Research and Development – Methodology reports from Statistics Sweden 2009:1

Statistiska centralbyrån Statistics Sweden

Design for estimation: Identifying auxiliary vectors to reduce nonresponse bias

Carl-Erik Särndal
Sixten Lundström

The series entitled "**Research and Development – Methodology Reports from Statistics Sweden**" presents results from research activities within Statistics Sweden. The focus of the series is on development of methods and techniques for statistics production. Contributions from all departments of Statistics Sweden are published and papers can deal with a wide variety of methodological issues.

Previous publication:

2006:1 Quantifying the quality of macroeconomic variables

2006:2 Stochastic population projections for Sweden

2007:1 Jämförelse av röjanderiskmått för tabeller

2007:2 Assessing auxiliary vectors for control of nonresponse bias in the calibration estimator.

2007:3 Kartläggning av felkällor för bättre aktualitet

2008:1 Optimalt antal kontaktförsök i en telefonundersökning

.

Research and Development – Methodology reports from Statistics Sweden 2009:1

Design for estimation: Identifying auxiliary vectors to reduce nonresponse bias

Carl-Erik Särndal
Sixten Lundström

Statistiska centralbyrån
2009

Research and Development – Methodology reports from Statistics Sweden 2009:1

Design for estimation: Identifying auxiliary vectors to reduce nonresponse bias

Statistics Sweden
2009

Producer

Statistics Sweden, Research and Development Department
SE-701 89 ÖREBRO
+ 46 19 17 60 00

Inquiries

Carl-Erik Särndal, +46 19 17 60 43
carl.sarndal@rogers.com

Sixten Lundström, + 46 19 17 64 96
sixten.lundstrom@scb.se

Cover Ateljén, SCB

ISSN 1653-7149

URN:NBN:SE:SCB-2009-X103OP0901_pdf (pdf)

This publication is only published electronically on Statistics Sweden's website www.scb.se

Preface

Nonresponse occurs in essentially all sample surveys, and, seemingly, at ever increasing rates. As a result, the quality of the statistics produced in a survey is at stake, unless powerful adjustment procedures can be brought to bear. Statistics Sweden is in a relatively favourable position, because the many administrative registers that are available provide a rich source of auxiliary information useful for nonresponse adjustment in estimation by calibration.

The present article by Carl-Erik Särndal and Sixten Lundström, *Design for estimation: Identifying auxiliary vectors to reduce nonresponse bias* is a continuation of Statistics Sweden's commitment to research on nonresponse adjustment methods. It further develops the ideas in an earlier article by the same two authors, *Assessing auxiliary vectors for control of nonresponse bias in the calibration estimator*, Research and Development report 2007:2.

The indicators presented in the present article provide further useful tools for selecting the most powerful ones among the many auxiliary variables available in Sweden for purposes of nonresponse bias adjustment.

Statistics Sweden, June 2009

Folke Carlsson

Åke Bruhn

Contents

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 et al (2006). Weighing 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.

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 $\mathbf{x}$-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, one using data for a constructed population, and the other using data from a large survey at Statistics Sweden.

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. We assume $r \subset s \subset U$, and r non-empty. The (design weighted) response rate is

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

Ordinarily a survey has many study variables. A typical one, 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; we focus on a typical one. (If A is a set of units, $A \subseteq U$, a sum $\sum_{k \in A}$ will be written as $\sum_A$.)

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

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

Tied to the k^{th} unit is the vector $(y_k, \mathbf{x}_k, \pi_k, \theta_k)$. 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. The response probability $\theta_k = \Pr(k \in r | s)$ is unknown for all *k*. It is assumed positive and independent of s. Although called "response probability", θ_k is seen more generally as the probability that the value y_k gets recorded. With probability $1 - \theta_k$, it goes missing, for whatever reason. Apart from the notion of response probabilities we make no assumptions on the response mechanism.

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 in this paper and does not severely restrict $\mathbf{x}_k$. Most $\mathbf{x}$-vectors that are 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, $j = 1, 2, ..., J$; (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

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

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 $\hat{Y}_{CAL} = \sum_r w_k y_k$ 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).

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

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

where

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

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, $\hat{Y}_{CAL}$ is

usually preferred to $\widetilde{Y}_{CAL}$ 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 $\hat{Y}_{CAL}$ and $\widetilde{Y}_{CAL}$, and we focus on the latter. Under liberal conditions, the difference between the bias of $N^{-1}\hat{Y}_{CAL}$ and that of $N^{-1}\widetilde{Y}_{CAL}$ 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 $\widetilde{Y}_{CAL}$ defined by (2.5) is

$$\widetilde{Y}_{CAL} = (\sum_s d_k \mathbf{x}_k)' \mathbf{B}_{\mathbf{x}} \tag{2.7}$$

where

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

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 of these shows the set of units over which the quantity is computed and the second indicates the weighting, as for example in $\mathbf{B}_{\mathbf{x}|r;d}$ in (2.8), and in weighted means such as $\bar{y}_{r;d} = \sum_r d_k y_k / \sum_r d_k$.

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 $\widetilde{Y}_{CAL}$ is the expansion estimator:

$$\widetilde{Y}_{EXP} = (1/P)\sum_r d_k y_k = \hat{N}\,\bar{y}_{r;d} \tag{3.1}$$

where $\hat{N} = \sum_s d_k$ is design unbiased for the population size N. The bias of $\widetilde{Y}_{EXP}$ 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 $\sum_U \mathbf{x}_k^*$,

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

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

$$\widetilde{Y}_{FUL} = \sum_s d_k y_k = \hat{N}\,\bar{y}_{s;d} \tag{3.2}$$

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 $\hat{Y}_{FUL}$ and $\widetilde{Y}_{FUL}$. The difference in bias between the two is of little consequence, even for modest sample sizes. We can focus on $\widetilde{Y}_{FUL}$.

4 The bias ratio

Consider $\widetilde{Y}_{CAL}$, $\widetilde{Y}_{EXP}$ and $\widetilde{Y}_{FUL}$, defined respectively by (2.7), (3.1) and (3.2). The nearly unbiased $\widetilde{Y}_{FUL}$ represents an ideal that cannot be computed, since it depends on missing y-values. Both $\widetilde{Y}_{EXP}$ (generated by the primitive vector $\mathbf{x}_k = 1$) and $\widetilde{Y}_{CAL}$ (generated by a better $\mathbf{x}$-vector) are computable under nonresponse, but biased. As the $\mathbf{x}$-vector improves, $\widetilde{Y}_{CAL}$ will distance itself from $\widetilde{Y}_{EXP}$ and come near the nearly unbiased $\widetilde{Y}_{FUL}$.

