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

Survey Methodology
44-1

SURVEY
METHODOLOGY

A JOURNAL
PUBLISHED BY
STATISTICS CANADA

Canada

Release date: June 21, 2018

Statistics Canada
Statistique Canada

Canada

How to obtain more information

For information about this product or the wide range of services and data available from Statistics Canada, visit our website, www.statcan.gc.ca.

You can also contact us by

email at STATCAN.infostats-infostats.STATCAN@canada.ca

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

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

Depository Services Program

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

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 “Contact us” > “Standards of service to the public.”

Note of appreciation

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

Standard table symbols

The following symbols are used in Statistics Canada publications:

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

Published by authority of the Minister responsible for Statistics Canada

© Her Majesty the Queen in Right of Canada as represented by the Minister of Industry, 2018

All rights reserved. Use of this publication is governed by the Statistics Canada Open Licence Agreement.

An HTML version is also available.

Cette publication est aussi disponible en français.

Survey Methodology

Catalogue No. 12-001-XPB

A journal published by Statistics Canada

June 2018

•

Volume 44

•

Number 1

Statistics Canada Statistique Canada

Canada

SURVEY METHODOLOGY

A Journal Published by Statistics Canada

Survey Methodology is indexed in The ISI Web of knowledge (Web of science), The Survey Statistician, Statistical Theory and Methods Abstracts and SRM Database of Social Research Methodology, Erasmus University and is referenced in the Current Index to Statistics, and Journal Contents in Qualitative Methods. It is also covered by SCOPUS in the Elsevier Bibliographic Databases.

MANAGEMENT BOARD

Chairman C. Julien

Past Chairmen J. Kovar (2009-2013)
D. Royce (2006-2009)
G.J. Brackstone (1986-2005)
R. Platek (1975-1986)

Members G. Beaudoin
S. Fortier (Production Manager)
W. Yung

EDITORIAL BOARD

Editor W. Yung, *Statistics Canada*

Past Editor M.A. Hidiroglou (2010-2015)
J. Kovar (2006-2009)
M.P. Singh (1975-2005)

Associate Editors

J.-F. Beaumont, *Statistics Canada*
M. Brick, *Westat Inc.*
P. Brodie, *Office for National Statistics*
P.J. Cantwell, *U.S. Bureau of the Census*
J. Chipperfield, *Australian Bureau of Statistics*
J. Dever, *RTI International*
J.L. Eltinge, *U.S. Bureau of Labor Statistics*
W.A. Fuller, *Iowa State University*
J. Gambino, *Statistics Canada*
D. Haziza, *Université de Montréal*
M.A. Hidiroglou, *Statistics Canada*
B. Hulliger, *University of Applied Sciences Northwestern Switzerland*
D. Judkins, *Abt Associates*
J. Kim, *Iowa State University*
P. Kott, *RTI International*
P. Lahiri, *JPSM, University of Maryland*
P. Lavallée, *Statistics Canada*
I. Molina, *Universidad Carlos III de Madrid*
J. Opsomer, *Colorado State University*
D. Pfeffermann, *Hebrew University*
J.N.K. Rao, *Carleton University*
L.-P. Rivest, *Université Laval*
F. Scheuren, *National Opinion Research Center*
P.L.N.D. Silva, *Escola Nacional de Ciências Estatísticas*
P. Smith, *University of Southampton*
D. Steel, *University of Wollongong*
M. Thompson, *University of Waterloo*
D. Toth, *U.S. Bureau of Labor Statistics*
J. van den Brakel, *Statistics Netherlands*
C. Wu, *University of Waterloo*
A. Zaslavsky, *Harvard University*
L.-C. Zhang, *University of Southampton*

Assistant Editors C. Bocci, K. Bosa, C. Boulet, H. Mantel, S. Matthews, C.O. Nambeu, Z. Patak and Y. You, *Statistics Canada*

EDITORIAL POLICY

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

Submission of Manuscripts

Survey Methodology is published twice a year in electronic format. Authors are invited to submit their articles in English or French in electronic form, preferably in Word to the Editor, (statcan.smj-rte.statcan@canada.ca, Statistics Canada, 150 Tunney's Pasture Driveway, Ottawa, Ontario, Canada, K1A 0T6). For formatting instructions, please see the guidelines provided in the journal and on the web site (www.statcan.gc.ca/surveymethodology).

Survey Methodology

A Journal Published by Statistics Canada

Volume 44, Number 1, June 2018

Contents

Regular papers

Model based inference using ranked set samples

Omer Ozturk and Konul Bayramoglu Kavlak[1]

Abstract

This paper develops statistical inference based on super population model in a finite population setting using ranked set samples (RSS). The samples are constructed without replacement. It is shown that the sample mean of RSS is model unbiased and has smaller mean square prediction error (MSPE) than the MSPE of a simple random sample mean. Using an unbiased estimator of MSPE, the paper also constructs a prediction confidence interval for the population mean. A small scale simulation study shows that estimator is as good as a simple random sample (SRS) estimator for poor ranking information. On the other hand it has higher efficiency than SRS estimator when the quality of ranking information is good, and the cost ratio of obtaining a single unit in RSS and SRS is not very high. Simulation study also indicates that coverage probabilities of prediction intervals are very close to the nominal coverage probabilities. Proposed inferential procedure is applied to a real data set.

Key Words: Ranked set sampling; Finite population; Mean square prediction error; Sampling cost model; Coherent ranking; Concomitant ranking; Visual ranking.

1 Introduction

In many survey sampling studies, it is very common that the sampling frame has additional auxiliary information in addition to characteristic of interest. Under a fairly strong modeling assumption, this auxiliary information improves the statistical inference. For example, ratio and regression estimators use covariate information under a linearity assumption to estimate the population mean or total. The auxiliary information can also be used under a weaker assumption in a ranked set sample (RSS) and judgment post stratified (JPS) sample. These samples use auxiliary information to increase the information content of each measured unit through a ranking process. The ranking process is performed in a small set of size H formed by combining the measured unit with an additional $H-1$ unmeasured units from the population. Ranking process is performed either before or after measurement and determines the relative position of each measured unit. Ranking information can be obtained from either a visual inspection or some other form of ranking process. A reasonable ranking mechanism requires some sort of monotonic relationship between the ranking variable and response, which is much weaker than the strong linearity assumption of regression and ratio estimators.

A balanced ranked set sample of set size H and cycle size d can be constructed by first selecting $n = Hd$ simple random samples of size H from the population and ranking the units in each sample without measurement from smallest to largest. In these n ranked sets (samples), one then measures the units with rank 1 in the first d sets, the unit with rank 2 in the next d sets and so on. This yields samples of H different sets of judgment order statistics, each of which has d independent and identically distributed judgment order statistics.

A sharp contrast exists between an observation from SRS and RSS, where the observation from an SRS sample provides information only about the unit on which it was measured while the observation from an RSS sample, in addition to the information that the measured unit provides, also provides limited

[1] Omer Ozturk, Department of Statistics, The Ohio State University, 1958 Neil Avenue, Columbus, OH, 43210, U.S.A. E-mail: omer@stat.osu.edu; Konul Bayramoglu Kavlak, Department of Actuarial Sciences, Hacettepe University, Ankara, Turkey. E-mail: konul.bayramoglu@hacettepe.edu.tr.

information about the other $(H-1)$ unmeasured units in the set through the relative position (rank) of measured unit. Since ranking process does not require a formal measurement and is usually less expensive in comparison with formal measurement, the RSS sample provides substantial amount of reduction in sampling cost.

A JPS sample differs from an RSS sample in that the ranking step comes after the construction of an SRS sample. Construction of a JPS sample of size n requires a set size H. Once the set size H is determined, one first draws a simple random sample of size n and makes a measurement on each of the n units. For each measured unit in the sample, one then selects additional $H-1$ units to form a set of size H. The units in this set are ranked from smallest to largest without measurement and the rank of the measured unit in the set is recorded. The JPS sample then consists of n measured values, together with their ranks.

Both RSS and JPS samples induces a stochastic structure among measured units in which observations in judgment class h are usually smaller than the observations in judgment class h', $h < h'$. This stochastic ordering feature spreads the measured units in the support of the distribution and creates a better representative sample than a simple random sample. The nature of stochastic ordering in a JPS sample is significantly different from the stochastic ordering in an RSS sample. A JPS sample consists of a simple random sample and an associated rank vector. This rank vector is loosely related to the sample and may be ignored if desired. On the other hand, an RSS sample is measured as judgment order statistics, judgment ranks can not be separated from the observed values. An RSS sample can not be treated as an SRS sample.

Both JPS and RSS sampling designs have generated extensive research interest in a finite population setting. Patil, Sinha and Taillie (1995) used ranked set sample to estimate population mean for a population of size N when the sample is constructed without replacement. Takahasi and Futatsuya (1998) showed that the ranked set sample estimator of the population mean is more precise than the simple random sample estimator when samples are drawn without replacement from a finite population. Deshpande, Frey and Ozturk (2006) described three different sampling designs and constructed nonparametric confidence intervals for population quantiles. Al-Saleh and Samawi (2007), Ozdemir and Gokpinar (2007 and 2008), Gokpinar and Ozdemir (2010), Ozturk and Jozani (2013), Frey (2011) and Ozturk (2014, 2015, 2016a) computed inclusion probabilities and constructed Horwitz-Thompson type estimators for population mean and total based on a ranked set sample. These research papers show that an RSS design yields a substantial amount of improvement in efficiency over the usual simple random sampling design. Ozturk (2016b) developed estimators for population mean based on a JPS sample where he showed that the estimator needs a finite population correction factor similar to the one used in a simple random sample.

All available research in literature in JPS and RSS sampling designs in a finite population setting considers design-based approach. To our knowledge, super population model has not been used. In this paper, we develop a model-based statistical inference using RSS sampling design for population mean and total in a finite population setting. Similar results, with some additional variation due to random judgment class samples sizes, can also be established for a JPS sampling design. Because of the random judgment class sample sizes, the estimators based on a JPS sample are less efficient than the estimators based on an RSS sample. For this reason, the JPS sample is not considered further in this paper. Section 2 clearly defines the model and describes the sampling designs for RSS under super population model. We show that estimators of population mean and total are model-unbiased and their mean square prediction errors (MSPE) are smaller than the MSPE of the same estimators of an SRS sample. Section 3 constructs unbiased estimators for the MSPE and provides approximate confidence intervals for the population mean and total. Section 4 introduces cost models to account the effect of additional cost (excess of the cost of construction of SRS sample) in construction of RSS sample. Section 5 provides empirical evidence about the performance of the estimators. Section 6 applies the proposed estimators to an example in a finite population setting. Section 7 provides some concluding remarks.

2 Sampling designs

We consider RSS sampling designs from a super population model to draw statistical inference in a finite population setting. Let Y be the characteristic of interest. The copies of Y, $Y_1, \ldots, Y_N$, are considered as independent identically distributed (iid) random variables from a super population. Basic assumption for this super population model can be stated as

$$\text{Model: } Y_1, \ldots, Y_N \text{ independent identically distributed with } E_M(Y_i) = \mu, V_M(Y_i) = \sigma^2. \quad (2.1)$$

The subscript M in model (2.1) is used to highlight that the mean and variance are computed based on a super population model, not the randomization distribution as in Ozturk (2016b). In this super population model, μ and σ^2 represent unknown infinite population parameters.

In super population model, a particular realization, $y_1, \ldots, y_N$, of random variables $Y_1, \ldots, Y_N$ from model (2.1), is considered as a finite population. Let $P^N = \{y_1, \ldots, y_N\}$ denotes this finite population. Ranked set sample is constructed from P^N. Without loss of generality, we assume that $y_{(1)} < y_{(2)} < \ldots < y_{(N)}$ are ordered values of $y_1, \ldots, y_N$ where $y_{(i)}$ is the i^{th} largest value of Y in P^N. Throughout the paper, H and d are used to denote the set and cycle sizes, respectively.

To construct a ranked set sample, one selects a set of H experimental units, $y_{s_1}, \ldots, y_{s_H}$, at random from P^N and ranks them based on their Y values in an increasing magnitude without actual measurement. Ranking process can be performed either using visual inspection or some auxiliary variables and hence subjected to ranking error. The unit that corresponds to the smallest Y, $y_{[1]}$, is identified and measured where the square bracket in the subscript, [1], denotes the rank of the smallest unit (rank 1) in the set $\{y_{[1]}, y^*_{[2]}, \ldots, y^*_{[H]}\}$. The remaining unmeasured units are denoted with $\{y^*_{[2]}, \ldots, y^*_{[H]}\}$. After $y_{[1]}$ is measured, none of the H units in the set $\{y_{[1]}, y^*_{[2]}, \ldots, y^*_{[H]}\}$ are returned to the population. One then selects another set of H experimental units at random from the remaining population P^{N-H} and ranks them without measurement. This time, the unit that corresponds to the second smallest Y, $y_{[2]}$, is identified and measured in $\{y^*_{[1]}, y_{[2]}, y^*_{[3]}, \ldots, y^*_{[H]}\}$. This process is continued until a simple random sample of size H is taken from the reduced population $P^{N-H(H-1)}$ and the H^{th} smallest unit is identified and measured in the set $\{y^*_{[1]}, y^*_{[2]}, \ldots, y^*_{[H-1]}, y_{[H]}\}$. This is called a cycle. A cycle selects H disjoint sets, each of size H and only measures H units. The remaining $H(H-1)$ units are used only for ranking purposes. The cycles are repeated d times to yield a ranked set sample of size $n = dH$ units. A ranked set sample can then be represented as

$$W_{h,i,H} \;=\; \left\{y^*_{[1]i}, \ldots, y^*_{[h-1]i}, y_{[h]i}, y^*_{[h+1]i}, \ldots, y^*_{[H]i}\right\}, \;\; h = 1, \ldots, H, \;\; i = 1, \ldots, d, \qquad (2.2)$$

where only $y_{[h]i}$, $h = 1, \ldots, H$, $i = 1, \ldots, d$, are measured. The other values are used to obtain the rank of the measured values. Units in sets $W_{h,i,H}$ and $W_{h',i',H}$ are all independent if either $h \neq h'$ or $i \neq i'$, but the units in $W_{h,i,H}$ are all correlated since they are ranked in the same set. Under model (2.1), means, variances and covariances of judgment order statistics are given by

$$\begin{aligned} E_M(Y_{[h]i}) \;&=\; \mu_{[h]}, \mathrm{Var}_M(Y_{[h]i}) = \sigma^2_{[h]}, \\ \mathrm{Cov}_M(Y_{[h]i}, Y_{[h']i}) \;&=\; \begin{cases} \sigma_{[h,h']} & \text{if } Y_{[h]i}, Y_{[h']i} \text{ are from the same set} \\ 0 & \text{otherwise.} \end{cases} \end{aligned}$$

It should be noted that since all sets are disjoint no units can be used more than once in any one of the sets. Hence all sample units are distinct. Since the sets are independently ranked $Y_{[h]i}$'s are mutually independent. Observations having the same rank h, $Y_{[h]i}$, $i = 1, \ldots, d$ are identically distributed.

Estimator of the population mean μ based on RSS data in equation (2.2) can be defined as follows.

$$\bar{Y}_R \;=\; \frac{1}{dH} \sum_{h=1}^{H} \sum_{i=1}^{d} Y_{[h]i}. \qquad (2.3)$$

It can be immediately observed that the estimator $\bar{Y}_R$ is model unbiased. In other words, under the model (2.1), $E_M\left(\bar{Y}_R - \bar{Y}_N\right) = 0$, where $\bar{Y}_N = \frac{1}{N}\sum_{i=1}^{N} Y_i$.

We now consider the mean square prediction error (MSPE) of the estimator $\bar{Y}_R$ under model (2.1)

$$\mathrm{MSPE}_M(\bar{Y}_R) \;=\; E_M\left(\bar{Y}_R - \frac{1}{N}\sum_{i=1}^{N} Y_i\right)^2 = E_M\left(\bar{Y}_R - \bar{Y}_N\right)^2.$$

Since the predictor $\bar{Y}_R$ is model unbiased for $\bar{Y}_N$, $E_M\left(\bar{Y}_R - \bar{Y}_N\right) = 0$, the mean square prediction error (MSPE) of $\bar{Y}_R$ is the same as $\mathrm{Var}_M\left(\bar{Y}_R - \bar{Y}_N\right)$.

Theorem 1: *Let $Y_{[h]i}, h = 1, \ldots, H, i = 1, \ldots, d$, be a ranked set sample from a finite population P^N. Under a super population model in equation (2.1), the mean square prediction error of the estimator $\bar{Y}_R$ is given by*

$$\sigma^2_{\mathrm{RSS}} \;=\; \mathrm{MSPE}_M(\bar{Y}_R) = \frac{N-n}{Nn}\sigma^2 - \frac{1}{nH}\sum_{h=1}^{H}\left(\mu_{[h]} - \mu\right)^2. \qquad (2.4)$$

We note that expression on equation (2.4) is very similar to the sample variance of an infinite population RSS sample. Only difference is due to the coefficient $\frac{N-n}{Nn}$. In infinite population setting the fraction $\frac{N-n}{Nn}$ in equation (2.4) becomes $\frac{1}{n}$. Hence, $\left(1 - \frac{n}{N}\right)$ is the finite population correction (fpc) factor for the variance of RSS sample mean. If the sample size is not small in comparison with the population size N, the fpc, $\frac{N-n}{Nn}$, makes a correction on the variance of an RSS sample mean. This correction would be substantial if n is relatively large with respect to N. If n is small, fpc is close to 1 and the impact of finite population correction factor is minimal.

Corollary 1: *Assume that* n *and* N *increase in such a way that the ratio* $\frac{n}{N}$ *approaches to a limit at* a, $\lim_{n\to\infty} \frac{n}{N} = a$.

(i) *If* $a > 0$, σ_{RSS}^2 *converges to a simple form*

$$\lim_{n\to\infty} n\sigma_{RSS}^2 = (1-a)\,\sigma^2 - \frac{1}{H}\sum_{h=1}^{H}\left(\mu_{[h]} - \mu\right)^2,$$

(ii) *if* $a = 0$, $\lim_{n\to\infty} n\sigma_{RSS}^2 = \frac{1}{H}\sigma_{[h]}^2$, *which is the same as the variance of the sample mean of a balanced ranked set sample in an infinite population setting,*

(iii) *if* a *is strictly positive, then* $\lim_{n\to\infty} n\sigma_{RSS}^2 < \frac{1}{H}\sigma_{[h]}^2$.

The corollary indicates that when sample and population sizes grow at a certain rate, variance of sample mean of an RSS (σ_{RSS}^2) sample in a finite population setting reduces to simple form. If a is strictly positive, variance of an RSS sample mean is smaller than the variance of an RSS sample mean in an infinite population setting.

3 Unbiased estimators

In this section, we construct an unbiased estimator for σ_{RSS}^2. By rewriting the estimator for σ_{RSS}^2 in a slightly different form, we obtain

$$\begin{aligned}
\sigma_{\text{RSS}}^2 &= \left(\frac{N-n}{Nn}\right)\sigma^2 - \frac{1}{nH}\sum_{h=1}^{H}\left(\mu_{[h]} - \mu\right)^2 \\
&= \left(\frac{1}{n} - \frac{1}{N}\right)\sigma^2 - \frac{1}{nH}\left(H\sigma^2 - \sum_{h=1}^{H}\sigma_{[h]}^2\right) \\
&= \left(\frac{-1}{N}\right)\sigma^2 + \frac{1}{nH}\sum_{h=1}^{H}\sigma_{[h]}^2.
\end{aligned}$$

Let

$$\begin{aligned}
T_1^* &= \frac{1}{2d^2H^2}\sum_{h=1}^{H}\sum_{h\neq h'}^{H}\sum_{i=1}^{d}\sum_{j=1}^{d}\left(Y_{[h]i} - Y_{[h']j}\right)^2 \\
T_2^* &= \frac{1}{2d\,(d-1)\,H^2}\sum_{h=1}^{H}\sum_{i=1}^{d}\sum_{j\neq i}^{d}\left(Y_{[h]i} - Y_{[h]j}\right)^2.
\end{aligned}$$

Using these definitions, one can easily establish the following result.

Theorem 2: *Let* $Y_{[h]i}$, $i = 1, \ldots, n, h = 1, \ldots, H$ *be an RSS sample of set size* H *from a finite population. An unbiased estimator of* σ_{RSS}^2 *is given by*

$$\hat{\sigma}_{RSS}^2 = T_2^*\left(\frac{H}{n}\right) - \left(T_1^* + T_2^*\right)\frac{1}{N}. \tag{3.1}$$

Theorem 2 indicates that the variance estimator is unbiased for any sample and set sizes regardless of the quality of ranking information. Unbiased estimator of the variance of $\bar{Y}_R$ allows us to construct confidence interval for population mean and total. Using normal approximation, $(1-\alpha)100\%$ confidence interval for the population mean is given by

$$\bar{Y}_R \pm t_{n-H,\alpha/2}\hat{\sigma}^2_{\text{RSS}},$$

where $t_{df,a}$ is the a^{th} upper quantile of $t-$distribution with degrees of freedom df. The degrees of freedom $n-H$ is suggested to account the heterogeneity among H judgment classes. The choice of $df = n-H$ is also suggested in Ahn, Lim and Wang (2014) in infinite population setting.

4 Cost model

Efficiency improvement of the RSS estimator results from the relative position (rank) information of the measured observation among unmeasured $H-1$ units in a set. This extra information comes at the cost of sampling a set of size H and obtaining the subsequent ranking. Ranking can be performed either using concomitant (auxiliary) variable or visual inspection of the physical units in each set. Hence, these two approaches, visual and concomitant ranking models, may lead to different cost structures. In either case, there needs to be some sort of consistency in ranking process to develop a meaningful cost function. Patil, Sinha and Taillie (1997) defined a coherent ranking process in which ranking of a set is consistent for all subsets and supersets. Under a coherent ranking scheme, the rank order of H units would remain identical when ranking any of their subsets or supersets containing them. For further detail in coherent ranking, readers are referred to Patil et al. (1997) or Nahhas, Wolfe and Chen (2002).

Concomitant ranking uses an auxiliary variable to rank H units in a set. The quality of ranking depends on monotonic (not necessarily to be linear) relationship between the variable of interest and auxiliary variable. On the other hand, visual inspection can be performed in different ways. One of the strategy is to use pairwise comparison. Under coherent ranking, not all $\binom{H}{2}$ pairwise comparisons are necessary for a visual ranking. For example, in a set of size $H = 3,$ if unit 1 is judged to be smaller than unit 2 and unit 2 is smaller than unit 3, we reasonably assume unit 1 is less than unit 3 without a comparison. In order to differentiate the impact of the cost structures of the concomitant and visual ranking schemes, we denote the estimator in equation (2.3) with $\bar{Y}_{\text{RC}}$ for concomitant ranking and $\bar{Y}_{\text{RV}}$ for visual ranking.

For visual ranking, we use visual inspection model of Nahhas et al. (2002). This model always compares the selected unit with the largest element previously ranked. It chooses a unit at random and compares it with the unit previously judged to be largest. If it is judged to be larger, then it becomes the largest among all judged units. Otherwise, it is compared with the next largest previously judged unit until it is assigned a rank. The number of required pairwise comparisons under this ranking strategy with a coherent ranking scheme is an integer valued random variable having the support $H-1, H, H+1, \ldots, \binom{H}{2}.$ The expected number of pairwise comparison for this ranking scheme is approximately equal to $f(H) = (H+2)(H-1)/4.$ The reader is referred to Nahhas et al. (2002) for further development on expected number of pairwise comparisons.

We now introduce cost definitions for three models; concomitant, visual ranking and simple random sampling models: $C_C =$ total cost for concomitant ranking, $C_V =$ total cost for visual ranking, $C_S =$ total cost for simple random sampling, $c_i =$ cost of sampling a single unit, $c_{qy} =$ cost of quantification of the variable of interest (Y) for one unit, $c_{qx} =$ cost of quantification of concomitant (auxiliary X) variable for one unit, $c_r =$ cost of one pairwise comparison. We assume that overhead cost in SRS model to be zero, but the overhead cost (in excess of the overhead cost of SRS) of RSS concomitant (visual) ranking model is absorbed in $c_{qx}(c_r)$. Total cost for these three models are then given by

$$C_S = n_s\,(c_i + c_{qy}),\ C_C = n_c\,(Hc_i + Hc_{qx} + c_{qy}),\ C_V = n_V(Hc_i + f(H) + c_{qy}),$$

where n_S, n_C and n_V are the total (measured) observations in SRS, RSS concomitant and RSS visual ranking models. Readers are referred to Nahhas et al. (2002) for further details on these cost functions.

We now fix the total cost on these three models $C_S = C_C = C_V = C.$ Under this fixed cost, we look at the relative efficiency of $\overline{Y}_{\mathrm{RC}}$ and $\overline{Y}_{\mathrm{RV}}$ with respect to SRS sample mean $\overline{Y}_{\mathrm{SRS}}$. Let

$$\mathrm{RP} = \frac{1}{1-D}, \quad D = 1 - \frac{1}{H\sigma^2}\sum_{h=1}^{H}(\mu_{[h]} - \mu)^2,$$

where RP is the relative precision of RSS sample mean with respect to SRS sample mean in an infinite population setting. Under super population model, we can establish the following efficiency result.

Theorem 3: *Let $Y_{[h]i}$, $h = 1,\dots,H$, $i = 1,\dots,d$, be a ranked set sample from a finite population P^N. For a fixed cost, under super population model and coherent ranking scheme, the following efficiency results are established.*

$$\mathrm{RE}\left(\overline{Y}_{\mathrm{RC}}, \overline{Y}_{\mathrm{SRS}}\right) = \frac{\mathrm{Var}\left(\overline{Y}_{\mathrm{SRS}}\right)}{\mathrm{Var}\left(\overline{Y}_{\mathrm{RC}}\right)} \geq 1, \quad \text{if } \mathrm{RP} \geq \frac{Hc_i + Hc_{qx} + c_{qy}}{c_i + c_{qy}}$$

$$\mathrm{RE}\left(\overline{Y}_{\mathrm{RV}}, \overline{Y}_{\mathrm{SRS}}\right) = \frac{\mathrm{Var}\left(\overline{Y}_{\mathrm{SRS}}\right)}{\mathrm{Var}\left(\overline{Y}_{\mathrm{RV}}\right)} \geq 1, \quad \text{if } \mathrm{RP} \geq \frac{Hc_i + f(H)\,c_r + c_{qy}}{c_i + c_{qy}}.$$

The fractions on the right hand side of the inequalities in the above theorem is the ratio of the cost of selecting and measuring a single unit in RSS and SRS, respectively. If the cost of sampling a unit and cost of ranking a set are negligible (free), the cost ratio becomes 1. One of the basic assumptions, in settings where RSS is used, is that ranking cost of units is relatively cheap with respect to the cost of measurement. Hence, it is not unreasonable to assume that cost ratio will be very close to 1 for settings where use of RSS is appropriate. It is established in the literature that RP is always greater than or equal to 1 (see Dell and Clutter (1972), Patil et al. (1997), Nahhas et al. (2002)). It is equal to 1 only under random ranking. The values of RP for normal population for different values of ρ (correlation coefficient between response Y and auxiliary variable X) and set sizes are given in Table 4.1. It is now reasonable to say that RSS estimator under super population model is more efficient if the cost of sampling and ranking a unit is relatively cheap in comparison with measurement cost.

Table 4.1
Relative precision (RP) of RSS sample mean with respect to SRS sample mean under infinite population setting for normal distribution $N(0,1)$. ρ is the correlation coefficient between response and auxiliary variable, and H is the set size

ρ	$H = 2$	$H = 3$	$H = 4$	$H = 5$	$H = 6$
1.00	1.467	1.914	2.347	2.770	3.186
0.90	1.347	1.631	1.869	2.073	2.251
0.75	1.218	1.367	1.477	1.561	1.628
0.50	1.086	1.136	1.168	1.190	1.207

5 Empirical results

In this section, we conduct a simulation study to check the finite sample properties of the estimator for different values of simulation parameters. Data sets are generated from normal $(\mu = 10, \sigma = 4)$ and log normal $(\mu = 0, \sigma = 1)$ super populations. We consider two different finite populations with population sizes $N = 150$ and $N = 1{,}000$ to see the impact of population sizes on the estimators. Sample and set size combinations (n, H) are selected to be (10, 2), (15, 3), (20, 4), (25, 5). The quality of ranking information is modeled through a perceptual error model in Dell and Clutter (1972). The Dell and Clutter model considers two variables, the variable of interest Y and a correlated ranking variable X. The ranking variable is modeled through an additive model $X = Y + \epsilon,$ where ϵ is a random noise generated independently with respect to Y. To implement the perceptual error model, we generate a set (size H) of simple random sample, $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_H)$, from the true population of interest with mean μ and variance σ^2. Another set (size H) of random numbers are generated from a normal distribution with mean zero and variance, σ_ϵ^2, $\boldsymbol{\epsilon} = (\epsilon_1, \epsilon_2, \ldots, \epsilon_H)$. The perceptual error model is then defined by $X_i = Y_i + \epsilon_i$ $i = 1, 2, \ldots, H$. The random numbers (X_i, Y_i) are ranked with respect to the first components $(X_{(i)})$ and the second components are taken to be the judgment ranked order statistics $(Y_{[i]})$. The quality of the ranking information is controlled by the correlation coefficient between Y and X, $\rho = \text{corr}(Y, X) = \left(\frac{\sigma^2}{\sigma^2 + \sigma_\epsilon^2}\right)^{1/2}$. Since the units are ranked based on concomitant variable X, the ranking model is equivalent to concomitant ranking in Section 3. In the simulation study, we used $\rho = 1$ for perfect ranking and $\rho = 0.75, 0.50$ for imperfect ranking.

In each replication of the simulation, a finite population of size P^N is generated from the normal super population with specified mean μ and standard deviation σ, $P^N = \{y_1, \ldots, y_N\}$. A ranked set sample is then constructed from this finite population, a realization from normal super population, with specified set and cycle sizes. The quality of ranking information in each RSS sample is controlled generating random noise vector $\boldsymbol{\epsilon}$ with specified ρ (or equivalently σ_ϵ) in the perceptual error model. The simulation size is taken to be 50,000.

Simulation results are presented in Tables 5.1, 5.2, 5.3 and 5.4. There are several features that need to be discussed in these tables. For different ρ and sample size combinations (n, H), the relative efficiencies of the RSS estimator with respect to the SRS estimator are given by

$$\text{RE}_{\text{RC}} = \frac{V\left(\bar{Y}_{\text{SRS}}\right)}{V\left(\bar{Y}_{\text{RC}}\right)} \tag{5.1}$$

where $V\left(\overline{Y}_{\text{SRS}}\right)$ and $V\left(\overline{Y}_{\text{RC}}\right)$ are the MSPE of SRS and RSS sample means from the simulation study under a super population model in equation (2.1), respectively. In equation (5.1), the relative efficiency values $\left(\text{RE}_{\text{RC}}\right)$ greater than one indicate that the RSS estimator is more efficient than the SRS estimator. In all these tables, the RSS sample mean estimator performs better than the SRS estimator. Its efficiency is an increasing function of set size H and correlation coefficient ρ as expected. Under the concomitant cost model, if the cost ratio of obtaining a unit in RSS and a unit in SRS is less than the RP values in Table 4.1, the RSS sample mean has higher efficiency than the SRS sample mean.

The impact of the finite population size N can be observed by comparing the efficiency results in Tables 5.1 and 5.2 for the normal super population and Tables 5.3 and 5.4 for the lognormal super population. When $\rho > 0.50$, relative efficiencies $\left(\text{RE}_{\text{RC}}\right)$ are higher in Table 5.1 $(N = 150)$ than Table 5.2 $(N = 1{,}000)$. In Table 5.1, finite population correction factor is smaller than the finite population correction factor in Table 5.2. Hence, the reduction in MSPE is smaller in RSS estimator. Similar effect is also observed in Tables 5.3 and 5.4.

The simulation study also investigated the properties of the MSPE estimator of RSS sample mean estimator. Theoretical value of the MSPE estimator is given under the heading σ^2_{RSS} when $\rho = 1.0$. The simulated (unbiased) MSPE estimate is given in columns 5 (6) in Tables 5.1-5.4. It is very clear that simulated and unbiased MSPE estimates are almost identical when $\rho \neq 1$ as expected. Under perfect ranking $(\rho = 1)$ theoretical MSPE values, and the simulated and unbiased MSPE estimates are all close to each other within the simulation variation.

The coverage probabilities of the confidence intervals are given under the heading $C\left(\overline{Y}_{\text{RC}}\right)$ in column 7 in Tables 5.1-5.4. In Tables 5.1 and 5.2, the coverage probabilities of the confidence intervals based on $t-$ approximation are reasonably close to the nominal coverage probability 0.950. On the other hand, the coverage probabilities in Tables 5.3 and 5.4 are smaller than the nominal coverage probability 0.95 for lognormal super population. The coverage probabilities are getting closer to nominal values when the sample size increases. This indicates that for skewed populations, sample sizes should be large enough to have a reasonable coverage probability for the confidence intervals.

Table 5.1
MSPE estimate and relative efficiency of RSS sample estimator, and coverage probability of a 95% confidence interval of population mean. Data sets are generated from a normal super population with $\mu = 10$, $\sigma = 4$ and population size $N = 150$

		Est. from equations		**Est. from simu.**	**UE estimates**	**Coverage prb.**	**Relative eff.**
H	ρ	σ^2_{RSS}	σ^2_{SRS}	$V\left(\overline{Y}_{\text{RC}}\right)$	$\hat{\sigma}^2_{\text{RSS}}$	$C\left(\overline{Y}_{\text{RC}}\right)$	RE_{RC}
2.0	0.50	-	1.493	1.355	1.365	0.949	1.102
3.0	0.50	-	0.960	0.840	0.833	0.947	1.143
4.0	0.50	-	0.693	0.572	0.578	0.948	1.213
5.0	0.50	-	0.533	0.435	0.432	0.948	1.226
2.0	0.75	-	1.493	1.195	1.205	0.949	1.250
3.0	0.75	-	0.960	0.675	0.674	0.947	1.423
4.0	0.75	-	0.693	0.433	0.436	0.946	1.600
5.0	0.75	-	0.533	0.302	0.304	0.945	1.768
2.0	1.00	0.984	1.493	0.974	0.984	0.948	1.534
3.0	1.00	0.451	0.960	0.455	0.451	0.940	2.111
4.0	1.00	0.234	0.693	0.233	0.235	0.936	2.971
5.0	1.00	0.124	0.533	0.125	0.126	0.922	4.273

Table 5.2
MSPE estimate and relative efficiency of RSS sample estimator, and coverage probability of a 95% confidence interval of population mean. Data sets are generated from a normal super population with $\mu = 10$, $\sigma = 4$ and population size $N = 1{,}000$

		Est. from equations		Est. from simu.	UE estimate	Coverage prb.	Relative eff.
H	ρ	σ^2_{RSS}	σ^2_{SRS}	$V(\bar{Y}_{\text{RC}})$	$\hat{\sigma}^2_{\text{RSS}}$	$C(\bar{Y}_{\text{RC}})$	RE_{RC}
2.0	0.50	-	1.584	1.461	1.455	0.950	1.084
3.0	0.50	-	1.051	0.931	0.924	0.949	1.129
4.0	0.50	-	0.784	0.665	0.670	0.950	1.180
5.0	0.50	-	0.624	0.524	0.522	0.950	1.191
2.0	0.75	-	1.584	1.304	1.295	0.949	1.215
3.0	0.75	-	1.051	0.770	0.765	0.948	1.365
4.0	0.75	-	0.784	0.525	0.526	0.951	1.494
5.0	0.75	-	0.624	0.392	0.395	0.951	1.590
2.0	1.00	1.075	1.584	1.075	1.076	0.950	1.473
3.0	1.00	0.541	1.051	0.538	0.541	0.951	1.954
4.0	1.00	0.325	0.784	0.327	0.325	0.949	2.398
5.0	1.00	0.215	0.624	0.217	0.215	0.948	2.877

Table 5.3
MSPE estimate and relative efficiency of RSS sample estimator, and coverage probability of a 95% confidence interval of population mean. Data sets are generated from a log-normal super population with $\mu = 0$, $\sigma = 1$ and population size $N = 150$

		Est. from equations		Est. from simu.	UE estimate	Coverage prb.	Relative eff.
H	ρ	σ^2_{RSS}	σ^2_{SRS}	$V(\bar{Y}_{\text{RC}})$	$\hat{\sigma}^2_{\text{RSS}}$	$C(\bar{Y}_{\text{RC}})$	RE_{RC}
2.0	0.50	-	0.436	0.400	0.400	0.852	1.089
3.0	0.50	-	0.280	0.243	0.242	0.869	1.153
4.0	0.50	-	0.202	0.160	0.162	0.883	1.262
5.0	0.50	-	0.156	0.117	0.116	0.886	1.336
2.0	0.75	-	0.436	0.371	0.372	0.855	1.176
3.0	0.75	-	0.280	0.216	0.217	0.867	1.300
4.0	0.75	-	0.202	0.146	0.146	0.874	1.388
5.0	0.75	-	0.156	0.103	0.103	0.878	1.514
2.0	1.00	0.362	0.436	0.361	0.364	0.839	1.207
3.0	1.00	0.201	0.280	0.197	0.198	0.849	1.423
4.0	1.00	0.128	0.202	0.128	0.127	0.847	1.586
5.0	1.00	0.086	0.156	0.085	0.085	0.845	1.833

Table 5.4
MSPE estimate and relative efficiency of RSS sample estimator, and coverage probability of a 95% confidence interval of population mean. Data sets are generated from a log-normal super population with $\mu = 0$, $\sigma = 1$ and population size $N = 1{,}000$

		Est. from equations		Est. from simu.	UE estimate	Coverage prb.	Relative eff.
H	ρ	σ^2_{RSS}	σ^2_{SRS}	$V(\bar{Y}_{\text{RC}})$	$\hat{\sigma}^2_{\text{RSS}}$	$C(\bar{Y}_{\text{RC}})$	RE_{RC}
2.0	0.50	-	0.462	0.432	0.433	0.851	1.070
3.0	0.50	-	0.307	0.263	0.263	0.868	1.164
4.0	0.50	-	0.229	0.189	0.190	0.882	1.208
5.0	0.50	-	0.182	0.141	0.141	0.889	1.296
2.0	0.75	-	0.462	0.413	0.413	0.852	1.119
3.0	0.75	-	0.307	0.240	0.238	0.868	1.276
4.0	0.75	-	0.229	0.171	0.170	0.878	1.337
5.0	0.75	-	0.182	0.129	0.129	0.884	1.415
2.0	1.00	0.389	0.462	0.387	0.386	0.839	1.195
3.0	1.00	0.228	0.307	0.225	0.227	0.852	1.364
4.0	1.00	0.154	0.229	0.155	0.155	0.857	1.479
5.0	1.00	0.113	0.182	0.113	0.113	0.862	1.614

6 Example

In this section we apply the proposed estimators to a data set which contains a sheep population in a research farm at Ataturk University, Erzurum, Turkey. Data set contains birth weights, mothers' weights at mating and the weights at the 7th month after birth for 224 lambs. The entire data set is given in Hollander, Wolfe and Chicken (2014, page 709). Variable of interest is the weights (Y) at the 7th month after birth for 224 lambs. We use birth weights (X_1) and mothers' weights (X_2) at mating as auxiliary variables to perform ranking process. The ranking variables are positively correlated with the variable of interest Y. The correlation coefficient $(\rho = \text{corr}(X, Y))$ between X_1, Y and X_2, Y are 0.8425 and 0.5941, respectively. The histogram of the variable of interest, Y, is roughly symmetric. Mean and variance of Y are $\bar{Y}_N = 28.125$kg and $S_N^2 = 15.23$kg^2, respectively, where $S_n^2 = \sum_{i=1}^{224} (Y_i - \bar{Y}_N)^2 \big/ 223$. We treated these 224 lambs as a realization from a super population having finite mean μ and variance σ^2. We constructed samples based RSS sampling design using this finite population. Samples are generated for sample and set size combinations, (n, H), (10, 2), (15, 3), (20, 4), (25, 5). Simulation size is taken to be 50,000.

In this example, we incorporate the sampling cost to RSS and SRS sampling designs with concomitant ranking in RSS. We first need to determine reasonable costs associated with various aspects of RSS. Weight measurement is obtained from seven-month-old lambs. These animals are very active and measurement cost is substantial. The measurement process usually require three people for separating the lamb from the flock, bringing it to scale, holding it firm during the measurement. Suppose that the farm employs the workers in an annual salary of \$50,000. This corresponds to a rate of approximatley \$25 per hour per person. Assume that the measurement of a lamb takes about 5 minutes. The measurement cost for a lamb then would be about $c_{qy} = 3(25/12) \approx 6$ for three workers. Ranking will be performed using auxilairy variables X_1 and X_2. These variables are maintained in the data base for some other purposes. Only cost to sampling would be due to personal cost for ranking. Ranking will be performed in the office by selecting sets at random from the data base and ranking them based on auxiliary variables. Suppose that ranking a set of size H takes about 1/2 minute. This leads to ranking cost of $Hc_{qx} = \$0.21$. We may assume that cost related to identification of a unit in the population is negligible $(c_i = 0)$. Under these stipulations, the cost ratio of selecting and measuring a unit in RSS and SRS is given by ratio $= (Hc_i + Hc_{qx} + c_{qy}) / (c_i + c_{qy}) = (6 + 0.21)/6 = 1.035$. Since this ratio is less than all entries in Table 4.1, we anticipate that $\bar{Y}_{\text{RC}}$ provides higher efficiency than $\bar{Y}_{\text{SRS}}$.

Table 6.1 presents the estimated MSPE and relative efficiency of RSS esimator as well as the coverage probability of the confidence interval of μ for different ρ and sample size combinations. It is clear that the RSS estimator outperform the SRS estimator for all simulation parameter combinations. Estimated MSPEs and coverage probabilities also show similar behaviors as in Section 3. The estimated MSPE values are very close to the simulated MSPE values. The coverage probabilities of the confidence intervals based on $t-$ approximation appear to be very close to the nominal coverage probabiliy, 0.950.

Table 6.1
MSPE estimate and relative efficiency of RSS sample estimator, and coverage probability of a 95% confidence interval of population mean of a sheep population of size $N = 224$

		Est. from equation	Est. from simu.	UE estimate	Coverage prb.	Relative eff.
H	ρ	σ^2_{SRS}	$V(\bar{Y}_{\text{RC}})$	$\hat{\sigma}^2_{\text{RSS}}$	$C(\bar{Y}_{\text{RC}})$	RE_{RC}
2.0	0.59	1.453	1.279	1.275	0.946	1.136
3.0	0.59	0.946	0.776	0.774	0.948	1.219
4.0	0.59	0.693	0.536	0.537	0.948	1.293
5.0	0.59	0.540	0.399	0.402	0.948	1.353
2.0	0.84	1.453	1.107	1.105	0.945	1.312
3.0	0.84	0.946	0.600	0.602	0.946	1.576
4.0	0.84	0.693	0.377	0.382	0.946	1.839
5.0	0.84	0.540	0.263	0.264	0.944	2.056

7 Concluding remarks

We have developed a model based statistical inference for population mean and total based on RSS samples in a finite population setting where samples are constructed by using a without replacement sampling design. It is shown that the sample mean of RSS samples are model unbiased and they have smaller mean square prediction error (MSPE) than the MSPE of a simple random sample mean. We constructed unbiased estimator for the MSPE and prediction confidence interval for the population mean. A small scale simulation study showed that estimators are as good as or better than SRS estimators when the quality of ranking information in RSS sampling is low or high, respectively, and the cost ratio of obtaining a unit in RSS and a unit in SRS is not too high. The coverage probabilities of the prediction intervals are also very close to the nominal coverage probabilities. Proposed sampling designs and inferential procedures are applied to a data set containing a sheep population in an agricultural research farm.

Acknowledgements

The second author’s research is supported by The Scientific and Technological Research Council of Turkey (TUBITAK). The authors thank the Editor, Associate Editor and two referees for their constructive comments and suggestions.

Appendix

Proof of Theorem 1: We write mean square prediction error (MSPE) as

$$\text{MSPE}_M\left(\bar{Y}_R\right) = E_M\left\{\bar{Y}_R - \frac{1}{N}\sum_{i=1}^{N} Y_i\right\}^2 = E_M\left\{\frac{1}{dH}\sum_{h=1}^{H}\sum_{i=1}^{d} Y_{[h]i} - \frac{1}{N}\sum_{i=1}^{N} Y_i\right\}^2 .$$

Let $Z_i, \, i = 1, \ldots, N - nH,$ be the responses on $N - nH$ population units that are neither measured nor used in ranking in any one of the randomly selected sets of size H in the construction of the RSS sample. Then the MSPE can be written

$$\text{MSEP}_M \left(\bar{Y}_R\right) = E_M \left\{ \frac{1}{dH} \sum_{h=1}^{H} \sum_{i=1}^{d} Y_{[h]i} - \frac{1}{N} \sum_{h=1}^{H} \sum_{i=1}^{d} \left[Y_{[h]i} + \sum_{h \neq h'}^{H} Y_{[h']i}^{*} \right] - \frac{1}{N} \sum_{i=1}^{N-nH} Z_i \right\}^2$$

where $Y_{[h']i}^{*}, \, h' \neq, h,$ are responses on unmeasured units that are used in ranking of units in a set. Hence, $Y_{[h']i}^{*}$ and $Y_{[h]i}$ are correlated, but they are uncorrelated with Z_i. Let

$$c_{h,h'} = \begin{cases} \dfrac{N-n}{n} & h = h' \\ -1 & h \neq h'. \end{cases}$$

Using the definition of $c_{h,h'}$, we combine $Y_{[h']i}^{*}$ and $Y_{[h]i}$ under the same summation and write the MSPE as

$$\begin{aligned} \text{MSPE}_M \left(\bar{Y}_R\right) &= \frac{1}{N^2} \text{var} \left\{ \sum_{h=1}^{H} \sum_{i=1}^{d} \sum_{h'=1}^{H} c_{h,h'} Y_{[h]i} \right\} + \text{var} \left\{ \frac{1}{N} \sum_{i=1}^{N-nH} Z_i \right\} \\ &= \frac{1}{N^2} \sum_{h=1}^{H} \text{var} \left[\sum_{i=1}^{d} \sum_{h'=1}^{H} c_{h,h'} Y_{[h]i} \right] + \text{var} \left\{ \frac{1}{N} \sum_{i=1}^{N-nH} Z_i \right\} \\ &= \frac{d}{N^2} \sum_{h=1}^{H} \sum_{h'=1}^{H} \left(c_{h,h'}\right)^2 \sigma_{[h']}^2 + \frac{d}{N^2} \sum_{h=1}^{H} \left(\sum_{h'=1}^{H} \sum_{t \neq h'}^{H} c_{h,h'} c_{h,t} \sigma_{[h',t]} \right) \\ &\quad + \text{var} \left\{ \frac{1}{N} \sum_{i=1}^{N-nH} Z_i \right\} \\ &= A + B + \frac{(N - nH)\, \sigma^2}{N^2}. \end{aligned} \qquad \text{(A.1)}$$

The expression A reduces to

$$\begin{aligned} A &= \frac{d}{N^2} \sum_{h=1}^{H} \left(c_{h,h}\right)^2 \sigma_{[h]}^2 + \frac{d}{N^2} \sum_{h} \sum_{h' \neq h}^{H} \left(c_{h,h'}\right)^2 \sigma_{[h,h']} \\ &= \frac{d}{N^2} \left[\left(\frac{N-n}{n} \right)^2 + (H-1) \right] \sum_{h=1}^{H} \sigma_{[h]}^2. \end{aligned}$$

In a similar fashion, the expression B reduces to

$$\begin{aligned} B &= \frac{d}{N^2} \sum_{h=1}^{H} \left[\sum_{t \neq h', h}^{H} \left(\sum_{h'=1}^{H} c_{h,h'} c_{h,t} \sigma_{[h',t]} + c_{h,h} c_{h,t} \sigma_{[h,t]} \right) + \sum_{h' \neq h}^{H} c_{h,h'} c_{h,h} \sigma_{[h',h]} \right] \\ &= \frac{d}{N^2} \sum_{h=1}^{H} \left[\sum_{h' \neq h}^{H} \sum_{t \neq h', h}^{H} \sigma_{[h',t]} - 2 \left(\frac{N-n}{n} \right) \left(\sum_{t=1}^{H} \sigma_{[h,t]} - \sigma_{[h,h]} \right) \right] \\ &= \frac{d}{N^2} \left[\left(H^2 - 2H\right) \sigma^2 - (H-2) \sum_{h=1}^{H} \sigma_{[h]}^2 - 2 \left(\frac{N-n}{n} \right) \left(H \sigma^2 - \sum_{h=1}^{H} \sigma_{[h]}^2 \right) \right]. \end{aligned}$$

By inserting expressions A and B in equation (A.1), we conclude that

$$\begin{aligned}\mathrm{MSPE}_M\left(\bar{Y}_R\right) &= \frac{d}{N^2}\left[\left(\frac{N-n}{n}\right)^2 + (H-1)\right]\sum_{h=1}^{H}\sigma^2_{[h]} \\ &+ \frac{d}{N^2}\left[(H^2-2H)\sigma^2 - (H-2)\sum_{h=1}^{H}\sigma^2_{[h]} - 2\left(\frac{N-n}{n}\right)\left(H\sigma^2 - \sum_{h=1}^{H}\sigma^2_{[h]}\right)\right] \\ &+ \left(\frac{N-nH}{N^2}\right)\sigma^2 \\ &= \left(\frac{N-n}{Nn}\right)\sigma^2 - \frac{1}{nH}\sum_{h=1}^{H}\left(\mu_{[h]}-\mu\right)^2\end{aligned}$$

which completes the proof. Note that to establish the last equality we used the fact that $\sigma^2 = \sum_{h=1}^{H}\sigma^2_{[h]}\Big/H + \sum_{h=1}^{H}\left(\mu_{[h]}-\mu\right)^2\Big/H.$

Proof of Theorem 2: We first look at the expected values of T_1^* and T_2^* under the super population model in equation (2.1)

$$\begin{aligned}E\left(T_1^*\right) &= \frac{1}{H}\sum_{h=1}^{H}\left(\mu_{[h]}-\mu\right)^2 + \frac{H-1}{H^2}\sum_{h=1}^{H}\sigma^2_{[h]} = \sigma^2 - \frac{1}{H^2}\sum_{h=1}^{H}\sigma^2_{[h]} \\ E\left(T_2^*\right) &= \frac{1}{H^2}\sum_{h=1}^{H}\sigma^2_{[h]}.\end{aligned}$$

It is now easy to establish that $E\left(T_1^* + T_2^*\right) = \sigma^2$. The proof is then completed by inserting these expressions in equation (3.1).

Proof of Theorem 3: We sketch the proof for $\mathrm{RE}\left(\bar{Y}_{\mathrm{RC}}, \bar{Y}_{\mathrm{SRS}}\right)$. From the total cost function, we write

$$n_S = \frac{C}{c_i + c_{qy}} \quad \text{and} \quad n_R = \frac{C}{Hc_i + Hc_{qx} + c_{qy}},$$

where C is the fixed total cost. Using these expressions, we have

$$\begin{aligned}\mathrm{RE}\left(\bar{Y}_{\mathrm{RC}}, \mathrm{SRS}\right) &= \frac{n_R\left(N-n_S\right)}{n_S\left(N-n_r-ND\right)} \\ &= \frac{N\left(c_i+c_{qy}\right)-C}{N\left(Hc_i+Hc_{qx}+c_{qy}\right)-C-ND\left(Hc_i+Hc_{qx}+c_{qy}\right)}.\end{aligned}$$

We now establish that $\mathrm{RE}\left(\bar{Y}_{\mathrm{RC}}, \mathrm{SRS}\right) \geq 1$ if and only if

$$\begin{aligned}c_i + c_{qy} &\geq \left(Hc_i + Hc_{qx} + c_{qy}\right)(1-D) = \frac{Hc_i + Hc_{qx} + c_{qy}}{\mathrm{RP}} \\ \mathrm{RP} &\geq \frac{Hc_i + Hc_{qx} + c_{qy}}{c_i + c_{qy}}\end{aligned}$$

which completes the proof.

References

Ahn, S., Lim, J. and Wang, X. (2014). The students's *t* approximation to distributions of pivotal statistics from ranked set samples. *Journal of Korean Statistical Society*, 43, 4, 643-652.

Al-Saleh, M.F., and Samawi, H.M. (2007). A note on inclusion probability in ranked set sampling and some of its variations. *Test*, 16, 1, 198-209.

Dell, T.R., and Clutter, J.L. (1972). Ranked-set sampling theory with order statistics background. *Biometrics*, 28, 2, 545-555.

Deshpande, J.V., Frey, J. and Ozturk, O. (2006). Nonparametric ranked set-sampling confidence intervals for a finite population. *Environmental and Ecological Statistics*, 13, 1, 25-40.

Frey, J. (2011). A note on ranked-set sampling using a covariate. *Journal of Statistical Planning and Inference*, 141, 2, 809-816.

Gokpinar, F., and Ozdemir, Y.A. (2010). Generalization of inclusion probabilities in ranked set sampling. *Hacettepe Journal of Mathematics and Statistics*, 39, 1, 89-95.

Hollander, M., Wolfe, D.A. and Chicken, E. (2014). *Nonparametric Statistical Methods, 3rd Edition*. Wiley, 709.

Nahhas, R.W., Wolfe, D.A. and Chen, H. (2002). Ranked set sampling: Cost and optimal set size. *Biometrics*, 58, 964-971.

Ozdemir, Y.A., and Gokpinar, F. (2007). A generalized formula for inclusion probabilities in ranked set sampling. *Hacettepe Journal of Mathematics and Statistics*, 36, 1, 89-99.

Ozdemir, Y.A., and Gokpinar, F. (2008). A new formula for inclusion probabilities in median ranked set sampling. *Communications in Statistics - Theory and Methods*, 37, 13, 2022-2033.

Ozturk, O. (2014). Estimation of population mean and total in finite population setting using multiple auxilary variables. *Journal of Agricultural, Biological and Environmental Statistics*, 19, 2, 161-184.

Ozturk, O. (2015). Distribution free two-sample methods for judgment-post stratified data. *Statistica Sinica*, 25, 1691-1712.

Ozturk, O. (2016a). Estimation of finite population mean and total using population ranks of sample units. *Journal of Agricultural, Biological and Environmental Statistics*, 21, 1, 181-202.

Ozturk, O. (2016b). Statistical inference based on judgment post-stratifed samples in finite population. *Survey Methodology*, 42, 2, 239-262. Paper available at https://www150.statcan.gc.ca/n1/pub/12-001-x/2016002/article/14664-eng.pdf.

Ozturk, O., and Jozani, M.J. (2013). Inclusion probabilities in partially rank ordered set sampling. *Computational Statistics & Data Analysis*, 69, 122-132.

Patil, G.P., Sinha, A.K. and Taillie, C. (1995). Finite population corrections for ranked set sampling. *Annals of the Institute of Statistical Mathematics*, 47, 4, 621-636.

Patil, G.P., Sinha, A.K. and Taillie, C. (1997). Ranked set sampling, coherent rankings and size-biased permutations. *Journal of Statistical Planning and Inference*, 63, 2, 311-324.

Takahasi, K., and Futatsuya, M. (1998). Dependence between order statistics in samples from finite population and its application to ranked set sampling. *Annals of the Institute of Statistical Mathematics*, 50, 1, 49-70.

Survey Methodology, June 2018
Vol. 44, No. 1, pp. 17-42
Statistics Canada, Catalogue No. 12-001-X

Linearization versus bootstrap for variance estimation of the change between Gini indexes

Guillaume Chauvet and Camelia Goga[1]

Abstract

This paper investigates the linearization and bootstrap variance estimation for the Gini coefficient and the change between Gini indexes at two periods of time. For the one-sample case, we use the influence function linearization approach suggested by Deville (1999), the without-replacement bootstrap suggested by Gross (1980) for simple random sampling without replacement and the with-replacement of primary sampling units described in Rao and Wu (1988) for multistage sampling. To obtain a two-sample variance estimator, we use the linearization technique by means of partial influence functions (Goga, Deville and Ruiz-Gazen, 2009). We also develop an extension of the studied bootstrap procedures for two-dimensional sampling. The two approaches are compared on simulated data.

Key Words: Composite estimator; Horvitz-Thompson estimator; Influence function; Intersection estimator; Replication weights; Two-sample survey; Two-dimensional sampling design; Union estimator; Variance estimation.

1 Introduction

The Gini coefficient (Gini, 1914) is one of the best known concentration measure often desired in economical studies. If $\mathcal{Y}_1$ denotes a quantitative positive variable such as the income and $F_1(\cdot)$ denotes its distribution function defined on $]-\infty, \infty[$, the Gini coefficient is

$$G_1 = \frac{1}{2}\frac{\iint |v-u|\, dF(u)\, dF(v)}{\int u dF(u)},$$

provided $\int udF(u) \neq 0.$ The Gini coefficient measures the dispersion of a quantitative positive variable within a population. Statistical institutes generally make use of the Gini coefficient to evaluate the income inequalities of a country at different periods of time, or of different countries at the same time. In the last decades, the Gini coefficient has also been considered in economic and sociodemographic fields (see for example Navarro, Muntaner, Borrell, Benach, Quiroga, Rodriguez-Sanz, Vergès and Pasarin, 2006; Bhattacharya, 2007; Lai, Huang, Risser and Kapadia, 2008; Barrett and Donald, 2009), biology (Graczyk, 2007), environment (Druckman and Jackson, 2008; Groves-Kirkby, Denman and Phillips, 2009) or astrophysics (Lisker, 2008).

There is an extensive literature on variance estimation for the Gini coefficient with observations obtained from survey data, see Langel and Tillé (2013) for a review. Glasser (1962) and Sandström, Wretman and Waldèn (1985) considered the case of simple random sampling. Sandström, Wretman and Waldèn (1988) listed possible variance estimators for a general sampling design, including a jackknife variance estimator. This latter approach was further investigated by Yitzhaki (1991), Karagiannis and Kovačević (2000) and

[1] Guillaume Chauvet, Université Rennes, ENSAI, CNRS, IRMAR - UMR 6625, F-35000 Rennes, France. E-mail: chauvet@ensai.fr; Camelia Goga, Laboratoire des Mathématiques de Besançon, Université de Bourgogne Franche-Comté, UMR 6623, Besançon Cedex, France. E-mail: camelia.goga@univ-fcomte.fr.

Berger (2008). Linearization variance estimation was studied by Kovačević and Binder (1997), and Berger (2008) demonstrated the equivalence between linearization and a generalized jackknife technique first suggested by Campbell (1980). Qin, Rao and Wu (2010) proposed bootstrap and empirical likelihood based confidence intervals for the Gini coefficient. They studied these methods both theoretically and empirically in the particular case of stratified with replacement simple random sampling. However, bootstrap variance estimation has not been compared with alternative methods for the change between Gini indexes.

