}_{\text{int}}(\mathbf{A}_w) - 1}{\bar{n}_{\text{int}} - 1},$$

i.e., if and only if

$$\rho_C > \frac{\bar{n}_{\text{int}}(\mathbf{A}_w) - \bar{n}_{\text{int}}}{(\bar{n}_{\text{int}} - 1)(\bar{n}_{\text{psu}}(\mathbf{b}_w) - 1)} =: \rho_C^*, \text{ say.}$$

In Example 2 (see Table 3), $ieff$ is a conservative value for $ieff_{s,w}$ for a) to e) if $\rho_C > 0$. The same holds for f) to h) if $\rho_C > 0.004$.

Table 3
Average interviewer workloads for several parameter combinations (Example 2); *ieff* / *ieff*$_{s,w}$ for $\rho_{\text{int}} = 0.01$ and $\rho_C = 0.02$

					$IA = (1,3)$				$IA = (2,2)$				$IA = (3,1)$			
	n_i	$\bar{w}_i$	σ_i^2	$\bar{n}_{\text{int}}$	$\bar{n}_{\text{int}}(A_w)$	$\bar{n}_{\text{psu}}(\mathbf{b}_w)$	ρ_C^*	$\frac{ieff}{ieff_{s,w}}$	$\bar{n}_{\text{int}}(A_w)$	$\bar{n}_{\text{psu}}(\mathbf{b}_w)$	ρ_C^*	$\frac{ieff}{ieff_{s,w}}$	$\bar{n}_{\text{int}}(A_w)$	$\bar{n}_{\text{psu}}(\mathbf{b}_w)$	ρ_C^*	$\frac{ieff}{ieff_{s,w}}$
a)	25	1.022	0.299	25	19.202	47.528	-0.005	1.133	19.202	38.389	-0.006	1.123	19.202	49.039	-0.005	1.135
	25	1.036	0.375													
	25	0.998	0.276													
	25	0.945	0.260													
b)	10	1	1	25	15	41	-0.010	1.151	15	29	-0.015	1.138	15	26	-0.017	1.134
	20	1	1													
	30	1	1													
	40	1	1													
c)	10	1	1	25	7.5	20.5	-0.037	1.185	7.5	14.5	-0.054	1.180	7.5	13	-0.061	1.178
	20	1	2													
	30	1	3													
	40	1	4													
d)	10	1	4	25	10	27.333	-0.024	1.171	10	19.333	-0.034	1.163	10	17.333	-0.038	1.161
	20	1	3													
	30	1	2													
	40	1	1													
e)	10	4	144	25	1.801	2.755	-0.551	1.230	1.801	3.603	-0.371	1.231	1.801	4.344	-0.289	1.231
	20	2	9													
	30	0.333	0.555													
	40	0.250	0.125													
f)	10	0.333	0.025	25	31.820	75.685	0.004	1.104	31.820	58.427	0.005	1.084	31.820	40.629	0.007	1.058
	20	0.666	0.075													
	30	1	0.125													
	40	1.333	0.175													
g)	10	1	0.010	25	29.126	79.612	0.002	1.118	29.126	56.311	0.003	1.094	29.126	50.485	0.003	1.086
	20	1	0.020													
	30	1	0.030													
	40	1	0.040													
h)	10	1	0.004	25	29.940	81.836	0.003	1.117	29.940	57.884	0.004	1.092	29.940	51.896	0.004	1.084
	20	1	0.003													
	30	1	0.002													
	40	1	0.001													

If a household and a person within the household are selected at random, then the weights are often independent of the psu and the interviewer and depend only on the household sizes. In such a situation, the household sizes form the weighting classes. For weighting classes, we define

m_{pij}: number of sampling units in psu p assigned to interviewer i belonging to weighting class j,

$m_{pj} = \sum_{i=1}^{I_p} m_{pij}$: number of sampling units in psu p belonging to weighting class j,

$m_j = \sum_{p=1}^{P} \sum_{i=1}^{I_p} m_{pij}$: number of sampling units belonging to weighting class j.

Thus,

$n_{pi} = \sum_{j=1}^{J} m_{pij}$: number of sampling units in psu p assigned to interviewer i,

$n_p = \sum_{i=1}^{I_p} \sum_{j=1}^{J} m_{pij}$: number of sampling units in psu p,

$n = \sum_{p=1}^{P} \sum_{i=1}^{I_p} \sum_{j=1}^{J} m_{pij}$: sample size.

Furthermore,

$$\bar{n}_{\text{int}}(\mathbf{A}_w) = \frac{\sum_{p=1}^{P} \sum_{i=1}^{I_P} \left(\sum_{k=1}^{n_{pi}} w_{pik} \right)^2}{\sum_{p=1}^{P} \sum_{i=1}^{I_p} \sum_{k=1}^{n_{pi}} w_{pik}^2} = \frac{\sum_{p=1}^{P} \sum_{i=1}^{I_P} \left(\sum_{j=1}^{J} w_j m_{pij} \right)^2}{\sum_{j=1}^{J} w_j^2 m_j}$$

and

$$\bar{n}_{\text{psu}}(\mathbf{b}_w) = \frac{\sum_{p=1}^{P} \left(\sum_{i}^{I_P} \sum_{k}^{n_{pi}} w_{pik} \right)^2}{\sum_{p=1}^{P} \sum_{i=1}^{I_p} \sum_{k=1}^{n_{pi}} w_{pik}^2} = \frac{\sum_{p=1}^{P} \left(\sum_{j} w_j m_{pj} \right)^2}{\sum_{j=1}^{J} w_j^2 m_j},$$

are ratios of quadratic forms in $\mathbf{w} = (w_1, ..., w_J)$.

5. Overall effects

The overall effects take into account unequal weighting, spatial clustering, and the interview effects and can be viewed as a generalization to the traditional design effects. Multiplying the SRS variance for the unweighted sample mean by the overall effects will provide the total variance estimator.

$$eff = \frac{\mathrm{Var}_{M_4}(\bar{y}_w)}{\mathrm{Var}_{M_1^*}(\bar{y})} = eff_w \times eff_s \times eff_{\mathrm{int}},$$

where

$$eff_w = \frac{\mathrm{Var}_{M_1^*}(\bar{y}_w)}{\mathrm{Var}_{M_1^*}(\bar{y})},$$

$$eff_s = \frac{\mathrm{Var}_{M_3}(\bar{y}_w)}{\mathrm{Var}_{M_1^*}(\bar{y}_w)},$$

$$eff_{\mathrm{int}} = ieff_{s,w} = \frac{\mathrm{Var}_{M_4}(\bar{y}_w)}{\mathrm{Var}_{M_3}(\bar{y}_w)}.$$

In the above, $\mathrm{Var}_{M_1^*}$ is with respect to the following model:

$$M_1^*\text{: } \mathrm{Cov}(y_{pik}, y_{p'i'k'}) = \begin{cases} \sigma^2 & \text{if } p = p', i = i', k = k', \\ 0 & \text{otherwise.} \end{cases}$$

It can be shown that

$$eff = \frac{n\sum_{p=1}^{P}\sum_{i=1}^{I_p}\sum_{k=1}^{n_{pi}} w_{pik}^2}{\left(\sum_{p=1}^{P}\sum_{i=1}^{I_p}\sum_{k=1}^{n_{pi}} w_{pik}\right)^2}$$

$$\times\left[1+\rho_C\left(\frac{\sum_{p=1}^{P}\left(\sum_{i=1}^{I_p}\sum_{k=1}^{n_{pi}} w_{pik}\right)^2}{\sum_{p=1}^{P}\sum_{i=1}^{I_p}\sum_{k=1}^{n_{pi}} w_{pik}^2}-1\right)+\rho_{\mathrm{int}}\left(\frac{\sum_{p=1}^{P}\sum_{i}^{I_P}\left(\sum_{k}^{n_{pi}} w_{pik}\right)^2}{\sum_{p=1}^{P}\sum_{i=1}^{I_p}\sum_{k=1}^{n_{pi}} w_{pik}^2}-1\right)\right].$$

The relative contributions of weighting, spatial clustering, and interviewer effects to the overall effects are given by

$$\mathrm{Re}\,eff_w = \frac{\dfrac{n\sum_{p=1}^{P}\sum_{i=1}^{I_p}\sum_{k=1}^{n_{pi}} w_{pik}^2}{\left(\sum_{p=1}^{P}\sum_{i=1}^{I_p}\sum_{k=1}^{n_{pi}} w_{pik}\right)^2}}{eff},$$

$$\mathrm{Re}\,eff_s = \frac{1+\rho_C(\bar{n}_{\mathrm{psu}}(\mathbf{b}_w)-1)}{eff},$$

$$\mathrm{Re}\,eff_I = \frac{1+\rho_{\mathrm{int}}\dfrac{\bar{n}_{\mathrm{psu}}(A_w)-1}{1+\rho_C(\bar{n}_{\mathrm{psu}}(\mathbf{b}_w)-1)}}{eff}.$$

In Figure 2, we present three dimensional graphs of the relative contributions of weighting, spatial clustering, and interviewer effects to the overall effects for different combinations of intra-cluster and intra-interviewer correlations for different patterns of weights given in cases a), f) and h) of Table 3 with $IA = (1, 3)$, where $IA = (a, b)$ indicates that the first a of the four interviewers are in psu 1 and the last b interviewers are in psu 2.

Remark 5.1: From Result 6, we get

$$eff \geq 1 + \rho_C\left(\frac{n}{P}-1\right) + \rho_{\mathrm{int}}\left(\frac{n}{I}-1\right).$$

The right side is the overall effect if the same number of interviewers with equal workload is assigned to each psu. It is interesting to note the similarity between the right hand side of the above inequality and the design effects formula given in (3.1) of Hansen, Hurwitz and Madow (1953, Vol. I, page 370). To claim the similarity, we need to treat the secondary sampling units as the units belonging to an interviewer. In this connection, we also note the formula (3.7) given in Hansen *et al.* (1953, Vol. II, page 292) for the case $I = P$.

Remark 5.2: When we have the same weighting classes across psu × interviewer, we have

$$eff = \frac{n\sum_{j=1}^{J} w_j^2 m_j}{\left(\sum_{j=1}^{J} w_j m_j\right)^2}$$

$$\times\left[1+\rho_C\left(\frac{\sum_{p=1}^{P}\left(\sum_{j=1}^{J} w_j m_{pj}\right)^2}{\sum_{j=1}^{J} w_j^2 m_j}-1\right)+\rho_{\mathrm{int}}\left(\frac{\sum_{p=1}^{P}\sum_{i=1}^{I_P}\left(\sum_{j=1}^{J} w_j m_{pij}\right)^2}{\sum_{j=1}^{J} w_j^2 m_j}-1\right)\right].$$

Remark 5.3: Consider the special case

$$m_{pij} = \frac{n_{pi} m_j}{n}$$

in which we allow variation in weights within psu × interviewer classes, but we constrain the weights to have the same relative frequency distribution in each class, *i.e.*, the means and the variances of the weights within the classes do not depend on the class (Lynn and Gabler 2004). It is easy to see that in this case

$$eff = \frac{n\sum_{j=1}^{J} w_j^2 m_j}{\left(\sum_{j=1}^{J} w_j m_j\right)^2}$$

$$\times\left[1+\rho_C\left(\frac{\left(\sum_{j=1}^{J} w_j m_j\right)^2}{\sum_{j=1}^{J} w_j^2 m_j}\frac{\sum_{p=1}^{P} n_p^2}{n^2}-1\right)\right.$$

$$\left.+\rho_{\mathrm{int}}\left(\frac{\left(\sum_{j=1}^{J} w_j m_j\right)^2}{\sum_{j=1}^{J} w_j^2 m_j}\sum_{p=1}^{P}\sum_{i=1}^{I_P}\frac{n_{pi}^2}{n^2}-1\right)\right].$$

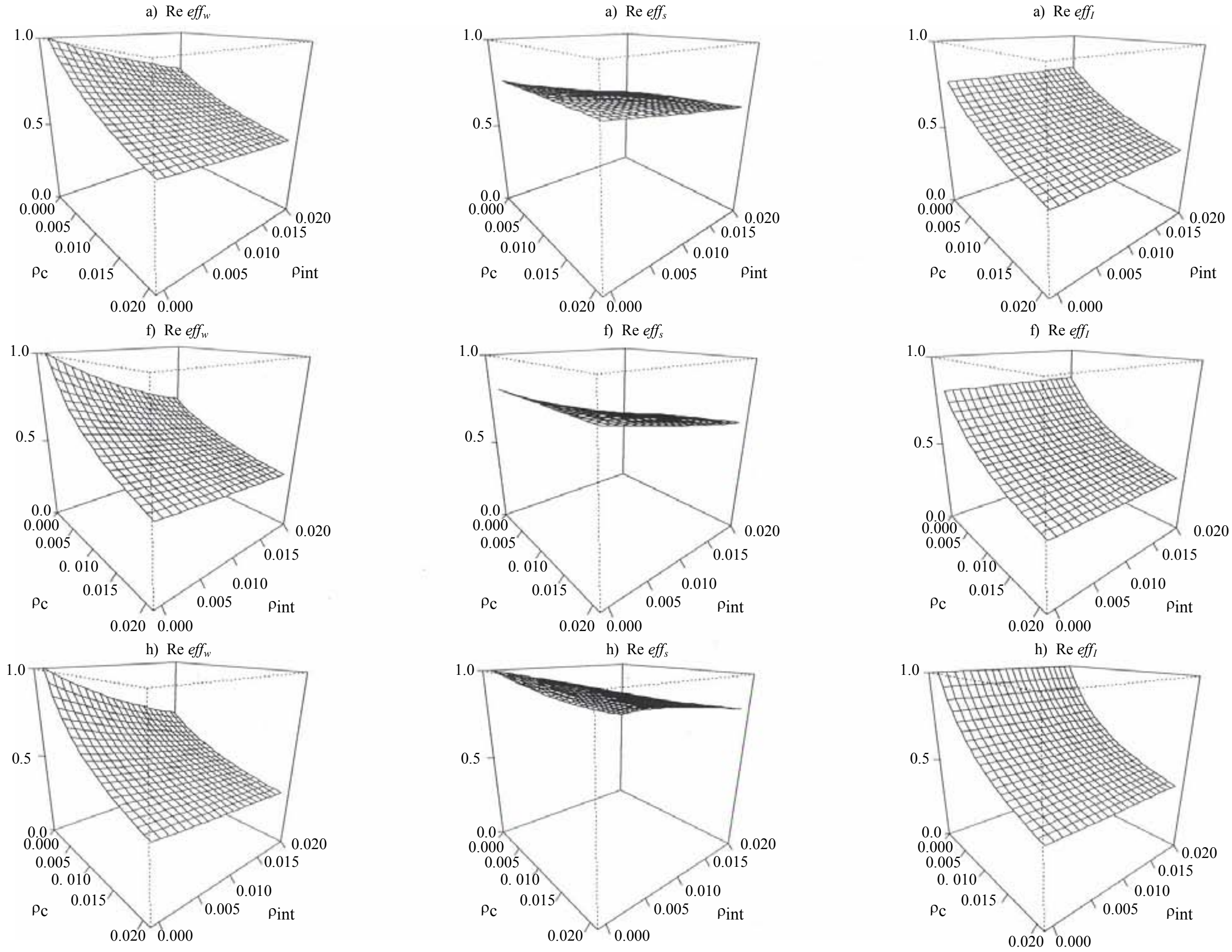


Figure 2 Relative contributions of weighting, design and interviewer effects to the overall effects for cases a), f) and h) in Example 2 for the case *IA* = (1, 3)

Using the same argument given in the proof of Result 6, we get

$$eff \geq 1 + \rho_C \left(\frac{\sum_{p=1}^{P} n_p^2}{n^2} - 1 \right) + \rho_{\text{int}} \left(\sum_{p=1}^{P} \sum_{i=1}^{I_P} \frac{n_{pi}^2}{n} - 1 \right)$$

$$= 1 + \rho_C (\bar{n}_{\text{psu}}(\mathbf{b}) - 1) + \rho_{\text{int}} (\bar{n}_{\text{int}}(\mathbf{A}) - 1).$$

This means that the overall effect is larger than the overall effect for an epsem design (see Remark 5.4).