This leads us to consider three deviations: $\widetilde{Y}_{EXP} - \widetilde{Y}_{FUL}$, $\widetilde{Y}_{EXP} - \widetilde{Y}_{CAL}$ and $\widetilde{Y}_{CAL} - \widetilde{Y}_{FUL}$, of which only the middle one is computable. The unknown "deviation total", $\widetilde{Y}_{EXP} - \widetilde{Y}_{FUL}$, is decomposable as "deviation accounted for" (through the choice of $\mathbf{x}$-vector) plus "deviation remaining":

$$\widetilde{Y}_{EXP} - \widetilde{Y}_{FUL} = (\widetilde{Y}_{EXP} - \widetilde{Y}_{CAL}) \; + \; (\widetilde{Y}_{CAL} - \widetilde{Y}_{FUL}) \tag{4.1}$$

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

$$\text{bias ratio} \; = \; \frac{\widetilde{Y}_{CAL} - \widetilde{Y}_{FUL}}{\widetilde{Y}_{EXP} - \widetilde{Y}_{FUL}} \tag{4.2}$$

The goal is to choose the auxiliary vector $\mathbf{x}_k$ used in $\widetilde{Y}_{CAL}$ so that the bias ratio is small. We scale the three deviations in (4.1) 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

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

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

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

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.

To summarize, for a given survey outcome (s,r) and a given y-variable, the three deviations have the following features: (i) Δ_T cannot be computed; it depends on unobserved (as well as observed) y_k-values, but not on any $\mathbf{x}_k$-values ; (ii) Δ_A is computable; it depends on y_k for $k \in r$ and on $\mathbf{x}_k$ for $k \in s$; the choice of $\mathbf{x}$-vector determines the value of Δ_A; (iii) Δ_R cannot be computed; it depends on unobserved y_k, and on $\mathbf{x}_k$ for $k \in s$.

For a given survey outcome (s,r) and a given y-variable, the total deviation $\Delta_T = \bar{y}_{r;d} - \bar{y}_{s;d}$ is an unknown constant value, not influenced by the choice of $\mathbf{x}$-vector. It can have either sign. Suppose $\Delta_T > 0$, indicating a positive bias in $\tilde{Y}_{EXP}$, as when large units respond with greater propensity than small ones. Then Δ_A and Δ_R are usually of the same positive sign, although not necessarily so for all choices of $\mathbf{x}$-vector. It can happen that $\Delta_A > \Delta_T$ so that $\Delta_R < 0$. When the $\mathbf{x}$-vector used for $\tilde{Y}_{CAL}$ becomes progressively more powerful, Δ_A tends to come near Δ_T, leaving a small Δ_R. If $\Delta_T < 0$, these tendencies are reversed.

The bias ratio (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 $\mathbf{x}_k = 1$. However, bias ratio ≈ 0 is a desirable goal. For the given outcome (s, r) and the given y-variable, we obtain this goal by finding an $\mathbf{x}$-vector that gives a large absolute value of the computable numerator $\Delta_A = (\bar{\mathbf{x}}_{r;d} - \bar{\mathbf{x}}_{s;d})'\mathbf{B}_\mathbf{x}$. This is within our reach. Whatever our final choice, the remaining bias of $\widetilde{Y}_{CAL}$ is unknown.

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 $\mathbf{x}$-vector is one capable of controlling bias in all estimators. This is usually not possible without compromise, as we discuss later.

The form of (4.3) may suggest a reasoning which is however misleading: Suppose a certain 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 the message, which may be false, that the postulated vector $\mathbf{x}_k$ is efficient. The weakness of the argument is that nonresponse causes $\mathbf{B}_\mathbf{x}$ to be biased for a regression vector that may perfectly well describe a y-to-$\mathbf{x}$ relationship in the population. More incisive analysis is required. Further comments on this issue are given in Section 8.

5 Expressing the deviation accounted for

The responding unit k receives the weight $d_k m_k$ in the estimator $\widetilde{Y}_{CAL} = \sum_r d_k m_k y_k$. The factor $m_k = (\sum_s d_k \mathbf{x}_k)'(\sum_r d_k \mathbf{x}_k \mathbf{x}_k')^{-1} \mathbf{x}_k$ brings a nonresponse adjustment to 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 $\widetilde{Y}_{CAL}$ only for $k \in r$). We have

$$\sum_r d_k m_k \mathbf{x}_k = \sum_s d_k \mathbf{x}_k \;\; ; \quad \sum_r d_k m_k = \sum_s d_k \;\; ;$$
$$\sum_r d_k m_k^2 = \sum_s d_k m_k \qquad (5.1)$$

Two weighted means are needed :

$$\overline{m}_{r;d} = \frac{\sum_r d_k m_k}{\sum_r d_k} = \frac{\sum_s d_k}{\sum_r d_k} = \frac{1}{P} \quad ; \quad \overline{m}_{s;d} = \frac{\sum_s d_k m_k}{\sum_s d_k} \qquad (5.2)$$

where P is the response rate (2.1). Thus the average adjustment factor in $\widetilde{Y}_{CAL} = \sum_r d_k m_k y_k$ is $1/P$, regardless of the choice of $\mathbf{x}$-vector. Whether a chosen $\mathbf{x}$-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

$$S_m^2 = S_{m|r;d}^2 = \frac{1}{\sum_r d_k} \sum_r d_k (m_k - \overline{m}_{r;d})^2 \qquad (5.3)$$

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

$$S_m^2 = \overline{m}_{r;d} (\overline{m}_{s;d} - \overline{m}_{r;d}) \qquad (5.4)$$

The coefficient of variation of the m_k is

$$cv_m = \frac{S_m}{\overline{m}_{r;d}} = \sqrt{\frac{\overline{m}_{s;d}}{\overline{m}_{r;d}} - 1} \tag{5.5}$$

The weighted variance of the study variable y is given by

$$S_y^2 = S_{y|r;d}^2 = \sum_r d_k (y_k - \overline{y}_{r;d})^2 / \sum_r d_k \tag{5.6}$$

(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

$$Cov(y,m) = Cov(y,m)_{r;d} = \frac{1}{\sum_r d_k} \sum_r d_k (m_k - \overline{m}_{r;d})(y_k - \overline{y}_{r;d}) \tag{5.7}$$

