Component of Statistics Canada
Catalogue no. 12-001-X Business Survey Methods Division

Article

Design for estimation: Identifying auxiliary vectors to reduce nonresponse bias

SURVEY
METHODOLOGY

A JOURNAL
PUBLISHED BY
STATISTICS CANADA

Canada

by Carl-Erik Särndal and Sixten Lundström

December 2010

Statistics Canada Statistique Canada

Canada

Survey Methodology, December 2010
Vol. 36, No. 2, pp. 131-144
Statistics Canada, Catalogue No. 12-001-X

Design for estimation:
Identifying auxiliary vectors to reduce nonresponse bias

Carl-Erik Särndal and Sixten Lundström [1]

Abstract

This article develops computational tools, called indicators, for judging the effectiveness of the auxiliary information used to control nonresponse bias in survey estimates, obtained in this article by calibration. This work is motivated by the survey environment in a number of countries, notably in northern Europe, where many potential auxiliary variables are derived from reliable administrative registers for household and individuals. Many auxiliary vectors can be composed. There is a need to compare these vectors to assess their potential for reducing bias. The indicators in this article are designed to meet that need. They are used in surveys at Statistics Sweden. General survey conditions are considered: There is probability sampling from the finite population, by an arbitrary sampling design; nonresponse occurs. The probability of inclusion in the sample is known for each population unit; the probability of response is unknown, causing bias. The study variable (the *y*-variable) is observed for the set of respondents only. No matter what auxiliary vector is used in a calibration estimator (or in any other estimation method), a residual bias will always remain. The choice of a "best possible" auxiliary vector is guided by the indicators proposed in the article. Their background and computational features are described in the early sections of the article. Their theoretical background is explained. The concluding sections are devoted to empirical studies. One of these illustrates the selection of auxiliary variables in a survey at Statistics Sweden. A second empirical illustration is a simulation with a constructed finite population; a number of potential auxiliary vectors are ranked in order of preference with the aid of the indicators.

Key Words: Calibration weighting; Nonresponse adjustment; Nonresponse bias; Auxiliary variables; Bias indicator.

1. Introduction

Large nonresponse is typical of many surveys today. This creates a need for techniques for reducing as much as possible the nonresponse bias in the estimates. Powerful auxiliary information is needed. Administrative data files are a source of such information. The Scandinavian countries and some other European countries, notably the Netherlands, are in an advantageous position. Many potential auxiliary variables (called *x*-variables) can be taken from high quality administrative registers where auxiliary variable values are specified for the entire population. Variables measuring aspects of the data collection is another useful type of auxiliary data. Effective action can be taken to control nonresponse bias. Beyond sampling design, *design for estimation* becomes, in these countries, an important component of the total design. Statistics Sweden has devoted considerable recourses to the development of techniques for selecting the best auxiliary variables.

Many articles discuss weighting in surveys with nonresponse and the selection of "best auxiliary variables". Examples include Eltinge and Yansaneh (1997), Kalton and Flores-Cervantes (2003), and Thomsen, Kleven, Wang and Zhang (2006). Weighting in panel surveys with attrition receives special attention in, for example, Rizzo, Kalton and Brick (1996), who suggest that "the choice of auxiliary variables is an important one, and probably more important than the choice of the weighting methodology". The review by Kalton and Flores-Cervantes (2003) provides many references to earlier work. As in this paper, a calibration approach to nonresponse weighting is favoured in Deville (2002) and Kott (2006).

Some earlier methods are special cases of the outlook in this article, which is based on a systematic use of auxiliary information by calibration at two levels. Recently the search for efficient weighting has emphasized two directions: (i) to provide a more general setting than the popular but limited cell weighting techniques, and (ii) to quantify the search for auxiliary variables with the aid of computable indicators. Särndal and Lundström (2005, 2008) propose such indicators, while Schouten (2007) uses a different perspective to motivate an indicator. An article of related interest is Schouten, Cobben and Bethlehem (2009).

This content of this article has four parts: The general background for estimation with nonresponse is stated in Sections 2 to 4. Indicators for preference ranking of **x**-vectors are presented in Sections 5 and 6, and the computational aspects are discussed. The linear algebra derivations behind the indicators is presented in Sections 7 and 8. The two concluding Sections 9 and 10 present two empirical illustrations. The first (Section 9) uses real data from a large survey at Statistics Sweden. The second (Section 10) reports a simulation carried out on a constructed finite population.

[1] Carl-Erik Särndal, Professor and Sixten Lundström, Senior Methodological Advisor, Statistics Sweden. E-mail: carl.sarndal@scb.se.

2. Calibration estimators for a survey with nonresponse

A probability sample s is drawn from the population $U = \{1, 2, ..., k, ..., N\}$. The sampling design gives unit k the known inclusion probability $\pi_k = \Pr(k \in s) > 0$ and the known design weight $d_k = 1/\pi_k$. Nonresponse occurs. The response set r is a subset of s; how it was generated is unknown. We assume $r \subset s \subset U$, and r non-empty. The (design weighted) response rate is

$$P = \frac{\sum_r d_k}{\sum_s d_k} \tag{2.1}$$

(if A is a set of units, $A \subseteq U$, a sum $\sum_{k \in A}$ will be written as $\sum_A$). Ordinarily a survey has many study variables. A typical one, whether continuous or categorical, is denoted y. Its value for unit k is y_k, recorded for $k \in r$, not available for $k \in U - r$. We seek to estimate the population y-total, $Y = \sum_U y_k$. Many parameters of interest in the finite population are functions of several totals, but we can focus on one such total.

The auxiliary information is of two kinds. To these correspond two vector types, $\mathbf{x}_k^*$ and $\mathbf{x}_k^\circ$. *Population auxiliary information* is transmitted by $\mathbf{x}_k^*$, a vector value known for every $k \in U$. Thus $\sum_U \mathbf{x}_k^*$ is a known population total. Alternatively, we allow that $\sum_U \mathbf{x}_k^*$ is imported from an exterior source and that $\mathbf{x}_k^*$ is a known (observed) vector value for every $k \in s$. *Sample auxiliary information* is transmitted by $\mathbf{x}_k^\circ$, a vector value known (observed) for every $k \in s$; the total $\sum_U \mathbf{x}_k^\circ$ is unknown but is estimated without bias by $\sum_s d_k \mathbf{x}_k^\circ$. The auxiliary vector value combining the two types is denoted $\mathbf{x}_k$. This vector and the associated information is

$$\mathbf{x}_k = \begin{pmatrix} \mathbf{x}_k^* \\ \mathbf{x}_k^\circ \end{pmatrix}; \mathbf{X} = \begin{pmatrix} \sum_U \mathbf{x}_k^* \\ \sum_s d_k \mathbf{x}_k^\circ \end{pmatrix}. \tag{2.2}$$

Tied to the k^{th} unit is the vector $(y_k, \mathbf{x}_k, \pi_k)$. Here, π_k is known for all $k \in U$, y_k for all $k \in r$, the component $\mathbf{x}_k^*$ of $\mathbf{x}_k$ carries population information, the component $\mathbf{x}_k^\circ$ of $\mathbf{x}_k$ carries sample information.

Many $\mathbf{x}$-vectors can be formed with the aid of variables from administrative registers, survey process data or other sources. Among all the vectors at our disposal, we wish to identify the one most likely to reduce the nonresponse bias, if not to zero, so at least to a near-zero value.

We consider vectors having the property that there exists a constant non-null vector $\boldsymbol{\mu}$ such that

$$\boldsymbol{\mu}'\mathbf{x}_k = 1 \text{ for all } k \in U \tag{2.3}$$

"Constant" means that $\boldsymbol{\mu} \neq \mathbf{0}$ does not depend on k, nor on s or r. Condition (2.3) simplifies the mathematical derivations and does not severely restrict $\mathbf{x}_k$. Most $\mathbf{x}$-vectors useful in practice are in fact covered. Examples include: (1) $\mathbf{x}_k = (1, x_k)'$, where x_k is the value for unit k of a continuous auxiliary variable x; (2) the vector representing a categorical x-variable with J mutually exclusive and exhaustive classes, $\mathbf{x}_k = \boldsymbol{\gamma}_k = (\gamma_{1k}, ..., \gamma_{jk}, ..., \gamma_{Jk})'$, where $\gamma_{jk} = 1$ if k belongs to group j, and $\gamma_{jk} = 0$ if not, $j = 1, 2, \ldots, J$; (3) the vector $\mathbf{x}_k$ used to codify two categorical variables, the dimension of $\mathbf{x}_k$ being $J_1 + J_2 - 1$, where J_1 and J_2 are the respective number of classes, and the 'minus-one' is to avoid a singularity in the computation of weights calibrated to the two arrays of marginal counts; (4) the extension of (3) to more than two categorical variables. Vectors of the type (3) and (4) are especially important in statistics production in statistical agencies (the choice $\mathbf{x}_k = x_k$, not covered by (2.3), leads to the nonresponse ratio estimator, known to be a usually poor choice for controlling nonresponse bias, compared with $\mathbf{x}_k = (1, x_k)'$, so excluding the ratio estimator is no great loss).

The calibration estimator of $Y = \sum_U y_k$, computed on the data y_k for $k \in r$, is

$$\hat{Y}_{\text{CAL}} = \sum_r w_k y_k \tag{2.4}$$

with $w_k = d_k\{1 + (\mathbf{X} - \sum_r d_k \mathbf{x}_k)'(\sum_r d_k \mathbf{x}_k \mathbf{x}_k')^{-1} \mathbf{x}_k\}$. The weights w_k are calibrated on both kinds of information: $\sum_r w_k \mathbf{x}_k = \mathbf{X}$, which implies $\sum_r w_k \mathbf{x}_k^* = \sum_U \mathbf{x}_k^*$ and $\sum_r w_k \mathbf{x}_k^\circ = \sum_s d_k \mathbf{x}_k^\circ$. We assume throughout that the symmetric matrix $\sum_r d_k \mathbf{x}_k \mathbf{x}_k'$ is nonsingular (for computational reasons, it is prudent to impose a stronger requirement: The matrix should not be ill-conditioned, or near-singular). In view of (2.3), we have $\hat{Y}_{\text{CAL}} = \sum_r w_k y_k$ with weights $w_k = d_k v_k$ where $v_k = \mathbf{X}'(\sum_r d_k \mathbf{x}_k \mathbf{x}_k')^{-1} \mathbf{x}_k$. The weights satisfy $\sum_r d_k v_k \mathbf{x}_k = \mathbf{X}$, where $\mathbf{X}$ has one or both of the components in (2.2).

A closely related calibration estimator is based on the same two-tiered vector $\mathbf{x}_k$ but with calibration only to the sample level:

$$\tilde{Y}_{\text{CAL}} = \sum_r d_k m_k y_k \tag{2.5}$$

where

$$m_k = \left(\sum_s d_k \mathbf{x}_k\right)' \left(\sum_r d_k \mathbf{x}_k \mathbf{x}_k'\right)^{-1} \mathbf{x}_k. \tag{2.6}$$

The calibration equation then reads $\sum_r d_k m_k \mathbf{x}_k = \sum_s d_k \mathbf{x}_k$, where $\mathbf{x}_k$ has the two components as in (2.2). The auxiliary vector $\mathbf{x}_k$ serves two purposes: To achieve a low variance and a low nonresponse bias. From the variance perspective alone, $\hat{Y}_{\text{CAL}}$ is usually preferred to $\tilde{Y}_{\text{CAL}}$ because the former profits from the input of a known population total $\sum_U \mathbf{x}_k^*$. But this paper studies the bias. From that perspective, we are virtually indifferent between $\hat{Y}_{\text{CAL}}$ and

$\tilde{Y}_{\text{CAL}}$, and we focus on the latter. Under liberal conditions, the difference between the bias of $N^{-1}\hat{Y}_{\text{CAL}}$ and that of $N^{-1}\tilde{Y}_{\text{CAL}}$ is of order n^{-1}, thereby of little practical consequence even for modest sample sizes *n*, as discussed for example in Särndal and Lundström (2005).