Remark 5.4: For an epsem design, we have

$$eff = 1 + \rho_C (\bar{n}_{\text{psu}}(\mathbf{b}) - 1) + \rho_{\text{int}} (\bar{n}_{\text{int}}(\mathbf{A}) - 1),$$

where

$$\bar{n}_{\text{psu}}(\mathbf{b}) = \frac{\sum_{p=1}^{P} n_p^2}{n} \text{ and } \bar{n}_{\text{int}}(\mathbf{A}) = \frac{\sum_{p=1}^{P} \sum_{i}^{I_P} n_{pi}^2}{n}.$$

Note that Davis and Scott (1995) obtained this formula for the special case of the following linear mixed model:

$$y_{pik} = \mu + \alpha_i + \beta_p + \varepsilon_{pik},$$

where μ is the overall effect, α_i, β_p are random effects due to the interviewer i, psu p and ε_{pik} is the pure error. They assumed that the random effects are independent with $\alpha_i \sim N(0, \sigma_\alpha^2)$, $\beta_p \sim N(0, \sigma_\beta^2)$ and $\varepsilon_{\text{pik}} \sim N(0, \sigma_\varepsilon^2)$.

For the above linear mixed model, it is easy to check that

$$\rho_{\text{int}} = \frac{\sigma_\alpha^2}{\sigma_\alpha^2 + \sigma_\beta^2 + \sigma_\varepsilon^2} \text{ and } \rho_c = \frac{\sigma_\beta^2}{\sigma_\alpha^2 + \sigma_\beta^2 + \sigma_\varepsilon^2}.$$

However, it is instructive to note that the definition *eff* does not require ρ_{int} and ρ_c to be strictly positive and the definition goes beyond the linear mixed model. For example, the definition applies to the following example:

Example 3: A simple model for binary data.

Assuming $0 < \min(\alpha,\ \beta) < \theta < 1$, we define the following model:

For all n_{pi} different respondents of interviewer i in psu p.

$P(Y_{pik} = x_1, Y_{pik'} = x_2)$

x_2 x_1	1	0	Total
1	α	$\theta - \alpha$	θ
0	$\theta - \alpha$	$1 - 2\theta + \alpha$	$1 - \theta$
Total	θ	$1 - \theta$	1

For all n_{pi} respondents of interviewer i and psu p and all $n_{pi'}$ respondents of interviewer i' and psu p.

$P(Y_{pik} = x_1, Y_{pi'k'} = x_2)$

x_2 x_1	1	0	Total
1	β	$\theta - \alpha$	θ
0	$\theta - \alpha$	$1 - 2\theta + \beta$	$1 - \theta$
Total	θ	$1 - \theta$	1

For all n_p respondents of psu p and all $n_{p'}$ respondents of psu p'.

$P(Y_{pik} = x_1, Y_{p'i'k'} = x_2)$

x_2 x_1	1	0	Total
1	θ^2	$\theta(1 - \theta)$	θ
0	$\theta(1 - \theta)$	$(1 - \theta)^2$	$1 - \theta$
Total	θ	$1 - \theta$	1

Therefore, we have

$$E(Y_{pik}) = \theta \text{ for all } p,\ i,\ k,$$

$$\operatorname{Var}(Y_{pik}) = \theta(1 - \theta) \text{ for all } p,\ i,\ k,$$

$$\rho = \frac{\operatorname{Cov}(Y_{pik}, Y_{pik'})}{\sqrt{\operatorname{Var}(Y_{pik})\operatorname{Var}(Y_{pik'})}}$$

$$= \frac{\alpha - \theta^2}{\theta(1 - \theta)} \text{ for all } p,\ i \text{ and } k \neq k',$$

$$\rho_C = \frac{\operatorname{Cov}(Y_{pik}, Y_{pi'k'})}{\sqrt{\operatorname{Var}(Y_{pik})\operatorname{Var}(Y_{pi'k'})}}$$

$$= \frac{\beta - \theta^2}{\theta(1 - \theta)} \text{ for all } p \text{ and } i \neq i',$$

which is a special case of Model M_4 with $\sigma^2 = \operatorname{Var}(Y_{pik}) = \theta(1 - \theta)$. Note that both ρ_C and ρ may be negative and $\rho_{\text{int}} = \rho - \rho_C$ is positive if and only if $\alpha > \beta$.

Remark 5.5: For an epsem design with common psu size $b = n / P,$ we have

$$\mathit{eff} = 1 + \rho_C(b - 1) + \rho_{\text{int}}(\overline{n}_{\text{int}}(A) - 1).$$

Remark 5.6: In discussing Verma *et al.* (1980), Holt considered the case when there is no interviewer variability and psu is the weighting class, *i.e.*, the case when $\rho_{\text{int}} = 0$ and $w_{pik} = w_p$ for all $p,\ i,\ k.$ In this case eff reduces to

$$\mathit{eff} = \frac{n\sum_{p=1}^{P} n_p w_p^2}{\left(\sum_{p=1}^{P} n_p w_p\right)^2} \times \left[1 + \rho_C\left(\frac{\sum_{p=1}^{P} n_p^2 w_p^2}{\sum_{p=1}^{P} n_p w_p^2} - 1\right)\right].$$

Note that the above formula can be obtained from equation (A4) of Holt in discussing Verma *et al.* (1980), after correcting an obvious typo (*i.e.*, deleting n in the denominator), choosing his choice of survey weight and some algebra. Design effect formulae in the absence of the interviewer effects were considered by many authors. See Kish (1965), Verma *et al.* (1980), Skinner (1986), Valliant (1987), Skinner *et al.* (1989), Gabler, Häder and Lahiri (1999), Lynn and Gabler (2004), Kalton, Brick and Lê (2005) and others.

6. Concluding remarks

We have noticed that the standard interviewer effects formula could have either an overestimation or underestimation problem depending on the situation. For example, it could severely underestimate the interviewer effects in an epsem sampling design with different interviewer workloads. Interestingly, spatial correlation can turn this underestimation to an overestimation. In the former case, the survey designer who uses the standard interviewer effect formula may pay little attention to control the interviewer effect. In the latter case, a high value of the interviewer effect may unnecessarily raise concerns about the quality of data connected with the interviewer. This may trigger allocation of a higher portion of budget than is necessary to reduce the interviewer effect, which may be already much lower than the value obtained by an application of the standard formula. The paper is an attempt to define and interpret interviewer effects that are appropriate in different complex survey situations.

We have considered the case when an interviewer is assigned only in one psu. The case when an interviewer works in different psu's is also important and will be considered in a later paper. The weights used in the proposed formulae only account for sampling weights as they are planned at the design stage, but do not necessarily reflect the actual weights attached to each case once the data are collected. In other words, our interviewer effect formulae do not incorporate the effects due to nonresponse and post-stratification adjustments. The formulae presented in the paper are mainly useful in the planning and design stage when we have some ideas about the intra-interviewer and spatial correlations.

Reliable estimation of ρ_{int} and ρ_{c} is important. Although there are some papers that deal with the estimation of ρ_{int} and ρ_{c}, there is certainly a need to advance research in this important area. In comparing the two sources of homogeneity, Hansen *et al.* (1961) found that the interviewer variability was often larger than the sampling variability. In many surveys, such an evaluation, which requires estimation of the intra-interviewer and intra-cluster correlations, is either difficult or even impossible because the interviewer effects are often confounded with the spatial clustering effects. The use of an interpenetrating design, first proposed by Mahalanobis (1946), where respondents are randomly assigned to the interviewers, is a way to get around the problem. In practice, the implementation of such a design in a large scale sample survey is difficult, but some approximated interpenetrated designs can be applied (Hansen *et al.* 1961, Bailar, Bailey and Stevens 1977, Bailey, Moore and Bailar 1978, Collins and Butcher 1982, O'Muircheartaigh and Campanelli 1998). Multi-level models have been used as a partial remedy to the problem (Hox and De Leeuw 1994, Davis and Scott 1995, O'Muircheartaigh and Campanelli 1998, Scott and Davis 2001). We have not considered the problem of the estimation of the intra-interviewer and intra-cluster correlations. This is an important problem and will be considered in a later paper.

In practice, interviewer or design effects are computed for many items using the same formula and a summary measure such as the median interviewer or design effect is taken for the planning and design of the survey. So far as the issues related to handling multiple items are concerned, one may continue to follow one's own protocol; the only change we may suggest is to use our new definitions for interviewer effects or overall effects whenever applicable. The use of our formula may suggest overall effects, which may be much lower than the standard formula. This, in turn, may suggest lower sample size and hence may save survey costs.

Appendix

Result 1. $ieff_w = \dfrac{\text{Var}_{M_2}(\bar{y}_w)}{\text{Var}_{M_1}(\bar{y}_w)} = 1 + \rho_{\text{int}} \left(\dfrac{\sum_i \left(\sum_k w_{ik} \right)^2}{\sum_i \sum_k w_{ik}^2} - 1 \right).$

Proof: The result follows by noting

$$\text{Var}_{M_1}(\bar{y}_w) = \text{Var}_{M_1} \left[\frac{\sum_i \sum_k w_{ik} y_{ik}}{\sum_i \sum_k w_{ik}} \right] = \frac{\sigma^2 \sum_i \sum_k w_{ik}^2}{\left(\sum_i \sum_k w_{ik} \right)^2},$$

and

$$\text{Var}_{M_2}(\bar{y}_w) = \frac{\sigma^2 \left[\sum_i \sum_k w_{ik}^2 + \rho_{\text{int}} \sum_i \sum_{k \neq k'} w_{ik} w_{ik'} \right]}{\left(\sum_i \sum_k w_{ik} \right)^2},$$

and some algebra.

Corollary: Assume $\rho_{\text{int}} > 0$ and $w_{ik} = 1/n.$ Using Result 1 and the Cauchy-Schwarz inequality, we get

$$ieff(\mathbf{a}_1) = 1 + \rho_{\text{int}} \left(\frac{\sum_i n_i^2}{n} - 1 \right) \geq 1 + \rho_{\text{int}} \left(\frac{n}{I} - 1 \right) = ieff.$$

Result 2. $ieff_w \leq ieff(\mathbf{a}_2),$ where

$$\mathbf{a}_2 = (a_{21}, \ldots, a_{2I}) \text{ with } a_{2i} = \frac{\sum_k w_{ik}^2}{\sum_i \sum_k w_{ik}^2}.$$

Proof: Using the Cauchy-Schwarz inequality, we have

$$\sum_i \left(\sum_k w_{ik} \right)^2 \leq \sum_i n_i \sum_k w_{ik}^2$$

with equality if and only if $w_{ik} = \bar{w}_i$ for all i and $k,$ where

$$\bar{w}_i = \frac{\sum_{k=1}^{n_i} w_{ik}}{n_i}$$

is the average survey weight for the i^{th} interviewer. Thus, we have $ieff_w \leq 1 + [\bar{n}_{\text{int}}(\mathbf{a}_2) - 1]\rho_{\text{int}} = ieff(\mathbf{a}_2).$

The equality holds if and only if $w_{ik} = \bar{w}_i$ for all i and k in which case $ieff_w = ieff(\mathbf{a}_2^*),$ where

$$\mathbf{a}_2^* = (a_{21}^*, \ldots, a_{2I}^*), \text{ with } a_{2i}^* = \frac{n_i \bar{w}_i^2}{\sum_i n_i \bar{w}_i^2}.$$

If all weights are non-negative, then

$$\sigma_i^2 = \frac{1}{n_i}\sum_k (w_{ik} - \overline{w}_i)^2 \leq (n_i - 1)\,\overline{w}_i^2,$$

since σ_i^2 is Schur-convex. Defining

$$x_i = \frac{1 + \dfrac{\sigma_i^2}{\overline{w}_i^2}}{n_i} \text{ implies } \frac{1}{n_i} \leq x_i \leq 1$$

and

$$\begin{aligned}
\overline{n}_{\text{int}}(\mathbf{a}_2) &= \frac{\sum_i n_i \sum_k w_{ik}^2}{\sum_i\sum_k w_{ik}^2} = \frac{\sum_i n_i^2\overline{w}_i^2 + \sum_i n_i^2\sigma_i^2}{\sum_i n_i\,\overline{w}_i^2 + \sum_i n_i\,\sigma_i^2} \\
&= \frac{\sum_i n_i^3\overline{w}_i^2 - \sum_i n_i^2((n_i - 1)\,\overline{w}_i^2 - \sigma_i^2)}{\sum_i n_i^2\overline{w}_i^2 - \sum_i n_i\,((n_i - 1)\,\overline{w}_i^2 - \sigma_i^2)} \\
&= \frac{\sum_i n_i^3\overline{w}_i^2 x_i}{\sum_i n_i^2\overline{w}_i^2 x_i} \leq \frac{\sum_i n_i^3\overline{w}_i^2}{\sum_i n_i^2\overline{w}_i^2} = \sum_i n_i\,\frac{n_i^2\overline{w}_i^2}{\sum_i n_i^2\overline{w}_i^2}
\end{aligned}$$

with equality if and only if $\sigma_i^2 = (n_i - 1)\,\overline{w}_i^2$ for all i or if all n_i are equal.

The inequality follows from the logarithmic concavity of $\overline{n}_{\text{int}}(\mathbf{a}_2)$ as function of $(x_1, \ldots, x_I)$.

Result 3. For $\mathbf{a}_2^* = (a_{21}^*, \ldots, a_{2I}^*)$ with $a_{2i}^* = \dfrac{n_i\overline{w}_i^2}{\sum_i n_i\overline{w}_i^2}$

and

$$\mathbf{a}_2 = (a_{21}, \ldots, a_{2I}) \text{ with } a_{2i} = \frac{\sum_k w_{ik}^2}{\sum_i\sum_k w_{ik}^2},$$

we have

$$ieff(\mathbf{a}_2^*) \overset{\leq}{\underset{\geq}{\;}} ieff(\mathbf{a}_2) \quad \text{if and only if}$$

$$\sum_i n_i\,\sigma_i^2 \sum_i n_i^2\overline{w}_i^2 \overset{\leq}{\underset{\geq}{\;}} \sum_i n_i^2\sigma_i^2 \sum_i n_i\,\overline{w}_i^2.$$

Proof. We have

$$ieff(\mathbf{a}_2^*) - ieff(\mathbf{a}_2) = \frac{\sum_i n_i\,\sigma_i^2\sum_i n_i^2\overline{w}_i^2 - \sum_i n_i^2\sigma_i^2\sum_i n_i\,\overline{w}_i^2}{\left(\sum_i n_i\,\sigma_i^2 + \sum_i n_i\,\overline{w}_i^2\right)\sum_i n_i\,\overline{w}_i^2}.$$

For $n_i = n/I$ for all i, we get

$$ieff(\mathbf{a}_2^*) = ieff(\mathbf{a}_2).$$

For $w_{ik} = \overline{w}_i$ for all i, *i.e.*, $\sigma_i^2 = 0$, we get

$$ieff(\mathbf{a}_2^*) = ieff(\mathbf{a}_2).$$

For $\overline{w}_i = \overline{w}$ for all i and $\sigma_i^2 = \sigma^2 > 0$ for all i, we get

$$ieff(\mathbf{a}_2^*) = ieff(\mathbf{a}_2).$$

For $\overline{w}_i = \overline{w}$ for all i, we get

$$ieff(\mathbf{a}_2^*) \overset{\leq}{\underset{\geq}{\;}} ieff(\mathbf{a}_2) \text{ iff } \sum_i n_i\,\sigma_i^2 \sum_i n_i^2 \overset{\leq}{\underset{\geq}{\;}} n\sum_i n_i^2\sigma_i^2.$$

For $\sigma_i^2 = \sigma^2 > 0$ for all i, we get

$$ieff(\mathbf{a}_2^*) \overset{\leq}{\underset{\geq}{\;}} ieff(\mathbf{a}_2) \text{ iff } n\sum_i n_i^2\overline{w}_i^2 \overset{\leq}{\underset{\geq}{\;}} \sum_i n_i\,\overline{w}_i^2 \sum_i n_i^2.$$

Result 4. We have

$$\begin{aligned}
ieff_w - ieff &= \frac{\overline{n}_{\text{int}}}{SST + n\overline{w}^2}\left[\sum_{i=1}^{I}\left(\frac{n_i}{\overline{n}_{\text{int}}} - 1\right)n_i\overline{w}_i^2 - SSW\right]\rho_{\text{int}} \\
&= \frac{\overline{n}_{\text{int}}}{(1 + CV_w^{-2})SST}\left[\sum_{i=1}^{I}\left(\frac{n_i}{\overline{n}_{\text{int}}} - 1\right)n_i\overline{w}_i^2 - SSW\right]\rho_{\text{int}} \\
&= \frac{\overline{n}_{\text{int}}\,\tau_w}{1 + CV_w^{-2}}\left(\frac{\sum_{i=1}^{I}\left(\frac{n_i}{\overline{n}_{\text{int}}} - 1\right)n_i\overline{w}_i^2}{SSW} - 1\right)\rho_{\text{int}}.
\end{aligned}$$

Proof.

$$ieff_w - ieff$$

$$= 1 + \left(\frac{\sum_i\left(\sum_k w_{ik}\right)^2}{\sum_i\sum_k w_{ik}^2} - 1\right)\rho_{\text{int}} - 1 - (\overline{n}_{\text{int}} - 1)\,\rho_{\text{int}}$$

$$= \left(\frac{\sum_i\left(\sum_k w_{ik}\right)^2}{\sum_i\sum_k w_{ik}^2} - \overline{n}_{\text{int}}\right)\rho_{\text{int}}$$

$$= \left(\frac{\sum_i n_i^2\overline{w}_i^2}{SST + n\overline{w}^2} - \overline{n}_{\text{int}}\right)\rho_{\text{int}}$$

$$= \frac{\overline{n}_{\text{int}}}{SST + n\overline{w}^2}\left(\frac{\sum_i n_i^2\overline{w}_i^2}{\overline{n}_{\text{int}}} - (SST + n\overline{w}^2)\right)\rho_{\text{int}}$$

$$= \frac{\overline{n}_{\text{int}}}{SST + n\overline{w}^2}\left(\sum_{i=1}^{I}\left(\frac{n_i}{\overline{n}_{\text{int}}} - 1\right)n_i\overline{w}_i^2 + \sum_{i=1}^{I} n_i\overline{w}_i^2 - (SST + n\overline{w}^2)\right)\rho_{\text{int}}.$$

Now the result follows using algebra.