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

The deviation $\Delta_A = (\overline{\mathbf{x}}_{r;d} - \overline{\mathbf{x}}_{s;d})'\mathbf{B}_{\mathbf{x}}$ is a crucial factor in the bias ratio (4.3). We seek an $\mathbf{x}$-vector that makes Δ_A large. Computational tools to assist this search are expressed in (5.8) to (5.10). Their derivation by linear algebra is deferred to Section 7. Their use in stepwise and other methods for selecting *x*-variables is explained in Section 6, and empirically illustrated in Sections 9 and 10. We can factorize Δ_A / S_y as

$$\Delta_A / S_y = -R_{y,m} \times cv_m \tag{5.8}$$

Two simple multiplicative factors determine Δ_A / S_y: The coefficient of variation cv_m, which is free of y_k and computed from the known $\mathbf{x}_k$ alone, and the (positive or negative) correlation coefficient $R_{y,m}$. Another simple representation of (5.8) is

$$\Delta_A / S_y = F \times R_{y,\mathbf{x}} \times cv_m \tag{5.9}$$

where $R_{y,\mathbf{x}} = \sqrt{R_{y,\mathbf{x}}^2}$ is the coefficient of multiple correlation between y and $\mathbf{x}$, $R_{y,\mathbf{x}}^2$ is the proportion of the y-variance S_y^2 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

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

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.

It follows from (5.8) that $0 \le |\Delta_A|/S_y \le cv_m$ whatever the y-variable. A shaper inequality is $|\Delta_A|/S_y \le R_{y,\mathbf{x}} \times cv_m$, 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 $|\Delta_A|/S_y \approx |F_0| \times R_{y,\mathbf{x}} \times cv_m$.

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

$$H_0 = \Delta_A/S_y = -R_{y,m} \times cv_m \qquad (5.11)$$

As borne out by theory in Section 8 and by the empirical work in Section 9, over a long run of outcomes (s,r), the average of H_0 tracks the average deviation $\hat{Y}_{CAL} - Y$ (which measures the bias of $\hat{Y}_{CAL}$) 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):

$$H_1 = |\Delta_A|/S_y = |R_{y,m}| \times cv_m \quad ; \quad H_2 = R_{y,\mathbf{x}} \times cv_m \quad ; \quad H_3 = cv_m \qquad (5.12)$$

Our main alternatives are H_1 and H_3 . Of these, H_1 is directly linked to Δ_A, which we want to maximize, 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 **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

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: The bias remaining in the calibration estimator depends on the y-variable. Some y-variables are more bias prone than others. An objective is to identify an $\mathbf{x}$-vector that succeeds in reducing the bias for the y-variables deemed to be the most important ones in the survey. For the discussion here we assume that one important y-variable has been singled out. (If more than one y-variable needs to be taken into account, a perhaps not so easy compromise must be struck, which suggests Scenario 2 below.) In the interest of bias reduction, we use the indicator $H_1 = |\Delta_A|/S_y = |R_{y,m}| \times cv_m$ and choose the $\mathbf{x}$-vector to make its value as large as possible. An ad hoc alternative is to use the indicator $H_2 = R_{y,\mathbf{x}} \times cv_m$, and strive to make it as large as possible.

Scenario 2. The objective is to identify a general purpose $\mathbf{x}$-vector, efficient for all or most y-variables in the survey. This suggests to use $H_3 = cv_m$ as a compromise indicator, and to choose the $\mathbf{x}$-vector that maximizes H_3. To the same effect, Särndal and Lundström (2005, 2008) used the indicator $S_m^2 = H_3^2 / P^2$, motivating it by showing that the derived variable value m_k given by (2.6) is a predictor of the unknown inverse response probability $1/\theta_k$, and that choosing the $\mathbf{x}$-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 $\mathbf{x}$-vectors is prepared, based on appropriate judgment. We compute the chosen indicator for *every* candidate $\mathbf{x}$-vector, and settle for the vector that gives the highest indicator value. The resulting $\mathbf{x}$-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 $\mathbf{x}$-vector by a stepwise forward (or stepwise backward) selection from among the available *x*-variables, one variable at a time, using the successive changes in the value of the chosen indicator to signal the inclusion (or exclusion) of a given *x*-variable at a given step. Suppose that we are comparing two $\mathbf{x}$-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 necessarily increase the value of H_2 and H_3, but that transition does not guarantee an increased value for the most appropriate indicator, H_1. These matters are illustrated in the empirical Sections 9 and 10.

7 Derivations

For given y-variable and outcome (s,r), we seek an $\mathbf{x}$-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) as large as possible, in absolute value. In this section we prove the factorizations $\Delta_A/S_y = -R_{y,m} \times cv_m = F \times R_{y,\mathbf{x}} \times cv_m$ in (5.8) and (5.9). First we express cv_m^2 as a quadratic form in the vector that contrasts the $\mathbf{x}$-mean in the response set r with the $\mathbf{x}$-mean in the sample s. Define

$$\mathbf{D} = \bar{\mathbf{x}}_{r;d} - \bar{\mathbf{x}}_{s;d} \quad ; \quad \boldsymbol{\Sigma} = \sum_r d_k \mathbf{x}_k \mathbf{x}_k' / \sum_r d_k \tag{7.1}$$

Then, with P given by (2.1),

$$cv_m^2 = P^2 \times S_m^2 = \mathbf{D}'\boldsymbol{\Sigma}^{-1}\mathbf{D} \tag{7.2}$$

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

$$\bar{\mathbf{x}}_{r;d}'\boldsymbol{\Sigma}^{-1}\bar{\mathbf{x}}_{r;d} = \bar{\mathbf{x}}_{r;d}'\boldsymbol{\Sigma}^{-1}\bar{\mathbf{x}}_{s;d} = 1 \tag{7.3}$$

Next we define the covariance vector as

$$\mathbf{C} = \left(\sum_r d_k (\mathbf{x}_k - \bar{\mathbf{x}}_{r;d})(y_k - \bar{y}_{r;d})\right)/(\sum_r d_k) \tag{7.4}$$

whereby we can write Δ_A as a bilinear form:

$$\Delta_A = \mathbf{D}'\mathbf{B}_{\mathbf{x}} = \mathbf{D}'\boldsymbol{\Sigma}^{-1}\mathbf{C} \tag{7.5}$$

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

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

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

with

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

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

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

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

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

The proportion of the y-variance S_y^2 explained by the regression of y on *m* is

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

The proportions $R^2_{y,\mathbf{x}}$ and $R^2_{y,m}$ satisfy $R^2_{y,m} \le R^2_{y,\mathbf{x}} \le 1$.