An alternative expression for (2.5) is

$$\tilde{Y}_{\text{CAL}} = \left(\sum\nolimits_s d_k \mathbf{x}_k\right)' \mathbf{B}_{\mathbf{x}} \qquad (2.7)$$

where

$$\mathbf{B}_{\mathbf{x}} = \mathbf{B}_{\mathbf{x}|r;d} = \left(\sum\nolimits_r d_k \mathbf{x}_k \mathbf{x}_k'\right)^{-1} \sum\nolimits_r d_k \mathbf{x}_k y_k \qquad (2.8)$$

is the regression coefficient vector arising from the (d_k-weighted) least squares fit based on the data $(y_k, \mathbf{x}_k)$ for $k \in r$.

A remark on the notation: When needed for emphasis, a symbol has two indices separated by a semicolon. The first shows the set of units over which the quantity is computed and the second indicates the weighting, as in $\mathbf{B}_{\mathbf{x}|r;d}$ given by (2.8), and in weighted means such as $\bar{y}_{r;d} = \sum_r d_k y_k / \sum_r d_k$. If the weighting is uniform, the second of the two indices is dropped as in $\bar{y}_U = (1/N)\sum_U y_k$.

3. Points of reference

The most primitive choice of vector is the constant one, $\mathbf{x}_k = 1$ for all *k*. Although inefficient for reducing nonresponse bias, it serves as a benchmark. Then $m_k = 1/P$ for all *k*, where *P* is the survey response rate (2.1), and $\tilde{Y}_{\text{CAL}}$ is the expansion estimator:

$$\tilde{Y}_{\text{EXP}} = (1/P)\sum\nolimits_r d_k y_k = \hat{N}\,\bar{y}_{r;d} \qquad (3.1)$$

where $\hat{N} = \sum_s d_k$ is design unbiased for the population size *N*. The bias of $\tilde{Y}_{\text{EXP}}$ can be large.

At the opposite end of the bias spectrum are the unbiased, or nearly unbiased, estimators obtainable under full response, when $r = s$. They are hypothetical, not computable in the presence of nonresponse. Among these are the GREG estimator with weights calibrated to the known population total $\sum_U \mathbf{x}_k^*$,

$$\hat{Y}_{\text{FUL}} = \sum\nolimits_s d_k g_k y_k$$

where $g_k = 1 + (\sum_U \mathbf{x}_k^* - \sum_s d_k \mathbf{x}_k^*)'(\sum_s d_k \mathbf{x}_k^* \mathbf{x}_k^{*'})^{-1}\mathbf{x}_k^*$, and FUL refers to full response. The unbiased HT estimator (obtained when $g_k = 1$ for all *k*) is

$$\tilde{Y}_{\text{FUL}} = \sum\nolimits_s d_k y_k = \hat{N}\,\bar{y}_{s;d}. \qquad (3.2)$$

It disregards the information $\sum_U \mathbf{x}_k^*$, which may be important for variance reduction. But for the study of bias in this paper, we are indifferent between $\hat{Y}_{\text{FUL}}$ and $\tilde{Y}_{\text{FUL}}$. The difference in bias between the two is of little consequence, even for modest sample sizes. We can focus on $\tilde{Y}_{\text{FUL}}$.

4. The bias ratio

For a given outcome (s, r), consider the estimates $\tilde{Y}_{\text{CAL}}$, $\tilde{Y}_{\text{EXP}}$ and $\tilde{Y}_{\text{FUL}}$ given by (2.5), (3.1) and (3.2) as three points on a horizontal axis. Both $\tilde{Y}_{\text{EXP}}$ (generated by the primitive $\mathbf{x}_k = 1$) and $\tilde{Y}_{\text{CAL}}$ (generated by a better **x**-vector) are computable, but biased. As the **x**-vector improves, $\tilde{Y}_{\text{CAL}}$ will distance itself from $\tilde{Y}_{\text{EXP}}$ and may come near the unbiased but unrealized ideal $\tilde{Y}_{\text{FUL}}$. We consider therefore three deviations: $\tilde{Y}_{\text{EXP}} - \tilde{Y}_{\text{FUL}}$, $\tilde{Y}_{\text{EXP}} - \tilde{Y}_{\text{CAL}}$ and $\tilde{Y}_{\text{CAL}} - \tilde{Y}_{\text{FUL}}$, of which only the middle one is computable. The unknown "deviation total", $\tilde{Y}_{\text{EXP}} - \tilde{Y}_{\text{FUL}}$, is decomposable as "deviation accounted for" (by the chosen **x**-vector) plus "deviation remaining":

$$\tilde{Y}_{\text{EXP}} - \tilde{Y}_{\text{FUL}} = (\tilde{Y}_{\text{EXP}} - \tilde{Y}_{\text{CAL}}) + (\tilde{Y}_{\text{CAL}} - \tilde{Y}_{\text{FUL}}). \qquad (4.1)$$

If computable, $\tilde{Y}_{\text{CAL}} - \tilde{Y}_{\text{FUL}}$ would be of particular interest, as an estimate of the bias remaining in $\tilde{Y}_{\text{CAL}}$ (and in $\hat{Y}_{\text{CAL}}$), whereas $\tilde{Y}_{\text{EXP}} - \tilde{Y}_{\text{FUL}}$ would estimate the usually much larger bias of the benchmark, $\tilde{Y}_{\text{EXP}}$. The bias ratio for a given outcome (*s*, *r*) sets the estimated bias of $\tilde{Y}_{\text{CAL}}$ in relation to that of $\tilde{Y}_{\text{EXP}}$:

$$\text{bias ratio} = \frac{\tilde{Y}_{\text{CAL}} - \tilde{Y}_{\text{FUL}}}{\tilde{Y}_{\text{EXP}} - \tilde{Y}_{\text{FUL}}}. \qquad (4.2)$$

We scale the three deviations by the estimated population size $\hat{N} = \sum_s d_k$ and use the notation $\Delta_T = \Delta_A + \Delta_R$, where *T* suggests "total", *A* "accounted for" and *R* "remaining". Noting that $\sum_r d_k (y_k - \mathbf{x}_k' \mathbf{B}_{\mathbf{x}}) = 0$, we have

$$\Delta_T = \hat{N}^{-1}(\tilde{Y}_{\text{EXP}} - \tilde{Y}_{\text{FUL}}) = \bar{y}_{r;d} - \bar{y}_{s;d};$$

$$\Delta_R = \hat{N}^{-1}(\tilde{Y}_{\text{CAL}} - \tilde{Y}_{\text{FUL}}) = \bar{\mathbf{x}}_{s;d}' \mathbf{B}_{\mathbf{x}} - \bar{y}_{s;d}$$

$$\Delta_A = \hat{N}^{-1}(\tilde{Y}_{\text{EXP}} - \tilde{Y}_{\text{CAL}}) = (\bar{\mathbf{x}}_{r;d} - \bar{\mathbf{x}}_{s;d})' \mathbf{B}_{\mathbf{x}}$$

where $\bar{\mathbf{x}}_{s;d} = \sum_s d_k \mathbf{x}_k / \sum_s d_k$, $\bar{\mathbf{x}}_{r;d} = \sum_r d_k \mathbf{x}_k / \sum_r d_k$, and $\bar{y}_{s;d}$ and $\bar{y}_{r;d}$ are the analogously defined means for the *y*-variable. Then (4.2) takes the form

$$\text{bias ratio} = \frac{\Delta_R}{\Delta_T} = 1 - \frac{\Delta_A}{\Delta_T} = 1 - \frac{(\bar{\mathbf{x}}_{r;d} - \bar{\mathbf{x}}_{s;d})' \mathbf{B}_{\mathbf{x}}}{\bar{y}_{r;d} - \bar{y}_{s;d}}. \qquad (4.3)$$

We have bias ratio = 1 for the primitive vector $\mathbf{x}_k = 1$. Ideally, we want the auxiliary vector $\mathbf{x}_k$ for $\tilde{Y}_{\text{CAL}}$ to give bias ratio ≈ 0. For a given outcome (*s*, *r*) and a given *y*-variable, we take steps in that direction by finding an **x**-vector that makes the computable numerator $\Delta_A = (\bar{\mathbf{x}}_{r;d} - \bar{\mathbf{x}}_{s;d})' \mathbf{B}_{\mathbf{x}}$ large (in absolute value). This is within our reach. But whatever our final choice of **x**-vector, the remaining bias of $\tilde{Y}_{\text{CAL}}$ is unknown. Even with the best available **x**-vector, considerable bias may remain. We have then attempted to do the best possible, under perhaps unfavourable circumstances.

To summarize, for a given outcome (s, r) and a given y-variable, the three deviations have the following features: (i) $\Delta_T = \bar{y}_{r;d} - \bar{y}_{s;d}$ is an unknown constant value, depending on both unobserved and observed y-values; (ii) Δ_A is computable; it depends on y_k for $k \in r$ and on the values $\mathbf{x}_k$ for $k \in s$ of the chosen **x**-vector; (iii) Δ_R cannot be computed; it depends on unobserved values y_k, and on $\mathbf{x}_k$ for $k \in s$.

To follow the progression of the estimates when the **x**-vector improves, consider a given outcome (s, r). The deviation Δ_T can have either sign. Suppose $\Delta_T > 0$, indicating a positive bias in $\tilde{Y}_{\text{EXP}}$, as when large units respond with greater propensity than small ones. When the **x**-vector in $\tilde{Y}_{\text{CAL}}$ becomes progressively more powerful by the inclusion of more and more x-variables, Δ_A tends to increase away from zero and will, ideally, come near Δ_T, indicating a desired closeness of $\tilde{Y}_{\text{CAL}}$ to the unbiased $\tilde{Y}_{\text{FUL}}$. As long as the **x**-vector remains relatively weak, $\Delta_A < \Delta_T$ is likely to hold. When the **x**-vector becomes increasingly powerful, Δ_A moves closer to the fixed Δ_T, a sign of bias nearing zero. It could even "move beyond", so that an "over-adjustment", $\Delta_A > \Delta_T$, has occurred. This not a detrimental feature; although $\Delta_R = \Delta_T - \Delta_A$ is then negative, it is ordinarily small (the analyst can only work with Δ_A; it is unknown to him/her whether Δ_A and Δ_T are close, or whether the over-adjustment $\Delta_A > \Delta_T$ has occurred). These points are illustrated by the simulation in Section 10. If $\Delta_T < 0$, these tendencies are reversed.