In this article, we consider linearization versus bootstrap to estimate the change between Gini indexes. The paper is structured as follows. In Section 2, we first consider the estimation of the Gini coefficient in the one-sample case. The notation is defined in Section 2.1, and the substitution estimator of the Gini coefficient is presented in Section 2.2. The linearization variance estimator is given in Section 2.3, with application to the simple random sampling (SI) design and to a multistage sampling design. The main principles of the weighted bootstrap are briefly reviewed at the beginning of Section 2.4, and the without-replacement bootstrap (BWO) suitable for SI sampling is introduced in Section 2.4.1, while the bootstrap of primary sampling units (BWR) suitable for multistage sampling is introduced in Section 2.4.2. In Section 3, we consider the estimation of the change between Gini indexes in the two-sample case. The notation is defined in Section 3.1, and we briefly review the principles of composite estimation which is applied in Section 3.1.1 for the two-dimensional SI design (SI2) and in Section 3.1.2 for a two-dimensional two-stage sampling design (MULT2). The composite estimator of the change between Gini indexes is presented in Section 3.2. The linearization variance estimator by means of the partial influence functions is given in Section 3.3, with application to the SI2 design and to the MULT2 design. An extension of the BWO for the SI2 design and of the BWR for the MULT2 design are then presented in Section 3.4. Linearization and the proposed bootstrap methods are compared in Section 4 through a simulation study. Section 5 concludes.

2 One sample case

2.1 Notation

Let U denote some finite population of size N whose units may be identified by the labels $k = 1, \ldots, N$. Suppose that the variable $\mathcal{Y}_1$ is measured on the population U, and let $y_{11}, \ldots, y_{1N}$ denote the values taken by $\mathcal{Y}_1$ on the units in the population. Let $M_1 = \sum_{k \in U} \delta_{y_{1k}}$ denote the discrete measure taking unit mass on any point y_{1k} in the population and 0 elsewhere, with $\delta_{y_{1k}}$ the Dirac mass at y_{1k}. Most of the parameters of interest θ_1 studied in surveys can be written as a functional T of M_1, namely $\theta_1 = T(M_1)$. For instance, the total $t_{y1} = \sum_{k \in U} y_{1k}$ equals $\int \mathcal{Y}_1 dM_1$. In practice, a sample s (with or without repetitions) is selected by means of a sampling design $p(\cdot)$, and we observe the values y_{1k} for $k \in s$ only. A substitution principle is used for estimation (see Deville, 1999, and Goga, Deville and Ruiz-Gazen, 2009). Let π_k denote the expected number of draws for unit k in the sample; in case of without-replacement sampling, this is the probability that unit k is selected in the sample. Let $\hat{M}_1 = \sum_{k \in s} w_k \delta_{y_{1k}}$ denote the discrete measure taking mass w_k on any point in the sample and 0 elsewhere, where $w_k = \pi_k^{-1}$ is the sampling weight. Substituting $\hat{M}_1$ into θ_1 yields the estimator $\hat{\theta}_1 = T(\hat{M}_1)$.

For a without-replacement sampling design, the substitution estimator for a total is the so-called Horvitz-Thompson (HT) estimator $\hat{t}_{y1}^{\mathrm{HT}} = \sum_{k \in s} w_k y_{1k}$. The HT variance estimator is

$$v^{\mathrm{HT}}\left(\hat{t}_{y1}^{\mathrm{HT}}\right) = \sum_{k \in s}\sum_{l \in s} \frac{\Delta_{kl}}{\pi_{kl}} \frac{y_{1k}}{\pi_k} \frac{y_{1l}}{\pi_l}, \tag{2.1}$$

where $\pi_{kl} = \Pr(k,\, l \in s)$ denotes the probability that units k and l are selected jointly in the sample, and $\Delta_{kl} = \pi_{kl} - \pi_k \pi_l$. In the particular case of simple random sampling without replacement (SI) of size n, we have $\hat{t}_{y1}^{\mathrm{HT}} = N\bar{y}_{1,s}$ with $\bar{y}_{1,s} = n^{-1}\sum_{k \in s} y_{1k}$, and formula (2.1) yields

$$v^{\mathrm{HT}}\left(\hat{t}_{y1}^{\mathrm{HT}}\right) = N^2\left(\frac{1}{n} - \frac{1}{N}\right) S_{y_1,s}^2 \quad \text{where} \quad S_{y_1,s}^2 = \frac{1}{n-1}\sum_{k \in s}\left(y_{1k} - \bar{y}_{1,s}\right)^2. \tag{2.2}$$

For a with-replacement sampling design, the substitution estimator for a total is the so-called Hansen-Hurwitz (HH) estimator $\hat{t}_{y1}^{\mathrm{HH}} = \sum_{k \in s} w_k y_{1k}$. We consider the important case of multistage sampling, where the N units are grouped inside N_I non-overlapping Primary Sampling Units (PSU) $U_1, \ldots, U_{N_I}$, and where a with-replacement first-stage sample s_I of size m is selected. Let π_{Ii} denote the expected number of draws for the PSU U_i in s_I. A second-stage sample s_i is then selected inside any $i \in s_I$ by means of some sampling design $p_i(\cdot)$. Let $\pi_{k|i}$ denote the expected number of draws for unit k in s_i. The estimated measure is then $\hat{M}_1 = \sum_{i \in s_I}\sum_{k \in s_i} \pi_{Ii}^{-1}\pi_{k|i}^{-1}\delta_{y_{1k}}$. We have $\hat{t}_{y1}^{\mathrm{HH}} = \sum_{i \in s_I} \pi_{Ii}^{-1}\hat{Y}_i$ where $\hat{Y}_i = \sum_{k \in s_i} \pi_{k|i}^{-1} y_{1k}$, and an unbiased variance estimator for $\hat{t}_{y1}^{\mathrm{HH}}$ is

$$v^{\mathrm{HH}}\left(\hat{t}_{y1}^{\mathrm{HH}}\right) = \frac{m}{m-1}\sum_{i \in s_I}\left(\frac{\hat{Y}_i}{\pi_{Ii}} - \frac{\hat{t}_{y1}^{\mathrm{HH}}}{m}\right)^2. \tag{2.3}$$

2.2 Estimating the Gini coefficient

If the variable $\mathcal{Y}_1$ is measured on the population U, the Gini coefficient is

$$G_1 = \frac{1}{2}\frac{\sum_{k \in U}\sum_{l \in U}\left|y_{1k} - y_{1l}\right|}{N\sum_{k \in U} y_{1k}},$$

see for example Nygård and Sandström (1985). It follows that G_1 is zero if $\mathcal{Y}_1$ is constant on the population, which occurs when the total of $\mathcal{Y}_1$ is equally distributed among all the population individuals. In the opposite case, when only one individual owns the whole amount of $\mathcal{Y}_1$, G_1 is maximized and equal to $1 - 1/N$: the total of $\mathcal{Y}_1$ is then concentrated in one point only, which means maximum inequality among members of the population.

If all individuals $k \neq l$ have different values for the variable $\mathcal{Y}_1$, the Gini coefficient G_1 is

$$G_1 = \frac{\sum_{k=1}^{N} y_{1(k)} \left(2k/N - 1\right)}{t_{y1}} - \frac{1}{N} = \frac{\sum_{k \in U} y_{1k} \left\{2F_{1N}\left(y_{1k}\right) - 1\right\}}{t_{y1}} - \frac{1}{N} \tag{2.4}$$

with $y_{1(1)} \leq \cdots \leq y_{1(N)}$ the ordered values and $F_{1N}(\cdot) = N^{-1}\sum_{k \in U} 1_{\{y_{1k} \leq \cdot\}}$ the finite population distribution function; see Sandström, Wretman and Waldèn (1988) and Deville (1997) for further details on the derivation of (2.4). Nygård and Sandström (1985) called the term $-1/N$ the Gini finite population correction and gave several reasons to make this correction, such as the non-negativity of the lower bound of G_1. As is frequently done in the literature (see for example Glasser, 1962), this correction is ignored in the sequel. We redefine the Gini coefficient as

$$G_1 = \frac{\sum_{k \in U} y_{1k} \left\{2F_{1N}\left(y_{1k}\right) - 1\right\}}{t_{y1}} = \frac{\int \left\{2F_{1N}(y) - 1\right\} y dM_1(y)}{\int y dM_1(y)} \tag{2.5}$$

where the finite population distribution function $F_{1N}(\cdot)$ is a functional family

$$F_{1N}(y) = \frac{1}{\int dM_1(y)} \int 1_{\{\xi \leq y\}} dM_1(\xi) \tag{2.6}$$

indexed by y. Substituting $\hat{M}_1$ into (2.5) and (2.6) yields the estimator

$$\hat{G}_1 = \frac{\int \left\{2\hat{F}_{1N}(y) - 1\right\} y d\hat{M}_1(y)}{\int y d\hat{M}_1(y)} = \frac{\sum_{k \in s} w_k \left\{2\hat{F}_{1N}(y_{1k}) - 1\right\} y_{1k}}{\sum_{k \in s} w_k y_{1k}}, \tag{2.7}$$

where

$$\hat{F}_{1N}(y) = \frac{1}{\int d\hat{M}_1(y)} \int 1_{\{\xi \leq y\}} d\hat{M}_1(\xi) = \frac{1}{\sum_{k \in s} w_k} \sum_{k \in s} w_k 1_{\{y_{1k} \leq y\}} \tag{2.8}$$

is the substitution estimator of the distribution function F_{1N}.

2.3 Linearization variance estimation

We give below some brief details about the influence function linearization (IFL) (Deville, 1999), which consists in giving a first-order expansion of the substitution estimator $\hat{\theta}_1 = T\left(\hat{M}_1\right)$ around the true value $\theta_1 = T(M_1)$, to approximate the error by a linear estimator of some artificial *linearized variable*. More precisely, the first derivatives of T with respect to M_1 are the influence functions

$$\mathrm{IT}(M_1; y) = \lim_{h \to 0} \frac{T(M_1 + h\delta_y) - T(M_1)}{h},$$

and $u_{1k} = \mathrm{IT}(M_1; y_{1k})$ is the linearized variable for all $k \in U$. Suppose that $T(\cdot)$ is homogeneous, namely there exists some positive number β dependent on T such that $T(rM_1) = r^{\beta} T(M_1)$ for any real $r > 0$.

Assume also that $\lim_{N\to\infty} N^{-\beta} T(M_1) < \infty$. Under some additional regularity assumptions upon $T(\cdot)$ and the sampling design (e.g., Goga and Ruiz-Gazen, 2014), Deville (1999) establishes that

$$\hat{\theta}_1 - \theta_1 = \left(\sum_{k\in s} w_k u_{1k} - \sum_{k\in U} u_{1k}\right) + o_p\left(N^{\beta} n^{-1/2}\right),$$

so that the error $\hat{\theta}_1 - \theta_1$ can be approximated by the error of the HT estimator for the total of the linearized variable u_{1k}. For a without-replacement sampling design, using a sample-based estimator $\hat{u}_{1k}$ of the linearized variable u_{1k} in the HT variance estimator yields the variance estimator

$$v_{\text{LIN}}^{\text{HT}}\left(\hat{\theta}_1\right) = \sum_{k\in s}\sum_{l\in s} \frac{\Delta_{kl}}{\pi_{kl}} \frac{\hat{u}_{1k}}{\pi_k} \frac{\hat{u}_{1l}}{\pi_l}, \tag{2.9}$$

where $\pi_{kl} = \Pr(k, l \in s)$ denotes the probability that units k and l are selected jointly in the sample, and $\Delta_{kl} = \pi_{kl} - \pi_k \pi_l$. Several results of asymptotic normality have been proved for specific sampling designs, see Hájek (1960, 1961, 1964), Rosén (1972), Sen (1980), Krewski and Rao (1981), Gordon (1983), Ohlsson (1986, 1989), Chen and Rao (2007), Brändén and Jonasson (2012), Saegusa and Wellner (2013) and Chauvet (2015), among others. If the sampling design is such that the substitution estimator $\hat{\theta}_1$ satisfies a central-limit theorem, an approximately $(1-2\alpha)\%$ confidence interval is $\left[\hat{\theta}_1 - z_\alpha\sqrt{v_{\text{lin}}\left(\hat{\theta}_1\right)}, \hat{\theta}_1 + z_\alpha\sqrt{v_{\text{lin}}\left(\hat{\theta}_1\right)}\right]$ where z_α is the upper $\alpha\%$ cutoff for the standard normal distribution.

In case of the Gini coefficient, we have $\beta = 0$ and the linearized variable is

$$u_{1k} = 2F_{1N}(y_{1k}) \frac{y_{1k} - \bar{y}_{1k,U<}}{t_{y1}} - y_{1k}\frac{G_1 + 1}{t_{y1}} + \frac{1 - G_1}{N}, \tag{2.10}$$

where $\bar{y}_{1k,U<} = \left(\sum_{l\in U} 1_{\{y_{1l}<y_{1k}\}}\right)^{-1} \sum_{j\in U} y_{1j} 1_{\{y_{1j}<y_{1k}\}}$ denotes the mean of the y_{1j} lower than y_{1k}, see Deville (1999). Kovačević and Binder (1997) derived the same expression by means of the estimating equations linearization method; using the Demnati and Rao (2004) linearization approach also leads to the same result. The estimated linearized variable is

$$\hat{u}_{1k} = 2\hat{F}_{1N}(y_{1k}) \frac{y_{1k} - \bar{y}_{1k,s<}}{\hat{t}_{y1}} - y_{1k}\frac{\hat{G}_1 + 1}{\hat{t}_{y1}} + \frac{1 - \hat{G}_1}{\hat{N}} \tag{2.11}$$

where $\bar{y}_{1k,s<} = \left(\sum_{l\in s} w_l 1_{\{y_{1l}<y_{1k}\}}\right)^{-1} \sum_{j\in s} w_j y_{1j} 1_{\{y_{1j}<y_{1k}\}}$.

In the particular SI case, the linearization variance estimator for the Gini coefficient is

$$v_{\text{LIN}}^{\text{HT}}\left(\hat{G}_1\right) = N^2\left(\frac{1}{n} - \frac{1}{N}\right) S_{\hat{u}_1,s}^2 \quad \text{where} \quad S_{\hat{u}_1,s}^2 = \frac{1}{n-1}\sum_{k\in s}\left(\hat{u}_{1k} - \bar{\hat{u}}_{1,s}\right)^2, \tag{2.12}$$

and where $\bar{\hat{u}}_{1,s} = n^{-1}\sum_{k\in s} \hat{u}_{1k}$. In the particular case of multistage sampling and with-replacement sampling of PSUs, the linearization variance estimator for the Gini coefficient is

$$v_{\text{LIN}}^{\text{HH}}\left(\hat{G}_1\right) = \frac{m}{m-1}\sum_{i\in s_I}\left(\frac{\hat{U}_{1i}}{\pi_{Ii}} - \frac{\hat{t}_{\hat{u}_1}^{\text{HH}}}{m}\right)^2 \quad \text{where} \quad \hat{U}_{1i} = \sum_{k\in s_i}\pi_{k|i}^{-1}\hat{u}_{1k}. \tag{2.13}$$

2.4 Bootstrap variance estimation

The use of bootstrap techniques in survey sampling has been extensively studied in the literature. The main bootstrap techniques may be thought as particular cases of the weighted bootstrap (Bertail and Combris, 1997; Antal and Tillé, 2011; Beaumont and Patak, 2012); see also Shao and Tu (1995, Chapter 6), Davison and Hinkley (1997, Section 3.7) and Davison and Sardy (2007) for detailed reviews. Under a weighted bootstrap procedure, the measure $\hat{M}_1 = \sum_s w_k \delta_{y_k}$ is estimated, conditionally on the sample s, by the bootstrap measure

$$\hat{M}_1^* = \sum_{k\in s} w_k D_k \delta_{y_k} \tag{2.14}$$

where $D = \{D_k\}_{k\in s}$ denotes a (random) vector of resampling weights. We note E_* and V_* for the expectation and variance with respect to the resampling scheme. In case of without-replacement sampling, the vector D is generated in such a way that

$$E_*\left(\sum_s w_k D_k y_k\right) \simeq \hat{t}_{y1}^{\text{HT}} \quad \text{and} \quad V_*\left(\sum_s w_k D_k y_k\right) \simeq v^{\text{HT}}\left(\hat{t}_{y1}^{\text{HT}}\right) \tag{2.15}$$

so that the two first moments of the HT-estimator are approximately matched. In case of with-replacement sampling, the vector D is generated in such a way that

$$E_*\left(\sum_s w_k D_k y_k\right) \simeq \hat{t}_{y1}^{\text{HH}} \quad \text{and} \quad V_*\left(\sum_s w_k D_k y_k\right) \simeq v^{\text{HH}}\left(\hat{t}_{y1}^{\text{HH}}\right) \tag{2.16}$$

so that the two first moments of the HH-estimator are approximately matched.

Under any weighted bootstrap technique, the plug-in estimator of $\theta_1 = T(M_1)$ is $\hat{\theta}_1^* = T\left(\hat{M}_1^*\right)$, and the variance of $\hat{\theta}_1 = T\left(\hat{M}_1\right)$ is estimated by

$$V_*\left(\hat{\theta}_1^*\right) = E_*\left\{\hat{\theta}_1^* - E_*\left(\hat{\theta}_1^*\right)\right\}^2. \tag{2.17}$$

Since the variance estimator (2.17) may be difficult to compute exactly, a simulation-based variance estimator may be used instead. More precisely, C independent realizations $D_1, \ldots, D_C$ of the vector D are generated, and we denote $\hat{\theta}_{1c}^* = T\left(\hat{M}_{1c}^*\right)$ with $\hat{M}_{1c}^*$ the Bootstrap measure associated to the vector D_c. Then $V\left(\hat{\theta}_1\right)$ is estimated by

$$v_B\left(\hat{\theta}_1\right) = \frac{1}{C-1}\sum_{c=1}^{C}\left\{\hat{\theta}_{1c}^* - \frac{1}{C}\sum_{c'=1}^{C}\hat{\theta}_{1c'}^*\right\}^2. \tag{2.18}$$

Two types of confidence intervals are usually computed. The percentile method makes use of the ordered bootstrap estimates $\hat{\theta}_{(1c)}^*$, $c = 1, \ldots, C$ to form a $(1-2\alpha)\%$ confidence interval $[\hat{\theta}_{(1L)}^*, \hat{\theta}_{(1U)}^*]$ with $L = \alpha C$

and $U = (1-\alpha)C.$ The bootstrap$-t$ involves the estimation of the pivotal statistic $t = (\hat{\theta}_1 - \theta_1)/\sqrt{v_{\mathrm{BWO}}(\hat{\theta}_1)}$ by its bootstrap counterpart $t^* = (\hat{\theta}_1^* - \hat{\theta}_1)/\sqrt{v_{\mathrm{BWO}}^*(\hat{\theta}_1^*)},$ where $v_{\mathrm{BWO}}^*(\hat{\theta}_1^*)$ is obtained by applying the bootstrap procedure to the resample $s^*.$ The bootstrap$-t$ is computationally very intensive since a double bootstrap is required, and is thus less attractive for a data user. Therefore, we do not pursue this approach further and we focus on the percentile method.

Linearization methods provide variance formulas applicable to general sampling designs, but involve possibly intricate computation of derivatives for complex parameters of interest such as the Gini coefficient. Unlike the linearization, the bootstrap avoids theoretical work by re-calculating the existing estimation system repeatedly. Replicate weights are supplied with the data set, and may be easily used to produce variance estimates for a wide range of statistics. However, a bootstrap technique is usually not suitable for general sampling designs. That is, a particular sampling design usually requires a tailor made resampling scheme. In this paper, we focus on two particular bootstrap techniques, which will be generalized in Section 3 to the two-sample context.

2.4.1 Without-replacement bootstrap for SI sampling

When the sample s is selected by means of SI, we consider the without replacement bootstrap (BWO) introduced by Gross (1980). The approach is readily extended to stratified simple random sampling (STSI) with a finite number of strata. Suppose that N/n is an integer. Then the vector D is obtained by, first creating a pseudo-population U^* of size N by duplicating N/n times each unit k in the original sample $s,$ and then by selecting a SI resample s^* of size n in $U^*.$

The bootstrap measure is given by (2.14), where the resampling weight D_k is the number of times unit $k \in s$ is selected in $s^*.$ The building of U^* may be avoided by noting that under the BWO procedure, the vector D follows a multivariate hypergeometric distribution. Therefore, the resampling weights may be directly generated. It can be shown that the BWO procedure leads to

$$E_*\left(\sum_s w_k D_k y_k\right) = \hat{t}_{y1}^{\mathrm{HT}} \quad \text{and} \quad V_*\left(\sum_s w_k D_k y_k\right) = \frac{1-n^{-1}}{1-N^{-1}} v^{\mathrm{HT}}\left(\hat{t}_{y1}^{\mathrm{HT}}\right), \tag{2.19}$$

where $v^{\mathrm{HT}}\left(\hat{t}_{y1}^{\mathrm{HT}}\right)$ is given in (2.2), so that equation (2.15) is approximately matched for a large sample size.

Several solutions have been proposed to handle the case when N/n is not an integer, see Chao and Lo (1985), Bickel and Freedman (1984), Sitter (1992b), Booth, Butler and Hall (1994), Presnell and Booth (1994), among others. The generalization of BWO variance estimation for unequal probability sampling designs is considered in Särndal, Swensson and Wretman (1992) and Chauvet (2007).

2.4.2 With-replacement bootstrap for multistage sampling

When the sample s is selected by means of multistage sampling and with-replacement unequal probability sampling of PSUs, we consider the bootstrap of PSUs (BWR) introduced by Rao and Wu (1988).

A with-replacement resample s_I^* of size $m-1$ is selected by means of simple random sampling with replacement (SIR) in the original first-stage sample s_I. The bootstrap measure is

$$\hat{M}_1^* = \frac{m}{m-1} \sum_{i \in s_I^*} \sum_{k \in s_i} \pi_{Ii}^{-1} \pi_{k|i}^{-1} \delta_{y_{1k}} = \sum_{k \in s} w_k D_k \delta_{y_k}, \tag{2.20}$$

where the resampling weight D_k equals $m(m-1)^{-1}$ multiplied by the number of times the PSU containing k is selected in s_I^*.

The resampling size $m-1$ is used to reproduce the usual unbiased variance estimator in the linear case (see Rao and Wu, 1988). It can be shown that the BWR procedure leads to

$$E_* \left(\sum_s w_k D_k y_k \right) = \hat{t}_{y1}^{\mathrm{HH}} \quad \text{and} \quad V_* \left(\sum_s w_k D_k y_k \right) = v^{\mathrm{HH}} \left(\hat{t}_{y1}^{\mathrm{HH}} \right), \tag{2.21}$$

where $v^{\mathrm{HH}} \left(\hat{t}_{y1}^{\mathrm{HH}} \right)$ is given in (2.3), so that equation (2.16) is exactly matched. The BWR procedure is particulary simple, since involving a resampling for the first-stage of sampling only, the sub-samples of Secondary sampling Units (SSUs) being left unchanged inside the resampled PSUs.

3 Two-sample case

3.1 Notation and composite estimation

Suppose now that two variables $\mathcal{Y}_1$ and $\mathcal{Y}_2$ are measured on the population U, and let $y_{d1}, \ldots, y_{dN}$ denote the values taken by $\mathcal{Y}_d, d = 1, 2,$ on the units in the population. The variables $\mathcal{Y}_1$ and $\mathcal{Y}_2$ may typically refer to some characteristic of interest collected at two different times τ_1 and τ_2. We consider the estimation of parameters $\Delta\theta$ that can be written as a functional $\Delta\theta = T(M_1, M_2)$, where $M_d = \sum_{k \in U} \delta_{\{y_{dk}\}}$. For instance, the linear case $\Delta t = t_{y2} - t_{y1}$ corresponds to the difference between the totals $t_{y2} = \sum_{k \in U} y_{2k}$ and $t_{y1} = \sum_{k \in U} y_{1k}$.

Let s_1 and s_2 be two samples of sizes n_1 and n_2, respectively, selected from the same population U according to some two-dimensional sampling design $p(\cdot, \cdot)$ (see Goga, 2003). The variable $\mathcal{Y}_1$ is measured on s_1, while the variable $\mathcal{Y}_2$ is measured on s_2. Plugging sample-based estimators $\hat{M}_d$ in $\Delta\theta$ yields the substitution estimator $\widehat{\Delta\theta} = T(\hat{M}_1, \hat{M}_2)$. Unlike the one-sample case, several estimators $\hat{M}_d$ are possible. In what follows, we focus on the general class of *composite estimators* introduced by Goga, Deville and Ruiz-Gazen (2009). We note $s_{1\bullet} = s_1 \setminus s_2$, $s_3 = s_1 \cap s_2$ and $s_{2\bullet} = s_2 \setminus s_1$. For $\Diamond \in \{1\bullet, 3, 2\,\bullet\}$, we note $\pi_{\Diamond,k}$ the expected number of draws for unit k in $s_\Diamond$ and $\hat{M}_{d,\Diamond} = \sum_{k \in s_\Diamond} w_{\Diamond,k} \delta_{y_{dk}}$, where $w_{\Diamond,k} = \pi_{\Diamond,k}^{-1}$. The composite estimators of M_1 and M_2 are

$$\hat{M}_1^{\mathrm{co}}(a) = a \hat{M}_{1,1\bullet} + (1-a) \hat{M}_{1,3} \quad \text{and} \quad \hat{M}_2^{\mathrm{co}}(b) = b \hat{M}_{2,2\bullet} + (1-b) \hat{M}_{2,3}, \tag{3.1}$$

where a and b are some known constants. The choice $a = b = 0$ leads to the *intersection estimator* with $\hat{M}_1^{\mathrm{int}} = \hat{M}_{1,3}$ and $\hat{M}_2^{\mathrm{int}} = \hat{M}_{2,3}$, where the overlapping sample s_3 only is used.

When estimating the parameter $\Delta t = t_{y2} - t_{y1},$ the composite estimator is

$$\widehat{\Delta t}^{\text{co}}(a,b) = \hat{t}_{y_2}^{\text{co}} - \hat{t}_{y_1}^{\text{co}}, \tag{3.2}$$

where $\hat{t}_{y_1}^{\text{co}} = \int y d\hat{M}_1^{\text{co}}(y)$ and $\hat{t}_{y_2}^{\text{co}} = \int y d\hat{M}_2^{\text{co}}(y)$. It may be rewritten as

$$\widehat{\Delta t}^{\text{co}}(a,b) = b\left(\hat{t}_{y_2,s_{2\bullet}} - \hat{t}_{y_2,s_3}\right) - a\left(\hat{t}_{y_1,s_{1\bullet}} - \hat{t}_{y_1,s_3}\right) + \left(\hat{t}_{y_2,s_3} - \hat{t}_{y_1,s_3}\right), \tag{3.3}$$

where $\hat{t}_{y_d,s_\Diamond} = \sum_{k \in s_\Diamond} w_{\Diamond,k} y_{dk}$. The variance of the composite estimator is

$$V\left\{\widehat{\Delta t}^{\text{co}}(a,b)\right\} = (b,-a,1) V\left\{\left(\hat{t}_{y_2,s_{2\bullet}} - \hat{t}_{y_2,s_3}, \hat{t}_{y_1,s_{1\bullet}} - \hat{t}_{y_1,s_3}, \hat{t}_{y_2,s_3} - \hat{t}_{y_1,s_3}\right)^\top\right\}(b,-a,1)^\top. \tag{3.4}$$

Finding the vector $(a_{\text{opt}}, b_{\text{opt}})^\top$ which minimizes the variance in (3.4) leads to the *optimal composite estimator* (Goga, Deville and Ruiz-Gazen, 2009, Section 3.6). Note that this is not an estimator per se, since it depends on unknown quantities which need to be estimated in practice. However, this is a useful benchmark which we will use for the appraisal of simpler composite estimators.

A variance estimator is obtained by substituting in (3.4) an estimator of the variance-covariance matrix. The derivation of variance estimators is detailed in Sections 3.1.1 and 3.1.2 for two examples of two-dimensional sampling designs.

3.1.1 Two-dimensional SI design

The two-dimensional SI design (SI2) of fixed size $(n_{1\bullet}, n_3, n_{2\bullet})$ assigns equal probabilities to all $s = (s_1, s_2)$ for which the associated subsamples $s_{1\bullet}$, s_3 and $s_{2\bullet}$ have the required sizes $n_{1\bullet}$, n_3 and $n_{2\bullet}$, see Goga (2003) and Qualité and Tillé (2008). The SI2 design has the attractive property that the marginal samples $s_{1\bullet}$, s_3 and $s_{2\bullet}$ are SI samples from the population U. Similarly, s_1 is a SI sample of size $n_1 = n_{1\bullet} + n_3$, and s_2 is a SI sample of size $n_2 = n_{2\bullet} + n_3$. For the SI2 sampling design, the composite estimator in (3.3) yields

$$\widehat{\Delta t}^{\text{co}}(a,b) = Nb\left(\bar{y}_{2,s_{2\bullet}} - \bar{y}_{2,s_3}\right) - Na\left(\bar{y}_{1,s_{1\bullet}} - \bar{y}_{1,s_3}\right) + N\left(\bar{y}_{2,s_3} - \bar{y}_{1,s_3}\right), \tag{3.5}$$

and the variance of the composite estimator is

$$V\left\{\widehat{\Delta t}^{\text{co}}(a,b)\right\} = N^2\left\{c_1(a) S_{y_1,U}^2 - 2c_{12}(a,b) S_{y_1y_2,U} + c_2(b) S_{y_2,U}^2\right\}, \tag{3.6}$$

with

$$\begin{aligned} c_1(a) &= \frac{(1-a)^2}{n_3} + \frac{a^2}{n_1 - n_3} - \frac{1}{N}, \\ c_2(b) &= \frac{(1-b)^2}{n_3} + \frac{b^2}{n_2 - n_3} - \frac{1}{N}, \\ c_{12}(a,b) &= \frac{(1-a)(1-b)}{n_3} - \frac{1}{N}, \end{aligned}$$

see Appendix for a proof.

We consider two examples. The choice $a = b = 0$ leads to the intersection estimator

$$\widehat{\Delta t}^{\text{int}} = \widehat{\Delta t}^{\text{co}}(0,0) = \frac{N}{n_3}\sum_{k\in s_3}(y_{2k} - y_{1k}), \tag{3.7}$$

and the variance simplifies as

$$V\left\{\widehat{\Delta t}^{\text{int}}\right\} = N^2\left(\frac{1}{n_3} - \frac{1}{N}\right)S^2_{y_2-y_1,U}. \tag{3.8}$$

The choice $a = n_1^{-1}n_{1\bullet}$ and $b = n_2^{-1}n_{2\bullet}$ leads to the *union estimator*

$$\widehat{\Delta t}^{\text{uni}} = \widehat{\Delta t}^{\text{co}}\left(n_1^{-1}n_{1\bullet}, n_2^{-1}n_{2\bullet}\right) = \frac{N}{n_2}\sum_{k\in s_2} y_{2k} - \frac{N}{n_1}\sum_{k\in s_1} y_{1k} \tag{3.9}$$

where the complete samples are used, and the variance may be written as

$$V\left\{\widehat{\Delta t}^{\text{uni}}\right\} = N^2\left\{\left(\frac{1}{n_1} - \frac{1}{N}\right)S^2_{y_1,U} - 2\left(\frac{n_3}{n_1\,n_2} - \frac{1}{N}\right)S_{y_1y_2,U} + \left(\frac{1}{n_2} - \frac{1}{N}\right)S^2_{y_2,U}\right\}. \tag{3.10}$$

The variances of the union estimator and of the intersection estimator were derived by Qualité and Tillé (2008), see also Tam (1984).

The choice of a and b is of practical importance to obtain an efficient composite estimator. After some algebra, the vector $(a_{\text{opt}}, b_{\text{opt}})^\top$ which minimizes the variance of $\widehat{\Delta t}^{\text{co}}(a,b)$ is given by

$$(a_{\text{opt}}, b_{\text{opt}})^\top = A^{-1}X \tag{3.11}$$

with

$$A = \begin{pmatrix} \dfrac{n_1}{n_1 - n_3} & -\dfrac{S_{y_1y_2,U}}{S^2_{y_1,U}} \\ -\dfrac{S_{y_1y_2,U}}{S^2_{y_2,U}} & \dfrac{n_2}{n_2 - n_3} \end{pmatrix} \quad \text{and} \quad X = \left(1 - \frac{S_{y_1y_2,U}}{S^2_{y_1,U}}, 1 - \frac{S_{y_1y_2,U}}{S^2_{y_2,U}}\right)^\top. \tag{3.12}$$

For two variables $\mathcal{Y}_1$ and $\mathcal{Y}_2$ related to a same characteristic collected at two different times, $S_{y_1y_2,U}$ is expected to be close to $S^2_{y_1,U}$ and $S^2_{y_2,U}$. The vector X in (3.12) is in turn close to the null vector, and if the size of the overlapping sample s_3 is comparable to that of $s_{1\bullet}$ and $s_{2\bullet}$ we obtain $a_{\text{opt}} \simeq 0$ and $b_{\text{opt}} \simeq 0$. Therefore, using the intersection estimator where $a = b = 0$ seems reasonable in practice. On the contrary, the union estimator can be very inefficient; see Section 4.2 for an illustration. These conclusions are consistent with that of Qualité and Tillé (2008), Section 2.2.2.

Several variance estimators may be used for the composite estimator. Estimating the dispersions on the overlapping sample only yields the unbiased variance estimator

$$v^{\text{HT}}_{\text{int}}\left\{\widehat{\Delta t}^{\text{co}}(a,b)\right\} = N^2\left\{c_1(a)\,S^2_{y_1,s_3} - 2c_{12}(a,b)\,S_{y_1y_2,s_3} + c_2(b)\,S^2_{y_2,s_3}\right\}, \tag{3.13}$$

while an estimation on the whole samples yields

$$v_{\text{uni}}^{\text{HT}}\left\{\widehat{\Delta t}^{\text{co}}(a,b)\right\} = N^2\left\{c_1(a)S_{y_1,s_1}^2 - 2c_{12}(a,b)S_{y_1y_2,s_3} + c_2(b)S_{y_2,s_2}^2\right\}. \tag{3.14}$$

Berger (2004) considered variance estimation for the union estimator under a maximum entropy rotating sampling scheme, by estimating separately the three components in (3.6).

3.1.2 Two-dimensional multistage design

We now consider a two-dimensional two-stage sampling design (MULT2). We assume that a with-replacement first-stage sample s_I of size m is first selected among the PSUs $U_1,\ldots,U_{N_I}$. Inside each PSU $i \in s_I$, a SI2 sample of size $(n_{1\bullet}^i, n_3^i, n_{2\bullet}^i)$ is then selected. This type of sampling design emerges in particular in case of a self-weighted two-stage design in two waves, with a partial replacement at the second wave of the SSUs selected at the first wave. The composite estimator in (3.3) yields

$$\widehat{\Delta t}^{\text{co}}(a,b) = \sum_{i\in s_I}\pi_{Ii}^{-1}\widehat{\Delta t}^{i,\text{co}}(a,b) \tag{3.15}$$

where

$$\widehat{\Delta t}^{i,\text{co}}(a,b) = N_i b\left(\overline{y}_{2,s_{2\bullet}^i} - \overline{y}_{2,s_3^i}\right) - N_i a\left(\overline{y}_{1,s_{1\bullet}^i} - \overline{y}_{1,s_3^i}\right) + N_i\left(\overline{y}_{2,s_3^i} - \overline{y}_{1,s_3^i}\right), \tag{3.16}$$

where $\overline{y}_{d,s_\Diamond^i} = (n_\Diamond^i)^{-1}\sum_{k\in s_\Diamond^i} y_{\Diamond k}$, where $s_\Diamond^i = s_\Diamond \cap U_i$, and where N_i denotes the number of SSUs inside the PSU u_i.

For example, using the overlapping samples only inside the PSUs yields the intersection estimator

$$\widehat{\Delta t}^{\text{int}} = \sum_{i\in s_I}\pi_{Ii}^{-1}\widehat{\Delta t}^{i,\text{int}} \quad \text{with} \quad \widehat{\Delta t}^{i,\text{int}} = N_i\left(\overline{y}_{2,s_3^i} - \overline{y}_{1,s_3^i}\right). \tag{3.17}$$

Using the complete samples inside the PSUs yields the union estimator

$$\widehat{\Delta t}^{\text{uni}} = \sum_{i\in s_I}\pi_{Ii}^{-1}\widehat{\Delta t}^{i,\text{uni}} \quad \text{with} \quad \widehat{\Delta t}^{i,\text{uni}} = N_i\left(\overline{y}_{2,s_2^i} - \overline{y}_{1,s_1^i}\right). \tag{3.18}$$

We note that for any vector of values $(a,b)^\top$, the variance due to the first-stage of sampling for $\widehat{\Delta t}^{\text{co}}(a,b)$ is the same. The possible composite estimators thus differ with respect to the second-stage variance only. In view of the discussion in Section 3.1.1, we therefore expect the intersection estimator to be close to the optimal composite estimator; see Section 4.2 for an illustration. An unbiased variance estimator for $\widehat{\Delta t}^{\text{co}}(a,b)$ is given by

$$v^{\text{HH}}\left\{\widehat{\Delta t}^{\text{co}}(a,b)\right\} = \frac{m}{m-1}\sum_{i\in s_I}\left(\frac{\widehat{\Delta t}^{i,\text{co}}(a,b)}{\pi_{Ii}} - \frac{\widehat{\Delta t}^{\text{co}}(a,b)}{m}\right)^2. \tag{3.19}$$

3.2 Estimation of the change between Gini indexes

The change between Gini indexes $\Delta G = G_2 - G_1$ may be written as

$$\Delta G = \frac{\int \{2F_{2N}(y) - 1\}\, y dM_2(y)}{\int y dM_2(y)} - \frac{\int \{2F_{1N}(y) - 1\}\, y dM_1(y)}{\int y dM_1(y)} \tag{3.20}$$

where $F_{dN}(y) = N^{-1} \sum_{k \in U} 1_{\{y_{dk} \le y\}},\ d = 1, 2.$ Using composite estimation leads to

$$\widehat{\Delta G}^{\text{co}}(a,b) = \frac{\int \{2\hat{F}_{2N}^{\text{co}}(y) - 1\}\, y d\hat{M}_2^{\text{co}}(y)}{\int y d\hat{M}_2^{\text{co}}(y)} - \frac{\int \{2\hat{F}_{1N}^{\text{co}}(y) - 1\}\, y d\hat{M}_1^{\text{co}}(y)}{\int y d\hat{M}_1^{\text{co}}(y)} \tag{3.21}$$

where $\hat{F}_{dN}^{\text{co}}(y) = \left\{ \int d\hat{M}_d^{\text{co}}(y) \right\}^{-1} \int 1_{\{\xi \le y\}}\, d\hat{M}_d^{\text{co}}(\xi).$

Usually, in a temporal sampling framework, the samples s_1 and s_2 are not independent. Consequently, our set-up differs from the usual estimation of functionals depending on distribution functions estimated with independent samples; see for example Pires and Branco (2002) and Reid (1981), who give the first-order expansion of a two-sample functional using the partial influence functions. Davison and Hinkley (1997, page 71) give bootstrap methods under a similar framework. Using a general two-dimensional sampling design $p(\cdot, \cdot),$ Goga, Deville and Ruiz-Gazen (2009) give a two-sample linearization technique of bivariate functionals that will be used in what follows.

3.3 Linearization variance estimation

To obtain the asymptotic variance of $\widehat{\Delta\theta}^{\text{co}}(a,b),$ we adopt the asymptotic framework introduced by Goga, Deville and Ruiz-Gazen (2009), which is an extension to the two-sample case of the asymptotic framework of Isaki and Fuller (1982). Define, when they exist, the *partial influence functions* of a functional $T(M_1, M_2)$ at point y as

$$\begin{aligned} I_1 T(M_1, M_2; y) &= \lim_{h \to 0} \frac{T(M_1 + h\delta_y, M_2) - T(M_1, M_2)}{h}, \\ I_2 T(M_1, M_2; y) &= \lim_{h \to 0} \frac{T(M_1, M_2 + h\delta_y) - T(M_1, M_2)}{h}. \end{aligned}$$

We define the *linearized variables* $u_{dk} = I_d T(M_1, M_2; y_{dk})$ for $d = 1, 2$ as the partial influence functions of T at (M_1, M_2) and $y = y_{dk}.$ For the change between Gini indexes $\Delta G,$ the linearized variables u_{dk} may be computed using (2.10), namely

$$u_{dk} = 2F_{dN}(y_{dk}) \frac{y_{dk} - \overline{y}_{dk,U<}}{t_{y_d}} - y_{dk} \frac{G_d + 1}{t_{y_d}} + \frac{1 - G_d}{N}, \tag{3.22}$$

where $\overline{y}_{dk,U<} = \left(\sum_{l \in U} 1_{\{y_{dl} < y_{dk}\}} \right)^{-1} \sum_{j \in U} y_{dj} 1_{\{y_{dj} < y_{dk}\}}.$ The estimated linearized variable is

$$\hat{u}_{dk} = 2\hat{F}_{dN}^{\text{co}}(y_{dk}) \frac{y_{dk} - \overline{y}_{dk,s<}^{\text{co}}}{\hat{t}_{y1}^{\text{co}}} - y_{dk} \frac{\hat{G}_d^{\text{co}} + 1}{\hat{t}_{y1}^{\text{co}}} + \frac{1 - \hat{G}_d^{\text{co}}}{\hat{N}}. \tag{3.23}$$

3.3.1 Two-dimensional SI design

In case of the SI2 design presented in Section 3.1.1, plugging the variables u_{dk} derived in (3.22) into the variance formula in (3.6) yields the variance approximation

$$V\left\{\widehat{\Delta G}^{\text{co}}(a,b)\right\} \simeq N^2\left\{c_1(a)\,S^2_{u_1,U} - 2c_{12}(a,b)\,S_{u_1u_2,U} + c_2(b)\,S^2_{u_2,U}\right\},$$

see Theorem 1 in Goga, Deville and Ruiz-Gazen (2009). To obtain a variance estimator, the linearized variables may be estimated in several ways. If the overlapping sample s_3 only is used, the estimated linearized variables $\hat{u}_d$ are obtained from (3.23) by taking $\hat{M}_1^{\text{co}} = \hat{M}_{1,3}$ and $\hat{M}_2^{\text{co}} = \hat{M}_{2,3}$. A variance estimator is then obtained by plugging these linearized variables into (3.13). This leads to

$$v_{\text{int}}^{\text{HT}}\left\{\widehat{\Delta G}^{\text{co}}(a,b)\right\} = N^2\left\{c_1(a)\,S^2_{\hat{u}_1,s_3} - 2c_{12}(a,b)\,S_{\hat{u}_1\hat{u}_2,s_3} + c_2(b)\,S^2_{\hat{u}_2,s_3}\right\}. \quad (3.24)$$

If the whole samples s_1 and s_2 are used, the estimated linearized variable $\hat{u}_d$ are obtained from (3.23) by taking $\hat{M}_1^{\text{co}} = \hat{M}_{1,1}$ and $\hat{M}_2^{\text{co}} = \hat{M}_{2,2}$. A variance estimator is then obtained by plugging these linearized variables into (3.14). This leads to

$$v_{\text{uni}}^{\text{HT}}\left\{\widehat{\Delta G}^{\text{co}}(a,b)\right\} = N^2\left\{c_1(a)\,S^2_{\hat{u}_1,s_1} - 2c_{12}(a,b)\,S_{\hat{u}_1\hat{u}_2,s_3} + c_2(b)\,S^2_{\hat{u}_2,s_2}\right\}. \quad (3.25)$$

3.3.2 Two-dimensional multistage design

In case of the MULT2 design presented in Section 3.1.2, the linearized variables may also be estimated in several ways. For the sake of simplicity, we consider using the overlapping sample s_3 only so that the estimated linearized variables $\hat{u}_d$ are obtained from (3.23) by taking $\hat{M}_1^{\text{co}} = \hat{M}_{1,3}$ and $\hat{M}_2^{\text{co}} = \hat{M}_{2,3}$. A variance estimator is then obtained by plugging these linearized variables into (3.19). This leads to

$$v^{\text{HH}}\left\{\widehat{\Delta G}^{\text{co}}(a,b)\right\} = \frac{m}{m-1}\sum_{i\in s_I}\left(\frac{\widehat{\Delta u}^{i,\text{co}}(a,b)}{\pi_{Ii}} - \frac{\widehat{\Delta u}^{\text{co}}(a,b)}{m}\right)^2, \quad (3.26)$$

where $\widehat{\Delta u}^{\text{co}}(a,b)$ and $\widehat{\Delta u}^{i,\text{co}}(a,b)$ are obtained from (3.15) and (3.16), respectively, by replacing y_{dk} with $\hat{u}_{dk}$.

3.4 Bootstrap variance estimation

Bootstrap methods have not yet been studied for the change between Gini indexes. The principles of the weighted bootstrap technique can be extended to the two-sample context, i.e. each measure $\hat{M}_{d,\Diamond}$ with $d = 1, 2$ and $\Diamond \in \{1\bullet, 3, 2\bullet\}$ is estimated, conditionally on the samples originally selected, by some weighted bootstrap measure $\hat{M}^*_{d,\Diamond}$ which enables to match, at least approximately, the two first moments of an unbiased estimator in the linear case. In Section 3.4.1, we consider a generalization of the BWO to the SI2 design. In Section 3.4.2, we propose a generalisation of the BWR to the MULT2 design.

3.4.1 A generalization of the BWO to the SI2 design

We first consider the SI2 design. Building a pseudo-population U^* is more intricate in the two-sample case, since the variables of interest measured at waves τ_1 and τ_2 need to be available for each unit in U^*. We therefore describe a bootstrap algorithm where the overlapping sample s_3 only is used to build the pseudo-population U^*, in the spirit of the intersection variance estimator in (3.24).

Suppose that N/n_3 is an integer. The vectors $D_\Diamond$ are obtained by, first creating a pseudo-population U^* of size N by duplicating N/n_3 times each unit k in the original sample s_3. A SI2 resample $s^* = (s_{1\bullet}^*, s_3^*, s_{\bullet 2}^*)$ of size $(n_{1\bullet}, n_3, n_{2\bullet})$ is then selected in U^*. The bootstrap measures are then

$$\hat{M}_{d,\Diamond}^* = \sum_{k \in s_3} w_{\Diamond,k} D_{\Diamond,k} \delta_{y_{dk}}, \tag{3.27}$$

with $D_{\Diamond,k}$ the number of times that unit k is selected in the resample $s_\Diamond^*$. In the linear case, the bootstrap estimator of the parameter Δt is then

$$\widehat{\Delta t}^{\text{co}*}(a,b) = b\left(\hat{t}_{y2,s_{2\bullet}^*} - \hat{t}_{y2,s_3^*}\right) - a\left(\hat{t}_{y1,s_{1\bullet}^*} - \hat{t}_{y1,s_3^*}\right) + \left(\hat{t}_{y2,s_3^*} - \hat{t}_{y1,s_3^*}\right), \tag{3.28}$$

where $\hat{t}_{y_d,s_\Diamond^*} = \sum_{k \in s_3} w_{\Diamond,k} D_{\Diamond,k} y_{dk}$. After some algebra, we obtain

$$E_*\left\{\widehat{\Delta t}^{\text{co}*}(a,b)\right\} = \widehat{\Delta t}^{\text{int}} \quad \text{and} \quad V_*\left\{\widehat{\Delta t}^{\text{co}*}(a,b)\right\} = \frac{1-n_3^{-1}}{1-N^{-1}} v_{\text{int}}^{\text{HT}}\left\{\widehat{\Delta t}^{\text{co}}(a,b)\right\}, \tag{3.29}$$

where $\widehat{\Delta t}^{\text{int}}$ is given in (3.7), and $v_{\text{int}}^{\text{HT}}\left(\hat{t}_{y1}^{\text{HT}}\right)$ is given in (3.13). The proposed generalization of the BWO therefore enables to exactly match the intersection estimator of the first moment, and to approximately match the intersection estimator of the second moment for a large n_3.

The building of U^* may be avoided by noting that under the BWO procedure, each vector $D_\Diamond$ follows a multivariate hypergeometric distribution. Therefore, the resampling weights may be directly generated. The algorithm may be adapted to the general case when N/n_3 is not an integer by means of any of the techniques mentioned in Section 2.4.

3.4.2 A generalization of the BWR for the two-dimensional multistage design

We now consider the two-dimensional two-stage sampling design with a common first-stage sample s_I presented in Section 3.1.2. The proposed bootstrap procedure is similar to that described in Rao and Wu (1988). A with-replacement resample s_I^* of size $m-1$ is selected by means of simple random sampling with replacement (SIR) in the original first-stage sample s_I. The bootstrap measures are then

$$\hat{M}_{d,\Diamond}^* = \frac{m}{m-1} \sum_{i \in s_I^*} \sum_{k \in s_\Diamond^i} \pi_{Ii}^{-1} \pi_{\Diamond k|i}^{-1} \delta_{y_{dk}} \quad \text{where} \quad \pi_{\Diamond k|i} = \frac{n_\Diamond^i}{N_i}. \tag{3.30}$$

It may be rewritten as

$$\hat{M}_{d,\Diamond}^* = \sum_{k \in s_\Diamond} w_{\Diamond,k} D_{\Diamond,k} \delta_{y_{dk}}, \tag{3.31}$$

with $s_{\Diamond}$ the union of the samples $s_{\Diamond}^{i}$ for $i \in s_I$, and where the resampling weight $D_{\Diamond,k}$ equals $m(m-1)^{-1}$ multiplied by the number of times the PSU containing k is selected in s_I^*.

In the linear case, the bootstrap estimator of the parameter Δt is then

$$\widehat{\Delta t}^{\text{co}*}(a,b) = \frac{m}{m-1} \sum_{i \in s_I^*} \pi_{Ii}^{-1} \widehat{\Delta t}^{i,\text{co}}(a,b) \tag{3.32}$$

where $\widehat{\Delta t}^{i,\text{co}}(a,b)$ is defined in (3.16). After some algebra, we obtain

$$E_* \left\{ \widehat{\Delta t}^{\text{co}*}(a,b) \right\} = \widehat{\Delta t}^{\text{co}}(a,b) \quad \text{and} \quad V_* \left\{ \widehat{\Delta t}^{\text{co}*}(a,b) \right\} = v^{\text{HH}} \left\{ \widehat{\Delta t}^{\text{co}}(a,b) \right\}, \tag{3.33}$$

where $\widehat{\Delta t}^{\text{co}}(a,b)$ is given in (3.15), and $v^{\text{HH}} \left\{ \widehat{\Delta t}^{\text{co}}(a,b) \right\}$ is given in (3.19). The proposed generalization of the BWR therefore enables to exactly match the composite estimator of the first moment, and the associated estimator of the second moment.

4 Simulation study

In this section, five artificial populations are first generated as described in Section 4.1. In Section 4.2, the union estimator is compared with the intersection estimator in terms of asymptotic variance. A Monte Carlo experiment is then presented in Section 4.3, and the performances of the linearization and the bootstrap are compared in case of a SI2 sampling design. A similar comparison is made in Section 4.4, in case of the bi-dimensional two-stage sampling design.

4.1 Simulation set-up

We generated 5 finite populations of size $N = 40{,}000$, each containing two study variables y_1 and y_2. The y_{1k} values and the y_{2k} values were generated according to the lognormal model

$$y_{dk} = \exp(\alpha_d \, \varepsilon_k). \tag{4.1}$$

The ε_k's were generated according to a standard normal distribution. The values of the Gini coefficients for the five populations are presented in Table 4.1.

Table 4.1
Gini coefficients for 5 populations

Population	Pop. 1	Pop. 2	Pop. 3	Pop. 4	Pop. 5
G_1	0.249	0.298	0.348	0.397	0.447
G_2	0.259	0.318	0.378	0.437	0.496
ΔG	0.010	0.020	0.030	0.040	0.049

In each of the 5 populations, the units were grouped into $M = 500$ clusters of equal size $N_0 = 80$. The clusters were built so that the intra-cluster correlation coefficient with respect to the variable y_1 was approximately equal to 0.20 in each population.

4.2 Comparison of the union estimator and of the intersection estimator

In this section, we compare the union estimator with the intersection estimator for the change between Gini indexes in terms of asymptotic variance. We consider two sampling designs: the SI2 design presented in Section 3.1.1 with $(n_{1\bullet}, n_3, n_{2\bullet}) =$ (1,000; 1,000; 1,000), (1,000; 2,000; 1,000) or (1,000; 4,000; 1,000); the MULT2 design presented in Section 3.1.2 with $m = 300$ and $(n_{1\bullet}^i, n_3^i, n_{2\bullet}^i) =$ (10; 10; 10), (10; 20; 10) or (10; 40; 10).

For each population, we compute the asymptotic variance $V_{\text{lin}}(\widehat{\Delta G}^{\text{uni}})$ of the union estimator, and the asymptotic variance $V_{\text{lin}}(\widehat{\Delta G}^{\text{int}})$ of the intersection estimator. So as to compare them, we compute the relative efficiency defined as

$$\text{RE}\left\{\widehat{\Delta G}^{\cdot}\right\} = \frac{V_{\text{lin}}\left\{\widehat{\Delta G}^{(\cdot)}\right\}}{V_{\text{lin}}\left\{\widehat{\Delta G}^{\text{opt}}\right\}}, \tag{4.2}$$

with $\widehat{\Delta G}^{\text{opt}}$ the optimal estimator.

The results are presented in Table 4.2. The union estimator is highly inefficient. Its asymptotic variance is 15 to 244 times higher than that of the intersection estimator for SI2, and 2 to 44 times higher than that of the intersection estimator for MULT2. The difference between both estimators tends to decrease when the sample size of the common sample increases and/or when ΔG increases. On the other hand, the intersection estimator is slightly less efficient than the optimal estimator for SI2, with RE ranging from 1.33 to 2.46, and approximately as efficient as the optimal estimator for MULT2, with RE ranging from 1.02 to 1.12. This supports the heuristic reasoning in Section 3.1.1. In view of the poor performance of the union estimator, and of the good performance of the intersection estimator, we confine our attention to the latter in the remainder of the simulation study.

Table 4.2
Relative efficiency of the union estimator and of the intersection variance estimator for 5 populations

Design	Sample size	Pop. 1		Pop. 2		Pop. 3		Pop. 4		Pop. 5	
		$\widehat{\Delta G}^{\text{uni}}$	$\widehat{\Delta G}^{\text{int}}$	$\widehat{\Delta G}^{\text{uni}}$	$\widehat{\Delta G}^{\text{int}}$	$\widehat{\Delta G}^{\text{uni}}$	$\widehat{\Delta G}^{\text{int}}$	$\widehat{\Delta G}^{\text{uni}}$	$\widehat{\Delta G}^{\text{int}}$	$\widehat{\Delta G}^{\text{uni}}$	$\widehat{\Delta G}^{\text{int}}$
SI2	$n_3 = 1{,}000$	600.22	2.46	200.23	2.27	96.72	2.10	58.73	1.96	39.35	1.85
	$n_3 = 2{,}000$	410.23	1.84	141.71	1.76	70.71	1.68	44.18	1.61	30.33	1.54
	$n_3 = 4{,}000$	250.02	1.47	88.40	1.43	45.17	1.40	28.86	1.36	20.23	1.33
MULT2	$n_3^i = 10$	49.10	1.12	19.89	1.13	11.83	1.14	8.84	1.15	7.28	1.16
	$n_3^i = 20$	23.08	1.05	9.75	1.05	6.08	1.05	4.73	1.06	4.04	1.07
	$n_3^i = 40$	9.15	1.02	4.25	1.02	2.90	1.02	2.41	1.02	2.16	1.02

4.3 Comparison of linearization and bootstrap for the SI2 design

In this section, we compare the linearization and bootstrap for variance estimation and for producing confidence intervals, in case of the intersection estimator for the change between Gini indexes under the SI2 sampling design. From each population, we selected $B = 10{,}000$ two-dimensional samples by means of the SI2 design indexed by $(n_{1\bullet}, n_3, n_{2\bullet}) = (1{,}000; 1{,}000; 1{,}000)$, $(n_{1\bullet}, n_3, n_{2\bullet}) = (1{,}000; 2{,}000; 1{,}000)$ or $(n_{1\bullet}, n_3, n_{2\bullet}) = (1{,}000; 4{,}000; 1{,}000)$. In each sample, we computed the intersection estimator $\widehat{\Delta G}^{\text{int}}$ of the change between Gini indexes. For this estimator, we computed (i) the linearization variance estimator $v_{\text{int}}(\widehat{\Delta G}^{\text{int}})$ given in (3.24), and (ii) the Bootstrap variance estimator $v_{\text{BWO}}(\widehat{\Delta G})$, following the Bootstrap procedure described in Section 3.4.1.

To measure the bias of a variance estimator $v(\widehat{\Delta G})$, we used the Monte Carlo Percent Relative Bias

$$\text{RB}\left\{v\left(\widehat{\Delta G}\right)\right\} = 100 \times \frac{B^{-1}\sum_{b=1}^{B} v\left(\widehat{\Delta G}_b\right) - \text{MSE}\left(\widehat{\Delta G}\right)}{\text{MSE}\left(\widehat{\Delta G}\right)}, \tag{4.3}$$

where $v(\widehat{\Delta G}_b)$ denotes the estimator $v(\widehat{\Delta G})$ in the b^{th} sample, and $\text{MSE}(\widehat{\Delta G})$ is a simulation-based approximation of the true mean square error of $\widehat{\Delta G}$, obtained from an independent run of 100,000 simulations. As a measure of stability of $v(\widehat{\Delta G})$, we used the Relative Stability

$$\text{RS}\left\{v\left(\widehat{\Delta G}\right)\right\} = \frac{\left[B^{-1}\sum_{b=1}^{B}\left\{v\left(\widehat{\Delta G}\right) - \text{MSE}\left(\widehat{\Delta G}\right)\right\}^2\right]^{1/2}}{\text{MSE}\left(\widehat{\Delta G}\right)}. \tag{4.4}$$

Finally, we compared the coverage rates of (i) the normality-based confidence interval with use of the linearization variance estimator and (ii) the confidence interval associated to the percentile Bootstrap. The bootstrap variance estimators and the bootstrap confidence intervals are based on $C = 1{,}000$ bootstrap replications. Error rates of the confidence intervals (with nominal one-tailed error rate of 2.5% in each tail) are compared. The comparison with nominal error rate of 5% gave no qualitative difference and is thus omitted.

