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

Survey Methodology

Reducing the response imbalance: Is the accuracy of the survey estimates improved?

by Carl-Erik Särndal, Kaur Lumiste and Imbi Traat

Release date: December 20, 2016

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

An HTML version is also available.

Cette publication est aussi disponible en français.

Survey Methodology, December 2016
Vol. 42, No. 2, pp. 219-238
Statistics Canada, Catalogue No. 12-001-X

Reducing the response imbalance: Is the accuracy of the survey estimates improved?

Carl-Erik Särndal, Kaur Lumiste and Imbi Traat[1]

Abstract

We present theoretical evidence that efforts during data collection to balance the survey response with respect to selected auxiliary variables will improve the chances for low nonresponse bias in the estimates that are ultimately produced by calibrated weighting. One of our results shows that the variance of the bias – measured here as the deviation of the calibration estimator from the (unrealized) full-sample unbiased estimator – decreases linearly as a function of the response imbalance that we assume measured and controlled continuously over the data collection period. An attractive prospect is thus a lower risk of bias if one can manage the data collection to get low imbalance. The theoretical results are validated in a simulation study with real data from an Estonian household survey.

Key Words: Survey nonresponse; Bias; Adaptive data collection; Calibration estimator; Auxiliary variables.

1 Introduction

The problem of accurate estimation despite considerable nonresponse needs to be examined from two time dependent angles: First, ways to handle the data collection, then ways to handle the estimation with the data that were finally collected. The first activity may require substantial resources. In a telephone survey, the daily scheduling of contact attempts, the interaction with the interviewers, and consideration for their workloads, can be expensive efforts. The estimation stage is administratively simpler; there is a search for the best auxiliary variables for a calibrated nonresponse adjustment weighting, whereupon the computation of estimates is usually carried out with existing software.

The data collection is in focus in the literature on Responsive Design; Groves (2006), Groves and Heeringa (2006) are early references. Adaptive survey designs are discussed in Wagner (2008). One idea in this research tradition is that a data collection that extends over a period of time might be inspected at suitable decision points, where action may be taken to realize in the end a well-balanced set of respondents. Schouten, Calinescu and Luiten (2013) explain how adaptive survey designs may be tailored to optimize response rates and reduce nonresponse selectivity, with cost aspects taken into account. Much exploratory work has been carried out on responsive (or adaptive) design. Seeking well balanced or representative response can be pursued as a goal in itself. Different avenues have been explored: Case prioritization, (Peytchev, Riley, Rosen, Murphy and Lindblad 2010); stopping rules to halt data collection attempts for specific sample units, (Rao, Glickman and Glynn 2008; Wagner and Raghunathan 2010); uses of paradata more generally to manage the survey response, (Couper and Wagner 2011).

Measuring and controlling the imbalance belongs in the data collection phase. The imbalance statistic (see Section 3) has a central role in this article; it was used for example in Särndal (2011), Lundquist and Särndal (2013), Särndal and Lundquist (2014a, 2014b). It is related to the R- indicator (R for

1. Carl-Erik Särndal, Ph.D., professor emeritus, Statistics Sweden. E-mail: carl.sarndal@telia.com; Kaur Lumiste, M.Sc., Institute of Mathematical Statistics, University of Tartu, Estonia; Imbi Traat, Ph.D., associate professor, Institute of Mathematical Statistics, University of Tartu, Estonia.

representativity); see Schouten, Cobben and Bethlehem (2009) and Bethlehem, Cobben and Schouten (2011).

The second time slice relies on estimation theory to resolve the challenge of nonresponse, primarily how to achieve low bias in the estimates. Viewed strictly as an estimation problem, it is an activity in itself, after a completed data collection. The set of responding units is fixed; the data on those units is a "frozen" supply. The choice of auxiliary variables plays a crucial role. The "best ones" should be selected. This aspect has been dealt with extensively, as in Särndal and Lundström (2005). Two factors are traditionally cited as important for the accuracy of the estimates: The degree to which the chosen auxiliary variables can explain the study variable and the degree to which these variables can explain the 0/1 response indicator showing presence or not in the set of respondents. Each of the two degrees of explanation is partial at best, not perfect. The two roles of the auxiliary variables interact, as recognized for example in Little and Vartivarian (2005). An extensive review of weighting adjustment procedures for nonresponse is given in Brick (2013).

The supply of auxiliary variables depends on the survey environment. In Scandinavia, surveys on individuals and households can draw on extensive sources – administrative registers – of auxiliary variables. This is increasingly so in other countries also.

One view holds that the estimation is the all-important step: Whatever may be accomplished at the data collection stage – balancing, improved representativeness – is perhaps superfluous; achieving best possible accuracy in the estimates can be dealt with effectively at the estimation stage, by clever use of the auxiliary variables in a nonresponse adjustment weighting or in other ways. This point of view is supported for example in Beaumont, Bocci and Haziza (2014).

Nevertheless, it is clear that measurable aspects of the data collection will influence the accuracy of the estimates that are ultimately produced. One such measure is the imbalance statistic defined in Section 3. In this article, the two time dependent activities are taken into account: Balancing the response should be combined with efficient estimation methods, to get in the end the best possible (most accurate) estimates. Such a view underlies, for example, Schouten, Cobben, Lundquist and Wagner (2014).

The motivation for this paper is as follows: Methods exist for different courses of action – stopping rules, case prioritization, and others – during data collection, so as to get in the end a favourable response set r. Särndal and Lundquist (2014a, 2014b) used the imbalance statistic IMB given in Section 3 as a tool to achieve low imbalance in the final response set. Considering that auxiliary variables will also be used in the estimation, to what extent, if any, will better accuracy in the estimates follow from low imbalance in the preceding data collection? There are encouraging signs, as in Särndal and Lundquist (2014a), that lower imbalance creates some accuracy improvement, although modest. That work was empirical; in this article we give mathematical/analytical support for a similar conclusion.

The contents are arranged as follows: The survey background (Section 2) and the imbalance statistic (Section 3) are presented. The regression relationship – that of the study variable on the auxiliary vector – is important (Section 4), notably for the estimator (called CAL) obtained by calibrated nonresponse weight adjustment (Section 5). The deviation of the calibration (CAL) estimator from the (unbiased) estimator requiring full response is analyzed (Sections 6, Section 7, Section 8), showing how deviation depends on imbalance. Two results are presented on statistical properties (mean and variance) of the CAL deviation. In

particular, the variance of that deviation is shown to be, approximately, a linear function of the imbalance statistic. Hence the deviation is likely to be smaller, and estimates more accurate, if the imbalance can be reduced during data collection. The theoretical results are empirically validated (Section 9) using data from an Estonian household survey. The statistical software R is used; R Core Team (2014). A discussion (Section 10) concludes the article. Three appendices provide the necessary proofs and derivations.

2 Background and notation

A probability sample s is drawn from the finite population $U = \{1, 2, \ldots, k, \ldots, N\}$. Unit k has the known inclusion probability $\pi_k = \Pr(k \in s)$ and the known design weight $d_k = 1/\pi_k$. Nonresponse occurs. The response set, denoted r, is that subset of s for which the study variable is observed. We do not know how r was generated from s; the response probabilities are unknown (if assumed to "exist", they are not needed in this article). The (design weighted) response rate is

$$P = \sum_r d_k \Big/ \sum_s d_k . \tag{2.1}$$

If A is a set of units, $A \subseteq U$, a sum $\sum_{k \in A}$ will be written as $\sum_A$. The survey may have many study variables. A typical one, denoted y (continuous or categorical), has value y_k recorded for $k \in r$ but missing for $k \in s - r$. Our objective is to estimate the population y-total, $Y = \sum_U y_k$. The response indicator I has value $I_k = 1$ for $k \in r$, $I_k = 0$ for $k \in s - r$. A goal for practice is to get a response r that is well balanced, in the sense specified later. We are led to consider the different r that may arise from a given s.

The auxiliary vector $\mathbf{x}$ of dimension $J \geq 1$ has value $\mathbf{x}_k$ known at least for all units $k \in s$. Auxiliary information can be used in the data collection (for monitoring the data inflow to achieve improved balance) and/or in the estimation (for calibrated weight computation). The auxiliary vector need not be the same for the two purposes, but this article assumes that they agree, and that the $\mathbf{x}$-information used is for $k \in s$. This includes the important case of paradata, that is, data about the data collection process.

An important type of auxiliary vector is a *group vector*. It identifies membership of every unit k in one of J mutually exclusive and exhaustive sample groups, so that $\mathbf{x}_k = (0, \ldots, 1, \ldots 0)'$, where the only "1" indicates the unique group (out of J possible) to which k belongs.

