Journal of Official Statistics, Vol. 24, No. 2, 2008, pp. 167–191

Assessing Auxiliary Vectors for Control of Nonresponse Bias in the Calibration Estimator

Carl-Erik Särndal[1] *and Sixten Lundström*[2]

This article develops a bias indicator, a computational tool useful for selecting auxiliary variables likely to be particularly effective for reducing nonresponse bias in estimates obtained by calibration.

The weights used in the calibration estimator are computed on information about a specified auxiliary vector. Even the best among the available auxiliary vectors will leave some bias remaining in the estimator. The objective is to reduce this remaining bias as far as possible, through the choice of a "best possible" auxiliary vector.

The theory in the article is inspired by the survey environment in the Nordic countries and in other North European countries, where many reliable administrative registers provide rich sources of auxiliary variables, in particular for surveys on individuals and households. The many potential auxiliary variables allow the statistician to compose a wide range of auxiliary vectors. There is a need to compare and rank these vectors to assess their effectiveness for bias reduction. The indicator examined in the article serves this purpose.

The indicator is computed on the auxiliary vector values known for the sampled units, responding and nonresponding. It has the advantage of being independent of the study variables, of which the survey may have many. A large value of the indicator suggests a low nonresponse bias, independently of the study variable.

The main body of the article explores the relationship existing between the indicator and the amount of bias in the estimator. The concluding sections are devoted to empirical studies. One of these involves a constructed finite population. The potential auxiliary vectors are ranked with the aid of the indicator. A second empirical illustration illustrates how the indicator is used for selecting auxiliary variables in a large survey at Statistics Sweden.

Key words: Calibration weighting; nonresponse adjustment; nonresponse bias; auxiliary variables; bias indicator.

1. Introduction

When nonresponse occurs in a survey, a pressing objective is to "cleanse" the survey estimates of bias, through an efficient weighting scheme. The bias must lie at the centre of our attention, because the squared bias component often dominates the mean squared error. Unlike the variance, the bias does not approach zero with increasing sample size. The use of auxiliary variables is important, but even in the presence of powerful auxiliary information the nonresponse adjusted estimator may have considerable remaining bias, whether it be constructed by calibration or by any other method.

[1] 2115 Erinbrook Crescent, no. 44, Ottawa, Ontario K1B 4J5, Canada. Email: carl.sarndal@rogers.com
[2] Statistics Sweden, DIH/LEDN, SE-701 89 Örebro, Sweden. Email: sixten.lundstrom@scb.se
Acknowledgments: The authors acknowledge helpful comments from four anonymous referees and the Associate Editor.

Adjustment weighting for nonresponse bias, with the use of auxiliary information, has been considered by several authors and from diverse angles, for example Bethlehem (1988), Deville (2002), Folsom and Singh (2000), Fuller, Loughin, and Baker (1994), Harms (2003), Lundström (1997), Rizzo, Kalton, and Brick (1996), and Thomsen et al. (2006). Some contributions focus on estimation by calibration, notably those of Deville (2002), Harms (2003), and Lundström (1997), and so does this article, which further develops results in the book by Särndal and Lundström (2005).

The calibrated weights are computed from information carried by an auxiliary vector, more or less powerful. In practice, it is impossible to designate a vector that will completely eliminate the bias. Even the best of auxiliary vectors leave some bias remaining in a calibration estimator (or in any other type of estimator). Nevertheless, if estimates are produced at all in the survey, one must settle for one auxiliary vector for computing the calibrated weights and the survey estimates.

A pool of potential auxiliary variables is identified in a preliminary step. The search may involve a matching of different administrative registers. The Nordic countries, The Netherlands and other countries in northern Europe are privileged, equipped as they are with many reliable registers. In a typical survey on individuals and households, a pool of potential auxiliary variables will typically include categorical variables such as sex, age group, income class, country of origin, region of residence, family size, education level, professional group and a variety of others.

With a given pool of auxiliary variables, a number of different auxiliary vectors can be formed. We need to compare these vectors to assess their effectiveness for bias reduction. A tool for this was proposed on intuitive grounds by Särndal and Lundström (2005). Sections 2 to 10 of this article give a more complete picture of this bias indicator and its properties. Section 11 illustrates its use in building an effective auxiliary vector via a stepwise selection of auxiliary variables.

Desirable features of an auxiliary vector are the following: (i) it should well explain the response pattern, (ii) it should well explain the study variable(s) in the survey, and (iii) it should identify the most important domains of interest in the survey. This article focuses on aspect (i).