The results are presented in Table 4.3. Both variance estimators are negatively biased. This bias is moderate (less than 5%) in most cases, except for the smaller sample size $n = 1{,}000$, and for the population U_5 with the highest value of ΔG. The bootstrap variance estimator is systematically slightly more biased than the linearization variance estimator, but the difference decreases as the sample size increases. For both variance estimators, the instability increases with ΔG. The Bootstrap variance estimator is slightly more stable for the smaller sample size $n = 1{,}000$, but the situation is reversed when the sample size increases. Turning to the coverage of the confidence intervals, both methods lead to under-coverage which is consistent with the negative bias of both variance estimators. The normality-based confidence intervals show a slightly better coverage than the bootstrap percentile confidence intervals. For both confidence intervals, the under-coverage is more acute when ΔG increases, and reduces when the sample size increases.

Table 4.3
Relative Bias, Relative Stability and Nominal One-Tailed Error Rates for linearization and Bootstrap variance estimation of the intersection estimator of the change between Gini indexes for 5 populations and with the SI2 sampling design

Pop.	Linearization					Bootstrap				
	RB	RS	L	U	L+U	RB	RS	L	U	L+U
				Sample size $(n_{1\bullet}, n_3, n_{2\bullet}) = (1,000; 1,000; 1,000)$						
Pop. 1	-1.41	24.6	1.8	4.5	6.3	-1.83	24.6	1.8	4.9	6.7
Pop. 2	-1.98	32.4	1.6	5.2	6.8	-2.64	32.1	1.7	5.9	7.6
Pop. 3	-2.80	41.9	1.3	6.3	7.7	-3.83	40.9	1.3	7.0	8.3
Pop. 4	-4.00	52.5	1.0	7.7	8.7	-5.57	50.6	1.1	8.2	9.3
Pop. 5	-5.80	64.0	1.0	9.2	10.1	-8.11	60.6	0.8	9.9	10.7
				Sample size $(n_{1\bullet}, n_3, n_{2\bullet}) = (1,000; 2,000; 1,000)$						
Pop. 1	-1.38	17.3	1.6	3.7	5.3	-1.67	17.8	1.8	4.1	5.9
Pop. 2	-1.64	23.0	1.4	4.3	5.8	-2.05	23.2	1.4	4.7	6.1
Pop. 3	-1.99	30.1	1.2	5.0	6.2	-2.58	30.0	1.1	5.3	6.4
Pop. 4	-2.50	38.4	1.0	6.0	6.9	-3.38	37.9	1.0	6.3	7.3
Pop. 5	-3.30	47.9	0.7	7.2	7.9	-4.62	46.7	0.7	7.5	8.2
				Sample size $(n_{1\bullet}, n_3, n_{2\bullet}) = (1,000; 4,000; 1,000)$						
Pop. 1	-0.60	11.9	2.0	3.4	5.3	-0.68	12.8	2.1	3.4	5.5
Pop. 2	-0.67	15.9	1.8	3.7	5.6	-0.80	16.5	2.0	3.9	5.9
Pop. 3	-0.83	20.8	1.8	4.4	6.2	-1.03	21.3	1.9	4.4	6.3
Pop. 4	-1.13	26.7	1.5	5.0	6.6	-1.46	26.9	1.6	5.0	6.6
Pop. 5	-1.64	33.4	1.4	5.8	7.1	-2.18	33.5	1.4	5.8	7.1

4.4 Comparison of linearization and bootstrap for the MULT2 design

In this section, we compare the linearization and bootstrap for variance estimation and for producing confidence intervals, in case of the intersection estimator for the change between Gini indexes under the MULT2 sampling design presented in Section 3.1.2. From each population, we selected B = 10,000 two-dimensional two-stage samples by means of the MULT2 design indexed by $m = 300$ and $(n_{1\bullet}^i, n_3^i, n_{2\bullet}^i) = (10; 10; 10)$, $(10; 20; 10)$ or $(10; 40; 10)$. In each sample, we computed the intersection estimator $\widehat{\Delta G}^{\text{int}}$ of the change between Gini indexes. For this estimator, we computed (i) the linearization variance estimator $v^{\text{HH}}\{\widehat{\Delta G}^{\text{co}}(a, b)\}$ given in (3.26), and (ii) the Bootstrap variance estimator $v_{\text{BWR}}(\widehat{\Delta G}^{\text{int}})$, following the Bootstrap procedure described in Section 3.4.2.

To measure the bias of a variance estimator $v(\widehat{\Delta G})$, we used the Monte Carlo Percent Relative Bias defined in equation (4.3), and the Relative Stability defined in equation (4.4). The true mean square error of $\widehat{\Delta G}$ was obtained from an independent run of 100,000 simulations. Also, we compared the coverage rates of (i) the normality-based confidence interval with use of the linearization variance estimator and (ii) the confidence interval associated to the percentile Bootstrap. The bootstrap variance estimators and the bootstrap confidence intervals are based on C = 1,000 bootstrap replications. Error rates of the confidence intervals (with nominal one-tailed error rate of 2.5% in each tail) are compared. The comparison with nominal error rate of 5% gave no qualitative difference and is thus omitted.

The results are presented in Table 4.4. Both variance estimators are approximately unbiased for small values of ΔG, but show a moderate negative bias which increases with ΔG. The bootstrap variance estimator is more biased than the linearization variance estimator. For both variance estimators, the instability increases with ΔG. The Bootstrap variance estimator is slightly more stable than the linearization variance estimator. Both methods lead to an under-coverage which is consistent with the negative bias of both variance estimators. The normality-based confidence intervals perform slightly better. For both confidence intervals, the under-coverage is more acute when ΔG increases, and reduces when the sample size increases.

Table 4.4
Relative Bias, Relative Stability and Nominal One-Tailed Error Rates for linearization and Bootstrap variance estimation of the intersection estimator of the Gini Coefficient Change for 5 populations and with the MULT2 sampling design

Pop.	**Linearization**					**Bootstrap**				
	RB	**RS**	**L**	**U**	**L+U**	**RB**	**RS**	**L**	**U**	**L+U**
Sample sizes $m = 300$ and $(n^i_{1\bullet}, n^i_3, n^i_{2\bullet}) = (10; 10; 10)$										
Pop. 1	1.23	33.8	0.6	4.9	5.5	1.09	33.2	0.6	6.0	6.6
Pop. 2	0.64	41.1	0.8	5.5	6.3	-0.20	39.7	0.6	6.5	7.1
Pop. 3	-0.42	48.7	0.7	7.1	7.8	-2.05	46.6	0.7	8.4	9.1
Pop. 4	-2.07	56.4	0.8	8.4	9.2	-4.47	53.3	0.6	9.6	10.2
Pop. 5	-4.44	63.7	0.9	9.2	10.1	-7.56	59.5	0.4	10.3	10.7
Sample sizes $m = 300$ and $(n^i_{1\bullet}, n^i_3, n^i_{2\bullet}) = (10; 20; 10)$										
Pop. 1	1.70	32.6	1.5	4.9	6.4	-1.70	32.3	1.5	6.0	7.5
Pop. 2	1.10	39.0	1.4	5.4	6.8	-1.91	38.3	1.5	6.9	8.4
Pop. 3	0.17	45.6	1.2	7.4	8.6	-2.49	44.4	1.1	7.7	8.8
Pop. 4	-1.17	52.0	1.0	9.0	10.0	-3.58	50.3	0.8	9.7	10.5
Pop. 5	-3.03	57.9	0.9	10.4	11.3	-5.35	55.4	0.7	11.0	11.7
Sample sizes $m = 300$ and $(n^i_{1\bullet}, n^i_3, n^i_{2\bullet}) = (10; 40; 10)$										
Pop. 1	-0.99	32.1	1.2	6.1	7.3	-3.21	32.2	1.7	6.7	8.4
Pop. 2	-1.68	38.3	1.4	6.7	8.1	-3.70	38.3	1.4	7.6	9.0
Pop. 3	-2.58	44.6	1.3	7.5	8.8	-4.40	44.5	1.2	8.9	10.1
Pop. 4	-3.78	50.6	1.1	8.9	10.0	-5.50	50.1	0.9	10.6	11.5
Pop. 5	-5.39	55.9	0.8	10.9	11.7	-7.16	54.8	0.6	12.8	13.4

5 Conclusion

In this paper, we considered the estimation of the change between Gini indexes. We presented the class of composite estimators introduced by Goga, Deville and Ruiz-Gazen (2009), and studied more particularly the intersection estimator which makes use of the common sample only, and the union estimator which makes use of the whole available samples. We justified both heuristically and through the simulation study in Section 4.2 that the intersection estimator can be close to the optimal estimator, while the union estimator exhibits poor performances in all the scenarios considered. The intersection estimator is also easy to compute, while the optimal estimator involves unknown quantities which need to be estimated in practice. We therefore advocate for the use of the intersection estimator for estimating the change between Gini indexes.

We also compared linearization and bootstrap for variance estimation and for producing confidence intervals. In the scenarios that we considered in the simulation study, the linearization performed better with usually smaller relative biases for the variance estimator, and better coverage rates with normality-based confidence intervals than with percentile confidence intervals. Bootstrap$-t$ confidence intervals (not considered in the simulation study) would be a competitor of interest, but due to the intensive computational work involved, they are less attractive for a data user. Linearization has also the advantage to offer a unified approach suitable for any sampling design, while a specific sampling design usually requires a specific bootstrap procedure, as illustrated with the BWO for SI sampling and the BWR for multistage sampling.

From the simulation study, we note that the coverage rates may not be well respected neither with linearization nor bootstrap, particularly in the multistage context and even with large sample sizes. There is a need for confidence intervals with better coverage rates under a reasonable computational burden. This is a matter for further research.

Acknowledgements

We thank Anne Ruiz-Gazen for helpful discussion. We also thank two referees and an Associate Editor for useful comments and suggestions which led to a significant improvement of the paper.

Appendix

Proof of equation (3.6)

From (3.3), we have $\widehat{\Delta t}^{\text{co}} = N(A^\top X)$, where $X = (\overline{y}_{2,s_{2\bullet}} - \overline{y}_{2,s_3}, \overline{y}_{1,s_{1\bullet}} - \overline{y}_{1,s_3}, \overline{y}_{2,s_3} - \overline{y}_{1,s_3})^\top$ and $A = (b, -a, 1)^\top$. This leads to

$$V\left\{\widehat{\Delta t}^{\text{co}}\right\} = N^2\{A^\top V(X)A\}. \tag{A.1}$$

We compute the elements in $V(X)$ separately. We have

$$\begin{aligned} V(\overline{y}_{2,s_3} - \overline{y}_{1,s_3}) &= \left(\frac{1}{n_3} - \frac{1}{N}\right) S^2_{y_2 - y_1, U} \\ &= \left(\frac{1}{n_3} - \frac{1}{N}\right)\left(S^2_{y_2, U} + S^2_{y_1, U} - 2S_{y_1 y_2, U}\right). \end{aligned}$$

Also, since $E(\overline{y}_{2,s_{2\bullet}} - \overline{y}_{2,s_3} \mid s_2) = 0$, we have

$$\begin{aligned} V(\overline{y}_{2,s_{2\bullet}} - \overline{y}_{2,s_3}) &= \mathrm{EV}(\overline{y}_{2,s_{2\bullet}} - \overline{y}_{2,s_3} \mid s_2), \\ &= \mathrm{EV}\left(\frac{n_2}{n_{2\bullet}}\overline{y}_{2,s_2} - \frac{n_3}{n_{2\bullet}}\overline{y}_{2,s_3} - \overline{y}_{2,s_3} \mid s_2\right) \\ &= \left(1 + \frac{n_3}{n_{2\bullet}}\right)^2 \mathrm{EV}(\overline{y}_{2,s_3} \mid s_2) \\ &= \left(1 + \frac{n_3}{n_{2\bullet}}\right)^2 \left(\frac{1}{n_3} - \frac{1}{n_2}\right) S^2_{y_2, U} \\ &= \frac{n_2}{n_3(n_2 - n_3)} S^2_{y_2, U} \end{aligned}$$

and

$$\begin{aligned}
\operatorname{Cov}(\bar{y}_{2,s_{2\bullet}} - \bar{y}_{2,s_3}, \bar{y}_{2,s_3} - \bar{y}_{1,s_3}) &= \operatorname{ECov}(\bar{y}_{2,s_{2\bullet}} - \bar{y}_{2,s_3}, \bar{y}_{2,s_3} - \bar{y}_{1,s_3} \mid s_2) \\
&= \operatorname{ECov}\left(\frac{n_2}{n_{2\bullet}}\bar{y}_{2,s_2} - \frac{n_3}{n_{2\bullet}}\bar{y}_{2,s_3} - \bar{y}_{2,s_3}, \bar{y}_{2,s_3} - \bar{y}_{1,s_3} \mid s_2\right) \\
&= -\left(1 + \frac{n_3}{n_{2\bullet}}\right)\operatorname{ECov}(\bar{y}_{2,s_3}, \bar{y}_{2,s_3} - \bar{y}_{1,s_3} \mid s_2) \\
&= \left(1 + \frac{n_3}{n_{2\bullet}}\right)\left(\frac{1}{n_3} - \frac{1}{n_2}\right)\left(S^2_{y_2,U} - S_{y_1y_2,U}\right) \\
&= -\frac{1}{n_3}\left(S^2_{y_2,U} - S_{y_1y_2,U}\right).
\end{aligned}$$

Similar arguments lead to

$$\begin{aligned}
V(\bar{y}_{1,s_{1\bullet}} - \bar{y}_{1,s_3}) &= \frac{n_1}{n_3(n_1 - n_3)}S^2_{y_1,U}, \\
\operatorname{Cov}(\bar{y}_{1,s_{1\bullet}} - \bar{y}_{1,s_3}, \bar{y}_{2,s_3} - \bar{y}_{1,s_3}) &= \frac{1}{n_3}\left(S^2_{y_1,U} - S_{y_1y_2,U}\right).
\end{aligned}$$

Finally, we consider $\operatorname{Cov}(\bar{y}_{2,s_{2\bullet}} - \bar{y}_{2,s_3}, \bar{y}_{1,s_{1\bullet}} - \bar{y}_{1,s_3})$. We first compute $\operatorname{Cov}(\bar{y}_{2,s_{2\bullet}}, \bar{y}_{1,s_{1\bullet}})$, which may be written as

$$\begin{aligned}
\operatorname{Cov}(\bar{y}_{2,s_{2\bullet}}, \bar{y}_{1,s_{1\bullet}}) &= \operatorname{Cov}(E(\bar{y}_{2,s_{2\bullet}} \mid s_{1\bullet}), E(\bar{y}_{1,s_{1\bullet}} \mid s_{1\bullet})) \\
&= \operatorname{Cov}(\bar{y}_{2,U\setminus s_{1\bullet}}, \bar{y}_{1,s_{1\bullet}}) \\
&= \operatorname{Cov}\left(\frac{N}{N - n_{1\bullet}}\bar{y}_{2,U} - \frac{n_{1\bullet}}{N - n_{1\bullet}}\bar{y}_{2,s_{1\bullet}}, \bar{y}_{1,s_{1\bullet}}\right) \\
&= -\frac{n_{1\bullet}}{N - n_{1\bullet}}\operatorname{Cov}(\bar{y}_{2,s_{1\bullet}}, \bar{y}_{1,s_{1\bullet}}) \\
&= -\frac{n_{1\bullet}}{N - n_{1\bullet}}\left(\frac{1}{n_{1\bullet}} - \frac{1}{N}\right)S_{y_1y_2,U} \\
&= -\frac{1}{N}S_{y_1y_2,U}.
\end{aligned}$$

Similar arguments lead to

$$\operatorname{Cov}(\bar{y}_{2,s_{2\bullet}}, \bar{y}_{1,s_3}) = \operatorname{Cov}(\bar{y}_{2,s_3}, \bar{y}_{1,s_{1\bullet}}) = -\frac{1}{N}S_{y_1y_2,U}.$$

We obtain

$$\begin{aligned}
\operatorname{Cov}\left(\overline{y}_{2,s_2\bullet} - \overline{y}_{2,s_3}, \overline{y}_{1,s_1\bullet} - \overline{y}_{1,s_3}\right) &= \frac{1}{N} S_{y_1y_2,U} + \operatorname{Cov}\left(\overline{y}_{2,s_3}, \overline{y}_{1,s_3}\right) \\
&= \frac{1}{N} S_{y_1y_2,U} + \left(\frac{1}{n_3} - \frac{1}{N}\right) S_{y_1y_2,U} \\
&= \frac{1}{n_3} S_{y_1y_2,U}.
\end{aligned}$$

In summary, we obtain

$$V(X) = \begin{pmatrix}
\frac{n_2}{n_3\left(n_2 - n_3\right)} S^2_{y_2,U} & \frac{1}{n_3} S_{y_1y_2,U} & -\frac{1}{n_3}\left(S^2_{y_2,U} - S_{y_1y_2,U}\right) \\
 & \frac{n_1}{n_3\left(n_1 - n_3\right)} S^2_{y_1,U} & \frac{1}{n_3}\left(S^2_{y_1,U} - S_{y_1y_2,U}\right) \\
 & & \left(\frac{1}{n_3} - \frac{1}{N}\right)\left(S^2_{y_2,U} + S^2_{y_1,U} - 2S_{y_1y_2,U}\right)
\end{pmatrix}$$

which, along with (A.1), leads to (3.6).

References

Antal, E., and Tillé, Y. (2011). A direct bootstrap method for complex sampling designs from a finite population. *Journal of the American Statistical Association*, 106, 534-543.

Barrett, G.F., and Donald, S.G. (2009). Statistical inference with generalized Gini indices of inequality, poverty, and welfare. *Journal of Business and Economic Statistics*, 27, 1-17.

Beaumont, J.-F., and Patak, Z. (2012). On the generalized bootstrap for sample surveys with special attention to Poisson sampling. *International Statistical Review*, 80, 127-148.

Berger, Y.G. (2004). Variance estimation for measures of change in probability sampling. *Canadian Journal of Statistics*, 32, 451-467.

Berger, Y.G. (2008). A note on the asymptotic equivalence of jackknife and linearization variance estimation for the Gini coefficient. *Journal of Official Statistics*, 24, 541-555.

Bertail, P., and Combris, P. (1997). Bootstrap généralisé d'un sondage. *Annales d'Économie et de Statistique*, 46, 49-83.

Bhattacharya, D. (2007). Inference on inequality from household survey data. *Journal of Econometrics*, 137, 674-707.

Bickel, P.J., and Freedman, D.A. (1984). Asymptotic normality and the bootstrap in stratified sampling. *The Annals of Statistics*, 12, 470-482.

Booth, J.G., Butler, R.W. and Hall, P. (1994). Bootstrap methods for finite populations. *Journal of the American Statistical Association*, 89, 1282-1289.

Brändén, P., and Jonasson, J. (2012). Negative dependence in sampling. *Scandinavian Journal of Statistics*, 39, 830-838.

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

Chao, M.-T., and Lo, S.-H. (1985). A Bootstrap method for finite population. *Sankhyā, Series A*, 47, 3, 399-405.

Chauvet, G. (2007). Méthodes de Bootstrap en population finie. Ph.D. dissertation, Université Rennes 2.

Chauvet, G. (2015). Coupling methods for multistage sampling. *The Annals of Statistics*, 43(6), 2484-2506.

Chen, J., and Rao, J.N.K. (2007). Asymptotic normality under two-phase sampling designs. *Statistica Sinica*, 17, 1047-1064.

Davison, A.C., and Hinkley, D.V. (1997). *Bootstrap Methods and their Application*. Cambridge University Press.

Davison, A.C., and Sardy, S. (2007). Resampling variance estimation in surveys with missing data. *Journal of Official Statistics*, 23, 3, 371-386.

Demnati, A., and Rao, J.N.K. (2004). Linearization variance estimators for survey data. *Survey Methodology*, 30, 1, 17-26. Paper available at https://www150.statcan.gc.ca/n1/pub/12-001-x/2004001/article/6991-eng.pdf.

Deville, J.-C. (1997). Estimation de la variance du coefficient de Gini mesurée par sondage. *Actes des Journées de Méthodologie Statistique*, *Insee Méthodes*.

Deville, J.-C. (1999). Variance estimation for complex statistics and estimators: Linearization and residual techniques. *Survey Methodology*, 25, 2, 193-203. Paper available at https://www150.statcan.gc.ca/n1/pub/12-001-x/1999002/article/4882-eng.pdf.

Druckman, A., and Jackson, T. (2008). Measuring resource inequalities: The concepts and methodology for an area-based Gini coefficient. *Ecological Economics*, 65, 242-252.

Gini, C. (1914). Sulla misura della concentrazione e della variabilità dei caratteri. *Atti del Reale Istituto Veneto di Scienze Lettere ed Arti.*

Glasser, G.J. (1962). Variance formulas for the mean difference and coefficient of concentration. *Journal of the American Statistical Association*, 57, 648-654.

Goga, C. (2003). Estimation de la variance dans les sondages à plusieurs échantillons et prise en compte de l'information auxiliaire par des modèles nonparamétriques. Ph.D. dissertation, Université Rennes 2.

Goga, C., and Ruiz-Gazen, A. (2014). Efficient estimation of nonlinear finite population parameters using nonparametrics. *Journal of the Royal Statistical Society B*, 76, 113-140.

Goga, C., Deville, J.-C. and Ruiz-Gazen, A. (2009). Composite estimation and linearization method for two-sample survey data. *Biometrika*, 96, 691-709.

Gordon, L. (1983). Successive sampling in large finite populations. *Annals of Statistics*, 11, 702-706.

Graczyk, P.P. (2007). Gini coefficient: A new way to express selectivity of kinase inhibitors against a family of Kinases. *Journal of Medicinal Chemistry*, 50, 5773-5779.

Gross, S.T. (1980). Median estimation in sample surveys. *ASA Proceedings of Survey Research*, 181-184.

Groves-Kirkby, C.J., Denman, A.R. and Phillips, P.S. (2009). Lorenz Curve and Gini coefficient: Novel tools for analysing seasonal variation of environmental radon gas. *Journal of Environmental Management*, 90, 2480-2487.

Hájek, J. (1960). Limiting distributions in simple random sampling from a finite population. *Tud. Akad. Mat. Kutatò Int. Közl.*, 5, 361-374.

Hájek, J. (1961). Some extensions of the Wald-Wolfowitz-Noether theorem. *Annals of Mathematical Statistics*, 32, 506-523.

Hájek, J. (1964). Asymptotic theory of rejective sampling with varying probabilities from a finite population. *Annals of Mathematical Statistics*, 35, 1491-1523.

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

Karagiannis, E., and Kovačević, M.S. (2000). A method to calculate the jackknife variance estimator for the Gini coefficient. *Oxford Bulletin of Economics and Statistics*, 62, 119-122.

Kovačević, M.S., and Binder, D.A. (1997). Variance estimation for measures of income inequality and polarization - The estimating equation approach. *Journal of Official Statistics,* 13, 41-58.

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

Lai, D., Huang, J., Risser, J.M. and Kapadia, A.S. (2008). Statistical properties of generalized Gini coefficient with application to health inequality measurement. *Social Indicator Research,* 87, 249-258.

Langel, M., and Tillé, Y. (2013). Variance estimation of the Gini index: Revisiting a result several times published. *Journal of the Royal Statistical Society, Series A,* 176, 521-540.

Lisker, T. (2008). Is the Gini coefficient a stable measure on galaxy structure? *The Astrophysical Journal Supplement Series,* 179, 319-325.

Navarro, V., Muntaner, C., Borrell, C., Benach, J., Quiroga, A., Rodríguez-Sanz, M., Vergès, N. and Pasarín, M.I. (2006). Politics and health outcomes. *The Lancet*, 18, 1033-1037.

Nygård, F., and Sandström, A. (1985). The estimation of the Gini and the entropy inequality parameters in finite populations. *Journal of Official Statistics*, 1, 4, 399-412.

Ohlsson, E. (1986). Asymptotic normality of the Rao-Hartley-Cochran estimator: An application of the martingale CLT. *Scandinavian Journal of Statistics*, 13, 17-28.

Ohlsson, E. (1989). Asymptotic normality for two-stage sampling from a finite population. *Probability Theory and Related Fields*, 81, 341-352.

Pires, A.M., and Branco, J.A. (2002). Partial influence functions. *Journal of Multivariate Analysis,* 83, 451-468.

Presnell, B., and Booth, J.G. (1994). *Resampling Methods for Sample Surveys*. Technical report.

Qin, Y., Rao, J.N.K. and Wu, C. (2010). Empirical likelihood confidence intervals for the Gini measure of income inequality. *Economic Modelling,* 27, 1429-1435.

Qualité, L., and Tillé, Y. (2008). Variance estimation of changes in repeated surveys and its application to the Swiss survey of value added. *Survey Methodology*, 34, 2, 173-181. Paper available at https://www150.statcan.gc.ca/n1/pub/12-001-x/2008002/article/10758-eng.pdf.

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.

Reid, N. (1981). Influence functions for censored data. *The Annals of Statistics*, 9, 78-92.

Rosén, B. (1972). Asymptotic theory for successive sampling with varying probabilities without replacement. I, II. *Annals of Mathematical Statistics*, 43, 373-397, 748-776.

Saegusa, T., and Wellner, J.A. (2013). Weighted likelihood estimation under two-phase sampling. *The Annals of Statistics*, 41, 269-295.

Sandström, A., Wretman, J.H. and Waldèn, B. (1985). Variance estimators of the Gini coefficient - Simple random sampling. *Metron,* 43, 41-70.

Sandström, A., Wretman, J.H. and Waldèn, B. (1988). Variance estimators of the Gini coefficient - Probability sampling. *Journal of Business and Economic Statistics,* 6, 113-119.

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

Sen, P.K. (1980). Limit theorems for an extended coupon collector's problem and for successive subsampling with varying probabilities. *Calcutta Statistical Association Bulletin,* 29, 113-132.

Shao, J., and Tu, D. (1995). *The Jackknife and the Bootstrap*. Springer.

Sitter, R.R. (1992a). A resampling procedure for complex survey data. *Journal of the American Statistical Association,* 87, 755-765.

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

Tam, S.M. (1984). On covariances from overlapping samples. *The American Statistician,* 38, 288-289.

Yitzhaki, S. (1991). Calculating jackknife variance estimators for parameters of the Gini method. *Journal of Business and Economic Statistics,* 9, 235-239.

Survey Methodology, June 2018
Vol. 44, No. 1, pp. 43-60
Statistics Canada, Catalogue No. 12-001-X

Growth Rates Preservation (GRP) temporal benchmarking: Drawbacks and alternative solutions

Jacco Daalmans, Tommaso Di Fonzo, Nino Mushkudiani and Reinier Bikker[1]

Abstract

Benchmarking monthly or quarterly series to annual data is a common practice in many National Statistical Institutes. The benchmarking problem arises when time series data for the same target variable are measured at different frequencies and there is a need to remove discrepancies between the sums of the sub-annual values and their annual benchmarks. Several benchmarking methods are available in the literature. The Growth Rates Preservation (GRP) benchmarking procedure is often considered the best method. It is often claimed that this procedure is grounded on an ideal movement preservation principle. However, we show that there are important drawbacks to GRP, relevant for practical applications, that are unknown in the literature. Alternative benchmarking models will be considered that do not suffer from some of GRP's side effects.

Key Words: Benchmarking; Growth rate preservation; Data reconciliation; Macro integration.

1 Introduction

Benchmarking monthly and quarterly series to annual data is a common practice in many National Statistical Institutes. For example, each year Statistics Netherlands aligns 12 quarterly Supply and Use Tables with the three most recent annual accounts (Eurostat, 2013, Annex 8C).

The benchmarking problem arises when time series data for the same target variable are measured at different frequencies with different levels of accuracy. One might expect that a temporal aggregation relationship between these time series is fulfilled, e.g., that four quarterly values add up to one annual value, but because of differences in data sources and processing methods, this is often not the case. Benchmarking is the process to remove such discrepancies. In this process the preliminary values are adjusted to achieve mathematical consistency between low-frequency (e.g., annual) and high-frequency (e.g., quarterly or monthly) time series.

There are two main principles of benchmarking. Firstly, low-frequency benchmarks are fixed, because these data sources describe levels and long-term trends better than high-frequency sources. Secondly, short-term movements of high-frequency time series are preserved as much as possible, as these data sources provide the only information on short-term movements.

Several benchmarking methods are available in the literature. These methods differ in the way short-term movements of high-frequency series are defined. A distinction can be made between multiplicative and additive methods. Multiplicative methods try to preserve relative changes of preliminary high-frequency time series, while additive methods aim to preserve changes in absolute terms. In this paper the focus will be solely on multiplicative variants.

[1] Jacco Daalmans, Statistics Netherlands, PO Box 24500, 2490 HA Den Haag, Netherlands. E-mail: j.daalmans@cbs.nl; Tommaso Di Fonzo, Department of Statistical Sciences, University of Padua, Via Cesare Battisti, 241, 35121 Padova PD, Italy. E-mail: difonzo@stat.unipd.it; Nino Mushkudiani, Statistics Netherlands, Den Haag. E-mail: n.mushkudiani@cbs.nl; Reinier Bikker, Statistics Netherlands, Den Haag, E-mail: r.bikker@cbs.nl.

Two well-known multiplicative methods are Denton Proportionate First Differences (PFD), by Denton (1971), and Growth Rates Preservation (GRP) by Causey and Trager (1981; see also Trager, 1982 and Bozik and Otto, 1988).

In the literature it is generally agreed that GRP is grounded on the strongest theoretical foundation (Bloem, Dippelsman and Maehle, 2001, page 100). It explicitly preserves the period-to-period rates of change of the preliminary series. However, Denton PFD is more popularly used, because it is technically easier to apply. Mathematically, the Denton method deals with a standard linearly constrained quadratic optimization problem, while GRP solves a more difficult linearly constrained nonlinear problem that can be efficiently solved by an interior-point-algorithm (Di Fonzo and Marini, 2015).

From a number of simulation studies it is known that Denton PFD and GRP lead to similar or close to similar results for the large majority of cases (Harvill Hood, 2005; Titova, Findley and Monsell, 2010; Di Fonzo and Marini, 2012 and Daalmans and Di Fonzo, 2014). Therefore Denton PFD can be used as an approximation of GRP.

The aim of this paper is to demonstrate that GRP suffers from drawbacks that are, to the best of our knowledge, not described in the literature. A first drawback is that it matters whether benchmarking is applied “forward” or “backward” in time. In this context, we will present a link with the time reversibility property from index number theory. A second drawback is that undesirable results may be obtained due to singularities in the GRP objective function.

A second aim of this paper is to present alternative benchmarking methods that do satisfy time reversibility. This paper may be valuable for practitioners who apply or consider to apply benchmarking techniques.

First, in Section 2, we will give a formal description of the Denton PFD and GRP benchmarking methods. Section 3 describes the drawbacks of the GRP method. In Section 4 two new benchmarking methods are proposed that can be used as an alternative for GRP. Results of an illustrative application to real-life data are given in Section 5. Finally, Section 6 concludes this paper.

2 Temporal benchmarking methods

This section explains the Denton PFD and GRP benchmarking procedures. Because temporal aggregation constraints are the same for Denton PFD and GRP, these are described first. Thereafter, the Denton PFD and GRP benchmarking procedures are explained.

We focus on univariate variants of these methods, in which temporal consistency is the main constraint of interest. The observations that are presented in the remainder of this paper are however also valid for the multivariate case, in which multiple time-series are reconciled simultaneously and additional constraints between time-series apply (see Di Fonzo and Marini, 2011 and Bikker, Daalmans and Mushkudiani, 2013).

2.1 General notation and temporal constraints

In general, temporal aggregation constraints can be expressed as a linear system of equalities $\mathbf{Ax} = \mathbf{b}$, where $\mathbf{x}$ is the target vector of high-frequency values, $\mathbf{b}$ is a vector of low-frequency values, and $\mathbf{A}$ is a temporal aggregation matrix converting high- into low-frequency values.

The specific form of these constraints depends on the nature of the variables involved. For flow variables, a sum of subannual values, e.g., four quarterly values, usually needs to be the same as one annual value. For stock variables, one of the subannual values, usually the first or the last, needs to be the same as the relevant annual value. For example, for quarterly/annual flow variables, assuming for the sake of simplicity that the available time span begins on the first quarter of the first year and ends on the fourth quarter of the last observed year, it is

$$\mathbf{A} = \begin{bmatrix} 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & \cdots & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \cdots & 1 & 1 & 1 & 1 \end{bmatrix}.$$

Denoting by $\mathbf{p}$ a vector of preliminary values, in general it is $\mathbf{Ap} \neq \mathbf{b}$, otherwise no adjustment would be needed. We look for a vector of benchmarked estimates $\mathbf{x}^*$, a particular outcome for $\mathbf{x}$, which should be "as close as possible" to the preliminary values and that satisfies $\mathbf{Ax}^* = \mathbf{b}$.

Not all sub annual periods need to be covered by a benchmark. Thus, the number of rows in $\mathbf{A}$ may be smaller than the total number of annual periods, see e.g., Dagum and Cholette (2006) for more details.

In a benchmarking operation, characteristics of the original series $\mathbf{p}$ should be considered. For example, in an economic time series framework, the preservation of the temporal dynamics (however defined) of the preliminary series is often a major interest of the practitioner.

2.2 Growth Rates Preservation (GRP) and Denton PFD

This section gives a formal description of GRP and Denton PFD.

Causey and Trager (1981; see also Monsour and Trager, 1979 and Trager, 1982) obtain the benchmarked values $x_t^*,\ t = 1, \ldots, n$ as a solution to the following optimization problem:

$$\min_{x_t} f_{\mathrm{F}}^{\mathrm{GRP}}(\mathbf{x}) \quad \text{subject to} \quad \mathbf{Ax} = \mathbf{b}, \quad \text{where} \quad f_{\mathrm{F}}^{\mathrm{GRP}}(\mathbf{x}) = \sum_{t=2}^{n} \left(\frac{x_t}{x_{t-1}} - \frac{p_t}{p_{t-1}} \right)^2. \tag{2.1}$$

The GRP criterion to be minimized, $f_{\mathrm{F}}^{\mathrm{GRP}}(\mathbf{x})$, explicitly relates to growth rates: it minimizes the sum of squared differences between growth rates of preliminary and benchmarked values. The subscript "F" in the minimization function stands for "Forward", later in this paper a "Backward" minimization function will be defined.

Denton (1971) proposed a benchmarking procedure grounded on the *Proportionate First Differences* (PFD) between target and original series. Cholette (1984) slightly modified the result of Denton, in order to correctly deal with the starting conditions of the problem. The PFD benchmarked estimates are thus obtained as the solution to the constrained quadratic minimization problem

$$\min_{x_t} f_{\mathrm{F}}^{\mathrm{PFD}}(\mathbf{x}) \quad \text{subject to} \quad \mathbf{Ax} = \mathbf{b}, \quad \text{where} \quad f_{\mathrm{F}}^{\mathrm{PFD}}(\mathbf{x}) = \sum_{t=2}^{n} \left(\frac{x_t}{p_t} - \frac{x_{t-1}}{p_{t-1}} \right)^2. \tag{2.2}$$

The Denton PFD criterion to be minimized, $f_{\mathrm{F}}^{\mathrm{PFD}}(\mathbf{x})$, is a sum of squared linear terms, which is easier to deal with than the nonlinear GRP objective function.

3 Two problems with GRP benchmarking

3.1 Time reversibility

Time reversibility means that it does not matter whether a method is applied forward or backward in time. This property can be of interest in many application areas.

In physics, it means that if time would run backwards, all motions are reversed. In index number theory, time reversibility was introduced in a classical work of Fisher (1922, page 64). It is stated that “if taking 1913 as a base and going forward to 1918, we find that, on the average, prices have doubled, then, by proceeding in the reverse direction, we ought to find the 1913 price level to be half that of 1918”. The motivation of this principle is that the direction of time can be considered arbitrary; it does not have any naturally preferred direction.

Time reversibility can also be applied in the context of benchmarking. It means that if we would reverse a time series, apply benchmarking, and reverse the benchmarked series back again, we get exactly the same results as for benchmarking the original series. In other words: from the benchmarked results it cannot be seen whether benchmarking has been applied forward or backward in time.

Benchmarking a reversed time series, according to GRP and Denton PFD, respectively, is equivalent to minimizing the following objective functions

$$f_{\mathrm{B}}^{\mathrm{GRP}}(\mathbf{x}) = \sum_{t=2}^{n} \left(\frac{x_{t-1}}{x_t} - \frac{p_{t-1}}{p_t} \right)^2 \tag{3.1}$$

and

$$f_{\mathrm{B}}^{\mathrm{PFD}}(\mathbf{x}) = \sum_{t=2}^{n} \left(\frac{x_{t-1}}{p_{t-1}} - \frac{x_t}{p_t} \right)^2, \tag{3.2}$$

where subscript “B” stands for backwards. These objective functions are obtained from the forward objective functions by interchanging t and $t-1$. From now on, the minimization of (3.1) or (3.2) will be called “backward benchmarking”, as opposed to standard, forward benchmarking.

As mentioned above, a benchmarking method satisfies the time reversibility property if forward and backward benchmarking lead to the same results. It can be easily seen that $f_{\mathrm{F}}^{\mathrm{GRP}}(\mathbf{x}) \neq f_{\mathrm{B}}^{\mathrm{GRP}}(\mathbf{x})$, while $f_{\mathrm{F}}^{\mathrm{PFD}}(\mathbf{x}) = f_{\mathrm{B}}^{\mathrm{PFD}}(\mathbf{x})$. From this it follows that Denton PFD satisfies the time reversibility property, but GRP does not.

More practically, in many production processes “forward” benchmarking is applied, for example for the reconciliation of the Dutch Supply and Use tables (Bikker et al., 2013). However, after a revision, revised time series may be constructed “back in time”, by using backward objective functions. It is highly undesirable that there are any differences in outcomes that can be purely attributed to a difference in “time direction”. Practitioners who are unaware of the time reversibility property, may apply forward and backward benchmarking and mistakenly assume that both methods lead to the same results.

Although it is true that any benchmarking application can be restricted to preserving forward growth rates, it is undesirable that results are affected by the irrelevant property of time direction. Therefore, any benchmarking method should preferably satisfy time reversibility. Moreover, Subsection 3.3 illustrates that a benchmarking method that is not symmetric in time may change the timing of the most important economic events, e.g., the peaks and troughs that demark the start and end of a crisis.

3.2 Singularity

A second problem of GRP is the singularity of its objective function. If x_{t-1} approaches to zero in case of forward benchmarking (or x_t for backward benchmarking) the objective function value tends to infinity. This causes several problems.

One complication is that the optimization problem becomes unstable, a small change in preliminary values can lead to a large shift in benchmarked values. Consequently, undesirably large revisions can be obtained when benchmarking updated data.

Another complication is that, since a correction of near zero values can be heavily penalised, growth rates of such values are strongly preserved. This may however come at the expense of relatively large corrections of other growth rates. On the other hand, one may argue that growth rates do not contain much information for extremely small (close-to-zero) values. Hence, growth rate preservation can be deemed inappropriate in this case. Subsection 5.3 shows a real-life example of this problem.

A third complication is that, as close-to-zero benchmarked values may cause a large objective function value, GRP methods tend to avoid such values. Consequently, irregular correction patterns can be obtained. In particular, negative benchmarked values may be obtained for a problem in which all preliminary values are positive. Consider an example in which two consecutive values are 100. Then, an adjustment of the first value from 100 to -100 is less costly in terms of GRP's objective function value than a correction from 100 to 30. The corresponding objective function values are $((100/-100)-(100/100))^2 = 4$ and $((100/30)-(100/100))^2 = 5.44.$ A value that goes from a large positive to a large negative will however usually not be considered good movement preservation. Therefore, the example also demonstrates the questionability of the use of growth rates when positive and negative values occur.

For this reason, it can be advisable to avoid negative outcomes by inclusion of non-negativity constraints, see Subsection 4.1 for more details. For Denton PFD negative values are less likely obtained. In the previous example, an adjustment from 100 to 30 is preferred to an adjustment from 100 to -100. A real-life example of this problem is shown in Subsection 5.3.

Although singularity of GRP's objective function may trigger negative benchmarked values, it is not the only cause. Denton PFD may also yield negative values. In general, there is a risk of negative benchmarked values, when the (relative) change from one benchmark to another significantly differs from the (relative) change from the underlying annualised preliminary values.

A fourth complication of GRP's singular objective function is that irregular peaks and throughs may occur in a benchmarked time series. The explanation is that in standard GRP a correction of large positive value to a close-to-zero value is less costly in terms of the objective function value than an opposite correction from close-to-zero to a large positive. That is, a correction of a growth rate g with a factor $c,$ where $c > 1,$ corresponds to a larger objective function value than a correction with $1/c,$ especially if c is large. The objective function values are $((c-1)g)^2$ and $\left(\frac{(c-1)}{c} g\right)^2$ respectively. Since large upward corrections from a close-to-zero value are relatively costly, these are avoided as much as possible. Thus, the GRP's benchmarked values move more gradually from a close-to-zero value than Denton's results do. To compensate for this, larger peaks may be necessary for the following time-periods to fulfill the temporal aggregation constraint. As benchmarking usually aims at as smooth as possible corrections over time, irregular peaks can be considered undesirable. Related to the relatively slow growth from a close to zero value is that the peaks tend to turn up later in time than for a time-symmetric method like Denton PFD. For the backward variant of GRP the opposite occurs, benchmarked time series move relatively quickly from a close to zero value, which gives rise to relatively early peaks. The example in Subsection 3.3 illustrates this problem.

3.3 Example

Below we present an example that illustrates the problems of GRP methods. In this example, a time series consisting of 15 months is reconciled with five quarterly values. The monthly series is constant: each monthly value is 10. The quarterly values are: 80, 250, 80, 400 and 100, respectively. Figure 3.1 compares the results of Denton PFD, GRPF and GRPB.

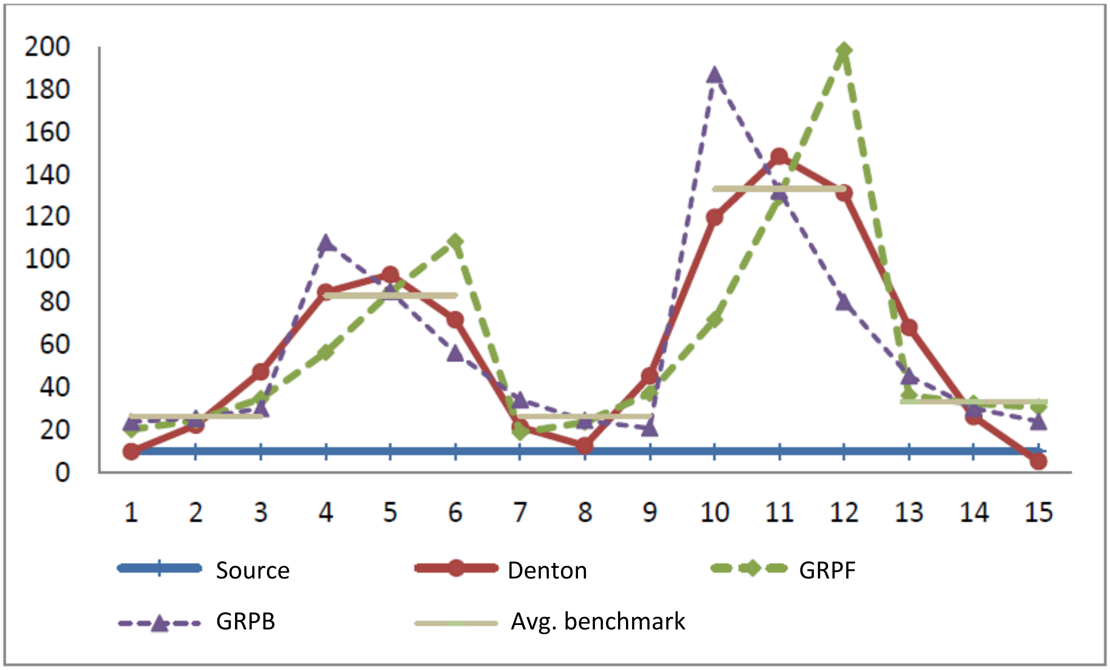


Figure 3.1 Example: Results of three benchmarking methods. "Avg. benchmark" stands for the average level of the monthly values that complies with the quarterly benchmarks and that is computed as one-third of its quarterly counterpart.

As the largest differences occur between both GRP methods, time reversibility is obviously not satisfied. The highest and lowest points appear at different months. The example clearly shows that the use of a different benchmarking method may lead to substantially different conclusions.

In accordance with Subsection 3.2, GRPF leads to relatively late peaks, i.e., at the last month of each quarter, while GRPB results in early peaks, i.e., at the first month of each quarter. Denton PFD's results are in between, peaks and troughs occur at the middle month of each quarter.

It needs however to be noted that the example cannot be considered representative for real life applications. In general, benchmarking methods are not meant to be used for reconciling large differences and for constant sub annual series. To explain the latter, a main assumption of Denton PFD is that the sub annual series provides information about short-term change. Constant series however cannot be considered very informative. Nevertheless, the problem of reconciling constant term series does occur in problems that are closely related to benchmarking, like interpolation and calenderization (see e.g., Dagum and Cholette, 2006 and Boot, Feibes and Lisman, 1967). The reason for choosing this example is purely educational. It provides good insight into properties of the different types of objective functions. The reader is referred to Subsection 5.3 for more realistic examples.

4 Alternative benchmarking techniques

In Section 3 we identified two problems with GRP methods. In this section we consider two alternative benchmarking techniques that solve the time irreversibility property.

4.1 Simultaneous growth rate preservation

Here, we propose two alternative objective functions for GRP. The first is a “time symmetric” variant of GRP, defined by

$$f_{\mathrm{S}}^{\mathrm{GRP}}(\mathbf{x}) = \frac{1}{2}\sum_{t=2}^{n}\left(\frac{x_t}{x_{t-1}} - \frac{p_t}{p_{t-1}}\right)^2 + \frac{1}{2}\sum_{t=2}^{n}\left(\frac{x_{t-1}}{x_t} - \frac{p_{t-1}}{p_t}\right)^2, \tag{4.1}$$

where subscript “S” stands for “simultaneous”. The method will be called GRPS in the remainder of this paper. The GRPS objective function both preserves forward and backward growth rates. As far as the authors know this method has not been mentioned elsewhere in the literature. It can be easily seen that GRPS satisfies time reversibility: interchanging t and $t - 1$ does not alter the objective function.

However, the second problem in Section 3 (singularity of objective function) is not considered. One of the consequences, negative benchmarked values, can be avoided by imposing lower bounds of zero on the benchmarked values. This can be done by including inequality constraints to an optimization problem, which is a well-established technique (e.g., Nocedal and Wright, 2006). The other problems related with singularity can however still occur.

4.2 Logarithmic growth rate preservation

Another “time symmetric” variant of GRP is given by the logarithmic form:

$$f_{\mathrm{L}}^{\mathrm{GRP}}(\mathbf{x}) = \sum_{t=2}^{n}\left[\log\left(\frac{x_t}{x_{t-1}}\right) - \log\left(\frac{p_t}{p_{t-1}}\right)\right]^2. \tag{4.2}$$

This function was firstly considered by Helfand, Monsour and Trager (1977). It is immediately verified that function (4.2) satisfies the time reversal property as well. The objective function can be considered the logarithmic version of GRP and equally well as the logarithmic version of Denton PFD. It will be denoted GRPL in the remainder of this paper, where “L” stands for “logarithmic”.

Note that (4.2) can be used for strictly positive preliminary values only, and that it produces benchmarked values that are larger than zero as well. This does not seem an important limitation, as Section 3 already mentioned that growth rate preservation can be considered inappropriate for problems with positive and negative values. Nevertheless, a potential solution for time series with negative values is to add a sufficiently large constant to the series prior to benchmarking and subtract that constant from the benchmarked series. A potential drawback of this solution is that adding a constant distorts initial growth-rates. Thus, it is unclear whether preliminary growth rates are actually preserved. Further research is necessary to better understand the implications of this solution.

Although GRPL necessarily produces positive values, other problems in Section 3.2, related to a singular objective function can still occur.

4.3 Comparison

When comparing GRPS and GRPL, it can be expected that GRPL behaves more like Denton PFD. Below we will give two reasons for this.

Firstly, because of the asymptotic properties of the log function, the problem that close-to-zero values are avoided is less severe for GRPL than for GRPS. Close-to-zero values are associated with large adjustments of growth rates. Very large adjustments of growth rates are penalised less in GRPL than in GRPS, since GRPS's objective function grows faster when corrections are large.

Secondly, the first-order Taylor linearization of GRPL's objective function corresponds to Denton PFD's function, whereas the approximation of GRPS leads to a different result. The linearization of the squared terms of the objective function in the preliminary values are given by $\left(\frac{x_t}{p_t}\right)-\left(\frac{x_{t-1}}{p_{t-1}}\right)$ and $\left(\frac{p_t}{p_{t-1}}+\frac{p_{t-1}}{p_t}\right)\left\{\left(\frac{x_t}{p_t}\right)-\left(\frac{x_{t-1}}{p_{t-1}}\right)\right\}$ for GRPL and GRPS respectively.

4.4 Example

In order to explore the properties of GRPL and GRPS, we will consider the example of Subsection 3.3 again. Figure 4.1 compares results of the symmetric $f_{\mathrm{S}}^{\mathrm{GRP}}$, $f_{\mathrm{L}}^{\mathrm{GRP}}$ and $f^{\mathrm{PFD}}(\mathbf{x})$ methods.

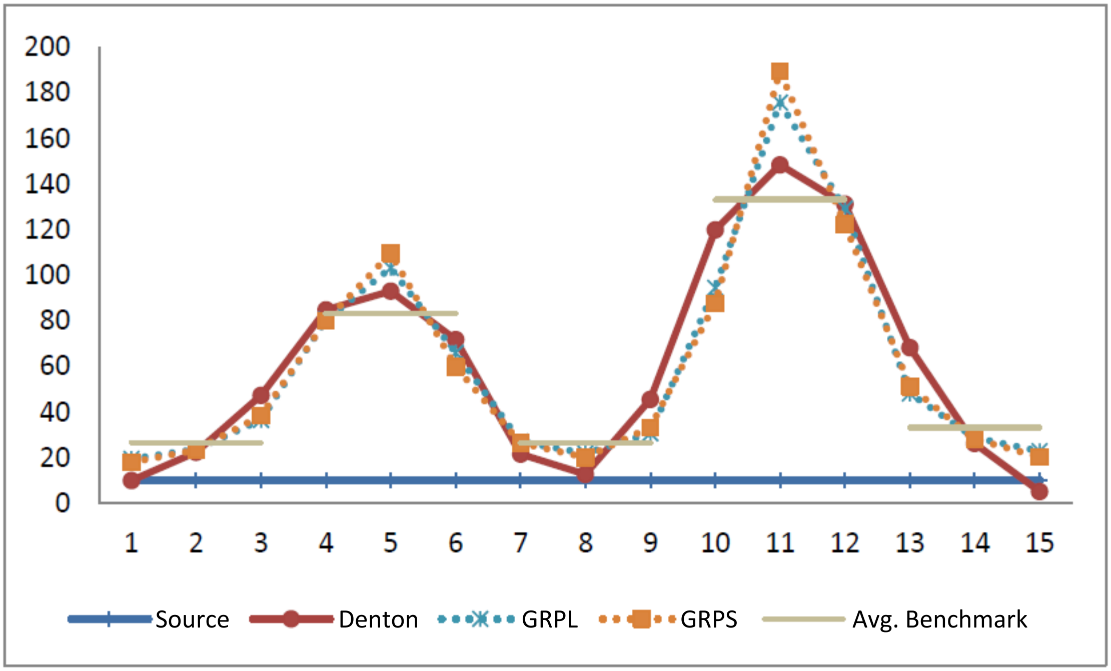


Figure 4.1 Example: results of three symmetric benchmarking methods. "Avg. Benchmark" stands for the average level of the monthly values that complies with the quarterly benchmarks and that is computed as one-third of its quarterly counterpart.

Firstly, it can be seen that the peaks and troughs occur at the same periods for all time symmetric methods.

Secondly, some of the drawbacks related to the singularity of the objective function still occur. When compared to Denton PFD, GRP methods tend to avoid close-to zero values, move away relatively slowly from low values (in both directions) and lead to irregular large peaks.

Thirdly, in accordance to Subsection 3.3, GRPL resembles Denton PFD more than GRPS, which follows from the slightly lower peaks of GRPL.

5 Empirical test

In this section an illustration exercise is conducted on real-life data, in order to find out whether or not the problems mentioned in Section 3 do occur in a realistic, practical application.

5.1 Data sets

The data set used for the illustration is obtained from quarterly and annual trade as published on the website of United Nations (UN).

The United Nations Commodity Trade Statistics Database (UN Comtrade) contains data from statistical authorities of reporting countries, or are received via partner organizations like the Organisation for Economic Co-operation and Development (OECD). The United Nations Totaltrade (UN Tottrade) data are mostly taken from the International Financial Statistics (IFS), published monthly by the International Monetary Fund (IMF). Differences between both sources emerge because of differences in data collection methods and purposes (United Nations, 2017). All data are publicly available at http://comtrade.un.org/.

We use UN Tottrade as data source for quarterly data and both UN Tottrade and UN Comtrade as sources for annual data. Both data sources include imports and exports for approximately 200 UN member states.

For our application all series were selected that include three annual totals and twelve quarterly values for 2002-2004. The variables of interest are total imports and exports. Series with quarterly or annual values smaller than 0.1 million dollars were deleted, as multiplicative benchmarking methods cannot be considered appropriate for zero or near zero values (see Subsection 3.2). Since the series are in million dollars, the cutoff value only excludes "extreme" cases and still leaves some real-life cases of singularity issues.

We end up with 238 time series for Comtrade and 253 series for Tottrade. The average year to year growth rates discrepancy between the annualized quarterly series and their benchmarks are 5.9%-point and 2.7%-point for Comtrade and Tottrade benchmarks, respectively. For the majority of series the discrepancy can be considered small. The percentage of series with a maximum discrepancy below 5%-point are 79% and 87%, respectively.

5.2 Results

Our first aim is to assess overall performance. We will compare the degree of preservation of the preliminary values and their growth rates for the various methods that are discussed in this paper.

Table 5.1 shows for the five methods the median values over all series, for the functions $f_{\mathrm{F}}^{\mathrm{GRP}}$, $f_{\mathrm{B}}^{\mathrm{GRP}}$, $f_{\mathrm{S}}^{\mathrm{GRP}}$ for forward, backward and simultaneous movement preservation and f^{Level} for preliminary value preservation. The latter function measures total squared relative adjustment, defined by

$$f^{\mathrm{Level}}(\mathbf{x}) := \sum_{t=1}^{n}\left(\frac{x_t}{p_t} - 1\right)^2. \tag{5.1}$$

Table 5.1
Median values of criteria in (2.1), (3.1), (4.1) and (5.1)

	COM data set				TOT data set			
	$f_{\mathrm{F}}^{\mathrm{GRP}}$	$f_{\mathrm{B}}^{\mathrm{GRP}}$	$f_{\mathrm{S}}^{\mathrm{GRP}}$	$f^{\text{ Level}}$	$f_{\mathrm{F}}^{\mathrm{GRP}}$	$f_{\mathrm{B}}^{\mathrm{GRP}}$	$f_{\mathrm{S}}^{\mathrm{GRP}}$	$f^{\text{ Level}}$
Denton PFD	0.87	0.88	0.88	26.42	0.33	0.41	0.37	2.07
GRPF	0.84	0.98	0.93	26.43	0.27	0.48	0.45	2.06
GRPB	1.00	0.82	0.91	26.47	0.48	0.28	0.45	2.07
GRPS	0.87	0.89	0.88	26.41	0.34	0.38	0.36	2.07
GRPL	0.87	0.88	0.88	26.42	0.33	0.41	0.37	2.07

The values for the COM and TOT data sets are *10^{-2} and *10^{-5}, respectively.

It can be seen from Table 5.1 that the GRPF method, that is designed to preserve forward growth rates, results in relatively poor backward movement preservation. The opposite is also true: GRPB does not preserve forward movements very well. From these results, we can conclude that time reversibility actually matters. Table 5.1 also demonstrates that the time symmetric methods, Denton PFD, GRPS and GRPL, perform well on all measures and that difference between those methods are only marginal.

To assess forward, backward and simultaneous growth rate preservation, a relative criterion is used that compares the values of the objective functions $f_{\mathrm{F}}^{\mathrm{GRP}}(\mathbf{x})$, $f_{\mathrm{B}}^{\mathrm{GRP}}(\mathbf{x})$ and $f_{\mathrm{S}}^{\mathrm{GRP}}(\mathbf{x})$ with their optimum values, which are obtained from GRPF, GRPB and GRPS, respectively. Analogous to the standards in Di Fonzo and Marini (2012), movement preservation is consided acceptable if it lies within 10% of the optimum value. That is, if $f^{\text{ method}}(\mathbf{x}) / f^{\text{ optimum}}(\mathbf{x}) \leq 1.1,$ where f is one of the previously mentioned objective functions.

For the five methods considered, Table 5.2 shows the percentage of time series with acceptable forward, backward and simultaneous movement preservation.

Table 5.2
Percentage of time series with acceptable movement preservation

	COM data set			TOT data set		
	Forward	**Backward**	**Simult.**	**Forward**	**Backward**	**Simult.**
Denton PFD	79.4	78.6	95.8	79.4	79.4	96.0
GRPF	100.0	48.7	81.5	100.0	47.8	82.6
GRPB	47.1	100.0	76.9	44.3	100.0	75.1
GRPS	82.4	77.3	100.0	80.6	79.4	100.0
GRPL	79.8	79.0	96.6	79.4	79.4	96.0

For Denton PFD an acceptable degree of simultaneous movement preservation is found for more than 95% of all cases. Thus, one can conclude that Denton PFD can be considered as a very good approximation for the optimal GRPS method; the approximation is even better than the GRPF and GRPB methods, for which acceptable performance is found for around 80% of all cases.

So far, we focused on performance for entire time series. Below we will consider the occurrence of large and extreme reconciliation adjustments made to single values and growth rates.

To measure the adjustments made to growth rates, the absolute difference $| g_{it}(\mathbf{x}) - g_{it}(\mathbf{p}) |$ * 100% is used, where g_{it} is a growth rate for series i and period *t*. Tables 5.3 and 5.4 compare the occurrence of large and extremely large adjustments to forward, backward and simultaneous growth rates.

Table 5.3
Percentage of large growth rate adjustments (> 10%-point difference)

	COM data set			TOT data set		
	Forward	**Backward**	**Simult.**	**Forward**	**Backward**	**Simult.**
Denton PFD	2.0	2.1	1.9	0.8	0.6	0.6
GRPF	1.9	2.4	2.3	0.6	0.9	0.7
GRPB	2.3	1.5	2.0	1.1	0.3	0.8
GRPS	1.9	1.9	1.8	0.8	0.6	0.6
GRPL	2.0	1.9	1.9	0.8	0.6	0.5

Table 5.4
Percentage of extreme growth rate adjustments (> 50%-point difference)

	COM data set			TOT data set		
	Forward	**Backward**	**Simult.**	**Forward**	**Backward**	**Simult.**
Denton PFD	0.3	0.2	0.4	0.1	0.1	0.1
GRPF	0.2	0.2	0.2	0.0	0.1	0.1
GRPB	0.3	0.0	0.2	0.2	0.0	0.1
GRPS	0.2	0.1	0.2	0.1	0.0	0.1
GRPL	0.2	0.1	0.2	0.1	0.0	0.1

These tables show minor differences between methods.