The form of (4.3) may suggest an argument which can however be misleading: Suppose that a vector $\mathbf{x}_k$ has been suggested, containing variables thought to be effective, along with an assumption that $y_k = \boldsymbol{\beta}'\mathbf{x}_k + \varepsilon_k$, where ε_k is a small residual. Then $\bar{y}_{r;d} - \bar{y}_{s;d} \approx (\bar{\mathbf{x}}_{r;d} - \bar{\mathbf{x}}_{s;d})'\mathbf{B}_{\mathbf{x}} \approx$ $(\bar{\mathbf{x}}_{r;d} - \bar{\mathbf{x}}_{s;d})'\boldsymbol{\beta}$, and consequently bias ratio ≈ 0, sending a message, often false, that the postulated vector $\mathbf{x}_k$ is efficient. One weakness of the argument stems from the well-known fact that nonresponse (unless completely random) will cause $\mathbf{B}_{\mathbf{x}}$ to be biased for a regression vector that describes the y-to-**x** relationship in the population. Further comments on this issue are given in Section 8.

Finally, there is the practical consideration that a typical survey has many y-variables. To every y-variable corresponds a calibration estimator, and a bias ratio given by (4.3). The ideal **x**-vector is one that would be capable of controlling bias in all those estimators. This is usually not possible without compromise, as we discuss later.

5. Expressing the deviation accounted for

The responding unit k receives the weight $d_k m_k$ in the estimator $\tilde{Y}_{\text{CAL}} = \sum_r d_k m_k y_k$. The nonresponse adjustment factor $m_k = (\sum_s d_k \mathbf{x}_k)'(\sum_r d_k \mathbf{x}_k \mathbf{x}_k')^{-1}\mathbf{x}_k$ expands the design weight d_k. We can view m_k as the value of a derived variable, defined for a particular outcome (r, s) and choice of $\mathbf{x}_k$, independent of all y-variables of interest, and computable for $k \in s$ (but used in $\tilde{Y}_{\text{CAL}}$ only for $k \in r$). Using (2.3), we have

$$\sum\nolimits_r d_k m_k \mathbf{x}_k = \sum\nolimits_s d_k \mathbf{x}_k; \sum\nolimits_r d_k m_k = \sum\nolimits_s d_k;$$

$$\sum\nolimits_r d_k m_k^2 = \sum\nolimits_s d_k m_k. \tag{5.1}$$

Two weighted means are needed:

$$\bar{m}_{r;d} = \frac{\sum_r d_k m_k}{\sum_r d_k} = \frac{\sum_s d_k}{\sum_r d_k} = \frac{1}{P}; \bar{m}_{s;d} = \frac{\sum_s d_k m_k}{\sum_s d_k} \tag{5.2}$$

where P is the response rate (2.1). Thus the average adjustment factor in $\tilde{Y}_{\text{CAL}} = \sum_r d_k m_k y_k$ is $1/P$, regardless of the choice of **x**-vector. Whether a chosen **x**-vector is efficient or not for reducing bias will depend on higher moments of the m_k. The weighted variance of the m_k is

$$S_m^2 = S_{m|r;d}^2 = \sum\nolimits_r d_k (m_k - \bar{m}_{r;d})^2 \big/ \sum\nolimits_r d_k. \tag{5.3}$$

The simpler notation S_m^2 will be used. A development of (5.3) and a use of (5.1) and (5.2) gives

$$S_m^2 = \bar{m}_{r;d}(\bar{m}_{s;d} - \bar{m}_{r;d}). \tag{5.4}$$

The coefficient of variation of the m_k is

$$\text{cv}_m = \frac{S_m}{\bar{m}_{r;d}} = \sqrt{\frac{\bar{m}_{s;d}}{\bar{m}_{r;d}} - 1}. \tag{5.5}$$

The weighted variance of the study variable y is given by

$$S_y^2 = S_{y|r;d}^2 = \sum\nolimits_r d_k (y_k - \bar{y}_{r;d})^2 \big/ \sum\nolimits_r d_k \tag{5.6}$$

(when the response probabilities are not all equal, $S_y^2 = S_{y|r;d}^2$ is not unbiased for the population variance $S_{y|U}^2$, but this is not an issue for the derivations that follow). We need the covariance

$$\text{Cov}(y, m) = \text{Cov}(y, m)_{r;d} =$$

$$\frac{1}{\sum_r d_k} \sum\nolimits_r d_k (m_k - \bar{m}_{r;d})(y_k - \bar{y}_{r;d}) \tag{5.7}$$

and the correlation coefficient, $R_{y,m} = \text{Cov}(y,m)/(S_y S_m)$, satisfying $-1 \le R_{y,m} \le 1$.

The deviation $\Delta_A = (\bar{\mathbf{x}}_{r;d} - \bar{\mathbf{x}}_{s;d})'\mathbf{B}_{\mathbf{x}}$ is a crucial component in the bias ratio (4.3). We seek an **x**-vector that makes Δ_A large. The factors that determine Δ_A are seen in (5.8) to (5.10). Computational tools (indicators) to assist the search for effective x-variables are given in (5.11) and (5.12). Their derivation by linear algebra is deferred to Section 7, which may be bypassed by readers more interested in the practical use of these tools in the search for x-variables, as illustrated in the empirical Sections 9 and 10. We can factorize Δ_A/S_y as

$$\Delta_A/S_y = -R_{y,m} \times \text{cv}_m. \qquad (5.8)$$

Two simple multiplicative factors determine Δ_A/S_y: The coefficient of variation cv_m, which is free of y_k and computed on the known $\mathbf{x}_k$ alone, and the (positive or negative) correlation coefficient $R_{y,m}$. Another factorization in terms of simple concepts is

$$\Delta_A / S_y = F \times R_{y,\mathbf{x}} \times \text{cv}_m \qquad (5.9)$$

where $R_{y,\mathbf{x}} = \sqrt{R^2_{y,\mathbf{x}}}$ is the coefficient of multiple correlation between y and $\mathbf{x}$, $R^2_{y,\mathbf{x}}$ is the proportion of the y-variance S^2_y explained by the predictor $\mathbf{x}$, and $F = -R_{y,m}/R_{y,\mathbf{x}}$ (formula (7.8) states the precise expression for $R^2_{y,\mathbf{x}}$). As Section 7 also shows, $|R_{y,m}| \le R_{y,\mathbf{x}}$ for any $\mathbf{x}$-vector and y-variable; consequently $-1 \le F \le 1$.

In (5.8) and (5.9), cv_m and $R_{y,\mathbf{x}}$ are non-negative terms, while $R_{y,m}$ and F can have either sign (or possibly be zero). Hence

$$|\Delta_A|/S_y = |R_{y,m}| \times \text{cv}_m = |F| \times R_{y,\mathbf{x}} \times \text{cv}_m. \qquad (5.10)$$

All of S_y, cv_m, $R_{y,\mathbf{x}}$, $R_{y,m}$ and F are easily computed in the survey. Both cv_m and $R_{y,\mathbf{x}}$ increase (or possibly stay unchanged) when further x-variables are added to the $\mathbf{x}$-vector; $R_{y,m}$ does not have this property.

To illustrate with the aid of fairly typical numbers, if $F = 0.5$; $R_{y,\mathbf{x}} = 0.6$ and $\text{cv}_m = 0.4$, then $\Delta_A/S_y = 0.12$, implying that $\tilde{Y}_{\text{CAL}}/N = \tilde{Y}_{\text{EXP}}/N - 0.12 \times S_y$. That is, the estimated y-mean $\tilde{Y}_{\text{CAL}}/\hat{N}$ has become adjusted by 0.12 standard deviations down from the primitive estimate $\tilde{Y}_{\text{EXP}}/\hat{N}$. The adjustment can be large compared to the standard deviation of the estimated y-mean, especially when the survey sample size is in the thousands. It remains unknown whether or not that adjustment has cured most of the biasing effect of nonresponse.

It follows from (5.8) that $0 \le |\Delta_A|/S_y \le \text{cv}_m$ whatever the y-variable. A shaper inequality is $|\Delta_A|/S_y \le R_{y,\mathbf{x}} \times \text{cv}_m$, but it depends on the y-variable. Further, if the correlation ratio F stays roughly constant when the $\mathbf{x}$-vector changes, so that $F \approx F_0$, then $|\Delta_A|/S_y \approx |F_0| \times R_{y,\mathbf{x}} \times \text{cv}_m$.

Although computable for any $\mathbf{x}$-vector and any outcome (s, r), Δ_A does not reveal the value of the bias ratio. But Δ_A suggests computational tools, called indicators, for comparing alternative $\mathbf{x}$-vectors. By (5.8), let

$$H_0 = \Delta_A / S_y = -R_{y,m} \times \text{cv}_m. \qquad (5.11)$$

As borne out by theory in Section 8 and by the empirical work in Section 10, over a long run of outcomes (s, r), the average of H_0 tracks the average deviation $\tilde{Y}_{\text{CAL}} - Y$ (which measures the bias of $\tilde{Y}_{\text{CAL}}$) in a nearly perfect linear manner when the $\mathbf{x}$-vector changes. This holds independently of the response distribution that generates r from s. Since H_0 can have either sign, it is practical to work with its absolute value denoted H_1; in addition we consider two other indicators, H_2 and H_3, inspired by (5.9) to (5.10):

$$H_1 = |\Delta_A| / S_y = |R_{y,m}| \times \text{cv}_m;$$

$$H_2 = R_{y,\mathbf{x}} \times \text{cv}_m; \; H_3 = \text{cv}_m. \qquad (5.12)$$

Our main alternatives are H_1 and H_3. Of these, H_1 is motivated by its direct link to Δ_A, which we want to make large, for a given y-variable. A strong reason to consider H_3 is its independence of all y-variables in the survey. The indicator H_2 is an *ad hoc* alternative; although H_2 contains a familiar concept, the multiple correlation coefficient $R_{y,\mathbf{x}}$, it is less appropriate than H_1 because the correlation coefficient ratio $F = -R_{y,m}/R_{y,\mathbf{x}}$ may vary considerably from one $\mathbf{x}$-vector to another. Both H_2 and H_3 increase when further x-variables are added to the $\mathbf{x}$-vector, something which does not hold in general for H_1. The use of these indicators is illustrated in the empirical Sections 9 and 10.

6. Preference ranking of auxiliary vectors

The methods in this paper are intended for use primarily with the large samples that characterize government surveys. The sample size is ordinarily much larger than the dimension of the $\mathbf{x}$-vector. The variance of estimates is ordinarily small, compared to the squared bias. However, for categorical auxiliary variables, no group size should be allowed to be "too small". It is recommended that all group sizes be at least 30, if not at least 50, in order to avoid instability. The crossing of categorical variables (to allow interactions) implies a certain risk of small groups. It is preferable to calibrate on marginal counts, rather than on frequencies for small crossed cells.

In a number of countries, the many available administrative registers provide a rich source of auxiliary information, particularly for surveys on individuals and households. These registers contain many potential x-variables from which to choose. Many different $\mathbf{x}$-vectors can be composed. The indicators in (5.12) provide computational tools for obtaining a preference ordering, or a ranking, of potential $\mathbf{x}$-vectors, with the objective to reduce as much as possible the bias remaining in the calibration estimator.

Scenario 1: Focus on a specific y-variable. The bias remaining in the calibration estimator depends on the y-variable; some are more bias prone than others. We identify one specific y-variable deemed to be highly important in the survey, and we seek to identify an **x**-vector that reduces the bias for this variable as much as possible (if more than one y-variable needs to be taken into account, a compromise must be struck, which suggests Scenario 2 below). For this purpose, we use the y-variable dependent indicator $H_1 = |\Delta_A|/S_y = |R_{y,m}| \times \mathrm{cv}_m$ and choose the **x**-vector so as to make H_1 large. An *ad hoc* alternative is to use the indicator $H_2 = R_{y,\mathbf{x}} \times \mathrm{cv}_m$, and strive to make it as large as possible.

