Design for estimation: Identifying auxiliary vectors to reduce nonresponse bias

Carl-Erik Särndal
Sixten Lundström

ISSN 1653-7149
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
Design for estimation: Identifying auxiliary vectors to reduce nonresponse bias

Carl-Erik Särndal
Sixten Lundström

Statistiska centralbyrån
2009
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
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

Preface .................................................................................................................. 3
1 Introduction ......................................................................................................... 7
2 Calibration estimators for a survey with nonresponse ............................... 9
3 Points of reference .......................................................................................... 13
4 The bias ratio .................................................................................................... 15
5 Expressing the deviation accounted for ......................................................... 19
6 Preference ranking of auxiliary vectors .......................................................... 23
7 Derivations ......................................................................................................... 25
8 Two remarks ...................................................................................................... 29
9 Empirical validation with a constructed population .................................... 33
10 Selection of auxiliary variables in the Swedish pilot survey on gaming and problem gambling ................................................................. 45
11 Concluding remarks ....................................................................................... 51
References ............................................................................................................. 53
Appendix .................................................................................................................. 55
Steps in constructing the population of size 6,000 used in Section 9 ... 55
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 $x$-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 \( U = \{1, 2, \ldots, k, \ldots, N\} \). 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

\[
P = \frac{\sum r_k d_k}{\sum s_k d_k}
\]

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, \( A \subset U \), 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, \( x_k^* \) and \( x_k^o \). Population auxiliary information is transmitted by \( x_k^* \), a vector value known for every \( k \in U \). Thus \( \sum_U x_k^* \) is a known population total. Alternatively, we allow that \( \sum_U x_k^* \) is imported from an exterior source and that \( x_k^* \) is a known (observed) vector value for every \( k \in s \). Sample auxiliary information is transmitted by \( x_k^o \), a vector value known (observed) for every \( k \in s \); the total \( \sum_U x_k^o \) is unknown but is estimated without bias by \( \sum_s d_k x_k^o \). The auxiliary vector value combining the two types is denoted \( x_k \). This vector and the associated information is
\[ \mathbf{x}_k = \begin{pmatrix} x_k^* \\ x_k^0 \end{pmatrix}; \quad \mathbf{X} = \begin{pmatrix} \sum_U x_k^* \\ \sum_s d_k x_k^* \end{pmatrix} \] (2.2)

Tied to the \( k \)th unit is the vector \((y_k, \mathbf{x}_k, \pi_k, \theta_k)\). Here, \( \pi_k \) is known for all \( k \in U \), \( y_k \) for all \( k \in r \), the component \( x_k^* \) of \( \mathbf{x}_k \) carries population information, the component \( x_k^0 \) 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”, \( \theta_k \) 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 \( \mathbf{\mu} \) such that

\[ \mathbf{\mu}'\mathbf{x}_k = 1 \text{ for all } k \in U \] (2.3)

“Constant” means that \( \mathbf{\mu} \neq \mathbf{0} \) 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,

\[ \mathbf{x}_k = \gamma_k = (\gamma_{jk1}, \ldots, \gamma_{jkJ}), \text{ where } \gamma_{jk} = 1 \text{ if } k \text{ belongs to group } j, \text{ and } \gamma_{jk} = 0 \text{ if not}, j = 1, 2, \ldots, J; \] (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 \( k = x_k \), not covered by (2.3), leads to the nonresponse ratio estimator, known to be a usually poor choice for controlling nonresponse bias, compared with \( x_k = (1, x_k)' \), so excluding the ratio estimator is no great loss.)

The calibration estimator of \( Y = \sum_{i \in r} y_k \), computed on the data \( y_k \) for \( k \in r \), is

\[
\hat{Y}_{Cal} = \sum_{i} w_k y_k
\]

with \( w_k = d_k \{1 + (X - \sum_{i} d_k x_k)'(\sum_{i} d_k x_k x_k')^{-1} x_k\} \). The weights \( w_k \) are calibrated on both kinds of information: \( \sum_{i} w_k x_k = X \), which implies \( \sum_{i} w_k x_k^* = \sum_{i} x_k^* \) and \( \sum_{i} w_k x_k^o = \sum_{i} d_k x_k^o \). We assume throughout that the symmetric matrix \( \sum_{i} d_k x_k x_k' \) 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 \( \hat{Y}_{Cal} = \sum_{i} w_k y_k \) with weights \( w_k = d_k v_k \) where \( v_k = X'(\sum_{i} d_k x_k x_k')^{-1} x_k \). The weights satisfy \( \sum_{i} d_k v_k x_k = X \), where \( X \) 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 \( x_k \) but with calibration only to the sample level:

\[
\tilde{Y}_{Cal} = \sum_{i} d_k m_k y_k
\]

where

\[
m_k = (\sum_{i} d_k x_k)'(\sum_{i} d_k x_k x_k')^{-1} x_k
\]

The calibration equation then reads \( \sum_{i} d_k m_k x_k = \sum_{i} d_k x_k \), where \( x_k \) has the two components as in (2.2). The auxiliary vector \( x_k \) 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 \( \tilde{Y}_{CAL} \) because the former profits from the input of a known population total \( \sum_U x_k^* \). But this paper studies the bias. From that perspective, we are virtually indifferent between \( \tilde{Y}_{CAL} \) and \( \tilde{Y}_{CAL'} \), and we focus on the latter. Under liberal conditions, the difference between the bias of \( N^{-1}\tilde{Y}_{CAL} \) and that of \( N^{-1}\tilde{Y}_{CAL} \) 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 \( \tilde{Y}_{CAL} \) defined by (2.5) is

\[
\tilde{Y}_{CAL} = (\sum_k d_k x_k) B_x
\]