Small differences between methods are also in observed in Table 5.5, which shows large and extreme corrections to preliminary values, as measured by the relative criterion $\left(x_{\mathrm{it}} / p_{it}\right) * 100\%$.

Hence, one can conclude that the problems caused by singularity do not translate into more often occurring large corrections.

Table 5.5
Percentage of large adjustments to preliminary values

	COM data set			TOT data set		
	Large (>10%)	**Extreme (>100%)**	**Negative (<0%)**	**Large (>10%)**	**Extreme (>100%)**	**Negative (<0%)**
Denton PFD	13.2	1.0	0.0	5.8	0.4	0.0
GRPF	13.0	1.0	0.0	5.8	0.3	0.1
GRPB	13.1	0.9	0.0	5.6	0.3	0.0
GRPS	13.1	0.9	0.0	5.8	0.4	0.0
GRPL	13.0	0.9	0.0	5.8	0.4	0.0

Most remarkable in Table 5.5 are the negative benchmarked values obtained for GRPF in the TOT data. An example of this is illustrated in Figure 5.3.

Despite the similar results of the five benchmarking methods in Tables 5.3-5.5, there are clear differences in smoothness of reconciliation adjustments. To demonstrate this, we will use the smoothness indicator (Temurshoev, 2012).

$$\text{Smoothness} \;=\; \sum_{t=2}^{n-2}\left[\mathrm{BI}_t - \overline{\mathrm{BI}_t}\,\right]^2, \tag{5.2}$$

where BI_t is the so-called benchmark-to-indicator ratio, i.e., x_t / p_t and $\overline{\mathrm{BI}_t}$ is the 5-terms moving average $\frac{1}{5}\sum_{k=t-2}^{k=t+2}\mathrm{BI}_k$.

According to this indicator, we find in Table 5.6 that the smoothest results are obtained for Denton PFD and GRPL. Conversely, the asymmetric GRPF and GRPB methods yield the most irregular adjustments. It follows that the time-symmetric method GRPS, but most so GRPL, suffers less from singularity than the asymmetric methods GRPF and GRPB do. These results most clearly illustrate the problems with the singularity of GRP's objective function that were described in Subsection 3.2.

Table 5.6
Smoothness indicator values (5.2), summed over all series

	COM data set	TOT data set
Denton PFD	3.4	0.3
GRPF	9.8	39.0
GRPB	8.2	2.9
GRPS	4.3	1.1
GRPL	3.3	0.5

5.3 Examples

Below we show two examples to demonstrate that the problems in Section 3 do occur in a real-life application.

The first example, in Figures 5.1 and 5.2, illustrates that non-symmetric GRP methods may change the timing of the most important economic events. When considering the first nine quarters, the two highest values occur at different time periods. GRPF's peak periods are at quarters 6 and 7 and those of GRPB are at quarters 5 and 6. Closely related to this, is that GRPF moves away relatively slowly from the close-to-zero values at quarters 1-4.

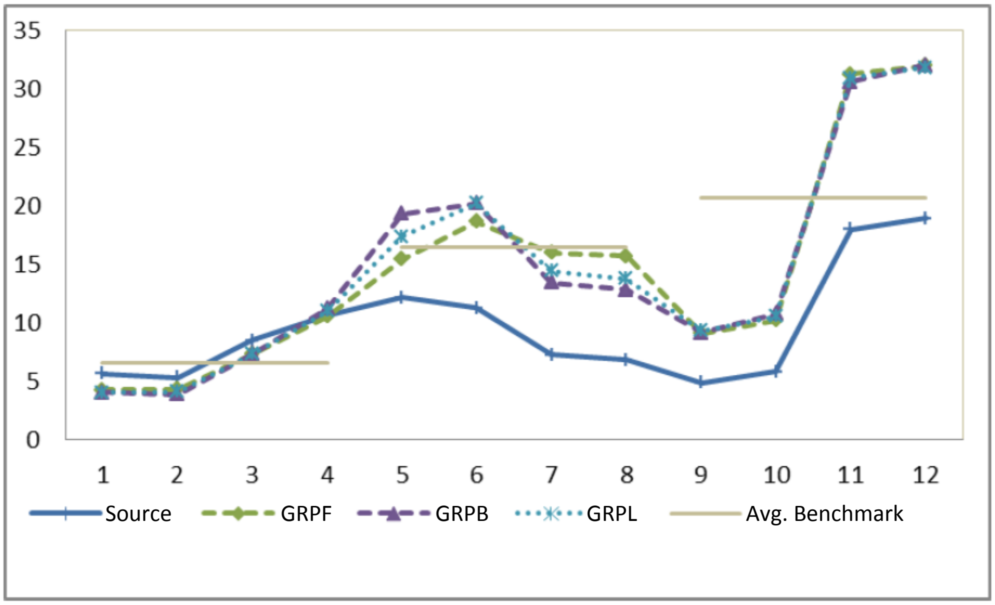


Figure 5.1 Exports Burundi, Comdata, 2002-2004, millions of US dollar. "Avg. Benchmark" stands for the average level of the quarterly data that complies with the annual benchmarks and that is computed as one-fourth of its annual counterpart.

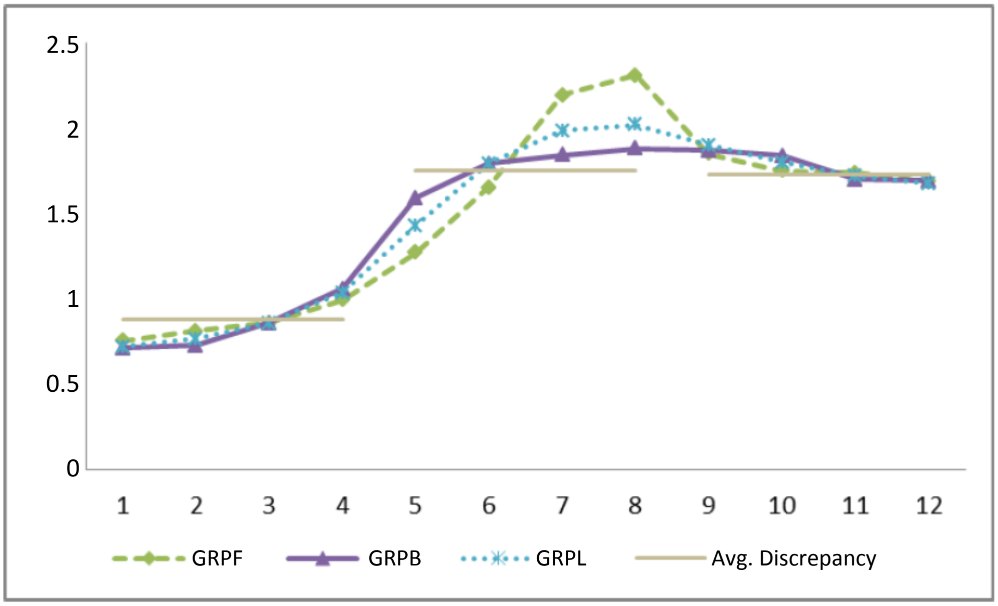


Figure 5.2 Benchmark to Indicator ratios, Exports Burundi, 2002-2004. "Avg. Discrepancy" stands for the annual BI-ratio, i.e., the ratio of an annual benchmark and the sum of the underlying quarterly indicators.

The second example illustrates the complications of a singular objective function. As shown in Figure 5.4, GRPF closely preserves growth rates of the quarters 6-10. This comes however at the expense of an irregular peak in quarter 5 and negative benchmarked values in the quarters 11 and 12.

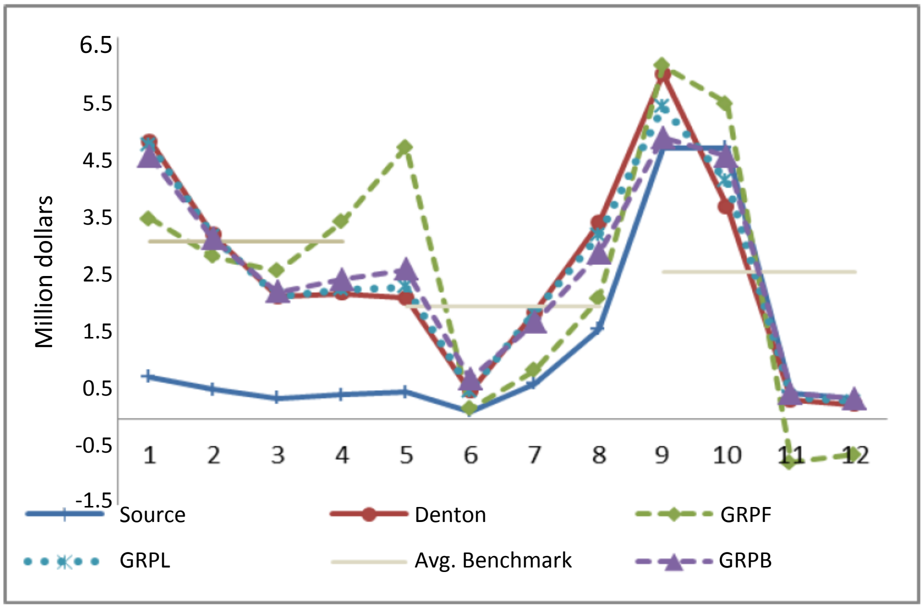


Figure 5.3 Exports Gambia, Totdata, 2002-2004, millions of US dollar. "Avg. Benchmark" stands for the average level of the quarterly data that complies with the annual benchmarks and that is computed as one-fourth of its annual counterpart.

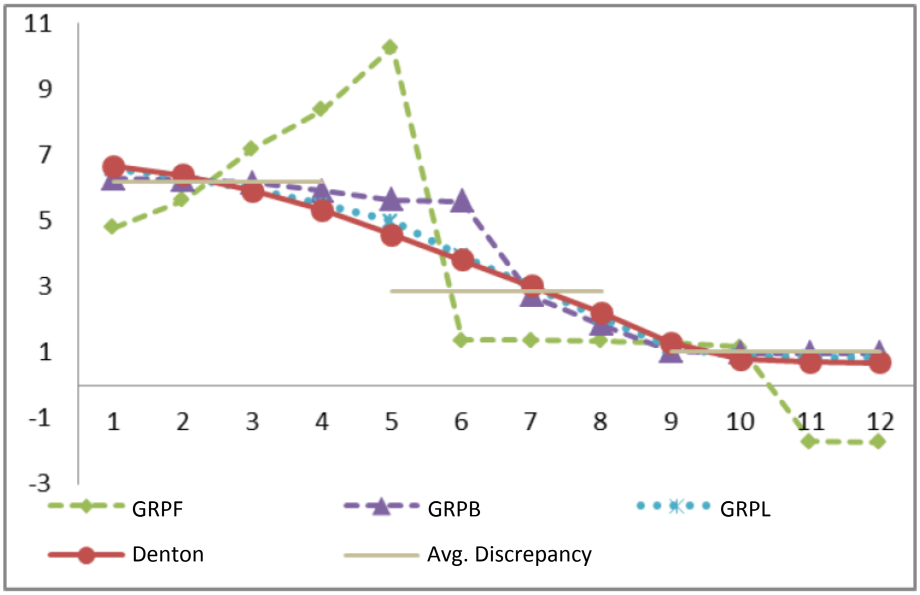


Figure 5.4 Exports Gambia, Totdata, 2002-2004, benchmark to indicator ratio. “Avg. Discrepancy” stands for the annual BI-ratio, i.e., the ratio of an annual benchmark and the sum of the underlying quarterly indicators.

6 Conclusions

Two well-known multiplicative benchmarking methods are Denton Proportionate First Differences (PFD) and Growth Rates Preservation (GRP). It is generally agreed that GRP has the strongest theoretical foundation. It better preserves initial growth rates than Denton PFD. However, from a technical point of view, Denton is the easiest method to apply. Because of this, and because Denton PFD is often a good approximation of GRP, Denton PFD is more popularly applied.

In this paper two drawbacks of GRP are demonstrated that, to the best knowledge of the authors, have not been mentioned elsewhere.

The first drawback is that GRP does not satisfy the time reversibility property. According to this property it should not matter for the results whether forward or backward growth rates are preserved. That is, benchmarking an original time series, $t = 1,\ldots,n$, or a “reversed” time series, $t = n,\ldots,1$ should lead to the same benchmarked series. Since direction of time is irrelevant for any benchmarking application, any benchmarking method should preferably satisfy time reversibility. Moreover, a benchmarking method that does not satisfy time reversibility may yield entirely difficult conclusions on the timing of economic events depending on the chosen time direction. For these reasons forward and backward GRP methods should preferably be discouraged.

In this paper two alternative GRP methods are presented that do satisfy time reversibility. The first alternative, a new GRPS method, preserves both forward and backward growth rates. The other alternative, an existing GRPL method, preserves logarithms of the forward growth rates.

A second drawback of all GRP methods in this paper are the singularities in its objective functions. Complications of this are: avoidance of close to zero outcomes, irregular peaks in results and unnecessary negative values in benchmarked results.

These problems actually occurred in an illustrative application on real-life data. Although unnecessary negative values only occasionally occurred, reconciliation adjustments are much more irregular than for Denton PFD. Since smoothness of reconciliation adjustments (BI ratios) is often the main interest of benchmarking, asymmetric GRP methods can be discouraged for many applications.

While the literature considers Denton PFD "a good approximation" of the ideal GRP method, our main conclusion is that Denton PFD is even more appropriate than standard GRP for many applications. Denton is computationally easier to apply, it does not suffer from the problems related to time irreversibility and a singular objective function. Furthermore, the approximation of Denton PFD's results is even more close for the time-symmetric versions of GRP than for standard GRP.

However, when growth rate preservation is the key point of interest, a time-symmetric version of GRP can also be a good choice, most in particular GRPL. Time symmetric methods preserve growth rates slightly better than Denton PFD, satisfy time reversibility and suffer less severe from the drawbacks of a singular objective function than standard GRP.

Acknowledgements

The authors would like to thank three reviewers and the associate editor for useful comments that greatly contributed to improving the paper.

References

Bikker, R.P., Daalmans, J.A. and Mushkudiani, N. (2013). Benchmarking large accounting frameworks: A generalized multivariate model. *Economic Systems Research*, 25, 390-408.

Bloem, A., Dippelsman, R. and Maehle, N. (2001). Quarterly National Accounts Manual. Concepts, Data Sources, and Compilation, International Monetary Fund, Washington DC.

Boot, J.C.G., Feibes, W. and Lisman, J.H.C. (1967). Further methods of derivation of quarterly figures from annual data. *Applied Statistics*, 16, 65-75.

Bozik, J.E., and Otto, M.C. (1988). Benchmarking: Evaluating methods that preserve month-to-month changes. Bureau of the Census - Statistical Research Division, RR-88/07, URL: http://www.census.gov/srd/papers/pdf/rr88-07.pdf.

Causey, B., and Trager, M.L. (1981). Derivation of Solution to the Benchmarking Problem: Trend Revision. Unpublished research notes, U.S. Census Bureau, Washington D.C. Available as an appendix in Bozik and Otto (1988).

Cholette, P.A. (1984). Adjusting sub-annual series to yearly benchmarks. *Survey Methodology*, 10, 1, 35-49. Paper available at https://www150.statcan.gc.ca/n1/pub/12-001-x/1984001/article/14348-eng.pdf.

Daalmans, J.A., and Di Fonzo, T. (2014). Denton PFD and GRP benchmarking are friends. An empirical evaluation on Dutch Supply and Use Tables. Paper presented to the IIOA-conference, Lisboa 14-18 July.

Dagum, E.B., and Cholette, P. (2006). *Benchmarking, Temporal Distribution and Reconciliation Methods for Time Series Data*. New York: Springer-Verlag.

Denton, F. (1971). Adjustment of monthly or quarterly series to annual totals: An approach based on quadratic minimization. *Journal of the American Statistical Association*, 66, 99-102.

Di Fonzo, T., and Marini, M. (2011). Simultaneous and two-step reconciliation of systems of time series: Methodological and practical issues. *Journal of the Royal Statistical Society C*, 60, 143-164.

Di Fonzo, T., and Marini, M. (2012). Benchmarking time series according to a growth rates preservation principle. *Journal of Economic and Social Measurement*, 37, 225-252.

Di Fonzo, T., and Marini, M. (2015). Reconciliation of systems of time series according to a growth rates preservation principle. *Statistical Methods & Applications*, 24, 651-669.

Eurostat (2013). *Handbook on Quarterly National Accounts – 2013 Edition*. http://ec.europa.eu/eurostat/web/products-manuals-and-guidelines/-/KS-GQ-13-004.

Fisher, I. (1922). *The Making of Index Numbers*, Houghton-Mifflin, Boston.

Harvill Hood, C.C. (2005). An empirical comparison of methods for benchmarking seasonally adjusted series to annual totals. Paper presented at the Workshop on Frontiers in Benchmarking Techniques and their Applications to Official Statistics, Luxembourg, 7-8 April 2005.

Helfand, S.D., Monsour, N.J. and Trager, M.L. (1977). Historical Revision of Current Business Survey Estimates. *Proceedings of the Business and Economic Statistics Section*, American Statistical Association, 246-250.

Monsour, N.J., and Trager, M.L. (1979). Revision and benchmarking of business time series. *Proceedings of the Business and Economic Statistics Section*, American Statistical Association, 333-337.

Nocedal, J., and Wright, S. (2006). *Numerical Optimization, 2nd Edition*. New York: Springer.

Temurshoev, U. (2012). *Entropy-Based Benchmarking Methods*. University of Groningen. http://www.ggdc.net/publications/memorandum/gd122.pdf.

Titova, N., Findley, D. and Monsell, B.C. (2010). Comparing the Causey-Trager method to the multiplicative Cholette-Dagum regression-based method of benchmarking sub-annual data to annual benchmarks. *Joint Statistical Meetings Proceedings, Business and Economic Statistics Section*, 3007-3021.

Trager, M.L. (1982). Derivation of Solution to the Benchmarking Problem: Relative Revision. Unpublished research notes, U.S. Census Bureau, Washington D.C. Available as an appendix in Bozik and Otto (1988).

United Nations (2017). *2016 International Trade Statistics Yearbook. Volume I Trade by Countries*. New York: United Nations. https://comtrade.un.org/pb/downloads/2016/ITSY2016VolI.pdf.

Investigating alternative estimators for the prevalence of serious mental illness based on a two-phase sample

Phillip S. Kott, Dan Liao, Jeremy Aldworth, Sarra L. Hedden, Joseph C. Gfroerer, Jonaki Bose and Lisa Colpe[1]

Abstract

A two-phase process was used by the Substance Abuse and Mental Health Services Administration to estimate the proportion of US adults with serious mental illness (SMI). The first phase was the annual National Survey on Drug Use and Health (NSDUH), while the second phase was a random subsample of adult respondents to the NSDUH. Respondents to the second phase of sampling were clinically evaluated for serious mental illness. A logistic prediction model was fit to this subsample with the SMI status (yes or no) determined by the second-phase instrument treated as the dependent variable and related variables collected on the NSDUH from all adults as the model's explanatory variables. Estimates were then computed for SMI prevalence among all adults and within adult subpopulations by assigning an SMI status to each NSDUH respondent based on comparing his (her) estimated probability of having SMI to a chosen cut point on the distribution of the predicted probabilities. We investigate alternatives to this standard cut point estimator such as the probability estimator. The latter assigns an estimated probability of having SMI to each NSDUH respondent. The estimated prevalence of SMI is the weighted mean of those estimated probabilities. Using data from NSDUH and its subsample, we show that, although the probability estimator has a smaller mean squared error when estimating SMI prevalence among all adults, it has a greater tendency to be biased at the subpopulation level than the standard cut point estimator.

Key Words: Bias; Bias-corrected estimator; Domain; Survey-sampling theory; Asymptotic.

1 Introduction

Serious mental illness is defined as currently or in the past year having a diagnosable mental, behavioral, or emotional disorder (excluding developmental and substance-use disorders) of sufficient duration to meet diagnostic criteria specified in the Diagnostic and Statistical Manual of Mental Disorders, 4th edition (American Psychiatric Association, 1994). The National Survey on Drug Use and Health (NSDUH), sponsored by the Substance Abuse and Mental Health Services Administration (SAMHSA), provides national and state-level estimates on the use of tobacco products, alcohol, and illicit drugs in the civilian, noninstitutionalized population of the United States aged 12 years or older. The Mental Health Surveillance Study (MHSS) was a follow-up study to the NSDUH main interview, designed to estimate the prevalence of serious mental illness (SMI) among adults 18 years of age or older at the national level and within particular subpopulations of interest. It was not practical to administer clinical interviews to the entire NSDUH sample of approximately 46,000 adults per year in order to obtain estimates of SMI due to financial and time constraints. Thus, a second phase of sampling was employed as in the National Comorbidity Survey Replication (See Kessler, Chiu, Demler and Walters, 2005). From 2008 through 2012, a clinical interview was administered to a randomly-selected (before nonresponse) subsample of NSDUH adult respondents within four weeks of completing the NSDUH main interview. For more information about the

[1] Phillip S. Kott, Dan Liao and Jeremy Aldworth, RTI International. E-mail: pkott@rti.org; Sarra L. Hedden, Joseph C. Gfroerer and Jonaki Bose, Substance Abuse and Mental Health Services Administration (SAMHSA), Center for Behavioral Health Statistics and Quality (CBHSQ), 5600 Fishers Lane, Rockville, MD 20857, U.S.A.; Lisa Colpe, National Institute of Mental Health, 6001 Executive Blvd., Bethesda, MD 20892.

NSDUH and its clinical subsample, the reader is referred to Center for Behavioral Health Statistics and Quality (CBHSQ), 2014.

The clinical evaluations were treated as the “gold standard”, despite the possibility of human error (this topic is discussed further in Section 4). A logistic prediction model was fit to the respondent subsample with the clinical SMI evaluations (yes or no) treated as the dependent variable and variables contained on the NSDUH as the model’s explanatory variables. The NSDUH variables included in the prediction model, obtained from all 46,000 adult respondents in the main survey, were measures of psychological distress and functional impairment derived from responses to NSDUH survey items, an age variable, the existence of a self-reported past-year major depressive episode, and the existence of past-year serious suicidal thoughts. The fitted prediction model was then applied to the NSDUH full sample to predict every adult respondent’s probability of having SMI based on their responses in the NSDUH main interview.

An adult NSDUH respondent with an estimated probability of having SMI that was greater than or equal to a cut point was assigned a predicted SMI value of 1 (i.e., has SMI); otherwise, he or she was assigned a predicted SMI value of 0 (i.e., does not have SMI). SMI prevalence rates for all adults and within various subpopulations were then estimated using these predicted values. The cut point was determined so that within the MHSS subsample the weighted fraction of false positives (subsample respondents predicted to have SMI based on the model but clinically diagnosed as *not* having SMI) and false negatives (subsample respondents predicted not to have SMI but clinically diagnosed as having SMI) came as close to being equal as possible. Strict equality was usually impossible because predicted probabilities of having SMI only take on a limited number of values depending on the covariates in the model.

The standard cut point estimator is derived from Receiver Operator Characteristics (ROC) theory. See, for example, Fawcett (2006). In Section 2, we use probability-sampling-based (also called “design-based”) survey-sampling theory to propose alternatives to this estimator. One such is the probability estimator, which simply assigns each NSDUH respondent his (her) estimated probability of having SMI, with no dichotomous designation. Also introduced are bias-corrected versions of both the standard cut point and probability estimators. These are similar to standard GREG estimators (see, for example, Särndal, Swensson and Wretman, 1989) and are nearly unbiased under survey-sampling theory whether or not the logistic model holds. The bias correction term in each serves as the basis of a test statistic for determining whether the associated model-based estimator – whether cut point or probability – is significantly biased.

Section 3 uses NSDUH/MHSS subsample data collected between 2008 and 2012 to evaluate the alternative estimators using the logistic model. We show that although the probability estimator has a smaller mean squared error when estimating SMI prevalence among all adults, it has a greater tendency to be biased at the subpopulation level than the standard cut point estimator. This leads us to propose a hybrid cut point estimator that is (at least) more efficient than the standard cut point estimator for all adults while not having the tendency to be biased at the subpopulation level.

Section 4 provides some concluding remarks. It is important to remember that SAMHSA planned to use the logistic model fit in the 2008 to 2012 clinical subsamples to help compute annual estimates of adult SMI prevalences based on NSDUH responses beyond 2012 without any new clinical subsamples.

2 Some estimators

2.1 Across all adults

Let S denotes the relevant NSDUH respondent sample (adults 18 years or older) from 2008 through 2012, and w_k the NSDUH analysis (first-phase) weight for an individual $k \in S$. Let S' denotes the subsample of S responding to a clinical evaluation of their SMI status. Let $y_k = 1$ when k is diagnosed to have serious mental illness, and $y_k = 0$ when k is diagnosed not to have serious mental illness. Let ω_k be the two-phase weight for an individual k in S'. For convenience, we set ω_k to 0 for individuals in S but not S'.

In actual practice, both sets of weights have been adjusted to account for nonresponse and undercoverage and to increase their efficiency, but we will ignore that fact here for simplicity. Instead, we will assume $1/w_k$ is the probability of selection for a NSDUH respondent, $1/\omega_k$ the probability of selection for a MHSS subsample respondent, and thus w_k/ω_k the conditional selection probability of a subsample respondent given (s)he was a NSDUH respondent. A nearly unbiased estimator for the prevalence of SMI among adults between 2008 and 2012 based on the two-phase sample is $\bar{y}_U = \sum_{S'} \omega_k y_k \Big/ \sum_{S'} \omega_k$, "nearly" because the denominator may contain some sampling error.

Suppose a ω_k-weighted logistic regression is run on the all-adult MHSS subsample respondents in S' with y_k as the dependent variable using a reasonable vector of explanatory covariates, $\mathbf{x}_k$, available for every respondent in the adult NSDUH sample. Exactly how the covariates have been chosen is beyond the scope of this investigation (for that, the reader is directed to Center for Behavioral Health Statistics and Quality, 2015; Chapter 4). Let the predictor for y_k from this weighted-logistic regression be $p_k = p(\mathbf{x}_k'\mathbf{b}) = [1 + \exp(-\mathbf{x}_k'\mathbf{b})]^{-1}$.

The use of weights in fitting the logistic-regression model protects against the possibility that the model residuals are correlated with the probabilities of selection. It is also consistent with how SMI prevalence was estimated; that estimate resulted from the weighted regression of y_k on the constant 1 and no covariates.

Sorting the subsample by the p_k values, one can find the cut point value p_C such that

$$\sum_{\substack{k \in S' \\ p_k \geq p_C}} \omega_k = \sum_{k \in S'} \omega_k y_k \tag{2.1}$$

holds exactly or as nearly so as possible. That is to say, the estimated number of adults in the population having p_k-values at or above the cut point approximately equals the estimated number of adults with SMI. Define an indicator random variable c_k to be 1 when $p_k \geq p_C$ and 0 otherwise. A cut point determined using equation (2.1) also comes as close as possible to equalizing the weighted false positives $\left(\sum_{S':c_k=1} \omega_k (1 - y_k)\right)$ and false negatives $\left(\sum_{S':c_k=0} \omega_k y_k\right)$ in S'.

Two alternative estimators for SMI prevalence among adults are the model-driven *cut point* and *probability* estimators:

$$\bar{y}_C = \frac{\sum_S w_k c_k}{\sum_S w_k}, \tag{2.2}$$

and

$$\bar{y}_P = \frac{\sum_S w_k p_k}{\sum_S w_k}, \tag{2.3}$$

these estimators are computed using the entire NSDUH sample rather than the smaller MHSS subsample as is $\bar{y}_U$.

We assume now that one of the covariates in the logistic model is 1 or the equivalent $(\mathbf{x}_k'\boldsymbol{\gamma} = 1$ for some $\boldsymbol{\gamma})$. Under this assumption, the probability estimator for SMI prevalence is exactly equal to a *bias-corrected* probability estimator given below:

$$\begin{aligned}\bar{y}_{P\text{-BC}} &= \frac{\sum_{S'} \omega_k y_k}{\sum_{S'} \omega_k} + \left(\frac{\sum_S w_k p_k}{\sum_S w_k} - \frac{\sum_{S'} \omega_k p_k}{\sum_{S'} \omega_k} \right) \\ &= \frac{\sum_S w_k p_k}{\sum_S w_k} + \frac{\sum_{S'} \omega_k (y_k - p_k)}{\sum_{S'} \omega_k}.\end{aligned} \tag{2.4}$$

The equality between $\bar{y}_P$ and $\bar{y}_{P\text{-BC}}$ results from the numerator of the *bias-correction term* in the second line of equation (2.4), $\sum_{S'} \omega_k (y_k - p_k) \big/ \sum_{S'} \omega_k$, equaling zero. Fitting a logistic regression forces $\sum_{S'} \omega_k (y_k - p_k)\, \mathbf{x}_k = \mathbf{0}$, and we have assumed $\mathbf{x}_k$ contains 1 or the equivalent.

Since the expectation of the term in parentheses in the *first* line of equation (2.4) is nearly zero under mild conditions, $\bar{y}_P = \bar{y}_{P\text{-BC}}$, like $\bar{y}_U$, is nearly unbiased under survey-sampling theory. *This is true whether or not the model used to determine the* p_k *is correct* so long as $\mathbf{b}$ in $p_k = p(\mathbf{x}_k'\mathbf{b}) = [1 + \exp(-\mathbf{x}_k'\mathbf{b})]^{-1}$ converges to *something* as the MHSS subsample and NSDUH sample sizes grow arbitrarily large.

The estimator $\bar{y}_{P\text{-BC}}$ is analogous to the popular GREG estimator. It follows Lehtonen and Veijanen (1998), and computes the p_k with a logistic rather than the linear model of the GREG.

A bias-corrected cut point estimator is

$$\begin{aligned}\bar{y}_{C\text{-BC}} &= \frac{\sum_{S'} \omega_k y_k}{\sum_{S'} \omega_k} + \left(\frac{\sum_S w_k c_k}{\sum_S w_k} - \frac{\sum_{S'} \omega_k c_k}{\sum_{S'} \omega_k} \right) \\ &= \frac{\sum_S w_k c_k}{\sum_S w_k} + \frac{\sum_{S'} \omega_k (y_k - c_k)}{\sum_{S'} \omega_k}.\end{aligned} \tag{2.5}$$

Using the same logic as above, this estimator is also nearly unbiased under mild conditions. It is close to the model-driven cut point estimator since the bias-correction term, $\sum_{S'} \omega_k (y_k - c_k) \big/ \sum_{S'} \omega_k$, is close to zero. The bias-correction term would be exactly zero if there were a cut point p_C that satisfied equation (2.1) exactly.

2.2 Domain estimation

Let us now turn our attention to a subpopulation of all adults, such as males or all adults who have received treatment for mental illness (or all adults who live in a particular state). We call such a subpopulation a "domain" of interest. To estimate SMI prevalence in a domain, we can simply insert an indicator for domain membership, d_k, equaling 1 when k is in the domain, 0 otherwise, into all our estimates:

$$\bar{y}_{U(d)} = \frac{\sum_{S'} \omega_k y_k d_k}{\sum_{S'} \omega_k d_k} \tag{2.6}$$

$$\bar{y}_{P(d)} = \frac{\sum_S w_k p_k d_k}{\sum_S w_k d_k} \tag{2.7}$$

$$\bar{y}_{C(d)} = \frac{\sum_S w_k c_k d_k}{\sum_S w_k d_k} \tag{2.8}$$

$$\begin{aligned} \bar{y}_{P-\mathrm{BC}(d)} &= \bar{y}_{U(d)} + \left(\frac{\sum_S w_k p_k d_k}{\sum_S w_k d_k} - \frac{\sum_{S'} \omega_k p_k d_k}{\sum_{S'} \omega_k d_k} \right) \\ &= \frac{\sum_S w_k p_k d_k}{\sum_S w_k d_k} + \frac{\sum_{S'} \omega_k (y_k - p_k) d_k}{\sum_{S'} \omega_k d_k} \end{aligned} \tag{2.9}$$

$$\begin{aligned} \bar{y}_{C-\mathrm{BC}(d)} &= \bar{y}_{U(d)} + \left(\frac{\sum_S w_k c_k d_k}{\sum_S w_k d_k} - \frac{\sum_{S'} \omega_k c_k d_k}{\sum_{S'} \omega_k d_k} \right) \\ &= \frac{\sum_S w_k c_k d_k}{\sum_S w_k d_k} + \frac{\sum_{S'} \omega_k (y_k - c_k) d_k}{\sum_{S'} \omega_k d_k}. \end{aligned} \tag{2.10}$$

It is here where the bias-correction terms serve an important purpose. If the logistic model, which was fit on the subsample of *all* adults, holds within the domain, then $\sum_{S'} \omega_k d_k (y_k - p_k) \big/ \sum_{S'} \omega_k d_k$ will be an estimate of zero, and the model-driven probability estimator, $\bar{y}_{P(d)}$ in equation (2.7), will be nearly unbiased. If the model does not hold in the domain (e.g., if males are more likely to have SMI than the model predicts), then the model-driven probability estimator can be significantly biased.

Adding the bias correction $\sum_{S'} \omega_k d_k (y_k - p_k) \big/ \sum_{S'} \omega_k d_k$ to $\bar{y}_{P(d)}$ produces an estimator that is nearly unbiased under survey-sampling theory. When the model holds in the domain, however, applying the correction will almost certainly result in a decrease in accuracy. A similar argument can be made about the appropriateness of adding the $\sum_{S'} \omega_k d_k (y_k - c_k) \big/ \sum_{S'} \omega_k d_k$ term in equation (2.10) to the cut point estimator, $\bar{y}_{C(d)}$, in equation (2.8).

Equations (2.4) and (2.5) can be viewed as special cases of (2.9) and (2.10), respectively, with $d_k \equiv 1$.

3 The MHSS subsample

3.1 About the MHSS subsample

The NSDUH is a stratified multi-stage probability survey. In 2008 through 2012, the MHSS subsample was drawn annually from adults responding to the corresponding NSDUH using Poisson sampling. Subsample selection probabilities were determined each year using an algorithm that tended to oversample adults with higher levels of psychological distress. The algorithm varied across the years. See Center for Behavioral Health Statistics and Quality (2014, Chapter 3) for more details.

A respondent subsample size of roughly 750 was targeted for 2008 while respondent subsamples of 500 each for the 2009 and 2010, and 1,500 each for the 2011 and 2012 were likewise targeted. A data set combining all the respondent from 2008 to 2012 was created for modeling SMI. Weights for modeling were developed assuming that the same model held across all the years. As a result, more weight was given to the samples from 2011 and 2012 than to earlier years (Center for Behavioral Health Statistics and Quality, 2014; Chapter 5).

For our purposes, we treat those subsample weights and associated NSDUH weights as given and based on survey-sampling theory. We also treat the strata and two variance primary sampling units (PSUs) per each of the 50 variance strata developed for the MHSS subsample variance estimator as if they were the NSDUH variance strata and variance PSUs. Finally, we treat the NSDUH PSUs as if they were selected with replacement.

3.2 Variance estimation under survey-sampling theory

Since the bias-corrected estimated domain totals in equations (2.9) and (2.10) are nearly unbiased under survey-sampling theory, one can use linearization to estimate their variances. In what follows, we use variants of the bias-corrected estimators in equation (2.9) and (2.10) to simplify the variance estimation.

Recalling that $\omega_k = 0$ when $k \notin S'$, a variance estimator for the sample mean

$$\bar{y}_{z(d)} = \frac{\sum_S w_k z_k d_k}{\sum_S w_k d_k}, \tag{3.1}$$

under a stratified, multistage sample, where $z_k = p_k + (\omega_k / w_k)(y_k - p_k)$ is

$$v(\bar{y}_{z(d)}) = \frac{\sum_{h=1}^{50} \left[\sum_{k \in S_{h1}} w_k d_k \left(z_k - \bar{y}_{z(d)}\right) - \sum_{k \in S_{h2}} w_k d_k \left(z_k - \bar{y}_{z(d)}\right)\right]^2}{\left(\sum_S w_k d_k\right)^2}, \tag{3.2}$$

where S_{hj} are the respondents in the j^{th} variance PSU and variance stratum $h.$ It is also a variance estimator for the following asymptotically-identical variant of the bias-corrected probability estimator:

$$\bar{y}_{P-\mathrm{BC}2(d)} = \frac{\sum_S w_k p_k d_k}{\sum_S w_k d_k} + \frac{\sum_S \omega_k (y_k - p_k) d_k}{\sum_S w_k d_k}. \tag{3.3}$$

This is because the MHSS subsample is Poisson (and thus independent across adults as well as PSUs) and the first stage of the NSDUH sample is treated as if it were drawn with replacement.

Similarly, by redefining $z_k = c_k + (\omega_k / w_k)(y_k - c_k)$, a variance estimator for the sample mean in equation (3.1) is also an estimator for that variance of this variant of the bias-corrected estimator:

$$\bar{y}_{C-\mathrm{BC2}(d)} = \frac{\sum_S w_k c_k d_k}{\sum_S w_k d_k} + \frac{\sum_S \omega_k (y_k - c_k) d_k}{\sum_S w_k d_k}. \tag{3.4}$$

The variance estimation approach taken above assumes that the domain respondent subsample sizes are such that p_k / P_k and c_k / C_k can be treated as unity, where P_k and C_k are the limits of p_k and c_k, respectively, as the subsample (along with the NSDUH sample and population) grows arbitrarily large. In fact, all these ratios are assumed to be $1 + \mathrm{O}_P\left(1/\sqrt{n}\right)$, where n is the MHSS subsample size.

Consider now a computed bias-correction term, say $\sum_S \omega_k (y_k - p_k) d_k / \sum_S w_k d_k$ or $\sum_S \omega_k (y_k - c_k) d_k / \sum_S w_k d_k$. To assess whether the term is significantly different from zero, one can create an asymptotic $t-$ statistic in the usual fashion, dividing the term by its standard error.

When evaluating the estimators in Section 3.3, we will instead use the asymptotically equivalent:

$$BiasMeasure\left(\bar{y}_{P(d)}\right) = \sum_{S'} \omega_k (y_k - p_k) d_k \Big/ \sum_{S'} \omega_k d_k, \tag{3.5}$$

and

$$BiasMeasure\left(\bar{y}_{C(d)}\right) = \sum_{S'} \omega_k (y_k - c_k) d_k \Big/ \sum_{S'} \omega_k d_k \tag{3.6}$$

to create asymptotic $t-$ statistics for evaluating domain-level biases so that the DESCRIPT procedure in SUDAAN (RTI International, 2012) can be employed treating the p_k and c_k as fixed (similarly, that variance estimator for (3.3) in equation (3.2) can be computed using DESCRIPT). Moreover, since virtually all the sampling error in the bias-correction terms comes from the MHSS subsampling phase (even in 2011 and 2012, subsample was only 3% of the NSDUH adult sample), we treat the standard errors of the bias measures as if they were computed for a Poisson sample with ignorably small sampling fractions, which is equivalent to a with-replacement element sample for variance-estimation purposes. For example, for the variance estimator of $BiasMeasure\left(\bar{y}_{P(d)}\right)$, we compute (using SUDAAN's DESCRIPT): $v\left[BiasMeasure\left(\bar{y}_{P(d)}\right)\right] = \frac{n}{n-1} \sum_{S'} \left[\omega_k \left\{[y_k - p_k] - BiasMeasure\left(\bar{y}_{P(d)}\right)\right\} d_k\right]^2 \Big/ \left(\sum_{S'} \omega_k d_k\right)^2$, where n is the sample size of S'.

3.3 Evaluating the estimators

The model used by SAMHSA to predict SMI from adult NSDUH respondents was a logistic model with five variables (Center for Behavioral Health Statistics and Quality, 2014; Chapter 7). Two of the variables were rescaled total scores from short forms that measure psychological distress and functional impairment due to distress. The third was a dichotomous $(0/1)$ variable created from the answers to a series of questions assessing whether the respondent had a major depressive episode in the previous year. The fourth was also $0/1$ and indicated whether the respondent seriously contemplated suicide in the past year, and the fifth was a linear function of age from 18 to 30 that stayed constant after 30. Details on how this model was selected can be found in Center for Behavioral Health Statistics and Quality (2015, Chapter 4).

We used that model to create a set of domain-level cut point and probability estimates from the combined 2008-2012 data sets and to evaluate their potential biases. Some of the results are displayed in Tables 3.1

and 3.2. These tables reviewed domain estimates based on personal characteristics rather than state of residence because it seemed more likely that significant biases would be found for the characteristics like these rather than for states. Moreover, sample sizes for characteristics tended to be larger than those for states.

Table 3.1 show that using the bias-corrected probability in equation (2.9) is usually slightly more efficient (has a smaller standard error) than the direct estimator $\bar{y}_{U(d)} = \sum_{S'} \omega_k d_k y_k \Big/ \sum_{S'} \omega_k d_k$. The bias-corrected cut point estimator in equation (2.10) is sometimes more efficient than the direct estimator, sometimes not. The standard errors in Table 3.1 are the square roots of linearization variance estimators for the direct estimator $\bar{y}_{U(d)}$ above or the bias-corrected estimator in equation (3.1) with the appropriated defined nonrandom z_k, each computed as a stratified with-replacement sample of primary sampling units and a probability subsample of individuals within each PSU; that is, with equation (3.2). For $v\left(\bar{y}_{U(d)}\right)$, $z_k - \bar{y}_{z(d)}$ is replaced by $y_k - \bar{y}_{U(d)}$.

Table 3.1
Nearly unbiased estimators with their standard errors

	Direct (eq. 2.6)		**Bias-Corrected Cut Point (eq. 2.10)**		**Bias-Corrected Probability (eq. 2.9)**	
	Estimate	**SE**	**Estimate**	**SE***	**Estimate**	**SE***
All Adults	3.93	0.29	3.96	0.26	3.91	0.23
Male	2.96	0.34	2.91	0.39	3.01	0.31
Female	4.84	0.46	4.93	0.39	4.74	0.36
Age: 18-25	3.77	0.62	3.97	0.48	3.66	0.52
Age: 26-34	4.35	0.68	4.29	0.61	4.37	0.57
Age: 35-49	5.74	0.57	6.15	0.52	5.87	0.50
Age: 50+	2.74	0.40	2.47	0.47	2.60	0.36
White, Not Hispanic	4.43	0.35	4.47	0.30	4.34	0.27
Black, Not Hispanic	3.28	0.54	3.62	0.42	3.38	0.40
Other, Not Hispanic	4.09	1.25	4.27	1.10	4.33	1.12
Hispanic	2.02	0.71	1.68	0.88	2.11	0.70
Northeast	2.80	0.51	3.59	0.49	3.25	0.47
North Central	4.17	0.49	3.99	0.53	4.13	0.37
South	3.74	0.49	3.93	0.51	3.65	0.45
West	5.04	0.84	4.26	0.57	4.62	0.57
Employed Full Time	2.36	0.29	2.36	0.28	2.32	0.25
Employed Part time	4.34	0.71	3.82	0.55	3.91	0.46
Unemployed	5.64	1.22	6.57	0.92	6.13	0.90
Other Employment Status	6.21	0.66	6.22	0.64	6.15	0.55
Less than High School	5.69	0.99	4.44	0.77	4.72	0.71
High School Graduate	4.05	0.57	4.08	0.57	4.14	0.44
Some College	4.14	0.57	4.31	0.44	4.18	0.40
College Graduate	2.88	0.52	3.27	0.46	3.01	0.46
Metro	3.78	0.45	3.96	0.39	3.74	0.37
Small Metro	4.15	0.47	3.60	0.44	3.96	0.29
Nonmetro	3.99	0.47	4.63	0.54	4.36	0.48
Health Insurance: Yes	3.57	0.31	3.83	0.26	3.65	0.24
Health Insurance: No	5.73	0.94	4.65	0.93	5.24	0.74
< 100% of Poverty Level	9.01	1.30	9.00	1.23	8.62	1.05
100%-199% of Poverty	5.61	0.85	4.72	0.63	4.88	0.52
100% of Poverty	2.59	0.28	2.64	0.28	2.61	0.23
Rec'd MH Treatment: Yes	18.84	1.57	19.69	1.29	19.00	1.32
Rec'd MH Treatment: No	1.54	0.18	1.42	0.20	1.46	0.15

* Standard error is the square root of variance estimate computed using equation (3.2) with appropriately defined z_k.

The table suggests that there is little gain to be had from model correction, and leads us back to the model-driven probability and cut point estimators in equations (2.7) and (2.8) unless they exhibit systematic biases. Table 3.2 (with bias measures and their standard errors computed as described in Section 3.2) strongly suggests that the probability estimator, although unbiased when estimating SMI prevalence among all adults, can be very biased at the domain level. The cut point estimator, by contrast, is significantly biased at the 0.1 level in only two domains and never at the 0.05 level. Since we computed two-sided $p-$values for 32 domains, finding two domains with $p-$values below 0.1 is about what one should expect under the null hypothesis that the cut point estimator is not biased at the domain level.

Table 3.2
Model-driven estimates and their bias measures

	Standard cut point (eq. 2.8)			**Probability (eq. 2.7)**		
	Estimate	**Bias Measure**	**SE of Bias Measure**	**Estimate**	**Bias Measure**	**SE of Bias Measure**
All Adults	3.95	-0.01	0.27	3.91	0.00	0.23
Male	2.99	0.08	0.42	3.18	0.17	0.34
Female	4.84	-0.10	0.34	4.58	-0.16	0.31
Age: 18-25	3.94	-0.02	0.55	3.59	-0.07	0.49
Age: 26-34	5.03	0.69	0.66	4.64	0.26	0.51
Age: 35-49	5.08	-1.10*	0.57	4.77	-1.15**	0.55
Age: 50+	2.84	0.37	0.42	3.21	0.61*	0.32
White, Not Hispanic	4.31	-0.17	0.33	4.18	-0.16	0.28
Black, Not Hispanic	3.14	-0.48	0.45	3.38	0.00	0.46
Other, Not Hispanic	3.14	-1.14	1.13	3.47	-0.86	1.08
Hispanic	3.31	1.63*	0.85	3.28	1.17	0.65
Northeast	3.55	-0.04	0.39	3.62	0.33	0.35
North Central	4.16	0.16	0.60	4.02	-0.10	0.40
South	3.80	-0.13	0.52	3.86	0.22	0.44
West	4.28	0.02	0.56	4.10	-0.56	0.55
Employed Full Time	2.76	0.38	0.33	3.09	0.75**	0.28
Employed Part time	4.19	0.39	0.59	4.05	0.15	0.47
Unemployed	6.61	0.03	0.75	5.48	-0.57	0.70
Other Employment Status	5.33	-0.93	0.66	4.91	-1.30	0.56
Less than High School	4.34	-0.11	0.90	4.15	-0.64	0.83
High School Graduate	4.09	0.01	0.59	3.92	-0.22	0.46
Some College	4.50	0.18	0.37	4.35	0.17	0.31
College Graduate	3.09	-0.16	0.46	3.36	0.33	0.40
Metro	3.63	-0.34	0.38	3.68	-0.06	0.35
Small Metro	4.35	0.73	0.49	4.20	0.23	0.35
Nonmetro	4.24	-0.38	0.59	4.09	-0.27	0.51
Health Insurance: Yes	3.67	-0.16	0.27	3.72	0.07	0.24
Health Insurance: No	5.39	0.72	0.86	4.89	-0.34	0.68
< 100% of Poverty Level	7.21	-2.07	1.27	6.13	-2.88**	1.16
100%-199% of Poverty	4.83	0.12	0.62	4.53	-0.38	0.55
100% of Poverty	2.98	0.32	0.28	3.24	0.61***	0.21
Rec’d MH Treatment: Yes	18.33	-1.37	1.31	13.97	-5.07***	1.26
Rec’d MH Treatment: No	1.62	0.20	0.23	2.28	0.81***	0.17

* Bias measure is significantly different from zero at the 0.1 level.
** Bias measure is significantly different from zero at the 0.05 level.
*** Bias measure is significantly different from zero at the 0.01 level.

One curious result bears a brief mention. The cut point estimator among all adults had very little bias (-0.01), so its estimated root mean squared error equaled the standard error of the bias-corrected cut point after rounding (0.26). Oddly, this value was less than the standard error of its bias measure (0.27). One possible reason for the difference between the two standard errors was that we used $\sum_S \omega_k (y_k - c_k) d_k \Big/ \sum_S w_k d_k$ as the bias-correction term and $\sum_S \omega_k (y_k - c_k) d_k \Big/ \sum_S \omega_k d_k$ as the bias measure within a domain; all adults being the special case where $d_k \equiv 1.$ Our analysis (not shown) was that the difference in the denominators had very little impact.

What has a greater impact was ignoring the stratification and clustering in the NSDUH sample when computing the standard errors of the bias measures. Unexpectedly, ignoring the clustering actually tended to increase standard errors. This may be because the clustering in the NSDUH has virtually no measurable impact on variance so that any difference between standard error estimates computed with and without clustering is attributable to random noise or to asymptotic biases that are not actually ignorable in finite estimates.

3.4 A hybrid cut point

Consider the following hybrid of the probability and standard cut point estimators. Suppose we sorted the NSDUH sample rather than just the MHSS subsample by the fitted p_k values, and established a cut point p_H such that

$$\sum_{\substack{k \in S \\ p_k \geq p_H}} w_k = \sum_{k \in S} w_k p_k \tag{3.7}$$

holds as closely as possible. Setting $h_k = 1$ when $p_k > p_H$ and 0 otherwise, the hybrid cut point estimator for SMI prevalence in a domain is

$$\bar{y}_{H(d)} = \frac{\sum_S w_k h_k d_k}{\sum_S w_k d_k}. \tag{3.8}$$

It is not hard to see that for all adults, if a p_H could be found that satisfied equation (3.7), then the hybrid cut point estimator would equal the probability estimator exactly. Failing that the hybrid cut point estimator for all adults would have a slight bias, which could be measured, squared, and then added to the standard error of the probability estimator to equal its root mean squared error. In this case, the hybrid SMI prevalence estimate for all adults rounded to 3.89. Its root mean squared error rounded to the same value as the standard error of the probability estimator (0.23).

Table 3.3 repeats much of Table 3.2 for the standard cut point but also displays analogous results for the hybrid

$$(BiasMeasure(\bar{y}_{H(d)})) = \sum_S \omega_k (y_k - h_k) d_k \Big/ \sum_S \omega_k d_k. \tag{3.9}$$

Its standard error is computed analogously to those of $\bar{y}_{C(d)}$ and $\bar{y}_{P(d)}$. The two sets of cut point outcomes are similar, but the bias measure for the hybrid estimator was significantly different from zero at the 0.05 level in two domains (both with $p-$ values of 0.043). Since there are 32 domains analyzed, this remains consistent with the null hypothesis of no bias at the domain level.

Table 3.3
The cut point estimators and their bias measures

	Standard Cut Point (eq. 2.8)			Hybrid Cut Point (eq. 3.8)		
	Estimate	Bias Measure	SE of Bias Measure	Estimate	Bias Measure	SE of Bias Measure
All Adults	3.95	-0.01	0.27	3.89	-0.10	0.27
Male	2.99	0.08	0.42	2.94	0.03	0.42
Female	4.84	-0.10	0.34	4.78	-0.21	0.33
Age: 18-25	3.94	-0.02	0.55	3.89	-0.03	0.55
Age: 26-34	5.03	0.69	0.66	4.97	0.68	0.66
Age: 35-49	5.08	-1.10*	0.57	5.02	-1.16**	0.57
Age: 50 or Older	2.84	0.37	0.42	2.79	0.22	0.41
White, Not Hispanic	4.31	-0.17	0.33	4.24	-0.22	0.33
Black, Not Hispanic	3.14	-0.48	0.45	3.10	-0.48	0.45
Other, Not Hispanic	3.14	-1.14	1.13	3.11	-1.14	1.13
Hispanic	3.31	1.63*	0.85	3.25	1.30	0.79
Northeast	3.55	-0.04	0.39	3.50	-0.05	0.39
North Central	4.16	0.16	0.60	4.12	0.07	0.59
South	3.80	-0.13	0.52	3.74	-0.29	0.51
West	4.28	0.02	0.56	4.23	0.01	0.56
Employed Full Time	2.76	0.38	0.33	2.71	0.36	0.33
Employed Part Time	4.19	0.39	0.59	4.16	0.37	0.59
Unemployed	6.61	0.03	0.75	6.43	-0.27	0.69
Other Employment Status	5.33	-0.93	0.66	5.27	-1.09*	0.65
Less than High School	4.34	-0.11	0.90	4.21	-0.14	0.90
High School Graduate	4.09	0.01	0.59	4.03	-0.26	0.56
Some College	4.50	0.18	0.37	4.45	0.18	0.37
College Graduate	3.09	-0.16	0.46	3.07	-0.17	0.46
Large Metro	3.63	-0.34	0.38	3.58	-0.36	0.38
Small Metro	4.35	0.73	0.49	4.27	0.58	0.47
Nonmetro	4.24	-0.38	0.59	4.19	-0.53	0.58
Health Insurance: Yes	3.67	-0.16	0.27	3.62	-0.20	0.27
Health Insurance: No	5.39	0.72	0.86	5.31	0.44	0.82
< 100% of Poverty Threshold	7.21	-2.07	1.27	7.12	-2.44**	1.21
100%-199% of Poverty	4.83	0.12	0.62	4.78	-0.01	0.61
> 200% of the Poverty	2.98	0.32	0.28	2.93	0.30	0.28
Rec'd MH Treatment: Yes	18.33	-1.37	1.31	18.19	-1.46	1.31
Rec'd MH Treatment: No	1.62	0.20	0.23	2.28	0.81	1.17

* Bias measure is significantly different from zero at the 0.1 level.
** Bias measure is significantly different from zero at the 0.05 level.

4 Some concluding remarks

Population mental health attributes have been estimated using data from a subsample selected from a large general survey that are given a clinical diagnostic assessment to develop prediction models that are then applied to the full sample (see, for example, Kessler, Abelson, Demler, Escobar, Gibbon, Guyer, Howes, Jin, Vega, Walters, Wang, Zaslavsky and Zheng, 2004). This is the methodology used by SAMHSA with the annual NSDUH and the adult MHSS subsample from 2008 through 2012.

We have shown with NSDUH/MHSS subsample data that when estimating the SMI prevalence for the full population, in this case adults, using the estimated probabilities of having SMI directly instead of the standard cut point methodology results in a lower standard error. Nevertheless, the so-called probability estimator of SMI prevalence can often be substantially biased at domain-level when the standard cut point is not.

We also investigated bias-corrected versions of the two estimators based on survey-sampling theory. Unfortunately, these estimators were only slightly more efficient than simply computing estimates from the MHSS subsample directly, especially for subpopulations.

We evaluated a hybrid cut point estimator that was slightly more efficient than the standard cut point in estimating SMI prevalence among all adults. It did not exhibit the large biases at the domain level that plagued the probability estimator, although whether or not it was free of domain-level bias was not completely clear.

In 2013, SAMHSA discontinued the collection of clinical interviews. Nevertheless, the agency continues to compute SMI prevalence estimates for adults based on the model and cutpoints developed using the 2008-2012 MHSS subsample. The standard cut point estimator which demonstrated smaller domain-level biases in our tables also continues to be used for SMI estimation at the state level as well as for the domains analyzed here.

A troubling question is how a standard cut point estimator for SMI derived from a logistic-model fit of MHSS subsample data can be less prone to bias than a probability estimator based on the sample model fit. We suspect that logistic regression does a reasonably good job at ordering the relative probabilities of adults having SMI but not at estimating individual probabilities, especially in the tails. Domains with unusually high and low prevalences, like adults receiving mental-health treatment (or not receiving treatment) were particularly prone to having biased probability estimates. It may be that the application of asymptotic theory is not appropriate in the tails.

The interested reader would likely want to know how one can compute standard errors for the cut point estimators at the domain level. Because a cut point estimator is not continuous, we attempted to compute standard errors for it using Fay's Balances repeated replications (BRR). Unfortunately, as explained in Center for Behavioral Health Statistics and Quality (2015, Chapter 2.4.2.), our variance estimators for domain estimates were not satisfactory in a modest simulation experiment.

Few MHSS subsample respondents were in the tails of the distribution (and even fewer realized values for the probability of having SMI given the covariates of the model). This frustrated our attempts at improving the probability estimator (for which domain-level standard errors can be measured via linearization) by adjusting tail probabilities (which did not fit very well under the logistic model). We will not discuss these attempts further here. More research is clearly needed to develop either good standard-error measures for standard cut point estimates or good estimates of SMI prevalence for which standard errors can be reasonably measured.

A final word about the "gold standard" is warranted. For our purposes, a clinical diagnosis of an adult having SMI has been treated as equivalent to the person actually having diagnosable seriously mental illness. In fact, diagnoses are more fluid. They may vary according to the clinician or the whim of the person answering the clinician's questions. CBHSQ (2014, Chapters 5 and 6) describes the effort expended on removing as much variation from the MHSS clinical diagnoses as possible. We assumed here that each clinical diagnosis was effectively unbiased, that is, the probability of a random diagnosis for an individual within a domain being a false positive equaled the probability of it being a false negative.

References

American Psychiatric Association (1994). *Diagnostic and Statistical Manual of Mental Disorders (DSM-IV)* (4th Ed.). Washington, DC: American Psychiatric Association.

Center for Behavioral Health Statistics and Quality (2015). *Estimating Mental Illness among Adults in the United States: Revisions to the 2008 Estimation Procedures*. Substance Abuse and Mental Health Services Administration, Rockville, MD: http://www.samhsa.gov/data/sites/default/files/NSDUH-N8-EstimatingMI-2012.pdf.

Center for Behavioral Health Statistics and Quality (2014). *2012 National Survey on Drug Use and Health: Methodological Resource Book (Section 16a, 2012 Mental Health Surveillance Study: Design and Estimation Report)*. Substance Abuse and Mental Health Services Administration, Rockville, MD: http://www.samhsa.gov/data/sites/default/files/NSDUH2012MRB-Ammended/NSDUHmrbMHSS-DesignEst2012.pdf.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, 27, 861-874.

Kessler, R.C., Abelson, J., Demler, O., Escobar, J., Gibbon, M., Guyer, M., Howes, M., Jin, R., Vega, W., Walters, E., Wang, P., Zaslavsky, A. and Zheng, H. (2004). Clinical calibration of DSM-IV diagnoses in the World Mental Health (WMH) version of the World Health Organization (WHO) Composite International Diagnostic Interview (WMHCIDI). *International Journal of Methods in Psychiatric Research*, 13, 122-139.