This article is organized as follows. In Section 2 we specify the calibration estimator and discuss the auxiliary information that goes into it. A close approximation to the bias remaining is given in Section 3. It is a theoretical quantity, depending on values for the entire population of (a) the study variable y_k, (b) the auxiliary vector $\mathbf{x}_k$, and (c) the response probability θ_k, where k is a typical unit. It is practical to work in terms of the inverse response probabilities, discussed in Section 4. We call the unknown $\phi_k = 1/\theta_k$ the *response influence* of population unit k. If the ϕ_k were known, nonresponse bias would cease to be a problem; they would provide the weights that allow unbiased estimation. Section 5 derives predictions of the unknown ϕ_k in terms of the known auxiliary vector values. These predictions, discussed in Sections 5 to 7 as theoretical quantities defined for all N population units and presented in Section 8 as computable, sample-based counterparts, play an important role in the following sections for developing an indicator of the nonresponse bias remaining.

The sample-based bias indicator, denoted q^2, is defined in Section 8 as the variance of the predicted influences for the responding units. An intuitive reason why this variance can

serve to indicate bias is that a variability in the predicted influences (which are surrogates for the true influences ϕ_k) is desirable in order to reflect the individual differences existing among the respondents. But more importantly, results in Sections 7 and 8 show that the nonresponse bias can be expected to decrease linearly, under certain conditions, when the computed value of q^2 increases as a consequence of adding further important variables to the auxiliary vector $\mathbf{x}_k$.

The computation of q^2 requires the auxiliary vector values $\mathbf{x}_k$ for the sampled units, respondents as well as nonrespondents, but is independent of the study variable values y_k. This independence is an advantage: a large survey involves many y-variables but time and resources seldom allow a special analysis for each particular y-variable.

The composition of the auxiliary vector becomes critically important. Like the coefficient of determination R^2 in regression theory, q^2 increases when further variables are added to the vector, and the prospects for reduced bias are improved. Section 9 discusses how q^2 is used as a diagnostic tool to identify the "best auxiliary vector," among those available in the survey.

A constructed population is used in Section 10 to validate the theoretical properties of the bias indicator. The concluding Section 11 shows the use of the bias indicator in the Swedish National Crime Victim and Security Study. The auxiliary vector is built through a stepwise selection of variables, with the aid of the indicator q^2.

2. Auxiliary Information for the Calibration Estimator

We consider a finite population $U = \{1, 2, \ldots, k, \ldots, N\}$. A probability sample s is drawn from U with a sampling design that gives unit k the known inclusion probability $\pi_k = \Pr(k \in s) > 0$. The known design weight of k is $d_k = 1/\pi_k$. Nonresponse occurs. A response set r is realized as a subset of s. We have $r \subseteq s \subseteq U$.

The target of estimation is the population total $Y = \sum_U y_k$, where y_k is the value for unit k of the study variable y, allowed to be continuous or categorical. An example of the latter is when $y_k = 1$ if k is unemployed and $y_k = 0$ otherwise; the target parameter Y is then the number of unemployed in the population. (If $A \subseteq U$ is a set of units, we write $\sum_A$ for $\sum_{k \in A}$.) The value y_k is recorded for $k \in r$ only. Refusal, not-at-home, and incapacity to respond are among the causes for a failure to record y_k for all $k \in s$.

Conceptually, the response set r results when the designated sample s is exposed to an unknown response distribution such that unit k has an unknown response probability $\theta_k = \Pr(k \in r|s)$, assumed positive and independent of s. Although called "response probability," θ_k is viewed here more generally as the probability that the value y_k will be recorded for the unit $k \in s$. With probability $1 - \theta_k$ it goes missing, for whatever reason.

Access to auxiliary information is essential for improved accuracy. Many surveys have information of two types, to which correspond two kinds of auxiliary vector, $\mathbf{x}_k^*$ and $\mathbf{x}_k^\circ$, with the following features. The vector $\mathbf{x}_k^*$ carries auxiliary information at the population level. The value $\mathbf{x}_k^*$ is known for every $k \in U$, as when it is specified in the frame; thus $\mathbf{x}_k^*$ is known also for every $k \in s$ and every $k \in r$. This situation is typical of surveys on individuals and households in Scandinavia and several other North European countries. Then the population total $\sum_U \mathbf{x}_k^*$ is obtained by simply adding the values $\mathbf{x}_k^*$. We allow also the case where $\sum_U \mathbf{x}_k^*$ is imported from a reliable outside source, as when $\sum_U \mathbf{x}_k^*$ includes

population counts, derived from reliable demographic sources, on age group by sex by region. In this article we assume that the individual value $\mathbf{x}_k^*$ is known for every $k \in s$ and consequently for every $k \in r$.