(2.7)

where

\[
B_x = B_{x_{r:d}} = (\sum_k d_k x_k x_k')^{-1} \sum_k d_k x_k y_k
\]

(2.8)

is the regression coefficient vector arising from the \( d_k \)-weighted least squares fit based on the data \( (y_k, x_k) \) 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 \( B_{x_{r:d}} \) 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, \( x_k = 1 \) 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 \( \tilde{Y}_{CAL} \) is the expansion estimator:

\[
\tilde{Y}_{EXP} = \frac{1}{P} \sum_k d_k y_k = \hat{N} \bar{y}_{r,d} \tag{3.1}
\]

where \( \hat{N} = \sum_k d_k \) is design unbiased for the population size \( N \). The bias of \( \tilde{Y}_{EXP} \) 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 \( \sum_U x_k^* \),

\[
\hat{Y}_{FUL} = \sum_k d_k g_k y_k
\]

where \( g_k = 1 + (\sum_U x_k^* - \sum_i d_i x_i^*)' (\sum_k d_k x_k^* x_k^*)^{-1} x_k^* \), and \( FUL \) refers to full response. The unbiased HT estimator (obtained when \( g_k = 1 \) for all \( k \)) is

\[
\tilde{Y}_{FUL} = \sum_k d_k y_k = \hat{N} \bar{y}_{r,d} \tag{3.2}
\]

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

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

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

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

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

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

The goal is to choose the auxiliary vector $x_k$ used in $\tilde{Y}_{CAL}$ so that the bias ratio is small. We scale the three deviations in (4.1) by the estimated population size $\hat{N} = \sum_k d_k$ and use the notation
The bias ratio $\Delta_T = \Delta_A + \Delta_R$, where $T$ suggests “total”, $A$ “accounted for” and $R$ “remaining”. Noting that $\sum d_k(y_k - x_k'B_s) = 0$, we have

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

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

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

where $\bar{x}_{s,d} = \sum d_kx_k / \sum d_k$, $\bar{x}_{r,d} = \sum d_kx_k / \sum d_k$, 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) $\Delta_T$ cannot be computed; it depends on unobserved (as well as observed) $y_k$-values, but not on any $x_k$-values; (ii) $\Delta_A$ is computable; it depends on $y_k$ for $k \in r$ and on $x_k$ for $k \in s$; the choice of $x$-vector determines the value of $\Delta_A$; (iii) $\Delta_R$ cannot be computed; it depends on unobserved $y_k$, and on $x_k$ 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 $x$-vector. It 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. Then $\Delta_A$ and $\Delta_R$ are usually of the same positive sign, although not necessarily so for all choices of $x$-vector. It can happen that $\Delta_A > \Delta_T$ so that $\Delta_R < 0$. When the $x$-vector used for $\tilde{Y}_{\text{CAL}}$ becomes progressively more powerful, $\Delta_A$ tends to come near $\Delta_T$, leaving a small $\Delta_R$. If $\Delta_T < 0$, these tendencies are reversed.

The bias ratio (4.2) takes the form

$$\text{bias ratio} = \frac{\Delta_R}{\Delta_T} = 1 - \frac{\Delta_A}{\Delta_T} = 1 - \frac{(\bar{x}_{r,d} - \bar{x}_{s,d})'B_s}{\bar{y}_{r,d} - \bar{y}_{s,d}} \quad \text{(4.3)}$$
We have \( \text{bias ratio} = 1 \) for the primitive vector \( \mathbf{x}_k = 1 \). However, bias ratio \( \approx 0 \) 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 \( \Delta_A = (\mathbf{x}_{r,d} - \mathbf{x}_{x,d})' \mathbf{B}_x \). This is within our reach. Whatever our final choice, the remaining bias of \( \tilde{\mathbf{Y}}_{CAL} \) 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 \( y_k = \mathbf{\beta}' \mathbf{x}_k + \mathbf{e}_k \), where \( \mathbf{e}_k \) is a small residual. Then \( \mathbf{y}_{r,d} - \mathbf{y}_{x,d} \approx (\mathbf{x}_{r,d} - \mathbf{x}_{x,d})' \mathbf{B}_x \approx (\mathbf{x}_{r,d} - \mathbf{x}_{x,d})' \mathbf{\beta} \), and consequently bias ratio \( \approx 0 \), 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 \( \mathbf{B}_x \) to be biased for a regression vector that may perfectly well describe a \( y \)-to-\( x \) 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 $\tilde{Y}_{CAL} = \sum_d d_k m_k y_k$. The factor $m_k = (\sum_d d_k x_k' y_k')(\sum_d d_k x_k x_k')^{-1} x_k$ 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 $x_k$, independent of all $y$-variables of interest, and computable for $k \in s$ (but used in $\tilde{Y}_{CAL}$ only for $k \in r$).

We have
\[
\sum_r d_k m_k x_k = \sum_r d_k x_k ; \quad \sum_r d_k m_k = \sum_r d_k ;
\]
\[
\sum_r d_k m_k^2 = \sum_r d_k m_k
\]
(5.1)

Two weighted means are needed:
\[
\bar{m}_{r,d} = \frac{\sum_r d_k m_k}{\sum_r d_k} = \sum_r d_k = \frac{1}{P} ; \quad \bar{m}_{s,d} = \frac{\sum_r d_k m_k}{\sum_r d_k}
\]
(5.2)

where $P$ is the response rate (2.1). Thus the average adjustment factor in $\tilde{Y}_{CAL} = \sum_d d_k m_k y_k$ 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
\[
S_m^2 = S_{m|r,d}^2 = \frac{1}{\sum_d d_k} \sum_r d_k (m_k - \bar{m}_{r,d})^2
\]
(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
\[
S_m^2 = \bar{m}_{r,d} (\bar{m}_{s,d} - \bar{m}_{r,d})
\]
(5.4)