Kessler, R.C., Chiu, W., Demler, O. and Walters, E. (2005). Prevalence, severity, and comorbidity of 12-month DSM-IV disorders in the National Comorbidity Survey Replication. *Archives of General Psychiatry*, 62, 617-627.

Lehtonen, R., and Veijanen, A. (1998). Logistic generalized regression estimators. *Survey Methodology*, 24, 1, 51-55. Paper available at https://www150.statcan.gc.ca/n1/pub/12-001-x/1998001/article/3909-eng.pdf.

RTI International (2012). *SUDAAN Language Manual, Volumes 1 and 2, Release 11*. Research Triangle Park, NC: Research Triangle Institute.

Särndal, C.-E., Swensson, B. and Wretman, J. (1989). The weighted residual technique for estimating the variance of the general regression estimator. *Biometrika*, 76, 527-537.

Survey Methodology, June 2018
Vol. 44, No. 1, pp. 75-99
Statistics Canada, Catalogue No. 12-001-X

Strategies for subsampling nonrespondents for economic programs

Katherine Jenny Thompson, Stephen Kaputa, and Laura Bechtel[1]

Abstract

The U.S. Census Bureau is investigating nonrespondent subsampling strategies for usage in the 2017 Economic Census. Design constraints include a mandated lower bound on the unit response rate, along with targeted industry-specific response rates. This paper presents research on allocation procedures for subsampling nonrespondents, conditional on the subsampling being systematic. We consider two approaches: (1) equal-probability sampling and (2) optimized allocation with constraints on unit response rates and sample size with the objective of selecting larger samples in industries that have initially lower response rates. We present a simulation study that examines the relative bias and mean squared error for the proposed allocations, assessing each procedure's sensitivity to the size of the subsample, the response propensities, and the estimation procedure.

Key Words: Quadratic program; Unit response rate; Nonresponse adjustment; Systematic sampling; Optimal allocation; Two-phase sampling.

1 Introduction

Many federal programs are simultaneously experiencing declining response rates and reductions in funding. At the same time, these programs are required to maintain predetermined reliability levels and are encouraged to collect an increased number of data items and to publish more statistics. Of course, as nonresponse increases, the precision of the survey estimates will decrease from the original design levels and can be sensitive to nonresponse bias. Consequently, many federal agencies are investigating adaptive collection design strategies, where the term "collection design" refers to protocol(s) for collecting data.

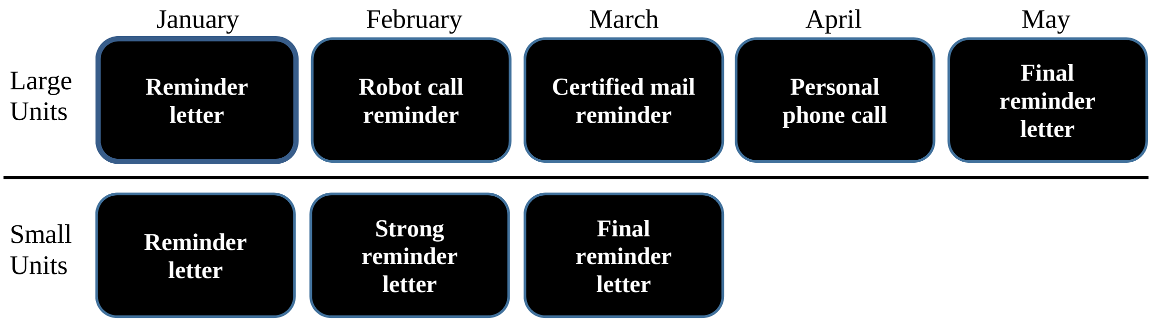
With business surveys, the collection design may vary by type of unit. These populations are generally highly skewed; the majority of a tabulated total in a given industry is often provided by a small number of large businesses. Because publication statistics are generally industry totals or percentage change, missing data from the largest cases can induce substantive nonresponse bias in the totals, whereas missing data from the smaller cases (even those with large sampling weights) often have little *apparent* effect on the tabulated levels (Thompson and Washington, 2013). Thus, the contact strategies are designed to ensure that the largest cases provide valid response data. Figure 1.1 illustrates nonresponse follow-up (NRFU) procedures that differ by a survey-specific unit size classification, where both collection designs have fixed calendar schedules and a fixed NRFU budget.

For the large unit category, the NRFU procedures become progressively more costly (per unit) with the exception of the final contact attempt. In contrast, with the smaller units, the NRFU procedures do not include personal contact and are therefore less expensive.

Selecting a probability subsample of nonrespondents is a strategic feature of many responsive and adaptive collection designs (Tourangeau, Brick, Lohr and Li, 2016). Of course, this is not a new practice

[1] Katherine Jenny Thompson, Economic Statistical Methods Division, U.S. Census Bureau, Washington, DC 20233. E-mail: Katherine.J.Thompson@census.gov; Stephen Kaputa, Economic Statistical Methods Division, U.S. Census Bureau, Washington, DC 20233. E-mail: Stephen.Kaputa@census.gov; Laura Bechtel, Economic Statistical Methods Division, U.S. Census Bureau, Washington, DC 20233. E-mail: Laura.Bechtel@census.gov.

for surveys. Indeed, nonrespondent subsampling has been a survey practice since first discussed in Hansen and Hurwitz (1946). Actually, the setting of the two-phase sample approach presented in Hansen and Hurwitz (1946) paper is quite similar to the business survey setting discussed here: an "inexpensive" mailed questionnaire to all sampled units (c.f. the "21st century design" that mails a letter containing a URL, user name, and password), followed by "expensive" personal interviews of subsampled nonrespondents (c.f. personal phone calls or certified reminder letters). Their proposed optimal allocation procedures are not entirely dissimilar either, with the final allocations being highly dependent on whether the response rates for each collection mode are known or estimated using auxiliary data rather than the previously collected responses.



	January	February	March	April	May
Large Units	**Reminder letter**	**Robot call reminder**	**Certified mail reminder**	**Personal phone call**	**Final reminder letter**
Small Units	**Reminder letter**	**Strong reminder letter**	**Final reminder letter**		

Figure 1.1 Nonresponse follow-up procedures for differing types of business in a fictional survey.

Fitting nonrespondent subsampling into a responsive or adaptive design framework is straightforward. As originally proposed by Groves and Herringa (2006), responsive designs require a minimum of two distinct phases of collection, with the second phase often being a probability subsample of nonrespondents that occurs at the "phase capacity" when the survey estimates are no longer changing, providing evidence the existing collection protocol is no longer cost-effective. Schouten, Calinescu and Luiten (2013) characterize responsive designs as a special case of adaptive collection designs. With an adaptive collection design, the data collection procedures can change (*adapt*) during the collection period. Paradata and sample data are used to determine whether to change the current procedures. The overall budget is fixed, but the implementation of a given strategy depends on (1) the realized sample of respondents at a point in time, (2) informative data obtained during data collection about the respondents and nonrespondents, and (3) information known in advance about the survey unit from the sampling frame. Consequently, selecting a probability sample of nonrespondents for NRFU – instead of attempting to contact all nonrespondents – falls under the adaptive design umbrella, with paradata (specifically response status) used to determine the sampling frame and frame data (e.g., the unit's size and industry classification) used as the basis of the sample design.

The U.S. Census Bureau is investigating nonrespondent subsampling strategies for the 2017 Economic Census (EC). Although a single program, the EC employs different sampling designs by sector (Probability proportional to size for the Construction sector, cut-off sampling for the Manufacturing and Mining sectors in collections prior to 2017, complete enumeration for the Wholesale Trade sector, and stratified simple random sampling without replacement (SRS-WOR) in the remaining sectors). Moreover, as is typical with many business programs, it is a multi-purpose collection, with the general statistics items collected from all surveyed units in a sector: examples include – but are not limited to – receipts/shipments, annual and first quarter payroll, and total employment. In addition, the EC collects information on product sales, types of which differ by sector and often industry. Imputation procedures differ by item, as do the estimators. Consequently, the subsampling design must be robust to sampling and estimator to the largest extent possible. We consider a *systematic sample* of nonrespondents sorted by a measure of size, a sampling design known to be as efficient as stratified simple random sampling without replacement (SRS-WOR) on average if the list is in random order and more efficient if the list is monotonic increasing or decreasing (Zhang, 2008; Lohr, 2010, Chapter 2, pages 50-51).

Ideally, the nonrespondent subsampling allocation procedure should be informed by properties of the respondent sample *during* the collection period. Of course, if the program is designed to collect one or two key items, then the allocation procedures should (at least attempt to) directly incorporate information on the survey design and estimation procedure, as well as detailed cost information, as proposed in Hansen and Hurwitz (1946) long-ago. In this case, one should use an optimal allocation procedure that minimizes costs subject to (estimable) reliability constraints. See Harter, Mach, Chaplin and Wolken (2007) and Beaumont, Bocci and Haziza (2014) for examples.

Such optimization is difficult to accomplish in the considered multi-purpose survey setting, especially when strongly correlated auxiliary variables are not available for all items. However, the OMB Statistical Standards for federal surveys require "survey (design) to achieve the highest practical rates of response, commensurate with the importance of survey uses, respondent burden, and data collection costs" and mandate nonresponse bias analyses for programs that fail to achieve these rates (Federal Register Notice, 2006). For nonrespondent subsampling occurring *during* the data collection cycle, imposing mandated lower bounds on the program-level response rate and in specified domains (examples include sampling strata or other post-strata such as industry code or type of government) is therefore a natural constraint to include in the allocation procedure.

In this paper, we explore allocation approaches that address such constraints, with an overall objective of selecting larger systematic subsamples in domains that have lower-than-targeted response rates. We introduce two optimized allocation procedures, both formulated as quadratic programs and solved with standard software packages: one that minimizes deviations between domain unit response rates and one that minimizes deviations between domain subsampling intervals. Our case study compares the statistical properties of subsamples obtained from each proposed allocation with three different estimators, considering two ratio estimators commonly used by business surveys along with the simple expansion (Horvitz-Thompson) estimator. The latter is not necessarily the most precise estimator when highly correlated auxiliary data are available, but gives an "upper bound" on the variance increase due to subsampling. The ratio estimators were selected to illustrate that the subsampling variance component can be reduced by incorporating correlated auxiliary data at the estimation stage.

Note that the presented allocation procedure is designed specifically for business surveys and implicitly assumes that largest units are excluded from the subsampling. In this case, the overall cost savings may not be substantial because the majority of a program's NRFU budget will be likely allocated to obtaining responses from the designated larger cases. However, the estimate quality can be improved. By equalizing response rates in considered domains, we hope to reduce the bias of the estimates by obtaining a respondent set that resembles the parent sample. Moreover, equalizing the subsampling intervals should help avoid overly increasing the sampling variance due to the second phase of selection, an unpleasant side effect of the additional stage of sampling that can completely offset any bias reduction obtained via the probability subsample (Biemer, 2010). And, it may be possible to further reduce both nonresponse bias and subsampling variance via an improved ratio or regression estimation procedure, if related covariates are available.

Section 2 provides context, briefly introduces the studied estimators, and presents our allocation procedures. Section 3 presents a simulation study that compares the statistical properties of the considered estimators for each realized allocation. We conclude in Section 4 with recommendations and suggestions for future research.

2 Methodology

2.1 Survey design and estimation

The general framework for our research is the two-phase sample design shown in Figure 2.1. The first stage is a stratified probability sample with a total sample size of n from a finite population (frame) of size N, performed *before* data collection begins. The survey is conducted, and units either respond or do not. During the data collection, response rates are monitored in H domains, where the domains do not necessarily equal the sampling strata. For example, total response rates might be monitored by three-digit industry classification, although these industry sampling strata are further broken down by size class. Furthermore, the domains could be independent of the original sampling strata e.g., race or sex categories (resembling post-strata). Hereafter, the term "domain" refers to the nonrespondent subsampling strata, indexed by $h\,(h = 1, 2, \ldots, H)$.

The second stage of probability sampling occurs at a predetermined point in the data collection cycle when we select an overall $1 - \text{in} - K$ subsample of size m_1 from the m nonrespondents (a two-phase sample); this predetermined point can be a fixed calendar date or via a responsive/adaptive design protocol. The value of K is determined by the program managers, who take into account the overall budget for NRFU (assumed fixed), mandated performance measures (e.g., response rates, coefficient of variation requirements), and other operational considerations such as length of collection period and available resources. Our allocation procedure determines the $1 - \text{in} - K_h$ systematic subsample of size m_{1h} from the m_h nonrespondents in each domain. Only the sampled m_{1h} units receive NRFU.

Our objective is to estimate Y, the population total of characteristic y. This estimate is $\hat{Y} = \hat{Y}_{R1} + \hat{Y}_{R2}$ where $\hat{Y}_{R1}$ is estimated from the r_{1h} first-stage sample respondents and $\hat{Y}_{R2}$ is estimated from the r_{2h} second-stage sample respondents (see Figure 2.1). Nonresponse adjustments to the r_{2h} subsampled (responding) units assume a missing at random response (MAR) mechanism, treated as a Bernoulli sample (Särndal, Swensson and Wretman, 1992, Chapter 15; Kott, 1994). We consider three different adjustment-to-sample reweighting estimators of $\hat{Y}_{R2}$ (Kalton and Flores-Cervantes, 2003): the double reweighted expansion (DE) estimator (Binder, Babyak, Brodeur, Hidiroglou and Wisner, 2000; Shao and Thompson, 2009; Haziza, Thompson and Yung, 2010), a separate ratio (SR) estimator that adjusts for unit nonresponse using a covariate that is highly correlated with both response propensity and the survey characteristic of interest (Shao and Thompson, 2009; Haziza et al., 2010), and a combined ratio (CR) estimator (Binder et al., 2000). Formulae are provided in the Appendix.

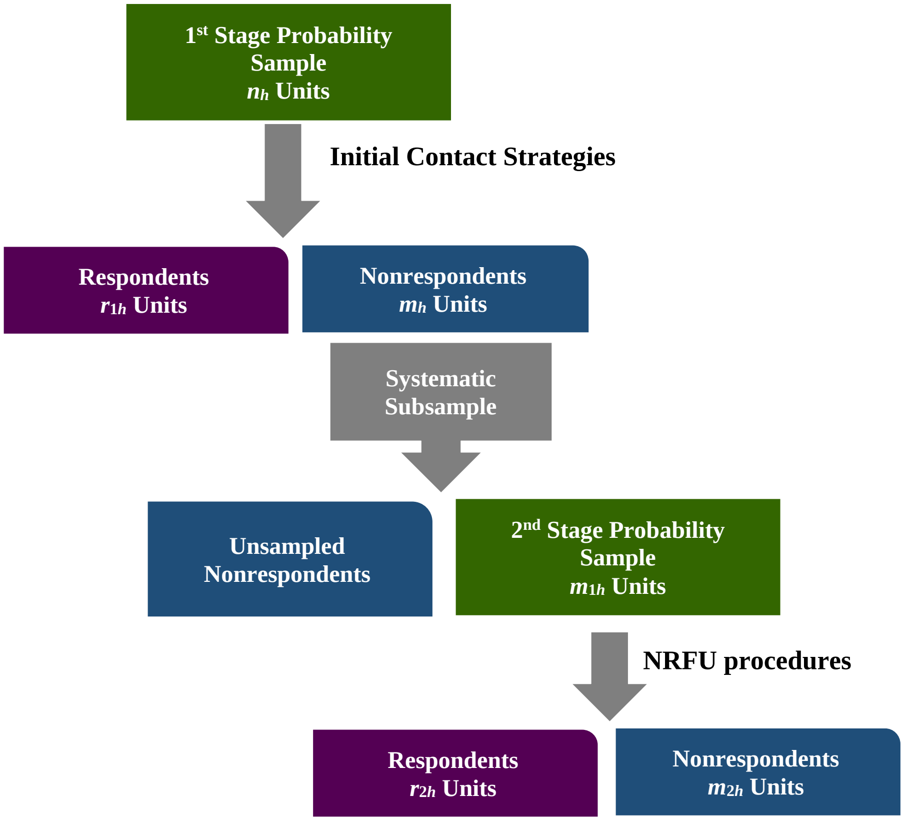


Figure 2.1 Nonrespondent subsample from probability sample, selected during data collection (two-phase sample design). Unsampled nonrespondents do not receive NRFU.

These estimators require a *minimum* of $r_{2h} = 1$ in each domain and a minimum of $r_{2h} = 2$ for variance estimation. These minimal conditions may not hold for several reasons. During the early stages of NRFU collection, an insufficient number of the subsampled units might respond in a given domain. Alternatively, the allocation procedure could determine that no subsampling is required in one or more domains. Lastly, the allocation procedure could require 100-percent follow-up (*all* units subsampled) in selected domains; henceforth, we refer to 100-percent follow-up/no subsampling as “full follow-up”. In these cases, the estimation procedure ignores the last stage of sampling as if it did not occur and produces estimates for domain h using the collapsed estimator formulae provided in the Appendix.

2.2 Allocation strategies

When *all* nonresponding cases are subjected to NRFU, respondent contact strategies focus on improving overall response rates. Analysts might focus primarily on obtaining responses from soft refusal cases that they believe have similar characteristics to previous respondents (“quick wins”), although this phenomenon is more likely when the survey collection is performed in the field, as with household or agricultural surveys, and perhaps is less likely for internet or mail collections. With business surveys, the size of the unit is a factor in the NRFU procedures as discussed in Section 1.

Our objective is to obtain a realized set of respondents that approximates a random subsample of the originally selected sample via a probability sample of nonrespondents. With a probability sample, the targeted cases represent a cross-section of the nonrespondent population. By focusing contact efforts on the subsample, we hope to decrease the effects of nonresponse bias on the estimated totals by obtaining data from all types of nonresponding units. Moreover, weighting or imputation methods may be more effective at reducing the nonresponse bias effects with a probability subsample of nonrespondents (Brick, 2013). Even though they do not receive additional NRFU, the unsampled nonrespondent cases may provide responses later in the collection cycle. If so, an unbiased estimation procedure would not include the unsampled late responses in the final estimate assuming that all subsampled units respond, as these units are represented by the subsampled cases. However, this procedure is extremely distasteful to many survey managers. Instead, we include their data in the tabulations as if they had responded *before* subsampling. This does induce bias in the estimate. In practice, we ensure that this situation occurs infrequently by subsampling late in the data collection cycle.

With a business survey that keeps track of little or no demographic information, most of the information on the nonrespondents such as industry and unit size (e.g., total payroll, total receipts) is obtained from the sampling frame. Sorting the nonrespondents within prespecified domains by unit size and selecting a systematic sample should yield a subsample that resembles the originally designed sample in terms of unit size composition. This is especially important for business surveys where responses tend to be obtained from the larger units (Thompson and Washington, 2013). The choice of subsampling domain is determined by overall survey objectives such as publication levels or by the adjustment cell design (e.g., weighting cells or imputation classes), although computations are considerably simplified when the domain of interest is the original sampling strata. In the EC, the industry is the domain of interest.

We consider two allocation approaches: (1) equal-probability sampling; and (2) optimized allocation with constraints on unit response rates and sample size in predetermined domains. Equal probability sampling is easy to implement and should have the lowest sampling variance among considered nonrespondent subsampling allocation strategies, since the subsampling weight adjustment will be a constant value in all domains. However, since the same proportion of nonrespondents is sampled in each domain, the subsample may not be large enough to offset nonresponse bias effects on totals in low-responding domains. We refer to the allocations obtained by equal probability sampling as $\text{Constant}-K$, where K refers to the overall sampling interval $(1-\text{in}-K)$.

Our optimized allocation methods address the above concern by concentrating NRFU efforts in domains that have low response rates, attempting to select sufficient cases to achieve the performance benchmarks. This strategy may decrease the nonresponse bias in the totals if the response mechanism is MAR, conditional on the auxiliary variables used to define the domains; see Wagner (2012). However, it can increase the variance, as the subsampling intervals will differ and the weights will become more variable. To minimize the additional sampling variance caused by differing sampling intervals, the domain *nonrespondent* *subsampling* intervals should be close to the overall nonrespondent subsampling interval. To control costs, the allocation should not select more units for NRFU than budgeted. Recall that the federal survey environment requires that target response rates be achieved or nearly achieved, which makes all domains "equally" important from a data collection viewpoint.

To describe the allocation procedures, we introduce additional notation:

Unit response rate: $$\mathrm{URR} = \frac{\sum_h (r_{1h} + r_{2h})}{\sum_h n_h}$$

Target response rate: $$\mathrm{URR}^{\mathrm{T}} = \frac{\sum_h r_{1h} + (q_h m_h / K)}{\sum_h n_h}$$

Target domain response rate: $$\mathrm{URR}_h^{\mathrm{T}} = \frac{r_{1h} + (q_h m_h / K_h)}{n_h}$$

with r_{1h} units of the n_h originally sampled units responding *before* subsampling, leaving m_h units available for subsampling in each domain. The unit response rate (URR) is the *actual* proportion of responding sampled units (Thompson and Oliver, 2012) and *does not* include an adjustment for subsampling. The target response rate $(\mathrm{URR}^{\mathrm{T}})$ used for allocation is the expected maximum obtainable URR for a given overall subsampling rate $K,$ with q_h representing the conditional probability of ultimately responding to the census/survey in domain $h,$ given that the unit did not respond prior to subsampling. In the allocation procedure, q_h can be modeled from historical data if available or can be assumed constant for a new survey or for sensitivity analyses.

We formulate optimized allocation as a quadratic program and consider two different objective functions. The first quadratic program minimizes the squared deviation of the target response rate in each domain $\mathrm{URR}_h^{\mathrm{T}}$ from the overall target unit response rate $\mathrm{URR}^{\mathrm{T}},$ subject to linear constraints on the size of nonrespondent sample. This objective function is analogous to the numerator of the Pearson chi-square goodness-of-fit test.

The second quadratic program minimizes the squared deviation in domain sampling intervals from the *overall sampling interval* (K) subject to linear constraints on the unit response rates in each domain and on the number of sampled nonrespondents. Thus, although the optimization procedure allows the sampling intervals to vary by domain, the program tries to avoid potentially large increases in variance caused by the deliberately introduced "disproportionate sampling fractions" referred to in Kish (1992). We refer to the allocations obtained from these quadratic programs as $\mathrm{Min}-\mathrm{URR}$ and $\mathrm{Min}-K$ respectively.

Both quadratic programs are primarily deterministic. However, recall that at the allocation stage, we must estimate the number of subsampled units that will eventually respond in each domain. Both quadratic programs use Constraints (1) through (3) in Table 2.1. Constraint (4) is included in the $\mathrm{Min}-K$ allocation to ensure that the optimization solution is not $K_h = K$ for all domain $h.$ There are two limiting scenarios (preconditions) that are addressed before the $\mathrm{Min}-K$ optimization. First, domains whose $\mathrm{URR}_h^{\mathrm{T}} \geq \mathrm{URR}^{\mathrm{T}}$ *before* subsampling must be removed from the optimization problem $(K_h = \infty).$ Second, if the estimated unit response rate cannot be possibly achieved in a given domain for an assumed $q_h,$ then all units in the domain are selected for NRFU $(K_h = 1)$. The $\text{Min}-K$ optimization is applied to the remaining domains, requiring that these subsampled domains have expected URRs that meet or exceed the target URRs.

Using sample data containing respondents and nonrespondents, along with different specified values for q_h, we use the SAS® PROC NLP (The data analysis for this paper was generated using SAS software. Copyright, SAS Institute Inc. SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc., Cary, NC, USA.) to solve the quadratic programs (obtaining the set of K_h). The realized allocations are not integer values, and the real valued intervals (K_h) were input to SAS® PROC SURVEYSELECT to select stratified systematic subsamples of nonrespondents. As noted by one reviewer, this yields a solution that is randomly rounded but constrained at the overall required sample size, and there may be some impact on reliability due to rounding error. Such effects were not studied in this paper.

Table 2.1
Optimized allocation quadratic programs

		Min−URR	**Min−*K***	**Purpose**
Objective Function		$\min\sum_h (\text{URR}_h^T - \text{URR}^T)^2$	$\min\sum_h (K_h - K)^2 = \min\sum_h \left(\frac{m_h}{m_{1h}} - K\right)^2$	
Constraints	(1)	$K \le \sum_h m_h \big/ \sum_h m_{1h}$		Selected sample size cannot exceed overall 1-in-K sample size
	(2)	$m_h / m_{1h} \ge 1$		Domain subsample cannot exceed number of nonrespondents in the strata
	(3)	$m_{1h} \ge 0$		Non-negativity constraint
	(4)	Not Applicable	$\frac{r_{1h} + q_h m_h}{n_h} < \text{URR}^T \quad K_h = 1$ $\frac{r_{1h}}{n_h} \ge \text{URR}^T \quad K_h = \infty$ $\text{URR}_h^T \ge \text{URR}^T \quad$ otherwise	Ensures that all domains achieve target URR as feasible.

3 Case study

This section presents the results of a simulation study that evaluates the considered allocation procedures on empirical sample data from the Annual Survey of Manufactures (ASM) from the 2010 and 2011 data collections. For more information on the ASM, see http://www.census.gov/manufacturing/asm.

The ASM is an establishment survey designed to produce "sample estimates of statistics for all manufacturing establishments with one or more paid employee(s)" (http://www.census.gov/manufacturing/asm/); it is a Pareto-PPS sample of approximately 50,000 establishments selected from a universe of 328,500. Approximately 20,000 establishments are included with certainty, and the remaining establishments are selected with probability proportional to a composite measure of size. Selected units are in the sample for the four years between censuses. Sampling strata are defined by six-digit industry code using the North American Industry Classification System.

The ASM estimates totals with a difference estimator (Särndal et al., 1992). To reduce respondent burden, units below a certain threshold are dropped from the sampling frame entirely. Instead, their data are imputed using administrative data values for selected items and industry-level regression models for the remaining items. Similarly, the ASM imputes complete records for unit nonrespondents. See http://www.census.gov/manufacturing/asm/ for additional information on the ASM methodology.

Because the items collected by the ASM questionnaire are a subset of the EC's manufacturing sector items, the ASM is often used to pretest new EC processing or data collection procedures. With the ASM and the EC, implementing a probability subsample of nonrespondents for NRFU represents a major procedural change. The ASM NRFU procedures are very similar to the EC procedures. Because a given company can comprise several establishments, the sets of multi-unit (MU) establishments corresponding to the company can be designated for phone follow-up as well as other company completeness checks. In contrast, the NRFU procedures for the single unit (SU) establishments – establishments with one location and parent company – differ. The largest SU establishments are included with certainty (sampled with probability = 1) and may receive a personal phone call in selected domains. The sampled SU establishments ("SU noncertainty establishments") receive some reminders, but are very unlikely to receive a personal phone call.

Our simulation study examines one of the fourteen key ASM items and employs the double expansion estimate and the two ratio estimators described in the Appendix, not the difference estimator used in ASM production estimates. Consequently, our results should not be extrapolated to the ASM.

3.1 Simulation study design

Our simulation study compares the statistical properties of total shipment estimates obtained from the three considered nonrespondent subsampling designs over repeated samples, using three different estimators. Our *sampling frame of nonrespondents* is derived from the fully imputed 2011 ASM sample and is limited to the SU noncertainty establishments so that the overall ASM publication reliability requirements are maintained. The ratio estimators employ the sample-based values of annual payroll as an auxiliary variable. This variable is highly correlated with total shipments, but is subject to imputation. Note that we use the *complete* ASM sample (all MU and SU establishments) for the allocations but present the relative bias and MSE results for the subsampled domains (SU noncertainty establishments) only.

For the SU noncertainty establishments, the first NRFU attempt – consisting of a reminder letter – is historically very effective, so nonrespondent subsample selection occurs before the second NRFU attempt. The second NRFU attempt is generally more expensive (historically a package re-mail, although reminder letters via certified mail will be used in future collections). Nonrespondent subsampling of SU noncertainty establishments occurs *after* the second contact attempt (i.e., after the first NRFU attempt).

To perform the simulation, we removed all MU establishments and SU certainty establishments from the ASM sample data to create a frame, and then independently repeated the following procedure 5,000 times for each allocation procedure:

1. Using the estimated response propensities provided in Table 3.1, randomly induce nonresponse into the sample using a MAR response mechanism.

2. Sort the induced nonrespondents by sampling weight.
3. Select a stratified systematic sample using the nonrespondent domain subsampling rates for a given allocation strategy.
4. Simulate unit response for each round of NRFU. Table 3.1 provides the conditional response propensities used for each distinct NRFU contact phase. These statistics use paradata from the 2010 and 2011 ASM collections (Fink and Lineback, 2013). Hereafter, we refer to these conditional probabilities as "nonrespondent conversion rates". If the unit responded, the mode of response is randomly assigned using historical frequencies provided by subject matter experts. After assigning response status/response mode to each unit, compute cumulative collection cost, URR, and estimates.
5. For each allocation, repeat Step 4 until either ten rounds of follow-up have been conducted or the total budget has been expended. If funds are exhausted within a round, then NRFU ceases. Given that the fixed budget assumes that $1/K$ of the original set of nonrespondents will receive NRFU, the budget can be exhausted under full follow-up. The total budget is never expended before ten rounds of NRFU with nonrespondent subsampling, as the cost-per-unit of mailing a reminder letter is quite low. Our choice of a maximum of ten rounds of NRFU in the simulation was subjective; the purpose was to demonstrate that subsampling would facilitate additional contact efforts at no additional cost.

Table 3.1
Nonrespondent conversion rates for noncertainty single unit establishments by NRFU contact round used for simulation

Domain	Initial Response Probability	Nonrespondent Conversion Rates for a given Round of Nonresponse Follow-up									
		1	2	3	4	5	6	7	8	9	10
1	0.31	0.27	0.15	0.17	0.24	0.12	0.06	0.03	0.03	0.03	0.03
2	0.44	0.32	0.24	0.15	0.36	0.18	0.09	0.05	0.05	0.05	0.05
3	0.39	0.28	0.24	0.18	0.11	0.06	0.03	0.02	0.02	0.02	0.02
4	0.35	0.36	0.17	0.19	0.18	0.09	0.05	0.02	0.02	0.02	0.02
5	0.25	0.19	0.13	0.10	0.17	0.09	0.04	0.02	0.02	0.02	0.02
6	0.27	0.13	0.29	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02
7	0.44	0.34	0.23	0.20	0.25	0.13	0.06	0.03	0.03	0.03	0.03
8	0.38	0.45	0.12	0.33	0.25	0.13	0.06	0.03	0.03	0.03	0.03
9	0.39	0.30	0.23	0.13	0.25	0.13	0.06	0.03	0.03	0.03	0.03
10	0.75	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
11	0.28	0.23	0.12	0.18	0.15	0.07	0.04	0.02	0.02	0.02	0.02
12	0.36	0.30	0.21	0.15	0.31	0.16	0.08	0.04	0.04	0.04	0.04
13	0.39	0.22	0.19	0.13	0.23	0.12	0.06	0.03	0.03	0.03	0.03
14	0.37	0.36	0.16	0.06	0.45	0.22	0.11	0.06	0.06	0.06	0.06
15	0.41	0.32	0.22	0.19	0.26	0.13	0.06	0.03	0.03	0.03	0.03
16	0.40	0.34	0.22	0.23	0.32	0.16	0.08	0.04	0.04	0.04	0.04
17	0.34	0.26	0.18	0.10	0.21	0.11	0.05	0.03	0.03	0.03	0.03
18	0.40	0.31	0.18	0.10	0.18	0.09	0.04	0.02	0.02	0.02	0.02
19	0.37	0.29	0.20	0.19	0.23	0.11	0.06	0.03	0.03	0.03	0.03
20	0.40	0.28	0.21	0.15	0.18	0.09	0.04	0.02	0.02	0.02	0.02
21	0.36	0.27	0.20	0.14	0.23	0.11	0.06	0.03	0.03	0.03	0.03

The nonrespondent conversion rates in the majority of domains follow the same pattern: a decaying response probability followed by a slight increase in the fourth round due to a longer collection period. Domain 10 does not follow this pattern; it contained only four units that all responded before subsampling began. After the 4th round of NRFU, the nonrespondent conversion rates are reduced by half until they achieve the minimum allowable value of 0.02. The pattern reflects the findings of Olson and Groves (2012) (Olson and Groves (2012) postulate that the response propensities change over the collection cycle, especially as data collection protocols are modified. With the ASM, the reminder letters become more stringent at each NRFU contact phase. Likewise, the authors demonstrate that response propensities decline over the collection phase when a stable data collection protocol is used, as reflected in nonrespondent conversion rates). Mail and phone response propensity estimates were provided by subject matter experts, as were approximate costs by mode and an overall budget figure.

To evaluate the statistical properties of each allocation method for each estimator, we computed the relative bias and the mean squared error. The relative bias (RBE) for each estimate of total shipments at NRFU phase t for a given sampling overall interval (K), allocation method a (Constant$-K$, Min$-K$, Min$-$URR), eventual response probability q, and estimator e (DE, SR, CR) is

$$\text{RBE}\,(Y)^{e}_{Kaqt} = 100 * \left[\left(\frac{\sum_{s=1}^{5,000} \hat{Y}^{e}_{Kaqts}}{5,000} \Big/ Y\right) - 1\right]$$

where $\hat{Y}^{e}_{Kaqts}$ is the estimated total and Y is the population total shipments value.

The mean squared error at NRFU phase t for a given sampling interval, allocation method and estimator is

$$\text{MSE}\,(Y)^{e}_{Kaqt} = \left[\sum_{s=1}^{5,000} \left(\hat{Y}^{e}_{Kaqts} - Y\right)^2\right] \Big/ 5,000.$$

Since our simulation induces MAR response, the DE estimates should be approximately unbiased over repeated samples, whereas the two ratio estimates should not be. However, the DE estimates are expected to have large variance; using ratio estimators with a positively correlated auxiliary variable is expected to reduce this variance (i.e., increase the precision). Thus, examining the MSE provides insight into the bias-variance tradeoff.

3.2 Allocation

The simulation study uses data from the 2011 ASM collection. Input parameters for allocation were estimated from 2010 ASM collection data. Recall that the target URR applies to the entire ASM program and is not restricted to the subsampling domains - in our case, SU noncertainty establishments. Consequently, the certainty SU and MU unit counts obtained from the 2010 ASM data are included in the allocation programs in the r_{1h} as constants; the remainder of the r_{1h} represents the estimated count of responding SU noncertainty establishments after the first round of NRFU is completed. To ensure that each nonrespondent sampling domain contained sufficient numbers of units to obtain a feasible solution, we used three-digit industry as NRFU sampling domain instead of the six-digit industry used for the ASM sample design [Note: the determination subsampling domain was determined collaborative with the ASM program managers and methodologists].

Both quadratic programs require an estimated probability of eventually responding to follow-up (q_h) to compute the URR^{T} (overall and by domain). To assess the sensitivity of the allocation procedure, we tested ten different constant values $(q_h = 0.10, 0.20, \ldots, 1)$, keeping the value constant across all domains. A similar approach can be taken when historic paradata are not available. In addition, we estimate the q_h directly from the 2010 ASM data. These estimates vary by 20-percent at three-digit industry level. However, the median of these is nearly 50-percent. Consequently, the allocation obtained using the estimated (historic-data) q_h values are very similar to those obtained with $q = 0.50.$

Approximately $21,000 was allotted for NRFU of SU noncertainty establishments after subsampling. With full follow-up, the expected final unit response rate was approximately 79%. Using data from the 2007 EC, Bechtel and Thompson (2013) found that the target industry unit response rates of 70% could only be achieved in a $1 - \text{in} - 3$ subsample if the average unit response rate in the majority of EC industries was 60% or larger *before* follow-up begins. With the ASM, the response rate prior to subsampling was approximately 57%. Instead, we select an overall $1 - \text{in} - 2$ subsample, which would save approximately 50-percent of the allotted budget after five completed rounds of NRFU at the cost of a decrease expected response rate (69%). The additional five rounds of NRFU added approximately $4,000 to the total cost without commensurate increases in response rate (70%). A larger subsample would be preferable in terms of quality, but is not cost effective.

For allocation, we obtain the URR^{T}, allowing the q_h to vary by domain. The maximum URR is always achieved with the $\text{Min}-\text{URR}$ quadratic program. Table 3.2 presents the target URRs and the allocation subsampling rates obtained from the $\text{Min}-\text{URR}$ quadratic program. A dash (-) indicates no subsample is selected for NRFU (a sampling interval of ∞). If $K = 1$, all units in the domain are selected for NRFU (full follow-up). A label of $q =$ <value> indicates that the eventual probability of respondent is the same constant value in all domains; values estimated from historical data are labeled as $q_h =$ Est. Recall that URR^{T} includes all respondent units in the ASM sample, not just the noncertainty single units that are eligible for subsampling. Consequently, selected domains have achieved their target URRs *before* subsampling and are not considered as subsampling candidates in the allocation programs.

As the probability of eventually responding increases, this allocation tends to select smaller subsamples in increasing numbers of domains. When the probability of an eventual response (q_h) is small (20-percent or less), then the allocations sensibly tend towards no subsampling or full follow-up, focusing on obtaining sample from the few domains with the poorest response rates. As the probability of an eventual response increases, the amount of subsampling tends to increase as well. At 70-percent, almost half of the domains are allocated at least one sampled unit, thus spreading the allocated sample across several domains instead of concentrating in a few domains that have exceptionally poor response rates. Note that rates below 20-percent are (hopefully) unrealistic as are rates greater than 70-percent. Domain 10 has highly variable sampling rates regardless; because all four units responded before subsampling, the quadratic program selected *any* sampling rate because, in effect, it always subsamples zero cases.

Table 3.2
Min−URR Allocations (Sampling Intervals) (Program Level $K = 2$)

	Min−URR									
Domain	***q* = 10**	***q* = 20**	***q* = 30**	***q* = 40**	***q* = 50**	***q* = 60**	***q* = 70**	***q* = 80**	***q* = 90**	***q* = 100**	**q_h = Est**
1	-	-	-	-	-	-	-	81.63	9.23	5.40	-
2	-	-	-	-	-	-	3.88	2.26	1.71	1.44	-
3	-	-	9.32	3.40	2.58	2.19	1.98	1.86	1.77	1.71	2.12
4	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.06	1.13	1.00
5	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
6	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	-
7	-	-	-	-	-	-	-	14.95	9.26	7.10	-
8	-	-	-	-	-	-	-	-	-	-	-
9	1.00	1.00	1.22	1.44	1.61	1.76	1.89	2.01	2.12	2.22	1.62
10	1.03	30.26	30.37	30.26	30.46	29.90	30.51	29.04	1.00	10.03	10.04
11	-	-	-	-	-	-	-	-	-	-	-
12	-	-	-	-	-	-	-	-	-	-	-
13	-	-	-	-	5.00	2.94	2.29	2.01	1.88	1.78	2.91
14	-	-	-	-	-	-	-	-	-	-	-
15	7.86	4.45	3.22	2.57	2.42	2.32	2.28	2.30	2.35	2.40	2.38
16	1.00	1.35	1.46	1.46	1.49	1.51	1.53	1.57	1.62	1.66	1.66
17	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
18	-	-	-	-	-	-	-	-	-	37.95	-
19	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
20	1.00	1.00	1.00	1.16	1.34	1.49	1.63	1.75	1.87	1.97	1.35
21	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
URR^{T}	72.5%	72.9%	73.3%	73.7%	74.1%	74.4%	74.8%	75.2%	75.6%	76.0%	74.3%

Unlike the Min−URR quadratic program, the Min−K quadratic program did not always obtain a solution for a given target URR because of the domain-level constraints on the target URRs. When this occurred, we incrementally lowered the target response rate until a feasible solution could be obtained. Table 3.3 presents the target URRs and the allocations obtained from the Min−K quadratic program.

Both the allocation methods tend to designate the same domains for either no subsampling or for full follow-up. However, the two methods produce very different allocations for the *same* q_h in the *subsampled* domains. The Min−K allocations avoid subsampling in domains that have already achieved their maximum estimated target URR, regardless of the probability of eventually obtaining a response, with 40- to 50-percent of the domains not being subsampled when $0.30 \le q_h \le 0.50.$ Otherwise, the subsampling tends to be split between full follow-up (all units selected) or subsampling at an approximately $1 - \text{in} - 2$ sampling rate. In short, the Min−URR allocations yield domain subsampling intervals that can differ considerably from the overall interval, as the allocation seeks to equalize the target URR in each domain. The resultant variability in sampling intervals can lead to large increases in sampling variance. Because the Min−K objective function seeks to equalize sampling intervals, the domain subsampling intervals tend to be less variable and are generally close to the overall sampling interval.

Table 3.3
Min$-K$ Allocations (Sampling Intervals) (Program Level $K = 2$)

	Min$-K$ (Target $K = 2$)										
Domain	***q* = 10**	***q* = 20**	***q* = 30**	***q* = 40**	***q* = 50**	***q* = 60**	***q* = 70**	***q* = 80**	***q* = 90**	***q* = 100**	***q* = Est**
1	-	-	-	-	-	-	-	-	-	-	-
2	-	-	-	-	-	-	-	-	2.00	2.00	-
3	-	-	-	-	1.99	2.00	2.00	2.00	2.00	2.01	1.99
4	1.00	1.00	1.00	1.00	1.00	1.00	1.01	1.10	1.18	1.26	1.00
5	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
6	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
7	-	-	-	-	-	-	-	-	-	2.06	-
8	-	-	-	-	-	-	-	-	-	-	-
9	1.00	1.32	1.44	1.72	1.90	1.99	1.97	1.96	1.96	2.09	1.90
10	-	-	-	-	-	-	-	-	-	-	-
11	-	-	-	-	-	-	-	-	-	-	-
12	-	-	-	-	-	-	-	-	-	-	-
13	-	-	-	-	-	1.99	1.99	1.98	1.98	2.04	-
14	-	-	-	-	-	-	-	-	-	-	-
15	2.52	2.23	2.36	1.90	1.76	1.97	1.92	1.90	1.90	2.27	1.76
16	2.17	2.08	1.71	1.83	1.90	1.97	1.97	1.96	1.96	2.09	1.90
17	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
18	-	-	-	-	-	-	-	-	-	-	-
19	-	2.01	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
20	1.00	1.00	1.11	1.36	1.57	1.75	1.90	1.97	1.97	2.06	1.59
21	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
URR^T	71.0%	71.4%	72.3%	72.7%	73.1%	73.4%	73.8%	74.2%	74.6%	75.0%	73.3%

3.3 Results

Our baseline closely mimics the NRFU procedures used in the 2012 ASM NRFU – four phases of full follow-up $(K = 1) -$ but can include an additional incomplete fifth round when the planned budget was not depleted to retain programming consistency. For other values of K, NRFU is concluded after ten rounds regardless of the remaining funds.

Table 3.4 presents the relative bias of the estimates (RBE) and the mean squared error (MSE) results obtained with full NRFU and the Constant$-K$ allocation for each considered estimator. In all cases, the unbiased double expansion (DE) estimator yields unbiased estimates, whereas the ratio estimators are slightly biased as expected. With subsampling, the relative bias of the ratio estimators increases, whereas the DE estimator remains unbiased. Regardless of estimator, the additional stage of subsampling increases the sampling variance and consequently the MSE; the bias tends to remain unaffected because the subsampled units are a representative subsample at each round of follow-up.

With equal probability subsampling (Constant$-K$), a subsample may contain a few sampled cases in one or more domains. Although the subsampling weighting adjustment is not variable, the nonresponse adjustment factors can be quite large. The optimal allocations are designed to equalize response rates across domains, which can lead to occasionally "oversampling" in low-responding domains. Table 3.5 presents the RBE and the MSE for the Min$-$URR optimal allocations, using three different constant values of

$q\,(q = 0.30, 0.50, 0.70)$ and the domain specific rates estimated from historical data $(q_h = \text{Estimated})$. In all scenarios, the DE estimates are unbiased, the CR estimates are slightly biased, and the SR estimates are the most biased. This repeats the RBE pattern shown in the $\text{Constant}-K$ allocation results. Moreover, the RBE estimates do not appear to be overly sensitive to values of q_h used in allocation. Again, even with the additional rounds of NRFU, the bias of the subsamples' estimates is larger than that obtained with full follow-up of nonrespondents. In all cases, the MSE of the estimates obtained from the optimal allocations are smaller than those obtained with the $\text{Constant}-K$ allocations.

Regardless of estimator, the bias decreases when eventually probability of responding is low. This seems a bit counterintuitive but is in fact a direct consequence of the subsampling allocation procedure. When the probability of obtaining an eventual response is low, the $\text{Min}-\text{URR}$ allocation tends to subsample all or no units in a domain. With full follow-up, all responding units within the same domain have the same nonresponse adjustment. With a subsample, *only* the responding subsampled units' weights are adjusted for nonresponse and subsampling, in turn occasionally creating extremely variable weights within domain. As the probability of an eventual response increases, then the optimal allocation has sample in more domains, and finer adjustments are possible. With that said, the CR estimators tend to produce the lowest MSEs, regardless of allocation.

Table 3.4
Summary of relative bias in percent of the estimate and MSE for $\text{Constant}-K$ allocations in $x10^{12}$

Constant−*K* Relative Bias of the Estimate						
Percent	***K* = 1 (Full)**			***K* = 2**		
Contact	**DE**	**CR**	**SR**	**DE**	**CR**	**SR**
2	0.01%	0.03%	0.10%	0.00%	0.51%	1.43%
3	0.00%	0.03%	0.08%	-0.01%	0.29%	0.77%
4	0.00%	0.01%	0.06%	-0.02%	0.14%	0.40%
5	0.01%	0.02%	0.04%	-0.01%	0.12%	0.32%
6				-0.01%	0.11%	0.29%
7				0.00%	0.11%	0.28%
8				0.00%	0.11%	0.27%
9				0.00%	0.10%	0.25%
10				0.00%	0.10%	0.25%
Constant−*K* Mean Squared Error						
x10^12	***K* = 1 (Full)**			***K* = 2**		
Contact	**DE**	**CR**	**SR**	**DE**	**CR**	**SR**
2	4.96	2.60	5.56	37.53	26.34	70.49
3	3.67	1.96	4.17	19.82	13.80	28.88
4	2.55	1.39	3.03	11.75	8.30	14.87
5	2.48	1.39	2.87	9.94	7.10	12.12
6				9.36	6.75	11.16
7				9.09	6.63	10.63
8				8.80	6.48	10.23
9				8.51	6.32	9.95
10				8.27	6.19	9.74

Table 3.5
Summary of relative bias of the estimate and MSE for $\text{Min}-\text{URR}$ optimal allocations

Min URR RBE (Target $K = 2$)												
Percent	***q* = 0.30**			***q* = 0.50**			***q* = 0.70**			***q* = Estimated**		
Contact	**DE**	**CR**	**SR**	**DE**	**CR**	**SR**	**DE**	**CR**	**SR**	**DE**	**CR**	**SR**
2	-0.01%	0.06%	0.20%	0.01%	0.07%	0.36%	0.01%	0.08%	0.31%	-0.01%	0.08%	0.32%
3	0.00%	0.05%	0.16%	0.01%	0.05%	0.26%	0.01%	0.07%	0.23%	0.01%	0.07%	0.23%
4	0.00%	0.05%	0.14%	0.01%	0.05%	0.18%	0.01%	0.06%	0.19%	0.01%	0.06%	0.17%
5	0.00%	0.05%	0.14%	0.01%	0.05%	0.18%	0.01%	0.05%	0.17%	0.00%	0.06%	0.15%
6	0.01%	0.05%	0.13%	0.02%	0.05%	0.17%	0.01%	0.05%	0.17%	0.00%	0.06%	0.15%
7	0.01%	0.05%	0.13%	0.01%	0.05%	0.17%	0.01%	0.05%	0.16%	0.00%	0.06%	0.15%
8	0.01%	0.05%	0.13%	0.01%	0.05%	0.16%	0.01%	0.05%	0.16%	0.00%	0.05%	0.14%
9	0.01%	0.05%	0.13%	0.01%	0.05%	0.16%	0.01%	0.05%	0.16%	0.00%	0.05%	0.14%
10	0.00%	0.05%	0.13%	0.01%	0.05%	0.15%	0.01%	0.05%	0.15%	0.00%	0.05%	0.14%
Min URR MSE (Target $K = 2$)												
x10^12	***q* = 0.30**			***q* = 0.50**			***q* = 0.70**			***q* = Estimated**		
Contact	**DE**	**CR**	**SR**	**DE**	**CR**	**SR**	**DE**	**CR**	**SR**	**DE**	**CR**	**SR**
2	12.55	6.77	16.03	14.35	7.48	17.60	14.31	7.44	17.15	14.39	7.79	17.05
3	8.88	5.13	10.87	9.80	5.43	11.32	9.57	5.42	11.36	9.75	5.47	10.98
4	7.00	4.28	8.15	7.61	4.45	8.28	7.31	4.43	8.44	7.53	4.44	8.05
5	6.61	4.07	7.41	7.02	4.17	7.44	6.80	4.13	7.60	6.94	4.14	7.42
6	6.45	3.97	7.16	6.78	4.09	7.18	6.62	4.03	7.25	6.75	4.05	7.15
7	6.37	3.92	7.05	6.68	4.06	7.08	6.55	3.97	7.07	6.67	4.02	7.03
8	6.34	3.90	6.94	6.57	4.01	6.97	6.45	3.93	6.95	6.57	3.98	6.93
9	6.28	3.87	6.86	6.50	3.98	6.89	6.39	3.90	6.86	6.47	3.94	6.84
10	6.23	3.85	6.78	6.40	3.91	6.76	6.35	3.87	6.75	6.42	3.89	6.73

The $\text{Min}-K$ allocation procedure is designed to reduce the variability in the subsampled units' adjustment weights. Table 3.6 presents the relative bias of the estimate and MSE for the $\text{Min}-K$ optimal allocation method. The $\text{Min}-K$ estimators display the same pattern as before. The DE estimates are unbiased, the CR estimates are nearly unbiased and the SR estimates are slightly biased.

The MSE estimates for the $\text{Min}-K$ method follow a similar pattern as the $\text{Min}-\text{URR}$ method, as expected due to the similarities between corresponding $\text{Min}-\text{URR}$ and $\text{Min}-K$ allocations. These results appear to be relatively insensitive to assumed eventual probability of response (q). The historical-data estimated conversion rates produce similar results to an assumed $q =$ 0.50. In many cases, the $\text{Min}-\text{URR}$ method produces the least biased estimates. However, bias is only a single component of the MSE, and the $\text{Min}-\text{URR}$ allocations tend to have smaller expected number of respondents in several strata than their $\text{Min}-K$ counterparts. Moreover, the $\text{Min}-K$ allocations have smaller sampling variances by design, ultimately yielding estimates with lower MSEs than their $\text{Min}-\text{URR}$ counterparts.

Figures 3.1 and 3.2 plot the RBEs and MSEs obtained at each round of NRFU for the CR estimator (our "best" estimator) using the q_h obtained from historical data for each of the considered optimal allocation methods along with the benchmark values (labeled as "Full Follow-up"). In Figure 3.1, the benchmark estimates are the least biased. However, this extremely low bias is in part a consequence of our nonresponse model, which is uniform within domain and NRFU phase. Neither of the optimal allocation estimates attained the benchmark estimate levels, but they become very close after seven rounds of NRFU and the

RBEs of the Min−URR and Min−K CR estimates are less than *one tenth* of one percent (0.06% and 0.05% respectively). In summary, subsampling with either optimal allocation strategy yielded trivial biases increases over full follow-up.

Table 3.6
Summary of relative bias of the estimate and MSE for Min−K optimal allocations

Min−K RBE (Target K = 2)												
Percent	***q* = 0.30**			***q* = 0.50**			***q* = 0.70**			***q* = Estimated**		
Contact	**DE**	**CR**	**SR**	**DE**	**CR**	**SR**	**DE**	**CR**	**SR**	**DE**	**CR**	**SR**
2	0.03%	0.08%	0.24%	0.03%	0.09%	0.31%	0.00%	0.08%	0.33%	0.01%	0.07%	0.30%
3	0.03%	0.05%	0.20%	0.03%	0.08%	0.22%	0.00%	0.05%	0.22%	0.01%	0.06%	0.21%
4	0.02%	0.04%	0.16%	0.03%	0.07%	0.18%	0.00%	0.05%	0.17%	0.01%	0.05%	0.17%
5	0.02%	0.04%	0.15%	0.03%	0.06%	0.17%	0.01%	0.05%	0.16%	0.01%	0.05%	0.16%
6	0.02%	0.05%	0.14%	0.02%	0.06%	0.16%	0.01%	0.05%	0.15%	0.00%	0.05%	0.15%
7	0.02%	0.05%	0.14%	0.02%	0.05%	0.16%	0.01%	0.05%	0.15%	0.01%	0.05%	0.15%
8	0.02%	0.05%	0.14%	0.02%	0.05%	0.16%	0.01%	0.05%	0.15%	0.01%	0.04%	0.14%
9	0.02%	0.05%	0.14%	0.02%	0.05%	0.15%	0.01%	0.05%	0.14%	0.01%	0.04%	0.14%
10	0.02%	0.05%	0.14%	0.02%	0.05%	0.16%	0.01%	0.05%	0.15%	0.01%	0.04%	0.14%

Min−K MSE (Target K = 2)												
x10^12	***q* = 0.30**			***q* = 0.50**			***q* = 0.70**			***q* = Estimated**		
Contact	**DE**	**CR**	**SR**	**DE**	**CR**	**SR**	**DE**	**CR**	**SR**	**DE**	**CR**	**SR**
2	12.86	7.19	15.85	13.81	7.42	16.80	15.09	8.34	18.00	13.43	7.19	16.07
3	8.74	5.04	10.26	9.32	5.38	10.82	10.45	5.89	11.38	9.25	5.30	10.69
4	6.92	4.07	7.65	7.26	4.26	7.92	7.84	4.60	8.19	7.22	4.33	7.93
5	6.50	3.85	7.07	6.77	4.05	7.33	7.23	4.28	7.47	6.65	4.06	7.21
6	6.32	3.80	6.80	6.57	3.94	7.02	7.02	4.19	7.28	6.45	3.95	6.91
7	6.23	3.76	6.69	6.49	3.88	6.91	6.90	4.15	7.16	6.31	3.91	6.78
8	6.21	3.73	6.61	6.39	3.84	6.82	6.78	4.10	7.06	6.23	3.87	6.68
9	6.16	3.70	6.54	6.35	3.79	6.71	6.68	4.05	6.93	6.15	3.83	6.57
10	6.10	3.66	6.43	6.24	3.74	6.62	6.60	3.98	6.87	6.11	3.80	6.48

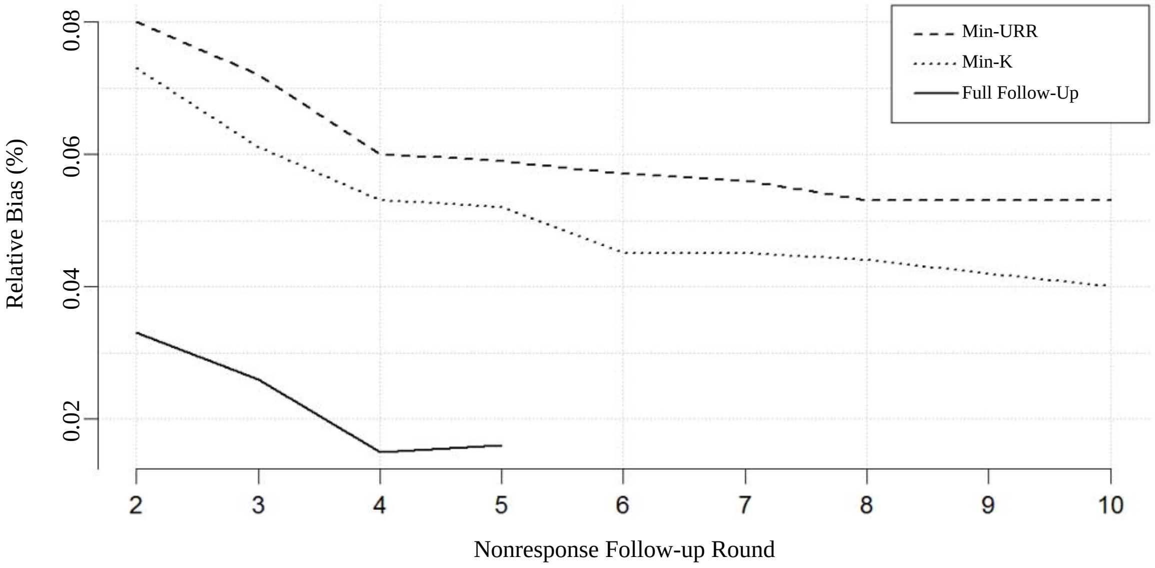


Figure 3.1 Relative bias of the estimates (Historic q_h) for the CR estimator.

Figure 3.2 plots MSE values by NRFU round using the CR estimator. The targeted nonresponse sampling strategy used for the $\text{Min}-K$ allocation appears to reduce the overall error. We believe that this is due to two factors. First, the $\text{Min}-K$ allocation procedure samples larger proportions of nonrespondents in low responding areas than obtained with the $\text{Min}-\text{URR}$ allocations. Second, the quadratic formula for the $\text{Min}-K$ allocation includes a constraint on the domain response rates, lowering the overall target response but reducing the variability in the proportion of respondents by domain. Ultimately, this approach yields similar response rates across sampling domains, indicative of a representative sample (Wagner, 2012; Schouten, Cobben and Bethlehem, 2009). Note that the increased MSE is not trivial with nonrespondent subsampling, even when using an adjustment procedure that benefits from a strong covariate in the ratio adjustment procedure. This is an acknowledged price paid for nonrespondent subsampling (Biemer, 2010). However, this additional variance component is measurable. If the measured component is too large, the program managers can subsample less (use a larger K).

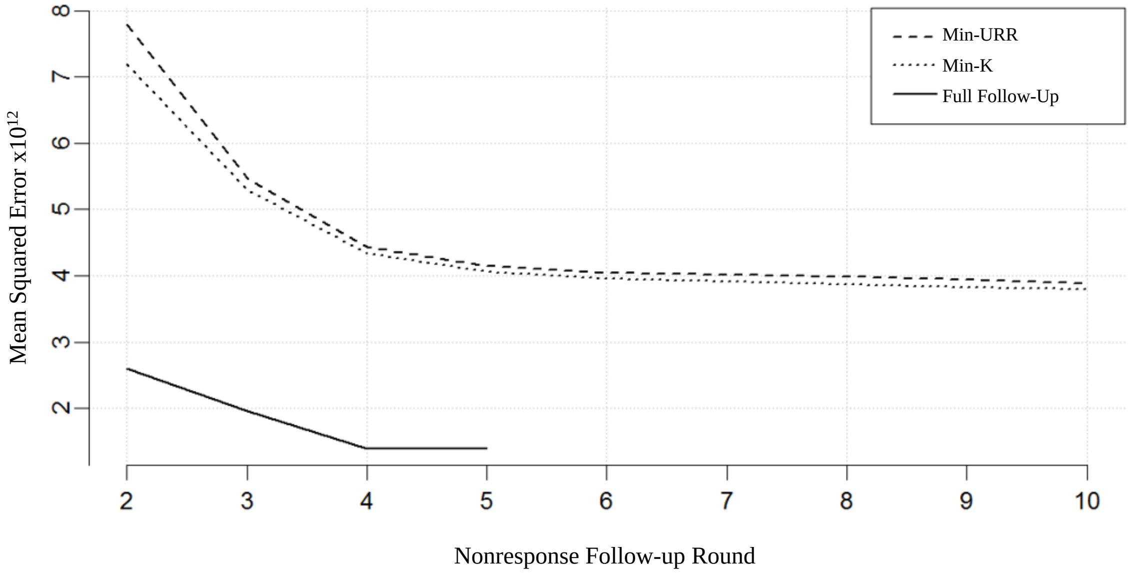


Figure 3.2 Mean squared error (Historic q_h) for the CR estimator.