A group vector occurs when several categorical auxiliary variables are completely crossed. To illustrate, if $\mathbf{x} = (sex \times education \times age)$ represents a crossing of 2 sexes, 3 exhaustive education categories and 4 exhaustive age categories, then $\mathbf{x}$ is a group vector with dimension $J = 2 \times 3 \times 4 = 24$ and equally many possible values $\mathbf{x}_k$. When several categorical variables are used although not in completely crossed manner – important for practice in statistical agencies – then the dimension J of $\mathbf{x}_k$ can be kept relatively modest (say less than 15) while still coding a much larger number (say more than one hundred) of possible properties $\mathbf{x}_k$ of the units k. For a study of the Swedish Living Conditions Survey, Särndal and Lundquist

(2014a) used an $\mathbf{x}$-vector of dimension 14 with 256 possible values. The group vector case and the non-group vector case give important differences in the results that follow.

All auxiliary vectors used here satisfy a requirement that grants mathematical convenience without severely restricting the choice of vector: There exists a constant vector $\boldsymbol{\mu}$ such that

$$\boldsymbol{\mu}'\mathbf{x}_k = 1 \text{ for all } k. \tag{2.2}$$

For example, when $J = 2$ and $\mathbf{x}_k = (1, x_k)'$, then $\boldsymbol{\mu} = (1,0)'$ satisfies the requirement. In the group vector case where $\mathbf{x}_k = (0,\ldots,1,\ldots,0)'$, then $\boldsymbol{\mu} = (1,\ldots,1,\ldots,1)'$ satisfies the requirement. If $\mathbf{x}$ is not a group vector, say, one used to code "education" with three mutually exclusive and exhaustive categories and "gender" as a univariate variable equal to 1 or 0, then $J = 3+1 = 4$ (education and gender not crossed), and $\boldsymbol{\mu} = (1,1,1,0)'$ satisfies the requirement.

3 Imbalance

The concept of *balance* has been often used in statistical literature with reference to an equality of means of specified variables for two sets of units, one a subset of the other. One method to realize a probability sample s from U that is balanced with respect to a vector $\mathbf{x}$ is the Cube Method, see Deville and Tillé (2004). In the context with nonresponse, we want to know how well balanced a response r is, compared with the probability sample s that would have given unbiased estimates. A given auxiliary vector $\mathbf{x}$ has computable means $\bar{\mathbf{x}}_r = \sum_r d_k \mathbf{x}_k \big/ \sum_r d_k$ for the response and $\bar{\mathbf{x}}_s = \sum_s d_k \mathbf{x}_k \big/ \sum_s d_k$ for the sample. If they are equal, an unlikely outcome, the response is perfectly balanced with respect to $\mathbf{x}$. The contrast between response r and sample s can be measured by the scalar quantities

$$Q_s = \left(\bar{\mathbf{x}}_r - \bar{\mathbf{x}}_s\right)' \boldsymbol{\Sigma}_s^{-1} \left(\bar{\mathbf{x}}_r - \bar{\mathbf{x}}_s\right); \quad Q_r = \left(\bar{\mathbf{x}}_r - \bar{\mathbf{x}}_s\right)' \boldsymbol{\Sigma}_r^{-1} \left(\bar{\mathbf{x}}_r - \bar{\mathbf{x}}_s\right). \tag{3.1}$$

They differ only in the $J \times J$ weighting matrix, $\boldsymbol{\Sigma}_s = \sum_s d_k \mathbf{x}_k \mathbf{x}_k' \big/ \sum_s d_k$ as opposed to $\boldsymbol{\Sigma}_r = \sum_r d_k \mathbf{x}_k \mathbf{x}_k' \big/ \sum_r d_k$, both assumed non-singular. In particular, Q_s is important for the statistic called *imbalance* of r with respect to the specified $\mathbf{x}$-vector:

$$IMB(r, \mathbf{x} \mid s) = P^2 \left(\bar{\mathbf{x}}_r - \bar{\mathbf{x}}_s\right)' \boldsymbol{\Sigma}_s^{-1} \left(\bar{\mathbf{x}}_r - \bar{\mathbf{x}}_s\right) = P^2 Q_s, \tag{3.2}$$

where P is the response rate (2.1); see for example Särndal and Lundquist (2014a). The full notation $IMB(r, \mathbf{x} \mid s)$ emphasizes that imbalance depends on the realized response r and on the choice of $\mathbf{x}$-vector. Unless required for emphasis, we use the simpler notation IMB. We have $0 \le IMB \le P(1-P)$ for any r and vector formulation $\mathbf{x}$, given s. IMB is a descriptive measure of the response r. It is related to a special case of the R-indicator, whose motivation lies instead in the estimation of (the unknown) response probabilities for the population units, see for example Bethlehem et al. (2011).

The IMB statistic (3.2) can be continuously computed and monitored in a data collection extending over a period of time, say several days or weeks, during which contact attempts continue with a sample unit until desired data are obtained, or, if this fails, until the unit is declared a non-respondent. As the response rate P grows, IMB serves as a tool for monitoring and managing the data collection to achieve in the end a response set r which, if not perfectly balanced to satisfy $\bar{\mathbf{x}}_r = \bar{\mathbf{x}}_s$, will at least have considerably lower IMB than if balancing had not been attempted in data collection. There are methods for balancing based on response propensity, such as the Threshold method and the Equal proportions method in Särndal and Lundquist (2014a, 2014b).

We consider later the particular case where s is a self-weighting sample (as when s is a simple random sample), the response r has fixed size m, and $\mathbf{x}$ is a group vector of dimension J as defined in Section 2. Then both s and r are split into J non-overlapping groups. For the sample group s_j, denote by n_j the size and by $W_{js} = n_j / n$ the relative size; $\sum_{j=1}^{J} n_j = n$. For the response group r_j, let $m_j \le n_j$ be the size; $\sum_{j=1}^{J} m_j = m$. The imbalance (3.2) is then

$$IMB = \sum_{j=1}^{J} W_{js} \left(p_j - p \right)^2, \tag{3.3}$$

where the response rates are $p_j = m_j / n_j$ in group j and $p = m/n$ overall. (The response rate P is defined in (2.1) with general design weights d_k; for a self-weighting sample, where d_k is constant, we use small p for the response rate.) If $IMB = 0$, we have perfect balance; all group response rates p_j are then equal.

4 The regression aspect

The imbalance (*IMB*) is determined by the auxiliary vector $\mathbf{x}$ with no attention paid to the study variable y. But the relation of $\mathbf{x}$ to y is also important for the bias of estimated y-totals. Strong regression of y on $\mathbf{x}$ is likely to give small bias, intuitively because regression predicted y-values can then give close substitutes for those missing. For some survey data, the strength of the regression may be modest but nevertheless important in its effect on bias. The ordinary linear regression coefficient vectors for the whole sample s and for the response r are, respectively,

$$\mathbf{b}_s = \left(\sum_s d_k \mathbf{x}_k \mathbf{x}_k' \right)^{-1} \sum_s d_k \mathbf{x}_k y_k; \quad \mathbf{b}_r = \left(\sum_r d_k \mathbf{x}_k \mathbf{x}_k' \right)^{-1} \sum_r d_k \mathbf{x}_k y_k. \tag{4.1}$$

Under nonresponse, $\mathbf{b}_r$ is computable but not $\mathbf{b}_s$. The $J \times J$ matrices to invert are assumed non-singular. Normally $\mathbf{b}_r \ne \mathbf{b}_s$, perhaps with considerable (but unknown) difference. The regression based on the response is inconsistent.

The imbalance in the y-variable is $\bar{y}_r - \bar{y}_s$, where the means are $\bar{y}_s = \sum_s d_k y_k / \sum_s d_k$ for the sample (unknown) and $\bar{y}_r = \sum_r d_k y_k / \sum_r d_k$ for the response (computable). The decomposition

$$\bar{y}_r - \bar{y}_s = \left(\bar{\mathbf{x}}_r - \bar{\mathbf{x}}_s\right)' \mathbf{b}_r + \left(\mathbf{b}_r - \mathbf{b}_s\right)' \bar{\mathbf{x}}_s, \tag{4.2}$$

highlights two undesirable differences, $\bar{\mathbf{x}}_r - \bar{\mathbf{x}}_s$ (due to imbalance in the $\mathbf{x}$-vector), and $\mathbf{b}_r - \mathbf{b}_s$ (due to inconsistent regression); to obtain (4.2) note that $\bar{\mathbf{x}}_r'\mathbf{b}_r = \bar{y}_r$ and $\bar{\mathbf{x}}_s'\mathbf{b}_s = \bar{y}_s$, which are consequences of the $\mathbf{x}$-vector condition (2.2).