Result 5.

$$ieff_{s,w} = \frac{\mathrm{Var}_{M_4}(\bar{y}_w)}{\mathrm{Var}_{M_3}(\bar{y}_w)}$$

$$= 1 + \rho_{\mathrm{int}} \frac{\dfrac{\sum_{p=1}^{P}\sum_{i}^{I_P}\left(\sum_{k}^{n_{pi}} w_{pik}\right)^2}{\sum_{p=1}^{P}\sum_{i=1}^{I_p}\sum_{k=1}^{n_{pi}} w_{pik}^2} - 1}{1 + \rho_C\left(\dfrac{\sum_{p=1}^{P}\left(\sum_{i}^{I_P}\sum_{k}^{n_{pi}} w_{pik}\right)^2}{\sum_{p=1}^{P}\sum_{i=1}^{I_p}\sum_{k=1}^{n_{pi}} w_{pik}^2} - 1\right)}.$$

Proof. The result follows by noting that

$$\frac{\mathrm{Var}_{M_4}(\bar{y}_w)}{\mathrm{Var}_{M_3}(\bar{y}_w)} =$$

$$\frac{\sum_{p=1}^{P}\sum_{i=1}^{I_P}\sum_{k=1}^{n_{pi}} w_{pik}^2 + \rho_C \sum_{p=1}^{P}\sum_{i\neq i'}\sum_{k}^{n_{pi}}\sum_{k'}^{n_{pi'}} w_{pik}w_{pi'k'} + \rho\sum_{p=1}^{P}\sum_{i}^{I_P}\sum_{k\neq k'}^{n_{pi}} w_{pik}w_{pik'}}{\sum_{p=1}^{P}\left(\sum_{i=1}^{I_p}\sum_{k=1}^{n_{pi}} w_{pik}^2 + \rho_C\sum_{i,i'}^{I_p}\sum_{k\neq k'} w_{pik}w_{pi'k'}\right)}$$

and some algebra.

Result 6. For $0 < \rho_C < 1$ and $0 < \rho_{\mathrm{int}} < 1$,

$$eff \geq 1 + \rho_C\left(\frac{n}{P} - 1\right) + \rho_{\mathrm{int}}\left(\frac{n}{I} - 1\right),$$

with equality if and only if the weights are all equal and each interviewer has the same workload.

If we have in each psu only one interviewer, then

$$eff \geq 1 + (\rho_C + \rho_{\mathrm{int}})\left(\frac{n}{P} - 1\right).$$

Proof: Using some algebra and the general inequality,

$$\sum_{j}^{J} p_j x_j^2 \geq \left(\sum_{j}^{J} p_j x_j\right)^2$$

with

$$p_j \geq 0 \text{ and } \sum_{j=1}^{J} p_j = 1,$$

we have

$$eff = \frac{n\sum_{p=1}^{P}\sum_{i=1}^{I_p}\sum_{k=1}^{n_{pi}} w_{pik}^2}{\left(\sum_{p=1}^{P}\sum_{i=1}^{I_p}\sum_{k=1}^{n_{pi}} w_{pik}\right)^2}(1 - \rho_C - \rho_{\mathrm{int}})$$

$$+ n\rho_C \frac{\sum_{p=1}^{P}\left(\sum_{i=1}^{I_p}\sum_{k=1}^{n_{pi}} w_{pik}\right)^2}{\left(\sum_{p=1}^{P}\sum_{i=1}^{I_p}\sum_{k=1}^{n_{pi}} w_{pik}\right)^2} + n\rho_{\mathrm{int}} \frac{\sum_{p=1}^{P}\sum_{i}^{I_P}\left(\sum_{k}^{n_{pi}} w_{pik}\right)^2}{\left(\sum_{p=1}^{P}\sum_{i=1}^{I_p}\sum_{k=1}^{n_{pi}} w_{pik}\right)^2}$$

$$\geq 1 - \rho_C - \rho_{\mathrm{int}} + \rho_C\frac{n}{P} + \rho_{\mathrm{int}} n \frac{I\sum_{p=1}^{P}\frac{I_p}{I}\left(\frac{1}{I_p}\sum_{i}^{I_P}\sum_{k}^{n_{pi}} w_{pik}\right)^2}{\left(\sum_{p=1}^{P}\sum_{i=1}^{I_p}\sum_{k=1}^{n_{pi}} w_{pik}\right)^2}$$

$$\geq 1 - \rho_C - \rho_{\mathrm{int}} + \rho_C\frac{n}{P} + \rho_{\mathrm{int}}\frac{n}{I}$$

$$= 1 + \rho_C\left(\frac{n}{P} - 1\right) + \rho_{\mathrm{int}}\left(\frac{n}{I} - 1\right).$$

Acknowledgements

We thank the editor and referees for constructive comments and suggestions which have improved the original version of this paper.

References

Bailar, B.A., Bailey, L. and Stevens, J. (1977). Measures of interviewer bias and variance. *Journal of Marketing Research*, 14, 337-343.

Bailey, L., Moore, T.F. and Bailar, B.A. (1978). An interviewer variance study for eight impact cities of the National Crime Survey Cities Sample. *Journal of the American Statistical Association*, 73, 16-23.

Biemer, P.P., and Stokes, S.L. (1985). Optimal design of interviewer variance experiments in complex surveys. *Journal of the American Statistical Association*, 80, 369, 158-166.

Biemer, P., and Trewin, D. (1997). A review of measurement error effects on the analysis of survey data. 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., 603-632.

Brewer, K.R.W. (1963). Ratio estimation and finite populations: Some results deductible from the assumption of an underlying stochastic process, *Australian Journal of Statistics*, 5, 93-105.

Chambers, R.L., and Skinner, C.J. (Eds.) (2003). *Analysis of Survey Data*. Wiley, Chichester.

Collins, M., and Butcher, B. (1982). Interviewer and clustering effects in an attitude survey. *Journal of the Market Research Society*, 25, 39-58.

Davis, P.D., and Scott, A.J. (1995). The effect of interviewer variance on domain comparisons. *Survey Methodology*, 21, 99-106.

Feather, J. (1973). A study of interviewer variance. WHO International Collaborative Study of Medical Care Utilization, Saskatchewan Study Area Reports, Series II, Monograph No. 3.

Fellegi, I.P. (1964). Response variance and its estimation. *Journal of the American Statistical Association*, 59, 1016-1041.

Gabler, S., Häder, S. and Lahiri, P. (1999). A model based justification of Kish's formula for design effects for weighting and clustering. *Survey Methodology*, 25, 105-106.

Gray, P.G. (1956). Examples of interviewer variability taken from two sample surveys. *Applied Statistics*, V, 73-85.

Groves, R.M. (1989). *Survey Errors and Survey Costs*. New York: John Wiley & Sons, Inc.

Groves, R.M., and Magilavy, L.J. (1986). Measuring and explaining interviewer effects in centralized telephone surveys. *Public Opinion Quarterly*, 50, 251-266.

Hansen, M.H., Hurwitz, W.N. and Madow, W.G. (1953). *Sample Survey Methods and Theory*. Vol I, II. New York: John Wiley & Sons, Inc.

Hansen, M.H., Hurwitz, W.N. and Bershad, M.A. (1961). Measurement errors in census and surveys. *Bulletin of the ISI 38*, 2, 351-374.

Hanson, R.H., and Marks, E.S. (1958). Influence of the interviewer on the accuracy of survey results. *Journal of the American Statistical Association*, 53, 635-655.

Hájek, J. (1971). Comments, In Foundations of Statistical Inference, (Eds. V.P. Godambe and D.A. Sprott). Toronto: Holt, Rinchart, and Winston.

Heeb, J.-L., and Gmel, G. (2001). Interviewers and respondents effects on self-reported alcohol consumption in Swiss Health Survey. *Journal of Studies on Alcohol*, 62, 434-442.

Hox, J.J., and De Leeuw, E.D. (1994). A comparison of nonresponse in mail, telephone, and face-to-face surveys. Applying multilevel modeling to meta-analysis. *Quality & Quantity*, 329-344.

Kalton, G., Brick, J.M. and Lê, T.h. (2005). Estimating components of design effects for use in sample design. In: *Household Sample Surveys in Developing and Transition Countries*, Chapter VI. Available from http://unstats.un.org/unsd/hhsurveys/pdf/Chapter_6.pdf.

Kish, L. (1962). Studies of interviewer variance for attitudinal variables. *Journal of the American Statistical Association*, 57, 92-115.

Kish, L. (1965). *Survey Sampling*. New York: John Wiley & Sons, Inc.

Lynn, P., and Gabler, S. (2004). Approximations to *b** in the estimation of design effects due to clustering. *Working Papers of the Institute for Social and Economic Research*, paper 2004-07. Colchester: University of Essex. Available from http://www.iser.essex.ac.uk/pubs/workpaps/pdf/2004-07.pdf.

Mahalanobis, P.C. (1946). Recent experiments in statistical sampling in the Indian Statistical Institute. *Journal of the Royal Statistical Society*, Serie A, 109, 325-378, reprinted in *Sankhyā* (1958), 1-68.

O'Muircheartaigh, C., and Campanelli, P. (1998). The relative impact of interviewer effects and sample design effects on survey precision. *Journal of the Royal Statistical Society*, Series A, 161, 63-77.

Philippens, M., and Loosveldt, G. (2004). Interviewer-related variance in the European Social Survey. Paper presented at the sixth international conference on social science methodology, August 17-20 in Amsterdam.

Rice, S.A. (1929). Contagious bias in the interview: A methodological note. *American Journal of Sociology*, 35, 420-423.

Rao, J.N.K., and Scott, A.J. (1984). On chi-squared tests for multi-way contingency tables with proportions estimated from survey data. *Annals of Statistics*, 12, 46-60.

Schnell, R., and Kreuter, F. (2005). Separating interviewer and sampling-point effects. *Journal of Official Statistics*, 21, 389-410.

Scott, A.J., and Davis, P.D. (2001). Estimating interviewer effects for binary responses. *Proceedings*: *Symposium 2001*, Achieving Data Quality in a Statistical Agency.

Skinner, C.J. (1986). Design effect of two-stage sampling, *Journal of the Royal Statistical Society*, Serie B, 48, 89-99.

Skinner, C.J., Holt, D. and Smith, T.M.F. (eds.) (1989). *Analysis of Complex Surveys*. New York: John Wiley & Sons, Inc.

Tucker, C. (1983). Interviewer effects in telephone surveys. *Public Opinion Quarterly*, 47, 84-95.

Valliant, R.M. (1987). Generalized variance functions in stratified two-stage sampling, 82, 499-508.

Verma, V., Scott, C. and O'Muircheartaigh, C. (1980). Sample designs and sampling errors for the World Fertility Survey. *Journal of the Royal Statistical Society*, Serie A, 143, 431-473.

Survey Methodology, June 2009
Vol. 35, No. 1, pp. 101-113
Statistics Canada, Catalogue No. 12-001-X

Indicators for the representativeness of survey response

Barry Schouten, Fannie Cobben and Jelke Bethlehem [1]

Abstract

Many survey organisations focus on the response rate as being the quality indicator for the impact of non-response bias. As a consequence, they implement a variety of measures to reduce non-response or to maintain response at some acceptable level. However, response rates alone are not good indicators of non-response bias. In general, higher response rates do not imply smaller non-response bias. The literature gives many examples of this (*e.g.*, Groves and Peytcheva 2006, Keeter, Miller, Kohut, Groves and Presser 2000, Schouten 2004).

We introduce a number of concepts and an indicator to assess the similarity between the response and the sample of a survey. Such quality indicators, which we call R-indicators, may serve as counterparts to survey response rates and are primarily directed at evaluating the non-response bias. These indicators may facilitate analysis of survey response over time, between various fieldwork strategies or data collection modes. We apply the R-indicators to two practical examples.

Key Words: Quality; Non-response; Non-response reduction; Non-response adjustment.

1. Introduction

It is a well-developed finding in the survey methodological literature that response rates by themselves are poor indicators of non-response bias, see *e.g.*, Curtin, Presser and Singer (2000), Groves, Presser and Dipko (2004), Groves (2006), Groves and Peytcheva (2006), Keeter *et al.* (2000), Merkle and Edelman (2002), Heerwegh, Abts and Loosveldt (2007) and Schouten (2004). However, the field has yet to propose alternative indicators of non-response that may be less ambiguous as indicators of survey quality.

We propose an indicator, which we call an R-indicator (‘R’ for representativeness), for the similarity between the response to a survey and the sample or the population under investigation. This similarity can be referred to as “representative response”. In the literature, there are many different interpretations of the ‘representativeness’ concept. See Kruskal and Mosteller (1979a, b and c) for a thorough investigation of the statistical and non-statistical literature. Rubin (1976) introduced the concept of ignorable non-response; the minimal conditions that allow for unbiased estimation of a statistic. Some authors explicitly define representativeness. Hájek (1981) links “representative” to the estimation of population parameters; the pair formed by an estimator and a missing-data mechanism are representative when, with probability one, the estimator is equal to the population parameter. Following Hajèk’s definition, calibration estimators (*e.g.*, Särndal, Swensson and Wretman 2003) are representative for the auxiliary variables that are calibrated. Bertino (2006) defines a so-called univariate representativeness index for continuous random variables. This index is a distribution-free measure based on the Cramér – Von Mises statistic. Kohler (2007) defines what he calls an internal criterion for representativeness. His univariate criterion resembles the Z-statistic for population means.

We separate the concept of representativeness from the estimation of a specific population parameter but relate this concept to the impact on the overall composition of response. By separating indicators from a specific parameter, they can be used as tools for comparing different surveys and surveys over time, and for a comparison of different data collection strategies and modes. Also, the measure gives a multivariate perspective of the dissimilarity between sample and response.

The R-indicator that we propose employs estimated response probabilities. The estimation of response probabilities implies that the R-indicator itself is a random variable, and, consequently, has a precision and possibly a bias. The sample size of a survey, therefore, plays an important role in the assessment of the R-indicator as we will show. However, this dependence exists for any measure; small surveys simply do not allow for strong conclusions about the missing-data mechanism.

We show that the proposed R-indicator relates to Cramèr’s V measure for the association between response and auxiliary variables. In fact, we view the R-indicator as a lack-of-association measure. The weaker the association the better, as this implies there is no evidence that non-response has affected the composition of the observed data.

In order to be able to use R-indicators as tools for monitoring and comparing survey quality in the future, they need to have the features of a measure. That is, we want an R-indicator to be interpretable, measurable, able to be normalized and also to satisfy the mathematical properties of a measure. Especially since the interpretation and normalization are not straightforward features.