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

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

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

$R^2_{y,\mathbf{x}} = \sum_r d_k (\hat{y}(\mathbf{x})_k - \bar{y}_{r;d})^2 / [(\sum_r d_k)\, S_y^2] = \mathbf{C}'\mathbf{\Sigma}^{-1}\mathbf{C} / S_y^2$, 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

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

It then follows from (7.2) that $R_{y,m} = Cov(y,m)/(S_y S_m)$ 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 $B_m = Cov(y,m)/S_m^2 = -P\,(\mathbf{D}'\mathbf{\Sigma}^{-1}\mathbf{C})/(\mathbf{D}'\mathbf{\Sigma}^{-1}\mathbf{D})$. 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 $R^2_{y,m}$ in (7.10). Finally, $R^2_{y,m} \le R^2_{y,\mathbf{x}}$ follows from the Cauchy-Schwarz inequality for a bilinear form: $(\mathbf{D}'\mathbf{\Sigma}^{-1}\mathbf{C})^2 \le (\mathbf{D}'\mathbf{\Sigma}^{-1}\mathbf{D})(\mathbf{C}'\mathbf{\Sigma}^{-1}\mathbf{C})$. □

The fact that $R^2_{y,m} \le R^2_{y,\mathbf{x}} \le 1$ illustrates 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 alternative 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), $-R_{y,m} cv_m = \mathbf{D}'\mathbf{\Sigma}^{-1}\mathbf{C} / S_y = \Delta_A / S_y$, as claimed by formula (5.8). Moreover, (7.8), (7.9) and (7.7) 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 Two remarks

Two issues are examined in this section. The first concerns the relationship between bias and goodness of fit. The second establishes the essentially perfect relation existing between the long-run average of $\Delta_A = \hat{N}^{-1}(\widetilde{Y}_{EXP} - \widetilde{Y}_{CAL})$ and the bias of $\widetilde{Y}_{CAL}$ or $\hat{Y}_{CAL}$.

The total deviation defined 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, $\hat{N}\Delta_R = \widetilde{Y}_{CAL} - \widetilde{Y}_{FUL}$ would have been an estimate of the bias of $\widetilde{Y}_{CAL}$ (and of that of $\hat{Y}_{CAL}$). How is "small bias" (a small value of Δ_R) related to the goodness of fit of the model $y_k = \boldsymbol{\beta}'\mathbf{x}_k + \varepsilon_k$? The residuals from that fit determine the value of $\Delta_R = \bar{\mathbf{x}}'_{s;d}\mathbf{B}_{\mathbf{x}|r;d} - \bar{y}_{s;d}$, where $\mathbf{B}_{\mathbf{x}|r;d}$ is given by (2.8). (In this section the more precise notation $\mathbf{B}_{\mathbf{x}|r;d}$ is preferable to the simpler $\mathbf{B}_{\mathbf{x}}$ used earlier.) Two aspects of the fit of that model are: (i) The computable fit to the data $(y_k, \mathbf{x}_k)$ observed for $k \in r$; (ii) The hypothetical fit to the data $(y_k, \mathbf{x}_k)$ for $k \in s$, some observed, some not. Let us consider the two cases.

Weighted LSQ fit on the observed data $(y_k, \mathbf{x}_k)$ for $k \in r$ gives the residuals $e_{k|r;d} = y_k - \mathbf{x}'_k\mathbf{B}_{\mathbf{x}|r;d}$, defined and computable for $k \in r$, with the property $\sum_r d_k e_{k|r;d} = 0$. For $k \in s - r$, define $K_k = y_k - \mathbf{x}'_k\mathbf{B}_{\mathbf{x}|r;d}$. Although $e_{k|r;d}$ and K_k agree in form, different notation is required, because, in contrast to $e_{k|r;d}$, K_k is not a regression residual, is not computable, and has an unknown non-zero mean $\bar{K}_{s-r;d} = \sum_{s-r} d_k K_k / \sum_{s-r} d_k$. We have

$$\Delta_R = -(1-P)\bar{K}_{s-r;d} \tag{8.1}$$

The expression (8.1), which may be far from zero, does not depend on the residuals $e_{k|r;d}$. Regardless of whether the fit is good (residuals small; $R^2_{y,\mathbf{x}}$ near one) or poor (residuals large; $R^2_{y,\mathbf{x}}$ near

zero), the remaining deviation Δ_R is still given by (8.1), and the term Δ_A remains unchanged at $\Delta_A = \mathbf{D}'\mathbf{B}_{\mathbf{x}|r;d} = \mathbf{D}'\boldsymbol{\Sigma}^{-1}\mathbf{C}$. Even if the fit is perfect for the respondents, so that $e_{k|r,d} = 0$ for all $k \in r$ and $R^2_{y,\mathbf{x}} = 1$, there is no indication that the bias is small. Given the respondent data $(y_k, \mathbf{x}_k)$, $k \in r$, the remaining deviation Δ_R is unchanged, given by (8.1).

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

$$\hat{Y}_{imp} = \sum_r d_k y_k + \sum_{s-r} d_k \hat{y}_k$$

As is easily verified, $\hat{Y}_{imp} = \widetilde{Y}_{CAL}$, so $\hat{Y}_{imp}$ has the same exposure to bias as $\widetilde{Y}_{CAL}$. When the nonresponse causes a skewed selection of y-values, the imputed values computed from that skewed selection will misrepresent the y-values in the sample *s* or in the whole population U.

(ii) The weighted LSQ regression fit to the data points $(y_k, \mathbf{x}_k)$ for $k \in s$ is hypothetical, because y_k is missing for $k \in s - r$. The hypothetical regression coefficient vector is $\mathbf{B}_{\mathbf{x}|s;d} = \left(\sum_s d_k \mathbf{x}_k \mathbf{x}'_k\right)^{-1} \sum_s d_k \mathbf{x}_k y_k$, and the residuals $e_{k|s;d} = y_k - \mathbf{x}'_k\mathbf{B}_{\mathbf{x}|s;d}$ for $k \in s$ would have the property $\sum_s d_k e_{k|s;d} = 0$. 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

$$\Delta_R = (1/\hat{N}) \sum_r d_k m_k e_{k|s;d} \qquad (8.2)$$

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 $\Delta_R = \hat{N}^{-1}(\widetilde{Y}_{CAL} - \widetilde{Y}_{FUL}) = 0$. A belief that the bias is small hinges on an unverifiable assumption.