5 Estimating the population total under nonresponse

The equation (4.2), when multiplied by $\hat{N} = \sum_s d_k$, can be expressed in terms of three common estimators of the population total $Y = \sum_U y_k$. Two are possible under nonresponse,

$$\hat{Y}_{EXP} = \hat{N} \frac{\sum_r d_k y_k}{\sum_r d_k}, \quad \hat{Y}_{CAL} = \hat{N} \frac{\sum_r d_k g_k y_k}{\sum_r d_k}, \tag{5.1}$$

with $g_k = \bar{\mathbf{x}}_s' \boldsymbol{\Sigma}_r^{-1} \mathbf{x}_k$. Of these, $\hat{Y}_{EXP}$ is just a simple expansion of the response mean of y and often considerably biased. The calibration estimator $\hat{Y}_{CAL}$ gives y_k the weight $d_k g_k / P$. The calibration property is $\sum_r \left(d_k g_k / P\right) \mathbf{x}_k = \sum_s d_k \mathbf{x}_k$, where the right hand side is unbiased for the population $\mathbf{x}$-total $\sum_U \mathbf{x}_k$, which explains why $\hat{Y}_{CAL}$ can be considerably less biased than $\hat{Y}_{EXP}$ when $\mathbf{x}$ and y are well related. If y-values had been recorded for the full sample s, unbiased estimation would be carried out with the Horvitz-Thompson estimator

$$\hat{Y}_{FUL} = \sum_s d_k y_k.$$

The three estimator types will be referred to as EXP, CAL and FUL. Now (4.2) multiplied by $\hat{N} = \sum_s d_k$ reads

$$\hat{Y}_{EXP} - \hat{Y}_{FUL} = \left(\hat{Y}_{EXP} - \hat{Y}_{CAL}\right) + \left(\hat{Y}_{CAL} - \hat{Y}_{FUL}\right). \tag{5.2}$$

In words, Deviation of $EXP =$ Bias adjustment term + Deviation of CAL. The computable adjustment is $\hat{Y}_{EXP} - \hat{Y}_{CAL} = \hat{N}\left(\bar{\mathbf{x}}_r - \bar{\mathbf{x}}_s\right)' \mathbf{b}_r$. The two deviations from the unbiased estimate, $\hat{Y}_{CAL} - \hat{Y}_{FUL} = \hat{N}\left(\mathbf{b}_r - \mathbf{b}_s\right)' \bar{\mathbf{x}}_s$ for CAL and $\hat{Y}_{EXP} - \hat{Y}_{FUL} = \hat{N}\left(\bar{y}_r - \bar{y}_s\right)$ for EXP, are not computable under nonresponse, because they require y-values for the full sample.

As mentioned, we have methods to reduce the imbalance IMB during data collection. Low imbalance is intuitively attractive, but does it yield better accuracy in estimates? Or is it enough to involve the auxiliary variables at the estimation stage, through a calibrated weight adjustment as in the CAL estimator? The adjustment term $\hat{Y}_{EXP} - \hat{Y}_{CAL} = \hat{N}\left(\bar{\mathbf{x}}_r - \bar{\mathbf{x}}_s\right)' \mathbf{b}_r$ can clearly be reduced by constructing r to have low imbalance; it is zero for the perfect balance $\bar{\mathbf{x}}_r = \bar{\mathbf{x}}_s$. In practice, the CAL estimator is preferred to the EXP estimator, the former being usually more accurate because of the auxiliary information. But is the deviation $\hat{Y}_{CAL} - \hat{Y}_{FUL} = \hat{N}\left(\mathbf{b}_r - \mathbf{b}_s\right)' \bar{\mathbf{x}}_s$ smaller if the response r had been built to have low IMB? Asked

differently, is it worth the (perhaps costly) effort to manage the data collection to get $\bar{\mathbf{x}}_r$ closer to $\bar{\mathbf{x}}_s$ and therefore reduced IMB? The question is essentially whether this would also make $\mathbf{b}_r$ and $\mathbf{b}_s$ move closer.

6 Statistical properties of the CAL estimator deviation

In the decomposition (5.2), the deviation of CAL from the unbiased FUL is $\hat{Y}_{CAL} - \hat{Y}_{FUL} = \hat{N}\Delta_r$ where $\boldsymbol{\Delta}_r = (\mathbf{b}_r - \mathbf{b}_s)' \bar{\mathbf{x}}_s$. To see if $\boldsymbol{\Delta}_r$ is smaller, or likely to be so, by realizing low imbalance in data collection, we seek analytic results about statistical properties, such as mean and variance, of $\boldsymbol{\Delta}_r$ as a function of the IMB statistic (3.2). Highly general results of this kind are hard to obtain. Several factors complicate the analysis, such as the sampling design used to draw s, the probability distribution of the response sets r given s, the make-up of the auxiliary vector $\mathbf{x}$, and so on. Results for special situations are obtained in Sections 7 and 8.

Result 1 in Section 7 gives properties – expected value and variance – of $\boldsymbol{\Delta}_r = (\mathbf{b}_r - \mathbf{b}_s)' \bar{\mathbf{x}}_s$ over response outcomes r with fixed size m and fixed mean $\bar{\mathbf{x}}_r$ when $\mathbf{x}$ is a group vector, and s is a simple random sample. The mean of $\boldsymbol{\Delta}_r$ over such outcomes is zero. The imbalance appears in the variance of $\boldsymbol{\Delta}_r$, which is linearly increasing in IMB, approximately. A reason for taking $\mathbf{x}$ to be a group vector is that conditioning on $\bar{\mathbf{x}}_r$ grants relatively simple derivations. A fixed $\bar{\mathbf{x}}_r$ implies a fixed value IMB. (But the opposite is not true; several $\bar{\mathbf{x}}_r$ can give the same IMB.) Another simplification when $\mathbf{x}$ is a group vector is due to diagonal matrices $\boldsymbol{\Sigma}_r$ and $\boldsymbol{\Sigma}_s$. The empirical test in Section 9.1 addresses Result 1.

Simple derivations for the group vector are at the expense of generality. The $\mathbf{x}$-vectors used in production at Statistics Sweden, for example, are often not group vectors. To get transparent mathematical results about $\boldsymbol{\Delta}_r$ is then more difficult.

Result 2 in Section 8 is derived under a model of linear regression between y and $\mathbf{x}$. The y_k are then considered random, with properties stated by the model. A group vector feature for $\mathbf{x}$ is no longer necessary. The conclusions are in some respects similar to those in Result 1. The empirical Test situation 2 in Section 9.2 refers to both Results 1 and 2.

7 The first result

Result 1 refers to the following survey context: A self-weighting sample s of size n is drawn from $U = \{1,\ldots,k,\ldots,N\}$; d_k is the same for all k. The auxiliary vector $\mathbf{x}$ is a group vector of dimension J, so the sample s and the response set r, assumed to be of fixed size $m < n$, are split into J non-overlapping groups. The notation for these is given at the end of Section 3. The values y_k are treated as fixed, non-random, as is usual in the design-based tradition. If y_k were observed for all $k \in s$, then $\hat{Y}_{FUL} = N\,\bar{y}_s$ with $\bar{y}_s = \sum_s y_k \big/ n$ would be design unbiased for the population y-total $Y = \sum_U y_k$. But y_k is available for

$k \in r$ only; the CAL estimator (5.1) becomes $\hat{Y}_{CAL} = N\sum_{j=1}^{J} W_{js}\bar{y}_{r_j}$, where $\bar{y}_{r_j}$ is the mean of respondent values y_k in group j. Statistical properties – expected value and variance – of $\left(\hat{Y}_{CAL} - \hat{Y}_{FUL}\right)/N = \Delta_r$ with $\Delta_r = \left(\mathbf{b}_r - \mathbf{b}_s\right)' \bar{\mathbf{x}}_s = \sum_{j=1}^{J} W_{js}\bar{y}_{r_j} - \bar{y}_s$ are given in Result 1 for the following probabilistic setting: All $\binom{n}{m}$ response sets r of fixed size m are assumed a priori equally probable. Given s, the imbalance IMB is determined by $\bar{\mathbf{x}}_r = (1/m)\left(m_1,\ldots,m_j,\ldots,m_J\right)'$. Given $\bar{\mathbf{x}}_r$, we are left with $R = \prod_{j=1}^{J}\binom{n_j}{m_j}$ sets r, all with the same non-zero probability $1/R$ and the same IMB, given by (3.3). The other sets r of size m are no longer in scope. Conditioning on $\bar{\mathbf{x}}_r$ allows us to study the properties of CAL as a function of IMB. Result 1 involves the variance of the study variable y, within-group and combined over groups:

$$S_{yj}^2 = \sum\nolimits_{s_j}\left(y_k - \bar{y}_{s_j}\right)^2 \Big/ \left(n_j - 1\right); \quad S_y^2 = \sum_{j=1}^{J} W_{js}S_{yj}^2. \tag{7.1}$$