3.4 Discussion

Given a sophisticated allocation method, a ratio estimator employing a highly correlated auxiliary variable, and a fairly large subsample, this case study shows that nonrespondent subsampling does not overly penalize quality to save cost. The additional stage of sampling increased the MSE for the studied variable, but the level was reduced by the judicious choice of estimator. Of course, we consider only one variable in our simulation, and this variable may or may not “behave” similarly to other survey items. One referee suggested the usage of an R-indicator (Schouten et al., 2009) or balance indicator (Särndal and Lundquist, 2014) to assess the overall representativeness of the respondent sets in a field survey setting. This might be useful at later stages of data collection (after nonrespondent subsampling and during NRFU), but would not provide any further insight into the degree of bias reduction on any collected item, as we can do in this simulation setting.

Of the three considered allocation methods, the $\text{Constant}-K$ method had the worst performance, often selecting a very small probability subsample when not needed and consequently increasing the sampling variance without reducing the bias. Of the three considered allocation methods, the $\text{Min}-K$ allocation was the most effective in realizing acceptable response rates and achieving reliable estimates; the larger bias caused by the varying domain sampling intervals is generally offset by the reduced sampling variance. However, implementation of the $\text{Min}-K$ allocation can be more challenging than the $\text{Min}-\text{URR}$.

For both optimal allocation procedures, we tested four different eventual probabilities of response to assess the sensitivity of the allocation procedures to these inputs. By comparing allocations obtained with a constant assumed input value to those obtained using the empirical estimates, we found that the realized allocations could over- or under- sample in selected domain, and the domain response rates could vary more than expected when the actual (survey) values are quite different from the input values. Consequently, we recommend using values estimated from historic paradata whenever possible.

If reducing cost is the overall goal, then we note that additional NRFU contact attempts beyond the fifth contact did not improve the bias or MSE of the subsampled estimates in our case study. Of course, if the achieved cost reduction for a $1-\text{in}-2$ subsample with up to ten NRFU contact attempts is acceptable, the funds allocated to these final contact attempts might be better expended earlier in the collection cycles using other contact strategies.

4 Conclusion

In general, the NRFU procedures for economic programs conducted by the U.S. Census Bureau follow a calendar schedule. Budget is tied to the fiscal year, and contact strategies are budgeted accordingly. Since economic populations are highly skewed and the statistics of interest are totals, a large fraction of the NRFU budget is allocated to the larger units. The smaller units are believed to be homogeneous – at least in size. However, it is difficult to validate that belief in the absence of collected respondent data. Given that the NRFU procedures rely on obtaining response data from the larger units, the response rates from smaller units tend to be much lower. It is quite likely that the realized respondent set is neither "balanced…which means (the selected sample has) the same or almost the same characteristics as the whole population" for selected items (Särndal, 2011) nor "representative… with respect to the sample if the response propensities ρ_i are the same for all units in the population" (Schouten et al., 2009). The emphasis on obtaining responses from the larger units at the cost of the lower unit response in turn creates a bias in the estimates, as imputed or adjusted values for smaller units resemble the large unit values (Thompson and Washington, 2013).

By limiting the target domain for nonrespondent subsampling to the smaller units, we can reduce this unmeasurable bias. Our allocation method increases the potential of obtaining a balanced and representative sample by targeting the low responding areas that usually would not receive any special treatment. It can be implemented at any stage of the data collection process and with any sample design, making it quite flexible although not necessarily optimal for specific sample designs and estimators. It is a "safe" approach for a multi-purpose survey, presumably designed to obtain reliable estimates for a variety of items. Moreover, selecting a systematic subsample from a list sorted by a unit measure of size avoids incidence of additional nonresponse bias incurred by focusing NRFU efforts on high response propensity cases (Tourangeau et al., 2016; Beaumont et al., 2014). We acknowledge that the increased variability in design weights and reduction in response rates are less than desirable effects caused by subsampling. However, these effects can be lessened via the choice of estimator, as demonstrated by our improved results with a ratio estimator. More sophisticated calibration estimators or other collapsed estimators could likewise be considered at the estimation stage.

Without probability subsampling, the contention that the realized respondent set of small businesses remains a probability sample is debatable. Several discussions of the summary report of the AAPOR Task Force on non-probability sampling (Baker, Brick, Bates, Battaglia, Couper, Dever, Gile and Tourangeau, 2013) specifically question whether “a probability sample with less than full coverage and high nonresponse should still be considered a probability sample”. That question is certainly relevant in our studied context, where sampled smaller units truly “opt in” to respond. Selecting a probability subsample of nonrespondents and instructing survey analysts to limit NRFU contact to these cases may limit this phenomenon. In addition, with a probability subsample, one can use accepted quality measures such as sampling error or response rates for evaluation.

All of the results presented for our case study assume that the existing NRFU contact strategies are used with the subsampled designs. However, subsampling nonrespondents without changing the data collection procedure may have minimal tangible benefits besides cost reduction. The reverse is also true: for example, Kirgis and Lepkowski (2013) present improved response data results for targeted small domains obtained with probability samples and revised contact strategies.

Tourangeau et al. (2016) note that “it is not always clear how to intervene to obtain cases, particularly cases with low underlying propensities, to respond”. This is especially relevant in the business survey context. Business surveys can draw on a wealth of cognitive research on data collection strategies for large companies: see Paxson, Dillman and Tarnai, 1995; Tuttle, Morrison and Willimack, 2010; Willimack and Nichols, 2010; Snijkers, Haraldsen, Jones and Willimack, 2013. In contrast, the smaller businesses receive very little personal contact (if any) and there is limited cognitive research on preferable contact strategies to draw upon. That said, the literature suggests that there are differences in collected data quality between large and small businesses: see Thompson and Washington (2013), Willimack and Nichols (2010), Bavdaž (2010), Torres van Grinsven, Bolko and Bavdaž (2014), and Thompson, Oliver and Beck (2015). Additional cognitive research for small establishments combined with field tests could yield better contact strategies. Subsampling nonrespondents paired with a new contact strategy for these “hard to reach” establishments would create a truly adaptive approach for all units, not just the larger ones. To this point, in response to these presented analyses, the Census Bureau conducted an embedded field experiment to test alternative NRFU strategies for selected small units in the 2014 ASM (Thompson and Kaputa, 2017). The outcome of that study was a new NRFU protocol implemented in the 2015 ASM and a second embedded field experiment that paired our proposed nonrespondent subsampling design with the most effective follow-up procedures determined from the 2014 test (Kaputa, Thompson and Beck, 2017).

Acknowledgements

Any views expressed are those of the author(s) and not necessarily those of the U.S. Census Bureau. The authors thank Eric Fink, Xijian Liu, Jared Martin, Edward Watkins III, Hannah Thaw, the Associate Editor, and two referees for their review of an earlier version of the manuscript, David Haziza for his thoughtful discussion of the paper, and Barry Schouten for his useful suggestions on the optimization problems.

Appendix

Our objective is to estimate Y, population total of characteristic y, from the realized sample of respondents. Let

S_{hi} = 1 if unit i in domain h was in original sample; 0 otherwise.

θ_{hi} = the probability of sampling unit i in domain h into the original sample $(w_{hi} = 1/\theta_{hi})$.

R_{hi} = 1 if unit i in domain h provided a response before subsampling time t (value for y); 0 otherwise.

I_{hi} = 1 if unit i in domain h was selected for NRFU (i.e., was a subsampled nonrespondent); 0 otherwise.

J_{hi} = 1 if unit i in domain h responds, given selected into nonrespondent subsample; 0 otherwise.

f_{hi} = adjustment factor for nonrespondent subsampling and unit nonresponse after NRFU.

y_{hi} = value of characteristic y for unit i in domain h, available only for respondents.

x_{hi} = value of characteristic x for unit i in domain h, available for all sampled units considered for nonrespondent subsampling (i.e., the nonrespondent subsampling frame). Then $\hat{Y} = \sum_h \sum_i w_{hi} y_{hi} S_{hi} R_{hi} + \sum_h \sum_i w_{hi} f_{hi} y_{hi} S_{hi} (1 - R_{hi}) I_{hi} J_{hi} = \hat{Y}_{R1} + \hat{Y}_{R2}$.

We consider three different adjustment-to-sample reweighting estimators of $\hat{Y}_{R2}$:

Double Expansion (DE): $$\hat{Y}_{R2}^{\text{DE}} = \sum_h \sum_{i \in h} w_{hi} K_h \left(\frac{m_{1h}}{r_{2h}}\right) y_{hi} S_{hi} (1 - R_{hi}) I_{hi} J_{hi}$$

Separate Ratio (SR): $$\hat{Y}_{R2}^{\text{SR}} = \sum_h \sum_{i \in h} w_{hi} K_h \left(\frac{\sum_{i \in m_{1h}} x_{hi}}{\sum_{i \in r_{2h}} x_{hi}}\right) y_{hi} S_{hi} (1 - R_{hi}) I_{hi} J_{hi}$$

Combined Ratio (CR): $$\hat{Y}_{R2}^{\text{CR}} = \sum_h \sum_{i \in h} w_{hi} K_h \left(\frac{m_{1h}}{r_{2h}}\right) \left(\frac{\sum_{i \in m_{1h}} w_{hi} K_h x_{hi}}{\sum_{i \in r_{2h}} w_{hi} K_h \left(\frac{m_{1h}}{r_{2h}}\right) x_{hi}}\right) y_{hi} S_{hi} (1 - R_{hi}) I_{hi} J_{hi}.$$

Note that the DE and CR estimators are variations of the recommended reweighting procedure described in Brick (2013) and are discussed in Binder et al. (2000) among others. The DE estimator is the InfoS estimator presented in Särndal and Lundström (2005), studied in Shao and Thompson (2009), among others;

the SR estimator is a variation of the InfoP estimator presented in Särndal and Lundström (2005), treating the realized sample as the "population". Sampling weights were not included in the SR so that the adjustment reduces to the DE adjustment when $x_{hi} \equiv 1 \forall i \in h$; note that this unweighted response rate adjustment is recommended in Little and Vartivarian (2005). The CR estimator is presented in Binder et al. (2000), and is also studied in Shao and Thompson (2009). In our case study, a better choice might have been the quasi-randomization estimator from Oh and Scheuren (1983), which incorporates sampling weights in the adjustment factor, thus reducing their variability.

Collapsed estimators are used in three scenarios: (1) All units in the domain receive NRFU (no subsampling); (2) No units in the domain receive NRFU because response rate targets have been achieved (no subsampling); and (3) A single subsampled unit responded to NRFU (subsampling). The collapsed estimators analogues are given as follows:

Collapsed DE: $$\hat{Y}_h^{\mathrm{DE,C}} = \sum_{i \in h} w_{hi} \left(\frac{n_h}{r_{1h} + r_{2h}} \right) y_{hi} S_{hi} R_{hi}$$

Collapsed SR: $$\hat{Y}_h^{\mathrm{SR,C}} = \sum_{i \in h} w_{hi} \left(\frac{\sum_{i \in n_h} x_{hi}}{\sum_{i \in r_{1h} + r_{2h}} x_{hi}} \right) y_{hi} S_{hi} \, (1 - R_{hi}) \, I_{hi} J_{hi}$$

Collapsed CR: $$\hat{Y}_h^{\mathrm{CR,C}} = \sum_{i \in h} w_{hi} \left(\frac{n_h}{r_{1h} + r_{2h}} \right) \left(\frac{\sum_{i \in n_h} w_{hi} x_{hi}}{\sum_{i \in r_{1h} + r_{2h}} w_{hi} \left(\frac{n_h}{r_{1h} + r_{2h}} \right) x_{hi}} \right) y_{hi} S_{hi} R_{hi} \, .$$

References

Baker, R., Brick, J.M., Bates, N., Battaglia, M., Couper, M., Dever, J., Gile, K. and Tourangeau, R. (2013). Summary report of the AAPOR task force on non-probability sampling – Report and rejoinder. *Journal of Survey Statistics and Methodology*, 1, 90-137.

Bavdaž, M. (2010). The multidimensional integral business survey response model. *Survey Methodology*, 36, 1, 81- 93. Paper available at https://www150.statcan.gc.ca/n1/pub/12-001-x/2010001/article/11245-eng.pdf.

Beaumont, J.-F., Bocci, C. and Haziza, D. (2014). An adaptive data collection procedure for call prioritization. *Journal of Official Statistics*, 30(4), 607-621.

Bechtel, L., and Thompson, K.J. (2013). Optimizing unit nonresponse adjustment procedures after subsampling nonrespondents in the Economic Census. *Proceedings of the Federal Committee on Statistical Methods Research Conference*, https://nces.ed.gov/FCSM/index.asp.

Biemer, P. (2010). Total survey error: Design, implementation, and evaluation. *The Public Opinion Quarterly*, 74(5), 817-848.

Binder, D., Babyak, C., Brodeur, M., Hidiroglou, M. and Wisner, J. (2000). Variance estimation for two-phase stratified sampling. *The Canadian Journal of Statistics*, 28, 751-764.

Brick, J.M. (2013). Unit nonresponse and weighting adjustments: A critical review. *Journal of Official Statistics*, 29, 329-353.

Federal Register Notice (2006). OMB Standards and Guidelines for Statistical Surveys, Washington, DC.

Fink, E., and Lineback, J.F. (2013). Using paradata to understand business survey reporting patterns. *Proceedings of the Federal Committee on Statistical Methods Research Conference*, https://nces.ed.gov/FCSM/index.asp.

Groves, R., and Herringa, S. (2006). Responsive design for household surveys: Tools for actively controlling survey errors and costs. *Journal of the Royal Statistical Society Series A*, 169(3), 439-57.

Hansen, M.H., and Hurwitz, W.N. (1946). The problem of non-response in sample surveys. *Journal of the American Statistical Association*, 41, 517-529.

Harter, R.M., Mach, T.L., Chaplin, J.F. and Wolken, J.D. (2007). Determining subsampling rates for nonrespondents. *Proceedings of the Third International Conference on Establishment Surveys*, American Statistical Association.

Haziza, D., Thompson, K.J. and Yung, W. (2010). The effect of nonresponse adjustments on variance estimation. *Survey Methodology*, 36, 1, 35-43. Paper available at https://www150.statcan.gc.ca/n1/pub/12-001-x/2010001/article/11246-eng.pdf.

Kalton, G., and Flores-Cervantes, I. (2003). Weighting methods. *Journal of Official Statistics*, 19, 2, 81-97.

Kaputa, S., Thompson, K.J. and Beck, J. (2017). An embedded experiment for targeted nonresponse follow-up in establishment surveys. *Proceedings of the Section on Survey Research Methods*, American Statistical Association.

Kirgis, N., and Lepkowski, J. (2013). Design and management strategies for paradata-driven responsive design: Illustrations for the 2006-2010 National Survey of Family Growth. *Improving Surveys with Paradata*, (Ed., Frauke Kreuter). Hoboken, NJ: John Wiley & Sons, Inc.

Kish, L. (1992). Weighting for unequal P_i. *Journal of Official Statistics*, 8(2), 183-200.

Kott, P. (1994). A note on handling nonresponse in sample surveys. *Journal of the American Statistical Association*, 89, 693-696.

Little, R.J., and Vartivarian, S. (2005). Does weighting for nonresponse increase the variance of survey means? *Survey Methodology*, 31, 2, 161-168. Paper available at https://www150.statcan.gc.ca/n1/pub/12-001-x/2005002/article/9046-eng.pdf.

Lohr, S.L. (2010). *Sampling: Design and Analysis*. Boston: Brooks/Cole.

Oh, H.L., and Scheuren, F.J. (1983). Weighting adjustment of unit nonresponse. *Incomplete Data in Sample Surveys*. New York: Academic Press, 20, 143-184.

Olson, K., and Groves, R.M. (2012). An examination of within-person variation in response propensity over the data collection field period. *Journal of Official Statistics*, 28, 29-51.

Paxson, M.C., Dillman, D.A. and Tarnai, J. (1995). Improving response to business mail surveys. In *Business Survey Methods*, (Eds., B.G. Cox, D. Binder, B. Nanajamma Chinnappa, M. Colledge and P. Kott). New York: John Wiley & Sons, Inc.

Särndal, C.-E. (2011). The 2010 Morris Hansen lecture: Dealing with survey nonresponse in data collection, in estimation. *Journal of Official Statistics*, 27, 1-21.

Särndal, C., and Lundquist, P. (2014). Accuracy in estimation with nonresponse: A function of degree of imbalance and degree of explanation. *Journal of Survey Statistics and Methodology*, 2(4), 361-387.

Särndal, C.-E., and Lundström, S. (2005). *Estimation in Surveys with Nonresponse*. Hoboken, NJ: John Wiley & Sons, Inc.

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

Schouten, B., Calinescu, M. and Luiten, A. (2013). Optimizing quality of response through adaptive survey designs. *Survey Methodology*, 39, 1, 29-58. Paper available at https://www150.statcan.gc.ca/n1/pub/12-001-x/2013001/article/11824-eng.pdf.

Schouten, B., Cobben, F. and Bethlehem, J. (2009). Indicators for the representativeness of survey response. *Survey Methodology*, 35, 1, 101-113. Paper available at https://www150.statcan.gc.ca/n1/pub/12-001-x/2009001/article/10887-eng.pdf.

Shao, J., and Thompson, K.J. (2009). Variance estimation in the presence of nonrespondents and certainty strata. *Survey Methodology*, 35, 2, 215-225. Paper available at https://www150.statcan.gc.ca/n1/pub/12-001-x/2009002/article/11043-eng.pdf.

Snijkers, G., Haraldsen, G., Jones, J. and Willimack, D.K. (2013). *Designing and Conducting Business Surveys*. Hoboken, NJ: John Wiley & Sons, Inc.

Thompson, K.J., and Kaputa, S. (2017). Investigating adaptive nonresponse follow-up strategies for small businesses through embedded experiments. *Journal of Official Statistics*, 33(3), 1-23.

Thompson, K.J., and Oliver, B. (2012). Response rates in business surveys: Going beyond the usual performance measure. *Journal of Official Statistics*, 27, 221-237.

Thompson, K.J., Oliver, B. and Beck, J. (2015). An analysis of the mixed collection modes for two business surveys conducted by the US Census Bureau. *Public Opinion Quarterly*, 79(3), 769-789.

Thompson, K.J., and Washington, K.T. (2013). Challenges in the treatment of unit nonresponse for selected business surveys: A case study. *Survey Methods: Insights from the Field*. Retrieved from http://surveyinsights.org/?p=2991.

Torres van Grinsven, V., Bolko, I. and Bavdaž, M. (2014). In search of motivation for the business survey response task. *Journal of Official Statistics*, 30(4), 579-606.

Tourangeau, R., Brick, J.M., Lohr, S. and Li, J. (2016). Adaptive and responsive survey designs: A review and assessment. *Journal of the Royal Statistical Society A*, 180, 203-223.

Tuttle, A., Morrison, R. and Willimack, D. (2010). From start to pilot: A multi-method approach to the comprehensive redesign of an economic survey questionnaire. *Journal of Official Statistics*, 26, 87-103.

Wagner, J. (2012). Research synthesis: A comparison of alternative indicators for the risk of nonresponse bias. *Public Opinion Quarterly*, 76(3), 555-575.

Willimack, D., and Nichols, E. (2010). A hybrid response process model for business surveys. *Journal of Official Statistics*, 26, 3-24.

Zhang, L.C. (2008). On some common practices of systematic sampling. *Journal of Official Statistics*, 24, 557-569.

Survey Methodology, June 2018
Vol. 44, No. 1, pp. 101-115
Statistics Canada, Catalogue No. 12-001-X

Robust Bayesian small area estimation

Malay Ghosh, Jiyoun Myung and Fernando A.S. Moura[1]

Abstract

Small area models handling area level data typically assume normality of random effects. This assumption does not always work. The present paper introduces a new small area model with *t* random effects. Along with this, this paper also considers joint modeling of small area means and variances. The present approach is shown to perform better than other methods.

Key Words: Random effects model; Student's *t*-distribution; Non-subjective priors; MCMC; Gibbs sampling; Metropolis-Hastings algorithm.

1 Introduction

The classic paper of Fay and Herriot (1979) has become a cornerstone of research in small area estimation for nearly four decades. The Fay-Herriot model is a random effects model with a normality assumption for both the random effects and the errors. Moreover, the error variances are assumed to be known. The latter is almost imperative due to an identifiability issue. With availability of only the area level direct small area estimates plus nonavailability of microdata, any effective modeling of the error variances is near impossible.

Some valiant remedial attempts were made by W.R. Bell and his colleagues at the US Census Bureau (Bell and Huang, 2006; Bell, 2008) for handling some census data, but questions remain regarding the universal application of their approach. Additionally, nonavailability of microdata for secondary survey users is primarily due to confidentiality reasons, especially from the Federal Agencies. If microdata becomes available, unit level models are more appropriate than area level models. A classic example is the well-cited article of Battese, Harter and Fuller (1988). However, area level models are widely used due to their simplicity of implementation in a complex survey setting when compared to unit level models.

As the field developed and more data started getting analyzed, researchers found the inappropriateness of the assumption of normality as well as that of known error variances. As mentioned in the previous paragraph, the latter is hard to rectify without any extra information. One of the first attempts in this regard is due to Lahiri and Rao (1995) who replaced the normality assumption of random effects by the finiteness of their eighth moments. Datta and Lahiri (1995) considered a general mixture of normal distributions for random effects that includes the $t-$ distribution. There are papers, dispensing fully with the normality, but maintaining linearity of the model, and using ANOVA estimators of the variances. One may refer to Butar and Lahiri (2002) and Jiang, Lahiri and Wan (2002) who calculated the corresponding uncertainty measures either via jackknife or bootstrap. Bell and Huang (2006) used $t-$ distributions for random effects or sampling errors to diminish the effects of outliers.

[1] Malay Ghosh, Department of Statistics, University of Florida, 102 Griffin-Floyd Hall, P.O. Box 118545, Gainesville, Florida, U.S.A., 32611. E-mail: ghoshm@stat.ufl.edu; Jiyoun Myung, Department of Statistics and Applied Probability, University of California, Santa Barbara, 5513 South Hall, Santa Barbara, California, U.S.A., 93106. E-mail: myung@pstat.ucsb.edu; Fernando A.S. Moura, Instituto de Matemática, Universidade Federal do Rio de Janeiro, Caixa Postal 68530, CEP: 21941-909, RJ, Brazil. E-mail: fmoura@im.ufrj.br.

The objective of the present article is to address these two important issues in the context of small area estimation. First, we consider small area modeling of both the population means and population variances. This is possible due to the availability of additional data purported to estimate the error variances. Second, in order to induce some robustness of our procedure, we consider $t-$ priors for the random effects.

The data set considered in this paper came from a test demographic census carried out in one municipality in Brazil consisting of 140 enumeration districts, hereafter referred to as small areas. The response variable was the average income of the heads of households for each small area, and the goal was to make predictions for the 140 population means of the heads of household's income. The auxiliary variables were the respective small area population means of the educational attainment of the heads of households, and the respective population means of the number of rooms in the households for each small area. Only area level data was provided to us.

We propose a full non-subjective Bayesian analysis for the general small area problem, where we model both the population means and variances. The initial idea was to use Jeffreys' general rule prior, treating all the parameters including the degrees of freedom of the Student's $t-$ distribution as unknown. However, the resultant prior yielded an improper posterior, which led to a modified Jeffreys' prior resulting in a proper posterior.

The outline of the remaining sections is as follows. Section 2 introduces the model, the Fisher information matrix, Jeffreys' prior and its modification. The impropriety of the posterior under the former, and its propriety under the latter are also included in this section. Section 3 contains a real data analysis as well as a simulation study. Some final remarks are made in Section 4.

The fact that error variances are really random has been recognized for a long time. The work of Otto and Bell (1995), Arora and Lahiri (1997), Wang and Fuller (2003), Rivest and Vandal (2003) and others have tried to account for this in different ways. Slud and Maiti (2006), Dass, Maiti, Ren and Sinha (2012) and Maiti, Ren and Sinha (2014) used an empirical Bayes approach towards this end by estimating the hyper-parameters. Full Bayesian analysis using hierarchical Bayesian methods with normality of area level effects has been considered in You and Chapman (2006) and Sugasawa, Tamae and Kubokawa (2017). We will demonstrate that $t-$ priors for random effects often perform better than the methods of the last two papers via data analysis and simulations.

The use of $t-$ priors for the errors in the standard normal regression models, but not in mixed effects models, was proposed in Lange, Little and Taylor (1989), Fernandez and Steel (1998), Vrontos, Dellaportas and Politis (2000), Jacquier, Polson and Rossi (2004), and Fonseca, Ferreira and Migon (2008) primarily for protection against outliers. However, there are situations where normality of errors is a reasonable assumption, mainly because of the central limit theorem. Also, there are model diagnostic techniques to check this. The normality assumption of random effects, however, does not always work well. For the Brazilian data that we have on hand, joint modeling of both sample means and variances along with $t-$ priors for random effects yields better performance than some of the other area level models. As suggested by referee, we used the data to compute the residuals fitting a regression with the true area means and covariates to investigate the distribution of the random effects for this application. See Section 3.

2 The model

A typical area level model is given by $y_i = \mathbf{x}_i^T \boldsymbol{\beta} + u_i + e_i, \; (i = 1, \ldots, m)$, where m denotes the number of small areas, $\mathbf{x}_1, \ldots, \mathbf{x}_m$ are $p\,(< m)$ dimensional covariates, and $\boldsymbol{\beta}\,(p \times 1)$ is the vector of regression coefficients. The random effects u_i and the sampling errors e_i are assumed to be independently distributed with the $u_i \overset{\text{iid}}{\sim} N\,(0, \sigma_u^2)$ and the $e_i \overset{\text{ind}}{\sim} N\,(0, v_i)$. That is, the classic area level model is

$$\begin{aligned} y_i \,|\, \theta_i \; &\overset{\text{ind}}{\sim} \; N\,(\theta_i, v_i), \\ \theta_i \,|\, \boldsymbol{\beta}, \sigma_u^2 \; &\overset{\text{ind}}{\sim} \; N\,(\mathbf{x}_i^T \boldsymbol{\beta}, \sigma_u^2), \quad i = 1, \ldots, m. \end{aligned}$$

The v_i are assumed to be known in order to avoid non-identifiability. The assumption of known v_i almost becomes mandatory for secondary users of survey data who do not have access to any micro data for modeling the v_i. However, in reality they are random, based on sampled data. In situations when one has additional data to model the v_i, the data can be used efficiently for estimating the v_i. Moreover, in such situations, it is possible to have shrinkage estimators of the small area means θ_i as well as of the variances v_i.

We address small area estimation problems where we have additional data to model the v_i. Also, for robustification, we assume $t-$ distribution of the random effects instead of the normal distribution. We state our model as follows,

$$\begin{aligned} y_i \,|\, \theta_i, v_i \; &\overset{\text{ind}}{\sim} \; N\,(\theta_i, v_i), \; s_i^2 \,|\, v_i \overset{\text{ind}}{\sim} G\left(\frac{n_i - 1}{2}, \frac{1}{2v_i}\right) \\ \theta_i \,|\, \boldsymbol{\beta}, \sigma_\delta^2, \nu \; &\overset{\text{ind.}}{\sim} \; t_\nu\,(\mathbf{x}_i^T \boldsymbol{\beta}, \sigma_\delta), \quad i = 1, \ldots, m, \end{aligned} \tag{2.1}$$

where n_i is the sample size in the i^{th} area, $t_\nu\,(\mu, \sigma)$ denotes the Student's $t-$ distribution with location μ, scale σ and degrees of freedom ν, and $G\,(c, d)$ denotes the gamma distribution with the kernel density $x^{c-1}\exp\,(-dx)$ for $x > 0$.

For a full Bayesian analysis, our objective is to find the posterior distribution of $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m)^T$, given $\mathbf{y} = (y_1, \ldots, y_m)^T$ and $\mathbf{s}^2 = (s_1^2, \ldots, s_m^2)^T$. To this end, we first need to find the prior distributions for all the hyper-parameters, $\boldsymbol{\beta}$, $\mathbf{v} = (v_1, \ldots, v_m)^T$, σ_δ^2, and ν. The usual first try is Jeffreys' prior which is proportional to the positive square root of the determinant of the Fisher information matrix. The Fisher information matrix in our case is

$$\mathbf{I}\,(\boldsymbol{\beta}, \mathbf{v}, \sigma_\delta^2, \nu) = \begin{bmatrix} \dfrac{(\nu + 1)}{\sigma_\delta^2\,(\nu + 3)} \mathbf{X}^T \mathbf{X} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{D} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \dfrac{m\nu}{2\,(\sigma_\delta^2)^2\,(\nu + 3)} & \dfrac{-m}{\sigma_\delta^2\,(\nu + 1)(\nu + 3)} \\ \mathbf{0} & \mathbf{0} & \dfrac{-m}{\sigma_\delta^2\,(\nu + 1)(\nu + 3)} & mg\,(\nu) \end{bmatrix},$$

where $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_m)^T, \mathbf{D} = \mathrm{Diag}\left(\frac{n_1}{2v_1^2}, \ldots, \frac{n_m}{2v_m^2}\right),$ and $g(\nu) = \{\Psi'(\frac{\nu}{2}) - \Psi'(\frac{\nu+1}{2})\}/4 - (\nu+5)/\{2\nu(\nu+1)(\nu+3)\},$ with $\Psi(z) = \Gamma'(z)/\Gamma(z)$ and $\Psi'(z) = d\Psi(z)/dz$ which are the digamma and the trigamma functions. Thus, Jeffreys' prior is

$$\boldsymbol{\pi}_J\,(\boldsymbol{\beta},\ \mathbf{v},\ \sigma_\delta^2,\ \nu) \propto (\sigma_\delta^2)^{-\frac{p}{2}-1}\left|\mathbf{X}^T\mathbf{X}\right|^{\frac{1}{2}}\left|\mathbf{D}\right|^{\frac{1}{2}}\left(\frac{\nu+1}{\nu+3}\right)^{\frac{p}{2}}\left[\frac{\nu g(\nu)}{2(\nu+3)} - \frac{1}{(\nu+1)^2(\nu+3)^2}\right]^{\frac{1}{2}}. \tag{2.2}$$

However, Jeffreys' prior leads to an improper posterior due to the factor of $(\sigma_\delta^2)^{-\frac{p}{2}-1}$ in (2.2).

Theorem 1. Jeffreys' prior (2.2) leads an improper posterior.

Proof. Let $\boldsymbol{\pi}_J \equiv \boldsymbol{\pi}_J\,(\boldsymbol{\beta},\ \mathbf{v},\ \sigma_\delta^2,\ \nu\,|\,\mathbf{y},\ \mathbf{s}^2)$ be the posterior density with Jeffreys' prior (2.2). Considering the terms that contain σ_δ^2 in $\boldsymbol{\pi}_J$ and taking the transformation $w_i = (\theta_i - \mathbf{x}_i^T\boldsymbol{\beta})/\sigma_\delta,$ i.e., $\theta_i = \mathbf{x}_i^T\boldsymbol{\beta} + \sigma_\delta w_i,$ we have

$$\begin{aligned}
&\int_0^\infty (\sigma_\delta^2)^{-\frac{p}{2}-1}\exp\left[-\frac{1}{2}\sum_{i=1}^m \frac{1}{v_i}(y_i - \mathbf{x}_i^T\boldsymbol{\beta} - \sigma_\delta w_i)^2\right]d\sigma_\delta^2\\
&\qquad \ge \int_0^\infty (\sigma_\delta^2)^{-\frac{p}{2}-1}\exp\left[-\sum_{i=1}^m\left\{\frac{(y_i - \mathbf{x}_i^T\boldsymbol{\beta})^2}{v_i} + \frac{\sigma_\delta^2 w_i^2}{v_i}\right\}\right]d\sigma_\delta^2\\
&\qquad \ge \exp\left\{-\sum_{i=1}^m\frac{(y_i - \mathbf{x}_i^T\boldsymbol{\beta})^2}{v_i}\right\}\int_0^k (\sigma_\delta^2)^{-\frac{p}{2}-1}\exp\left(-\sum_{i=1}^m\frac{\sigma_\delta^2 w_i^2}{v_i}\right)d\sigma_\delta^2 \text{ for a constant } k > 0,\\
&\qquad \ge \exp\left[-\sum_{i=1}^m\left\{\frac{(y_i - \mathbf{x}_i^T\boldsymbol{\beta})^2}{v_i} + \frac{k w_i^2}{v_i}\right\}\right]\int_0^k (\sigma_\delta^2)^{-\frac{p}{2}-1}\,d\sigma_\delta^2 = \infty.
\end{aligned}$$

Therefore, Jeffreys' prior leads to an improper posterior.

However, once the component $(\sigma_\delta^2)^{-\frac{p}{2}-1}$ in (2.2) is replaced by $(\sigma_\delta^2)^{-\frac{p}{2}-1}\exp(-a/2\sigma_\delta^2)$ for some $a > 0,$ this modified version of Jeffreys' prior will lead a proper posterior under the condition, $\min(n_1,\ \ldots,\ n_m) > p.$ Therefore, we suggest a modified Jeffreys' prior for our model as follows:

$$\begin{aligned}
\boldsymbol{\pi}_{\mathrm{MJ}}\,(\boldsymbol{\beta},\ \mathbf{v},\ \sigma_\delta^2,\ \nu)\ &\propto (\sigma_\delta^2)^{-\frac{p}{2}-1}\exp\left(-\frac{a}{2\sigma_\delta^2}\right)\left(\prod_{i=1}^m\frac{1}{v_i}\right)\left(\frac{\nu+1}{\nu+3}\right)^{\frac{p}{2}}\\
&\qquad \times\left[\frac{\nu g(\nu)}{2(\nu+3)} - \frac{1}{(\nu+1)^2(\nu+3)^2}\right]^{\frac{1}{2}} \text{ where } a > 0.
\end{aligned}\tag{2.3}$$

By combining the likelihood of (2.1) and modified Jeffreys' prior (2.3), the full posterior of parameters given the data is

$$\boldsymbol{\pi}_{\text{MJ}}\left(\boldsymbol{\theta}, \boldsymbol{\beta}, \mathbf{v}, \sigma_\delta^2, \nu \,|\, \mathbf{y}, \mathbf{s}^2\right) \propto \left(\sigma_\delta^2\right)^{-\frac{p+m}{2}-1} \exp\left(-\frac{a}{2\sigma_\delta^2}\right)\left(\prod_{i=1}^m v_i^{-\frac{n_i}{2}-1}\right)\exp\left[-\frac{1}{2}\sum_{i=1}^m \frac{1}{v_i}\left(y_i - \theta_i\right)^2\right]$$

$$\times \exp\left(-\frac{1}{2}\sum_{i=1}^m \frac{s_i^2}{v_i}\right)\left[\prod_{i=1}^m \left\{1 + \frac{\left(\theta_i - \mathbf{x}_i^T \boldsymbol{\beta}\right)^2}{\nu \sigma_\delta^2}\right\}^{-\frac{\nu+1}{2}}\right]\left[\frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\sqrt{\nu}}\right]^m$$

$$\times \left(\frac{\nu+1}{\nu+3}\right)^{\frac{p}{2}}\left[\frac{\nu g(\nu)}{2(\nu+3)} - \frac{1}{(\nu+1)^2(\nu+3)^2}\right]^{\frac{1}{2}}. \tag{2.4}$$

Theorem 2. Under the model (2.1), the posterior $\boldsymbol{\pi}_{\text{MJ}}\left(\boldsymbol{\theta}, \boldsymbol{\beta}, \mathbf{v}, \sigma_\delta^2, \nu \,|\, \mathbf{y}, \mathbf{s}^2\right)$ in (2.4) is proper, provided $\min\left(n_1, \ldots, n_m\right) > p$.

Proof. See Appendix A.

Theorem 2 shows that modified Jeffreys' prior (2.3) leads to a proper posterior (2.4). The key idea is that we need a prior for σ_δ^2 such that $\int_0^\infty \boldsymbol{\pi}\left(\sigma_\delta^2\right)\left(\sigma_\delta^2\right)^{-\frac{p}{2}-1} d\sigma_\delta^2 < \infty$.

Remark 1. $\boldsymbol{\pi}_{\text{MJ}}\left(\boldsymbol{\beta}, \mathbf{v}, \sigma_\delta^2, \nu\right)$ can be factored into four independent priors for each parameter.

$$\boldsymbol{\pi}_{\text{MJ}}\left(\boldsymbol{\beta}, \mathbf{v}, \sigma_\delta^2, \nu\right) \propto \boldsymbol{\pi}(\boldsymbol{\beta})\,\boldsymbol{\pi}(\mathbf{v})\,\boldsymbol{\pi}\left(\sigma_\delta^2\right)\boldsymbol{\pi}(\nu)$$

where

$$\boldsymbol{\pi}(\boldsymbol{\beta}) \propto 1,\ \boldsymbol{\pi}(v_i) \propto \frac{1}{v_i} \quad \text{for} \quad i = 1, \ldots, m,$$

$$\boldsymbol{\pi}\left(\sigma_\delta^2\right) \sim \text{IG}\left(\frac{p}{2}, \frac{a}{2}\right)$$

and

$$\boldsymbol{\pi}(\nu) \propto \left(\frac{\nu+1}{\nu+3}\right)^{\frac{p}{2}}\left[\frac{\nu g(\nu)}{2(\nu+3)} - \frac{1}{(\nu+1)^2(\nu+3)^2}\right]^{\frac{1}{2}}.$$

Here $\text{IG}(c, d)$ denotes the inverse gamma distribution with the kernel density $x^{-c-1}\exp(-d/x)$ for $x > 0$.

The full conditional distributions to implement the Markov chain Monte Carlo (MCMC) are given in details in Appendix B. To generate samples, we use Gibbs sampling with Metropolis-Hastings algorithm where the conditional distribution of a parameter is known only up to a multiplicative constant. We provide details on how to apply a result of Chib and Greenberg (1995) for the Metropolis-Hastings algorithm to generate samples.

3 Application

3.1 Real data analysis

The data set is selected by a 10% random sampling of households in each area from a test demographic census completed in one municipality in Brazil. The municipality consists of 38,740 households in 140 small areas in total, and the number of households per area in the population ranges from 57 to 588. Thus the area sample sizes in the data set range from 6 to 59. We are interested in estimating the 140 population means of the head of household's income. The response variable y_i denotes the average income of the heads of households in i^{th} area.

This data set includes two centered auxiliary covariates which are the respective small area population means of the educational attainment of the head of households (ordinal scale of $0-5$) and the average number of rooms in households $(1-11+)$. Lastly, the data set contains the respective sampling variances which are calculated in the usual way. Since only area level data were provided to us and the true area means are known, we can compare the 140 small area predictions with the true area means respectively. The analysis suggests that our model performs better than other models where random effects are based on the normal distribution. For comparison, we use three other models.

The first one is the Fay-Herriot model, referred to as FH, with known sampling variances.

$$\begin{aligned}&y_i \mid \theta_i, v_i \overset{\text{ind.}}{\sim} N(\theta_i, v_i), \quad \theta_i \mid \boldsymbol{\beta}, \sigma_\delta^2 \overset{\text{ind.}}{\sim} N(\mathbf{x}_i^T \boldsymbol{\beta}, \sigma_\delta^2)\\ &\pi(\boldsymbol{\beta}) \propto 1, \quad \pi(\sigma_\delta^2) \sim \text{IG}(a_0, b_0),\end{aligned}$$

where a_0 and b_0 are chosen to be 0.0001 (a small constant) to reflect the vague knowledge of σ_δ^2.

The second model suggested by You and Chapman (2006), referred to as YC, is a hierarchical Bayesian model given by

$$\begin{aligned}&y_i \mid \theta_i, v_i \overset{\text{ind.}}{\sim} N(\theta_i, v_i), \quad \theta_i \mid \boldsymbol{\beta}, \sigma_\delta^2 \overset{\text{ind.}}{\sim} N(\mathbf{x}_i^T \boldsymbol{\beta}, \sigma_\delta^2)\\ &s_i^2 \mid v_i \overset{\text{ind}}{\sim} G\left(\frac{n_i - 1}{2}, \frac{1}{2v_i}\right)\\ &\pi(v_i) \sim \text{IG}(a_i, b_i), \quad \pi(\boldsymbol{\beta}) \propto 1, \quad \pi(\sigma_\delta^2) \sim \text{IG}(a_0, b_0),\end{aligned}$$

where a_i, b_i, a_0 and b_0 are also chosen to be 0.0001.

The third model is a Bayesian multi-stage small area model proposed by Sugasawa et al. (2017), referred to as STK. The STK model produces shrinkage estimation of both means and variances.

$$\begin{aligned}&y_i \mid \theta_i, v_i \overset{\text{ind.}}{\sim} N(\theta_i, v_i), \quad \theta_i \mid \boldsymbol{\beta}, \sigma_\delta^2 \overset{\text{ind.}}{\sim} N(\mathbf{x}_i^T \boldsymbol{\beta}, \sigma_\delta^2)\\ &s_i^2 \mid v_i \overset{\text{ind}}{\sim} G\left(\frac{n_i - 1}{2}, \frac{1}{2v_i}\right), \quad v_i \mid \gamma \sim \text{IG}(a_i, b_i\gamma),\\ &\pi(\boldsymbol{\beta}, \sigma_\delta^2, \gamma) = 1,\end{aligned}$$

where $a_i = 2$ and $b_i = 1/n_i$ as suggested by authors for a reasonable choice.

We compare the small area means predicted by FH, YC, STK, and our model, hereafter referred to as RTS model. For the MCMC implementation, we generate a chain with a burn-in length of 50,000 and the sampling size of $G = 50{,}000$. The estimates of the θ_i are given by

$$\hat{\theta}_i = \frac{1}{G}\sum_{g=1}^{G}\left(\gamma_i^{(g)} y_i + \left(1-\gamma_i^{(g)}\right)\mathbf{x}_i^T \boldsymbol{\beta}^{(g)}\right)$$

where

$$\gamma_i^{(g)} = \frac{v_i^{-1(g)}}{v_i^{-1(g)} + \sigma_\delta^{-2(g)}\eta_i^{(g)}}.$$

The comparison criteria are the average squared deviation (ASD), average absolute bias (AAB), average squared relative bias (ASRB), and average relative bias (ARB). They are defined as follows;

$$\text{ASD} = \frac{1}{m}\sum_{i=1}^{m}\left(\hat{\theta}_i - \theta_i\right)^2, \quad \text{AAB} = \frac{1}{m}\sum_{i=1}^{m}\left|\hat{\theta}_i - \theta_i\right|, \quad \text{ASRB} = \frac{1}{m}\sum_{i=1}^{m}\left(\frac{\hat{\theta}_i - \theta_i}{\theta_i}\right)^2,$$

and

$$\text{ARB} = \frac{1}{m}\sum_{i=1}^{m}\left|\frac{\hat{\theta}_i - \theta_i}{\theta_i}\right|,$$

where $\hat{\theta}_i$ and θ_i are the estimated and true values respectively in the i^{th} area. Table 3.1 compares the four models. Recall the prior distribution of σ_δ^2, which is $\boldsymbol{\pi}\left(\sigma_\delta^2\right) \sim \text{IG}\left(\frac{p}{2}, \frac{a}{2}\right)$. With the shape parameter, $\frac{p}{2} = 1$, we consider several values of a. If we choose a to be close to p, RTS model fits better than the rest under all four criteria. When we choose $a = 1$, RTS model performs best. YC model performs worse than the other three models. If we choose very small a, such as 0.01 or 0.001, then RTS model performs the worst.

Table 3.1
Comparison between RTS model, FH model, YC model, and STK model

Model	ASD	AAB	ASRB	ARB
RTS model (a = 0.0001)	57.297	6.152	0.395	0.589
RTS model (a = 0.01)	16.546	2.741	0.090	0.244
RTS model (a = 0.5)	3.244	1.249	0.020	0.118
RTS model (a = 0.2)	4.185	1.439	0.025	0.133
RTS model (a = 1)	2.745	1.164	0.019	0.113
RTS model (a = 2)	3.080	1.231	0.020	0.117
RTS model (a = 3)	3.079	1.229	0.020	0.117
RTS model (a = 5)	2.994	1.213	0.019	0.116
RTS model (a = 10)	3.377	1.278	0.020	0.119
RTS model (a = 50)	2.905	1.180	0.018	0.112
RTS model (a = 100)	2.799	1.154	0.018	0.109
FH model	4.484	1.448	0.026	0.133
YC model	4.983	1.543	0.029	0.141
STK model	3.199	1.257	0.021	0.121

Additionally, as suggested by a referee, we compute the residuals fitting a regression with the true area means and covariates to see the distribution of the random effects for this real data. Figure 3.1 shows that the distribution departs from the normal distribution.

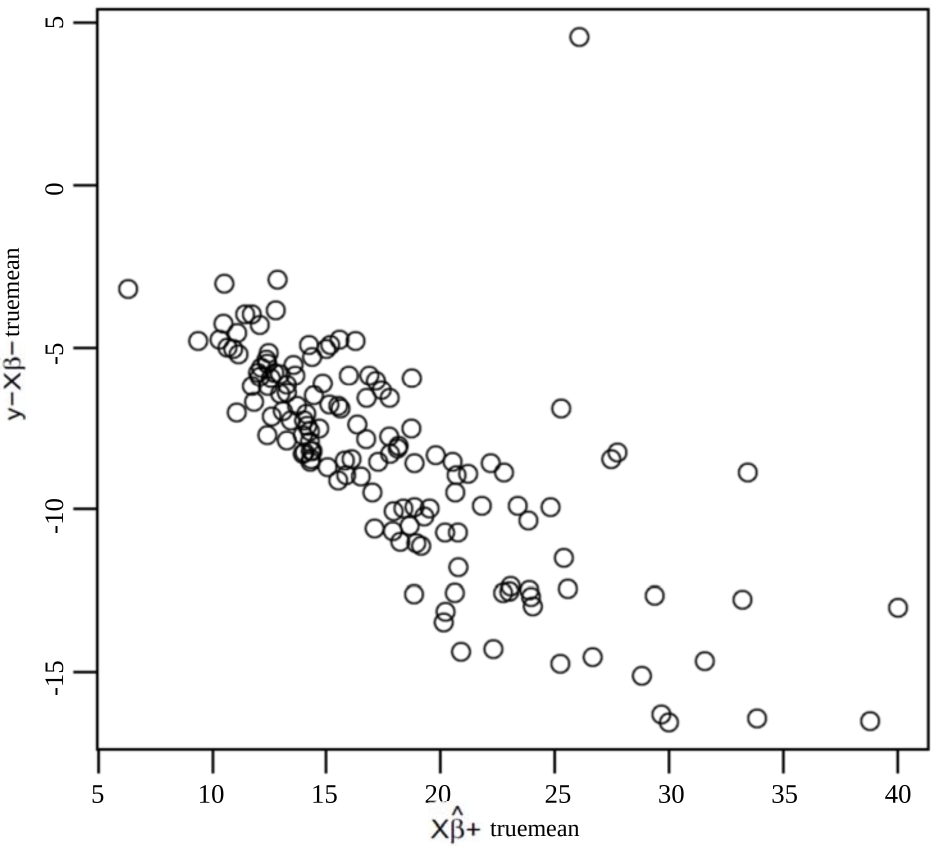


Figure 3.1 Residuals fitting a regression with the true mean and covariates.

3.2 Simulation study

In this section, we set up a simulation close to Maiti et al. (2014) (or Sugasawa et al. (2017)) to compare the accuracy of our estimators to other estimators, specifically those from You and Chapman (2006) and Sugasawa et al. (2017). We generate observations for each small area from the model

$$y_{ij} = \beta_0 + \beta_1 x_i + u_i + e_{ij}, \quad j = 1, \ldots, n_i, \, i = 1, \ldots, m,$$

where $u_i \sim t_\nu\,(0, \sigma_\delta)$ and $e_{ij} \sim N\,(0, n_i v_i)$. Then the random effects model for the small area mean is

$$y_i = \beta_0 + \beta_1 x_i + u_i + e_i, \quad i = 1, \ldots, m,$$

where $y_i = \bar{y}_i = n_i^{-1}\sum_{j=1}^{n_i} y_{ij}$ and $e_i = n_i^{-1}\sum_{j=1}^{n_i} e_{ij}$. Hence, $y_i \,|\, \theta_i \sim N\,(\theta_i, v_i)$ where $\theta_i = \beta_0 + \beta_1 x_i + u_i$; $\theta_i \sim t_\nu\,(\beta_0 + \beta_1 x_i, \sigma_\delta)$, and $e_i \sim N\,(0, v_i)$. The interest parameter is the mean θ_i, for the i^{th} small area. Also, the direct estimator of v_i is

$$\frac{1}{n_i\,(n_i-1)}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_i)^2.$$

We set $m = 30$ and $n_i = 7$ for all areas, and generate covariates x_i from the uniform distribution on (2, 8). The true parameter values are set as $\beta_0 = 0.5$, $\beta_1 = 0.8$, $\sigma_\delta = 1$, $\nu = 3$ and $v_i \sim \mathrm{IG}(10, 5\exp(0.3x_i))$. Also, we chose $a = 3$ for all simulations.

For the MCMC implementation, we generated 5,000 posterior samples after discarding the first 1,000 for $R = 2{,}000$ simulation runs. Table 3.2 provides comparison among the four models. The comparison criteria are ASD, AAB, and BIAS, the latter being defined as

$$\mathrm{BIAS} = \frac{1}{mR}\sum_{i=1}^{m}\left|\sum_{r=1}^{R}\left(\hat{\theta}_i^{(r)} - \theta_i^{(r)}\right)\right|.$$

Table 3.2
Simulation result for t random effects with $\nu = 3$

Model	Mean			Variance		
	ASD	AAB	BIAS	ASD	AAB	BIAS
RTS	1.393	0.895	0.021	5.048	1.608	1.324
STK	1.821	0.933	0.025	5.042	1.514	1.367
FH	1.540	0.942	0.022			
YC	2.165	0.974	0.030	5.970	1.803	1.689

RTS model performs better than others under ASD, AAB, and BIAS criteria for the mean. While RTS shows small improvements over other models for AAB and BIAS criteria, it shows approximate 23.5%, 10%, and 35.7% improvements over STK, FH and YC models respectively for ASD criteria. For the variance, RTS and STK models perform better than YC model.

The following two tables provide the simulation results when one sets the degrees of freedom as $\nu = 2$ and $\nu = 4$.

Table 3.3
Simulation result for t random effects with $\nu = 2$

Model	Mean			Variance		
	ASD	AAB	BIAS	ASD	AAB	BIAS
RTS	1.617	0.949	0.020	6.569	1.610	1.070
STK	7.566	1.107	0.038	11.144	1.441	0.996
FH	1.921	1.035	0.022			
YC	9.063	1.187	0.038	7.072	1.685	1.340

Table 3.4
Simulation result for t random effects with $\nu = 4$

Model	Mean			Variance		
	ASD	AAB	BIAS	ASD	AAB	BIAS
RTS	1.265	0.862	0.019	4.876	1.619	1.428
STK	1.322	0.874	0.019	5.077	1.577	1.489
FH	1.350	0.894	0.020			
YC	1.509	0.905	0.022	6.201	1.869	1.802

With $\nu = 2,$ RTS model performs better than others under ASD, AAB, and BIAS criteria for the mean. In this simulation, the ASD values for STK and YC models are very large compared with RTS and FH models. RTS model shows improvements of about 78.6% over STK model, 82.2% over YC model, and 15.8% over FH model. For AAB and BIAS, the values of RTS model are smaller than those of other models. When considering the variance, ASD for RTS model gives smallest value.

With $\nu = 4,$ RTS model also shows better performance over others. Especially, ASD and BIAS values indicate that RTS model improves results when compared with STK and YC model.

The next two tables consider the situation where one assumes normality of the random effects. Here RTS model performs slightly worse than the other models.

Table 3.5
Simulation result for normal random effects with $N(0, 5^2)$

Model	Mean			Variance		
	ASD	AAB	BIAS	ASD	AAB	BIAS
RTS	2.896	1.305	0.038	6.036	1.512	0.514
STK	2.560	1.229	0.051	1.851	0.961	0.114
FH	2.597	1.240	0.036			
YC	2.735	1.259	0.048	3.674	1.305	0.463

Table 3.6
Simulation result for normal random effects with $N(0, 10^2)$

Model	Mean			Variance		
	ASD	AAB	BIAS	ASD	AAB	BIAS
RTS	3.007	1.316	0.032	10.117	1.895	1.202
STK	2.784	1.272	0.031	2.221	1.038	0.155
FH	2.765	1.272	0.048			
YC	2.873	1.285	0.033	9.166	1.798	1.129

4 Final remarks

The paper considers small area models for handling area level data. The new feature of this article is modeling both small area means and variances along with the use of $t-$ distribution of random effects. It is shown via both data analysis and simulation that the proposed method performs mostly better than the models of You and Chapman (2006) and Sugasawa et al. (2017) in most situations.

Acknowledgements

Ghosh's research was partially supported in part by NSF Grant SES-1327359.

Appendix A

Proof

Theorem 2. Under the model (2.1) with modified Jeffreys' prior (2.3), the posterior distribution (2.4) is proper, provided $\min(n_1, \ldots, n_m) > p.$

Proof. Recall the posterior distribution (2.4),

$$\boldsymbol{\pi}_{\mathrm{MJ}}\left(\boldsymbol{\theta},\, \boldsymbol{\beta},\, \mathbf{v},\, \sigma_\delta^2,\, \nu \,|\, \mathbf{y},\, \mathbf{s}^2\right) \propto \left(\sigma_\delta^2\right)^{-\frac{p+m}{2}-1} \exp\left(-\frac{a}{2\sigma_\delta^2}\right)\left(\prod_{i=1}^m v_i^{-\frac{n_i}{2}-1}\right)\exp\left[-\frac{1}{2}\sum_{i=1}^m \frac{1}{v_i}\left(y_i-\theta_i\right)^2\right]$$

$$\times \exp\left(-\frac{1}{2}\sum_{i=1}^m \frac{s_i^2}{v_i}\right)\left[\prod_{i=1}^m \left\{1+\frac{\left(\theta_i-\mathbf{x}_i^T\boldsymbol{\beta}\right)^2}{\nu\sigma_\delta^2}\right\}^{-\frac{\nu+1}{2}}\right]\left[\frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\sqrt{\nu}}\right]^m$$

$$\times\left(\frac{\nu+1}{\nu+3}\right)^{\frac{p}{2}}\left[\frac{\nu g(\nu)}{2(\nu+3)}-\frac{1}{(\nu+1)^2(\nu+3)^2}\right]^{\frac{1}{2}}$$

where $g(\nu) = \left\{\Psi'\left(\frac{\nu}{2}\right) - \Psi'\left(\frac{\nu+1}{2}\right)\right\}/4 - (\nu+5)/\{2\nu(\nu+1)(\nu+3)\}.$

First of all, since $\log\left[\Gamma\left\{\frac{(\nu+1)}{2}\right\}/\Gamma\left(\frac{\nu}{2}\right)\right] \doteq \{-\log(2e)+\nu\log(\nu+1)-(\nu-1)\log\nu\}/2$ by Stirling's approximation, we have

$$\frac{1}{2}\left[\Psi\left(\frac{\nu+1}{2}\right)-\Psi\left(\frac{\nu}{2}\right)\right] \doteq \frac{d}{d\nu}\log\frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)} = \frac{1}{2}\left[\log(\nu+1)-\log\nu\right]+\frac{1}{2\nu(\nu+1)}$$

and

$$\frac{1}{4}\left[\Psi'\left(\frac{\nu+1}{2}\right)-\Psi'\left(\frac{\nu}{2}\right)\right] \doteq \frac{1}{2}\left(\frac{1}{\nu+1}-\frac{1}{\nu}\right)-\frac{1}{2\nu^2}+\frac{1}{2(\nu+1)^2}.$$

This leads to

$$g(\nu) \doteq \frac{5\nu+3}{2\nu^2(\nu+1)^2(\nu+3)}$$

and

$$\left[\frac{\nu g(\nu)}{2(\nu+3)}-\frac{1}{(\nu+1)^2(\nu+3)^2}\right] \doteq \frac{1}{4\nu(\nu+1)^2(\nu+3)}.$$

Hence this approximation simplifies the last term in (2.4). The corresponding posterior is

$$\boldsymbol{\pi}_{\mathrm{MJ}}\left(\boldsymbol{\theta},\, \boldsymbol{\beta},\, \mathbf{v},\, \sigma_\delta^2,\, \nu \,|\, \mathbf{y},\, \mathbf{s}^2\right) \propto \left(\sigma_\delta^2\right)^{-\frac{p+m}{2}-1} \exp\left(-\frac{a}{2\sigma_\delta^2}\right)\left(\prod_{i=1}^m v_i^{-\frac{n_i}{2}-1}\right)\exp\left[-\frac{1}{2}\sum_{i=1}^m \frac{1}{v_i}\left(y_i-\theta_i\right)^2\right]$$

$$\times \exp\left(-\frac{1}{2}\sum_{i=1}^m \frac{s_i^2}{v_i}\right)\left[\prod_{i=1}^m \left\{1+\frac{\left(\theta_i-\mathbf{x}_i^T\boldsymbol{\beta}\right)^2}{\nu\sigma_\delta^2}\right\}^{-\frac{\nu+1}{2}}\right]\left[\frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\sqrt{\nu}}\right]^m$$

$$\times\left(\frac{\nu+1}{\nu+3}\right)^{\frac{p}{2}}\left[\frac{1}{\nu(\nu+1)^2(\nu+3)}\right]^{\frac{1}{2}}. \qquad \text{(A.1)}$$

First, integrating out with respect to $\boldsymbol{\beta}$. By letting $w_i = \left(\theta_i - \mathbf{x}_i^T\boldsymbol{\beta}\right)/\sigma_\delta$, i.e., $\theta_i = \mathbf{x}_i^T\boldsymbol{\beta} + \sigma_\delta w_i$, we have

$$\boldsymbol{\pi}_{\mathrm{MJ}}\,(\mathbf{w},\,\mathbf{v},\,\sigma_\delta^2,\,\nu\,|\,\mathbf{y},\,\mathbf{s}^2) \propto (\sigma_\delta^2)^{-\frac{p}{2}-1}\exp\left(-\frac{a}{2\sigma_\delta^2}\right)\left(\prod_{i=1}^m v_i^{-\frac{n_i}{2}-1}\right)\exp\left(-\frac{1}{2}\sum_{i=1}^m\frac{s_i^2}{v_i}\right)$$

$$\times\left[\prod_{i=1}^m\left(1+\frac{w_i^2}{\nu}\right)^{-\frac{\nu+1}{2}}\right]\left[\frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\sqrt{\nu}}\right]^m\left(\frac{\nu+1}{\nu+3}\right)^{\frac{p}{2}}\left[\frac{1}{\nu\,(\nu+1)^2\,(\nu+3)}\right]^{\frac{1}{2}}|\,\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X}\,|^{-\frac{1}{2}}$$

$$\times\exp\left[-\frac{1}{2}(\mathbf{Y}-\sigma_\delta\mathbf{w})^T\left\{\mathbf{V}^{-1}-\mathbf{V}^{-1}\mathbf{X}\,(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\,\mathbf{X}^T\mathbf{V}^{-1}\right\}(\mathbf{Y}-\sigma_\delta\mathbf{w})\right].$$

where $\mathbf{w}=(w_1,\,\ldots,\,w_m)^T$. Note that $\mathbf{V}^{-1}-\mathbf{V}^{-1}\mathbf{X}\,(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\,\mathbf{X}^T\mathbf{V}^{-1}$ is non-negative definite. Using $K\,(>0)$ as a generic constant which depends only on the data $(\mathbf{y},\,\mathbf{s}^2)$, we integrate out with respect to $\mathbf{w}$ and σ_δ^2 respectively in order, we have

$$\boldsymbol{\pi}_{\mathrm{MJ}}\,(\mathbf{v},\,\sigma_\delta^2,\,\nu\,|\,\mathbf{y},\,\mathbf{s}^2)\ \le K\,(\sigma_\delta^2)^{-\frac{p}{2}-1}\exp\left(-\frac{a}{2\sigma_\delta^2}\right)\left(\prod_{i=1}^m v_i^{-\frac{n_i}{2}-1}\right)$$

$$\times\exp\left(-\frac{1}{2}\sum_{i=1}^m\frac{s_i^2}{v_i}\right)\left[\frac{1}{\nu\,(\nu+1)^2\,(\nu+3)}\right]^{\frac{1}{2}}|\,\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X}\,|^{-\frac{1}{2}},$$

and then

$$\boldsymbol{\pi}_{\mathrm{MJ}}\,(\mathbf{v},\,\nu\,|\,\mathbf{y},\,\mathbf{s}^2)\le K\left(\prod_{i=1}^m v_i^{-\frac{n_i}{2}-1}\right)\exp\left(-\frac{1}{2}\sum_{i=1}^m\frac{s_i^2}{v_i}\right)\times\left[\frac{1}{\nu\,(\nu+1)^2\,(\nu+3)}\right]^{\frac{1}{2}}|\,\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X}\,|^{-\frac{1}{2}}.$$

Note that

$$\int_0^\infty \nu^{-\frac{1}{2}}\,(\nu+1)^{-1}\,(\nu+3)^{-\frac{1}{2}}\,d\nu=\int_0^c \nu^{-\frac{1}{2}}\,(\nu+1)^{-1}\,(\nu+3)^{-\frac{1}{2}}\,d\nu+\int_c^\infty \nu^{-\frac{1}{2}}\,(\nu+1)^{-1}\,(\nu+3)^{-\frac{1}{2}}\,d\nu$$

$$\le\int_0^c \nu^{-\frac{1}{2}}d\nu+\int_c^\infty \nu^{-2}d\nu = 2c^{\frac{1}{2}}+c^{-1}<\infty\ \text{for any}\ c>0.$$

After integrating out with respect to ν, we have

$$\boldsymbol{\pi}_{\mathrm{MJ}}\,(\mathbf{v}\,|\,\mathbf{y},\,\mathbf{s}^2)\le K\left(\prod_{i=1}^m v_i^{-\frac{n_i}{2}-1}\right)\exp\left(-\frac{1}{2}\sum_{i=1}^m\frac{s_i^2}{v_i}\right)|\,\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X}\,|^{-\frac{1}{2}}.$$

Finally, since $|\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X}|^{-\frac{1}{2}}\le(v_{\max})^{\frac{p}{2}}\,|\mathbf{X}^T\mathbf{X}|^{-\frac{1}{2}}$, where $v_{\max}=\max(v_1,\ldots,v_m)$, if $\min(n_1,\,\ldots,\,n_m)>p$, modified Jeffrey's prior leads to a proper posterior.