The second issue concerns the relation between the bias of $\tilde{Y}_{CAL}$ and the expected value of the indicator $H_0 = \Delta_A / S_y$ $= (\tilde{Y}_{EXP} - \tilde{Y}_{CAL}) / \hat{N} S_y = - R_{y,m} \times cv_m$. For fixed y-variable and $\mathbf{x}$-vector we have

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

Let *E* denote the expectation operator with respect to all outcomes (s, r), and denote $bias(\tilde{Y}_{CAL}) = E(\tilde{Y}_{CAL}) - Y$, $bias(\tilde{Y}_{EXP}) = E(\tilde{Y}_{EXP}) - Y$ and $C = E(\hat{N} S_y)$. Then

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

Here $bias(\tilde{Y}_{CAL})$ and $E(H_0)$ depend on the $\mathbf{x}$-vector; $bias(\tilde{Y}_{EXP})$ and C do not. When the $\mathbf{x}$-vector changes, (8.3) states that $bias(\tilde{Y}_{CAL})$ and $E(H_0)$ are essentially linearly related. If $\mathbf{x}_{1k}$ and $\mathbf{x}_{2k}$ are two possible $\mathbf{x}$-vectors, the bias differential is proportional to the difference in the expected value of H_0:

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

where $E_i = E\big(H_0(\mathbf{x}_{ik})\big)$ for $i = 1, 2$. The near-perfect linearity was confirmed by the Monte Carlo study in Section 9.

9 Empirical validation with a constructed population

The Monte Carlo study reported in this section was carried out to test how well the different indicators succeed in ranking potential $\mathbf{x}$-vectors with respect to the amount of bias that they leave remaining in the calibration estimator $\hat{Y}_{CAL}$. We use a constructed population of size $N = 6{,}000$ with one continuous y-variable and two continuous x-variables. The two x-variables were used to form 16 alternative auxiliary $\mathbf{x}$-vectors of the categorical type. We study the indicators H_0 to H_3 defined in (5.11) and (5.12). We confirm that, over a long run of outcomes (s,r), the average of $H_0 = \Delta_A / S_y = -R_{y,m} \times cv_m$ tracks the bias of the calibration estimator, as measured by the average of $\hat{Y}_{CAL} - Y$, in an essentially perfect linear manner, when the $\mathbf{x}$-vector moves through its 16 different formulations. This property holds, as it should, for the several response distributions in the experiment. We find that H_1 in particular, but H_2 and H_3 also, have strong relationship to the bias of $\hat{Y}_{CAL}$.

For this study we created values $(y_k, \mathbf{x}_k, \theta_k)$ for $k = 1, 2, ..., N = 6{,}000$. All 16 $\mathbf{x}$-vectors in the experiment are categorical, obtained by grouping the values x_{1k} and x_{2k} of two generated continuous auxiliary variables, x_1 and x_2. Four different response distributions are used, each with response probabilities $\theta_k = \Pr(k \in r | s)$ specified for all 6,000 units. We assume θ_k independent of *s*. The steps in creating (y_k, x_{1k}, x_{2k}) for $k = 1, 2, ..., 6{,}000$ is described in the appendix at the end. We experimented with several populations; the conclusions were similar.

Each of the two *x*-variables was transformed into four alternative group modes, denoted 8G, 4G, 2G and NG, yielding 4 × 4 = 16 different auxiliary vectors $\mathbf{x}_k$. The procedure for the variable x_1 was: The 6,000 values x_{1k} were size ordered, and 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 variable 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. For example, $\boldsymbol{\gamma}_{(x_1;8)k} = (0,0,0,0,1,0,0,0)'$ states that k is one of the 750 units in size group 5 of the x_1-variable. Next, successive group mergers are carried out, so that two adjoining groups always define a new group, every time doubling the group size and causing loss of information. 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, the merger of groups 3 and 4 forms the second new group of 1,500, and so on, and the vector value associated with unit k is $\boldsymbol{\gamma}_{(x_1;4)k}$. In mode 2G, unit k has the indicator vector $\boldsymbol{\gamma}_{(x_1;2)k}$ with 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, NG (for no grouping), 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 NG. The corresponding memberships 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 $\boldsymbol{\gamma}_{(x_2;1)k} = 1$. Finally, 4 × 4 = 16 different auxiliary vectors $\mathbf{x}_k$ are formed by combining the two kinds of group information; the 16 cells and their notation are shown in the following display.

Groups based on x_{1k}	Groups based on x_{2k}			
	Eight	Four	Two	None
Eight	8G+8G	8G+4G	8G+2G	8G+NG
Four	4G+8G	4G+4G	4G+2G	4G+NG
Two	2G+8G	2G+4G	2G+2G	2G+NG
None	NG+8G	NG+4G	NG+2G	NG+NG

The "+" indicates that the $\mathbf{x}_k$-vector has the two $\boldsymbol{\gamma}$-vectors placed side by side, the result being a calibration on the two margins, and any interaction effects are relinquished. Thus for the cell 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, giving $\mathbf{x}_k$ the dimension 8+8-1 = 15. The cell 8G+8G has the highest information content. At the other extreme, the cell NG+NG disregards all the *x*-information and $\mathbf{x}_k = 1$ for all k. There are 14 intermediate cases. For example, the cell 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; the cell 4G+NG 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 for simplicity and to avoid small cells we consider only $\mathbf{x}$-vectors that ignore that interaction.)

The four response distributions and their response probabilities θ_k, $k = 1, 2, ..., N = 6{,}000$, were specified as follows:

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

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

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

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

The constant *c* was adjusted in each option to give a mean response probability of $\bar{\theta}_U = \sum_U \theta_k / N = 0.70$. In (i) and (ii), 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 options (ii) and (iv), the response is entirely y-variable dependent, hence "purely non-ignorable".