Result 1. Let s be a self-weighting sample of size n and let $\mathbf{x}_k$ be a group vector of dimension J. Assume that all $\binom{n}{m}$ response sets r of fixed size m are a priori equally probable. Then

$$\bar{\Delta} = \mathrm{E}\left(\Delta_r \,|\, \bar{\mathbf{x}}_r, m, s\right) = 0 \tag{7.2}$$

$$S_\Delta^2 = \mathrm{E}\left(\left(\Delta_r - \bar{\Delta}\right)^2 \,\middle|\, \bar{\mathbf{x}}_r, m, s\right) = \left(\frac{1}{m} - \frac{1}{n}\right)S_y^2 + \frac{1}{m}\sum_{j=1}^{J} W_{js}\left(\frac{p}{p_j} - 1\right)S_{yj}^2 \tag{7.3}$$

where $W_{js} = n_j/n$ and $p_j = m_j/n_j$ are relative size and response rate, respectively, for group j, $p = m/n$ is the overall response rate, and S_y^2 and S_{yj}^2 are given in (7.1). If response rates p_j and variances S_{yj}^2 vary by little only over the groups, then

$$S_\Delta^2 \approx \left(1 - p + \frac{IMB}{p^2}\right)\frac{S_y^2}{m} \tag{7.4}$$

where IMB is given by (3.3).

For full response, when $r = s$, the right hand sides of (7.3) and (7.4) are zero; the approximation in (7.4) is exact: $S_\Delta^2 = 0$. To interpret Result 1, note that the first term on the right hand side of (7.3) is a constant, given m. It states the conditional variance for a perfectly balanced response, where p_j is the same for all groups. The second is the *penalty term*, namely the penalty for failing to get perfect balance in data collection. Its size depends on how well an adaptive design succeeds in generating group response rates p_j that vary little only. It is zero if all p_j can be made equal.

Formula (7.4) states that the variance S_Δ^2 is decreasing with IMB in a roughly linear fashion. Thus low imbalance brings improved chances for a small deviation $\hat{Y}_{CAL} - \hat{Y}_{FUL} = \hat{N}\Delta_r$. This is important for practice. To illustrate, for a nonresponse of $1 - p = 40$ per cent, $S_\Delta^2 \approx 0.57S_y^2/m$ if $IMB = 0.06$, but if $IMB = 0$, as in perfect balance, that variance is reduced to $S_\Delta^2 \approx 0.40S_y^2/m$. The improvement is clear but cannot be claimed to be very large. This is because with most data, IMB/p^2 is small compared with a nonresponse $1 - p$ of the order of 30 to 60 per cent, cases that we are mainly concerned with here. Thus taking action to reduce imbalance has a desirable effect, although modest rather than strong.

In (7.2) and (7.3), the expectation $\mathrm{E}(\cdot)$ is taken by averaging over the $R = \prod_{j=1}^{J} \binom{n_j}{m_j}$ equi-probable sets r that remain out of $\binom{n}{m}$ after fixing $\bar{\mathbf{x}}_r$. It should also be noted that more than one $\bar{\mathbf{x}}_r$ can give the same value IMB. Hence there may be more than one value S_Δ^2 for the same IMB. The linearly increasing function of IMB in (7.4) is nevertheless their common approximation.

8 The second result

In Result 1, the survey variable values y_k are treated as fixed, nonrandom. In Result 2, they are random with properties as stated in a linear regression model ξ with residuals $\varepsilon_k = y_k - \mathbf{x}_k'\boldsymbol{\beta}$ for some unknown $\boldsymbol{\beta}$:

$$E_\xi(y_k \mid \mathbf{x}_k) = \mathbf{x}_k'\boldsymbol{\beta}; \quad E_\xi(\varepsilon_k^2 \mid \mathbf{x}_k) = \sigma_\varepsilon^2, \text{ all } k \in s; \quad E_\xi(\varepsilon_k \varepsilon_\ell \mid \mathbf{x}_k, \mathbf{x}_\ell) = 0, \text{ all } k \neq \ell \in s. \tag{8.1}$$

The properties in (8.1) apply also to units k and ℓ belonging in any subset r of s. Result 2 presents expected value and approximate variance of $\boldsymbol{\Delta}_r = (\mathbf{b}_r - \mathbf{b}_s)'\bar{\mathbf{x}}_s$ conditionally on a fixed self-weighting sample s and a fixed response set r with respective sizes n and m.

Result 2: Let s of size n be a self-weighting sample. Let $\mathbf{X}$ be the $J \times n$ $\mathbf{x}$-data matrix with columns $\mathbf{x}_k$, $k \in s$. Then, under the model ξ in (8.1),

$$E_\xi(\Delta_r \mid \mathbf{X}, r, s) = 0; \quad E_\xi(\Delta_r^2 \mid \mathbf{X}, r, s) \approx \left(1 - p + \frac{IMB}{p^2}\right)\frac{\sigma_\varepsilon^2}{m}, \tag{8.2}$$

where m is the size of the fixed response set r, $p = m/n$ is the response rate and IMB is given by (3.2).

Result 2 (for arbitrary $\mathbf{x}$-vector and random y_k) mirrors Result 1 (for group $\mathbf{x}$-vector and non-random y_k) in that both give conditional mean zero and the same linearly increasing form for the conditional variance approximation.

The derivation in Appendix 3 of Result 2 relies on a comparison of the two quadratic forms in $\bar{\mathbf{x}}_r - \bar{\mathbf{x}}_s$ given in (3.1), Q_s and Q_r. The former is used in the imbalance statistic (3.2), $IMB = P^2 Q_s$; the latter determines the weight factors g_k for the CAL estimator (5.1). The approximation $Q_r \approx Q_s$, needed for Result 2, is justified in Appendix 2.

9 Empirical testing

Results 1 and 2 give the basis for testing empirically in this section how mean and variance of the deviation $\hat{Y}_{CAL} - \hat{Y}_{FUL} = \hat{N}\Delta_r = \hat{N}(\mathbf{b}_r - \mathbf{b}_s)'\bar{\mathbf{x}}_s$ depend on the imbalance IMB. Both results state that the variance of $\boldsymbol{\Delta}_r$ increases in a roughly linear fashion as IMB increases, without being small even if IMB is near zero.

We use real data from an Estonian survey with 17,540 households. The following variables are available for every household: Household net income, used here as the study variable y, and three categorical variables referring to the designated head of household, used here as auxiliary variables: (i) Gender (1 for male, 0 for female), (ii) Economic activity (1 for employed, 0 for not employed) and (iii) Education, with three exhaustive levels: low, medium, high.

We compute the mean $\bar{\Delta}$ of $\mathbf{\Delta}_r$ and the variance S_Δ^2 of $\mathbf{\Delta}_r$ by averaging over the sets r with fixed mean $\bar{\mathbf{x}}_r$, given s.

9.1 Test situation 1

In line with Result 1, we want to consider the response sets r with fixed size m arising from a given sample s of size n. The computational volume is prohibitive even for rather small n. We drew s as a simple random sample of size $n = 20$ from 17,540. The d_k are then constant. The sample mean for the y-variable (household income) was $\bar{y}_s = 10,386.65$. We define $\mathbf{x}_k$ as the group vector of dimension $J = 3$ that identifies the three exhaustive levels of Education; low, medium, high. For the realized sample s, we have $n\bar{\mathbf{x}}_s = (5,8,7)'$.