Scenario 2: The objective is to identify a general purpose **x**-vector, efficient for all or most y-variables in the survey. This suggests $H_3 = \mathrm{cv}_m$ as a compromise indicator, and to choose the **x**-vector that maximizes H_3. To that same effect, Särndal and Lundström (2005, 2008) used the indicator $S_m^2 = H_3^2 / P^2$. They showed that the derived variable m_k in (2.6) can be seen as a predictor of the inverse of the unknown response probability and that choosing the **x**-vector to make S_m^2 large signals a bias reduction in the calibration estimator, irrespective of the y-variable.

For each scenario we can distinguish two procedures:

All vectors procedure: A list of candidate **x**-vectors is prepared, based on appropriate judgment. We compute the chosen indicator for *every* candidate **x**-vector, and settle for the vector that gives the highest indicator value. The resulting **x**-vector may not be the same for H_1 (which targets a specific y-variable) as for H_3 (which seeks a compromise for all y-variables in the survey).

Stepwise procedure: There is a pool of available x-variables. We build the **x**-vector by a stepwise forward (or stepwise backward) selection from among the available x-variables, one variable at a time, using the successive changes (if considered large enough) in the value of the chosen indicator to signal the inclusion (or exclusion) of a given x-variable at a given step. The indicators H_1, H_2 and H_3 do not in general give the same selection of variables. Consider two **x**-vectors, $\mathbf{x}_{1k}$ and $\mathbf{x}_{2k}$, such that $\mathbf{x}_{2k}$ is made up of $\mathbf{x}_{1k}$ and an additional vector $\mathbf{x}_{+k}$: $\mathbf{x}_{2k} = (\mathbf{x}'_{1k}, \mathbf{x}'_{+k})'$. The transition from $\mathbf{x}_{1k}$ to $\mathbf{x}_{2k}$ will increase the value of H_2 and H_3. In each step of a forward selection procedure we select the variable bringing the largest increase in H_2 or H_3. But the transition does not guarantee an increased value for the most appropriate indicator, H_1. However, H_1 may be used in stepwise selection in the manner described in Section 9.

7. Derivations

For given y-variable and outcome (s, r), we seek an **x**-vector to make the computable numerator $\Delta_A = (\bar{\mathbf{x}}_{r;d} - \bar{\mathbf{x}}_{s;d})'\mathbf{B}_{\mathbf{x}}$ in the bias ratio (4.3) large, in absolute value. In this section we prove the factorizations $\Delta_A/S_y = -R_{y,m} \times \mathrm{cv}_m = F \times R_{y,\mathbf{x}} \times \mathrm{cv}_m$ in (5.8) and (5.9). We note first that cv_m^2 is a quadratic form in the vector that contrasts the **x**-mean in the response set r with the **x**-mean in the sample s. Let

$$\mathbf{D} = \bar{\mathbf{x}}_{r;d} - \bar{\mathbf{x}}_{s;d};\ \mathbf{\Sigma} = \sum\nolimits_r d_k \mathbf{x}_k \mathbf{x}'_k \Big/ \sum\nolimits_r d_k. \quad (7.1)$$

Then, with P given by (2.1),

$$\mathrm{cv}_m^2 = P^2 \times S_m^2 = \mathbf{D}'\mathbf{\Sigma}^{-1}\mathbf{D}. \quad (7.2)$$

This expression follows from (5.3) and a consequence of (2.3), namely,

$$\bar{\mathbf{x}}'_{r;d}\mathbf{\Sigma}^{-1}\bar{\mathbf{x}}_{r;d} = \bar{\mathbf{x}}'_{r;d}\mathbf{\Sigma}^{-1}\bar{\mathbf{x}}_{s;d} = 1. \quad (7.3)$$

The vector of covariances with the study variable y is

$$\mathbf{C} = \left(\sum\nolimits_r d_k (\mathbf{x}_k - \bar{\mathbf{x}}_{r;d})(y_k - \bar{y}_{r;d})\right) \Big/ \left(\sum\nolimits_r d_k\right). \quad (7.4)$$

We can then write Δ_A as a bilinear form:

$$\Delta_A = \mathbf{D}'\mathbf{B}_{\mathbf{x}} = \mathbf{D}'\mathbf{\Sigma}^{-1}\mathbf{C} \quad (7.5)$$

using that $\mathbf{D}'\mathbf{\Sigma}^{-1}\bar{\mathbf{x}}_{r;d} = (\bar{\mathbf{x}}_{r;d} - \bar{\mathbf{x}}_{s;d})'\mathbf{\Sigma}^{-1}\bar{\mathbf{x}}_{r;d} = 0$ by (7.3).

A useful perspective on Δ_A is gained from the geometric interpretation of **C** and **D** in (7.5) as vectors in the space whose dimension is that of $\mathbf{x}_k$. We have

$$\Delta_A = \Lambda\,(\mathbf{D}'\mathbf{\Sigma}^{-1}\mathbf{D})^{1/2}(\mathbf{C}'\mathbf{\Sigma}^{-1}\mathbf{C})^{1/2} \quad (7.6)$$

where

$$\Lambda = \frac{\mathbf{D}'\mathbf{\Sigma}^{-1}\mathbf{C}}{(\mathbf{D}'\mathbf{\Sigma}^{-1}\mathbf{D})^{1/2}(\mathbf{C}'\mathbf{\Sigma}^{-1}\mathbf{C})^{1/2}}. \quad (7.7)$$

For a specific y-variable and a specific **x**-vector, the scalar quantities $(\mathbf{D}'\mathbf{\Sigma}^{-1}\mathbf{D})^{1/2}$ and $(\mathbf{C}'\mathbf{\Sigma}^{-1}\mathbf{C})^{1/2}$ represent the respective vector lengths of **D** and **C** (following an orthogonal transformation based on the eigenvectors and eigenvalues of $\mathbf{\Sigma}^{-1}$). The scalar quantity Λ represents the cosine of the angle between **D** (which is independent of y) and **C** (which depends on y); hence $-1 \le \Lambda \le 1$.

When the auxiliary vector $\mathbf{x}_k$ is allowed to expand by adding further available x-variables, both vector lengths $(\mathbf{D}'\mathbf{\Sigma}^{-1}\mathbf{D})^{1/2}$ and $(\mathbf{C}'\mathbf{\Sigma}^{-1}\mathbf{C})^{1/2}$ increase. The change in the angle Λ may be in either direction; if $|\Lambda|$ stays roughly constant, (7.6) shows that $|\Delta_A|$ will increase.

A second useful perspective on Δ_A follows by decomposing the total variability of the study variable y, $\sum_r d_k (y_k - \bar{y}_{r;d})^2 = (\sum_r d_k)\, S_y^2$. Two regression fits need to be examined, the one of y on the auxiliary vector $\mathbf{x}$, and the one of y on the derived variable m defined by (2.6). To each fit corresponds a decomposition of S_y^2 into explained y-variation and residual y-variation. The two explained portions have important links to the bias ratio (4.3). Result 7.1 summarizes the two decompositions.

Result 7.1. For a given survey outcome (s, r), let $\mathbf{D}, \mathbf{\Sigma}$ and $\mathbf{C}$ be given by (7.1) and (7.4). Then the proportion of the y-variance S_y^2 explained by the regression of y on $\mathbf{x}$ is

$$R_{y,\mathbf{x}}^2 = (\mathbf{C}'\mathbf{\Sigma}^{-1}\mathbf{C}) / S_y^2. \tag{7.8}$$

The coefficient of correlation between y and the univariate predictor m is

$$R_{y,m} = -(\mathbf{D}'\mathbf{\Sigma}^{-1}\mathbf{C}) / [(\mathbf{D}'\mathbf{\Sigma}^{-1}\mathbf{D})^{1/2} \times S_y]. \tag{7.9}$$

Consequently, the proportion of S_y^2 explained by m is

$$R_{y,m}^2 = (\mathbf{D}'\mathbf{\Sigma}^{-1}\mathbf{C})^2 / [(\mathbf{D}'\mathbf{\Sigma}^{-1}\mathbf{D}) \times S_y^2]. \tag{7.10}$$

The proportions $R_{y,\mathbf{x}}^2$ and $R_{y,m}^2$ satisfy $R_{y,m}^2 \leq R_{y,\mathbf{x}}^2 \leq 1$.

Proof. The proof of (7.8) uses the weighted least squares regression of y on $\mathbf{x}$ fitted over r. The residuals are $y_k - \hat{y}(\mathbf{x})_k$, where $\hat{y}(\mathbf{x})_k = \mathbf{x}_k'\mathbf{B}_\mathbf{x}$ with $\mathbf{B}_\mathbf{x}$ given by (2.8). The decomposition is

$$\sum\nolimits_r d_k (y_k - \bar{y}_{r;d})^2 = \sum\nolimits_r d_k (\hat{y}(\mathbf{x})_k - \bar{y}_{r;d})^2$$
$$+ \sum\nolimits_r d_k (y_k - \hat{y}(\mathbf{x})_k)^2.$$

The mixed term is zero. A development of the term "variation explained" gives $\sum_r d_k (\hat{y}(\mathbf{x})_k - \bar{y}_{r;d})^2 = (\sum_r d_k)$ $\mathbf{C}'\mathbf{\Sigma}^{-1}\mathbf{C}$. Thus the proportion of variance explained is $R_{y,\mathbf{x}}^2 = \sum_r d_k (\hat{y}(\mathbf{x})_k - \bar{y}_{r;d})^2 / [(\sum_r d_k)\, S_y^2] = \mathbf{C}'\mathbf{\Sigma}^{-1}\mathbf{C} / S_y^2$, as claimed in (7.8). To show (7.9) we note that the covariance (5.7) can be written with the aid of (7.5) as

$$\mathrm{Cov}(y, m) = -\Delta_A / P = -\mathbf{D}'\mathbf{\Sigma}^{-1}\mathbf{C} / P.$$

It then follows from (7.2) that $R_{y,m} = \mathrm{Cov}(y, m) / (S_y S_m)$ has the expression (7.9). The residuals from the regression (with intercept) of y on the univariate explanatory variable m are $\hat{y}(m)_k = \bar{y}_{r;d} + B_m(m_k - \bar{m}_{r;d})$ with $B_m = \mathrm{Cov}(y, m)/S_m^2 = -P\,(\mathbf{D}'\mathbf{\Sigma}^{-1}\mathbf{C})/(\mathbf{D}'\mathbf{\Sigma}^{-1}\mathbf{D})$. The proportion of variance explained is $\sum_r d_k (\hat{y}(m)_k - \bar{y}_{r;d})^2 / [(\sum_r d_k)\, S_y^2]$, which upon development gives the expression for $R_{y,m}^2$ in (7.10). Finally, $R_{y,m}^2 \leq R_{y,\mathbf{x}}^2$ follows from the Cauchy-Schwarz inequality for a bilinear form: $(\mathbf{D}'\mathbf{\Sigma}^{-1}\mathbf{C})^2 \leq (\mathbf{D}'\mathbf{\Sigma}^{-1}\mathbf{D})$ $(\mathbf{C}'\mathbf{\Sigma}^{-1}\mathbf{C})$.