From the constructed population of size $N = 6{,}000$ we generated $J = 5{,}000$ outcomes (s, r), where s of size $n = 1{,}000$ is drawn by simple random sampling and, for every given *s*, a response set r is realized by each of the four response distributions. For unit *k*, if included in *s*, a Bernoulli trial was carried out with the specified probability θ_k of "success", which stands for 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 $RD=(\hat{Y}_{CAL}-Y)/Y$, where $\hat{Y}_{CAL}$ is given by (2.4) and $Y=\sum_U y_k$ is the targeted y-total, known in this experimental setting. (Alternatively, we used $\tilde{Y}_{CAL}$ given by (2.5) but, as expected, the difference in bias compared with $\hat{Y}_{CAL}$ is negligible.) We also computed the indicators H_i, $i=0,1,2,3$, given by (5.11) and (5.12). Summary measures were then computed as

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

where the index j serves to indicate the computed value for the j th outcome, $j=1,2,...,5{,}000=J$. For each response distribution, this gives 16 values for each of the five *Av*-quantities. The quantity $relbias=Av(RD)$ is the Monte Carlo measure of the relative bias of $\hat{Y}_{CAL}$, $(E(\hat{Y}_{CAL})-Y)/Y$, where expectation is jointly with respect to sampling design and response distribution.

Table 9.1. Relbias in % and, within parenthesis, the value of $Av(H_1)\times 10^3$ 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}							
	Eight		Four		Two		None	
Eight	0.2	(101)	0.5	(99)	1.3	(93)	3.4	(76)
Four	0.5	(98)	0.9	(96)	1.8	(89)	4.1	(70)
Two	1.5	(91)	1.9	(88)	3.2	(78)	6.5	(52)
None	4.1	(70)	5.0	(64)	7.3	(46)	13.2	(0)

To illustrate the layout of the experiment, Table 9.1 shows, for IncExp(10+ x_1+x_2), *relbias* in % and $Av(H_1)\times 10^3$ for the 16 $\mathbf{x}$-vectors. To save space, we do not show $Av(H_0)$, $Av(H_2)$ and $Av(H_3)$. For the cell NG+NG, corresponding to the primitive vector $\mathbf{x}_k=1$, all four *Av*-quantities are zero, and *relbias* is at its highest

level. At the opposite extreme, the cell 8G+8G, representing the maximum use of auxiliary information, gives the highest value for all four *Av*-quantities, and *relbias* is at its smallest value. A mathematical property of both $Av(H_2)$ and $Av(H_3)$ is that their value increases in nested transitions, as when we move upwards within a column, or from right to left within a row. This experiment happened to be one in which *relbias* and $Av(H_1)$ also follow this row-wise and column-wise monotonic pattern. More generally, however, $Av(H_1)$ can increase or decrease in a comparison of nested $\mathbf{x}$-vectors. Not shown are the counterparts of Table 9.1 for the other three response distributions. The patterns are similar.

The summary Tables 9.2 to 9.5 show (i) that $Av(H_1)$ gives a perfect ranking of the 16 $\mathbf{x}$-vectors, and (ii) that $Av(H_1)$ tracks the value of *relbias* linearly. For these data, the ranking obtained by $Av(H_2)$ and $Av(H_3)$ is not perfect but nearly so, for these data. We computed the Spearman rank correlation coefficient, denoted *rancor*, between *relbias* and each of the three indicators, based on the 16 data points. The bottom line of Tables 9.2 to 9.5 shows that $|rancor| = 1$ for $Av(H_1)$, and $|rancor|$ is close to one for $Av(H_2)$ and $Av(H_3)$. (For Tables 9.4 and 9.5 *relbias* is always negative and shown in absolute value.)

A comparison of Tables 9.2 to 9.5 shows that the most powerful of the $\mathbf{x}$-vector (cell 8G+8G) leaves a considerably greater bias remaining for the *y*-dependent response distributions,

IncExp(10+ y) and DecExp(y), than for the two depending solely on the *x*-variables. The value of $|relbias|$ for cell 8G+8G is 8.2% in Table 9.5 for DecExp(y), contrasting with only 0.2% in Table 9.2 for IncExp(10+ $x_1 + x_2$). Important, however, is that large bias reduction is obtained for the y -dependent cases as well, in the transition from the primitive to the best $\mathbf{x}$-vector.

Table 9.2. Value, in ascending order, of relbias in %, and corresponding value and rank of $Av(H_1)\times 10^3$, $Av(H_2)\times 10^3$ and $Av(H_3)\times 10^3$, for 16 auxiliary vectors. Bottom line: Value of Spearman rank correlation, *rancor*. Response distribution IncExp(10+ x_1 + x_2)

relbias	$Av(H_1)\times 10^3$		$Av(H_2)\times 10^3$		$Av(H_3)\times 10^3$	
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	(10)	192	(9)
3.4	76	(10)	90	(11)	188	(11)
4.1	70	(11)	84	(9)	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

Table 9.3. Value, in ascending order, of relbias in %, and corresponding value and rank of $Av(H_1)\times 10^3$, $Av(H_2)\times 10^3$ and $Av(H_3)\times 10^3$, for 16 auxiliary vectors. Bottom line: Value of Spearman rank correlation, *rancor*. Response distribution IncExp(10+ y).

relbias	$Av(H_1)\times 10^3$		$Av(H_2)\times 10^3$		$Av(H_3)\times 10^3$	
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	(7)	142	(7)
4.9	63	(7)	72	(8)	146	(8)
5.3	60	(8)	69	(6)	143	(6)
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

Table 9.4. Value, in ascending order, of $|\mathbf{relbias}|$ in %, and corresponding value and rank of $Av(H_1)\times 10^3$, $Av(H_2)\times 10^3$ and $Av(H_3)\times 10^3$, for 16 auxiliary vectors. Bottom line: Value of Spearman rank correlation, *rancor*. Response distribution DecExp($x_1 + x_2$)

relbias	$Av(H_1)\times 10^3$		$Av(H_2)\times 10^3$		$Av(H_3)\times 10^3$	
2.7	160	(1)	179	(1)	329	(1)
3.5	152	(2)	168	(2)	318	(2)
3.9	148	(3)	160	(3)	300	(5)
4.7	138	(4)	148	(5)	286	(3)
4.9	137	(5)	150	(4)	306	(9)
5.5	130	(6)	138	(6)	267	(4)
6.4	121	(7)	128	(7)	270	(7)
6.6	119	(8)	123	(8)	250	(6)
7.1	113	(9)	119	(9)	291	(8)
7.6	108	(10)	113	(10)	233	(13)
8.7	97	(11)	99	(11)	224	(10)
8.8	95	(12)	97	(12)	211	(11)
9.1	92	(13)	94	(13)	249	(12)
11.6	66	(14)	66	(14)	169	(15)
12.6	55	(15)	55	(15)	182	(14)
17.7	0	(16)	0	(16)	0	(16)
rancor		1.00		1.00		0.94