Appendix B

Full conditional distributions

The full posterior of the parameters given the data is specified in (A.1). For the MCMC implementation, it is convenient to use the latent parameters η_i $(i=1,\,\ldots,\,m)$ such that

$$\begin{aligned}\theta_i \mid \eta_i &\overset{\text{ind}}{\sim} \mathrm{N}\left(\mathbf{x}_i^T \boldsymbol{\beta},\, \eta_i^{-1} \sigma_\delta^2\right)\\ \eta_i &\overset{\text{iid}}{\sim} \mathrm{G}\left(\frac{\nu}{2}, \frac{\nu}{2}\right).\end{aligned}$$

Let $\boldsymbol{\Lambda} = \operatorname{Diag}(\eta_1, \ldots, \eta_m)$. The full conditional distributions are

I. $\left[\sigma_\delta^2 \mid \boldsymbol{\theta}, \boldsymbol{\Lambda}, \boldsymbol{\beta}, \mathbf{v}, \nu, \mathbf{y}, \mathbf{s}^2\right] \sim \mathrm{IG}\left[\frac{p+m}{2}, \frac{1}{2}\left\{a + \sum_{i=1}^m \eta_i \left(\theta_i - \mathbf{x}_i^T \boldsymbol{\beta}\right)^2\right\}\right];$

II. $\left[v_i \mid \boldsymbol{\theta}, \boldsymbol{\Lambda}, \boldsymbol{\beta}, \sigma_\delta^2, \nu, \mathbf{y}, \mathbf{s}^2\right] \overset{\text{ind}}{\sim} \mathrm{IG}\left[\frac{n_i}{2}, \frac{1}{2}\left\{\left(y_i - \theta_i\right)^2 + s_i^2\right\}\right];$

III. $\left[\boldsymbol{\beta} \mid \boldsymbol{\theta}, \boldsymbol{\Lambda}, \mathbf{v}, \sigma_\delta^2, \nu, \mathbf{y}, \mathbf{s}^2\right] \sim \mathrm{N}\left[\left(\mathbf{X}^T \boldsymbol{\Lambda} \mathbf{X}\right)^{-1} \mathbf{X}^T \boldsymbol{\Lambda} \boldsymbol{\theta}, \sigma_\delta^2 \left(\mathbf{X}^T \boldsymbol{\Lambda} \mathbf{X}\right)^{-1}\right];$

IV. $\left[\eta_i \mid \boldsymbol{\theta}, \boldsymbol{\beta}, \mathbf{v}, \sigma_\delta^2, \nu, \mathbf{y}, \mathbf{s}^2\right] \overset{\text{ind}}{\sim} \mathrm{G}\left[\frac{\nu+1}{2}, \frac{1}{2}\left\{\nu + \frac{\left(\theta_i - \mathbf{x}_i^T \boldsymbol{\beta}\right)^2}{\sigma_\delta^2}\right\}\right];$

V. $\left[\theta_i \mid \boldsymbol{\Lambda}, \boldsymbol{\beta}, \mathbf{v}, \sigma_\delta^2, \nu, \mathbf{y}, \mathbf{s}^2\right] \overset{\text{ind}}{\sim} \mathrm{N}\left[\left(v_i^{-1} + \sigma_\delta^{-2} \eta_i\right)^{-1} \left(v_i^{-1} y_i + \sigma_\delta^{-2} \eta_i \mathbf{x}_i^T \boldsymbol{\beta}\right), \left(v_i^{-1} + \sigma_\delta^{-2} \eta_i\right)^{-1}\right];$

and

$$\begin{aligned}\text{IV.}\quad \left[\nu \mid \boldsymbol{\theta}, \boldsymbol{\Lambda}, \boldsymbol{\beta}, \mathbf{v}, \sigma_\delta^2, \mathbf{y}, \mathbf{s}^2\right] &\propto \nu^{\frac{m-1}{2}} \exp\left\{-\tfrac{\nu}{2} \sum_{i=1}^m \left(\eta_i - \log \eta_i - 1\right)\right\}\\ &\times \left\{\frac{\nu^{\frac{m\nu - m}{2}} \exp\left(-\frac{m\nu}{2}\right)}{2^{\frac{m\nu}{2}} \Gamma^m\left(\frac{\nu}{2}\right)} \left(\frac{\nu+1}{\nu+3}\right)^{\frac{p}{2}} (\nu+1)^{-1} (\nu+3)^{-\frac{1}{2}}\right\}. \qquad \text{(B.1)}\end{aligned}$$

All but (VI) requires the generation of samples from standard distributions. While we use the Gibbs sampling method for (I)-(V), we use the Metropolis-Hastings algorithm for generating samples from (VI) as given in Chib and Greenberg (1995).

How to apply the result of Chib and Greenberg (1995) to (IV) in (B.1).

If the target density $\pi(t)$ can be written as $\pi(t) \propto \psi(t)\, h(t)$, where $h(t)$ is a density that can be sampled and $\psi(t)$ is uniformly bounded, then one can set $h(t)$ as a candidate density to draw samples and use $\psi(t)$ in $\alpha(x, y) = \min\{\psi(y)/\psi(x), 1\}$ which is the probability of move.

Recall that the full conditional distribution of ν is

$$\boldsymbol{\pi}(\nu \mid \cdot) \propto \nu^{\frac{m-1}{2}} \exp\left\{-\frac{\nu}{2} \sum_{i=1}^m \left(\eta_i - \log \eta_i - 1\right)\right\} \times \left\{\frac{\nu^{\frac{m\nu - m}{2}} \exp\left(-\frac{m\nu}{2}\right)}{2^{\frac{m\nu}{2}} \Gamma^m\left(\frac{\nu}{2}\right)} \left(\frac{\nu+1}{\nu+3}\right)^{\frac{p}{2}} (\nu+1)^{-1} (\nu+3)^{-\frac{1}{2}}\right\}.$$

Since $\Gamma\left(\frac{\nu}{2}\right) \geq \sqrt{2\pi}\, \exp\left(-\frac{\nu}{2}\right) \left(\frac{\nu}{2}\right)^{\frac{\nu-1}{2}}$, the second $\{\cdot\}$ term above is bounded by $\left(\sqrt{2\pi}\right)^{-m} (\nu+1)^{\frac{p}{2}-1} (\nu+3)^{-\frac{p}{2}-\frac{1}{2}} \leq \left(\sqrt{2\pi}\right)^{-m} \Big/ \sqrt{3}$. Hence, we can apply third method in the Section 5 of Chib and Greenberg (1995) with

$$h(\nu) \sim \mathrm{G}\left(\frac{m+1}{2}, \frac{1}{2} \sum_{i=1}^m \left(\eta_i - \log \eta_i - 1\right)\right)$$

and

$$\psi(\nu) = \frac{\nu^{\frac{m\nu - m}{2}} \exp\left(-\frac{m\nu}{2}\right)}{2^{\frac{m\nu}{2}} \Gamma^m\left(\frac{\nu}{2}\right)} \left(\frac{\nu+1}{\nu+3}\right)^{\frac{p}{2}} (\nu+1)^{-1} (\nu+3)^{-\frac{1}{2}}.$$

Here $h(v)$ is a candidate-generating density, and $\psi(v)$ is uniformly bounded.

References

Arora, V., and Lahiri, P. (1997). On the superiority of the Bayesian method over the BLUP in small area estimation problems. *Stastica Sinica*, 7, 1053-1063.

Battese, G.E., Harter, R.M. and Fuller, W.A. (1988). An error component model for prediction of mean crop areas using survey and satellite data. *Journal of the American Statistical Association*, 95, 28-36.

Bell, W.R. (2008). Examining sensitivity of small area inferences to uncertainty about sampling error variances. In *Proceedings of the Survey Research Methods Section*, American Statistical Association, 327-333.

Bell, W.R., and Huang, E.T. (2006). Using the $t-$distribution to deal with outliers in small area estimation. Proceedings: *Symposium 2006, Methodological Issues in Measuring Population Health*, Statistics Canada.

Butar, F., and Lahiri, P. (2002). On the measures of uncertainty of empirical Bayes small-area estimators. *Journal of Statistical Planning and Inference*, 112, 63-76.

Chib, S., and Greenberg, E. (1995). Understanding the Metropolis-Hastings Algorithm. *The American Statistician*, 49, 327-335.

Dass, S.C., Maiti, T., Ren, H. and Sinha, S. (2012). Confidence interval estimation of small area parameters shrinking both means and variances. *Survey Methodology*, 38, 2, 173-187. Paper available at https://www150.statcan.gc.ca/n1/pub/12-001-x/2012002/article/11756-eng.pdf.

Datta, G.S., and Lahiri, P. (1995). Robust hierarchical Bayes estimation of small area characteristics in the presence of outliers. *Journal of Multivariate Analysis*, 54, 310-328.

Fay, R.E., and Herriot, R.A. (1979). Estimates of income for small places: An application of James-Stein procedures to census data. *Journal of the American Statistical Association*, 74, 269-277.

Fernandez, C., and Steel, M.F.J. (1998). On Bayesian modeling of fat tails and skewness. *Journal of the American Statistical Association*, 93, 359-371.

Fonseca, T.C.O., Ferreira, M.A.R. and Migon, H.S. (2008). Objective Bayesian analysis for the student$-t$ regression model. *Biometrika*, 95, 325-333.

Jacquier, E., Polson, N.G. and Rossi, P.E. (2004). Bayesian analysis of stochastic volatility models with fat-tails and correlated errors. *Journal of Econometrics*, 122, 185-212.

Jiang, J., Lahiri, P. and Wan, S. (2002). A unified jackknife theory for empirical best prediction with M-estimation. *Annals of Statistics*, 30, 1782-1810.

Lahiri, P., and Rao, J.N.K. (1995). Robust estimation of mean square error of small area estimators. *Journal of the American Statistical Association*, 90, 758-766.

Lange, K.L., Little, R.J.A. and Taylor, J.M.G. (1989). Robust statistical modeling using the t distribution. *Journal of the American Statistical Association*, 84, 881-896.

Maiti, T., Ren, H. and Sinha, A. (2014). Prediction error of small area predictors shrinking both means and variances. *Scandinavian Journal of Statistics*, 41, 775-790.

Otto, M.C., and Bell, W.R. (1995). Sampling error modelling of poverty and income statistics for states. In *Proceedings of the Section on Government Statistics*, American Statistical Association, 160-165.

Rivest, L.-P., and Vandal, N. (2003). Mean squared error estimation for small areas when the small area variances are estimated. In *Proceedings of the International Conference on Recent Advances in Survey Sampling.*

Slud, E.V., and Maiti, T. (2006). Mean-squared error estimation in transformed Fay-Herriot models. *Journal of Royal Statistical Society, Series B*, 68, 239-257.

Sugasawa, S., Tamae, H. and Kubokawa, T. (2017). Bayesian estimators for small area models shrinking both means and variances. *Scandinavian Journal of Statistics*, 44, 150-167.

Vrontos, I.D., Dellaportas, P. and Politis, D.N. (2000). Full Bayesian inference for GARCH and EGARCH models. *Journal of Business and Economic Statistics*, 18, 187-198.

Wang, J., and Fuller, W. (2003). The mean squared error of small area predictors constructed with estimated error variances. *Journal of the American Statistical Association*, 98, 716-723.

You, Y., and Chapman, B. (2006). Small area estimation using area level models and estimated sampling variances. *Survey Methodology*, 32, 1, 97-103. Paper available at https://www150.statcan.gc.ca/n1/pub/12-001-x/2006001/article/9263-eng.pdf.

Vol. 44, No. 1, pp. 117-144
Statistics Canada, Catalogue No. 12-001-X

Model-assisted calibration of non-probability sample survey data using adaptive LASSO

Jack Kuang Tsung Chen, Richard L. Valliant and Michael R. Elliott[1]

Abstract

The probability-sampling-based framework has dominated survey research because it provides precise mathematical tools to assess sampling variability. However increasing costs and declining response rates are expanding the use of non-probability samples, particularly in general population settings, where samples of individuals pulled from web surveys are becoming increasingly cheap and easy to access. But non-probability samples are at risk for selection bias due to differential access, degrees of interest, and other factors. Calibration to known statistical totals in the population provide a means of potentially diminishing the effect of selection bias in non-probability samples. Here we show that model calibration using adaptive LASSO can yield a consistent estimator of a population total as long as a subset of the true predictors is included in the prediction model, thus allowing large numbers of possible covariates to be included without risk of overfitting. We show that the model calibration using adaptive LASSO provides improved estimation with respect to mean square error relative to standard competitors such as generalized regression (GREG) estimators when a large number of covariates are required to determine the true model, with effectively no loss in efficiency over GREG when smaller models will suffice. We also derive closed form variance estimators of population totals, and compare their behavior with bootstrap estimators. We conclude with a real world example using data from the National Health Interview Survey.

Key Words: Adaptive LASSO estimators; Generalized regression estimator; Non-representative sample; Over-fitting; Variable selection; Oracle property.

1 Introduction

Probability-based sampling has dominated survey research for the greater part of the past century (Stephan, 1948; Frankel and Frankel, 1987). Given complete measures on sampled units with known selection probabilities, randomization theory removes selection bias by generating representative samples of the target population. On the other hand, non-probability samples generated without known selection probabilities are automatically at risk for selection bias, as samples can differ from the target population on key statistics (Groves, 2006). Well-documented failures in 1936 and 1948 presidential election polls highlight the potential downfalls in making population inference from non-probability samples (Mosteller, 1949).

Although the probability-sampling-based framework provides survey practitioners precise mathematical tools to assess and correct sampling errors, declining response rates among traditional data collection methods raise concerns over the potentially high nonresponse bias of probability samples. Pew Research reported that their response rates (RRs) in typical telephone surveys dropped from 36% in 1997 to 9% in 2012 (Kohut, Keeter, Doherty, Dimock and Christian, 2012), suggesting that telephone-based probability sampling may no longer be a viable methodology for general population surveys. In addition, obtaining data without exercising much control over the set of units for which it is collected is often cheaper and quicker than probability sampling. For these reasons non-probability sampling is currently staging a kind of

[1] Jack Kuang Tsung Chen is Statistician, Survey Monkey Inc., Palo Alto, CA. E-mail: jjkkttcc@gmail.com; Richard L. Valliant is Research Professor, Survey Research Center, Institute for Social Research, University of Michigan, Ann Arbor, MI. E-mail: valliant@umich.edu; Michael R. Elliott is Research Professor, Survey Research Center, Institute for Social Research, and Professor, Department of Biostatistics, University of Michigan School of Public Health, Ann Arbor, MI. E-mail: mrelliot@umich.edu.

renascence (Baker, Brick, Bates, Battaglia, Couper, Dever, Gile and Tourangeau, 2013; Elliott and Valliant, 2017). Online data collection, a platform without a universal sampling frame to conduct probability-based sampling, was estimated to comprise nearly half of all U.S. survey research spending in 2012 (Terhanian and Bremer, 2012), and has almost certainly grown since then.

For many survey agencies, adjusting survey weights to known auxiliary information is the final and most crucial step in the weight construction process. Standard approaches include poststratification, in which weights are adjusted so that the weighted sample distribution of categorical auxiliary variables matches that of the population, and its extention to generalized regression estimation (GREG), which ensures that the weighted sum of each auxiliary variable (continuous or categorical) equals to its corresponding total in the population (Deville and Särndal, 1992). Calibration plays an important role in official statistics because it can generate weights such that the weighted demographic estimates across different surveys are consistent.

In probability samples, when design weights are equal to the inverse of selection probabilities, weighted estimates of sample totals are design-unbiased for the population total. Calibration adjusts design weights by a minimal degree so that the weighted sample totals for auxiliary variables match their known population totals (Särndal, Swensson and Wretman, 1992). In the probability sampling setting, calibration is introduced to reduce variance and/or correct for bias by adjusting for undercoverage or overcoverage of sub-groups of the sample. For large samples, the final calibrated weights can be applied to all variables in the survey, because they approximately maintain the unbiased property of original design weights. In non-probability samples, however, there are no selection probabilities to construct initial design weights that can produce unbiased estimates. Thus, there is no guarantee that the traditional calibrated weights can work for all variables in the non-probability sample. To make inference from non-probability samples, one practical approach is to construct a set of weights that can lower the root-mean-square error (RMSE) of weighted estimates with respect to a specific outcome of interest. Model-assisted calibration provides the framework to construct calibrated weights targeting an outcome variable, given a model that can approximate the expected values of the outcome (Wu and Sitter, 2001). The key to successful model-assisted calibration is a model with strong predictive properties: model parameters estimated from one sample can be used to reliably predict values in a different sample of the same population. Of course, such predictors are not always available; Tourangeau, Conrad and Couper (2013) provide an example where the lack of predictive covariates prevent weighting adjustments from performing well. However, Tourangeau, et al. (2013) had in mind household surveys. Predictors can be more powerful in establishment or institutional surveys or in some specialized surveys like election polls. For example, Wang, Rothschild, Goel and Gelman (2015) use party affiliation and candidate voted in the previous election to make accurate predictions of the outcome of the 2012 US presidential election based on a non-probability sample that was distributed much different from that of all voters.

Clearly, then, model-assisted calibration might be expected to be most effective when there is a relatively rich set of auxiliary population covariates and consequently an extremely large set of models to be considered. In these settings, obtaining balance between structure – to minimize model misspecification and thus bias – and parsimony – to stabilize estimates and thus minimize variance – can be challenging. The

Least Angle Shrinkage and Selection Operator, LASSO, is a regularized regression that can perform both variable selection and parameter estimation (Tibshirani, 1996). A wide range of applications have demonstrated that LASSO is effective in preventing model over-fitting by automatically selecting more accurate and parsimonious models. Kamarianakis, Shen and Wynter (2012) found success with LASSO in predicting average traffic speed in the presence of severe multi-collinearity due to aggregated area-level regressors. Kohannim, Hibar, Stein, Jahanshad, Hua, Rajagopalan, Toga, Jack Jr, Weiner, de Zubicaray and McMahon (2012) applied LASSO regression to identify subsets of high-dimensional and correlated single nucleotide polymorphisms (SNPs) that are related to brain structure measures. In a review of challenges in ecological analysis with collinear covariates, Dormann, Elith, Bacher, Buchmann, Carl, Carre, Marquez, Gruber, Lafourcade, Leitao and Mnkemller (2013) found that LASSO is one of the methods to consistently produce low root-mean-square-errors. In the fields of genetics and finance, LASSO has been used effectively in prediction modeling with hundreds or thousands of predictors (Wu, Chen, Hastie, Sobel and Lange, 2009).

There is a literature that considers stabilizing forms of traditional calibration. Park and Yang (2008) considered a ridge regression form of a generalized regression estimator that used a penalty term to stabilize the calibration estimators, proving design consistency and showing reduction in variance in simulation studies. Goga, Muhammad-Shehzad and Vanheuverzwyn (2011) and Cardot, Goga and Shehzad (2017) considered calibration to principle components of population totals rather than the population totals themselves, allowing large numbers of auxiliary variables to be collapsed into a manageable subset. Perhaps most relevant to this work, McConville (2011) and McConville, Breidt, Lee and Moisen (2017) developed, again under traditional calibration, the theoretical framework to show approximate design unbiasedness and consistency of LASSO calibration estimator of a total, given LASSO regression parameter estimates. Although model-assisted calibration with LASSO holds great promise in constructing a set of weights that can result in small RMSE of weighted estimates for an outcome variable in a non-probability sample, there is no theoretical framework established for the bias and consistency properties of model-assisted LASSO calibration estimators for non-probability sample.

Thus the main objectives of this article are:

(1) Develop the theoretical framework for model-assisted calibration with LASSO for both continuous and binary outcome variables: derive the point estimate of the total, its asymptotic expectation, and asymptotic theoretical variance estimate.

(2) Investigate relative performances, in terms of root-mean-square-error, of LASSO calibration to traditional calibration under different outcome types, sampling schemes, sample sizes, and calibration variable covariance structures.

While our development of the asymptotic theory assumes known design weights, a key finding is that LASSO calibration yields consistent estimators of a population total regardless of whether the design weights are correctly specified as long as the regression model includes all superpopulation parameters as a subset of the parameters in the model. Hence, we focus estimation in the simulation studies in the non-probability-based setting, where initial design weights taken to be the same as those for simple-random-sampling (SRS), $d_i = N/n$ for population and sample sizes N and n, regardless of how the samples are formed (which in practice would be unknown). We also apply LASSO calibration to estimation of the total number of adults diagnosed with cancer in the US population, using data on cancer incidence from the 2013 National Health Interview Survey (NHIS) and auxiliary population data from the US Census American Communities Survey, ignoring sample design weight to approximate a non-probability sample and comparing results to the fully-weighted (representative) estimates.

The organization of this article is as follows. Section 2 provides the definition and notations for calibration and LASSO regression. Section 3 develops the LASSO calibration estimator of population total, its model expectation, and asymptotic variances. Section 4 describes the simulation and results for evaluating the root-mean-square-error and variance estimates of the LASSO-calibrated estimator. Section 5 considers the NHIS example. We conclude with Section 6 summarizing the findings.

2 Calibration

2.1 Traditional calibration

For an analytical sample s_A (the sample which requires weight calibration) of size n drawn from sample design $\mathcal{A}$ with design weights $\underset{n\times 1}{\mathbf{d}}$, and the diagonal matrix of design weights $\mathbf{D}$, calibrated weights $\underset{n\times 1}{\mathbf{w}}$ minimize a distance measure

$$E_{\mathcal{A}}\left[\sum_{i\in s_A} g(w_i,\, d_i)\Big/q_i\right] \tag{2.1}$$

under the constraint:

$$\sum_{i\in s_A} w_i\mathbf{x}_i^T = \mathbf{T}^{\mathbf{X}} \tag{2.2}$$

where $E_{\mathcal{A}}$ is expectation with respect to the analytic (probability) design, $g(w_i,\, d_i)$ is a differentiable function with respect to w_i, strictly convex on an interval containing d_i, and $g(d_i,\, d_i) = 0$, and where $\mathbf{T}^X$ is a row vector of known population totals of sample calibration variables $\mathbf{X}$ (Deville and Särndal, 1992). The constant q_i is independent of design weight d_i. The commonly used generalized regression (GREG) estimator uses the chi-square distance: $g(w_i,\, d_i) = (w_i - d_i)^2/d_i$ with $q_i = 1$. Under this distance measure:

$$\mathbf{w}^{\text{GREG}} = \mathbf{d} + \mathbf{DX}\left(\mathbf{X}^T\mathbf{DX}\right)^{-1}\left(\mathbf{T}^{\mathbf{X}} - \mathbf{d}^T\mathbf{X}\right)^T. \tag{2.3}$$

The estimate of population total of outcome $\mathbf{y}$ is based on calibrated weights:

$$\begin{aligned}\hat{T}_y^{\text{GREG}} &= \mathbf{w}^{(\text{GREG})T}\mathbf{y}\\ &= \mathbf{d}^T\mathbf{y} + \left(\mathbf{T}^{\mathbf{X}} - \mathbf{d}^T\mathbf{X}\right)\left(\mathbf{X}^T\mathbf{DX}\right)^{-1}\mathbf{X}^T\mathbf{Dy}\\ &= \hat{T}_y^{\text{HT}} + \left(\mathbf{T}^{\mathbf{X}} - \mathbf{d}^T\mathbf{X}\right)\hat{\boldsymbol{\beta}}\end{aligned} \tag{2.4}$$

where $\hat{T}_y^{\text{HT}} = \sum_{i\in s_A} d_i y_i$ is the standard (weighted) design-based estimator, $\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^T\mathbf{DX}\right)^{-1}\mathbf{X}^T\mathbf{Dy}$ is the weighted least squares estimate of the linear regression $E_\xi\left[y_i\,|\,\mathbf{x}_i,\, \boldsymbol{\beta}\right] = \mathbf{x}_i^T\boldsymbol{\beta}$, given weights $\mathbf{D}$. (This corresponds to the poststratified estimator when $\mathbf{X}$ consists entirely of cell totals for categorical variables.)

The calibrated weights defined in equation (2.3) do not rely on any outcome variable. Thus the same set of weights can be applied to all variables in the survey. Note that GREG assumes a working model that is linear. Although $\hat{T}_y^{\text{GREG}}$ is asymptotically design-unbiased for T_y, when the relationship between $\mathbf{y}$ and $\mathbf{X}$ is non-linear, such as in the case when $\mathbf{y}$ is binary, the design variance of $\hat{T}_y^{\text{GREG}}$ can be larger than the design variance $\hat{T}_y^{\text{HT}}$.

2.2 Model-assisted calibration

Model-assisted calibration estimators can have significant advantage over $\hat{T}_y^{\text{GREG}}$ because model-assisted calibration allows for non-linear models to assist in the construction of calibrated weights. In model-assisted calibration, we assume a relationship between an outcome $\mathbf{y}$ and $\mathbf{X}$ through first two moments (Wu and Sitter, 2001):

$$E_\xi\left(y_i \,|\, \mathbf{x}_i\right) = \mu\left(\mathbf{x}_i, \boldsymbol{\beta}\right), V_\xi\left(y_i \,|\, \mathbf{x}_i\right) = \nu_i^2 \sigma^2 \tag{2.5}$$

where $\boldsymbol{\beta} = \left(\beta_1, \ldots, \beta_p\right)^T$ and σ are unknown superpopulation parameters, $\mu\left(x_i, \boldsymbol{\beta}\right)$ is a known function of $\mathbf{x}_i$ and $\boldsymbol{\beta}$, and ν_i is a known function of $\mathbf{x}_i$ or $\mu\left(\mathbf{x}_i, \boldsymbol{\beta}\right)$. E_ξ and V_ξ are expectation and variance with respect to the model ξ. Let $\mathbf{B}$ be the finite population (or census) estimate of $\boldsymbol{\beta}$ (i.e., the quasilikelihood estimator of $\boldsymbol{\beta}$ based on the entire finite population), and $\hat{\mu}_i = \mu\left(\mathbf{x}_i, \hat{\mathbf{B}}\right)$, where $\hat{\mathbf{B}}$ is the sample estimate of $\mathbf{B}$. The model-assisted calibrated weights $\mathbf{w}$ then minimize a distance measure $E_{\mathcal{A}}\left[\sum_{i \in s_A} g\left(w_i, d_i\right) / q_i\right]$ under the constraints $\sum_{i \in s_A} w_i = N$ and $\sum_{i \in s_A} w_i \hat{\mu}_i = \sum_i^N \hat{\mu}_i$. The main conceptual difference between traditional calibration and model-assisted calibration is that in model-assisted calibration, the constraints are based on two quantities: (1) population size, and (2) population total of predicted values $\hat{\mu}_i$. In traditional calibration, the constraint is a vector of population totals of $\mathbf{X}$ (see equation (2.2)). Under chi-square distance measure with $q_i = 1$, the model-assisted calibrated weights are:

$$\mathbf{w}^{\text{MC}} = \mathbf{d} + \mathbf{DM}\left(\mathbf{M}^T \mathbf{DM}\right)^{-1}\left(\mathbf{T}^M - \mathbf{d}^T \mathbf{M}\right)^T \tag{2.6}$$

where $\mathbf{T}^M = \left[N, \sum_i^N \hat{\mu}_i\right]$ and $\mathbf{M} = \left[\mathbf{d}, \left(\hat{\mu}_i\right)_{i \in s_A}\right]$. (In the non-probability setting the vector of design weights $\mathbf{d}$ can be replaced with $(N/n)\,\mathbf{1}$.) The estimate for the population total based on model-assisted calibrated weights is then:

$$\begin{aligned} \hat{T}_y^{\text{MC}} &= \left(\mathbf{w}^{\text{MC}}\right)^T \mathbf{y} \\ &= \mathbf{d}^T \mathbf{y} + \left(\mathbf{T}^{\mathbf{M}} - \mathbf{d}^T \mathbf{M}\right)\left(\mathbf{X}^T \mathbf{DX}\right)^{-1} \mathbf{X}^T \mathbf{Dy} \\ &= \hat{T}_y^{\text{HT}} + \left(\sum_i^N \hat{\mu}_i - \sum_{i \in s_A} d_i \hat{\mu}_i\right) \hat{B}^{\text{MC}} \end{aligned} \tag{2.7}$$

where $\hat{B}^{\text{MC}}$ is the calibration slope to satisfy the calibration constraints (different from the model parameter estimates $\hat{\mathbf{B}}$) :

$$\hat{B}^{\text{MC}} = \frac{\sum_{i \in s_A} d_i \left(\hat{\mu}_i - \bar{\hat{\mu}}\right)\left(y_i - \bar{y}\right)}{\sum_{i \in s_A} d_i \left(\hat{\mu}_i - \bar{\hat{\mu}}\right)^2}, \ \bar{\hat{\mu}} = \sum_{i \in s_A} d_i \hat{\mu}_i \Big/ \sum_{i \in s_A} d_i, \ \bar{y} = \sum_{i \in s_A} d_i y_i \Big/ \sum_{i \in s_A} d_i .$$

Unbiasedness and small variances of $\hat{T}_y^{\text{MC}}$ both rely on how well the $\hat{\mu}_i$ approximates the true expected value of y_i.

3 Model selection and robust calibration using adaptive LASSO

3.1 Adaptive LASSO background

3.1.1 Definition and parameters

The adaptive LASSO regression coefficients are obtained by solving a penalized regression equation. For linear adaptive LASSO regression (Zou, 2006):

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}}\left(\sum_{i \in s_A}\left(y_i - \mathbf{x}_i^T \boldsymbol{\beta}\right)^2 + \lambda_n \sum_{j=1}^{p} \alpha_j^{\gamma} \,|\, \beta_j \,|\right) \tag{3.1}$$

where α_j^{γ} is an adjustable weight and λ_n is a penalty used to optimize a model fit measure. Similarly for logistic adaptive LASSO:

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}}\left(\sum_{i \in s_A}\left[-y_i\left(\mathbf{x}_i^T \beta\right) + \log\left(1 + \exp\left(\mathbf{x}_i^T \boldsymbol{\beta}\right)\right)\right] + \lambda_n \sum_{j=1}^{p} \alpha_j^{\gamma} \,|\, \beta_j \,|\right). \tag{3.2}$$

Given λ_n and $\gamma,$ we can calculate $\hat{\boldsymbol{\beta}}$ through iterative procedures. The R package *glmnet* will compute both the linear and logistic adaptive LASSO (Friedman, Hastie and Tibshirani, 2010).

The role of the weight parameter, $\alpha_j,$ is to prevent LASSO from selecting covariates with large effect sizes in favor of lowering prediction error when the sample size is small. Thus the weights are inversely proportional to effect sizes of regression parameters: $\alpha_j \propto 1/|\beta_j|.$ A common choice of α_j is $1/|\hat{\beta}_j^{\text{MLE}}|,$ where $\hat{\beta}_j^{\text{MLE}}$ is the maximum likelihood estimate of $\beta_j.$ The power of the weight parameter, $\gamma,$ is a constant greater than 0 that interacts with α_j to control LASSO from selecting or excluding parameters. For example, if we still want LASSO to favor large effect covariates when the sample size is small, we should set γ small. If we want to de-emphasize effect sizes further, we should set γ large.

3.1.2 Oracle property

An important concept in measuring the performance of a model selection and estimation method is called the “oracle property”. The optimal method selects the correct variables and provides unbiased estimates of selected parameters. Suppose the parameters in a full regression model have both zero and non-zero components. Without loss of generality, let the first p be non-zero and the last q zero:

$$\boldsymbol{\beta}^F = \begin{pmatrix} \boldsymbol{\beta}_{(p \times 1)}^{(1)} \\ \boldsymbol{\beta}_{(q \times 1)}^{(2)} = \mathbf{0} \end{pmatrix}.$$

A regression model has the oracle property if it satisfies the following conditions (Fan and Li, 2001):

- The probability of estimating 0 for zero-valued parameters tends to one: $\Pr\left(\hat{\boldsymbol{\beta}}^{(2)} = \mathbf{0}\right) \rightarrow 1$ as $n \rightarrow \infty.$

- The estimates of non-zero parameters are as good as if the true sub-model is known: $\sqrt{n}\left(\hat{\boldsymbol{\beta}}^{(1)} - \boldsymbol{\beta}^{(1)}\right) \rightarrow N(\mathbf{0}, \mathbf{C})$ where $\mathbf{C} = \Sigma(\boldsymbol{\beta}^{(1)})$ is the covariance matrix of $\boldsymbol{\beta}^{(1)}$ under linear model, and $\mathbf{C} = I^{-1}(\boldsymbol{\beta}^{(1)})$ is the inverse of the Fisher information matrix of $\boldsymbol{\beta}^{(1)}$ under the generalized linear model.

For finite-population inference, suppose ν indexes a population with size N_ν, let $\mathbf{B}_\nu$ be the quasilikelihood estimates of $\boldsymbol{\beta}$ in population ν, and $\hat{\mathbf{B}}_\nu$ is the estimate of $\mathbf{B}_\nu$ based on a sample with size $n_\nu \leq N_\nu$. We assume that $N_\nu \rightarrow \infty, n_\nu \rightarrow \infty,$ and $n_\nu / N_\nu \rightarrow 0$ as $\nu \rightarrow \infty$. The finite-population equivalent of the oracle property is then:

$$\begin{aligned} \Pr\left(\hat{\mathbf{B}}_\nu^{(2)} = \mathbf{0}\right) &\rightarrow 1 \\ \sqrt{n_\nu}\left(\hat{\mathbf{B}}_\nu^{(1)} - \mathbf{B}_\nu^{(1)}\right) &\rightarrow N_\nu(\mathbf{0}, \mathbf{C}_\nu) \\ \mathbf{B}_\nu &\rightarrow \boldsymbol{\beta} \\ \text{as } &\nu \rightarrow \infty \end{aligned}$$

where $\mathbf{C}_\nu = \Sigma\left(\mathbf{B}_\nu^{(1)}\right)$ is the covariance matrix of $\mathbf{B}_\nu^{(1)}$ if the model is linear, and $\mathbf{C}_\nu = I^{-1}\left(\mathbf{B}_\nu^{(1)}\right)$ is the inverse of Fisher information matrix of $\mathbf{B}_\nu^{(1)}$ under the generalized linear model.

Zou (2006) has shown that if $\lambda_n \Big/ \left(\sqrt{n} \Big/ \left(\sqrt{n}\right)^\gamma\right) \rightarrow \infty$ and $\lambda_n / \sqrt{n} \rightarrow 0$, then the adaptive LASSO satisfies the oracle property. The conditions require that λ_n grow at least at the rate of $\sqrt{n} \Big/ \left(\sqrt{n}\right)^\gamma$, but not faster than $\sqrt{n}$. The choice of λ_n and γ, and R code for implementing it, are discussed in the Appendix.

3.2 LASSO calibration

This section derives the analytical formula for a LASSO estimator of total, its model expectation, and estimators of the asymptotic design variance. We make the following assumptions:

1. The samples are drawn from a single-stage sample design $\mathcal{A}$, allowing for unequal probabilities of selection. The selection probability for unit i is denoted by π_i^A, and the joint selection probability of units i and j is denoted by π_{ij}^A. We denote the design weight for unit i by $d_i^A = 1/\pi_i^A$, the vector of design weights by $\mathbf{d}^A$, and the diagonal matrix of design weights by $\mathbf{D}^A$.
2. Population-level auxiliary data are known, denoted by $\mathbf{X} = (\mathbf{x}_i^T), i = 1, \ldots, N$.
3. A superpopulation model is assumed, as is described in Section 2.2:
$$\begin{aligned} E_\xi(y_i \mid \mathbf{x}_i) &= \mu(\mathbf{x}_i, \boldsymbol{\beta}) \\ V_\xi(y_i \mid \mathbf{x}_i) &= \nu_i^2 \sigma^2. \end{aligned}$$
4. The true superpopulation parameters are a subset of the full regression model for LASSO:
$$\boldsymbol{\beta}^F = \begin{pmatrix} \boldsymbol{\beta}_{(p\times 1)} \\ \boldsymbol{\beta}_{(q\times 1)}^{(2)} \end{pmatrix}.$$

5. The full-range of $\mathbf{X}$ in the population has non-zero probability of being observed in the analytical sample.

3.2.1 Point estimate: $\hat{T}_y^{\text{LASSO}}$

The LASSO calibration estimate of total can be obtained following the steps:

1. Obtain LASSO regression coefficients $\hat{\mathbf{B}}$ as described in the Appendix.
2. Use $\hat{\mathbf{B}}$ to calculate $\hat{\mu}_i = \mu\left(\mathbf{x}_i, \hat{\mathbf{B}}\right)$ in the population.
3. Define $\mathbf{T}^M = \left(N, \sum_i^N \hat{\mu}_i\right)$ and $\mathbf{M} = \left[\mathbf{d}^A, \sum_{i \in s_A} \hat{\mu}_i\right]$, under chi-square distance measure with $q_i = 1$:

$$\mathbf{w}^{\text{LASSO}} = \mathbf{d}^A + \mathbf{D}^A \mathbf{M} \left(\mathbf{M}^T \mathbf{D}^A \mathbf{M}\right)^{-1} \left(\mathbf{T}^M - \left(\mathbf{d}^A\right)^T \mathbf{M}\right)^T. \tag{3.3}$$

4. Determine the LASSO calibration estimator of total:

$$\begin{aligned} \hat{T}_y^{\text{LASSO}} &= \left(\mathbf{w}^{\text{LASSO}}\right)^T \mathbf{y} \\ &= \left(\mathbf{d}^A\right)^T \mathbf{y} + \left(\mathbf{T}^{\mathbf{M}} - \left(\mathbf{d}^A\right)^T \mathbf{M}\right) \left(\mathbf{X}^T \mathbf{D}^A \mathbf{X}\right)^{-1} \mathbf{X}^T \mathbf{D}^A \mathbf{y} \\ &= \left(\mathbf{d}^A\right)^T \mathbf{y} + \left(\sum_i^N \hat{\mu}_i - \sum_{i \in s_A} d_i^A \hat{\mu}_i\right) \hat{B}^{\text{MC}} \end{aligned} \tag{3.4}$$

where $\hat{B}^{\text{MC}}$ is the calibration slope to satisfy the calibration constraints:

$$\hat{B}^{\text{MC}} = \frac{\sum_{i \in s_A} d_i^A \left(\hat{\mu}_i - \hat{\bar{\mu}}\right)\left(y_i - \bar{y}\right)}{\sum_{i \in s_A} d_i^A \left(\hat{\mu}_i - \hat{\bar{\mu}}\right)^2}, \quad \hat{\bar{\mu}} = \sum_{i \in s_A} d_i^A \hat{\mu}_i \Big/ \sum_{i \in s_A} d_i^A, \quad \bar{y} = \sum_{i \in s_A} d_i^A y_i \Big/ \sum_{i \in s_A} d_i^A.$$

3.2.2 Asymptotic behavior of $\hat{T}_y^{\text{LASSO}}$

Wu and Sitter (2001) established the conditions to derive an asymptotic model-assisted calibration estimator. We state the conditions here with slight modification in notations to be consistent with the current research. Let $\boldsymbol{\beta}$ be the true superpopulation parameter for the model defined in equation (2.5), and $\mathbf{B}$ be the finite-population quasilikelihood estimator of $\boldsymbol{\beta}$. The following conditions are used for deriving LASSO calibration asymptotic estimator of total:

1. $\hat{\mathbf{B}} = \mathbf{B} + O_p\left(1/\sqrt{n}\right)$, $\mathbf{B}$ is the finite-population regression slope of $\boldsymbol{\beta}$, $\mathbf{B} \to \boldsymbol{\beta}$.
2. For each $\mathbf{x}_i$, $\partial \mu\left(\mathbf{x}_i, \mathbf{t}\right)/\partial \mathbf{t}$ is continuous in $\mathbf{t}$, and $\max_i \left| \partial \mu\left(\mathbf{x}_i, \mathbf{t}\right)/\partial \mathbf{t} \right| \le h\left(\mathbf{x}_i, \boldsymbol{\beta}\right)$ for $\mathbf{t}$ in a neighborhood of $\boldsymbol{\beta}$, and $N^{-1} \sum_{i \in U} h\left(\mathbf{x}_i, \boldsymbol{\beta}\right) = O(1)$.
3. For each $\mathbf{x}_i$, $\partial^2 \mu\left(\mathbf{x}_i, \mathbf{t}\right)/\partial \mathbf{t} \partial \mathbf{t}^T$ is continuous in $\mathbf{t}$, and $\max_{j,k} \left| \partial^2 \mu\left(\mathbf{x}_i, \mathbf{t}\right)/\partial t_j \partial t_k \right| \le k\left(\mathbf{x}_i, \boldsymbol{\beta}\right)$ for $\mathbf{t}$ in a neighborhood of $\boldsymbol{\beta}$, and $N^{-1} \sum_{i \in U} k\left(\mathbf{x}_i, \boldsymbol{\beta}\right) = O(1)$.
4. The Horvitz-Thompson (HT) estimators of certain population means are asymptotically normally distributed (Fuller, 2009; pages 47-57).
5. $\lambda_n \Big/ \left(\sqrt{n} \big/ \left(\sqrt{n}\right)^\gamma\right) \to \infty$ and $\lambda_n \big/ \sqrt{n} \to 0$.

Lemma 1: Assume that superpopulation model (2.5) holds. Let $\mathbf{B}$ be the finite-population quasilikelihood estimate of $\boldsymbol{\beta}$, $\mathbf{B} \to \boldsymbol{\beta}$. Under conditions (1)-(5), the model-assisted asymptotic estimator of population total is:

$$\hat{T}_y^{\mathrm{MC}} = \sum_{i \in s_A} d_i^A \left(y_i - \mu_i B^{\mathrm{MC}}\right) + \sum_{i=1}^{N} \mu_i B^{\mathrm{MC}} + o_p\left(\frac{N}{\sqrt{n}}\right) \tag{3.5}$$

where

$$\mu_i = \mu\left(\mathbf{x}_i, \mathbf{B}\right)$$

$$B^{\mathrm{MC}} = \frac{\sum_{i=1}^{N}\left(\mu_i - \bar{\mu}\right)\left(y_i - \bar{y}\right)}{\sum_{i=1}^{N}\left(\mu_i - \bar{\mu}\right)^2}.$$

Proof. See Appendix.

Given Lemma 1, we derive $\hat{T}_y^{\mathrm{LASSO}}$ the asymptotic LASSO estimator of total in Theorem 1. We show $\hat{T}_y^{\mathrm{LASSO}}$ is model unbiased for the population total in Theorem 2. Finally, Theorem 3 determines variance estimates for the LASSO calibration estimator of a total.

Theorem 1: Suppose the parameters in a full regression model have both zero and non-zero components. Without loss of generality, let the first p be non-zero and the last q be zero:

$$\boldsymbol{\beta}^F = \begin{pmatrix} \boldsymbol{\beta}_{(p \times 1)}^{(1)} \\ \boldsymbol{\beta}_{(q \times 1)}^{(2)} \end{pmatrix}, \quad \boldsymbol{\beta}^{(1)} = \boldsymbol{\beta} \quad \text{and} \quad \boldsymbol{\beta}^{(2)} = \mathbf{0}_{(q \times 1)},$$

under conditions (1)-(5), the asymptotic LASSO calibration estimator of total is:

$$\hat{T}_y^{\mathrm{LASSO}} = \sum\nolimits_{i \in s_A} d_i^A \left(y_i - \mu_i B^{\mathrm{MC}}\right) + \sum\nolimits_{i=1}^{N} \mu_i B^{\mathrm{MC}} + o_p\left(\frac{N}{\sqrt{n}}\right).$$

Proof. See Appendix.

Theorem 2: $\hat{T}_y^{\mathrm{LASSO}}$ is model-unbiased, that is $E_\xi\left(\hat{T}_y^{\mathrm{LASSO}}\right) = T$.

Proof. See Appendix.

Thus, as long as LASSO regression parameters include the superpopulation parameters, $\hat{T}_y^{\mathrm{LASSO}}$ is model-unbiased regardless of design weights. (Note that this is a quality that $\hat{T}_y^{\mathrm{GREG}}$ shares with $\hat{T}_y^{\mathrm{LASSO}}$. However, $\hat{T}_y^{\mathrm{LASSO}}$ can assume models with much larger numbers of covariates than $\hat{T}_y^{\mathrm{GREG}}$.) This property is essential in non-probability samples, where there are no initial design weights to guarantee unbiasedness.

Theorem 3: The estimator of the asymptotic variance of $\hat{T}_y^{\mathrm{LASSO}}$ is given by

$$\begin{aligned} v_{\mathcal{A}}\left(\hat{T}_y^{\mathrm{LASSO}}\right) = & \sum_{i \in s_A}\left(\frac{y_i - \hat{\mu}_i \hat{B}^{\mathrm{MC}}}{\pi_i}\right)^2 \left(1 - \pi_i\right) \\ & + \sum_{i \in s_A}\sum_{j \neq i} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_{ij}} \frac{\left(y_i - \hat{\mu}_i \hat{B}^{\mathrm{MC}}\right)}{\pi_i} \frac{\left(y_j - \hat{\mu}_j \hat{B}^{\mathrm{MC}}\right)}{\pi_j}. \end{aligned} \tag{3.6}$$

Proof. The theoretical design variance of the LASSO estimator is

$$\begin{aligned}
V_{\mathcal{A}}\left(\hat{T}_y^{\text{LASSO}}\right) &= V_{\mathcal{A}}\left(\sum_{i\in s_A} d_i^A\left(y_i - \mu_i B^{\text{MC}}\right) + \sum_{i=1}^{N} \mu_i B^{\text{MC}}\right)\\
&= V_{\mathcal{A}}\left(\sum_{i\in s_A} d_i^A\left(y_i - \mu_i B^{\text{MC}}\right)\right)\\
&= \sum_{i\in U}\left(\frac{y_i - \mu_i B^{\text{MC}}}{\pi_i}\right)^2 \pi_i\left(1-\pi_i\right)\\
&\quad + \sum_{i\in U}\sum_{j\neq i}\left(\pi_{ij} - \pi_i\pi_j\right)\frac{\left(y_i - \mu_i B^{\text{MC}}\right)}{\pi_i}\frac{\left(y_j - \mu_j B^{\text{MC}}\right)}{\pi_j}
\end{aligned} \tag{3.7}$$

which follows from equation (3.30) derived for the variance of traditional LASSO calibration estimator of total in McConville (2011). Equation (3.6) then follows from replacing estimates for population quantities.

An alternative variance estimate, suggested by Särndal, Swensson and Wretman (1989), multiplies $\left(y_i - \hat{\mu}_i\hat{B}^{\text{MC}}\right)$ by g – weights, which are the ratios of calibrated weights to the original design weights:

$$\mathbf{g} = \mathbf{1}_{(n\times 1)} + \mathbf{M}\left(\mathbf{M}^T\mathbf{D}^A\mathbf{M}\right)^{-1}\left(\mathbf{T}^M - \left(\mathbf{d}^A\right)^T\mathbf{M}\right)^T$$

$$\begin{aligned}
v.g_{\mathcal{A}}\left(\hat{T}_y^{\text{LASSO}}\right) &= \sum_{i\in s_A}\left(\frac{g_i\left(y_i - \hat{\mu}_i\hat{B}^{\text{MC}}\right)}{\pi_i}\right)^2\left(1-\pi_i\right)\\
&\quad + \sum_{i\in s_A}\sum_{j\neq i}\frac{\pi_{ij} - \pi_i\pi_j}{\pi_{ij}}\frac{g_i\left(y_i - \hat{\mu}_i\hat{B}^{\text{MC}}\right)}{\pi_i}\frac{g_j\left(y_j - \hat{\mu}_j\hat{B}^{\text{MC}}\right)}{\pi_j}.
\end{aligned} \tag{3.8}$$

To simplify notations, we refer to $v_{\mathcal{A}}\left(\hat{T}_y^{\text{LASSO}}\right)$ as v^{LASSO} and $v.g_{\mathcal{A}}\left(\hat{T}_y^{\text{LASSO}}\right)$ as v_g^{LASSO}.

4 Simulation study

We design a simulation to evaluate the finite sample properties of $\hat{T}_y^{\text{LASSO}}$ and the asymptotic variance estimates of $\hat{T}_y^{\text{LASSO}}$, v^{LASSO} and v_g^{LASSO}. We also consider a naive bootstrap estimator $v_{\text{boot}}^{\text{LASSO}}$, obtained by drawing 500 samples with replacement from each simulation sample, as an alternative variance estimator of $\hat{T}_y^{\text{LASSO}}$.

To simulate non-probability samples, we generate samples with unequal selection probabilities, but set design weights to $\mathbf{d}^A = N/n$. We also consider $\hat{T}_y^{\text{GREG}}$ (traditional calibration estimator) and $\hat{T}_y^{\text{HT}}$ (pure design-based Horvitz-Thompson estimator). Because $\hat{T}_y^{\text{LASSO}}$ performs both variable selection and estimation, we implement a backward stepwise selection to select the working model for GREG. Although there is no theoretical justification for using stepwise variable selection, Skinner and Silva (1997) have shown that given two auxiliary variables, a stepwise procedure can result in improved efficiency of GREG estimator. We are interested in knowing the performance of each estimator under (1) populations with different signal-to-noise-ratios (SNR), (2) independent, informative, and biased sampling schemes, and (3) small and large sample sizes. The signal-to-noise ratio is calculated according to definitions in Czanner, Sarma, Eden and Brown (2008). We set two levels of correlations (low/high) between covariates, crossed with two levels of effect sizes (low/high) of the covariates. We set the low/high and high/low populations to have the same SNR in order to understand the influence of correlation and effect size on estimator's performance given the same SNR. Three sampling schemes are used to draw samples: simple-random-sampling without replacement, SRS, Poisson sampling with selection probabilities proportional to covariates, $\mathrm{POI}(\mathrm{X})$, and Poisson sampling with selection probabilities proportional to covariates and the outcome, $\mathrm{POI}(\mathrm{X+Y})$. $\mathrm{POI}(\mathrm{X+Y})$ sampling simulates self-selection bias of non-probability samples, where the propensity of a respondent to participate in a study relates to the analysis variable. We consider two sample sizes: 250 and 1,000. Thus we have a total of $2 \times 2 \times 3 \times 2 = 24$ experimental groups.

4.1 Population

To create collinearity among covariates, we follow an auto-decay correlation structure commonly used in LASSO-related simulations (Tibshirani, 1996): $\mathrm{cor}(X_i, X_j) = \rho^{|i-j|}$, $i = 1, \ldots, p$. We generate a population of size $N = 100{,}000$ from a multivariate normal distribution with mean $\mathbf{0}_{(p\times 1)}$ and covariance Σ^{ρ}, $p = 40$. The continuous outcome variable is generated by the regression model:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_{40} x_{i40} + N(0,3).$$

The binary outcome variable is generated by the logistic regression model:

$$\begin{aligned}\phi_i &= \mathrm{expit}(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_{40} x_{i40}), \qquad \mathrm{expit}(u) = (1+\exp(u))^{-1}\\ y_i &= \mathrm{bernoulli}(\phi_i).\end{aligned}$$

We set $\rho = 0.15$ for low correlation population, and $\rho = 0.73$ for high correlation population. For both continuous and binary outcome variables:

$$\begin{aligned}\text{Low effect-size } \boldsymbol{\beta}^{(1)} &:= \beta_{12} \ldots \beta_{19}, \beta_{32} \ldots \beta_{39} = 0.45\\ \text{High effect-size } \boldsymbol{\beta}^{(1)} &:= \beta_{12} \ldots \beta_{19}, \beta_{32} \ldots \beta_{39} = 0.74.\end{aligned}$$

For continuous $\mathbf{y}$: $\beta_0 = 1$, for binary $\mathbf{y}$: $\beta_0 = 0.4$. The rest of $\beta_i = 0$. Out of 41 regression parameters, 16 are non-zero and 25 are zero.

4.2 Sampling schemes

Three sampling schemes are used to generate the sample:

1. Simple-Random-Sampling (SRS): selection probabilities $= n/N$.
2. Poisson sampling with probabilities proportional to $\mathbf{X}$, $\mathrm{POI}(\mathrm{X})$

$$\begin{cases}\text{continuous } \mathbf{y}: \ \pi_i \propto 0.4 + 0.4x_{i5} + 0.4x_{i15} + 0.4x_{i25} + 0.4x_{i35}\\ \quad \text{binary } \mathbf{y}: \ \mathrm{logit}(\pi_i) = 0.4 + 0.4x_{i5} + 0.4x_{i15} + 0.4x_{i25} + 0.4x_{i35}.\end{cases}$$

3. Poisson sampling with probabilities proportional to $\mathbf{X}$ and $\mathbf{y}$, $\mathrm{POI}(\mathrm{X{+}Y})$

$$\begin{cases} \text{continuous } \ \mathbf{y}: \ \pi_i \propto 0.4 + 0.4x_{i5} + 0.4x_{i15} + 0.4x_{i25} + 0.4x_{i35} + 0.5y_i \\ \quad\ \text{binary } \ \mathbf{y}: \ \operatorname{logit}(\pi_i) \propto 0.4 + 0.4x_{i5} + 0.4x_{i15} + 0.4x_{i25} + 0.4x_{i35} + y_i. \end{cases}$$

4.3 Evaluation metrics

We evaluate empirical bias, variance, and RMSE for each estimator of total. We evaluate the asymptotic variance estimates and bootstrap variance estimates by their 95% nominal coverage and %bias relative to empirical variance. We use the normal approximation to generate confidence intervals. We calculate %bias as $\%\text{bias} = 100[v - \operatorname{var}(\hat{T}_y^{\text{LASSO}})] \big/ \operatorname{var}(\hat{T}_y^{\text{LASSO}})$, where $\operatorname{var}(\hat{T}_y^{\text{LASSO}})$ is the empirical variance obtained from the simulation samples.

4.4 Simulation results

The simulation results are based on $S = 1{,}000$ simulated samples per each experimental group. Table 4.1 lists the numerical results of bias, variance, and root-mean-square-error of each estimator under different experimental designs for estimating the total of a continuous outcome variable. Table 4.2 lists the numerical results for estimating the total of a binary outcome variable.