[1] Barry Schouten, Fannie Cobben and Jelke Bethlehem, Statistics Netherlands, Department of Methodology and Quality, PO Box 4000, 2370 JM Voorburg, The Netherlands. E-mail: bstn@cbs.nl.

We apply the R-indicator to two studies that were conducted at Statistics Netherlands in 2005 and 2006. The objectives of those studies were the comparison of different data collection strategies. The studies involved different data collection modes and different non-response follow-up strategies. For each of the studies, a detailed analysis was done and documented. These studies are, therefore, suited to an empirical validation of the R-indicator. We compare the values of the R-indicator to the conclusions in the analyses. We refer to Schouten and Cobben (2007) and Cobben and Schouten (2007) for more illustrations and empirical investigations.

In section 2, we start with a discussion of the concept of representative response. Next, in section 3, we define the mathematical notation for our R-indicator. Section 4 is devoted to the features of the R-indicator. Section 5 describes the application of the R-indicator to the field studies. Finally, section 6 contains a discussion.

2. The concept of representative response

We, first, discuss what it means when a survey respondent pool is representative of the sample. Next, we make the concept of representativeness mathematically rigorous by giving it a definition.

2.1 What does representative mean?

Literature warns us not to single-mindedly focus on response rates as an indicator of survey quality. This can easily be illustrated by an example from the 1998 Dutch survey POLS (short for Permanent Onderzoek Leefsituatie or Integrated Survey on Household Living Conditions in English).

Table 1 contains the one- and two-month POLS survey estimates for the proportion of the Dutch population that receives a form of social allowance and the proportion that has at least one parent that was born outside the Netherlands. Both variables are taken from registry data and are artificially treated as survey items by deleting their values for non-respondents The sample proportions are also given in Table 1. After one month, the response rate was 47.2%, while after the full two-month interview period, the rate was 59.7%. In the 1998 POLS, the first month was CAPI (Computer Assisted Personal Interview). Non-respondents after the first month were allocated to CATI (Computer Assisted Telephone Interview) when they had a listed, land-line phone. Otherwise, they were allocated once more to CAPI. Hence, the second interview month gave another 12.5% of response. However, from table 1 we can see that after the second month, the survey estimates have a larger bias than after the first month.

Table 1
Response means in POLS for the first month of interviews and the full two-month interview period

Variable	After 1 month	After 2 months	Sample
Receiving social allowance	10.5%	10.4%	12.1%
Non-native	12.9%	12.5%	15.0%
Response rate	47.2%	59.7%	100%

From the example, it seems clear that the increased effort led to a less representative response with respect to both auxiliary variables. But what do we mean by representative in general?

It turns out that the term "representative" is often used with hesitation in the statistical literature. Kruskal and Mosteller (1979a, b and c) make an extensive inventory of the use of the word "representative" in the literature and identify nine interpretations. A number of interpretations they have found are omnipresent in the statistical literature. The statistical interpretations that Kruskal and Mosteller named 'absence of selective forces', 'miniature of the population', and 'typical or ideal cases' relate to probability sampling, quota sampling and purposive sampling. In the next section, we will propose a definition that corresponds to the 'absence of selective forces' interpretation. First, we will explain why we make this choice.

The concept of representative response is also closely related to the missing-data mechanisms Missing-Completely-at-Random (MCAR), Missing-at-Random (MAR) and Not-Missing-at-Random (NMAR) that are often referred to in the literature, see Little and Rubin (2002). A missing-data mechanism is MCAR when the probability of response does not depend on the survey topic of interest. The mechanism is MAR if the response probability depends on observed data only, which is, hence, a weaker assumption than MCAR. If the probability depends on missing data also, then the mechanism is said to be NMAR. These mechanisms, in fact, find their origin in model-based statistical theory. Somewhat loosely interpreted with respect to a survey topic, MCAR means that respondents are on average the same as non-respondents, MAR means that within known subpopulations, respondents are on average the same as non-respondents, and NMAR implies that even within subpopulations, respondents are different. The addition of the survey topic is essential. Within one questionnaire, some survey items can be MCAR, while other items are MAR or NMAR. Furthermore, the MAR assumption for one survey item holds for a particular stratification of the population. A different item may need a different stratification.

Given that we wish to monitor and compare the response to different surveys in topic or time, it is not appealing to define a representative response as dependent on the survey topic itself nor as dependent on the estimator used. We focus instead on the quality of data collection and not on the estimation. This setting leads us to compare the response composition to that of the sample. Clearly, the survey topics influence the probability that households participate in the survey, but the influence cannot be measured or tested and, hence, from our perspective, this influence cannot be the input for assessing response quality. We propose to judge the composition of response by pre-defined sets of variables that are observed outside of the survey and can be employed for each survey under investigation. We want the respondent selection to be as close as possible to a 'simple random sample of the survey sample', *i.e.*, with as little relation as possible between response and characteristics that distinguish units from each other. The latter can be interpreted as having selective forces which are absent in the selection of respondents, or as MCAR with respect to all possible survey variables.

2.2 Definition of a representative response subset

Let $i = 1, 2, 3, \ldots, N$ be the unit labels for the population. By s_i we denote the 0-1-sample indicator, *i.e.*, when unit i is sampled, it takes the value 1 and 0 otherwise. By r_i we denote the 0-1-response indicator for unit i. If unit i is sampled and did respond then $r_i = 1$. It is 0 otherwise. The sample size is n. Finally, π_i denotes the first-order inclusion probability of unit i.

The key to our definitions lies in the individual response propensities. Let ρ_i be the probability that unit i responds when it is sampled.

The interpretation of a response propensity is not straightforward by itself. We follow a model-assisted approach, *i.e.*, the only randomness is in the sample and response indicators. A response probability is a feature of a labelled and identifiable unit, a biased coin that the unit carries in a pocket, so to speak, and is, therefore, inseparable from that unit. With a little effort, however, all concepts can be translated into a model-based context.

First, we give a strong definition.

Definition (*strong*): *A response subset is representative with respect to the sample if the response propensities* ρ_i *are the same for all units in the population*

$$\rho_i = P[r_i = 1 \mid s_i = 1] = \rho, \quad \forall i, \tag{1}$$

and if the response of a unit is independent of the response of all other units.

If a missing-data mechanism would satisfy the strong definition, then the mechanism would correspond to Missing-Completely-at-Random (MCAR) with respect to all possible survey questions. Although the definition is appealing, the validity of it can never be tested in practice. We have no replicates of the response of one single unit. We, therefore, also construct a weak definition that can be tested in practice.

Definition (*weak*): *A response subset is representative of a categorical variable* X *with* H *categories if the average response propensity over the categories is constant*

$$\bar{\rho}_h = \frac{1}{N_h} \sum_{k=1}^{N_h} \rho_{hk} = \rho, \quad \text{for} \quad h = 1, 2, \ldots, H, \tag{2}$$

where N_h *is the population size of category* h, ρ_{hk} *is the response propensity of unit* k *in class* h *and summation is over all units in this category.*

The weak definition corresponds to a missing-data mechanism that is MCAR with respect to X, as MCAR states that we cannot distinguish respondents from nonrespondents based on knowledge of X.

3. R-indicators

In the previous section, we defined strong and weak representative response. Both definitions make use of individual response probabilities that are unknown in practice. First, we start with a population R-indicator. From there on, we base the same R-indicator on a sample and on estimated response propensities.

3.1 Population R-indicators

We first consider the hypothetical situation where the individual response propensities are known. Clearly, in that case we can even test the strong definition and we simply want to measure the amount of variation in the response propensities; the more variation, the less representative in the strong sense. Let $\rho = (\rho_1, \rho_2, \ldots, \rho_N)'$ be a vector of response propensities, let $\mathbf{1} = (1, 1, \ldots, 1)'$ be the N-vector of ones, and let $\rho_0 = \mathbf{1} \times \bar{\rho}$ be the vector consisting of the average population propensity.

Any distance function d in $[0, 1]^N$ would suffice in order to measure the deviation from a strong representative response by calculating $d(\rho, \rho_0)$. Note that the height of the overall response does not play a role. The Euclidean distance is a straightforward distance function. When applied to a distance between ρ and ρ_0, this measure is proportional to the standard deviation of the response probabilities

$$S(\rho) = \sqrt{\frac{1}{N - 1} \sum_{i=1}^{N} (\rho_i - \bar{\rho})^2}. \tag{3}$$

It is not difficult to show that

$$S(\rho) \leq \sqrt{\bar{\rho}(1-\bar{\rho})} \leq \frac{1}{2}. \quad (4)$$

We want the R-indicator to take values on the interval [0, 1] with the value 1 being strong representativeness and the value 0 being the maximum deviation from strong representativeness. We propose the R-indicator R, which is defined by

$$R(\rho) = 1 - 2S(\rho). \quad (5)$$

Note that the minimum value of (5) depends on the response rate, see figure 1. For $\bar{\rho} = 1/2$, it has a minimum value of 0. For $\bar{\rho} = 0$ and $\bar{\rho} = 1$, clearly no variation is possible and the minimum value is 1. Paradoxically, the lower bound increases when the response rate decreases from $1/2$ to 0. For a low response rate, there is less room for individual response propensities to have a large variation.

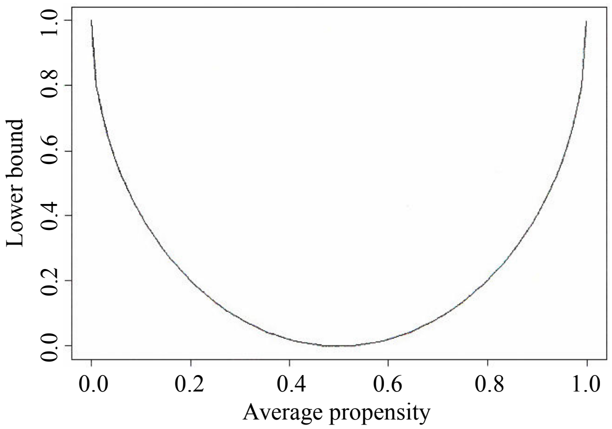


Figure 1 Minimum value of R-indicator (5) as a function of the average response propensity

One may view R as a lack of association measure. When $R(\rho) = 1$ there is no relation between any survey item and the missing-data mechanism. We show that R in fact has a close relation to the well-known χ^2-statistic that is often used to test independence and goodness-of-fit.

Suppose that the response propensities are only different for classes h defined by a categorical variable X. Let $\bar{\rho}_h$ and f_h be, respectively, the response propensity and the population function of class h, *i.e.*,

$$f_h = \frac{N_h}{N}, \quad \text{for} \quad h = 1, 2, \ldots, H. \quad (6)$$

Hence, for all i with $X_i = h$ the response propensity is $\rho_i = \bar{\rho}_h$.

Since the variance of the response propensities is the sum of the 'between' and 'within' variances over classes h, and the within variances are assumed to be zero, it holds that

$$\begin{aligned} S^2(\tilde{\rho}) &= \frac{1}{N-1}\sum_{h=1}^{H} N_h(\bar{\rho}_h - \bar{\rho})^2 \\ &= \frac{N}{N-1}\sum_{h=1}^{H} f_h(\bar{\rho}_h - \bar{\rho})^2 \approx \sum_{h=1}^{H} f_h(\bar{\rho}_h - \bar{\rho})^2. \quad (7) \end{aligned}$$

The χ^2-statistic measures the distance between observed and expected proportions. However, it is only a true distance function in the mathematical sense for fixed marginal distributions f_h and $\bar{\rho}$. We can apply the χ^2-statistic to X in order to 'measure' the distance between the true response behaviour and the response behaviour that is expected when response is independent of X. In other words, we measure the deviation from weak representativeness with respect to X.

We can rewrite the χ^2-statistic to get

$$\begin{aligned} \chi^2 &= \sum_{h=1}^{H} \frac{(N_h\bar{\rho}_h - N_h\bar{\rho})^2}{N_h\bar{\rho}} \\ &\quad + \sum_{h=1}^{H} \frac{(N_h(1-\bar{\rho}_h) - N_h(1-\bar{\rho}))^2}{N_h(1-\bar{\rho})} \\ &= \sum_{h=1}^{H} \frac{Nf_h(\bar{\rho}_h - \bar{\rho})^2}{\bar{\rho}} + \sum_{h=1}^{H} \frac{Nf_h(\bar{\rho}_h - \bar{\rho})^2}{(1-\bar{\rho})} \\ &= \frac{N}{\bar{\rho}(1-\bar{\rho})}\sum_{h=1}^{H} f_h(\bar{\rho}_h - \bar{\rho})^2 \\ &= \frac{N-1}{\bar{\rho}(1-\bar{\rho})} S^2(\tilde{\rho}). \quad (8) \end{aligned}$$

An association measure that transforms the χ^2-statistic to the [0, 1] interval, see *e.g.*, Agresti (2002), is Cramèr's V

$$V = \sqrt{\frac{\chi^2}{N(\min\{C, R\} - 1)}}, \quad (9)$$

where C and R are, respectively, the number of columns and rows in the underlying contingency table. Cramèr's V attains a value 0 if observed proportions exactly match expected proportions and its maximum is 1. In our case, the denominator equals N since the response indicator has only two categories: response and non-response. As a consequence, (9) changes into

$$V = \sqrt{\frac{\chi^2}{N}} = \sqrt{\frac{N-1}{N\bar{\rho}(1-\bar{\rho})}}\, S(\tilde{\rho}). \quad (10)$$

From (10) we can see that for large N, Cramèr's V is approximately equal to the standard deviation of the response propensities standardized by the maximal standard deviation $\sqrt{\bar{\rho}(1-\bar{\rho})}$ for a fixed average response propensity $\bar{\rho}$.

3.2 Response-based R-indicators

In section 3.1, we assumed that we know the individual response propensities. Of course, in practice these propensities are unknown. Furthermore, in a survey, we only have information about the response behaviour of sample units. We, therefore, have to find alternatives to the indicators R. An obvious way to do this is to use response-based estimators for the individual response propensities and the average response propensity.

We let $\hat{\rho}_i$ denote an estimator for ρ_i which uses all or a subset of the available auxiliary variables. Methods that support such estimation are, for instance, logistic or probit regression models (Agresti 2002) and CHAID classification trees (Kass 1980). By $\hat{\bar{\rho}}$ we denote the weighted sample average of the estimated response propensities, *i.e.*,

$$\hat{\bar{\rho}} = \frac{1}{N} \sum_{i=1}^{N} \hat{\rho}_i \frac{s_i}{\pi_i}, \tag{11}$$

where we use the inclusion weights.

We replace R by the estimators $\hat{R}$

$$\hat{R}(\rho) = 1 - 2 \sqrt{\frac{1}{N-1} \sum_{i=1}^{N} \frac{s_i}{\pi_i} (\hat{\rho}_i - \hat{\bar{\rho}})^2}. \tag{12}$$

Note that in (12) there are in fact two estimation steps based on different probability mechanisms. The response propensities themselves are estimated and the variation in the propensities is estimated. We return to the consequences of the two estimation steps in section 4.

3.3 Example

We apply the proposed R-indicators to the survey data from the 1998 POLS that we described in section 2.1. Recall that the survey was a combination of face-to-face and telephone interviewing in which the first month was CAPI only. The sample size was close to 40,000 and the response rate was approximately 60%. We linked the fieldwork administration to the sample and deduced whether each contact attempt resulted in a response. This way, we can monitor the pattern of the R-indicator during the fieldwork period.

For the estimation of response rates we used a logistic regression model with region, ethnic background and age as independent variables. Region was a classification with 16 categories, the 12 provinces and the four largest cities – Amsterdam, Rotterdam, The Hague and Utrecht – as separate categories. Ethnic background has seven categories: native, Moroccan, Turkish, Surinam, Dutch Antilles, other non-western non-native and other western non-native. The classification is based on the country of birth of the parents of the selected person. The age variable has three categories: 0 – 34 years, 35 – 54 years, and 55 years and older.

In figure 2, $\hat{R}$ is plotted against the response rate for the first six contact attempts in POLS. The leftmost value corresponds to the respondent pool after one attempt was made. For each additional attempt, the response rate increases but the indicator shows a drop in representativeness. This result confirms the findings in Schouten (2004).