We fixed the size of the response sets r to be $m = 12$. The response rate is 60 per cent for every one of the $\binom{20}{12} \approx 1.26 \times 10^5$ possible response sets r. From these, we excluded all those for which the response count vector $m\bar{\mathbf{x}}_r$ contained a zero, to avoid a singular $\mathbf{\Sigma}_r$. This left 31 configurations (m_1, m_2, m_3) such that $m_1 + m_2 + m_3 = 12$ and all three counts $m_j \geq 1$. For each of the 31 possibilities, we computed $\bar{\Delta}$ and S_Δ^2 by averaging over the response sets r satisfying the fixed configuration. For example, $(m_1, m_2, m_3) = (3,4,5)$ is satisfied by 14,700 response sets r, so mean and variance of Δ_r are computed over those. Other configurations give much fewer response sets, for example, only 70 for the configuration (3, 8, 1); a few of those can then be very influential in the computations. For every one of the 31 cases, $\bar{\Delta}$ is theoretically zero, by Result 1. The computations confirmed this; a plot of $\bar{\Delta}$ against IMB is unnecessary. Figure 9.1 shows the 31 point plot of S_Δ^2 against IMB. Because of the non-uniqueness of IMB noted earlier, it happens several times that more than one S_Δ^2 occurs at the same IMB value. Figure 9.1 shows that S_Δ^2 has a clear upward trend as IMB increases. Figure 9.1 also shows the approximation $S_\Delta^2 \approx S_{\Delta approx}^2 = \left(S_y^2 / m\right)(1 - p + IMB / p^2)$ from Result 1. We have $p = 0.6, m = 12$ and $S_y^2 = 26.3 \times 10^6$, so the computed approximation, linear in IMB, is $S_{\Delta approx}^2 = a + b\, IMB$ with $a = 0.879 \times 10^6$ and $b = 6.102 \times 10^6$. For points with low IMB, S_Δ^2 agrees closely with the linearly increasing $S_{\Delta approx}^2$. A contributing reason is that when IMB is low, the group response rates p_j vary little, and this is one of the conditions for close approximation, as the derivation of Result 1 in Appendix 1 explains. For higher IMB values, the increasing trend in S_Δ^2 is still evident, but the scatter around the theoretical line is more pronounced. Five outlying points in Figure 9.1 have very large S_Δ^2; three of them occur when one component of (m_1, m_2, m_3) is equal to the maximal count (5 or 8 or 7). For those, less accurate linear approximation is expected, the p_j being far from equal.

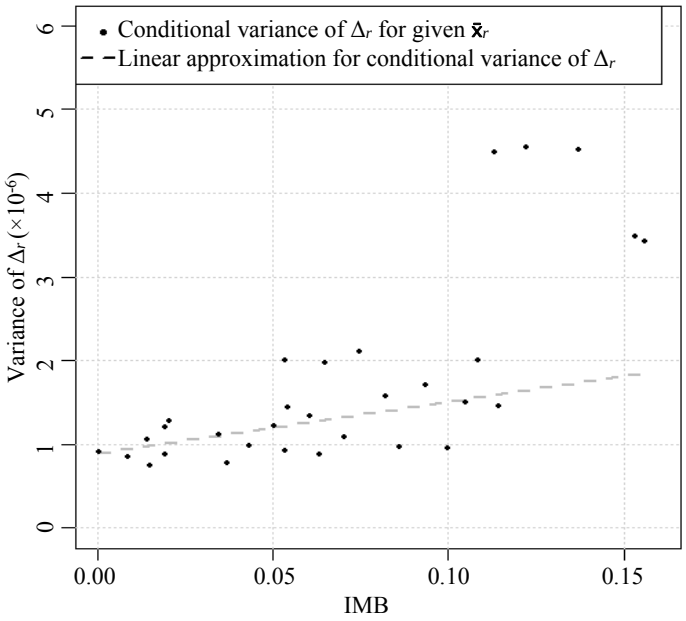

Figure 9.1 Conditional variance of Δ_r as a function of imbalance *IMB*; $\mathbf{x}_k$ a group vector of dimension 3; response sets r of fixed size 12 from a fixed sample s of size 20.

9.2 Test situation 2

The setup and the computational steps are similar to those in Test situation 1, but $\mathbf{x}_k$ is no longer a group vector; some results change considerably, compared with Test situation 1.

A new simple random sample s of size $n = 20$ was drawn from the 17,540 households. For this sample, $\bar{y}_s = 9{,}618.4$. We let $\mathbf{x}_k$ incorporate all three auxiliary variables (i), (ii) and (iii), but not completely crossed: Gender (univariate coded 0 or 1), Economic activity (univariate coded 0 or 1) and Education level (three exhaustive categories coded (1,0,0) or (0,1,0) or (0,0,1)). This $\mathbf{x}_k$ is not a group vector; it has dimension $1+1+3=5$ and $2\times 2\times 3 = 12$ possible values; $\boldsymbol{\Sigma}_r$ and $\boldsymbol{\Sigma}_s$ are not diagonal. We have $n\bar{\mathbf{x}}_s = (9,11,4,7,9)'$. For this sample s we considered the response sets r of fixed size $m = 12$ excepting those where one or more of the five components of the count vector $m\bar{\mathbf{x}}_r$ are zero. This left 658 different vectors $m\bar{\mathbf{x}}_r$, each composed of five non-zero counts, and satisfied by a certain number of response sets r over which we computed, by simple averaging, the mean $\bar{\Delta}$ and variance S^2_{Δ}. These are thus moments conditionally on $\bar{\mathbf{x}}_r$.

Figure 9.2 shows the 658 point plot of $\bar{\Delta}$ against *IMB*. In Test situation 1, $\bar{\Delta}$ was zero for every point because $\mathbf{x}_k$ was a group vector. This is not so in Figure 9.2, where the means $\bar{\Delta}$ fan out when *IMB* increases. They are much more concentrated around zero for low *IMB* than for large *IMB*. Several points (that is several means $\bar{\mathbf{x}}_r$) can give the same or nearly the same *IMB*. Figure 9.2 shows that in a small neighborhood of a fixed value IMB_0 on the *IMB* axis, the mean of the means $\bar{\Delta}$ is roughly zero. With reference to Result 2, we can expect to see the average of $\bar{\Delta}$ for fixed *IMB* to be near zero: Under model

(8.1) for y_k, Result 2 says that $E_\xi(\Delta_r \mid \mathbf{X}, r, s) = 0$. When $\mathbf{X}$ and r are fixed, so is IMB. If the model is a reasonably good representation, the average of Δ_r for fixed IMB should be close to zero, as Figure 9.2 indicates.

Figure 9.3 shows the plot of the conditional variance S^2_Δ against IMB. The pattern with a variance S^2_Δ that increases linearly in IMB prevails, even though $\mathbf{x}_k$ is not a group vector here. Figure 9.3 shows the computed approximating line $S^2_{\Delta approx} = (\hat{\sigma}^2_\varepsilon / m)(1 - p + IMB/p^2)$ derived from Result 2, with $\hat{\sigma}^2_\varepsilon = \sum_s (y_k - \mathbf{x}'_k \mathbf{b}_s)^2 / (n - J)$ used to estimate σ^2_ε. We have $J = 5$, $p = 0.6$, $m = 12$ and $\hat{\sigma}^2_\varepsilon = 33.6 \times 10^6$, so the line in Figure 9.3 is $S^2_{\Delta approx} = a + b\, IMB$ with $a = 1.12 \times 10^6$ and $b = 7.78 \times 10^6$. The linear approximation holds particularly well for small IMB, say less than 0.1. For large IMB, there is much scatter; S^2_Δ has some very large values, and some very low values as well. Figure 9.4 shows the joint behavior of $\bar{\Delta}$ and S^2_Δ for the 658 points. The size of a dot is proportional to IMB^2; the reason for squaring is to better contrast larger and smaller IMB values. Response sets r with small IMB are found to give small $\bar{\Delta}$ and S^2_Δ, a favourable sign because the CAL and FUL estimators are then close. To illustrate, for points satisfying $IMB \le 0.1$, $\bar{\Delta}$ is in the interval (-1,390; 1,447) and S^2_Δ in (0.846×10^6; 4.86×10^6). These are narrow intervals; this is even more pronounced for $IMB \le 0.05$. But when IMB is large, this advantageous situation no longer holds. For example, $\bar{\Delta}$ can be very small and at the same time S^2_Δ very large (points in the middle and right side of the figure). On the other hand, S^2_Δ can be near zero while $\bar{\Delta}$ is very large in absolute value (points in the top and bottom left parts of the figure.) Test situation 2 illustrates that a non-group vector $\mathbf{x}_k$ can give both a distinctly non-zero mean of Δ_r and a high variance of Δ_r, and that these tendencies are accentuated by large imbalance.