The coefficient of variation of the $m_k$ is
Expressing the deviation accounted for Design for estimation

$$cv_m = \frac{S_m}{m_{r,d}} = \sqrt{\frac{m_{r,d}}{m_{r,d}} - 1} \quad (5.5)$$

The weighted variance of the study variable $y$ is given by

$$S^2_y = S^2_{y|r,d} = \frac{\sum_i d_i (y_i - \bar{y}_{r,d})^2}{\sum_i d_i} \quad (5.6)$$

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

$$\text{Cov}(y, m) = \text{Cov}(y, m)_{r,d} = \frac{1}{\sum_i d_i} \sum_i d_i (m_i - \bar{m}_{r,d})(y_i - \bar{y}_{r,d}) \quad (5.7)$$

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

The deviation $\Delta_d = (x_{r,d} - \bar{x}_{r,d})'B_k$ is a crucial factor in the bias ratio (4.3). We seek an $x$-vector that makes $\Delta_d$ 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 $\Delta_d/S_y$ as

$$\Delta_d/S_y = -R_{y,m} \times cv_m \quad (5.8)$$

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

Another simple representation of (5.8) is

$$\Delta_d/S_y = F \times R_{y,x} \times cv_m \quad (5.9)$$

where $R_{y,x} = \sqrt{R^2_{y,x}}$ is the coefficient of multiple correlation between $y$ and $x$, $R^2_{y,x}$ is the proportion of the $y$-variance $S^2_y$ explained by the predictor $x$, and $F = -R_{y,m}/R_{y,x}$ (Formula (7.8) states the
Design for estimation

Expressing the deviation accounted for

precise expression for $R^2_{y,x}$.) As Section 7 also shows, $|R_{y,m}| \leq R_{y,x}$ for any $x$-vector and $y$-variable; consequently $-1 \leq F \leq 1$.

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

$$\left| \Delta_A \right| / S_y = |R_{y,m}| \times cv_m = |F| \times R_{y,x} \times cv_m$$

(5.10)

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

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

$$\left| \Delta_A \right| / S_y \approx |F_0| \times R_{y,x} \times cv_m$$

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

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

(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 $x$-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):

$$H_1 = \left| \Delta_A \right| / S_y = |R_{y,m}| \times cv_m ; \quad H_2 = R_{y,x} \times cv_m ; \quad H_3 = cv_m$$

(5.12)
Our main alternatives are $H_1$ and $H_3$. Of these, $H_1$ is directly linked to $\Delta_A$, 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 $R_{y,x}$, it is less appropriate than $H_1$ because the correlation coefficient ratio $F = -R_{y,m}/R_{y,x}$ may vary considerably from one $x$-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 x-vectors can be composed. The indicators in (5.12) provide computational tools for obtaining a preference ordering, or a ranking, of potential 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 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 \( H_1 = \frac{\Delta_y}{S_y} = \frac{R_{y,jn}}{cV_{jn}} \) and choose the x-vector to make its value as large as possible. An ad hoc alternative is to use the indicator \( H_2 = R_{y,x} \times cV_{jn} \), 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 to use \( H_3 = cV_{jn} \) as a compromise indicator, and to choose the 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 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:
Preference ranking of auxiliary vectors  

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 \( x \)-vector to make the computable numerator \( \Delta_d = (\bar{x}_{r,d} - \bar{x}_{r,d})' B_x \) in the bias ratio (4.3) as large as possible, in absolute value. In this section we prove the factorizations \( \Delta_d/S_y = -R_{y,m} \times cv_m = F \times R_{y,x} \times cv_m \) in (5.8) and (5.9). First we express \( cv_m^2 \) as a quadratic form in the vector that contrasts the \( x \)-mean in the response set \( r \) with the \( x \)-mean in the sample \( s \). Define

\[
D = \bar{x}_{r,d} - \bar{x}_{s,d} ; \quad \Sigma = \sum_k d_k x_k x'_k / \sum_k d_k \quad (7.1)
\]

Then, with \( P \) given by (2.1),

\[
(cv_m^2 = P^2 \times S_m^2 = D' \Sigma^{-1} D \quad (7.2)
\]

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

\[
\bar{x}_{r,d} \Sigma^{-1} x_{r,d} = \bar{x}_{r,d} \Sigma^{-1} x_{s,d} = 1 \quad (7.3)
\]

Next we define the covariance vector as

\[
C = (\sum_k d_k (x_k - \bar{x}_{r,d})(y_k - \bar{y}_{r,d}) / (\sum_k d_k) \quad (7.4)
\]

whereby we can write \( \Delta_d \) as a bilinear form:

\[
\Delta_d = D' B_x = D' \Sigma^{-1} C \quad (7.5)
\]

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

A useful perspective on \( \Delta_d \) is gained from the geometric interpretation of \( C \) and \( D \) in (7.5) as vectors in the space whose dimension is that of \( x_k \). We have

\[
\Delta_d = \Lambda (D' \Sigma^{-1} D)^{1/2} (C' \Sigma^{-1} C)^{1/2} \quad (7.6)
\]

with

\[
\Lambda = \frac{D' \Sigma^{-1} C}{(D' \Sigma^{-1} D)^{1/2} (C' \Sigma^{-1} C)^{1/2}} \quad (7.7)
\]
For a specific $y$-variable and a specific $x$-vector, the scalar quantities $(D'\Sigma^{-1}D)^{1/2}$ and $(C'\Sigma^{-1}C)^{1/2}$ represent the respective vector lengths of $D$ and $C$ (following an orthogonal transformation based on the eigenvectors and eigenvalues of $\Sigma^{-1}$). The scalar quantity $\Lambda$ represents the cosine of the angle between $D$ (which is independent of $y$) and $C$ (which depends on $y$); hence $-1 \leq \Lambda \leq 1$.