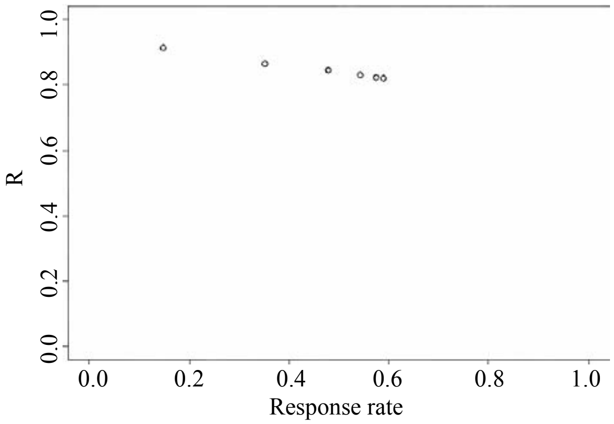


Figure 2 R-indicator for first six contact attempts in POLS 1998

4. Features of R-indicators

In section 3, we propose a candidate indicator for representativeness. However, other indicators can be constructed. There are many association measures or fit indexes, *e.g.*, Goodman and Kruskal (1979), Bentler (1990) and Marsh, Balla and McDonald (1988). Association measures have a strong relation to R-indicators. Essentially, R-indicators attempt to measure in a multivariate setting the lack of association. In this section, we discuss the desired features of R-indicators. We show that the proposed R-indicator R allows for a straightforward upper bound on the non-response bias.

4.1 Features in general

We want R-indicators to be based on a distance function or metric in the mathematical sense. The triangle inequality property of a distance function allows for a partial ordering of the variation in response propensities which enables interpretation. A distance function can easily be derived from any mathematical norm. In section 3, we chose to use the Euclidean norm as this norm is commonly used. The Euclidean norm led us to an R-indicator that uses the standard deviation of response propensities. Other norms, like the supremum norm, would lead us to alternative distance functions. In section 4.3, however, we show that the Euclidean norm based R-indicators have interesting normalization features.

We must make a subtle distinction between R-indicators and distance functions. Distance functions are symmetric while an R-indicator measures a deviation with respect to a specific point, namely the situation where all response propensities are equal. If we change the vector of individual propensities, then this point is in most cases shifted. However, if we fix the average response propensity, then the distance function facilitates interpretation.

Apart from a relation to a distance function, we want to be able to measure, interpret and normalize the R-indicators. In section 3.2, we already derived response-based estimators for 'population' R-indicators that are not measurable when response propensities are unknown and all we have is the response to a survey. Hence, we made R-indicators measurable by switching to estimators. The other two features are discussed separately in the next two sections.

4.2 Interpretation

The second feature of R-indicators is the ease with which we can interpret their values and the concept they are measuring. We moved to an estimator for an R-indicator that is based on the samples of surveys and on estimators of individual response probabilities. Both have far-reaching consequences for the interpretation and comparison of the R-indicator.

Since the R-indicator is an estimator itself, it is also a random variable. This means that it depends on the sample, *i.e.*, it is potentially biased and has a certain accuracy. But what is it estimating?

Let us first assume that the sample size is arbitrarily large so that precision does not play a role and also suppose the selection of a model for response propensities is no issue. In other words, we are able to fit any model for any fixed set of auxiliary variables.

There is a strong relation between the R-indicator and the availability and use of auxiliary variables. In section 2, we defined strong and weak representativeness. Even in the case where we are able to fit any model, we are not able to estimate response propensities beyond the 'resolution' of the available auxiliary variables. Hence, we can only draw conclusions about weak representativeness with respect to the set of auxiliary variables. This implies that whenever an R-indicator is used, it is necessary to complement its value by the set of covariates that served as a grid to estimate individual response propensities. If the R-indicator is used for comparative purposes, then those sets must be the same. We must add that it is not necessary for all auxiliary variables to be used for the estimation of propensities, since they may not add any explanatory power to the model. However, the same sets should be available. The R-indicator then measures a deviation from weak representativeness.

The R-indicator does not capture differences in response probabilities within subgroups of the population other than the subgroups defined by the classes of X. If we let $h = 1, 2, \ldots, H$ again denote strata defined by X, N_h be the size of stratum h, and $\bar{\rho}_h$ be the population average of the response probabilities in stratum h, then it is not difficult to show that $\hat{R}$ is a consistent estimator of

$$R_X(\rho) = 1 - 2\sqrt{\frac{1}{N-1}\sum_{h=1}^{H} N_h(\bar{\rho}_h - \bar{\rho})^2}, \qquad (13)$$

when standard models like logistic regression or linear regression are used to estimate the response probabilities. Of course, (13) and (5) may be different.

In practice, the sample size is not arbitrarily large. The sample size affects both estimation steps; the estimation of response propensities and the estimation of the R-indicator using a sample.

If we knew the individual response propensities, then the sample-based estimation of the R-indicator would only lead to variance and not to bias. We would be able to estimate the population R-indicator without bias. Hence, for small sample sizes, the estimators would have a small precision which could be accounted for by using confidence intervals instead of merely point estimators.

The implications for the estimation of response probabilities are, however, different because of model selection and model fit. There are two alternatives. Either one imposes a model to estimate propensities fixing the covariates beforehand, or one lets the model be dependent on the significant contribution of covariates with respect to some predefined level. In the first case, again no bias is introduced but the standard error may be affected by over fitting. In the second case, the model for the estimation of response propensities depends on the size of the sample; the larger the sample, the more interactions that are accepted as significant. Although it is standard statistical practice to fit models based on a significance level, model selection may introduce bias and variance to the estimation of any R-indicator. This can be easily understood by going to the extreme of a sample of, say, size 10. For such a small sample, no interaction between response behaviour and auxiliary characteristics will be accepted, leaving an empty model and an estimated R-indicator of 1. Small samples simply do not allow for the estimation of response propensities. In general, a smaller sample size will, thus, lead to a more optimistic view on representativeness.

We should make a further subtle distinction. It is possible that, for one survey, a lot of interactions contribute to the prediction of response propensities but each one contributes very little, while in another survey there is only one but it strongly contributes a single interaction. None of the small contributions may be significant, but together they are as strong as the one large contribution that is significant. Hence, we would be more optimistic in the first example even if sample sizes would be comparable.

These observations show that one should always use an R-indicator with some care. It cannot be viewed as separate from the auxiliary variables that were used to compute it. Furthermore, the sample size has an impact on both bias and precision.

4.3 Normalization

The third important feature is the normalization of an R-indicator. We want to be able to attach bounds to an R-indicator so that the scale of an R-indicator, and, hence, changes in the R-indicator get a meaning. Clearly, the interpretation issues that we raised in the previous section also affect the normalization of the R-indicator. Therefore, in this section we assume the ideal situation where we can estimate response propensities without bias. This assumption holds for large surveys. We discuss the normalization of the R-indicator $\hat{R}$.

4.3.1 Maximal absolute bias and maximal root mean square error

We show that for any survey item Y, the R-indicator can be used to set upper bounds to the non-response bias and to the root mean square error (RMSE) of adjusted response means. We use these bounds of the R-indicator to show the impact under worst-case scenarios.

Let Y be some variable that is measured in a survey and let $\hat{\bar{y}}_{\mathrm{HT}}$ be the Horvitz-Thompson estimator for the population mean based on the survey response. It can be shown (*e.g.*, Bethlehem 1988, Särndal and Lundström 2005) that its bias $B(\hat{\bar{y}}_{\mathrm{HT}})$ is approximately equal to

$$B(\hat{\bar{y}}_{\mathrm{HT}}) = \frac{C(y,\, \rho)}{\bar{\rho}}, \tag{14}$$

with $C(y,\, \rho) = 1/N \sum_{i=1}^{N} (y_i - \bar{y})(\rho_i - \bar{\rho})$ the population covariance between the survey items and the response probabilities. For a close approximation of the variance $s^2(\hat{\bar{y}}_{\mathrm{HT}})$ of $\hat{\bar{y}}_{\mathrm{HT}}$ we refer to Bethlehem (1988).

A normalization of R is found by the Cauchy-Schwarz inequality. This inequality states that the covariance between any two variables is bounded in absolute sense by the product of the standard deviations of the two variables. We can translate this to bounds for the bias (14) of $\hat{\bar{y}}_{\mathrm{HT}}$

$$\begin{aligned} |B(\hat{\bar{y}}_{\mathrm{HT}})| \le \frac{S(\rho)\, S(y)}{\bar{\rho}} &= \frac{(1 - R(\rho))\, S(y)}{2\bar{\rho}} \\ &= B_m(\rho,\, y). \end{aligned} \tag{15}$$

Clearly, we do not know the upper bound $B_m(\rho,\, y)$ in (15) but we can estimate it using the sample and the estimated response probabilities. We denote the estimator by $\hat{B}_m(\hat{\rho},\, y)$.

In a similar way, we can set a bound to the root mean square error (RMSE) of $\hat{\bar{y}}_{\mathrm{HT}}$. It holds approximately that

$$\begin{aligned} \mathrm{RMSE}(\hat{\bar{y}}_{\mathrm{HT}}) &= \sqrt{B^2(\hat{\bar{y}}_{\mathrm{HT}}) + s^2(\hat{\bar{y}}_{\mathrm{HT}})} \\ &\le \sqrt{B_m^2(\rho,\, y) + s^2(\hat{\bar{y}}_{\mathrm{HT}})} \\ &= E_m(\rho,\, y). \end{aligned} \tag{16}$$

Again, we do not know $E_m(\rho,\, y)$. Instead, we use the sample-based estimator that employs the estimated response probabilities, denoted by $\hat{E}_m(\hat{\rho},\, y)$.

The bounds $\hat{B}_m(\hat{\rho},\, y)$ and $\hat{E}_m(\hat{\rho},\, y)$ are different for each survey item y. For comparison purposes it is, therefore, convenient to define a hypothetical survey item. We suppose that $\hat{S}(y) = 0.5$. The corresponding bounds we denote by $\hat{B}_m(\hat{\rho})$ and $\hat{E}_m(\hat{\rho})$. They are equal to

$$\hat{B}_m(\hat{\rho}) = \frac{(1 - \hat{R}(\hat{\rho}))}{4\hat{\bar{\rho}}} \tag{17}$$

$$\hat{E}_m(\hat{\rho}) = \sqrt{\hat{B}_m^2(\hat{\rho}) + \hat{s}^2(\hat{\bar{y}}_{\mathrm{HT}})}. \tag{18}$$

We compute (17) and (18) in all studies described in section 5. We have to note that (17) and (18) are again random variables that have a certain precision and that are potentially biased.

4.3.2 Response-representativeness functions

In the previous section, we used the R-indicator to set upper bounds to the non-response bias and to the root mean square error of the (adjusted) response mean. Conversely, we may set a lower bound to the R-indicator by demanding that either the absolute non-response bias or the root mean square error is smaller than some prescribed value. Such a lower bound may be chosen as one of the ingredients of quality restrictions put upon the survey data by a user of the survey. If a user does not want the non-response bias or root mean square to exceed a certain value, then the R-indicator must be bigger than the corresponding bound.

Clearly, lower bounds to the R-indicator depend on the survey item. Therefore, again we restrict ourselves a hypothetical survey item for which $\hat{S}(y) = 0.5$.

It is not difficult to show from (17) that if we demand that

$$\hat{B}_m(\hat{\rho}) \le \gamma, \tag{19}$$

then it must hold that

$$\hat{R} \geq 1 - 4\hat{\bar{\rho}}\gamma = r_1(\gamma, \hat{\rho}). \tag{20}$$

Analogously, using (18) and demanding that

$$\hat{E}_m(\hat{\rho}) \leq \gamma, \tag{21}$$

we arrive at

$$\hat{R} \geq 1 - 4\hat{\bar{\rho}}\sqrt{\gamma^2 - \hat{s}^2(\hat{\bar{y}}_{\mathrm{HT}})} = r_2(\gamma, \hat{\rho}). \tag{22}$$

In (20) and (22) we let $r_1(\gamma, \hat{\rho})$ and $r_2(\gamma, \hat{\rho})$ denote lower limits to the R-indicator. In the following section, we refer to $r_1(\gamma, \hat{\rho})$ and $r_2(\gamma, \hat{\rho})$ as response-representativeness functions. We compute them for the studies in section 5.

4.3.3 Example

We again illustrate the normalization with the same example used in sections 2.1 and 3.3. Figure 3 contains the response-representativeness function $r_1(\gamma, \hat{\rho})$ and the observed R-indicators $\hat{R}$ for the six contact attempts in POLS 1998. Three values of γ are chosen, $\gamma = 0.1$; $\gamma = 0.075$ and $\gamma = 0.05$.

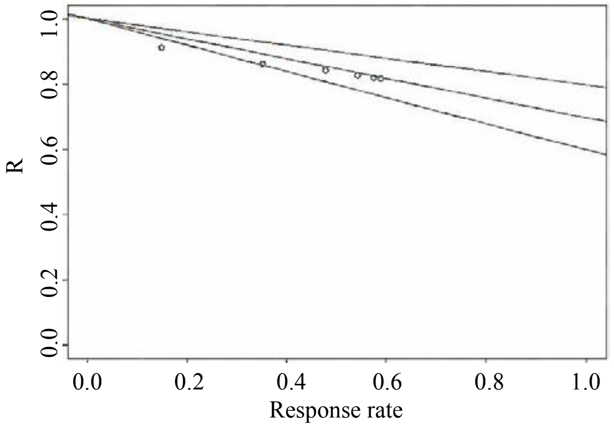


Figure 3 Lower bounds for R-indicator $\hat{R}$ for the first six contact attempts of POLS 1998. Lower bounds are based on $\gamma = 0.1$, $\gamma = 0.075$ and $\gamma = 0.05$

Figure 3 indicates that after the second contact attempt, the values of the R-indicator exceed the lower bound corresponding to the 10%-level. After four attempts, the R-indicator is close to the 7.5%-level. However, the values never exceed the other lower bound that is based on the 5%-level.

In figure 4, the maximal absolute bias $\hat{B}_m(\hat{\rho})$ is plotted against the response rate of the six contact attempts. After the third contact attempt, the R-indicator has converged on a value around 8%.

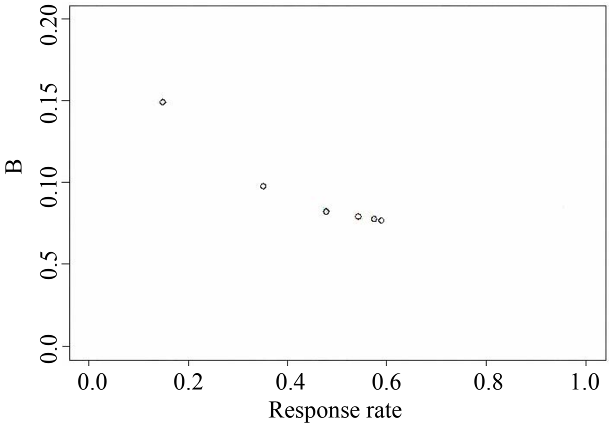


Figure 4 Maximal absolute bias for the first six contact attempts of POLS 1998

5. Application of the R-indicator

In this section, we apply the R-indicator to two studies that investigate different non-response follow-up strategies and different combinations of data collection modes. The first study involves the Dutch Labour Force Survey (LFS). The study is an investigation of both the call-back approach (Hansen and Hurwitz 1946) and the basic-question approach (Kersten and Bethlehem 1984). The second study deals with mixed-mode data collection designs applied to the Dutch Safety Monitor survey.

In sections 5.2 and 5.3 we take a closer look at the studies in connection with the representativeness of their different fieldwork strategies. First, in section 5.1 we describe how we approximate standard errors.

5.1 Standard error and confidence interval

If we want to compare the values of the R-indicator for different surveys or data collection strategies, we need to estimate their standard errors.

The R-indicator $\hat{R}$ involves the sample standard deviation of the estimated response probabilities. This means that there are two random processes involved. The first process is the sampling of the population. The second process is the response mechanism of the sampled units. If the true response probabilities were known, then drawing a sample would still introduce uncertainty about the population R-indicator and, hence, lead to a certain loss of precision. However, since we do not know the true response probabilities, these probabilities are estimated using the sample. This introduces additional precision loss.