Table 9.5. Value, in ascending order, of $|\mathbf{relbias}|$ in %, and corresponding value and rank of $Av(H_1)\times 10^3$, $Av(H_2)\times 10^3$ and $Av(H_3)\times 10^3$, for 16 auxiliary vectors. Bottom line: Value of Spearman rank correlation, *rancor*. Response distribution DecExp(y)

relbias	$Av(H_1)\times 10^3$		$Av(H_2)\times 10^3$		$Av(H_3)\times 10^3$	
8.2	135	(1)	146	(1)	264	(1)
8.9	128	(2)	135	(2)	250	(2)
9.0	126	(3)	133	(3)	249	(3)
9.8	117	(4)	121	(5)	233	(5)
9.8	117	(5)	123	(4)	237	(4)
10.5	110	(6)	115	(6)	230	(6)
10.9	105	(7)	108	(8)	217	(8)
11.0	105	(8)	110	(7)	224	(7)
11.5	99	(9)	101	(9)	210	(9)
12.2	91	(10)	93	(10)	202	(12)
12.9	83	(11)	84	(12)	187	(10)
12.9	83	(12)	87	(11)	204	(11)
14.4	68	(13)	69	(13)	176	(13)
14.8	63	(14)	63	(14)	162	(14)
16.8	41	(15)	41	(15)	131	(15)
20.5	0	(16)	0	(16)	0	(16)
rancor		1.00		0.99		0.99

An important question not addressed in Tables 9.2 to 9.5 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 $\hat{Y}_{CAL}(\mathbf{x}_{1k})$ is smaller than that of $\hat{Y}_{CAL}(\mathbf{x}_{2k})$, 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 $\mathbf{x}_{2k}$. Sample size plays a role in this. Because $H(\mathbf{x}_k)$ has sampling variability, its success rate (the rate of correct indication) will depend on the sample size, and we expect it to increase with sample size.

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 specified $\mathbf{x}$-vectors. Several pairwise comparisons of this kind were carried out. A few typical results are shown in Table 9.6, for the response distribution 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 relative bias, 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 9.6, 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 is generally higher when the relative bias values are further apart. The case 4G+8G vs. 8G+8G compares nested $\mathbf{x}$-vectors, so it is known beforehand that H_2 and H_3 give perfect success rates.

Table 9.6. 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
NG+8G (4.1) vs.	74.3	70.3	45.0
8G+NG (3.4)	82.8	78.0	43.3
4G+NG (4.1) vs.	90.6	61.4	83.9
2G+2G (3.2)	97.0	68.8	92.3
NG+2G (7.3) vs.	77.4	77.4	34.5
2G+NG (6.5)	85.9	85.9	28.8

10 Selection of auxiliary variables in the Swedish pilot survey on gaming and problem gambling

The experiment with the constructed population in Section 9 corroborates some of the theory in earlier sections. By construction, that population has regularity features ordinarily not present in real survey data, which typically contains outliers and other oddities. Therefore we selected a set of real survey data to illustrate the use of the indicators H_1, H_2 and H_3 in building the $\mathbf{x}$-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 = 2000 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 that are 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 to some degree 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 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 (that is, $k \in r$) 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 value of the study variable, $y_k = 1$ if k is *employed* and $y_k = 0$ otherwise, is known for $k \in s$, so the unbiased estimate $\widetilde{Y}_{FUL}$ defined by (3.2) can be computed and used as a reference. We also computed $\widetilde{Y}_{EXP}$ defined by (3.1), as well as $\widetilde{Y}_{CAL}$ defined by (2.5) for different $\mathbf{x}$-vectors built by stepwise selection from the pool of 12 *x*-variables. The selection was done with each of the indicators H_1, H_2 and H_3 defined by (5.12).

We carried out the forward selection as follows: The auxiliary vector in Step 0 is the trivial $\mathbf{x}_k = 1$, and the estimator is $\widetilde{Y}_{EXP}$. 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 will be different for each indicator.

The values of H_2 and H_3 that identify the next variable for inclusion are necessarily increasing in every step. Important to note is that this does not hold for H_1. In a certain step *j*, the *x*-variable with the largest of the computed H_1-values is included, but 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 10.1 illustrates.

The unbiased estimate is $\widetilde{Y}_{FUL} = 4265$; the primitive estimate is $\widetilde{Y}_{EXP} = 4719$ (both in thousands). This suggests a large positive bias in $\widetilde{Y}_{EXP}$, whose relative deviation (in %) from $\widetilde{Y}_{FUL}$ is $RDF = (\widetilde{Y}_{EXP} - \widetilde{Y}_{FUL}) / \widetilde{Y}_{FUL} \times 10^2 = 10.7$. Admitting *x*-variables one by one into the **x**-vector will successively change this deviation, although not always to a smaller value. In each step we computed the indicator, $\widetilde{Y}_{CAL}$ and $RDF = (\widetilde{Y}_{CAL} - \widetilde{Y}_{FUL}) / \widetilde{Y}_{FUL} \times 10^2$. Tables 10.1 and 10.3 show step by step results for H_1 and H_3. The number of categories is given in parenthesis for each selected variable.

Table 10.1 shows the stepwise selection with indicator H_1. 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 10.1: The value of H_1 then starts to decline, and $\widetilde{Y}_{CAL}$ 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 10.2 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 the estimate $\widetilde{Y}_{CAL}$ 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 = 0.6 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 the final value 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.