When the auxiliary vector $x_k$ is allowed to expand, by adding further available $x$-variables, both vector lengths $(D'\Sigma^{-1}D)^{1/2}$ and $(C'\Sigma^{-1}C)^{1/2}$ increase. The angle $\Lambda$ will ordinarily change, but if $|\Lambda|$ stays roughly constant, (7.6) shows that $|\Delta_d|$ will increase.

A second useful perspective on $\Delta_d$ follows by decomposing the total variability of the study variable $y$,

\[ \sum d_k (y_k - \bar{y}_d)^2 = (\sum d_k) S_y^2. \]

Two regression fits need to be examined, the one of $y$ on the auxiliary vector $x$, 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 $D$, $\Sigma$ and $C$ 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 $x$ is

\[ R_{y,x}^2 = (C'\Sigma^{-1}C) / S_y^2 \]  

(7.8)

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

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

(7.9)

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

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

(7.10)
The proportions $R^2_{y,x}$ and $R^2_{y,m}$ satisfy $R^2_{y,m} \leq R^2_{y,x} \leq 1$.

**Proof.** The proof of (7.8) uses the weighted least squares regression of $y$ on $x$ fitted over $r$. The residuals are $y_k - \hat{y}(x)_k$, where

$$\hat{y}(x)_k = x'_k B_x$$

with $B_x$ given by (2.8). The decomposition is

$$\sum_k d_k (y_k - \bar{y}_{r,ld})^2 = \sum_k d_k (\hat{y}(x)_k - \bar{y}_{r,ld})^2 + \sum_k d_k (y_k - \hat{y}(x)_k)^2$$

The mixed term is zero. A development of the term “variation explained” gives

$$\sum_k d_k (\hat{y}(x)_k - \bar{y}_{r,ld})^2 = (\sum_k d_k) C^T \Sigma^{-1} C .$$

Thus the proportion of variance explained is

$$R^2_{y,x} = \frac{\sum_k d_k (\hat{y}(x)_k - \bar{y}_{r,ld})^2}{[(\sum_k d_k) S_y^2]} = C^T \Sigma^{-1} C / S_y^2 ,$$

as claimed in (7.8). To show (7.9) we note that

$$\text{Cov}(y, m) = -\Delta A / P = -D^T \Sigma^{-1} C / P .$$

It then follows from (7.2) that

$$R^2_{y,m} = \text{Cov}(y, m) / (S_y S_m)$$

has the expression (7.9). The residuals from the regression (with intercept) of $y$ on the univariate explanatory variable $m$ are

$$\hat{y}(m)_k = \bar{y}_{r,ld} + B_m (m_k - m_{r,ld})$$

with $B_m = \text{Cov}(y, m) / S_m^2 = -P (D^T \Sigma^{-1} C) / (D^T \Sigma^{-1} D)$. The proportion of variance explained is

$$\sum_k d_k (\hat{y}(m)_k - \bar{y}_{r,ld})^2 / [(\sum_k d_k) S_y^2],$$

which upon development gives the expression for $R^2_{y,m}$ in (7.10). Finally, $R^2_{y,m} \leq R^2_{y,x}$ follows from the Cauchy-Schwarz inequality for a bilinear form:

$$(D^T \Sigma^{-1} C)^2 \leq (D^T \Sigma^{-1} D)(C^T \Sigma^{-1} C) .$$

The fact that $R^2_{y,m} \leq R^2_{y,x} \leq 1$ illustrates that, among all predictions $\hat{y}_k = x'_k \beta$ that are linear in the $x$-vector, those that maximize the variance explained are $\hat{y}(x)_k = x'_k B_x$, so the alternative predictions $\hat{y}(m)_k$, which are linear in $x_k$ via $m_k$, cannot yield a greater variance explained than that maximum.

Now from (7.9), (7.2) and (7.5),

$$-R_{y,m} cv_m = D^T \Sigma^{-1} C / S_y = \Delta A / S_y ,$$

as claimed by formula (5.8). Moreover, (7.8), (7.9) and (7.7) imply
\[-R_{y,n}/R_{y,x} = \Lambda, \text{ so the correlation coefficient ratio } F \text{ in (5.9) equals the angle } \Lambda \text{ defined by (7.7)}\]
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 \( \Delta_d = \hat{N}^{-1}(\hat{Y}_{\text{Exp}} - \hat{Y}_{\text{CAL}}) \) and the bias of \( \hat{Y}_{\text{CAL}} \) or \( \hat{Y}_{\text{CAL}} \).

The total deviation defined in Section 4 is \( \Delta_T = \Delta_y + \Delta_R \), where \( \Delta_y \) is computable but \( \Delta_R \) and \( \Delta_T \) are not. If computable, \( \hat{N}_R = \hat{Y}_{\text{CAL}} - \hat{Y}_{\text{FUL}} \) would have been an estimate of the bias of \( \hat{Y}_{\text{CAL}} \) (and of that of \( \hat{Y}_{\text{CAL}} \)). How is “small bias” (a small value of \( R \)) related to the goodness of fit of the model \( y_k = \beta' x_k + \epsilon_k ? \) The residuals from that fit determine the value of \( \Delta_R = \sum_{r \in \mathbb{R}_d} B_{x, d} y_k - \hat{y}_k \), where \( B_{x, d} \) is given by (2.8). (In this section the more precise notation \( B_{x, d} \) is preferable to the simpler \( B_x \) used earlier.) Two aspects of the fit of that model are: (i) The computable fit to the data \( (y_k, x_k) \) observed for \( k \in r \); (ii) The hypothetical fit to the data \( (y_k, x_k) \) for \( k \in s \), some observed, some not. Let us consider the two cases.