An analytical derivation of the standard error of $\hat{R}$ is not straightforward due to the estimation of the response probabilities. In this paper, we are resigned to naïve numerical approximations of the standard error. We estimate the standard error of the R-indicator by non-parametric bootstrapping (Efron and Tibshirani 1993). The non-parametric bootstrap estimates the standard error of the R-indicator by drawing a number $b = 1, 2, \ldots, B$ of so-called bootstrap samples. These are samples drawn independently and with replacement from the original dataset, of the same size n as the original dataset. The R-indicator is calculated for every bootstrap sample b. We thus obtain B replications of the R-indicator; $\hat{R}_b^{\text{BT}}$, $b = 1, 2, \ldots, B$. The standard error for the empirical distribution of these B replications is an estimate for the standard error of the R-indicator, that is

$$s_R^{\text{BT}} = \sqrt{\frac{1}{B-1}\sum_{b=1}^{B} (\hat{R}_b^{\text{BT}} - \hat{\bar{R}}^{\text{BT}})^2} \qquad (23)$$

where $\hat{\bar{R}}^{\text{BT}} = 1/B \sum_{b=1}^{B} \hat{R}_b^{\text{BT}}$ is the average estimated R-indicator.

In the approximations, we take $B = 200$ for all studies. We experimented with larger numbers of B of up to $B = 500$, but found that in all cases, the estimate of the standard error had converged by $B = 200$.

We determine $100(1 - \alpha)\%$ confidence intervals by assuming a normal approximation of the distribution of $\hat{R}$ employing the estimated standard errors using (23)

$$CI_{\alpha}^{\text{BT}} = (\hat{R} \pm \xi_{1-\alpha} \times s_R^{\text{BT}}) \qquad (24)$$

with $\xi_{1-\alpha}$ the $1 - \alpha$ quantile of the standard normal distribution.

5.2 Labour Force Survey; follow-up study 2005

From July to December 2005, Statistics Netherlands conducted a large-scale follow-up of non-respondents in the Dutch Labour Force Survey (LFS). In the study, two samples of non-respondents in the LFS were approached once more using either a call-back approach (Hansen and Hurwitz 1946) or a basic-question approach (Kersten and Bethlehem 1984). The samples consisted of LFS households that refused, were not processed or were not contacted in the LFS for the months July – October. In the design of the follow-up study, we used the recommendations in the studies by Stoop (2005) and Voogt (2004).

The main characteristics of the call-back and basic-question approaches applied to the LFS are given in Table 2. For more details, we refer to Schouten (2007) and Cobben and Schouten (2007). The call-back approach employed the original household questionnaire in CAPI, while the basic-question approach used short questionnaires in a mixed-mode setting. The mixed-mode design involved web, paper and CATI. CATI was used for all households with a listed phone number. Households without a listed phone number received an advance letter, a paper questionnaire and a login to a secure website containing the web questionnaire. Respondents were left the choice to fill in either the paper or web questionnaire.

Table 2
Characteristics of the two approaches in the follow-up study

Call-back approach	Basic-question approach
• LFS questionnaire to be answered by all members of the household in CAPI • 28 interviewers geographically selected from historically best-performing interviewers • Interviewer was different from interviewer that received non-response • Interviewers received additional training in doorstep interaction • Extended fieldwork period of two months • Interviewer could offer incentives • Interviewers could receive a bonus • A paper summary of the characteristics of the non-responding household was sent to the interviewer • Allocation of address one week after non-response	• A strongly condensed questionnaire with key questions of the LFS which takes between 1 and 3 minutes to answer or fill in • Mixed-mode data collection design using web, paper and CATI • The questionnaire was to be answered by one person per household following the next birthday method • The timing is one week after the household is processed as a non-response

The sample size of the LFS pilot was $n = 18{,}074$ households, of which 11,275 households responded. The non-responding households were stratified according to the cause of non-response. Households that were not processed or contacted, and households that refused were eligible for a follow-up. It was considered to be unethical to follow-up households that did not respond due to other causes like illness. In total, 6,171 households were eligible. From these households, two simple random samples were drawn of size 775. In the analyses, the non-sampled eligible households were left out. The sampled eligible households received a weight accordingly. The 11,275 LFS respondents and the 628 ineligible households all received a weight of one. This implies that the inclusion probabilities are unequal for this example.

Schouten (2007) compared the LFS respondents to the converted and persistent non-respondents in the call-back approach using a large set of demographic and socio-economic characteristics. He used logistic regression models to predict the type of response. He concluded that the converted non-respondents in the call-back approach are different from the LFS respondents with respect to the selected auxiliary variables. Furthermore, he found no evidence that the converted non-respondents were different from persistent non-respondents with respect to the same characteristics. These findings have led to the conclusion that the combined response of the LFS and call-back approach is more representative with respect to the selected auxiliary variables.

The additional response in the basic question approach was analyzed by Cobben and Schouten (2007) using the same set of auxiliary variables and employing the same logistic regression models. For this follow-up, the findings were different for households with and without a listed phone number. When restricted to listed households, they found the same results as for the call-back approach; the response becomes more representative after the addition of the listed basic-question respondents. However, for the overall population, *i.e.*, including the unlisted households, the inverse was found. The basic-question approach gives ‘more of the same’ and, hence, sharpens the contrast between respondents and non-respondents. Combining LFS response with basic-question response leads to a less representative composition. In the logistic regression models by Cobben and Schouten (2007) the 0-1 indicators for having a listed phone number and having a paid job gave a significant contribution.

Cobben and Schouten (2007) and Schouten (2007) used the set of auxiliary variables listed in Table 3. The auxiliary variables were linked to the sample from various registers and administrative data. The variables in logistic regression models for response probabilities were selected when the variables gave a significant contribution at the 5% level. Otherwise, they were excluded.

Table 3
The auxiliary variables in the studies by Schouten (2007) and Cobben and Schouten (2007). The household core is the head of the household and his or her partner if present

Variable
Household has a listed phone number
Region of the country in 4 classes
Province and 4 largest cities
Average age in 6 classes
Ethnic group in 4 classes
Degree of urbanization in 5 classes
Household type in 6 classes
Gender
Average house value at zip code level in 11 classes
At least one member of household core is self-employed
At least one member of household core has a subscription to the CWI
At least one member of household core receives social allowance
At least one member of household core has a paid job
At least one member of household core receives disability allowance

Table 4 shows the weighted sample size, response rate, $\hat{R}$, $\mathrm{CI}_{0.05}^{\mathrm{BT}}$, $\hat{B}_m$ and $\hat{E}_m$ for the response to the LFS, the response of the LFS combined with the call-back response and the response of the LFS combined with the basic-question response. The standard errors are relatively large with respect to the studies in subsequent sections due to the weighting. There is an increase in $\hat{R}$ when the call-back respondents are added to the LFS respondents. As both the response rate and the R-indicator increase, the maximal absolute bias $\hat{B}_m$ decreases. The confidence intervals $\mathrm{CI}_{0.05}^{\mathrm{BT}}$ for the LFS response and the combined LFS and call-back response overlap. However, the one-sided null hypothesis H_0: $R_{\mathrm{LFS}} - R_{\mathrm{LFS+CB}} \geq 0$ is rejected at the 5%-level.

Table 4
Weighted sample size, response rate, R-indicator, confidence interval, maximal bias and maximal RMSE for LFS, LFS plus call-back, and LFS plus basic-question for the extended set of auxiliary variables

Response	n	Rate	$\hat{R}$	$\mathrm{CI}_{0.05}^{\mathrm{BT}}$	$\hat{B}_m$	$\hat{E}_m$
LFS	18,074	62.2%	80.1%	(77.5-82.7)	8.0%	8.0%
LFS + call-back	18,074	76.9%	85.1%	(82.4-87.8)	4.8%	4.9%
LFS + basic-question	18,074	75.6%	78.0%	(75.6-80.4)	7.3%	7.3%

In Table 4, there is a decrease in $\hat{R}$ when we compare the LFS response to the combined response with the basic-question approach. This decrease is not significant. $\hat{B}_m$ slightly decreases. In Table 5, this comparison is restricted to households with a listed phone number. The R-indicator in general is much higher than for all the households. Because the sample size is now smaller, the estimated standard errors are larger as is reflected in the width of the confidence interval. $\hat{B}_m$ is decreased. For the combined response in the LFS and the basic-question approach, we see an increase of $\hat{R}$ but again this increase is not significant. $\hat{B}_m$ decreases.

Table 5
Sample size, response rate, R-indicator, confidence interval, maximal bias and maximal RMSE for LFS, and LFS plus basic-question restricted to households with listed phone numbers and for the extended set of auxiliary variables

Response	n	Rate	$\hat{R}$	$\mathrm{CI}_{0.05}^{\mathrm{BT}}$	$\hat{B}_m$	$\hat{E}_m$
LFS	10,135	68.5%	86.3%	(83.1-89.5)	5.0%	5.1%
LFS + basic-question	10,135	83.0%	87.5%	(84.3-90.7)	3.8%	3.8%

We find in the example of the LFS follow-up that the R-indicators confirm the conclusions for the call-back approach and the basic question approach. Furthermore, the increase in the R-indicator that follows by adding the call-back response is significant at the 5% level.

5.3 Safety Monitor; pilot mixed-mode 2006

In 2006, Statistics Netherlands conducted a pilot on the Safety Monitor to investigate mixed-mode data collection strategies. See Cobben, Janssen, Berkel and Brakel (2007) for details. The regular Safety Monitor surveys individuals of 15 years and older in the Netherlands about issues that relate to safety and police performance. The Safety Monitor is a mixed-mode survey. Persons with a listed phone number are approached by CATI. Persons that cannot be reached by telephone are approached by CAPI. In the 2006 pilot, the possibility of using the Internet as one of the modes in a mixed-mode strategy was evaluated. Persons in the pilot were first approached with a web survey. Non-respondents to the web survey were re-approached by CATI when they had a listed phone number and by CAPI otherwise. In Table 6 we give the response rates for the normal survey, the pilot response to the web only, and the response to the pilot as a whole. The response to the web survey alone is low. Only 30% of the persons filled in the web questionnaire. This implied that close to 70% of the sampled units were re-allocated to either CAPI or CATI. This resulted in an additional response of approximately 35%. The overall response rate is slightly lower than that of the normal survey.

Fouwels, Janssen and Wetzels (2006) performed a univariate analysis of response compositions. They argue that the response rate is lower for the pilot but that this decrease is quite stable over various demographic subgroups. They observe a univariate decline in response rate for the auxiliary variables age, ethnic group, degree of urbanization and type of household. However, they do find indications that the response becomes less representative when the comparison is restricted to the web respondents only. This holds, not surprisingly, especially for the age of the sampled persons.

Table 6 contains the sample size, response rate, $\hat{R}$, $\mathrm{CI}^{\mathrm{BT}}_{0.05}$, $\hat{B}_m$ and $\hat{E}_m$ for three groups: the regular survey, the pilot survey restricted to web and the pilot survey as a whole. The auxiliary variables age, ethnic group, degree of urbanization and type of household were linked from registers and were selected in the logistic model for the response probabilities. Table 6 shows that the R-indicator for the web response is lower than that of the regular survey. The corresponding p-value is close to 5%. As a consequence of both a low response rate and a low R-indicator, the maximal absolute bias $\hat{B}_m$ is more than twice as high as for the regular survey. However, for the pilot as a whole, both the R-indicator and $\hat{B}_m$ are close to the values of the regular survey. Due to the smaller sample size of the pilot, the estimated standard errors are larger than in the regular survey.

Table 6
Sample size, response rate, R-indicator, confidence interval, maximal bias and maximal RMSE for response for the regular Safety Monitor, the pilot with web only and the pilot with web and CAPI/CATI follow-up

Response	n	Rate	$\hat{R}$	$\mathrm{CI}^{\mathrm{BT}}_{0.05}$	$\hat{B}_m$	$\hat{E}_m$
Regular	30,139	68.9%	81.4%	(80.3-82.4)	6.8%	6.8%
Pilot - web	3,615	30.2%	77.8%	(75.1-80.5)	18.3%	18.4%
Pilot - web plus	3,615	64.7%	81.2%	(78.3-84.0)	7.3%	7.4%

The findings in Table 6 do not contradict those of Fouwels *et al.* (2006). We also find that the web response in the pilot has a less balanced composition, whereas the composition of the full pilot response is not markedly worse than that of the Safety Monitor itself.

6. Discussion

We have three main objectives in this paper: a mathematically rigorous definition and perception of representative response, the construction of a potential indicator for representativeness, and the empirical illustrations of the indicator's use. As we saw, the proposed indicator is an example of what we call R-indicators, where 'R' stands for representativeness. With the empirical illustration, we want to find support for the idea that such R-indicators are valuable tools in the comparison of different surveys and data collection strategies. R-indicators are useful if they confirm findings in elaborate analyses of studies that involve multiple surveys in time or on a topic.

The R-indicator in this paper is promising because it can easily be computed and allows for interpretation and normalization when response propensities can be estimated without error. The application to real survey data shows that the R-indicator confirms earlier analyses of the non-response composition. Other R-indicators can, of course, simply be constructed by choosing different distance functions between vectors of response propensities. The R-indicator and graphical displays showed in this paper can be computed using most standard statistical software packages.

The computation of R-indicators is sample-based and employs models for individual response propensities. Hence, R-indicators are random variables themselves and there are two estimation steps that influence their bias and variance. However, it is mostly the modelling of response propensities that has important implications. The restriction to the sample for the estimation of R-indicators implies that those indicators are less precise, but this restriction does not introduce a bias asymptotically. Model selection and model fit usually are performed by choosing a significance level and adding only those interactions to the model that give a significant contribution. The latter means that the size of the sample and the availability of auxiliary variables play an important role in the estimation of response propensities. Bias may be introduced by the model selection strategy. There are various obvious approaches for dealing with the dependence on the size of the sample. One may not do a model selection but fix a stratification beforehand. That way, bias is avoided but standard errors are not controlled and may be considerable. One may also let empirical validation be the input to develop 'best practices' for R-indicators.

We applied the proposed R-indicator to two studies that were conducted at Statistics Netherlands in recent years, and that were thoroughly investigated by other authors. The increase or decrease in the R-indicator conforms to the more detailed analyses done by these authors. We, therefore, conclude that R-indicators can be valuable tools. However, more empirical evidence is clearly needed.

The application of the R-indicator showed that there is no clear relation between response rate and representativeness of response. Larger response rates do not necessarily lead to a more balanced response. Not surprisingly, we do find that higher response rates reduce the risk of non-response bias. The higher the response rate, the smaller the maximal absolute bias of survey items.

Application to the selected studies showed that standard errors do decrease with increasing sample size as expected, but they are still relatively large for modest sample sizes. For example, for a sample size of 3,600, we found a standard error of approximately 1.3%. Hence, if we assume a normal distribution, then the 95% confidence interval has an approximate width of 5.4%. The sample size of the LFS is about 30,000 units. The standard error is approximately 0.5% and the corresponding 95% confidence interval is approximately 2% wide. The standard errors are larger than we expected.

This paper contains a first empirical study of an R-indicator and its standard error. Much more theoretical and empirical research is necessary to fully understand R-indicators and their properties. First, we did not consider survey items at all. Clearly, it is imperative that we do this in the future. However, as we already argued, R-indicators are dependent on the set of auxiliary variables. It can, therefore, be conjectured that, as for non-response adjustment methods, the extent to which R-indicators predict non-response bias of survey items is dependent on the missing-data mechanism. In a missing-data mechanism that is strongly non-ignorable, R-indicators will not do a good job. However, without knowledge about the missing-data mechanism, no other indicator would either. For this reason, we constructed the notion of maximal absolute bias, as this gives a limit to non-response bias under the worst-case scenario. A second topic of future research is a theoretical derivation of the standard error of the R-indicator used in this paper. The non-parametric bootstrap errors only give naïve approximations. However, if we want R-indicators to play a more active role in the comparison of different strategies, then we need (approximate) closed forms. Third, we will need to investigate the relation between the selection and number of auxiliary variables and the standard errors of the R-indicator.

Acknowledgements

The authors would like to thank Bob Groves, Björn Janssen, Geert Loosveldt, and the associate editor and two referees for their useful comments and suggestions.

References

Agresti, A. (2002). Categorical data analysis. *Wiley Series in Probability and Statistics*. New York: John Wiley & Sons, Inc., NY, USA.

Bentler, P.M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin*, 107, 2, 238-246.

Bertino, S. (2006). A measure of representativeness of a sample for inferential purposes. *International Statistical Review*, 74, 149-159.

Bethlehem, J.G. (1988). Reduction of nonresponse bias through regression estimation. *Journal of Official Statistics*, 4, 3, 251-260.