Table 10.1. Stepwise forward selection, indicator H_1, dichotomous study variable *Employed*. Successive values of $H_1 \times 10^3$, of $\widetilde{Y}_{CAL}$ in thousands, and of $RDF = (\widetilde{Y}_{CAL} - \widetilde{Y}_{FUL})/\widetilde{Y}_{FUL} \times 10^2$. For comparison, $\widetilde{Y}_{EXP} \times 10^{-3} = 4719$; $\widetilde{Y}_{FUL} \times 10^{-3} = 4265$

Auxiliary variable entered	$H_1 \times 10^3$	$\widetilde{Y}_{CAL} \times 10^{-3}$	*RDF*
Income class (3)	76	4458	4.5
Education level (3)	107	4350	2.0
Presence of children (2)	114	4326	1.4
Urban centre dwelling (2)	118	4310	1.1
Sex (2)	123	4296	0.7
Marital status (2)	125	4286	0.5
Days unemployed (3)	121	4301	0.9
Months with sickness benefits (3)	120	4305	1.0
Level of debt (3)	115	4322	1.3
Cluster of postal codes (6)	109	4343	1.8
Country of birth (2)	103	4363	2.3
Age class (4)	99	4377	2.6

Table 10.2. Stepwise forward selection, indicator H_3, dichotomous study variable *Employed*. Successive values of $H_3 \times 10^3$, of $\widetilde{Y}_{CAL}$ in thousands, of $RDF = (\widetilde{Y}_{CAL} - \widetilde{Y}_{FUL})/\widetilde{Y}_{FUL} \times 10^2$. For comparison, $\widetilde{Y}_{EXP} \times 10^{-3} = 4719$; $\widetilde{Y}_{FUL} \times 10^{-3} = 4265$

Auxiliary variable entered	$H_3 \times 10^3$	$\widetilde{Y}_{CAL} \times 10^3$	*RDF*
Education level (3)	186	4520	6.0
Cluster of postcode areas (6)	250	4505	5.6
Country of birth (2)	281	4498	5.5
Income class (3)	298	4369	2.4
Age class (4)	354	4399	3.1
Sex (2)	364	4384	2.8
Urban centre dwelling (2)	374	4378	2.6
Level of debt (3)	381	4364	2.3
Months with sickness benefits (3)	384	4380	2.7
Presence of children (2)	387	4379	2.7
Marital status (2)	388	4379	2.7
Days unemployed (3)	388	4377	2.6

11 Concluding remarks

In this article, we discuss a survey situation where alternative auxiliary vectors ($\mathbf{x}$ -vectors) can be created and considered for use in the calibration estimator $\hat{Y}_{CAL}$ of the population total $Y = \sum_U y_k$.

For every specified $\mathbf{x}$ -vector, a certain bias remains in $\hat{Y}_{CAL}$. This bias is a function of the choice of $\mathbf{x}$ -vector. Our analysis is based on the bias ratio defined by (4.2) and (4.3). We have shown alternative ways, (5.8) to (5.10), to factorize the bias ratio in terms of simple, easily interpreted statistical measures.

To select "the best one" out of a number of available $\mathbf{x}$ -vectors, we can use the indicator H_1 given by (5.12). If we focus on minimizing bias for a fixed study variable (y -variable), we are led to maximize H_1 .

However, a typical government survey has many study variables. For practical reasons, it may be desirable to use the same $\mathbf{x}$ -vector for estimating all y -variable totals. A compromise is necessary. We have argued that the indicator H_3 in (5.12) is then a suitable one; this statistical measure does not depend on any y -data. Better indicators for the "many y -variable situation" can perhaps be developed; this is a topic for further research.

Another topic for further work is to examine alternative algorithms for stepwise selection of *x*-variables with the indicator H_1 , other than the one used in Section 10.

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 Journal 22, 43-53.

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

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.

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.

Appendix

Steps in constructing the population of size 6,000 used in Section 9

The values (y_k, x_{1k}, x_{2k}) for $k = 1, 2, ..., 6{,}000$ were created in three steps as follows:

Step 1: The continuous auxiliary variable x_1. The 6,000 values x_{1k} were created as independent outcomes of the gamma distributed random variable $\Gamma(a, b)$ with parameter values $a = 2$, $b = 5$. The theoretical mean and variance are $\mu_{x_1} = a\,b = 10$ and $\sigma^2_{x_1} = a\,b^2 = 50$ respectively. The mean and variance of the 6,000 realized values x_{1k} was 10.0 and 49.9, respectively.

Step 2: The continuous auxiliary variable x_2. For unit k, with the value x_{1k} fixed by Step 1, a value x_{2k} is realized as an outcome of the gamma random variable $\Gamma(A_k, B_k)$, with parameters $A_k = (\mu_{x_{2k}|x_{1k}})^2 / \sigma^2_{x_{2k}|x_{1k}}$ and $B_k = \sigma^2_{x_{2k}|x_{1k}} / \mu_{x_{2k}|x_{1k}}$, such that

$$\mu_{x_{2k}|x_{1k}} = A_k B_k = \alpha + \beta\, x_{1k} + K\, h(x_{1k}) \,;\, \sigma^2_{x_{2k}|x_{1k}} = A_k B_k^2 = \sigma^2 x_{1k}$$

with $h(x_{1k}) = x_{1k}(x_{1k} - \mu_{x_1})(x_{1k} - 3\mu_{x_1})$ where $\mu_{x_1} = 10$ Suitable values were assigned to the constants α, β, K and σ^2. The polynomial term $K\,h(x_{1k})$ gives a mild non-linear appearance to the plotted points (x_{2k}, x_{1k}). This was done in order to avoid a perfect linear relationship between x_1 and x_2. We used $\alpha = 1$, $\beta = 1$, $K = 0.001$, $\mu_{x_1} = 10$ and $\sigma^2 = 25$. 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 continuous 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 $\Gamma(a_k, b_k)$ with $a_k = (\mu_{y_k|x_{1k},x_{2k}})^2 / \sigma^2_{y_k|x_{1k},x_{2k}}$ and $b_k = \sigma^2_{y_k|x_{1k},x_{2k}} / \mu_{y_k|x_{1k},x_{2k}}$, such that

$$\mu_{y_k|x_{1k},x_{2k}} = a_k b_k = c_0 + c_1 x_{1k} + c_2 x_{2k} \ ;$$
$$\sigma^2_{y_k|x_{1k},x_{2k}} = a_k b_k^2 = \sigma_0^2 (c_1 x_{1k} + c_2 x_{2k})$$

It follows that the conditional expectation of y_k given x_{1k} is $c_0 + c_1 x_{1k} + c_2(\alpha + \beta x_{1k} + K h(x_{1k}))$. We used $c_0 = 1$, $c_1 = 0.7$, $c_2 = 0.3$ and $\sigma_0^2 = 2$. The values of α, β, K and σ^2 are fixed by Step 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. The correlation coefficient between y and x_2, computed on the 6,000 couples (y_k, x_{2k}), was 0.73.

ISSN 1653-7149

All officiell statistik finns på: **www.scb.se**
Kundservice: tfn 08-506 948 01

All official statistics can be found at: **www.scb.se**
Customer service, phone +46 8 506 948 01