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

\[
\Delta_R = -(1 - P) \bar{K}_{s-r, d}
\]  

(8.1)

The expression (8.1), which may be far from zero, does not depend on the residuals \( e_{x, d} \). Regardless of whether the fit is good (residuals small; \( R_{y, s} \) near one) or poor (residuals large; \( R_{y, s} \) near...
zero), the remaining deviation $\Delta_R$ is still given by (8.1), and the
term $\Delta_A$ remains unchanged at $\Delta_A = D'B_{x|d} = D'\Sigma^{-1}C$. Even if
the fit is perfect for the respondents, so that $e_{k|d} = 0$ for all $k \in r$
and $R^2_{y,s} = 1$, there is no indication that the bias is small. Given the
respondent data $(y_k, x_k)$, $k \in r$, the remaining deviation $\Delta_R$ is
unchanged, given by (8.1).

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

$$\hat{Y}_{imp} = \sum_d d_k y_k + \sum_{d \neq r} d_k \hat{y}_k$$

As is easily verified, $\hat{Y}_{imp} = \hat{Y}_{CAL}$, so $\hat{Y}_{imp}$ has the same exposure to
bias as $\hat{Y}_{CAL}$. 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 $(y_k, x_k)$ for
$k \in s$ is hypothetical, because $y_k$ is missing for $k \in s - r$. The
hypothetical regression coefficient vector is

$$B_{x|d} = (\sum_s d_k x_k x_k')^{-1} \sum_s d_k x_k y_k,$$

and the residuals

$$e_{k|d} = y_k - \hat{x}_k' B_{x|d}$$

for $k \in s$ would have the property

$$\sum s d_k e_{k|d} = 0.$$ Using that

$$\sum s d_k m_k y_k / \hat{N} = \bar{x}_{x|d}$$

and

$$\sum s d_k m_k y_k / \hat{N} = \bar{x}_{x|d} B_{x|d}$$,

we have

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

Suppose the model is “true for the sample $s$”, with a perfect fit, so
that $e_{k|d} = 0$ for all $k \in s$. Then, by (8.2) we do have

$$\Delta_R = \hat{N}^{-1}(\hat{Y}_{CAL} - \hat{Y}_{FUL}) = 0.$$ A belief that the bias is small hinges on
an unverifiable assumption.
The second issue concerns the relation between the bias of $\hat{Y}_{\text{CAL}}$ and the expected value of the indicator $H_0 = \Delta_{\Delta}/S_y$

$= (\bar{Y}_{\text{EXP}} - \bar{Y}_{\text{CAL}})/\hat{N}S_y = -R_{y;\eta} \times \text{cv}_m$. For fixed $y$ -variable and $x$ -vector we have

$\frac{(\bar{Y}_{\text{CAL}} - Y)}{\hat{N}S_y} = \frac{(\bar{Y}_{\text{EXP}} - Y)}{\hat{N}S_y} - H_0$

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

$\text{bias}(\bar{Y}_{\text{CAL}}) \approx \text{bias}(\bar{Y}_{\text{EXP}}) - C \times E(H_0)$ \hspace{1cm} (8.3)

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

$\text{bias}(\bar{Y}_{\text{CAL}}(\mathbf{x}_{1k})) - \text{bias}(\bar{Y}_{\text{CAL}}(\mathbf{x}_{2k})) \approx -C \left( E_{1} - E_{2} \right)$

where $E_{i} = E(H_0(\mathbf{x}_{ik}))$ 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 \( \hat{Y}_{\text{CAL}} \). 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 \( H_0 = \Delta_{s} / S_{y} = -R_{y,m} \times cv_{m} \) tracks the bias of the calibration estimator, as measured by the average of \( \hat{Y}_{\text{CAL}} - Y \), 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 \( \hat{Y}_{\text{CAL}} \).

For this study we created values \((y_{k}, \mathbf{x}_{k}, \theta_{k})\) for \( k = 1, 2, \ldots, N = 6,000 \). 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 \( \theta_{k} \) independent of \( s \). The steps in creating \((y_{k}, \mathbf{x}_{k}, x_{2k})\) for \( k = 1, 2, \ldots, 6,000 \) 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 \( 4 \times 4 = 16 \) 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 \( \gamma_{(x_1,8)k} \) of dimension eight with seven entries “0” and a single entry “1” to code the group membership of \( k \). For example, \( \gamma_{(x_1,8)k} = (0,0,0,1,0,0,0)' \) 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 \( \gamma_{(x_1,4)k} \). In mode 2G, unit \( k \) has the indicator vector \( \gamma_{(x_1,2)k} \) with value \( \gamma_{(x_1,2)k} = (1,0)' \) for the 3,000 largest \( x_1 \)-value units and \( \gamma_{(x_1,2)k} = (0,1)' \) 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 \( \gamma_{(x_1,1)k} = 1 \) 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 \( \gamma_{(x_2,8)k} \), \( \gamma_{(x_2,4)k} \), \( \gamma_{(x_2,2)k} \) and \( \gamma_{(x_2,1)k} = 1 \). Finally, \( 4 \times 4 = 16 \) different auxiliary vectors \( x_k \) are formed by combining the two kinds of group information; the 16 cells and their notation are shown in the following display.