Cobben, F., Janssen, B., Berkel, K. van and Brakel, J. van den (2007). Statistical inference in a mixed-mode data collection setting. Paper presented at ISI 2007, August 23-29, 2007, Lisbon, Portugal.

Cobben, F., and Schouten, B. (2007). An empirical validation of R-indicators. Discussion paper, CBS, Voorburg.

Curtin, R., Presser, S. and Singer, E. (2000). The effects of response rate changes on the index of consumer sentiment. *Public Opinion Quarterly*, 64, 413-428.

Efron, B., and Tibshirani, R.J. (1993). An introduction to the bootstrap. Chapman & Hall/CRC.

Fouwels, S., Janssen, B. and Wetzels, W. (2006). Experiment mixed-mode waarneming bij de VMR. Technical paper SOO-2007-H53, CBS, Heerlen.

Goodman, L.A., and Kruskal, W.H. (1979). Measures of association for cross-classifications. Springer-Verlag, Berlijn, Duitsland.

Groves, R.M. (2006). Nonresponse rates and nonresponse bias in household surveys. *Public Opinion Quarterly*, 70, 5, 646-675.

Groves, R.M., and Peytcheva, E. (2006). The impact of nonresponse rates on nonresponse bias: A meta-analysis. Paper presented at 17th International Workshop on Household Survey Nonresponse, August 28-30, Omaha, NE, USA.

Groves, R.M., Presser, S. and Dipko, S. (2004). The role of topic interest in survey participation decisions. *Public Opinion Quarterly*, 68, 2-31.

Hájek, J. (1981). Sampling from finite populations. New York: Marcel Dekker, USA.

Hansen, M.H., and Hurwitz, W.H. (1946). The problem of nonresponse in sample surveys. *Journal of the American Statistical Association*, 41, 517-529.

Heerwegh, D., Abts, K. and Loosveldt, G. (2007). Minimizing survey refusal and noncontact rates: Do our efforts pay off? *Survey Research Methods*, 1, 1, 3-10.

Kass, G.V. (1980). An exploratory technique for investigating large quantities of categorical data. *Applied Statistics*, 29, 2, 119-127.

Keeter, S., Miller, C., Kohut, A., Groves, R.M. and Presser, S. (2000). Consequences of reducing nonresponse in a national telephone survey. *Public Opinion Quarterly*, 64, 125-148.

Kersten, H.M.P., and Bethlehem, J.G. (1984). Exploring an reducing the nonresponse bias by asking the basic question. *Statistical Journal of the United Nations*, ECE 2, 369-380.

Kohler, U. (2007). Surveys from inside: An assessment of unit nonresponse bias with internal criteria. *Survey Research Methods*, 1, 2, 55-67.

Kruskal, W., and Mosteller, F. (1979a). Representative sampling I: Non-scientific literature. *International Statistical Review*, 47, 13-24.

Kruskal, W., and Mosteller, F. (1979b). Representative sampling II: Scientific literature excluding statistics. *International Statistical Review*, 47, 111-123.

Kruskal, W., and Mosteller, F. (1979c). Representative sampling III: Current statistical literature. *International Statistical Review*, 47, 245-265.

Little, R.J.A., and Rubin, D.B. (2002). Statistical analysis with missing data. Wiley Series in Probability and Statistics. New York: John Wiley & Sons, Inc., NY, USA.

Marsh, H.W., Balla, J.R. and McDonald, R.P. (1988). Goodness-of-fit indexes in confirmatory factor analysis: The effect of sample size. *Psychological Bulletin*, 103, 3, 391-410.

Merkle, D.M., and Edelman, M. (2002). Nonresponse in exit polls: A comprehensive analysis. In *Survey Nonresponse* (Eds. R.M. Groves, D.A. Dillman, J.L. Eltinge and R.J.A Little). New York: John Wiley & Sons, Inc., 243-258.

Rubin, D.B. (1976). Inference and missing data. *Biometrika*, 63, 581-592.

Särndal, C., and Lundström, S. (2005). Estimation in Surveys with Nonresponse. Wiley Series in Survey Methodology, John Wiley & Sons, Chichester, England.

Särndal, C., Swensson, B. and Wretman, J. (2003). Model-assisted survey sampling. Springer Series in Statistics, Springer, New York.

Schouten, B. (2004). Adjustment for bias in the Integrated Survey on Household Living Conditions (POLS) 1998. Discussion paper 04001, CBS, Voorburg, available at website http://www.cbs.nl/nl-NL/menu/methoden/research/discussionpapers/archief/2004/default.htm.

Schouten, B., and Cobben, F. (2007). R-indicators for the comparison of different fieldwork strategies and data collection modes, Discussion paper 07002, CBS, Voorburg. Available at website http://www.cbs.nl/nl-NL/menu/methoden/research/discussionpapers/archief/2007/default.htm.

Stoop, I. (2005). Surveying nonrespondents. *Field Methods*, 16, 23-54.

Voogt, R. (2004). I am not interested: Nonresponse bias, response bias and stimulus effects in election research. PhD dissertation, University of Amsterdam, Amsterdam.

Stratified balanced sampling

Guillaume Chauvet [1]

Abstract

In the selection of a sample, a current practice is to define a sampling design stratified on subpopulations. This reduces the variance of the Horvitz-Thompson estimator in comparison with direct sampling if the strata are highly homogeneous with respect to the variable of interest. If auxiliary variables are available for each individual, sampling can be improved through balanced sampling within each stratum, and the Horvitz-Thompson estimator will be more precise if the auxiliary variables are strongly correlated with the variable of interest. However, if the sample allocation is small in some strata, balanced sampling will be only very approximate. In this paper, we propose a method of selecting a sample that is balanced across the entire population while maintaining a fixed allocation within each stratum. We show that in the important special case of size-2 sampling in each stratum, the precision of the Horvitz-Thompson estimator is improved if the variable of interest is well explained by balancing variables over the entire population. An application to rotational sampling is also presented.

Key Words: Rotational sampling; Maximum entropy; Cube method; Stratification; Unequal probability sampling.

1. Introduction

In the case of stratified sampling, a population U is partitioned into H subpopulations U_h, $h = 1, ..., H$ called strata, in which samples S_h, $h = 1, ..., H$ are selected according to independent sampling designs p_h, $h = 1, ..., H,$ respectively. The inclusion probability of unit k is the probability π_k that unit k is in the sample, and the joint inclusion probability is the probability π_{kl} that two distinct units k and l are jointly in the sample. We will write $\boldsymbol{\pi} = (\pi_k)_{k \in U}$ and $\boldsymbol{\pi}^{\mathbf{h}} = (\pi_k)_{k \in U_h}$. We assume that within each stratum U_h, design $p_h(.)$ is of fixed size. In particular, then, we have $\sum_{k \in U_h} \pi_k = n_h$, $h = 1, ..., H,$ where n_h denotes the allocation in stratum U_h. In the rest of the paper, we assume that all sample sizes for stratum n_h are integers.

The Horvitz-Thompson estimator $\hat{t}_{\mathbf{z}\pi} = \sum_{k \in S} \mathbf{z}_k / \pi_k = \sum_{h=1}^{H} \hat{t}_{\mathbf{z}\pi}^{h}$, where $\hat{t}_{\mathbf{z}\pi}^{h} = \sum_{k \in S_h} \mathbf{z}_k / \pi_k$, provides an unbiased estimate of $t_{\mathbf{z}} = \sum_{k=1}^{H} t_{\mathbf{z}}^{h}$, where $t_{\mathbf{z}}^{h} = \sum_{k \in U_h} \mathbf{z}_k$ denotes the total of the variable (vector) $\mathbf{z}$ over U_h. In the particular case where $\mathbf{z}_k = y_k$ is scalar, the variance of the Horvitz-Thompson estimator is given by the Sen-Yates-Grundy variance formula:

$$\operatorname{Var}(\hat{t}_{y\pi}) = \frac{1}{2} \sum_{h=1}^{H} \sum_{k \neq l \in U_h} (\pi_k \pi_l - \pi_{kl}) \left(\frac{y_k}{\pi_k} - \frac{y_l}{\pi_l} \right)^2. \qquad (1)$$

This variance is small if the strata are homogeneous with respect to the variable of interest, specifically if y_k / π_k is approximately constant within each stratum.

If a vector $\mathbf{x} = (x_1, ..., x_q)$ of q auxiliary variables is available prior to sample selection for each individual in the population, the sampling within each stratum can be improved with the cube algorithm (Deville and Tillé 2004), which selects balanced samples. Sampling design $p_h(.)$ is said to be balanced on the $\mathbf{x}$ variables if the equations

$$\hat{t}_{\mathbf{x}\pi}^{h} = t_{\mathbf{x}}^{h} \qquad (2)$$

are exactly satisfied. The variance of the Horvitz-Thompson estimator is therefore zero for the estimate of the total of the balancing variables. In the particular case where $\mathbf{x} = \boldsymbol{\pi}$, *i.e.*, if the inclusion probability is the only balancing variable, (2) reduces to

$$\sum_{k \in S_h} 1 = \sum_{k \in U_h} \pi_k = n_h. \qquad (3)$$

Hence, stratified sampling of fixed size in each stratum is a particular case of balanced sampling. For any given number of constraints, an exactly balanced sample generally cannot be found. Suppose, for example, that population U_h contains 100 individuals on whom is defined a variable x with two possible values, 0 and 1, and that 53 individuals in the population have the value 0 for that variable. Selecting a size-10 equal-probability sample balanced on variable x would mean selecting a sample containing 5,3 individuals for whom $x = 0$ and 4,7 individuals for whom $x = 1$, which is impossible. Consequently, the goal is generally to select an approximately balanced sample, such that

$$\hat{t}_{\mathbf{x}\pi}^{h} \simeq t_{\mathbf{x}}^{h}. \qquad (4)$$

With the cube method (Deville and Tillé 2004), we can select approximately balanced samples on any number of variables, maintaining exactly a predetermined set of inclusion probabilities $\boldsymbol{\pi}$. The method is composed of two phases: the flight phase and the landing phase. At each step in the flight phase, we decide at random to either select or permanently discard one of the population units. At the end of the flight phase, we have, in each stratum U_h, a vector

[1] Guillaume Chauvet, Laboratoire de Statistique d'Enquête, CREST/ENSAI, rue Blaise Pascal, Campus de Ker Lann, 35 170 Bruz, France. E-mail: chauvet@ensai.fr.

$\boldsymbol{\pi}^{\mathbf{h}*} = (\pi_k^*)_{k\in U_h} \in [0, 1]^N$ that satisfies the following conditions:

$$E(\boldsymbol{\pi}^{\mathbf{h}*}) = \boldsymbol{\pi}^{\mathbf{h}}, \quad (5)$$

$$\sum_{k\in U_h} \frac{\mathbf{x}_k}{\pi_k}\pi_k^* = \sum_{k\in U_h} \mathbf{x}_k, \quad (6)$$

$$\text{Card}\{k \in U_h; 0 < \pi_k^* < 1\} \leq q, \quad (7)$$

where E denotes the expectation for the sampling method used in the flight phase. The vector $\boldsymbol{\pi}^{\mathbf{h}*}$ gives the outcome of the flight phase: π_k^* is 1 if unit k is selected, 0 if it is rejected, and between 0 and 1 only if the decision has not been made for unit k after the flight phase. Equations (5) and (6) ensure that the inclusion probabilities and balancing constraints are maintained perfectly at the end of the flight phase. Equation (7) ensures that a decision remains to be made for no more than q individuals in each stratum U_h, where q is the number of balancing variables. The flight phase ends when the balancing constraints can no longer be exactly satisfied. The landing phase consists in defining, conditionally on the outcome of the flight phase, an optimal sampling design defined on the remaining population V. This design is optimal in that it makes it possible to complete the sampling while minimizing the variance, conditionally on the outcome of the flight phase, of the Horvitz-Thompson estimator of the balancing variables. The remaining units are sampled, conditionally on the outcome of the flight phase, with inclusion probabilities $(\pi_k^*)_{k\in V}$, so that the units' unconditional inclusion probabilities $(\pi_k)_{k\in V}$ are maintained exactly.

The measure of entropy associated with a sampling design $p(.)$ defined on population U is given by

$$I(p) = -\sum_{s\subset U} p(s)\log(p(s)),$$

with the convention $0\log(0) = 0$. Deville and Tillé (2005) have shown that the balanced design with maximum entropy compared with other sampling designs balanced on the same variables and with the same inclusion probabilities can be regarded as the conditional of a Poisson design. Assuming the asymptotic normality of a multivariate Horvitz-Thompson estimator in the case of a Poisson design, they derived a variance approximation formula for the Horvitz-Thompson estimator for a balanced sampling design. In the case of stratified balanced sampling, we have

$$\text{Var}(\hat{t}_{y\pi}) \simeq \sum_{h=1}^{H}\sum_{k\in U_h}\frac{b_k}{\pi_k^2}(y_k - \beta_{\mathbf{h}}\mathbf{x}_k)^2 \quad (8)$$

where $\beta_{\mathbf{h}} = (\sum_{l\in U_h} b_l\ \mathbf{x}_l/\pi_l\ \mathbf{x}_l'/\pi_l)^{-1}\sum_{l\in U_h} b_l\ \mathbf{x}_l/\pi_l\ y_l/\pi_l$. Deville and Tillé (2005) offer several approximations for the b_k. The simplest is $b_k = \pi_k(1 - \pi_k)$. The variance of the Horvitz-Thompson estimator will be small if, in each stratum, variable of interest y is well explained by balancing variables $\mathbf{x}$.

Sampling will be balanced in each stratum if the number of balancing variables remains small relative to the sample size. In some cases, however, the allocation to each stratum is too small for balanced sampling: if the stratification of the population is very granular, a current practice is to select a size-2 sample in each stratum. In that case, the only condition that can be imposed is a fixed sample size in each stratum.

In the next section, we propose an algorithm based on the cube method that ensures balanced sampling across the entire population for selected variables and exactly maintains the desired allocation within each stratum. Hence, the samples are no longer selected independently in each stratum. Precision is improved in comparison with stratified sampling with fixed sample size in each stratum if the balancing variables are strongly correlated with the variable of interest across the entire population. The algorithm also has the advantage of ensuring approximate balancing in each stratum, and the larger the sample size allocated to the stratum, the more balanced the sampling will be.

2. Stratified balanced sampling with pooling of landing phases

If sample S is selected from U in accordance with the stratified balanced sampling procedure described in section 1, sampling will be balanced in each stratum as long as the landing phase affects a small number of individuals relative to the sample size. Specifically, equation (7) shows that the number of balancing variables must be small relative to the sample allocation in each stratum. In some cases, that constraint cannot be satisfied. The population is often partitioned into very small groups to make the results more relevant, which means decreasing the sample selected in each stratum; the limit generally used is a size-2 sample, which produces an unbiased variance estimator.

Again, we take the case of a population U divided into H strata $U_1, ..., U_H$, for which a vector $\mathbf{x}_k = (\pi_k, \mathbf{z}_k')'$ of auxiliary variables is known. We assume that the variable π_k is one of the balancing constraints, to ensure fixed-size sampling. Where the allocation to each stratum is too small for balanced sampling to apply constraints other than fixed size in each stratum, algorithm 1 provides an alternative sampling method. A flight phase is carried out independently in each of the H strata: we write $\boldsymbol{\pi}^{\mathbf{h}*} = (\pi_k^*)_{k\in U_h}$, $h = 1, ... H$ for the probability vectors obtained at the end of those flight phases, $\boldsymbol{\pi}^* = (\pi_k^*)_{k\in V}$, where V denotes the units that have not yet been sampled or rejected, and $\mathbf{x}_k^* = (\pi_k^* 1_{k\in U_1}, \ldots, \pi_k^* 1_{k\in U_H}, \mathbf{z}_k' \ \pi_k^*/\pi_k)'$. The probability vector obtained after a final flight phase over the set of remaining units is written $\boldsymbol{\pi}^{**} = (\pi_k^{**})_{k\in V}$. The set of units in stratum U_h that have not yet been sampled or rejected at the end of this new flight phase is denoted W_h.

Algorithm 1: Stratified balanced sampling with pooling of landing phases

Step 1. Carry out a flight phase, with balancing variables $\mathbf{x}_k$ and inclusion probabilities π_k, independently in each stratum U_h.

Step 2. Carry out a flight phase, with balancing variables $\mathbf{x}_k^*$ and inclusion probabilities π_k^*, on the set V of units remaining at the end of step 1.