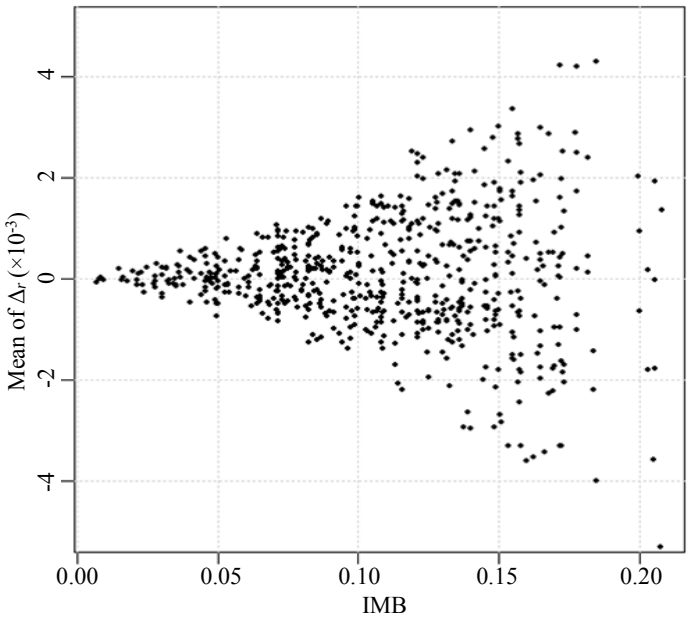

Figure 9.2 Conditional mean of Δ_r as a function of imbalance IMB; $\mathbf{x}_k$ is a non-group vector of dimension 5; response sets r of fixed size 12 from a fixed sample of size 20.

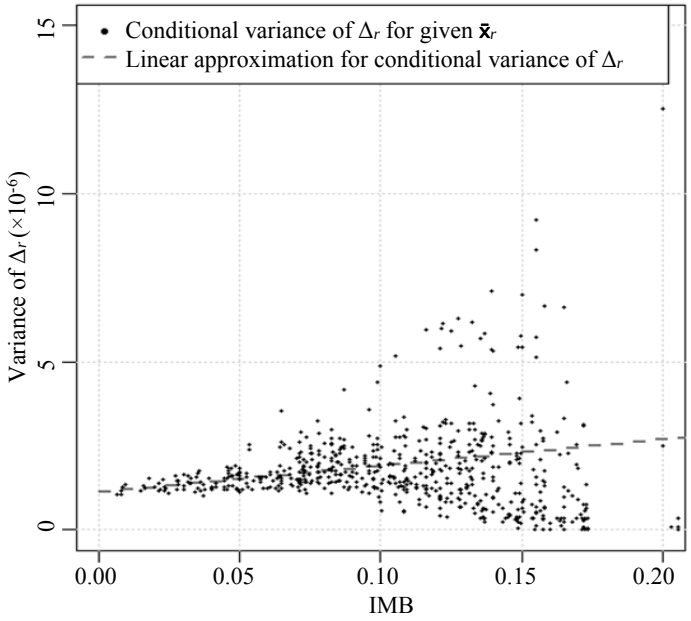

Figure 9.3 Conditional variance of Δ_r as a function of imbalance IMB; $\mathbf{x}_k$ is a non-group vector of dimension 5; response sets r of fixed size 12 from a fixed sample of size 20.

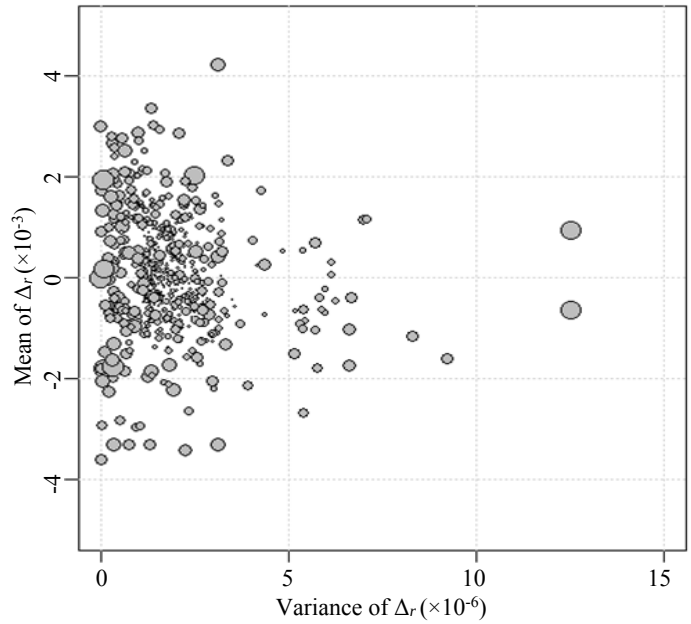

Figure 9.4 Plot of conditional mean of Δ_r against conditional variance of Δ_r; $\mathbf{x}_k$ is a non-group vector of dimension 5; response sets r of fixed size 12 from a fixed sample of size 20. Dot size proportional to imbalance squared.

10 Discussion

We comment on several issues arising and indicate limitations of our study.

1. Choice of variables for the auxiliary vector. The auxiliary variables for the vector $\mathbf{x}$ is treated as a fixed choice in this article. That choice is important when a perhaps large supply of such variables is available. Which ones should be chosen to meet the ultimate objective, which is best possible accuracy in the estimates? Result 1 shows that in the group vector case two factors are important for S_Δ^2 (which determines the conditional variance of CAL): The response imbalance IMB and the variance S_y^2 of the survey variable y. The fact that S_Δ^2 is (approximately) linearly decreasing with IMB gives incentive to try to reduce IMB in data collection. But allowing more variables in $\mathbf{x}$ increases IMB (because agreement is sought on more $\mathbf{x}$-means). As for the y-variance S_y^2, the trend is the opposite. By (7.1), S_y^2 is an averaged residual variance around group means; allowing additional variables in $\mathbf{x}$ will, especially if they explain y well, reduce S_y^2. The two factors work in opposite directions: More auxiliary variables give greater IMB but lower y-variance. It suggests a possible trade-off, a question not examined in this article. A particularity of a group vector $\mathbf{x}$ plays a role: When more categorical variables enter, the vector dimension grows in multiplicative bounds. The risk of small or empty cells restricts the expansion. To illustrate, if $\mathbf{x} = (sex \times education \times age)$ of dimension $J = 2 \times 3 \times 4 = 24$ is expanded to also include *occupation* with 4 categories, in completely crossed fashion, the new dimension (equal to the new number of groups) is $J = 24 \times 4 = 96$. In principle, S_y^2 decreases, but risk of small cells is a good reason to abstain from completely crossing all the variables and instead involve them in a non-group $\mathbf{x}$-vector. That case is addressed in Result 2, which says that if $\mathbf{x}$ explains y well, then σ_ε^2 is small and will give a desired low variance for Δ_r.

2. Auxiliary information at different levels. In this article, the imbalance IMB and the calibration estimator $\hat{Y}_{CAL}$ use the same $\mathbf{x}$-vector, and more particularly one that has auxiliary data for the sample units only. It is a realistic case. But in more general formulations, the data collection would use a monitoring vector $\mathbf{x}_{MV}$ possibly different from the calibration vector $\mathbf{x}_{CAL}$ used later in the estimation. The first is an instrument to get low imbalance IMB in the response, the second serves to get good calibrated weights for $\hat{Y}_{CAL}$. One reason why $\mathbf{x}_{MV}$ and $\mathbf{x}_{CAL}$ may differ in practice is that auxiliary variables for the estimation may be updated versions of the same variables available in the data collection. There may be other reasons to choose $\mathbf{x}_{MV}$ and $\mathbf{x}_{CAL}$ to be different. Also, they can contain information (if available) at the population level. Extensions of our approach to such situations are possible.

3. Uncertain benefit from reduced imbalance. Schouten et al. (2014) find evidence that balancing response reduces bias. We also find that there is incentive to strive, in data collection, for an ultimate response set with low imbalance IMB. As Results 1 and 2 show theoretically, and as test situations 1 and 2 confirm empirically, low imbalance gives a deviation $\hat{Y}_{CAL} - \hat{Y}_{FUL} = \hat{N}\Delta_r$ with zero or almost zero expected value and a small variance. This is our protection against large bias. If IMB were to increase, the variance

tends to increase. The zero expected value of the deviation $\hat{Y}_{CAL} - \hat{Y}_{FUL}$ is an average property. There is no guarantee that the deviation is small for any particular response r with low IMB.

4. Perfect balance does not eliminate the bias. Zero imbalance, $IMB = 0$, implies an equality of means for response and full sample, $\overline{\mathbf{x}}_r = \overline{\mathbf{x}}_s$. If that perfect balance were achieved, the bias adjustment term in (5.2) would be zero; the calibration (CAL) estimator and the expansion (EXP) estimator are identically equal. One can say that if perfect balance is achieved, the power of the auxiliary vector is exhausted, not in its potential for explaining the study variable, but in its potential for distancing itself from the crude EXP estimator, which, although it uses no auxiliary information at all, is as good as the otherwise better choice CAL. However, CAL ≡ EXP is still not ideal. As Result 1 shows, the variance of the CAL deviation is not near zero even if the imbalance IMB is near zero. Perfect balance does not eliminate the deviation of CAL, but small IMB protects against large deviation.