<table>
<thead>
<tr>
<th>Groups based on ( x_{1k} )</th>
<th></th>
<th>Groups based on ( x_{2k} )</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Eight</td>
<td>Four</td>
<td>Two</td>
<td>None</td>
</tr>
<tr>
<td>Eight</td>
<td>8G+8G</td>
<td>8G+4G</td>
<td>8G+2G</td>
</tr>
<tr>
<td>Four</td>
<td>4G+8G</td>
<td>4G+4G</td>
<td>4G+2G</td>
</tr>
<tr>
<td>Two</td>
<td>2G+8G</td>
<td>2G+4G</td>
<td>2G+2G</td>
</tr>
<tr>
<td>None</td>
<td>NG+8G</td>
<td>NG+4G</td>
<td>NG+2G</td>
</tr>
</tbody>
</table>

The “+” indicates that the \( x_k \)-vector has the two \( \gamma \)-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 \( x_k = (\gamma'_{(x_2,8)k};\gamma'_{(x_2,8)k})_{(-1)}' \), where
(-1) indicates that one category is excluded in either $\gamma_{(x;8)k}$ or $\gamma_{(x';8)k}$ to avoid a singular matrix, giving $x_k$ 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 $x_k = 1$ for all $k$. There are 14 intermediate cases. For example, the cell $4G+2G$ has $x_k = (\gamma'_{(x;4)k}, \gamma'_{(x';2)k})'(\cdot -1)$ of dimension $4+2-1 = 5$; the cell $4G+NG$ has $x_k = (\gamma'_{(x;4)k}, 1)'(\cdot -1) = \gamma_{(x;4)k}$ 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 $x$-vectors that ignore that interaction.)

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

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

(ii) IncExp($10+y$), with $\theta_k = 1 - e^{-c(10+y_2)}$ where $c = 0.06217$

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

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

The constant $c$ was adjusted in each option to give a mean response probability of $\bar{\theta}_U = \sum_k \theta_k / N = 0.70$. In (i) and (ii), the value 10 (rather than 0) was used to avoid a high incidence of small response probabilities $\theta_k$. 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 $\theta_k$ 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 \( x \)-vectors, and for every outcome \((s,r)\), we computed the relative deviation

\[
RD = (\hat{Y}_{\text{CAL}} - Y) / Y ,
\]

where \( \hat{Y}_{\text{CAL}} \) is given by (2.4) and \( Y = \sum y_k \) 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 then computed as

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

where the index \( j \) serves to indicate the computed value for the \( j \)th outcome, \( j = 1,2,\ldots, 5,000 = J \). 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 \( \hat{Y}_{\text{CAL}}, (E(\hat{Y}_{\text{CAL}}) - Y) / Y \), where expectation is jointly with respect to sampling design and response distribution.

**Table 9.1. Relbias in % and, within parenthesis, the value of \( Av(H_i) \times 10^3 \) for 16 auxiliary vectors \( x_k \). Response distribution IncExp(10+ x_1 + x_2).**

<table>
<thead>
<tr>
<th>Groups based on ( x_{1k} )</th>
<th>Eight</th>
<th>Four</th>
<th>Two</th>
<th>None</th>
</tr>
</thead>
<tbody>
<tr>
<td>Eight</td>
<td>0.2 (101)</td>
<td>0.5 (99)</td>
<td>1.3 (93)</td>
<td>3.4 (76)</td>
</tr>
<tr>
<td>Four</td>
<td>0.5 (98)</td>
<td>0.9 (96)</td>
<td>1.8 (89)</td>
<td>4.1 (70)</td>
</tr>
<tr>
<td>Two</td>
<td>1.5 (91)</td>
<td>1.9 (88)</td>
<td>3.2 (78)</td>
<td>6.5 (52)</td>
</tr>
<tr>
<td>None</td>
<td>4.1 (70)</td>
<td>5.0 (64)</td>
<td>7.3 (46)</td>
<td>13.2 (0)</td>
</tr>
</tbody>
</table>

To illustrate the layout of the experiment, Table 9.1 shows, for IncExp(10+ x_1 + x_2), relbias in % and \( Av(H_i) \times 10^3 \) for the 16 \( x \)-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 \( x_k = 1 \), 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 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 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 $\left|rancor\right| = 1$ for $Av(H_1)$, and $\left|rancor\right|$ 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 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 $\left|relbias\right|$ 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 x-vector.
Table 9.2. Value, in ascending order, of relbias in %, and corresponding value and rank of $A\nu(H_1) \times 10^3$, $A\nu(H_2) \times 10^3$ and $A\nu(H_3) \times 10^3$, for 16 auxiliary vectors. Bottom line: Value of Spearman rank correlation, $\text{rancor}$. Response distribution IncExp(10$^+$, $x_1 + x_2$)

<table>
<thead>
<tr>
<th>relbias</th>
<th>$A\nu(H_1) \times 10^3$</th>
<th>$A\nu(H_2) \times 10^3$</th>
<th>$A\nu(H_3) \times 10^3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.2</td>
<td>101 (1)</td>
<td>127 (1)</td>
<td>232 (1)</td>
</tr>
<tr>
<td>0.5</td>
<td>99 (2)</td>
<td>119 (2)</td>
<td>225 (2)</td>
</tr>
<tr>
<td>0.8</td>
<td>96 (4)</td>
<td>109 (4)</td>
<td>217 (4)</td>
</tr>
<tr>
<td>1.3</td>
<td>93 (5)</td>
<td>109 (5)</td>
<td>216 (5)</td>
</tr>
<tr>
<td>1.5</td>
<td>91 (6)</td>
<td>105 (6)</td>
<td>213 (6)</td>
</tr>
<tr>
<td>1.8</td>
<td>89 (7)</td>
<td>98 (7)</td>
<td>207 (7)</td>
</tr>
<tr>
<td>1.9</td>
<td>88 (8)</td>
<td>94 (8)</td>
<td>205 (8)</td>
</tr>
<tr>
<td>3.2</td>
<td>78 (9)</td>
<td>80 (10)</td>
<td>192 (9)</td>
</tr>
<tr>
<td>3.4</td>
<td>76 (10)</td>
<td>90 (11)</td>
<td>188 (11)</td>
</tr>
<tr>
<td>4.1</td>
<td>70 (11)</td>
<td>84 (9)</td>
<td>190 (10)</td>
</tr>
<tr>
<td>4.1</td>
<td>70 (12)</td>
<td>77 (12)</td>
<td>175 (13)</td>
</tr>
<tr>
<td>5.0</td>
<td>64 (13)</td>
<td>70 (13)</td>
<td>179 (12)</td>
</tr>
<tr>
<td>6.4</td>
<td>52 (14)</td>
<td>52 (14)</td>
<td>146 (15)</td>
</tr>
<tr>
<td>7.3</td>
<td>46 (15)</td>
<td>46 (15)</td>
<td>156 (14)</td>
</tr>
<tr>
<td>13.2</td>
<td>0 (16)</td>
<td>0 (16)</td>
<td>0 (16)</td>
</tr>
<tr>
<td>$\text{rancor}$</td>
<td>-1.00</td>
<td>-0.99</td>
<td>-0.99</td>
</tr>
</tbody>
</table>
Table 9.3. 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 correlation, rancor. Response distribution IncExp(10+ y).