The inequality $R_{y,m}^2 \leq R_{y,\mathbf{x}}^2 \leq 1$ can also be deduced by the fact that, among all predictions $\hat{y}_k = \mathbf{x}_k'\boldsymbol{\beta}$ that are linear in the $\mathbf{x}$-vector, those that maximize the variance explained are $\hat{y}(\mathbf{x})_k = \mathbf{x}_k'\mathbf{B}_\mathbf{x}$, so the predictions $\hat{y}(m)_k$, which are linear in $\mathbf{x}_k$ via m_k, cannot yield a greater variance explained than that maximum.

Now from (7.9), (7.2) and (7.5), $-R_{y,m}\mathrm{cv}_m = \mathbf{D}'\mathbf{\Sigma}^{-1}\mathbf{C}/S_y = \Delta_A/S_y$, as claimed by formula (5.8). Moreover, (7.7), (7.8) and (7.9) imply $-R_{y,m}/R_{y,\mathbf{x}} = \Lambda$, so the correlation coefficient ratio F in (5.9) equals the angle Λ defined by (7.7).

8. Comments: Goodness of fit, properties of the bias and a related selection procedure

Three issues are examined in this section: (i) The relationship between bias and goodness of fit, (ii) the linear relation between the expected value of $\Delta_A = \hat{N}^{-1}(\tilde{Y}_{\mathrm{EXP}} - \tilde{Y}_{\mathrm{CAL}})$ and the bias of $\tilde{Y}_{\mathrm{CAL}}$ or $\hat{Y}_{\mathrm{CAL}}$, and (iii) the alternative method for selection of auxiliary variables proposed by Schouten (2007).

For the issue (i), recall that the total deviation in Section 4 is $\Delta_T = \Delta_A + \Delta_R$, where Δ_A is computable but Δ_T and Δ_R are not. If computable, $\hat{N}\Delta_R = \tilde{Y}_{\mathrm{CAL}} - \tilde{Y}_{\mathrm{FUL}}$ would be an estimate of the bias of $\tilde{Y}_{\mathrm{CAL}}$ (and of that of $\hat{Y}_{\mathrm{CAL}}$). A small Δ_R is desirable. The question arises: Is this achieved when $y_k = \boldsymbol{\beta}'\mathbf{x}_k + \varepsilon_k$ (with a given vector $\mathbf{x}_k$) fits the data well? We need to distinguish two aspects: (a) The computable fit to the data $(y_k, \mathbf{x}_k)$ observed for $k \in r$; and (b) The hypothetical fit to the data $(y_k, \mathbf{x}_k)$ for $k \in s$, some observed, some not.

A good fit for the respondents, $k \in r$, does not guarantee a small Δ_R: The weighted LSQ fit using the observed data $(y_k, \mathbf{x}_k)$ for $k \in r$ gives the residuals $e_{k|r;d} = y_k - \mathbf{x}_k'\mathbf{B}_{\mathbf{x}|r;d}$, computable for $k \in r$, with the property $\sum_r d_k e_{k|r;d} = 0$ (here, the detailed notation $\mathbf{B}_{\mathbf{x}|r;d}$ specified in (2.8) is preferable to the simplified notation $\mathbf{B}_\mathbf{x}$). For $k \in s - r$, $e_{k|r;d}$ is not computable; it has an unknown non-zero mean $\bar{e}_{s-r;d} = \sum_{s-r} d_k e_{k|r;d} / \sum_{s-r} d_k$. We have

$$\Delta_R = (\tilde{Y}_{\mathrm{CAL}} - \tilde{Y}_{\mathrm{FUL}}) / \hat{N} = -(1 - P)\,\bar{e}_{s-r;d} \neq 0. \tag{8.1}$$

Regardless of whether the fit is good (small residuals $e_{k|r;d}$; $R_{y,\mathbf{x}}^2$ near one) or poor (large residuals $e_{k|r;d}$; $R_{y,\mathbf{x}}^2$ near zero), the deviation Δ_R given by (8.1) may be large, and $\tilde{Y}_{\mathrm{CAL}}$ far from unbiased. Even with a perfect fit for the respondents ($e_{k|r,d} = 0$ for all $k \in r$, and $R_{y,\mathbf{x}}^2 = 1$), there is no guarantee that the bias is small.

A similar inadequacy affects imputation based on the respondent data. If the regression imputations $\hat{y}_k = \mathbf{x}_k'\mathbf{B}_{\mathbf{x}|r;d}$ are used to fill in for the values y_k missing for $k \in s - r$, the imputed estimator is

$$\hat{Y}_{\mathrm{imp}} = \sum\nolimits_r d_k y_k + \sum\nolimits_{s-r} d_k \hat{y}_k.$$

Then $\hat{Y}_{\mathrm{imp}} = \tilde{Y}_{\mathrm{CAL}}$, so $\hat{Y}_{\mathrm{imp}}$ has the same exposure to bias as $\tilde{Y}_{\mathrm{CAL}}$, as is easily understood: When the nonresponse causes a skewed selection of y-values, the imputed values computed on that skewed selection will misrepresent the unknown y-values that characterize the sample s or the population U.

Consider now the aspect (b) of the fit, that is, the hypothetical weighted LSQ regression fit to the data $(y_k, \mathbf{x}_k)$ for $k \in s$. The regression coefficient vector would be $\mathbf{B}_{\mathbf{x}|s;d} = (\sum_s d_k \mathbf{x}_k \mathbf{x}_k')^{-1} \sum_s d_k \mathbf{x}_k y_k$, and the residuals $e_{k|s;d} = y_k - \mathbf{x}_k' \mathbf{B}_{\mathbf{x}|s;d}$ for $k \in s$ satisfy $\sum_s d_k e_{k|s;d} = 0$. Using that $\sum_r d_k m_k \mathbf{x}_k / \hat{N} = \bar{\mathbf{x}}_{s;d}$ and $\sum_r d_k m_k y_k / \hat{N} = \bar{\mathbf{x}}_{s;d}' \mathbf{B}_{\mathbf{x}|r;d}$, we have

$$\Delta_R = \hat{N}^{-1}(\tilde{Y}_{\text{CAL}} - \tilde{Y}_{\text{FUL}}) = (1/\hat{N}) \sum\nolimits_r d_k m_k e_{k|s;d}. \quad (8.2)$$

Suppose the model is "true for the sample s", with a perfect fit, so that $e_{k|s;d} = 0$ for all $k \in s$. Then, by (8.2) we do have $\Delta_R = 0$, so the nonresponse adjusted estimator $\tilde{Y}_{\text{CAL}}$ agrees with the unbiased estimator $\tilde{Y}_{\text{FUL}}$. A belief that the bias is small hinges on an unverifiable assumption.

Turning to the issue (ii), we now explain the essentially linear relation between the bias of $\tilde{Y}_{\text{CAL}}$ and the expected value of the indicator $H_0 = \Delta_A / S_y = (\tilde{Y}_{\text{EXP}} - \tilde{Y}_{\text{CAL}}) / \hat{N} S_y$. For a given outcome (s, r), a fixed y-variable and a fixed $\mathbf{x}$-vector we have

$$(\tilde{Y}_{\text{CAL}} - Y) / \hat{N} S_y = (\tilde{Y}_{\text{EXP}} - Y) / \hat{N} S_y - H_0.$$

Let E_{pq} denote the expectation operator with respect to all outcomes (s, r), that is, $E_{pq}(\cdot) = E_p(E_q(\cdot|s))$, where $p(s)$ and $q(r|s)$ are, respectively, the known sampling design and the unknown response distribution. We denote $\text{bias}(\tilde{Y}_{\text{CAL}}) = E_{pq}(\tilde{Y}_{\text{CAL}}) - Y$, $\text{bias}(\tilde{Y}_{\text{EXP}}) = E_{pq}(\tilde{Y}_{\text{EXP}}) - Y$ and $C = E_{pq}(\hat{N} S_y)$. Using the usual large sample replacement of the expected value of a ratio by the ratio of the expected values, we have $E_{pq}[(\tilde{Y}_{\text{CAL}} - Y) / \hat{N} S_y] \approx [E_{pq}(\tilde{Y}_{\text{CAL}}) - Y] / E_{pq}(\hat{N} S_y)$ and analogously for $\tilde{Y}_{\text{EXP}}$, so

$$\text{bias}(\tilde{Y}_{\text{CAL}}) \approx \text{bias}(\tilde{Y}_{\text{EXP}}) - C \times E(H_0). \quad (8.3)$$

Here $\text{bias}(\tilde{Y}_{\text{EXP}})$ and C do not depend on the choice of $\mathbf{x}$-vector, whereas $\text{bias}(\tilde{Y}_{\text{CAL}})$ and $E(H_0)$ do. Therefore, as the $\mathbf{x}$-vector changes, $\text{bias}(\tilde{Y}_{\text{CAL}})$ and $E(H_0)$ are essentially linearly related. No particular forms of $p(s)$ and $q(r|s)$ need to be specified for (8.3) to hold. As a consequence, when two auxiliary vectors, $\mathbf{x}_{1k}$ and $\mathbf{x}_{2k}$, are compared, the difference in bias is, to close approximation, proportional to the change in the expected value of H_0:

$$\text{bias}(\tilde{Y}_{\text{CAL}}(\mathbf{x}_{1k})) - \text{bias}(\tilde{Y}_{\text{CAL}}(\mathbf{x}_{2k})) \approx -C(E_1 - E_2) \quad (8.4)$$

where $E_i = E_{pq}(H_0(\mathbf{x}_{ik}))$ for $i = 1, 2$. The properties (8.3) and (8.4) are validated by the Monte Carlo study in Section 10.

Note that formula (8.3) does not guarantee that $\tilde{Y}_{\text{CAL}}$ based on a certain vector $\mathbf{x}_k$ will have zero or near-zero bias. It does not state that a comparatively large value of $|\Delta_A|$ guarantees a small bias in $\tilde{Y}_{\text{CAL}}$. What (8.3) says is that $\text{bias}(\tilde{Y}_{\text{CAL}})$ is linearly related to the expectation of the indicator $H_0 = \Delta_A / S_y$. Therefore, to assess available $\mathbf{x}$-vectors in terms of the indicator H_0 (or the indicator $H_1 = |\Delta_A| / S_y$) is consistent with the objective of bias reduction.

Turning to the issue (iii), we comment on the alternative method for selection of auxiliary variables proposed by Schouten (2007). His indicator for the step-by-step selection of variables differs from our indicators; it will usually not select exactly the same set of variables. In a list of say 30 available categorical x-variables, the first ten to enter will not be the same set of ten as with our indicators H_0 to H_3. The order in which variables are selected will not necessarily be the same either. For comparison, we compared, in some of our empirical work, with the variable selection realized by Schouten's method. In some cases we noted a considerable congruence between the two sets of "first ten" picked in the two procedures.

The differences between the two approaches are best appreciated by a comparison of their background and derivation. Our indicators H_0 and H_1 originate in the notion of separation (or distance), for a given outcome (s, r), between the adjusted estimator $\tilde{Y}_{\text{CAL}}$ and the primitive one, $\tilde{Y}_{\text{EXP}}$, and in the idea that this separation will ordinarily increase when the $\mathbf{x}$-vector becomes more powerful. The probability sampling design is taken into consideration; no assumptions are made on the response distribution.

Schouten uses a superpopulation argument; sampling weights do not appear to enter into consideration. An expression for the model-expected bias of an estimator of the population mean is found to be proportional to the correlation (at the level of the population) between the y-variable and the 0-1 indicator for response. It is shown that this correlation (and consequently the bias) can be bounded inside an interval. In particular, the generalized regression estimator is considered and it is shown that its maximum absolute bias equals the width of the bias interval. This width depends on the true unknown regression vector $\boldsymbol{\beta}$ for the regression (at the population level) between y and $\mathbf{x}$. This unknown $\boldsymbol{\beta}$ is replaced by an estimate based on the respondents, thus subject to some bias because of the nonresponse. Schouten emphasizes that a missing-at-random assumption is not needed for his method, which is in that respect similar to our method.