The vector $\mathbf{x}_k^\circ$ carries auxiliary information at the sample level: its value is observed or otherwise known for every $k \in s$ and thus for every $k \in r$. One example occurs when $\mathbf{x}_k^\circ$ expresses features of the data collection process, such as the identity of the interviewer assigned to unit k. As another example, in the case of refusal the interviewer may try the basic question "with the foot in the door," as Kersten and Bethlehem (1984) put it. Yet another example occurs in countries equipped with several administrative registers: it is cumbersome to match at the level of the population with several million records, but more manageable to match at the level of the sample; if the register information is transcribed to the sample data file instead of to the entire population file, it provides material for the $\mathbf{x}_k^\circ$-vector. The vectors $\mathbf{x}_k^\circ$ and $\mathbf{x}_k^*$ differ in that $\sum_U \mathbf{x}_k^*$ is known while $\sum_U \mathbf{x}_k^\circ$ is unknown. Still, the computable unbiased estimate $\sum_s d_k \mathbf{x}_k^\circ$ is an important input of information to the computation of calibrated weights.

For a survey admitting both kinds of information, the auxiliary vector and the information to which we calibrate are

$$\mathbf{x}_k = \begin{pmatrix} \mathbf{x}_k^* \\ \mathbf{x}_k^\circ \end{pmatrix}; \qquad \mathbf{X} = \begin{pmatrix} \sum_U \mathbf{x}_k^* \\ \sum_s d_k \mathbf{x}_k^\circ \end{pmatrix} \tag{1}$$

When the survey has only the first kind of information, then $\mathbf{x}_k = \mathbf{x}_k^*$ and $\mathbf{X} = \sum_U \mathbf{x}_k^*$. When only the second kind is available, $\mathbf{x}_k = \mathbf{x}_k^\circ$ and $\mathbf{X} = \sum_s d_k \mathbf{x}_k^\circ$.

The basis for estimation is as follows. With the kth population unit is associated the quadruplet $(y_k, \mathbf{x}_k, \pi_k, \theta_k)$. Here, π_k is recorded for all $k \in U$, y_k for all $k \in r$, the component $\mathbf{x}_k^*$ of $\mathbf{x}_k$ for all $k \in U$, and the component $\mathbf{x}_k^\circ$ of $\mathbf{x}_k$ for all $k \in s$. The response probability θ_k is defined conceptually; although unobservable for all $k \in U$, it can be estimated; the response indicator values are observed: $I_k = 1$ for $k \in r$, $I_k = 0$ for $k \in s - r$.

The calibration estimator of Y, based on the information $\mathbf{X}$ in (1), is given in Särndal and Lundström (2005) as $\hat{Y}_W = \sum_r w_k y_k$ with weights $w_k = d_k v_k$, where $d_k = 1/\pi_k$ is the design weight and the factor $v_k = 1 + (\mathbf{X} - \sum_r d_k \mathbf{x}_k)'(\sum_r d_k \mathbf{x}_k \mathbf{x}_k')^{-1} \mathbf{x}_k$ serves the double objective of bias reduction and variance reduction. The weights are calibrated to the given information: $\sum_r w_k \mathbf{x}_k = \mathbf{X}$. In this article we consider vectors $\mathbf{x}_k$ with the following property.

There exists a constant vector $\boldsymbol{\mu}$ such that

$$\boldsymbol{\mu}' \mathbf{x}_k = 1 \quad \text{for all } k \in U \tag{2}$$

"Constant" means that $\boldsymbol{\mu}$ must not depend on k, nor on s or on r. Condition (2) is not a major restriction on $\mathbf{x}_k$. A majority of $\mathbf{x}$-vectors of interest in practice are covered. Examples include the following: (1) $\mathbf{x}_k = (1, x_k)'$, where x_k is the value for unit k of a continuous auxiliary variable x; (2) $\mathbf{x}_k = (1, \mathbf{x}_{0k})'$, where the vector $\mathbf{x}_{0k}$ contains two or more continuous variables; (3) the vector representing a categorical x-variable with J mutually exclusive and exhaustive classes, $\mathbf{x}_k = \boldsymbol{\gamma}_k = (\gamma_{1k}, \ldots, \gamma_{jk}, \ldots, \gamma_{Jk})'$, where $\gamma_{jk} = 1$ if k

belongs to group j, and $\gamma_{jk} = 0$ if not, $j = 1, 2, \ldots, J$; (4) the combination of (1) and (3), $\mathbf{x}_k = (\boldsymbol{\gamma}_k', x_k\boldsymbol{\gamma}_k')'$; (5) 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 numbers of classes, and the "minus-one" avoids a singular matrix in the computation of weights calibrated to the two arrays of marginal counts; (6) the extension of (5) to more than two categorical variables. Vectors of Types (5) and (6) are especially important in statistics production at statistical agencies.