5. Practical implications. In this article we have primarily in mind surveys with a "substantial and non-eradicable nonresponse" that cannot realistically (under time and budget constraints for the survey) be brought to single-digit per cent levels even if large resources are spent. Surveys with 30 per cent or more nonresponse are common today. This is far from an ideal with near 100 per cent response, where imbalance and nonresponse would essentially cease to be issues; the EXP, CAL and FUL estimators would be close.

6. Directions for generalization. Results 1 and 2 show properties of the CAL deviation among response sets under a given formulation of the auxiliary vector. It would be desirable to generalize the results to other situations. Our proofs assume the existence of certain inverse matrices. Extensions to other cases would be possible with the aid of Moore-Penrose generalized inverse.

Acknowledgements

This work was supported by the Estonian Science Foundation grant 9127 and by the Institutional Research Funding IUT34-5 of Estonia. The authors gratefully acknowledge constructive comments from an Associate Editor and a Referee, both anonymous.

Appendix 1

Derivation of Result 1

We derive (7.2) to (7.4) under the conditions and notation in Section 7. The sample s is self-weighting, of size n, and $\mathbf{x}$ is a group vector of dimension J. We assume probability $\binom{n}{m}^{-1}$ for every response set r with fixed size m. We derive the expected value and the variance of $\Delta_r = \left(\mathbf{b}_r - \mathbf{b}_s\right)' \overline{\mathbf{x}}_s = \sum_{j=1}^{J} W_{js} \overline{y}_{r_j} - \overline{y}_s$, where $W_{js} = n_j / n$, conditionally on fixed m and mean $\overline{\mathbf{x}}_r = (1/m)\left(m_1, \ldots, m_j, \ldots, m_J\right)$; $\sum_{j=1}^{J} m_j = m$.

Under that conditioning, $R = \prod_{j=1}^{J} \binom{n_j}{m_j}$ sets r have the same probability, where n_j is the size of sample group s_j; $\sum_{j=1}^{J} n_j = n$. This is identical to the probability structure for stratified simple random sampling of m_j from n_j in stratum j; $j = 1,\ldots,J$. Given m and $\bar{\mathbf{x}}_r$, the expected value and variance of $\bar{y}_{r_j}$ are, respectively, $\bar{y}_{s_j} = \sum_{s_j} y_k \big/ n_j$ and $\left(1/m_j - 1/n_j\right) S_{yj}^2$ with S_{yj}^2 given in (7.1). Thus $\bar{\Delta} = \sum_{j=1}^{J} W_{js} \bar{y}_{s_j} - \bar{y}_s = 0$, which proves (7.2), and $S_\Delta^2 = \sum_{j=1}^{J} W_{js}^2 \left(1/m_j - 1/n_j\right) S_{yj}^2$. Substituting $p_j = m_j / n_j$ and $p = m/n$, and using $S_y^2 = \sum_{j=1}^{J} W_{js} S_{yj}^2$ given in (7.1), we get

$$S_\Delta^2 = \frac{1}{n} \sum_{j=1}^{J} W_{js} \left(\frac{1}{p_j} - 1 \right) S_{yj}^2 = \left(\frac{1}{m} - \frac{1}{n} \right) S_y^2 + \frac{1}{m} \sum_{j=1}^{J} W_{js} \left(\frac{p}{p_j} - 1 \right) S_{yj}^2 . \tag{A.1}$$

This proves (7.3). To analyze the penalty term (second term on right hand side) in (A.1), suppose that the p_j vary little only around the overall rate p. Then $\delta_j = p_j / p - 1$, $j = 1,\ldots,J$, are small quantities, and $1/p_j = 1/p\left(1+\delta_j\right) = (1/p)\left(1 - \delta_j + \delta_j^2 - \delta_j^3 + \ldots\right)$. Keeping terms to second order, $p/p_j - 1 \approx -\delta_j + \delta_j^2$. The penalty term is then approximated as

$$\frac{1}{m} \sum_{j=1}^{J} W_{js} \left(\frac{p}{p_j} - 1 \right) S_{yj}^2 \approx -\frac{1}{m} \sum_{j=1}^{J} W_{js} \left(\frac{p_j}{p} - 1 \right) S_{yj}^2 + \frac{1}{m} \sum_{j=1}^{J} W_{js} \left(\frac{p_j}{p} - 1 \right)^2 S_{yj}^2 . \tag{A.2}$$

Let us further assume that the group variances S_{yj}^2, $j = 1,\ldots,J$, vary little only around their weighted mean S_y^2. Approximating $S_{yj}^2 \approx S_y^2$ in (A.2) we get

$$\frac{S_y^2}{m} \sum_{j=1}^{J} W_{js} \left(\frac{p}{p_j} - 1 \right) \approx -\frac{S_y^2}{m} \sum_{j=1}^{J} W_{js} \left(\frac{p_j}{p} - 1 \right) + \frac{S_y^2}{m} \sum_{j=1}^{J} W_{js} \left(\frac{p_j}{p} - 1 \right)^2 .$$

Here the first term on the right hand side is zero. The second term, equal to $(IMB/p^2)\left(S_y^2/m\right)$ with IMB given in (3.3), becomes a second approximation for the penalty term in (A.1). Therefore, $S_\Delta^2 \approx (1/m - 1/n) S_y^2 + (IMB/p^2)\left(S_y^2/m\right)$. This gives the desired result (7.4).

Appendix 2

Comparing two quadratic forms

We compare the two quadratic forms in $\bar{\mathbf{x}}_r - \bar{\mathbf{x}}_s$, Q_r and Q_s defined in (3.1), and justify the approximation $Q_r \approx Q_s$ needed in the proof in Appendix 3 of Result 2. The respective weighting matrices, $\boldsymbol{\Sigma}_r$ and $\boldsymbol{\Sigma}_s$, are positive definite. Therefore Q_r (or Q_s) can be equal to zero only under the perfect balance $\bar{\mathbf{x}}_r = \bar{\mathbf{x}}_s$. Since $Q_r = Q_s$ for perfect balance, the continuity argument implies that $Q_r \approx Q_s$ for nearly balanced response sets. How close are they more generally?

The CAL estimator (5.1) uses the weight factors $g_k = \bar{\mathbf{x}}_s' \boldsymbol{\Sigma}_r^{-1} \mathbf{x}_k$, defined for all $k \in s$. Their link to Q_r is shown in the second and third expressions in (A.3) below. Consider also the factors $f_k = \bar{\mathbf{x}}_r' \boldsymbol{\Sigma}_s^{-1} \mathbf{x}_k$ for

$k \in s$. They are instrumental for Q_s, and for $IMB = P^2 Q_s$, as the last two expressions in (A.3) show. The following moments of g_k and f_k are verified with the aid of the $\mathbf{x}$-vector condition (2.2):

$$\bar{g}_r = 1, \mathrm{var}_r(g) = Q_r, \ \bar{g}_s = 1 + Q_r; \quad \bar{f}_s = 1, \mathrm{var}_s(f) = Q_s, \ \bar{f}_r = 1 + Q_s. \tag{A.3}$$

For g_k, the means are defined as $\bar{g}_s = \sum_s d_k g_k \big/ \sum_s d_k$, $\bar{g}_r = \sum_r d_k g_k \big/ \sum_r d_k$, and the variances are $\mathrm{var}_s(g) = \sum_s d_k (g_k - \bar{g}_s)^2 \big/ \sum_s d_k$, $\mathrm{var}_r(g) = \sum_r d_k (g_k - \bar{g}_r)^2 \big/ \sum_r d_k$. For the corresponding moments of f_k, replace g_k by f_k. The variances $\mathrm{var}_s(g)$ and $\mathrm{var}_r(f)$ do not have an equally transparent form and will be approximated. Another important property following from (2.2) is $\sum_s d_k f_k g_k \big/ \sum_s d_k = \sum_r d_k f_k g_k \big/ \sum_r d_k = 1$. Those equations and appropriate expressions in (A.3) give

$$\mathrm{cov}_s(f,g) = \sum_s d_k (f_k - \bar{f}_s)(g_k - \bar{g}_s) \big/ \sum_s d_k = 1 - \bar{f}_s \bar{g}_s = -Q_r,$$

$$\mathrm{cov}_r(f,g) = \sum_r d_k (f_k - \bar{f}_r)(g_k - \bar{g}_r) \big/ \sum_r d_k = 1 - \bar{f}_r \bar{g}_r = -Q_s.$$