9. Auxiliary variable choice for the Swedish pilot survey on gaming and problem gambling

We identified a real survey data set to illustrate the use of the indicators H_1, H_2 and H_3 in building the **x**-vector. In 2008, The Swedish National Institute of Public Health (*Svenska Folkhälsoinstitutet*) conducted a pilot survey to study the extent of gambling participation and the characteristics of persons with gambling problems. Sampling and weight calibration was carried out by Statistics Sweden. We illustrate the use of the indicators in this survey, for which a stratified simple random sample s of n = 2,000 persons was drawn from the Swedish Register of Total Population (RTP). The strata were defined by the cross classification of region of residence by age group. Each of the six regions was defined as a cluster of postal code areas deemed similar in regard to variables such as education level, purchasing power, type of housing, foreign background. The four age groups were defined by the brackets 16-24; 25-34; 35-64 and 65-84.

The overall unweighted response rate was 50.8%. The nonresponse, more or less pronounced in the different domains of interest, interferes with the accuracy objective. An extensive pool of potential auxiliary variables was available for this survey, including variables in the RTP, in the Education Register and a subset of those in another extensive Statistics Sweden data base, LISA. For this illustration, we prepared a data file consisting of 13 selected categorical variables. Twelve of these were designated as x-variables, and one, the dichotomous variable *Employed*, played the role of the study variable. The values of all variables are available for all units $k \in s.$ Response $(k \in r)$ or not $(k \in s - r)$ to the survey is also indicated in the data file.

Variables that are continuous by nature were used as grouped; all 12 x-variables are thus categorical and of the $\mathbf{x}_k^{\circ}$ type, as defined in Section 2 (because most of the variables are available for the full population, they are potentially of the type $\mathbf{x}_k^{*}$, but since the effect on bias is of little consequence, we used them as $\mathbf{x}_k^{\circ}$-variables). The study variable value, $y_k = 1$ if k is *employed* and $y_k = 0$ otherwise, is known for $k \in s,$ so the unbiased estimate $\tilde{Y}_{\text{FUL}}$ defined by (3.2) can be computed and used as a reference. We also computed $\tilde{Y}_{\text{EXP}}$ defined by (3.1), as well as $\tilde{Y}_{\text{CAL}}$ defined by (2.5) for different **x**-vectors built by stepwise selection from the pool of 12 x-variables with the aid of the indicators H_1, H_2 and H_3 defined by (5.12).

We carried out forward selection as follows: The auxiliary vector in Step 0 is the trivial $\mathbf{x}_k = 1,$ and the estimator is $\tilde{Y}_{\text{EXP}}$. In Step 1, the indicator value is computed for every one of 12 presumptive auxiliary variables; the variable producing the largest value of the indicator is selected. In Step 2, the indicator value is computed for all 11 vectors of dimension two that contain the variable selected in Step 1 and one of the remaining variables. The variable that gives the largest value for the indicator is selected in Step 2, and so on, in the following steps. A new variable always joins already entered variables in the "side-by-side" (or "+") manner. Interactions are thereby relinquished. The order of selection is different for each indicator.

The values of H_2 and H_3 that identify the next variable for inclusion are by mathematical necessity increasing in every step. This does not hold for H_1. In a certain step j, we used the rule to include the x-variable with the largest of computed H_1-values. That value can be smaller than the H_1-value that identified the variable entering in the preceding step, $j-1$. The series of H_1-values for inclusion will increase up to a certain step, then begin to decline, as Table 9.1 illustrates.

The unbiased estimate is $\tilde{Y}_{\text{FUL}} = 4,265;$ the primitive estimate is $\tilde{Y}_{\text{EXP}} = 4,719$ (both in thousands). This suggests a large positive bias in $\tilde{Y}_{\text{EXP}},$ whose relative deviation (in %) from $\tilde{Y}_{\text{FUL}}$ is $\text{RDF} = (\tilde{Y}_{\text{EXP}} - \tilde{Y}_{\text{FUL}}) / \tilde{Y}_{\text{FUL}} \times 10^2 = 10.7.$ Adding categorical x-variables one by one into the **x**-vector will successively change this deviation, although when a few variables have been admitted, the change is not always in the direction of a smaller value. In each step we computed the indicator, $\tilde{Y}_{\text{CAL}}$ and $\text{RDF} = (\tilde{Y}_{\text{CAL}} - \tilde{Y}_{\text{FUL}}) / \tilde{Y}_{\text{FUL}} \times 10^2.$

Table 9.1 shows the stepwise selection with the indicator H_1 (the number of categories is given in parenthesis for each selected variable). First to enter is the variable *Income class*; this brings a large reduction in RDF from 10.7 to 4.5. The next five selections take place with increased H_1-values, and the value of RDF is reduced, but by successively smaller amounts. Step six, where *Marital status* is selected, brings about a turning point, indicated by the double line in Table 9.1: The value of H_1 then starts to decline, and $\tilde{Y}_{\text{CAL}}$ and RDF start to increase. At step 6, RDF is at its lowest value, 0.5, then starts to rise, illustrating that inclusion of all available x-variables may not be best. The turning point of H_1 and the point at which RDF is closest to zero happen to agree in this example. This is not generally the case. Moreover, in a real survey setting, RDF is unknown, as is the step at which RDF is closest to zero.

Table 9.2 shows the stepwise selection with indicator H_3. Its value increases at every step, but at a rate that levels off, and successive changes in $\tilde{Y}_{\text{CAL}}$ become negligible. This suggests to stop after six steps, at which point RDF = 2.8. In none of the 12 steps does RDF come as close to zero as the value RDF = 0.5 obtained with H_1 after six steps. In this respect H_1 is better than H_3, in this example. With all 12 x-variables selected, RDF attains in both tables the final value 2.6.

Table 9.1
Stepwise forward selection, indicator H_1, dichotomous study variable *Employed*. Successive values of $H_1 \times 10^3$, of $\tilde{Y}_{\text{CAL}}$ in thousands, and of RDF = $(\tilde{Y}_{\text{CAL}} - \tilde{Y}_{\text{FUL}})/\tilde{Y}_{\text{FUL}} \times 10^2$. For comparison, $\tilde{Y}_{\text{EXP}} \times 10^{-3}$= 4,719; $\tilde{Y}_{\text{FUL}} \times 10^{-3}$= 4,265

Auxiliary variable entered	$H_1 \times 10^3$	$\tilde{Y}_{\text{CAL}} \times 10^{-3}$	RDF
Income class (3)	76	4,458	4.5
Education level (3)	107	4,350	2.0
Presence of children (2)	114	4,326	1.4
Urban centre dwelling (2)	118	4,310	1.1
Sex (2)	123	4,296	0.7
Marital status (2)	125	4,286	0.5
Days unemployed (3)	121	4,301	0.9
Months with sickness benefits (3)	120	4,305	1.0
Level of debt (3)	115	4,322	1.3
Cluster of postal codes (6)	109	4,343	1.8
Country of birth (2)	103	4,363	2.3
Age class (4)	99	4,377	2.6

Table 9.2
Stepwise forward selection, indicator H_3, dichotomous study variable *Employed*. Successive values of $H_3 \times 10^3$, of $\tilde{Y}_{\text{CAL}}$ in thousands, of RDF = $(\tilde{Y}_{\text{CAL}} - \tilde{Y}_{\text{FUL}})/\tilde{Y}_{\text{FUL}} \times 10^2$. For comparison, $\tilde{Y}_{\text{EXP}} \times 10^{-3}$= 4,719; $\tilde{Y}_{\text{FUL}} \times 10^{-3}$= 4,265

Auxiliary variable entered	$H_3 \times 10^3$	$\tilde{Y}_{\text{CAL}} \times 10^3$	RDF
Education level (3)	186	4,520	6.0
Cluster of postcode areas (6)	250	4,505	5.6
Country of birth (2)	281	4,498	5.5
Income Class (3)	298	4,369	2.4
Age class (4)	354	4,399	3.1
Sex (2)	364	4,384	2.8
Urban centre dwelling (2)	374	4,378	2.6
Level of debt (3)	381	4,364	2.3
Months with sickness benefits (3)	384	4,380	2.7
Presence of children (2)	387	4,379	2.7
Marital status (2)	388	4,379	2.7
Days unemployed (3)	388	4,377	2.6

The set of the first six variables to enter with H_3 has three in common with the corresponding set of six with H_1. There is no contradiction in the quite different selection patterns, because H_1 is geared to the specific y-variable *Employed*, while H_3 is a compromise indicator, independent of any y-variable. To save space, the step-by-step results for indicator H_2 are not shown. Its selection pattern resembles more that of H_3 than that of H_1. Out of the first six variables to enter with H_2, four are among the first six with H_3. As a general comment, we believe that in many practical situations the use of more than six variables is unnecessary, and the selection of the first few becomes crucially important.

10. Empirical validation by simulation for a constructed population

The theory presented in earlier sections makes no assumptions on the response distribution. It is unknown. The sampling design is arbitrary; its known inclusion probabilities are taken into account. For the experiment in this section, we specify several different response distributions with a specified positive value for the response probability θ_k for every $k \in U$. That is, with specified probability θ_k, the value y_k gets recorded in the experiment; with probability $1-\theta_k$, it goes missing. We find that the indicators H_0 (or $H_1 = |H_0|$) defined in (5.11) ranks the different **x**-vectors in the correct order of preference for all participating response distributions, consistent with the theoretical results (8.3) and (8.4). We confirm that, over a long run of outcomes (s, r), the average of $H_0 = \Delta_A / S_y = -R_{y,m} \times \text{cv}_m$ tracks the bias of the calibration estimator, measured by the average of $\tilde{Y}_{\text{CAL}} - Y$, in an essentially perfectly linear manner, when the **x**-vector moves through 16 different formulations. We also examine the indicators H_2 and H_3 defined in (5.12), and find in this experiment that they also have strong relationship to the bias of $\tilde{Y}_{\text{CAL}}$.

We experimented with several created populations; the conclusions were similar. We report here results for one constructed population of size $N = 6{,}000$, with created values $(y_k, \mathbf{x}_k, \theta_k)$ for $k = 1, 2, \ldots, N = 6{,}000$, for 16 alternative categorical formulations of $\mathbf{x}_k$, and four different ways to assign the θ_k.