4.4.1 Root mean square error

Under SRS, all estimators are unbiased, and LASSO and GREG perform approximately equally well relative to HT. $\mathrm{POI}(\mathrm{X})$ and $\mathrm{POI}(\mathrm{X{+}Y})$ induce biased samples by selecting cases with larger covariate values with higher probabilities. Under $\mathrm{POI}(\mathrm{X{+}Y})$, the selection also favors cases with larger outcome values. The absolute bias of LASSO decreases relative to GREG as SNR increases. This improvement is more dramatic in the binary case than the continuous case, especially for $\mathrm{POI}(\mathrm{X{+}Y})$. In terms of RMSE, LASSO has marginal improvement over GREG for estimating totals of continuous outcome variables. The improvement is slightly noticeable, about 3%, when there are highly correlated predictors in the model. For the binary setting, there is substantial improvement in MSE for LASSO over GREG as SNR increases, with reductions of 20% for the $\mathrm{POI}(\mathrm{X})$ and nearly 50% for the $\mathrm{POI}(\mathrm{X{+}Y})$ setting when SNR is large. In particular, under Low/High and High/Low population types, the SNR is the same, thus the difference in performance between LASSO and GREG is attributed to correlation or effect size. LASSO performs better in both bias and RMSE in High/Low population type, suggesting that LASSO has stronger advantage over GREG when there are highly correlated predictors in the model. This suggests that LASSO has a better variable selection capability in the presence of multicollinearity relative to stepwise variable selection procedure used in GREG.

Table 4.1
Simulation summary for continuous outcome: total, bias, and RMSE × 10^3; variance × 10^6

Population	n	Sampling scheme	HT			GREG			LASSO		
			bias	var	rmse	bias	var	rmse	bias	var	rmse
low/low T = 100.8 SNR = 0.47	250	SRS	0.5	546	23.3	0.9	425	20.6	0.9	428	20.7
		POI(X)	12.4	525	26.0	-0.6	446	21.1	-0.4	441	21.0
		POI(X+Y)	19.4	519	29.9	4.6	443	21.5	4.7	431	21.3
	1,000	SRS	0.2	129	11.4	0.3	94	9.6	0.3	94	9.7
		POI(X)	12.6	129	17.0	-0.1	91	9.5	-0.2	92	9.6
		POI(X+Y)	19.7	128	22.7	4.9	91	10.7	5.0	91	10.7
low/high T = 101.4 SNR = 1.26	250	SRS	0.4	849	29.1	0.9	415	20.4	1.0	417	20.4
		POI(X)	21.1	818	35.6	-1.3	434	20.9	-1.0	432	20.8
		POI(X+Y)	31.7	817	42.7	3.7	427	21.0	4.0	427	21.1
	1,000	SRS	0.0	200	14.1	0.3	94	10.0	0.3	93	9.7
		POI(X)	21.1	199	25.4	-0.1	91	9.6	-0.2	90	9.6
		POI(X+Y)	31.7	196	34.6	4.9	91	10.7	4.8	89	10.6
high/low T = 101.8 SNR = 1.26	250	SRS	0.1	941	30.7	1.0	421	20.6	1.0	399	20.0
		POI(X)	50.2	895	58.5	-0.7	434	20.8	-1.6	402	20.1
		POI(X+Y)	57.8	872	64.9	4.0	435	21.2	3.0	399	20.2
	1,000	SRS	0.0	218	14.8	0.3	94	9.7	0.3	93	9.6
		POI(X)	50.6	210	53.0	-0.1	93	9.7	-0.5	91	9.6
		POI(X+Y)	58.2	209	59.9	4.7	95	10.8	4.2	92	10.5
high/high T = 103.1 SNR = 3.41	250	SRS	-0.4	1,897	43.6	0.8	436	20.9	1.0	407	20.2
		POI(X)	83.3	1,826	93.7	-0.8	435	20.9	-1.5	406	20.2
		POI(X+Y)	96.4	1,779	105.3	3.7	428	21.0	3.0	404	20.3
	1,000	SRS	-0.2	444	21.0	0.3	93	9.7	0.3	93	9.7
		POI(X)	83.6	424	86.1	-0.2	93	9.7	-0.5	91	9.6
		POI(X+Y)	96.9	423	99.0	4.4	94	10.6	4.1	92	10.4

Table 4.2
Simulation summary for binary outcome: total, bias, and RMSE × 10^3; variance × 10^6

Population	n	Sampling scheme	HT			GREG			LASSO		
			bias	var	rmse	bias	var	rmse	bias	var	rmse
low/low T = 56.2 SNR = 0.51	250	SRS	0.0	10.2	3.2	0.0	7.2	2.7	0.0	7.0	2.7
		POI(X)	2.6	10.0	4.1	0.2	8.0	2.8	0.1	7.8	2.8
		POI(X+Y)	4.9	9.8	5.8	2.0	8.1	3.5	1.8	7.8	3.3
	1,000	SRS	-0.0	2.7	1.6	0.0	1.7	1.3	0.0	1.6	1.3
		POI(X)	2.5	2.4	2.9	0.0	1.8	1.3	-0.0	1.7	1.3
		POI(X+Y)	4.7	2.3	5.0	1.8	1.8	2.2	1.6	1.7	2.1
low/high T = 54.4 SNR = 1.10	250	SRS	-0.0	10.8	3.3	0.0	6.1	2.5	0.1	5.4	2.3
		POI(X)	3.0	10.2	4.4	0.1	6.1	2.5	0.1	5.8	2.4
		POI(X+Y)	5.3	9.8	6.2	1.6	6.2	2.9	1.3	5.8	2.8
	1,000	SRS	-0.0	2.7	1.6	0.0	1.3	1.1	0.0	1.1	1.0
		POI(X)	2.9	2.4	3.3	0.0	1.4	1.2	-0.1	1.2	1.1
		POI(X+Y)	5.2	2.2	5.4	1.4	1.4	1.8	1.1	1.2	1.6
high/low T = 54.2 SNR = 1.10	250	SRS	-0.0	10.3	3.2	0.0	5.8	2.4	0.1	4.9	2.2
		POI(X)	6.6	9.6	7.3	0.3	6.2	2.5	-0.2	4.8	2.2
		POI(X+Y)	8.6	9.3	9.1	1.8	6.3	3.1	0.9	4.9	2.4
	1,000	SRS	-0.0	2.5	1.6	0.0	1.2	1.1	0.0	1.0	1.0
		POI(X)	6.6	2.2	6.7	0.2	1.4	1.2	-0.2	1.1	1.1
		POI(X+Y)	8.5	2.1	8.7	1.6	1.4	2.0	1.0	1.0	1.4
high/high T = 52.8 SNR = 2.75	250	SRS	-0.1	10.2	3.1	-0.0	5.2	2.3	0.1	3.8	1.9
		POI(X)	7.1	9.8	7.8	0.3	5.7	2.4	-0.2	3.6	1.9
		POI(X+Y)	9.1	9.4	9.6	1.5	5.7	2.8	0.5	3.7	2.0
	1,000	SRS	-0.1	2.5	1.6	-0.0	1.1	1.0	0.0	0.6	0.8
		POI(X)	7.1	2.2	7.2	0.2	1.3	1.1	-0.2	0.7	0.9
		POI(X+Y)	9.1	2.2	9.2	1.4	1.2	1.8	0.5	0.7	1.0

4.4.2 LASSO variance estimates

Tables 4.3 and 4.4 list the 95% nominal coverage and percent-bias for each of the two asymptotic closed-form variance estimators developed in this research, as well as the naive bootstrap variance estimate of the LASSO calibration estimator.

For continuous outcomes, bootstrap variances have coverages that are consistently close to 95% under SRS and $\mathrm{POI}(\mathrm{X})$ sampling schemes for both sample sizes. Under $\mathrm{POI}(\mathrm{X+Y})$ sampling scheme, there is very modest undercoverage in Table 4.3. The closed-form variances have coverages that are sensitive to both sample size and sampling scheme, with smaller samples tending to undercover, particularly for the $\mathrm{POI}(\mathrm{X+Y})$ sampling scheme. The difference in coverage of variance estimates between small and large sample sizes is expected, since the variance estimates are asymptotic and improve over larger samples. In terms of bias of variance estimators, there is evidence that bias reduces as SNR increases. With the same SNR, both asymptotic closed-form and bootstrap variances have smaller bias given predictors with high correlations relative to predictors with high effect sizes. Closed-form variances tend to underestimate the empirical variance, especially when the sample size is small. Overall, there is very little difference between the two closed-form variance estimates. Bootstrap variance tends to overestimate the empirical variance, but the absolute bias is generally smaller than those of the closed-form variance estimates.

For binary outcomes, both asymptotic closed-form and bootstrap variance estimates are sensitive to sample size, sampling scheme, and SNR. Bootstrap variance coverages are consistently close to 95% under SRS and $\mathrm{POI}(\mathrm{X})$ for both sample sizes and all population types, but coverages range from 75% to 94% under $\mathrm{POI}(\mathrm{X+Y})$. Under $\mathrm{POI}(\mathrm{X+Y})$, the bootstrap variance coverages are better with sample size 250 than with sample size 1,000 when the bias becomes a larger part of the RMSE, and better with high-correlation populations than with low-correlation populations. In terms of coverage, closed-form variances show a similar trend under $\mathrm{POI}(\mathrm{X+Y})$ as bootstrap: better coverage with smaller samples than bigger samples, and better coverage with high-correlation populations than with low-correlation populations. Under SRS and $\mathrm{POI}(\mathrm{X})$, closed-form variance coverage improves as sample size increases. In terms of bias, both bootstrap and closed-form variances have smaller bias with larger sample sizes. Holding sample size fixed, closed-form variance estimates have larger bias as SNR increases. The same trend is not observed in bootstrap variance estimates. Similar to continuous outcome results, closed-form variance tends to underestimate the empirical variance, especially when the sample size is small. Unlike continuous outcome results, there is evidence that the $g-$weighted closed-form variance estimates have better bias-properties than unweighted closed-form variance estimates. The bootstrap variance tends to overestimate the empirical variance. However, the biases are much smaller than for the closed-form variance estimates.

Table 4.3
95% nominal coverage and %bias of variance estimates for LASSO

Continuous outcome			coverage			%bias		
Population	**n**	**scheme**	v^{LASSO}	v_g^{LASSO}	$v_{\text{boot}}^{\text{LASSO}}$	v^{LASSO}	v_g^{LASSO}	$v_{\text{boot}}^{\text{LASSO}}$
low/low	250	SRS	91.7%	91.8%	95.4%	-22.6%	-22.3%	2.9%
		POI(X)	91.2%	91.2%	96.1%	-25.1%	-24.5%	5.7%
		POI(X+Y)	89.6%	89.9%	95.4%	-23.5%	-22.8%	7.9%
	1,000	SRS	93.2%	93.2%	93.8%	-7.3%	-7.2%	-0.3%
		POI(X)	94.0%	93.9%	95.5%	-5.7%	-5.3%	6.6%
		POI(X+Y)	90.0%	90.1%	92.1%	-4.9%	-4.4%	7.9%
low/high	250	SRS	91.5%	91.5%	95.7%	-22.6%	-22.3%	6.2%
		POI(X)	90.9%	91.2%	96.4%	-25.4%	-24.9%	8.8%
		POI(X+Y)	90.0%	90.2%	95.1%	-24.5%	-23.7%	9.9%
	1,000	SRS	93.4%	93.5%	94.3%	-6.6%	-6.5%	-0.1%
		POI(X)	94.1%	94.2%	95.9%	-4.0%	-3.5%	7.6%
		POI(X+Y)	90.7%	90.7%	92.7%	-2.9%	-2.3%	9.6%
high/low	250	SRS	92.3%	92.2%	95.4%	-17.4%	-17.1%	2.0%
		POI(X)	92.5%	92.6%	95.8%	-17.9%	-16.1%	6.4%
		POI(X+Y)	91.2%	91.8%	96.5%	-17.4%	-15.4%	7.1%
	1,000	SRS	93.5%	93.5%	94.4%	-6.5%	-6.4%	-0.9%
		POI(X)	94.1%	94.0%	95.4%	-5.0%	-3.1%	5.7%
		POI(X+Y)	91.9%	92.3%	93.4%	-6.0%	-3.9%	5.0%
high/high	250	SRS	92.3%	92.3%	95.2%	-19.6%	-19.3%	2.2%
		POI(X)	92.0%	92.3%	96.1%	-19.6%	-17.8%	7.4%
		POI(X+Y)	91.2%	91.8%	95.6%	-19.1%	-16.9%	8.3%
	1,000	SRS	93.4%	93.4%	94.5%	-6.5%	-6.4%	-0.7%
		POI(X)	94.0%	94.5%	95.6%	-4.7%	-2.8%	6.7%
		POI(X+Y)	92.2%	92.4%	93.4%	-5.6%	-3.3%	6.1%

Table 4.4
95% nominal coverage and %bias of variance estimates for LASSO

Binary outcome			coverage			%bias		
Population	**n**	**scheme**	v^{LASSO}	v_g^{LASSO}	$v_{\text{boot}}^{\text{LASSO}}$	v^{LASSO}	v_g^{LASSO}	$v_{\text{boot}}^{\text{LASSO}}$
low/low	250	SRS	89.8%	90.0%	95.9%	-28.1%	-27.8%	9.2%
		POI(X)	88.1%	88.6%	96.7%	-37.3%	-35.3%	9.2%
		POI(X+Y)	79.0%	79.9%	91.2%	-38.7%	-35.9%	8.0%
	1,000	SRS	92.8%	92.8%	93.5%	-11.9%	-11.8%	-3.5%
		POI(X)	92.0%	92.8%	95.7%	-17.9%	-15.5%	1.0%
		POI(X+Y)	68.6%	69.6%	74.6%	-18.5%	-14.9%	0.5%
low/high	250	SRS	86.8%	87.0%	94.9%	-37.7%	-37.3%	11.3%
		POI(X)	85.4%	86.1%	95.5%	-42.9%	-41.2%	14.4%
		POI(X+Y)	78.7%	80.1%	92.6%	-44.0%	-41.3%	14.4%
	1,000	SRS	94.4%	94.3%	95.2%	-5.5%	-5.4%	5.8%
		POI(X)	91.8%	92.1%	94.9%	-20.5%	-18.6%	-1.8%
		POI(X+Y)	76.8%	77.8%	82.9%	-20.4%	-16.9%	-1.3%
high/low	250	SRS	89.2%	89.1%	94.4%	-28.5%	-28.1%	0.4%
		POI(X)	89.0%	90.1%	95.5%	-31.9%	-25.3%	12.7%
		POI(X+Y)	85.7%	88.4%	93.8%	-33.9%	-25.4%	10.9%
	1,000	SRS	93.9%	93.9%	95.6%	-6.3%	-6.2%	3.5%
		POI(X)	92.6%	93.4%	94.8%	-16.5%	-9.2%	1.9%
		POI(X+Y)	83.3%	85.4%	88.1%	-15.0%	-5.0%	5.2%
high/high	250	SRS	82.8%	82.8%	93.8%	-44.6%	-44.3%	-6.4%
		POI(X)	83.6%	85.5%	95.1%	-44.3%	-39.4%	3.8%
		POI(X+Y)	82.9%	85.1%	93.8%	-45.1%	-38.4%	4.6%
	1,000	SRS	94.3%	94.4%	96.1%	-7.8%	-7.6%	6.3%
		POI(X)	91.3%	92.2%	94.0%	-20.0%	-13.8%	0.2%
		POI(X+Y)	86.3%	88.6%	91.5%	-18.1%	-9.2%	2.8%

5 Application to National Health Interview Survey (NHIS)

5.1 NHIS and ACS data

We next apply LASSO calibration to National Health Interview Survey (NHIS) 2013 to estimate the total number of adults (age 18 or older) diagnosed with cancer in the population. The National Health Interview Survey is a nationally representative sample of non-institutionalized civilian households collected by a multi-stage area-probability sampling (Centers for Disease Control and Prevention, 2005). Each month, health-related data on a cross-sectional sample of people in selected households are obtained by face-to-face interviews. The data provides pseudo-primary-sampling-unit (PSU), pseudo-strata, and sampling weights to allow for weighted estimates with complex survey design. In addition to health-related measures, NHIS also collects demographic data. Our goal is to assess our LASSO estimator by treating the unweighted NHIS sample as reflective of a non-probability sample, and explore how GREG and LASSO calibration compare with the design-weighted estimator.

To calibrate NHIS on a set of demographic and income-related variables, we use the American Community Survey (ACS) 2013 micro-data as the benchmark data. ACS samples are households selected through multi-stage area-probability sampling from 3,143 counties of the U.S. The design of ACS is to improve estimates of small areas between the decennial census long-form samples. Around three million households are selected each year, with measures collected on household types and individual demographics within the households. ACS also collects data from group-quarters, which are excluded from this analysis. For ACS 2013, the sample size for adults is 2,317,301. The NHIS 2013 sample size is 34,201 after removing 242 cases with missing values on demographic variables. For the purposes of this analysis, we treat the weighted estimates from the ACS as known population totals, a reasonable assumption given the differences in sample size.

5.2 Estimators

The outcome variable of interest is whether a person has been diagnosed with cancer. Define the binary indicator for the outcome variable:

$$y_i = \begin{cases} 1: & \text{if person } i \text{ has been diagnosed with cancer} \\ 0: & \text{otherwise.} \end{cases}$$

We first use the NHIS 2013 sampling weights, $\mathbf{w}^{\text{NHIS}}$, and design variables to obtain an unbiased estimate of the population total, $T_y = \sum_{i=1}^{N} y_i$. Then we assume that the NHIS 2013 sample is collected from a simple-random-sampling, with initial design weights $\mathbf{d}^A = N/n$, where N is the population total obtained from ACS, and n is the sample size of NHIS. We calibrate $\mathbf{d}^A$ by a set of demographic and income variables with traditional GREG calibration and LASSO calibration. Finally, as a compromise between GREG and LASSO, we consider model-assisted calibration to a linear model for y_i instead of the LASSO using (2.7);

note that, when $\hat{\mu}_i$ is computed using the same linear model as in GREG, the point estimates of the total will correspond, even though the calibration weights will differ. Thus, we generate seven estimates:

1. $\hat{T}_y^{\text{NHIS}} = \sum_{i \in s_A} w_i^{\text{NHIS}} y_i$: Estimate obtained with NHIS weights.
2. $\hat{T}_y^{\text{HTSRS}} = \sum_{i \in s_A} (N/n)\, y_i$: Estimate obtained with weights $\mathbf{d}^A = N/n$.
3. $\hat{T}_y^{\text{GREG1}} = \sum_{i \in s_A} w_i^{\text{GREG1}} y_i$: Estimate obtained by calibrating $\mathbf{d}^A$ with GREG using all calibration variables.
4. $\hat{t}_y^{\text{GREG1MC}} = \sum_{i \in s_A} w_i^{\text{GREG1MC}} y_i$: Estimate obtained by model-assisted calibration to linear model using predictors in GREG1.
5. $\hat{T}_y^{\text{GREG2}} = \sum_{i \in s_A} w_i^{\text{GREG2}} y_i$: Estimate obtained by calibrating $\mathbf{d}^A$ with GREG using only calibration variables chosen using backward stepwise variable selection.
6. $\hat{t}_y^{\text{GREG2MC}} = \sum_{i \in s_A} w_i^{\text{GREG2MC}} y_i$: Estimate obtained by model-assisted calibration to linear model using predictors in GREG2.
7. $\hat{T}_y^{\text{LASSO}} = \sum_{i \in s_A} w_i^{\text{LASSO}} y_i$: Estimate obtained by model-assisted calibration with LASSO.

The variance of $\hat{T}_y^{\text{NHIS}}$ is the linearization variance estimate of total, accounting for sampling-stratum, primary-sampling-units, and survey weights in the NHIS 2013 sample. Variances of HTSRS, GREG1, and GREG2 are linearization variance estimates with weights $\mathbf{d}^A$, $\mathbf{w}^{\text{GREG1}}$, and $\mathbf{w}^{\text{GREG2}}$ respectively. We obtain the variance of LASSO estimator through naive bootstrap.

5.3 Working models

Table 5.1 lists calibration variable names, labels, values, and distributions in this analysis. The first column is the unweighted distribution of variables in the NHIS sample. The second column contains variable distributions in the NHIS sample, weighted by $\mathbf{w}^{\text{NHIS}}$ person-level weights. The third column is the distribution of variables in the population obtained from the ACS benchmark data. Missing income category is included as a separate category to capture the difference in missing patterns between NHIS and ACS. Including a missing category also allows us to maintain the analytic sample size. Relative to ACS, the unweighted NHIS sample has higher proportions of females, widowed/divorced/separated, and fewer proportion of non-Hispanic whites. After weighting, the NHIS distributions of gender and race are close to the benchmark's, and only marital status categories show some differences.

We use an unweighted linear model with backward-stepwise variable selection to determine the working model for GREG2. The final variables included in the model for GREG2 are age, education, race, employment status (yes/no), and family income. For standard GREG and LASSO calibration, we use all available variables.

Table 5.1
Calibration variables

			NHIS	ACS
		No weights	**Person-level weights**	**Person-level weights**
Region	Northeast	16%	18%	18%
	Midwest	20%	23%	21%
	South	37%	37%	37%
	West	26%	23%	23%
Age	18-29	19%	21%	21%
	30-39	17%	17%	17%
	40-49	16%	18%	18%
	50-59	17%	18%	18%
	60-69	15%	14%	14%
	70-79	9%	8%	8%
	80+	6%	4%	5%
Gender	Male	45%	48%	48%
	Female	55%	52%	52%
Education	Less than high school	16%	14%	13%
	High school or less	26%	26%	28%
	Some college	20%	20%	23%
	College graduate	29%	30%	25%
	Post-graduate	10%	10%	10%
Race/Ethnicity	Non-Hispanic white	60%	66%	66%
	Non-Hispanic black	15%	12%	12%
	Hispanic	17%	15%	15%
	Other	8%	7%	7%
Marital Status	Married/partnered	49%	60%	52%
	Widowed/divorced/separated	27%	18%	20%
	Never married	24%	22%	28%
Employed	Yes	35%	33%	39%
	No	65%	67%	61%
Income	1st quartile	22%	15%	19%
	2nd quartile	20%	17%	20%
	3rd quartile	21%	22%	20%
	4th quartile	21%	28%	19%
	missing	17%	19%	22%

5.4 Results

Table 5.2 lists the estimates, standard errors (SE), root mean square error treating the correctly weighted NHIS as the true value (RMSE), percent-deviate from the NHIS estimate: %deviate = $100\left(\hat{T} - \hat{T}_y^{\text{NHIS}}\right)\Big/\hat{T}_y^{\text{NHIS}},$ and the standard deviation and minimum and maximum of the weights associated with a given estimator. We treat NHIS estimate as the unbiased estimate because it is calculated with probability-based sampling weights provided by NHIS. Without any weighting adjustment, HTSRS shows a positive bias of 5.9%. The GREG2 estimator reduces this bias from 5.9% to 2.0%, the GREG1 estimator reduces bias to 1.8%, while LASSO estimator reduces the bias to 0.9%. By definition, use of the model-assisted estimator using linear predictors will yield the same estimator as the GREG model; however the variability is substantially reduced. In this analysis, if NHIS were a non-probability sample, without weighting adjustment, we would have over-counted the number of adults with cancer by 1.18 million. With traditional calibration, the error is reduced to an over-count of 365 thousand (without variable selection) or 392 thousand (with variable selection). LASSO calibration further reduces the over-count to 175 thousand.

Table 5.2
Results for estimating total number of individuals with cancer. % deviate is the difference to NHIS estimate divided by the NHIS estimate

Estimator	$\hat{T}$	SE	RMSE	% deviate from NHIS	SD (min, max) of weights
NHIS	19,889,327	492,263	492,263	0.00%	5,913 (168; 93,244)
HTSRS	21,070,498	362,883	1,235,657	5.94%	0 (6,866; 6,866)
GREG1	20,254,449	375,064	523,438	1.84%	2,474 (-2,409; 16,679)
GREG1 MC	20,254,449	349,100	505,158	1.84%	269 (6,181; 7,326)
GREG2	20,281,603	367,900	537,802	1.97%	2,039 (-626; 13,947)
GREG2 MC	20,281,603	349,552	525,421	1.97%	260 (6,215; 7,291)
LASSO	20,064,671	347,586	389,309	0.88%	323 (5,786; 7,168)

As expected, the standard error of the NHIS estimate is the largest, as it properly incorporates complex survey design. If the calibration working model correctly captures the relationship between the outcome variable and the calibration variables, we anticipate that the calibration estimator standard errors to be smaller than HTSRS estimator’s. This is not the case for either of the GREG estimator, where the standard error is larger than HTSRS’s, although the RMSE is smaller due to the reduction in bias. In addition, the standard GREG estimator has a standard error about 2.0% greater than the backward selection GREG estimator, a feature offset by its estimated 6.6% reduction in bias (although this is insufficient to reduce RMSE); use of the model-assisted GREG estimator does reduce the standard error, and the root mean square error, by 5-7% and 2-3% respectively, over the standard GREG estimates. For LASSO calibration, we do observe a smaller standard error than HTSRS’s, even with the bootstrap variance estimate that tends to overestimate. Without using the correct design weights, LASSO calibration produced the most accurate estimate of a population total while providing the smallest standard error among the estimators in this application. This is in spite of the fact that the standard deviation of the LASSO calibration weights were only about one-seventh as variable as the GREG weights, reflected in the smaller standard error of the estimator itself and greatly reduced RMSE.

6 Conclusion

In this manuscript, we developed the LASSO calibration estimator of population totals, $\hat{T}_y^{\mathrm{LASSO}}$, given population auxiliary data. We also derived closed-form variance estimates for $\hat{T}_y^{\mathrm{LASSO}}$. Simulation results show that the point estimates are approximately unbiased under simple-random sampling and informative sampling. For sample selections that are related to analysis variables, LASSO was able to significantly reduce sample bias even without the correct design weights. LASSO tends to outperform stepwise-selected working models when covariates are highly collinear. For analysis with many categorical variables, where there are natural correlations between the categories, LASSO calibration estimator can perform better than traditional calibration estimators, even if the effect sizes are small. The improvement is modest in the continuous variable setting, but substantial when the outcome of interest is binary, as shown in simulations and in the NHIS data example. We have demonstrated theoretically and through simulations that LASSO calibration holds great promise in making unbiased inference of population totals from non-probability samples. Although asymptotic closed-form variance estimates did not produce very accurate nominal coverage, the naive bootstrap is a viable alternative approach. In an application to estimate population total of individuals diagnosed with cancer, without correct design weights, the LASSO calibration estimator was able to produce an estimate that is the closest to the estimate based on correct survey weights. LASSO calibration estimator also has the smallest standard error of all the estimators considered, although the bootstrap variance estimate that was used did not fully account for the clustering in the NHIS, which generally increases standard errors. The application shows that LASSO calibration can generate inference to the population for a specific outcome variable, and the inference is both more accurate and precise than traditional calibration estimators.

The question arises when use of LASSO model-assisted calibration should be used instead of traditional calibration methods such as GREG. Both theoretical and empirical results in this paper suggest that there is little to be lost in terms of statistical efficiency to use LASSO model-assisted calibration, it does require additional effort on the part of the analyst to implement. While we cannot give specific cutoffs, our analysis suggests that this effort will be worthwhile when a) there are large numbers of potential calibration variables, b) many of these calibration variables are likely to be highly correlated, and c) the outcome is binary rather than continuous. We believe that conditions a) and b), at least, are increasing likely to be encountered in non-probability settings, where administrative datasets might provide these types of calibration variables and subsets of data obtained through various means will contain the core variable of interest.

While LASSO provides particularly convenient and rapid implementation, there are, of course, other modern regression methods that could be considered in addition to LASSO to develop penalized regression models for high-dimensional model-assisted regression, including approaches such as ridge regression, principle components, or Bayesian additive regression trees (Chipman, George and McCulloch, 2010). These approaches provide opportunity for further research in this area.

Finally, we note that this work is only a part of a larger and rapidly expanding literature on inference from non-probability surveys. In addition to the work of McConville et al. (2017), the “Mr. P” (multi-level regression and poststratification or MRP) approach of Wang et al. (2015) also uses high dimensional covariates to adjust non-probabilities samples, by use of a hierarchical model rather than penalized regression. Quasi-randomization (Elliott, 2009; Elliott, Resler, Flannagan and Rupp, 2010; Elliott and Valliant, 2017) and sample matching (Rivers, 2007; Vavreck and Rivers, 2008) also provide alternatives that use data from either known population quantities or probability sampling estimates to deal with selection bias issues in non-probability samples. Each have their strengths and weaknesses relative to each other and to model-assisted LASSO. The MRP approach makes distributional assumptions that might improve efficiency, but might reduce robustness, and is non-trivial to implement in its fully Bayesian form. Quasi-randomization forfeits the link to a particular outcome variable, making the weights it develops general purpose but likely less effective, while sample matching requires intervention at the design stage to sample elements from the non-probability frame that match elements from the population, ala quota sampling. The decision to use model-assisted LASSO calibration should be made in the context of these tradeoffs.

Acknowledgements

The authors thank Professor Fred M. Feinberg and Associate Research Scientist Sunghee Lee for their helpful review and suggestions, as well as the Editor, Associate Editor, and three referees whose suggestions greatly contributed to improving this paper.

Appendix

Determining estimates for adaptive LASSO

In practice, we do not observe the theoretical rate of growth of λ_n, which optimize model fit measures such as AIC or BIC, unless we have obtained many samples of the same population with various sample sizes. Given a sample, the choices of λ_n and γ depend on the modeler. In R *glmnet* implementation (Friedman et al., 2010), a range of λ_n is determined by the following scheme:

1. Set $\gamma = 0$.
2. Determine $\lambda_n^{\max}$ by finding the smallest λ_n that sets all coefficients to 0.
3. If sample size n is larger than the number of parameters in the regression model, set $\lambda_n^{\min} = 0.0001\lambda_n^{\max}$. If sample size n is smaller than the number of parameters, set $\lambda_n^{\min} = 0.01\lambda_n^{\max}$ (to set parameters to 0 sooner).
4. Generate a grid of λ_n, typically 100 equally spaced points between $\lambda_n^{\min}$ and $\lambda_n^{\max}$.

The initial range of values of λ_n is determined independently of γ. Choices of γ are less data-driven. Some modelers choose one of $\gamma = 0.1, 0.5, 1, 2$. Here we determine (λ_n, γ) through cross-validation as follows:

1. Obtain $\alpha_j = 1 / |\hat{\beta}_j^{\text{MLE}}|$.
2. Determine 100 equally spaced values of λ_n based on R *glmnet* 's implementation.
3. For each pair (λ_n, γ), λ_n from Step 2, and $\gamma = 0.1,\ 0.5,\ 1,\ 2,$ split data into 5 folds. Use 4 folds to obtain $\hat{\boldsymbol{\beta}}$.

4. Apply $\hat{\boldsymbol{\beta}}$ to the last fold not used to estimate $\hat{\boldsymbol{\beta}}$ and calculate a metric. For continuous $\mathbf{y}$, we calculate the mean-absolute-error (MAE), $\sum_{i \in s_{A(k)}} |\,\hat{\mu}_i - y_i\,|$. For binary $\mathbf{y}$, we calculate the area under curve (AUC) (calculated through R *glmnet* :: *auc* function).
5. Average the 5 metrics for each pair of $(\lambda_n,\, \gamma)$, and choose the pair with the best average metric: minimum MAE for continuous $\mathbf{y}$, maximum AUC for binary $\mathbf{y}$.

The adaptive LASSO coefficient estimates are then obtained by solving equations (3.1) or (3.2) in Section 3.1 given the selected $(\lambda_n,\, \gamma)$. The R code used to perform cross-validation is provided in the online supplemental material.

Asymptotic unbiasedness and variance of model-assisted LASSO calibration estimator of a population total

Lemma 1: *Assume the superpopulation model:*

$$E_\xi\,(y_k\,|\,\mathbf{x}_k) \;=\; \mu\,(\mathbf{x}_k,\,\boldsymbol{\beta}), V_\xi\,(y_k\,|\,\mathbf{x}_k) = v_k^2\sigma^2.$$

Let $\mathbf{B}$ *be the finite-population quasilikelihood estimate of* $\boldsymbol{\beta}$, $\mathbf{B} \to \boldsymbol{\beta}$. *Under conditions (1)-(5) in Section 3.2, the model-assisted asymptotic estimator of population total is:*

$$\hat{T}_y^{MC} \;=\; \sum_{i \in s_A} d_i^A\,(y_i - \mu_i B^{MC}) + \sum_{i=1}^{N} \mu_i B^{MC} + o_p\left(\frac{N}{\sqrt{n}}\right) \tag{A.1}$$

where

$$\mu_i \;=\; \mu\,(\mathbf{x}_i,\,\mathbf{B})$$

$$B^{MC} \;=\; \frac{\sum_{i=1}^{N}(\mu_i - \bar{\mu})\,(y_i - \bar{y})}{\sum_{i=1}^{N}(\mu_i - \bar{\mu})^2}.$$

Proof. The proof is adapted and expanded from the proof of Theorem 1 in Wu and Sitter (2001), with slight modifications in notations to be consistent with this paper. We begin by deriving the asymptotic model-assisted estimator for a population mean, $\hat{\bar{y}}^{\text{MC}} = N^{-1}\hat{T}_y^{\text{MC}}$ *(see equation (2.7)). By conditions (2) and (3), the second order Taylor series expansion of* $\mu\left(\mathbf{x}_i,\,\hat{\mathbf{B}}\right)$ *around* $\mathbf{B}$ *is:*

$$\mu\left(\mathbf{x}_i, \hat{\mathbf{B}}\right) = \mu\,(\mathbf{x}_i, \mathbf{B}) + \left\{\frac{\partial \mu\,(\mathbf{x}_i,\,\mathbf{t})}{\partial \mathbf{t}}\bigg|_{\mathbf{t}=\mathbf{B}}\right\}^T \left(\hat{\mathbf{B}} - \mathbf{B}\right) + \left(\hat{\mathbf{B}} - \mathbf{B}\right)^T \left\{\frac{\partial^2 \mu\,(\mathbf{x}_i,\,\mathbf{t})}{\partial \mathbf{t}\partial \mathbf{t}^T}\,|_{\mathbf{t}=\mathbf{B}^*}\right\}\left(\hat{\mathbf{B}} - \mathbf{B}\right) \tag{A.2}$$

for $\mathbf{B}^* \in \left(\hat{\mathbf{B}}, \mathbf{B}\right)$ *or* $\left(\mathbf{B},\, \hat{\mathbf{B}}\right)$. *Let*

$$\mathbf{h}\,(\mathbf{x}_i,\,\mathbf{B}) \;=\; \frac{\partial \mu\,(\mathbf{x}_i,\,\mathbf{t})}{\partial \mathbf{t}}\,|_{\mathbf{t}=\mathbf{B}}$$

$$\mathbf{k}\,(\mathbf{x}_i,\,\mathbf{B}^*) \;=\; \frac{\partial^2 \mu\,(\mathbf{x}_i,\,\mathbf{t})}{\partial \mathbf{t}\partial \mathbf{t}^T}\bigg|_{\mathbf{t}=\mathbf{B}^*}$$

Note that $\mathbf{h}$ *is a vector of length* m *and* $\mathbf{k}$ *is a matrix of size* $m \times m$, *where* m *is the number of parameters in* $\boldsymbol{\beta}$. *By conditions (2) and (3),*

$$\max_i \left| \mathbf{h}(\mathbf{x}_i, \mathbf{B}) \right| \le h(\mathbf{x}_i, \mathbf{B}) \tag{A.3}$$

$$\max_{k,\,j} \left| \mathbf{k}(\mathbf{x}_i, \mathbf{B}^*) \right| \le k(\mathbf{x}_i, \mathbf{B}^*). \tag{A.4}$$

Conditions (1) and (3) imply that

$$\mu\left(\mathbf{x}_i, \hat{\mathbf{B}}\right) = \mu(\mathbf{x}_i, \mathbf{B}) + O_p\left(1\big/\sqrt{n}\right) \tag{A.5}$$

$$\equiv \mu_i + O_p\left(1\big/\sqrt{n}\right). \tag{A.6}$$

By equation (2.2) in Section 2.1 and the boundedness conditions of (2) and (3) in Section 3.2.2 imply

$$\begin{aligned} N^{-1}\sum_{i\in s_A} d_i^A \mu\left(\mathbf{x}_i, \hat{\mathbf{B}}\right) &= N^{-1}\sum_{i\in s_A} d_i^A \mu(\mathbf{x}_i, \mathbf{B}) + N^{-1}\left(\sum_{i\in s_A} d_i^A \mathbf{h}(\mathbf{x}_i, \mathbf{B})\right)^T \left(\hat{\mathbf{B}} - \mathbf{B}\right) \\ &\quad + \left(\hat{\mathbf{B}} - \mathbf{B}\right)^T N^{-1}\left(\sum_{i\in s_A} d_i^A \mathbf{k}(\mathbf{x}_i, \mathbf{B}^*)\right)\left(\hat{\mathbf{B}} - \mathbf{B}\right) \\ &= N^{-1}\sum_{i\in s_A} d_i^A \mu(\mathbf{x}_i, \mathbf{B}) + N^{-1}\left(\sum_{i\in s_A} d_i^A \mathbf{h}(\mathbf{x}_i, \mathbf{B})\right)^T \left(\hat{\mathbf{B}} - \mathbf{B}\right) + O_p\left(\frac{1}{n}\right). \end{aligned} \tag{A.7}$$

By conditions (1), (4), and equation (A.7):

$$\begin{aligned} &N^{-1}\sum_{k=1}^{N} \mu\left(\mathbf{x}_k, \hat{\mathbf{B}}\right) - N^{-1}\sum_{i\in s_A} d_i^A \mu\left(\mathbf{x}_i, \hat{\mathbf{B}}\right) \\ &\qquad = N^{-1}\sum_{k=1}^{N} \mu(\mathbf{x}_i, \mathbf{B}) - N^{-1}\sum_{i\in s_A} d_i^A \mu(\mathbf{x}_i, \mathbf{B}) + O_p\left(\frac{1}{\sqrt{n}}\right). \end{aligned} \tag{A.8}$$

Using conditions (1) and (3),

$$\begin{aligned} \bar{\hat{\mu}} &= \sum_{i\in s_A} d_i^A \mu\left(\mathbf{x}_i, \hat{\mathbf{B}}\right) \Big/ \sum_{i\in s_A} d_i^A \\ &= \left(\sum_{i\in s_A} d_i^A\right)^{-1} \sum_{i\in s_A} d_i^A \left(\mu(\mathbf{x}_i, \mathbf{B}) + \mathbf{h}^T(\mathbf{x}_i, \mathbf{B})\left(\hat{\mathbf{B}} - \mathbf{B}\right) + \left(\hat{\mathbf{B}} - \mathbf{B}\right)^T \mathbf{k}(\mathbf{x}_i, \mathbf{B}^*)\left(\hat{\mathbf{B}} - \mathbf{B}\right)\right) \\ &= \left(\sum_{i\in s_A} d_i^A\right)^{-1} \sum_{i\in s_A} d_i^A \left(\mu(\mathbf{x}_i, \mathbf{B}) + \mathbf{h}^T(\mathbf{x}_i, \mathbf{B})\left(\hat{\mathbf{B}} - \mathbf{B}\right)\right) + O_p(1/n) \\ &= \bar{\mu} + \left(\sum_{i\in s_A} d_i^A\right)^{-1} \sum_{i\in s_A} d_i^A \mathbf{h}^T(\mathbf{x}_i, \mathbf{B})\left(\hat{\mathbf{B}} - \mathbf{B}\right) + O_p(1/n) \\ &\quad (\textit{by condition } (1) \textit{ and } (18)) \\ &= \bar{\mu} + O_p\left(1\big/\sqrt{n}\right) + O_p(1/n) \\ &= \bar{\mu} + O_p\left(1\big/\sqrt{n}\right) \end{aligned} \tag{A.9}$$

for $\bar{\mu} = \sum_{i\in s_A} d_i^A \mu_i \Big/ \sum_{i\in s_A} d_i^A$.

Then from (A.2) and (A.9) and using conditions (1)-(3), we have

$$\begin{aligned}
N^{-1}\sum_{i\in s_A} d_i^A\left(\hat{\mu}_i-\hat{\bar{\mu}}\right) &= N^{-1}\sum_{i\in s_A} d_i^A\Big(\mu\left(\mathbf{x}_i,\mathbf{B}\right)+\mathbf{h}^T\left(\mathbf{x}_i,\mathbf{B}\right)\left(\hat{\mathbf{B}}-\mathbf{B}\right)+\left(\hat{\mathbf{B}}-\mathbf{B}\right)^T\mathbf{k}\left(\mathbf{x}_i,\mathbf{B}^*\right)\left(\hat{\mathbf{B}}-\mathbf{B}\right)-\bar{\mu}\Big)\\
&= N^{-1}\sum_{i\in s_A} d_i^A\left(\mu_i-\bar{\mu}\right)+N^{-1}\sum_{i\in s_A} d_i^A\mathbf{h}^T\left(\mathbf{x}_i,\mathbf{B}\right)\left(\hat{\mathbf{B}}-\mathbf{B}\right)\\
&\quad + N^{-1}\sum_{i\in s_A} d_i^A\left(\hat{\mathbf{B}}-\mathbf{B}\right)^T\mathbf{k}\left(\mathbf{x}_i,\mathbf{B}^*\right)\left(\hat{\mathbf{B}}-\mathbf{B}\right)-O_p\left(1/\sqrt{n}\right)\\
&= N^{-1}\sum_{i\in s_A} d_i^A\left(\mu_i-\bar{\mu}\right)+O_p\left(1/\sqrt{n}\right)+O_p\left(1/n\right)-O_p\left(1/\sqrt{n}\right)\\
&= N^{-1}\sum_{i\in s_A} d_i^A\left(\mu_i-\bar{\mu}\right)+O_p\left(1/\sqrt{n}\right). \qquad \text{(A.10)}
\end{aligned}$$

Similarly,

$$N^{-1}\sum_{i\in s_A} d_i^A\left(\hat{\mu}_i-\hat{\bar{\mu}}\right)^2 = N^{-1}\sum_{i\in s_A} d_i^A\left(\mu_i-\bar{\mu}\right)^2+O_p\left(1/n\right). \qquad \text{(A.11)}$$

From (A.10) and (A.11) we have:

$$\begin{aligned}
\hat{B}^{MC} &= \frac{\sum_{i\in s_A} d_i^A\left(\hat{\mu}_i-\hat{\bar{\mu}}\right)\left(y_i-\bar{y}\right)}{\sum_{i\in s_A} d_i^A\left(\hat{\mu}_i-\hat{\bar{\mu}}\right)^2} = \frac{N^{-1}\sum_{i\in s_A} d_i^A\left(\hat{\mu}_i-\hat{\bar{\mu}}\right)\left(y_i-\bar{y}\right)}{N^{-1}\sum_{i\in s_A} d_i^A\left(\hat{\mu}_i-\hat{\bar{\mu}}\right)^2}\\
&= \frac{\sum_{i\in s_A} d_i^A\left(\mu_i-\bar{\mu}\right)\left(y_i-\bar{y}\right)+O_p\left(1/\sqrt{n}\right)}{\sum_{i\in s_A} d_i^A\left(\mu_i-\bar{\mu}\right)^2+O_p\left(1/n\right)}\\
&\to B^{MC} \qquad as\ n\to\infty. \qquad \text{(A.12)}
\end{aligned}$$

Thus $\hat{B}^{MC} = B^{MC}+o_p\left(1\right)$, *and we have:*

$$\begin{aligned}
\hat{\bar{y}}^{MC} &= N^{-1}\hat{T}_y^{MC}\\
&= N^{-1}\mathbf{d}^A\mathbf{y}+\left(N^{-1}\sum_{k=1}^{N}\mu\left(\mathbf{x}_k,\hat{\mathbf{B}}\right)+\sum_{i\in s_A} N^{-1}d_i^A\mu\left(\mathbf{x}_i,\hat{\mathbf{B}}\right)\right)\hat{B}^{MC}\\
&= N^{-1}\mathbf{d}^A\mathbf{y}+\left(N^{-1}\sum_{k=1}^{N}\mu\left(\mathbf{x}_k,\mathbf{B}\right)-N^{-1}\sum_{i\in s_A} d_i^A\mu\left(\mathbf{x}_i,\mathbf{B}\right)+O_p\left(\frac{1}{\sqrt{n}}\right)\right)\left(B^{MC}+o_p\left(1\right)\right)\\
&= N^{-1}\mathbf{d}^A\mathbf{y}+\left(N^{-1}\sum_{k=1}^{N}\mu\left(\mathbf{x}_k,\mathbf{B}\right)-N^{-1}\sum_{i\in s_A} d_i^A\mu\left(\mathbf{x}_i,\mathbf{B}\right)\right)B^{MC}+o_p\left(\frac{1}{\sqrt{n}}\right).
\end{aligned}$$

Since $N = O_p\left(N\right)$, *we have* $N\cdot o_P\left(1/\sqrt{n}\right) = O_p\left(N\right)o_p\left(1/\sqrt{n}\right) = o_p\left(N/\sqrt{n}\right)$. *Thus,*

$$\begin{aligned}
\hat{T}_y^{MC} &= N\hat{\bar{y}}^{MC} = N\left(N^{-1}\mathbf{d}^A\mathbf{y}+\left(N^{-1}\sum_{k=1}^{N}\mu\left(\mathbf{x}_k,\mathbf{B}\right)-N^{-1}\sum_{i\in s_A}\mu\left(\mathbf{x}_i,\mathbf{B}\right)\right)B^{MC}+o_p\left(\frac{1}{\sqrt{n}}\right)\right)\\
&= \mathbf{d}^A\mathbf{y}+\left(\sum_{k=1}^{N}\mu\left(\mathbf{x}_k,\mathbf{B}\right)-\sum_{i\in s_A}\mu\left(\mathbf{x}_i,\mathbf{B}\right)\right)B^{MC}+o_p\left(\frac{N}{\sqrt{n}}\right)\\
&= \sum_{iins_A} d_i^A\left(y_i-\mu_i B^{MC}\right)+\sum_{i=1}^{N}\mu_i B^{MC}+o_p\left(\frac{N}{\sqrt{n}}\right). \qquad \text{(A.13)}
\end{aligned}$$

Theorem 2: *Suppose the parameters in a full regression model have both zero and non-zero components, without loss of generality, let the first* p *be non-zero and the last* q *be zero:*

$$\boldsymbol{\beta}^F = \begin{pmatrix} \boldsymbol{\beta}^{(1)}_{(p\times 1)} \\ \boldsymbol{\beta}^{(2)}_{(q\times 1)} \end{pmatrix}, \ \boldsymbol{\beta}^{(1)} = \boldsymbol{\beta} \quad and \quad \boldsymbol{\beta}^{(2)} = \mathbf{0}_{(q\times 1)}.$$

Under conditions (1)-(5), the asymptotic LASSO calibration estimator of total is:

$$\hat{T}_y^{LASSO} = \sum_{i\in s_A} d_i^A \left(y_i - \mu_i B^{MC}\right) + \sum_{i=1}^{N} \mu_i B^{MC} + o_p\left(\frac{N}{\sqrt{n}}\right). \tag{A.14}$$

Proof. Under condition (5), the adaptive LASSO regression satisfies the oracle property through Theorems 1 and 4 in Zou (2006):

$$\begin{aligned} Pr\left(\mathbf{B}^{(2)} = \mathbf{0}\right) &\rightarrow 1 \\ \sqrt{n}\left(\hat{\mathbf{B}}^{(1)} - \mathbf{B}\right) &\rightarrow N\left(\mathbf{0}, \mathbf{C}\right) \\ \mathbf{B} &\rightarrow \boldsymbol{\beta} \end{aligned}$$

where $\mathbf{C} = \Sigma\left(\mathbf{B}\right)$ *is the covariance matrix of* $\mathbf{B}^{(1)}$ *under the linear model, and* $\mathbf{C} = I^{-1}\left(\mathbf{B}\right)$ *is the inverse of Fisher information matrix of* $\mathbf{B}^{(1)}$ *under generalized linear model. By Slutsky's theorem, the oracle property implies* $\hat{\mathbf{B}}^{(1)} = \mathbf{B} + O_p\left(1/\sqrt{n}\right)$. *By condition (1) and Lemma 1:*

$$\begin{aligned} \hat{T}_y^{LASSO} &\approx \hat{T}_y^{MC} \\ &= \sum_{iins_A} d_i^A \left(y_i - \mu_i B^{MC}\right) + \sum_{i=1}^{N} \mu_i B^{MC} + o_p\left(\frac{N}{\sqrt{n}}\right). \end{aligned}$$

Theorem 3: $\hat{T}_y^{LASSO}$ *is model-unbiased.*

Proof. Under the assumption of our theoretical framework, the superpopulation parameters are a subset of the full LASSO regression parameters, we can prove the model-unbiasedness of $\hat{T}_y^{LASSO}$ *by taking expectations with respect to model* ξ. *First note that:*

$$E_\xi\left[B^{MC}\right] = E_\xi\left[\frac{\sum_{i=1}^{N}(\mu_i - \bar{\mu})(y_i - \bar{y})}{\sum_{i=1}^{N}(\mu_i - \bar{\mu})^2}\right] = \frac{\sum_{i=1}^{N}(\mu_i - \bar{\mu})(\mu_i - \bar{\mu})}{\sum_{i=1}^{N}(\mu_i - \bar{\mu})^2} = 1.$$

Thus

$$\begin{aligned} E_\xi\left[\hat{T}_y^{LASSO} - T\right] &\approx E_\xi\left[\sum_{i\in s_A} d_i^A \left(y_i - \mu_i B^{MC}\right) + \sum_{i=1}^{N} \mu_i B^{MC} - \sum_{i=1}^{N} y_i\right] \\ &= \sum_{i\in s_A} d_i^A \left(\mu_i - \mu_i\right) + \sum_{i=1}^{N} \mu_i - \sum_{i=1}^{N} \mu_i \qquad (since\ E_\xi\left[B^{MC}\right] = 1) \\ &= 0. \end{aligned}$$

Thus, as long as LASSO regression parameters include the superpopulation parameters, $\hat{T}_y^{\text{LASSO}}$ is model-unbiased regardless of design weights. This property is essential in non-probability samples, where there are no initial design weights to guarantee unbiasedness.

References

Baker, R., Brick, J.M., Bates, N.A., Battaglia, M., Couper, M.P., Dever, J.A., Gile, K. and Tourangeau, R. (2013). Summary report of the AAPOR task force on non-probability sampling. *Journal of Survey Statistics and Methodology*, 1, 90-143.

Chipman, H.A., George, E.I. and McCulloch, R.E. (2010). BART: Bayesian additive regression trees. *Annals of Applied Statistics*, 4, 266-298.

Cardot, H., Goga, C. and Shehzad, M.-A. (2017). Calibration and partial calibration on principal components when the number of auxiliary variables is large. *Statistica Sinica*, 27, 243-260.

Centers for Disease Control and Prevention (2005). *2004 National Health Interview Survey (NHIS) Public Use Data Release: NHIS Survey Description*. National Center for Health Statistics: Hyattsville, Maryland. www.cdc.gov/nchs/data/nhis/srvydesc.pdf.

Czanner, G., Sarma, S.V., Eden, U.T. and Brown, E.N. (2008). A signal-to-noise ratio estimator for generalized linear model systems. *Proceedings of the World Congress on Engineering*, vol. 2.

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

Dormann, C.F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carre, G., Marquez, J.R.G., Gruber, B., Lafourcade, B., Leitao, P.J. and Mnkemller, T. (2013). Collinearity: A review of methods to deal with it and a simulation study evaluating their performance. *Ecology*, 36, 27-46.

Elliott, M.R. (2009). Combining data from probability and nonprobability samples using pseudo-weights. *Survey Practice*, 2(6).

Elliott, M.R, Resler, A., Flannagan, C. and Rupp, J. (2010). Combining data from probability and non-probability samples using pseudo-weights. *Accident Analysis and Prevention*, 42, 530-539.

Elliott, M.R., and Valliant, R. (2017). Inference for non-probability samples. *Statistical Science*, 32, 249-264.

Fan, J., and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96, 1348-1360.

Friedman, J., Hastie, T. and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33, 1-22.

Frankel, M.R., and Frankel, L.R. (1987). Fifty years of survey sampling in the United States. *Public Opinion Quarterly*, S127-S138.

Fuller, W.A. (2009). *Sampling Statistics*. New York: John Wiley & Sons, Inc.

Goga, C., Muhammad-Shehzad, A. and Vanheuverzwyn, A. (2011). Principal component regression with survey data: Application on the French media audience. *Proceedings of the 58th World Statistics Congress of the International Statistical Institute*, 3847-3852.

Groves, R.M. (2006). Nonresponse rates and nonresponse bias in household surveys. *Public Opinion Quarterly*, 70, 646-675.

Kamarianakis, Y., Shen, W. and Wynter, L. (2012). Real-time road traffic forecasting using regime-switching space-time models and adaptive LASSO. *Applied Stochastic Models in Business and Industry*, 28, 297-315.

Kohannim, O., Hibar, D.P., Stein, J.L., Jahanshad, N., Hua, X., Rajagopalan, P., Toga, A., Jack Jr, C.R., Weiner, M.W., de Zubicaray, G.I. and McMahon, K.L. (2012). Discovery and replication of gene influences on brain structure using LASSO regression. *Frontiers in Neuroscience*, 6, 115.

Kohut, A., Keeter, S., Doherty, C., Dimock, M. and Christian, L. (2012). Assessing the representativeness of public opinion surveys. *Pew Research Center for The People & The Press*. http://www.people-press.org/2012/05/15/assessing-the-representativeness-of-public-opinion-surveys/.

Mosteller, F. (1949). *The Pre-Election Polls of 1948: The Report to the Committee on Analysis of Pre-Election Polls and Forecasts*, vol. 60, Social Science Research Council.

McConville, K. (2011). *Improved Estimation for Complex Surveys Using Modern Regression Techniques*. Unpublished PhD Thesis, Colorado State University.

McConville, K., Breidt, F.J., Lee, T.M. and Moisen, G.G. (2017). Model-assisted survey regression estimation with the LASSO. *Journal of Survey Statistics and Methodology*, 5, 131-158.

Park, M., and Yang, M. (2008). Ridge regression estimation for survey samples. *Communication in Statistics - Theory and Methods*, 37, 532-543.

Rivers, D. (2007). Sampling for web surveys. *Proceedings of the Joint Statistical Meetings*, American Statistical Association.

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

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

Skinner, C., and Silva, P. (1997). Variable selection for regression estimation in the presence of nonresponse. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 76-81.

Stephan, F.F. (1948). History of the uses of modern sampling procedures. *Journal of the American Statistical Association*, 43, 12-39.

Terhanian, G., and Bremer, J. (2012). A smarter way to select respondents for surveys? *International Journal of Market Research*, 54, 751-780.

Tibshirani, R. (1996). Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society*, 58, 267-288.

Tourangeau, R., Conrad, F.G. and Couper, M.P. (2013). *The Science of Web Surveys.* Oxford University Press, Oxford, UK.

Vavreck, L., and Rivers, D. (2008). The 2006 Cooperative Congressional Election Study. *Journal of Elections, Public Opinion, and Parties*, 355-366.

Wang, W., Rothschild, D., Goel, S. and Gelman, A. (2015). Forecasting elections with non-representative Polls. *International Journal of Forecasting*, 31, 980-991.

Wu, C., and Sitter, R.R. (2001). A model-calibration approach to using complete auxiliary information from survey data. *Journal of the American Statistical Association*, 96, 185-193.

Wu, T.T., Chen, Y.F., Hastie, T., Sobel, E. and Lange, K. (2009). Genome-wide association analysis by LASSO penalized logistic regression. *Bioinformatics*, 25, 714-721.

Zou, H. (2006). The adaptive LASSO and its oracle properties. *Journal of the American Statistical Association*, 101, 1418-1429.

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 33, No. 4, December 2017

All inquires about submissions and subscriptions should be directed to jos@scb.se

JOURNAL OF OFFICIAL STATISTICS

An International Review Published by Statistics Sweden

JOS is a scholarly quarterly that specializes in statistical methodology and applications. Survey methodology and other issues pertinent to the production of statistics at national offices and other statistical organizations are emphasized. All manuscripts are rigorously reviewed by independent referees and members of the Editorial Board.

Contents
Volume 34, No. 1, March 2018

All inquires about submissions and subscriptions should be directed to jos@scb.se

The Canadian Journal of Statistics **La revue canadienne de statistique**

CONTENTS TABLE DES MATIÈRES

Volume 46, No. 1, March/mars 2018

The Canadian Journal of Statistics **La revue canadienne de statistique**

CONTENTS TABLE DES MATIÈRES

Volume 46, No. 2, June/juin 2018

GUIDELINES FOR MANUSCRIPTS

Authors are invited to submit their articles in English or French in electronic form to the Editor, (statcan.smj-rte.statcan@canada.ca). Before submitting the article, please examine a recent issue of *Survey Methodology* (Vol. 39, No. 1 and onward) as a guide and note particularly the points below. Articles must be submitted in machine-readable form, preferably in Word with MathType for the mathematical expressions. A pdf or paper copy may be required for formulas and figures.

1. Layout

1.1 Documents should be typed entirely double spaced with margins of at least 1½ inches on all sides.

1.2 The documents should be divided into numbered sections with suitable verbal titles.

1.3 The name (fully spelled out) and address of each author should be given as a footnote on the first page of the manuscript.

1.4 Acknowledgements should appear at the end of the text.

1.5 Any appendix should be placed after the acknowledgements but before the list of references.

2. Abstract

The manuscript should begin with an abstract consisting of one paragraph followed by three to six key words. Avoid mathematical expressions in the abstract.

3. Style

3.1 Avoid footnotes, abbreviations, and acronyms.

3.2 Mathematical symbols will be italicized unless specified otherwise except for functional symbols such as "exp(·)" and "log(·)", etc.

3.3 Short formulae should be left in the text but everything in the text should fit in single spacing. Long and important equations should be separated from the text and numbered with arabic numerals on the right if they are to be referred to later. Use a two-level numbering system based on the section of the paper. For example, equation (4.2) is the second important equation in section 4.

3.4 Write fractions in the text using a solidus.

3.5 Distinguish between ambiguous characters, (e.g., w, ω; o, O, 0; l, 1).

3.6 If possible, avoid using bold characters in formulae.

4. Figures and Tables

4.1 All figures and tables should be numbered with arabic numerals, with titles that are as self explanatory as possible, at the bottom for figures and at the top for tables. Use a two-level numbering system based on the section of the paper. For example, table 3.1 is the first table in section 3.

4.2 A detailed textual description of figures may be required for accessibility purposes if the message conveyed by the image is not sufficiently explained in the text.

5. References

5.1 References in the text should be cited with authors' names and the date of publication. If part of a reference is cited, indicate after the reference, e.g., Cochran (1977, page 164).

5.2 The list of references at the end of the manuscript should be arranged alphabetically and for the same author chronologically. Distinguish publications of the same author in the same year by attaching a, b, c to the year of publication. Journal titles should not be abbreviated. Follow the same format used in recent issues.

6. Short Notes

6.1 Documents submitted for the short notes section must have a maximum of 3,000 words.