Now use $\mathrm{cov}_s^2(f,g) \le \mathrm{var}_s(f)\,\mathrm{var}_s(g)$ and the analogous inequality where r replaces s. Using also $\mathrm{var}_s(f) = Q_s$ and $\mathrm{var}_r(g) = Q_r$ from (A.3), we get bounds for the ratio Q_r / Q_s:

$$\frac{Q_s}{\mathrm{var}_r(f)} \le \frac{Q_r}{Q_s} \le \frac{\mathrm{var}_s(g)}{Q_r}. \tag{A.4}$$

For more transparent upper and lower bounds, approximate the two variances in (A.4) by assuming that the coefficient of variation (standard deviation divided by mean) is approximately the same for the response r as for the sample s, and this for both f and g. This assumes a certain stability of the coefficient of variation. Then $\mathrm{var}_s(g) \approx (\bar{g}_s)^2 \mathrm{var}_r(g) \big/ (\bar{g}_r)^2 = (1 + Q_r)^2 Q_r$, so the upper bound in (A.4) is approximately $(1 + Q_r)^2 > 1$. Similarly, $\mathrm{var}_r(f) \approx (\bar{f}_r)^2 \mathrm{var}_s(f) \big/ (\bar{f}_s)^2 = (1 + Q_s)^2 Q_s$, which gives $(1 + Q_s)^{-2} < 1$ as an approximate lower bound in (A.4). The interval approximation for the ratio Q_r / Q_s is therefore

$$Q_r / Q_s \in \left((1 + Q_s)^{-2}, (1 + Q_r)^2 \right).$$

This is to illustrate that the ratio is not far from 1, because for most data both Q_s and Q_r are small compared with 1, Q_r usually the somewhat bigger. Empirical work suggests however that the approximate upper bound $(1 + Q_r)^2$ can often be too low.

Appendix 3

Derivation of Result 2

We derive the expressions in (8.2) under the stated conditions. The sizes of r and s are m and n, respectively; the response rate is $p = m/n$. The deviation of CAL from the unbiased FUL is $\hat{Y}_{CAL} - \hat{Y}_{FUL} = \hat{N} \Delta_r$ where

$$\Delta_r = \left(\mathbf{b}_r - \mathbf{b}_s\right)' \bar{\mathbf{x}}_s = \Sigma_r d_k g_k y_k \,/ \Sigma_r d_k - \Sigma_s d_k y_k \,/ \Sigma_s d_k$$

with $\mathbf{b}_r$ and $\mathbf{b}_s$ given by (4.1), and $g_k = \bar{\mathbf{x}}_s' \boldsymbol{\Sigma}_r^{-1} \mathbf{x}_k$. Note that $\mathbf{b}_s' \bar{\mathbf{x}}_s = \bar{y}_s$ by (2.2). Now $\Sigma_r d_k g_k \mathbf{x}_k' \,/\, \Sigma_r d_k = \Sigma_s d_k \mathbf{x}_k' \,/\, \Sigma_s d_k = \bar{\mathbf{x}}_s'$. Post-multiply that equation by $\boldsymbol{\beta}$ and use the result to get $\Delta_r = \Sigma_r d_k g_k \left(y_k - \mathbf{x}_k'\boldsymbol{\beta}\right) \big/ \Sigma_r d_k - \Sigma_s d_k \left(y_k - \mathbf{x}_k'\boldsymbol{\beta}\right) \big/ \Sigma_s d_k$, which expresses Δ_r in terms of the residuals $\varepsilon_k = y_k - \mathbf{x}_k'\boldsymbol{\beta}$ of the model (8.1):

$$\Delta_r = \frac{\sum_r d_k g_k \varepsilon_k}{\sum_r d_k} - \frac{\sum_s d_k \varepsilon_k}{\sum_s d_k}.$$

Then use the model properties of ε_k in (8.1). From $E_\xi\left(\varepsilon_k \mid \mathbf{x}_k\right) = 0$ for all k it follows that $E_\xi\left(\Delta_r \mid \mathbf{X}, r, s\right) = 0$. To evaluate the variance, use $E_\xi\left(\varepsilon_k^2 \mid \mathbf{x}_k\right) = \sigma_\varepsilon^2$, for all $k \in s$, and $E_\xi\left(\varepsilon_k \varepsilon_\ell \mid \mathbf{x}_k, \mathbf{x}_\ell\right) = 0$, all $k \neq \ell \in s$. This gives

$$E_\xi\left(\Delta_r^2 \mid \mathbf{X}, r, s\right) = \sigma_\varepsilon^2 \frac{\sum_r d_k^2 g_k^2}{\left(\sum_r d_k\right)^2} + \sigma_\varepsilon^2 \frac{\sum_s d_k^2}{\left(\sum_s d_k\right)^2} - 2\sigma_\varepsilon^2 \frac{\sum_r d_k^2 g_k}{\left(\sum_r d_k\right)\left(\sum_s d_k\right)}.$$

Here the d_k cancel out, because constant. The first and second expressions in (A.3) hold for any d_k, in particular constant d_k, so we get $\sum_r g_k \,/ m = 1$ for the mean and $\sum_r g_k^2 \,/ m = Q_r + 1$ for variance plus squared mean. Therefore,

$$E_\xi\left(\Delta_r^2 \mid \mathbf{X}, r, s\right) = \left(\frac{1}{m}\left(1 + Q_r\right) + \frac{1}{n} - 2\frac{1}{n}\right)\sigma_\varepsilon^2 = \left(\frac{1}{m} - \frac{1}{n} + \frac{Q_r}{m}\right)\sigma_\varepsilon^2.$$

As a final step, use the approximation $Q_r \approx Q_s$ justified in Appendix 2, and $IMB = p^2 Q_s$. Then, as claimed in Result 2, $E_\xi\left(\Delta_r^2 \mid \mathbf{X}, r, s\right) \approx \left(1 - p + IMB / p^2\right)\left(\sigma_\varepsilon^2 / m\right)$.

References

Beaumont, J.-F., Bocci, C. and Haziza, D. (2014). An adaptive data collection procedure for call prioritization. *Journal of Official Statistics*, 30, 607-622.

Bethlehem, J., Cobben, F. and Schouten, B. (2011). *Handbook of nonresponse in households surveys*. New York: John Wiley & Sons, Inc.

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

Couper, M.P., and Wagner, J. (2011). Using paradata and responsive design to manage survey nonresponse. Proceedings, 58th World Statistics Congress, International Statistical Institute.

Deville, J.-C., and Tillé, Y. (2004). Efficient balanced sampling. The cube method. *Biometrika*, 91, 893-912.

Groves, R. (2006). Research synthesis: Nonresponse rates and nonresponse error in household surveys. *Public Opinion Quarterly*, 70, 646-675.

Groves, R.M., and Heeringa, S.G. (2006). Responsive design for household surveys: Tools for actively controlling survey errors and costs. *Journal of the Royal Statistical Society, Series A*, 169, 439-457.

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

Lundquist, P., and Särndal, C.-E. (2013). Aspects of responsive design. With applications to the Swedish Living Conditions Survey. *Journal of Official Statistics*, 29, 557-582.

Peytchev, A., Riley, S., Rosen, J., Murphy, J. and Lindblad, M. (2010). Reduction of nonresponse bias in surveys through case prioritization. *Survey Research Methods*, 4, 21-29.

R Core Team (2014). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Rao, R.S., Glickman, M.E. and Glynn, R.J. (2008). Stopping rules for surveys with multiple waves of nonrespondent follow-up. *Statistics in Medicine*, 27, 2196-2213.

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.-E., and Lundström, S. (2005). *Estimation in Surveys with Nonresponse.* New York: John Wiley & Sons, Inc.

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

Särndal, C.-E., and Lundquist, P. (2014b). Balancing the response and adjusting estimates for nonresponse bias: Complementary activities. *Journal de la Société Française de Statistique*, 155(4), 28-50.

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 http://www.statcan.gc.ca/pub/12-001-x/12-001-x2013001-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 http://www.statcan.gc.ca/pub/12-001-x/2009001/article/10887-eng.pdf.

Schouten, B., Cobben, F., Lundquist, P. and Wagner, J. (2016). Does more balanced survey response imply less non-response bias? *Journal of the Royal Statistical Society, Series A*, 179, 727-748.

Wagner, J. (2008). Adaptive survey design to reduce nonresponse bias. Ph. D. Thesis, University of Michigan, Ann Arbor.

Wagner, J., and Raghunathan, T.E. (2010). A new stopping rule for surveys. *Statistics in Medicine*, 29, 1014-1024.