Step 3. Select a fixed-size sample from each subpopulation W_h, with inclusion probabilities π_k^{**}.

The algorithm is based on a method used by the Institut National de la Statistique et des Études Économiques (INSEE) to select the primary units of the 1999 Master Sample. The Master Sample is a sample of dwellings selected in the 1999 Census for use as a sample frame for household surveys. A detailed description of the sampling design for the Master Sample is provided in Bourdalle, Christine and Wilms (2000). The dwellings are first grouped into urban units and rural units. In the subpopulation of units with fewer than 100,000 residents, a sample of about 6% is selected. We have four auxiliary variables (taxable net income and three age groups). The expected number of sample units is too small for stratified sampling by region, with balanced sampling on the four variables in each region. The regions were therefore grouped into eight super-regions, and the sampling processes were coordinated in such a way as to ensure both overall balanced sampling for the four auxiliary variables in each super-region and a fixed sample size in each region.

A similar method was proposed by Rousseau and Tardieu (2004) for the selection of balanced samples from large frames using the CUBE macro available on INSEE's Web site. The macro's run time is approximately proportional to the square of the population size. Note that Chauvet and Tillé (2006) proposed a fast method of balanced sampling whose run time depends only on the size of the population and which can select balanced samples directly from very large populations. The algorithm was programmed into an SAS macro (see Chauvet and Tillé, 2005) and is also available in the R Sampling Package prepared by Matei and Tillé (2006). In both programs, the second flight phase is performed by adding a constraint associated with each stratum to balancing variables $\mathbf{x}_k^*$ and maintaining the fixed-size condition in each stratum.

Using inclusion probabilities vector $\boldsymbol{\pi}^*$ conditionally on the outcome of step 1 ensures that inclusion probabilities vector $\boldsymbol{\pi}$ is maintained by deconditioning from the outcome of step 1. At the end of step 1, equation (6) implies that

$$\forall h = 1 \ldots H \quad \sum_{k\in U_h/0<\pi_k^*<1} \frac{\mathbf{x}_k}{\pi_k}\pi_k^* = \sum_{k\in U_h}\mathbf{x}_k - \sum_{k\in U_h/\pi_k^*=1}\frac{\mathbf{x}_k}{\pi_k},$$

and summing these expressions yields

$$\sum_{k\in V}\frac{\mathbf{x}_k}{\pi_k}\pi_k^* = \sum_{k\in U}\mathbf{x}_k - \sum_{k\in U/\pi_k^*=1}\frac{\mathbf{x}_k}{\pi_k}.$$

At the end of step 2, equation (6) leads to

$$\sum_{k\in V}\frac{\mathbf{x}_k}{\pi_k}\pi_k^{**} = \sum_{k\in V}\frac{\mathbf{x}_k}{\pi_k}\pi_k^*,$$

and combining the last two expressions, we get

$$\sum_{k\in V}\frac{\mathbf{x}_k}{\pi_k}\pi_k^{**} + \sum_{k\in U/\pi_k^*=1}\frac{\mathbf{x}_k}{\pi_k} = \sum_{k\in U}\mathbf{x}_k, \qquad (9)$$

which ensures that balanced sampling on the variables $\mathbf{x}_k$ is maintained exactly at the end of step 2. Step 3 completes the sampling process while maintaining the fixed-size constraint within each stratum U_h and can be carried out by means of a linear program to limit the lack of balance (see Deville and Tillé 2004).

The variance can be approximated with the variance formula proposed by Deville and Tillé (2005), if each flight phase in algorithm 1 is carried out with high entropy. Entropy can be increased substantially by performing a random sort on the population prior to sampling. In this case, the balancing variables are both the $\mathbf{z}_k$ variables and the variables given by the product of the inclusion probabilities and the stratum membership indicators, which ensure a fixed sample size in each stratum. We have

$$\mathrm{Var}(\hat{t}_{y\pi}) \simeq \sum_{k\in U}\frac{b_k}{\pi_k^2}(y_k - \gamma'\mathbf{a}_k)^2 \qquad (10)$$

with $\mathbf{a}_k = (\pi_k 1_{k\in U_1}, \ldots, \pi_k 1_{k\in U_H}, \mathbf{z}_k')'$ and

$$\gamma = \left(\sum_{l\in U} b_l \frac{\mathbf{a}_l}{\pi_l}\frac{\mathbf{a}_l'}{\pi_l}\right)^{-1}\sum_{l\in U} b_l \frac{\mathbf{a}_l}{\pi_l}\frac{y_l}{\pi_l}.$$

We can use the variance estimator

$$v(\hat{t}_{y\pi}) = \sum_{k\in S}\frac{b_k}{\pi_k^3}(y_k - \hat{\gamma}'\mathbf{a}_k)^2 \qquad (11)$$

proposed by Deville and Tillé (2005, page 578), with

$$\hat{\gamma} = \left(\sum_{l\in S}\frac{b_l}{\pi_l}\frac{\mathbf{a}_l}{\pi_l}\frac{\mathbf{a}_l'}{\pi_l}\right)^{-1}\sum_{l\in S}\frac{b_l}{\pi_l}\frac{\mathbf{a}_l}{\pi_l}\frac{y_l}{\pi_l}.$$

As shown in the variance approximation formula (10), it is important to note that the independence of the samples from the various strata is lost with the proposed stratified balanced sampling method. The samples from strata $U_1, ..., U_H$ are coordinated to ensure overall balance across the whole population, which strips them of their independence. The Horvitz-Thompson estimator $\hat{t}^h_{y\pi}$ of total t_{yh} remains unbiased. Its approximate variance is derived from equation (10) by replacing y_k with $y_k \, 1_{k \in U_h}$, and is given by

$$\mathrm{Var}(\hat{t}^h_{y\pi}) \approx \sum_{k \in U} \frac{b_k}{\pi_k^2} (y_k \, 1_{k \in U_h} - (\gamma^h)' \mathbf{a}_k)^2 \qquad (12)$$

with $\mathbf{a}_k = (\pi_k \, 1_{k \in U_1}, ..., \pi_k \, 1_{k \in U_H}, \mathbf{z}'_k)'$ and

$$\gamma^h = \left(\sum_{l \in U} b_l \frac{\mathbf{a}_l}{\pi_l} \frac{\mathbf{a}'_l}{\pi_l} \right)^{-1} \sum_{l \in U_h} b_l \frac{\mathbf{a}_l}{\pi_l} \frac{y_l}{\pi_l}.$$

In the particular case where the inference does not apply to the entire population but to a domain D that consists of a small number of strata, balanced sampling overall on the $\mathbf{z}$ variables will be of little benefit. The variance of the Horvitz-Thompson estimator $\hat{t}^D_{y\pi}$ of total t^D_y for variable y for that domain will be close to the variance for stratified sampling, which is given by equation (1).

3. Quantitative results

In this section, we carry out a brief simulation study to test the performance of our sampling algorithm. First, we generate a finite population of 1,000, partitioned into 25 strata of equal size containing four variables: two variables of interest, y_1 and y_2; and two auxiliary variables, x_1 and x_2. Variables x_1 and x_2 are generated with a gamma distribution with parameters 4 and 25. Variable y_1 is generated within stratum U_h using the model

$$y_1 = \alpha_{1h} + \varepsilon_h. \qquad (13)$$

The ε_h are generated with a normal distribution with mean 0 and variance σ_h^2. The model used to generate the values of y_1 is given by (13), with $\alpha_{1h} = 20\,h$ and variance σ_h^2 selected to produce a coefficient of determination R^2 approximately equal to 0.60 in each stratum. Variable y_2 is generated with the model

$$y_2 = \alpha_2 + \beta_2 \, x_1 + \gamma_2 \, x_2 + \eta. \qquad (14)$$

The η are generated with a normal distribution with mean 0 and variance ρ^2. The model used to generate the values of y_2 is given by (14), with $\alpha_2 = 500$, $\beta_2 = \gamma_2 = 5$, and variance ρ^2 selected to produce a coefficient of determination R^2 approximately equal to 0,60.

We are interested in estimating the total of variables y_1 and y_2. We select a sample of $n = 25$ ($n = 50$ respectively) units with equal probabilities using three sampling designs:

Design 1: Stratified simple random sampling in each stratum
Design 2: Sampling balanced on variables π, x_1 and x_2
Design 3: Stratified sampling balanced on variables π, x_1 and x_2, with pooling of the landing phases

In the case of stratified sampling, we have an allocation of size 1 (2 respectively) in each stratum. In the balanced designs, each flight phase is preceded by a random sort of the population. The variance associated with design 1 is calculated directly. The variance associated with designs 2 and 3 is approximated on the basis of 10,000 simulations. The results are presented in Table 1.

Table 1
Variance associated with the estimate of the total of two variables for a stratified design, a balanced design and a stratified balanced design with pooling of landing phases

	***n* = 25**		***n* = 50**	
	Total var.	**Total var.**	**Total var.**	**Total var.**
Method	y_1 **(×10⁸)**	y_2 **(×10⁹)**	y_1 **(×10⁸)**	y_2 **(×10⁹)**
Design 1	6.05	7.13	2.95	3.48
Design 2	14.31	3.05	7.02	1.40
Design 3	6.00	3.63	2.98	1.54

In each case, the proposed sampling design is comparable with the better of the two strategies. If the variable of interest is approximately constant across all strata, the proposed algorithm produces the same results as the stratified design. If the balancing variables are highly explanatory, the results produced by our algorithm and by direct balanced sampling are equivalent. The slight loss of precision comes from the landing phase: in the case of direct balanced sampling, we attempt to complete the sampling while limiting the lack of balance. With the proposed algorithm, the selected solution is suboptimal because we are imposing the additional constraint of a fixed size in each stratum.

In the case of stratified balanced sampling with pooling of the landing phases, Table 2 shows the variance given by 10,000 simulations and the variance given by the approximation formula (10).

Table 2
Comparison of the variance given by 10,000 simulations and the variance given by the approximation formula in the case of the estimation of two totals for a stratified balanced sampling design with pooling of landing phases

	***n* = 25**		***n* = 50**	
	Total y_1 **(×10⁸)**	**Total** y_2 **(×10⁹)**	**Total** y_1 **(×10⁸)**	**Total** y_2 **(×10⁹)**
Simulation var.	6.0	3.6	3.0	1.5
Approximated var.	5.9	2.7	2.9	1.3

The approximation formula proposed by Deville and Tillé (2005) is close to exact if the variance associated with the landing phase is small relative to the variance associated with the flight phase. In the case of the y_2 variable, the balancing variables are highly explanatory. The variance is therefore larger for the landing phase than for the flight phase, and the approximation formula understates the actual variance. The variance associated with the landing phase will be considered in future studies.

Acknowledgements

The author is grateful to the referees and an associate editor for their constructive comments and suggestions.

References

Bourdalle, G., Christine, M. and Wilms, L. (2000). Échantillons maître et emploi. Série INSEE Méthodes, Paris, France, 21, 139-173.

Chauvet, G., and Tillé, Y. (2005). New SAS macros for balanced sampling. INSEE, Journées de Méthodologie Statistique, Paris.

Chauvet, G., and Tillé, Y. (2006). A fast algorithm of balanced sampling. *Computational Statistics*, 21, 53-61.

Deville, J.-C., and Tillé, Y. (2004). Efficient balanced sampling: The cube method. *Biometrika*, 91, 893-912.

Deville, J.-C., and Tillé, Y. (2005). Variance approximation under balanced sampling. *Journal of Statistical Planning and Inference*, 128, 569-591.

Matei, A., and Tillé, Y. (2006). The R 'sampling' package. *European Conference on Quality in Survey Statistics*, Cardiff.

Rousseau, S., and Tardieu, F. (2004). *La macro SAS CUBE d'échantillonnage équilibré - Documentation de l'utilisateur*. Technical report, INSEE, France.

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 24, No. 4, 2008

Assessing the Value of Bayesian Methods for Inference about Finite Population Quantities
Joseph Sedransk 495

Discussion
Nathaniel Schenker, Trivellore E. Raghunathan 507

Discussion
David A. Binder 513

Model Averaging Methods for Weight Trimming
Michael R. Elliott 517

A Note on the Asymptotic Equivalence of Jackknife and Linearization Variance Estimation for the Gini Coefficient
Yves G. Berger 541

On Some Common Practices of Systematic Sampling
Li-Chun Zhang 557

Does a Final Coverage Check Identify and Reduce Census Coverage Errors?
Elizabeth Martin, Don A. Dillman 571

Privacy Concerns, Too Busy, or Just Not Interested: Using Doorstep Concerns to Predict Survey Nonresponse
Nancy Bates, James Dahlhamer, Eleanor Singer 591

Seasonality in Investment Plans and Their Revisions
Alex Teterukovsky 612

Editorial Collaborators 623

Contents
Volume 25, No. 1, 2009

Model Averaging Methods for Weight Trimming in Generalized Linear Regression Models
Michael R. Elliott 1

The Use of Sample Weights in Hot Deck Imputation
Rebecca R. Andridge, Roderick J. Little 21

Small Area Estimation in the Presence of Correlated Random Area Effects
Monica Pratesi, Nicola Salvati 37

Nonparametric Variance Estimation for Nearest Neighbor Imputation
Jun Shao 55

On the Inter-Regional Mover Problem in Panel Household Surveys
Takis Merkouris 63

Compensating for Noncoverage of Nontelephone Households in Random-Digit-Dialing Surveys:
A Comparison of Adjustments Based on Propensity Scores and Interruptions in Telephone Service
K.P. Srinath, Martin R. Frankel, David C. Hoaglin, Michael P. Battaglia 77

Population Shifts and Demographic Methods
C. Matthew Snipp, Juanita Tamayo Lott 99

Privacy-Preserving Analysis of Vertically Partitioned Data Using Secure Matrix Products
Alan F. Karr, Xiaodong Lin, Ashish P. Sanil, Jerome P. Reiter 125

Using a Weighted Average of Base Period Price Indexes to Approximate a Superlative Index
Janice Lent, Alan H. Dorfman 139

Book and Software Reviews
Jana Asher, Trent D. Buskirk, John Hall, Gordon Willis 151

In Other Journals 165

All inquires about submissions and subscriptions should be directed to jos@scb.se

CONTENTS TABLE DES MATIÈRES

Volume 36, No. 3, September/septembre 2008

Christopher A. FIELD
Modeling biological data: several vignettes 341

Marco A.R. FERREIRA & Marc A. SUCHARD
Bayesian analysis of elapsed times in continuous-time Markov chains 355

Amélie FILS-VILLETARD, Armelle GUILLOU & Johan SEGERS
Projection estimators of Pickands dependence functions 369

Simon GUILLOTTE & François PERRON
A Bayesian estimator for the dependence function of a bivariate extreme-value distribution....383

Jesúus LÓPEZ-FIDALGO, Raul MARTÍN-MARTÍN & Douglas P.WIENS
Marginally restricted sequential D-optimal designs 397

Kerrie P. NELSON & Don EDWARDS
On population-based measures of agreement for binary classifications 411

Lei NIE, Haitao CHU & Valeriy R. KOROSTYSHEVSKIY
Bias reduction for nonparametric correlation coefficients under the bivariate normal copula
assumption with known detection limits 427

Jean-François PLANTE
Nonparametric adaptive likelihood weights 443

Yichuan ZHAO & Hongkun WANG
Empirical likelihood inference for the regression model of mean quality-adjusted lifetime with censored data....463

Volume 36, No. 4, December/décembre 2008

Mylène BÉDARD & Jeffrey S. ROSENTHAL
Optimal scaling of Metropolis algorithms: heading toward general target distributions 483

Jörn DANNEMANN & Hajo HOLZMANN
Testing for two states in a hidden Markov model....505

Christopher A. FIELD, Zhen PANG & Alan H. WELSH
Bootstrapping data with multiple levels of variation 521

Feifang HU, Li-Xin ZHANG, Siu H. CHEUNG & Wai S. CHAN
Doubly adaptive biased coin designs with delayed responses 541

Sai Man Simon KWOK & Wai Keung LI
On diagnostic checking of the autoregressive conditional intensity model 561

Omer OZTURK
Inference in the presence of ranking error in ranked set sampling 577

Matías SALIBIÁN-BARRERA & Ying WEI
Weighted quantile regression with nonelliptically structured covariates....595

Ana-Maria STAICU & Nancy M. REID
On probability matching priors....613

Robin WILLINK
Shrinkage confidence intervals for the normal mean: using a guess for greater efficiency 623