The 16 alternative categorical auxiliary **x**-vectors were obtained by grouping the generated values x_{1k} and x_{2k} of two continuous auxiliary variables, x_1 and x_2. The values (y_k, x_{1k}, x_{2k}) for $k = 1, 2, \ldots, 6{,}000$ were created in three steps as follows. Step 1 (the variable x_1): The 6,000 values x_{1k} were obtained as independent outcomes of the gamma distributed random variable $\Gamma(a, b)$ with parameter values a = 2, b = 5. The mean and variance of the 6,000 realized values x_{1k} was 10.0 and 49.9, respectively. Step 2 (the variable x_2): For unit k, with value x_{1k} fixed by Step 1, a value x_{2k} is realized as an outcome of the gamma random variable with parameters such that the conditional expectation and variance of x_{2k} are $\alpha + \beta x_{1k} + K\,h(x_{1k})$ and $\sigma^2 x_{1k}$, respectively, where $h(x_{1k}) = x_{1k}(x_{1k} - \mu_{x_1})(x_{1k} - 3\mu_{x_1})$ with $\mu_{x_1} = 10$. We used the values $\alpha = 1$, $\beta = 1, k = 0.001$ and $\sigma^2 = 25$. The polynomial term $K\,h(x_{1k})$ gives a mild non-linear shape to the plot of (x_{2k}, x_{1k}), to avoid an exactly linear relationship. The mean and variance of the 6,000 realized values x_{2k} were 11.0 and 210.0, respectively. The correlation coefficient between x_1 and x_2, computed on the 6,000 couples (x_{1k}, x_{2k}), was 0.48. Step 3 (the study variable y): For unit k, with values x_{1k} and x_{2k} fixed by Steps 1 and 2, a value y_k is realized as an outcome of the gamma random variable with parameters such that the conditional expectation and variance of y_k are $c_0 + c_1 x_{1k} + c_2 x_{2k}$ and $\sigma_0^2(c_1 x_{1k} + c_2 x_{2k})$, respectively. We used $c_0 = 1, c_1 = 0.7$, $c_2 = 0.3$ and $\sigma_0^2 = 2$. The mean and the variance of the 6,000 realized values y_k were 11.4 and 86.5, respectively. The correlation coefficient between y and x_1, computed on the 6,000 couples (y_k, x_{1k}), was 0.76; that between y and x_2, computed on the 6,000 couples (y_k, x_{2k}), was 0.73.

Each of the two x-variables was then transformed into four alternative group modes, denoted 8G, 4G, 2G and 1G, yielding $4 \times 4 = 16$ different auxiliary vectors $\mathbf{x}_k$. The 6,000 values x_{1k} of variable x_1 were size ordered; eight equal-sized groups were formed. Group 1 consists of the units with the 750 largest values x_{1k}, group 2 consists of the next 750 units in the size ordering, and so on, ending with group 8. In this mode 8G of x_1, unit k is assigned the vector value $\boldsymbol{\gamma}_{(x_1;8)k}$, of dimension eight with seven entries "0" and a single entry "1" to code the group membership of k. Next, successive group mergers are carried out, so that two adjoining groups always define a new group, every time doubling the group size. Thus for mode 4G, the merger of groups 1 and 2 puts the units with the 1,500 largest x_{1k}-values into a first new group; groups 3 and 4 merge to form the second new group of 1,500, and so on; the vector value associated with unit k is $\boldsymbol{\gamma}_{(x_1;4)k}$. In mode 2G, unit k has the vector value $\boldsymbol{\gamma}_{(x_1;2)k} = (1,0)'$ for the 3,000 largest x_1-value units and $\boldsymbol{\gamma}_{(x_1;2)k} = (0,1)'$ for the rest. In the ultimate mode, 1G, all 6,000 units are put together, all x_1-information is relinquished, and $\boldsymbol{\gamma}_{(x_1;1)k} = 1$ for all k. The 6,000 values x_{2k} were transformed by the same procedure into the group modes 8G, 4G, 2G and 1G. Corresponding group membership of unit k is coded by the vectors $\boldsymbol{\gamma}_{(x_2;8)k}, \boldsymbol{\gamma}_{(x_2;4)k}, \boldsymbol{\gamma}_{(x_2;2)k}$ and $\boldsymbol{\gamma}_{(x_2;1)k} = 1$. The $4 \times 4 = 16$ different auxiliary vectors $\mathbf{x}_k$ take into account both kinds of group information; the two $\boldsymbol{\gamma}$-vectors are placed side by side (as opposed to crossed), the result being a calibration on two margins, as indicated by the "+" sign. Thus for the case denoted 8G + 8G, unit k has the auxiliary vector value $\mathbf{x}_k = (\boldsymbol{\gamma}'_{(x_1;8)k}, \boldsymbol{\gamma}'_{(x_2;8)k})'_{(-1)}$, where (-1) indicates that one category is excluded in either $\boldsymbol{\gamma}_{(x_1;8)k}$ or $\boldsymbol{\gamma}_{(x_2;8)k}$ to avoid a singular matrix in the computation, giving $\mathbf{x}_k$ the dimension 8 + 8 – 1 = 15. The case 8G + 8G has the highest information content. At the other extreme, the case 1G + 1G disregards all the x-information and $\mathbf{x}_k = 1$ for all k. There are 14 intermediate cases of information content. For example, 4G + 2G has $\mathbf{x}_k = (\boldsymbol{\gamma}'_{(x_1;4)k}, \boldsymbol{\gamma}'_{(x_2;2)k})'_{(-1)}$ of dimension 4 + 2 – 1 = 5; 4G + 1G has $\mathbf{x}_k = (\boldsymbol{\gamma}'_{(x_1;4)k}, 1)'_{(-1)} = \boldsymbol{\gamma}_{(x_1;4)k}$ of dimension 4 (there is non-negligible interaction between x_1 and x_2 in this experiment, but we restrict the experiment to $\mathbf{x}$-vectors without interactions, causing no risk of small group counts).

We discuss here the results for four response distributions. Their response probabilities θ_k, $k = 1, 2, ..., N =$ 6,000, were specified as follows:

IncExp($10 + x_1 + x_2$), with $\theta_k = 1 - e^{-c(10 + x_{1k} + x_{2k})}$
where $c = 0.04599$

IncExp($10 + y$), with $\theta_k = 1 - e^{-c(10 + y_k)}$
where $c = 0.06217$

DecExp($x_1 + x_2$), with $\theta_k = e^{-c(x_{1k} + x_{2k})}$
where $c = 0.01937$

DecExp(y), with $\theta_k = e^{-c y_k}$
where $c = 0.03534$.

The constant c was adjusted in all four cases to give a mean response probability of $\overline{\theta}_U = \sum_U \theta_k / N = 0.70$. In the first two, the value 10 (rather than 0) was used to avoid a high incidence of small response probabilities θ_k. These four options represent contrasting features for the response probabilities: increasing as opposed to decreasing, dependent on x-values only as opposed to dependent on y-values only. In the second and fourth option, the response is directly y-variable dependent, and could hence be called "purely non-ignorable".

We generated $J = 5{,}000$ outcomes (s, r), where s of size $n = 1{,}000$ is drawn from $N = 6{,}000$ by simple random sampling and, for every given s, the response set r is realized by each of the four response distributions. That is, for $k \in s$, a Bernoulli trial was carried out with the specified probability θ_k of inclusion in the response set r. The Bernoulli trials are independent.

For each response distribution, for each of the 16 $\mathbf{x}$-vectors, and for every outcome (s, r), we computed the relative deviation $\text{RD} = (\hat{Y}_{\text{CAL}} - Y)/Y$, where $\hat{Y}_{\text{CAL}}$ is given by (2.4) and $Y = \sum_U y_k$ is the targeted y-total, known in this experimental setting (alternatively, we used $\tilde{Y}_{\text{CAL}}$ given by (2.5) but, as expected, the difference in bias compared with $\hat{Y}_{\text{CAL}}$ is negligible). We also computed the indicators $H_i, i = 0, 1, 2, 3,$ given by (5.11) and (5.12). Summary measures were computed as

$$\text{relbias} = \text{Av(RD)} = \frac{1}{J} \sum_{j=1}^{J} \text{RD}_j;$$

$$\text{Av}(H_i) = \frac{1}{J} \sum_{j=1}^{J} H_{ij} \quad \text{for} \quad i = 0, 1, 2, 3$$

where j indicates the value computed for the j^{th} outcome, $j = 1, 2, \ldots, 5{,}000 = J$. For each response distribution, we thus obtain the value *relbias* (which is the Monte Carlo measure of the relative bias $(E_{pq}(\hat{Y}_{\text{CAL}}) - Y)/Y$) and 16 values of $\text{Av}(H_i)$ (which is the Monte Carlo measure of $E_{pq}(H_i)$), $i = 0, 1, 2, 3$, where p stands for simple random sampling, and q stands for one of the four response distributions.

Table 10.1 shows, for IncExp($10 + x_1 + x_2$), *relbias* in % and $\text{Av}(H_1) \times 10^3$ for the 16 $\mathbf{x}$-vectors. For the cell 1G + 1G, with vector $\mathbf{x}_k = 1$, all four Av-quantities are zero, and *relbias* is at its highest level, 13.2%. At the opposite extreme, the cell 8G + 8G represents the highest level of information; it gives the highest value for $\text{Av}(H_1)$, and *relbias* is at its lowest value, 0.2%; virtually all bias is removed (except for a possible sign difference, $\text{Av}(H_0)$ and $\text{Av}(H_1)$ were equal for all cells).

The result (8.4), holding for any response distribution and any sampling design, states that the indicator H_0 will rank the $4 \times 4 = 16$ auxiliary vectors correctly for any response distribution (with response probabilities not all constant, as noted below). Table 10.1 illustrates (8.4) in terms of $H_1 = |H_0|$: The change, from any one cell to any other, in the value of $\text{Av}(H_1)$ (the Monte-Carlo estimate of the expected value of (H_1) is accompanied by a proportional change in the value of *relbias*. The same proportionality was noted for the other three response distributions. We could have chosen other response distributions to illustrate the same property.

Table 10.1
Relbias in % and, within parenthesis, the value of $\text{Av}(H_1) \times 10^3$ for 16 auxiliary vectors $\mathbf{x}_k$. Response distribution IncExp$(10 + x_1 + x_2)$

Groups based on x_{1k}	Groups based on x_{2k}							
	8G		4G		2G		1G	
8G	0.2	(101)	0.5	(99)	1.3	(93)	3.4	(76)
4G	0.5	(98)	0.9	(96)	1.8	(89)	4.1	(70)
2G	1.5	(91)	1.9	(88)	3.2	(78)	6.5	(52)
1G	4.1	(70)	5.0	(64)	7.3	(46)	13.2	(0)

The response distribution with a constant response probability θ_k for all k is a special case. The calibration estimator $\tilde{Y}_{\text{CAL}}$ based on any vector $\mathbf{x}_k$ then has zero bias (very nearly), and this includes the primitive estimator $\tilde{Y}_{\text{EXP}}$ with $\mathbf{x}_k = 1$. Result 8.3 continues to be valid, stating in that case that $E_{pq}(H_0) \approx \text{bias}(\tilde{Y}_{\text{CAL}}) \approx \text{bias}(\tilde{Y}_{\text{EXP}}) \approx 0$. In the context of the simulation in this section, if $\theta_k = 0.70$ for all k is taken to be an additional response distribution, Table 10.1 will in all 16 cells show nearly zero values of both *relbias in %* and $\text{Av}(H_1) \times 10^3$, from the weakest cell (1G + 1G) all the way to the cell of the most powerful **x**-vector (8G + 8G). There is no bias to be removed by an improvement of the **x**-vector. If in practice the indicator (H_1) does not react to an enlargement of the **x**-vector, there is no incentive to seek beyond the simplest vector formulation. It could signify one of three possibilities: The y-variable in question is not subject to nonresponse bias, or that the response probability is almost constant, or that none of the available **x**-vectors is capable of reducing an existing bias.

To save space we do not show the corresponding tables for $\text{Av}(H_2)$ and $\text{Av}(H_3)$. By mathematical necessity, both quantities increase in the nested transitions. Not shown either are the counterparts of Table 10.1 for the other three response distributions. The patterns are similar.