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

rancor: -1.00, -0.99, -0.99
Table 9.4. Value, in ascending order, of $|\text{relbias}|$ 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 correlation, $\text{rancor}$. Response distribution $\text{DecExp}(x_1 + x_2)$.

<table>
<thead>
<tr>
<th>relbias</th>
<th>$\text{Av}(H_1) \times 10^3$</th>
<th>$\text{Av}(H_2) \times 10^3$</th>
<th>$\text{Av}(H_3) \times 10^3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.7</td>
<td>160 (1)</td>
<td>179 (1)</td>
<td>329 (1)</td>
</tr>
<tr>
<td>3.5</td>
<td>152 (2)</td>
<td>168 (2)</td>
<td>318 (2)</td>
</tr>
<tr>
<td>3.9</td>
<td>148 (3)</td>
<td>160 (3)</td>
<td>300 (5)</td>
</tr>
<tr>
<td>4.7</td>
<td>138 (4)</td>
<td>148 (5)</td>
<td>286 (3)</td>
</tr>
<tr>
<td>4.9</td>
<td>137 (5)</td>
<td>150 (4)</td>
<td>306 (9)</td>
</tr>
<tr>
<td>5.5</td>
<td>130 (6)</td>
<td>138 (6)</td>
<td>267 (4)</td>
</tr>
<tr>
<td>6.4</td>
<td>121 (7)</td>
<td>128 (7)</td>
<td>270 (7)</td>
</tr>
<tr>
<td>6.6</td>
<td>119 (8)</td>
<td>123 (8)</td>
<td>250 (6)</td>
</tr>
<tr>
<td>7.1</td>
<td>113 (9)</td>
<td>119 (9)</td>
<td>291 (8)</td>
</tr>
<tr>
<td>7.6</td>
<td>108 (10)</td>
<td>113 (10)</td>
<td>233 (13)</td>
</tr>
<tr>
<td>8.7</td>
<td>97 (11)</td>
<td>99 (11)</td>
<td>224 (10)</td>
</tr>
<tr>
<td>8.8</td>
<td>95 (12)</td>
<td>97 (12)</td>
<td>211 (11)</td>
</tr>
<tr>
<td>9.1</td>
<td>92 (13)</td>
<td>94 (13)</td>
<td>249 (12)</td>
</tr>
<tr>
<td>11.6</td>
<td>66 (14)</td>
<td>66 (14)</td>
<td>169 (15)</td>
</tr>
<tr>
<td>12.6</td>
<td>55 (15)</td>
<td>55 (15)</td>
<td>182 (14)</td>
</tr>
<tr>
<td>17.7</td>
<td>0 (16)</td>
<td>0 (16)</td>
<td>0 (16)</td>
</tr>
<tr>
<td>$\text{rancor}$</td>
<td>1.00</td>
<td>1.00</td>
<td>0.94</td>
</tr>
</tbody>
</table>
Table 9.5. Value, in ascending order, of $|\text{relbias}|$ 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 correlation, $r_{\text{ancor}}$. Response distribution $\text{DecExp}(y)$

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

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 $H(x)$ succeed in pointing correctly to the preferred $x$-vector? To answer this, let $x_{1k}$ and $x_{2k}$ be two vectors selected for comparison. If the absolute value of the bias of $\hat{Y}_{\text{CAL}}(x_{1k})$ is smaller than that of $\hat{Y}_{\text{CAL}}(x_{2k})$, we would like to see that $H(x_{2k}) \geq H(x_{1k})$ 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 $x_{2k}$.

Sample size plays a role in this. Because $H(x_k)$ 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 \(\theta_k\).) 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 \(x\)-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 \(x\)-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($10^+ x_1 + x_2$)

<table>
<thead>
<tr>
<th>Cells compared</th>
<th>Percent outcomes with correct indication</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$H_1$</td>
</tr>
<tr>
<td>4G+8G (0.5) vs.</td>
<td>90.0</td>
</tr>
<tr>
<td>8G+8G (0.2)</td>
<td>96.4</td>
</tr>
<tr>
<td>4G+2G (1.8) vs.</td>
<td>66.8</td>
</tr>
<tr>
<td>2G+8G (1.5)</td>
<td>74.2</td>
</tr>
<tr>
<td>NG+8G (4.1) vs.</td>
<td>74.3</td>
</tr>
<tr>
<td>8G+NG (3.4)</td>
<td>82.8</td>
</tr>
<tr>
<td>4G+NG (4.1) vs.</td>
<td>90.6</td>
</tr>
<tr>
<td>2G+2G (3.2)</td>
<td>97.0</td>
</tr>
<tr>
<td>NG+2G (7.3) vs.</td>
<td>77.4</td>
</tr>
<tr>
<td>2G+NG (6.5)</td>
<td>85.9</td>
</tr>
</tbody>
</table>
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 $x$-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 \( x^o \) type, as defined in Section 2. (Because most of the variables are available for the full population, they are potentially of the type \( x^e \), but since the effect on bias is of little consequence, we used them as \( x^o \)-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 \( \tilde{Y}_{FUL} \) defined by (3.2) can be computed and used as a reference. We also computed \( \tilde{Y}_{EXP} \) defined by (3.1), as well as \( \tilde{Y}_{CAL} \) defined by (2.5) for different \( x \)-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 \( x = 1 \), and the estimator is \( \tilde{Y}_{EXP} \). 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 $\tilde{Y}_{FUL} = 4265$; the primitive estimate is $\tilde{Y}_{EXP} = 4719$ (both in thousands). This suggests a large positive bias in $\tilde{Y}_{EXP}$, whose relative deviation (in %) from $\tilde{Y}_{FUL}$ is $RDF = (\tilde{Y}_{EXP} - \tilde{Y}_{FUL}) / \tilde{Y}_{FUL} \times 10^2 = 10.7$. 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, $\tilde{Y}_{CAL}$ and $RDF = (\tilde{Y}_{CAL} - \tilde{Y}_{FUL}) / \tilde{Y}_{FUL} \times 10^2$. 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 $\tilde{Y}_{CAL}$ 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 $\tilde{Y}_{CAL}$ 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 \( H_1 \). 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 \( \tilde{y}_{CAL} \) in thousands, and of \( RDF = (\tilde{y}_{CAL}-\tilde{y}_{FUL})\tilde{y}_{FUL} \times 10^3 \). For comparison, \( \tilde{y}_{EXP} \times 10^{-3} = 4719 \); \( \tilde{y}_{FUL} \times 10^{-3} = 4265 \)

<table>
<thead>
<tr>
<th>Auxiliary variable entered</th>
<th>( H_1 \times 10^3 )</th>
<th>( \tilde{y}_{CAL} \times 10^{-3} )</th>
<th>RDF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Income class (3)</td>
<td>76</td>
<td>4458</td>
<td>4.5</td>
</tr>
<tr>
<td>Education level (3)</td>
<td>107</td>
<td>4350</td>
<td>2.0</td>
</tr>
<tr>
<td>Presence of children (2)</td>
<td>114</td>
<td>4326</td>
<td>1.4</td>
</tr>
<tr>
<td>Urban centre dwelling (2)</td>
<td>118</td>
<td>4310</td>
<td>1.1</td>
</tr>
<tr>
<td>Sex (2)</td>
<td>123</td>
<td>4296</td>
<td>0.7</td>
</tr>
<tr>
<td>Marital status (2)</td>
<td>125</td>
<td>4286</td>
<td>0.5</td>
</tr>
<tr>
<td>Days unemployed (3)</td>
<td>121</td>
<td>4301</td>
<td>0.9</td>
</tr>
<tr>
<td>Months with sickness benefits (3)</td>
<td>120</td>
<td>4305</td>
<td>1.0</td>
</tr>
<tr>
<td>Level of debt (3)</td>
<td>115</td>
<td>4322</td>
<td>1.3</td>
</tr>
<tr>
<td>Cluster of postal codes (6)</td>
<td>109</td>
<td>4343</td>
<td>1.8</td>
</tr>
<tr>
<td>Country of birth (2)</td>
<td>103</td>
<td>4363</td>
<td>2.3</td>
</tr>
<tr>
<td>Age class (4)</td>
<td>99</td>
<td>4377</td>
<td>2.6</td>
</tr>
</tbody>
</table>
Table 10.2. 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, $\tilde{Y}_{\text{EXP}} \times 10^{-3} = 4719$; $\tilde{Y}_{\text{FUL}} \times 10^{-3} = 4265$

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

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

For every specified x-vector, a certain bias remains in $\hat{Y}_{\text{CAL}}$. This bias is a function of the choice of 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 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 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


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 \(\mu_{x_1} = ab = 10\) and \(\sigma^2_{x_1} = ab^2 = 50\) 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

\[
A_k = \left(\mu_{x_{2k}|x_{1k}}\right)^2 / \sigma^2_{x_{2k}|x_{1k}} \quad \text{and} \quad B_k = \sigma^2_{x_{2k}|x_{1k}} / \mu_{x_{2k}|x_{1k}},
\]

such that

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

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 \(\alpha, \beta, K\) and \(\sigma^2\). 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
\[
    a_k = \left( \frac{\mu_{y_k|x_{1k},x_{2k}}}{\sigma^2_{y_k|x_{1k},x_{2k}}} \right)^2 \text{ and } b_k = \sigma^2_{y_k|x_{1k},x_{2k}} \left/ \frac{\mu_{y_k|x_{1k},x_{2k}}}{\mu_{y_k|x_{1k},x_{2k}}} \right.,
\]
such that
\[
    \mu_{y_k|x_{1k},x_{2k}} = a_k b_k = c_0 + c_1 x_{1k} + c_2 x_{2k};
\]
\[
    \sigma^2_{y_k|x_{1k},x_{2k}} = a_k b_k^2 = \sigma_0^2 \left( c_1 x_{1k} + c_2 x_{2k} \right).
\]

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 $\alpha$, $\beta$, $K$ and $\sigma^2$ 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.
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 kontaktönsök i en telefonundersökning
Design for estimation: Identifying auxiliary vectors to reduce nonresponse bias

Carl-Erik Särndal
Sixten Lundström