Table 10.2 for IncExp$(10 + x_1 + x_2)$ and Table 10.3 for IncExp$(10 + y)$ show how $\text{Av}(H_1)$, $\text{Av}(H_2)$ and $\text{Av}(H_3)$ rank the 16 **x**-vectors, represented by their value of *relbias*. To measure the success of ranking, we computed the Spearman rank correlation coefficient, denoted *rancor*, between *relbias* and the value of the indicator, based on the 16 values of each. For $\text{Av}(H_1)$, the bottom line of the two tables shows $|rancor| = 1$, for perfect ranking. For these data, $|rancor|$ is near one also for $\text{Av}(H_2)$ and $\text{Av}(H_3)$ (more generally, the ranking obtained with H_2 and H_3 may be good, but is data dependent).

Table 10.2
Value, in ascending order, of relbias in %, and corresponding value and rank of $\text{Av}(H_1) \times 10^3$, $\text{Av}(H_2) \times 10^3$ and $\text{Av}(H_3) \times 10^3$, for 16 auxiliary vectors. Bottom line: Value of Spearman rank correlations, *rancor*. Response distribution IncExp$(10 + x_1 + x_2)$

relbias	$\text{Av}(H_1) \times 10^3$		$\text{Av}(H_2) \times 10^3$		$\text{Av}(H_3) \times 10^3$	
0.2	101	(1)	127	(1)	232	(1)
0.5	99	(2)	119	(2)	225	(2)
0.5	98	(3)	118	(3)	224	(3)
0.8	96	(4)	109	(4)	217	(4)
1.3	93	(5)	109	(5)	216	(5)
1.5	91	(6)	105	(6)	213	(6)
1.8	89	(7)	98	(7)	207	(7)
1.9	88	(8)	94	(8)	205	(8)
3.2	78	(9)	80	(11)	192	(9)
3.4	76	(10)	90	(9)	188	(11)
4.1	70	(11)	84	(10)	190	(10)
4.1	70	(12)	77	(12)	175	(13)
5.0	64	(13)	70	(13)	179	(12)
6.4	52	(14)	52	(14)	146	(15)
7.3	46	(15)	46	(15)	156	(14)
13.2	0	(16)	0	(16)	0	(16)
Rancor	-1.00		-0.99		-0.99	

There is one notable contrast between the results on *relbias* for the two response distributions in Tables 10.2 and 10.3. The best among the auxiliary vectors leave considerably more bias for the non-ignorable IncExp$(10 + y)$ than for IncExp$(10 + x_1 + x_2)$. This is not unexpected, and it is important to note that considerable bias reduction is obtained for the non-ignorable case as well.

In the simulation, the over-adjustment mentioned in Section 4, $\Delta_A > \Delta_T > 0$ (when $(\tilde{Y}_{\text{EXP}})$ has positive bias) or $\Delta_A < \Delta_T < 0$ (when $\tilde{Y}_{\text{EXP}}$ has negative bias), happens for some outcomes (s, r). The frequency varies with the strength of the auxiliary vector and is different for different response distributions. The cell for which this over-adjustment is most likely to occur is 8G + 8G, the most powerful of the 16 auxiliary vectors. For IncExp$(10 + x_1 + x_2)$, the bias is almost completely removed for cell 8G + 8G; *relbias* is only 0.2%. Hence $\tilde{Y}_{\text{CAL}}$ is close to the unbiased $\tilde{Y}_{\text{FUL}}$, Δ_A is near Δ_T, and $\Delta_A > \Delta_T$ happened for 45.6% of all outcome (s, r). By contrast, for the non-ignorable case IncExp$(10 + y)$, the incidence of $\Delta_A > \Delta_T$

was only 0.1% for the cell 8G + 8G. Although that cell brings considerable bias reduction (compared to the primitive 1G + 1G), there is bias remaining, and as a consequence, $\Delta_A > \Delta_T$ almost never happens.

We do not show the corresponding tables for DecExp($x_1 + x_2$) and DecExp(y). The lowest value of *rancor* was 0.94, recorded for Av(H_3) in the case of DecExp($x_1 + x_2$).

A question not addressed in Tables 10.2 and 10.3 is: How often, over a long series of outcomes (s, r), does a given indicator $H(\mathbf{x}_k)$ succeed in pointing correctly to the preferred **x**-vector? To answer this, let $\mathbf{x}_{1k}$ and $\mathbf{x}_{2k}$ be two vectors selected for comparison. If the absolute value of the bias of $\hat{Y}_{\text{CAL}}(\mathbf{x}_{2k})$ is smaller than that of $\hat{Y}_{\text{CAL}}(\mathbf{x}_{1k})$, we would like to see that $H(\mathbf{x}_{2k}) \geq H(\mathbf{x}_{1k})$ holds for a vast majority of all outcomes (s, r), because then the indicator $H(\cdot)$ delivers with high probability the correct decision to prefer $\mathbf{x}_{2k}$. Because $H(\mathbf{x}_k)$ has sampling variability, its success rate (the rate of correct indication) depends on the sample size, and we expect it to increase with sample size.

Table 10.3
Value, in ascending order, of relbias in %, and corresponding value and rank of $\text{Av}(H_1) \times 10^3$, $\text{Av}(H_2) \times 10^3$ and $\text{Av}(H_3) \times 10^3$, for 16 auxiliary vectors. Bottom line: Value of Spearman rank correlations, *rancor*. Response distribution IncExp (10 + y)

relbias	$\text{Av}(H_1)\times 10^3$		$\text{Av}(H_2)\times 10^3$		$\text{Av}(H_3)\times 10^3$	
3.6	74	(1)	91	(1)	165	(1)
3.9	71	(2)	84	(2)	158	(2)
4.0	71	(3)	83	(3)	156	(3)
4.3	68	(4)	76	(5)	149	(5)
4.4	68	(5)	78	(4)	153	(4)
4.9	64	(6)	68	(8)	142	(3)
4.9	63	(7)	72	(6)	146	(6)
5.3	60	(8)	69	(7)	143	(7)
5.4	60	(9)	64	(9)	137	(9)
6.0	55	(10)	59	(10)	132	(10)
6.2	53	(11)	54	(11)	128	(11)
7.2	46	(12)	54	(12)	122	(12)
7.9	41	(13)	41	(14)	111	(13)
7.9	40	(14)	43	(13)	109	(14)
9.6	27	(15)	27	(15)	90	(15)
13.1	0	(16)	0	(16)	0	(16)
Rancor		-1.00		-0.99		-0.99

We threw some light on this question by extending the Monte Carlo experiment: 5,000 outcomes (s, r) were realized, first with sample size $n = 1{,}000$, then with sample size $n = 2{,}000$ (the response set r is realized according to one of the four response distributions, declaring unit k "responding" as a result of a Bernoulli trial with the specified probability θ_k). We computed the success rate as the proportion of all outcomes (s, r) in which the correct indication materializes in a confrontation of two different **x**-vectors. Several pairwise comparisons of this kind were carried out. Typical results are shown in Table 10.4, for IncExp($10 + x_1 + x_2$). The upper entry in a table cell shows the success rate in % for $n = 1{,}000$, the lower entry shows that rate for $n = 2{,}000$. Shown in parenthesis is the value of *relbias* for the vectors in question.

"Severe tests" are preferred, that is, confrontations of vectors with a small difference in absolute *relbias*, because the correct decision is then harder to obtain. There is a priori no reason why one of the indicators should always outperform the others in this study. In the five severe tests in Table 10.4, H_1 has, on the whole, better success rates than H_2 and H_3. The success rate of H_1 improves by doubling the sample size, and tends as expected to be greater when the *relbias* values are further apart. The case 4G + 8G *vs.* 8G + 8G compares nested **x**-vectors, so it is known beforehand that H_2 and H_3 give perfect success rates.

Table 10.4
Selected pairwise comparisons of auxiliary vectors; percentage of outcomes with correct indication, for the indicators H_1, H_2 and H_3. Within parenthesis, relbias in %. Upper entry: n = 1,000 lower entry: $n = 2{,}000$. Response distribution IncExp (10 + $x_1 + x_2$)

Cells compared	Percent outcomes with correct indication		
	H_1	H_2	H_3
4G + 8G(0.5) *vs.*	90.0	100.0	100.0
8G + 8G(0.2)	96.4	100.0	100.0
4G + 2G(1.8) *vs.*	66.8	86.0	70.7
2G + 8G(1.5)	74.2	89.0	67.4
1G + 8G(4.1) *vs.*	74.3	70.3	45.0
8G + 1G(3.4)	82.8	78.0	43.3
4G + 1G(4.1) *vs.*	90.6	61.4	83.9
2G + 2G(3.2)	97.0	68.8	92.3
1G + 2G(7.3) *vs.*	77.4	77.4	34.5
2G + 1G(6.5)	85.9	85.9	28.8

11. Concluding remarks

In this article, we address survey situations where many alternative auxiliary vectors (**x**-vectors) can be created and considered for use in the calibration estimator $\tilde{Y}_{\text{CAL}}$. For any given **x**-vector, a certain unknown bias remains in $\tilde{Y}_{\text{CAL}}$; we wish by an appropriate choice of **x**-vector to make that bias as small as possible. Hence we examine the bias ratio defined by (4.2) and (4.3). The component Δ_A of the bias ratio was expressed, in (5.8) to (5.10), as product of easily interpreted statistical measures. This led us to suggest several alternative bias indicators, for use in evaluating different **x**-vectors in regard to their capacity to effectively reduce the bias. We studied in particular the indicator H_1 given by (5.12). It functions very well but is geared to a particular study variable y. However, a typical government survey has many study variables, and for practical reasons it is desirable to use the same **x**-vector in estimating all y-totals. A compromise becomes necessary. We argued that the indicator H_3 in (5.12) suits this purpose; it depends on the $\mathbf{x}_k$ but not on any y-data. A topic for further research is to develop other indicators (than H_3) for the "many y-variable situation". Another topic for further work is to examine algorithms for stepwise selection of x-variables with the indicator H_1, other than the one used in Section 9.

Acknowledgements

The authors are grateful to the referees and to the Associate Editor for comments contributing to an improvement of this paper.

References

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

Eltinge, J., and Yansaneh, I. (1997). Diagnostics for the formation of nonresponse adjustment cells with an application to income nonresponse in the US Consumer Expenditure Survey. *Survey Methodology*, 23, 33-40.

Kalton, G., and Flores-Cervantes, I. (2003). Weighting methods. *Journal of Official Statistics*, 19, 81-98.

Kott, P.S. (2006). Using calibration weighting to adjust for nonresponse and coverage errors. *Survey Methodology*, 32, 133-142.

Rizzo, L., Kalton, G. and Brick, J.M. (1996). A comparison of some weighting adjustment methods for panel nonresponse. *Survey Methodology*, 22, 43-53.

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

Särndal, C.-E., and Lundström, S. (2008). Assessing auxiliary vectors for control of nonresponse bias in the calibration estimator. *Journal of Official Statistics*, 4, 251-260.

Schouten, B. (2007). A selection strategy for weighting variables under a not-missing-at-random assumption. *Journal of Official Statistics*, 23, 51-68.

Schouten, B., Cobben, F. and Bethlehem, J. (2009). Indicators for the representativeness of survey response. *Survey Methodology*, 35, 101-113.

Thomsen, I., Kleven, Ø., Wang, J.H. and Zhang, L.C. (2006). Coping with deceasing response rates in Statistics Norway. Recommended practice for reducing the effect of nonresponse. Reports 2006/29. Oslo: Statistics Norway.