A notable case not covered by condition (2) is $\mathbf{x}_k = x_k$, for a one-dimensional continuous variable x with known total $\sum_U x_k$. Then $\hat{Y}_W = (\sum_U x_k)(\sum_r d_k y_k)/(\sum_r d_k x_k)$. Because it has ratio estimator form, it is potentially of interest. But a preferred choice, for reducing nonresponse bias, is to base $\hat{Y}_W$ on the vector that includes the constant one, $\mathbf{x}_k = (1, x_k)'$.

In view of (2), the calibration estimator is

$$\hat{Y}_W = \sum\nolimits_r w_k y_k = \sum\nolimits_r d_k v_k y_k \tag{3}$$

where $d_k = 1/\pi_k$ and, with $\mathbf{X}$ given by (1),

$$v_k = \mathbf{X}'\Big(\sum\nolimits_r d_k \mathbf{x}_k \mathbf{x}_k'\Big)^{-1}\mathbf{x}_k \tag{4}$$

3. Expressions for the Bias Remaining

Even with the "best possible" calibration, some bias remains in $\hat{Y}_W$. The bias is defined jointly with respect to the sampling design $p(s)$ with its (known) inclusion probabilities π_k and the response distribution $q(r|s)$ with its (unknown) response probabilities θ_k. The exact expression, $\text{bias}(\hat{Y}_W) = E_p E_q(\hat{Y}_W|s) - Y$, is intractable, because $\hat{Y}_W$ is nonlinear. We focus on the approximation obtained by Taylor expansion, denoted $\text{nearbias}(\hat{Y}_W)$. Särndal and Lundström (2005), Chapter 9, show that

$$\text{nearbias}(\hat{Y}_W) = (\sum\nolimits_U \mathbf{x}_k)'(\mathbf{B}_{U;\theta} - \mathbf{B}_U) \tag{5}$$

$$\text{where } \mathbf{x}_k = \begin{pmatrix} \mathbf{x}_k^* \\ \mathbf{x}_k^\circ \end{pmatrix},\ \sum\nolimits_U \mathbf{x}_k = \begin{pmatrix} \sum_U \mathbf{x}_k^* \\ \sum_U \mathbf{x}_k^\circ \end{pmatrix},\ \mathbf{B}_{U;\theta} = \Big(\sum\nolimits_U \theta_k \mathbf{x}_k \mathbf{x}_k'\Big)^{-1}\sum\nolimits_U \theta_k \mathbf{x}_k y_k,$$

$$\text{and } \mathbf{B}_U = \Big(\sum\nolimits_U \mathbf{x}_k \mathbf{x}_k'\Big)^{-1}\Big(\sum\nolimits_U \mathbf{x}_k y_k\Big)$$

The derivation of (5) need not be reproduced here. Similar expressions are given in, or can be derived from, sources such as Bethlehem (1988) and Fuller (2002), although their conditions differ from ours. The approximation is close, even for rather modest sample sizes. Under mild conditions, $(1/N)(\text{bias}(\hat{Y}_W) - \text{nearbias}(\hat{Y}_W))$ is of order $n^{-1/2}$, where n is the sample size.

We note that $\text{nearbias}(\hat{Y}_W)$ is an unknown theoretical quantity. It depends on $(y_k, \mathbf{x}_k, \theta_k)$ for all $k \in U$ but is independent of the inclusion probabilities π_k. The nearbias is not reduced by the choice of a more variance-efficient sampling design.

The presence in (5) of the difference $\mathbf{B}_{U;\theta} - \mathbf{B}_U$ underlines one of the predicaments with nonresponse: we end up estimating not the desired regression coefficient $\mathbf{B}_U$, but the "tainted" regression coefficient $\mathbf{B}_{U;\theta}$. Their difference causes a more or less pronounced bias in $\hat{Y}_W$. (The following principle of notation applies for several symbols with two indices separated by a semicolon: the first index shows the set of units over which the quantity is defined, and the second shows the weighting, as in $\mathbf{B}_{U;\theta}$. In the case of equal weighting, the second index is suppressed, as in $\mathbf{B}_U$.)

To achieve nearbias$(\hat{Y}_W) = 0$ is a farfetched possibility. It would happen if all θ_k were equal, an unrealistic hope. As Result 4.1 will show, nearbias$(\hat{Y}_W) = 0$ holds under yet another condition, also one that is unlikely to hold in practice. No matter how good the auxiliary information, some bias remains; what we can do is to try to reduce it.

The approximation (5) of bias$(\hat{Y}_W)$ shows that the nearbias is the same whether a given auxiliary variable is of the $\mathbf{x}_k^*$ kind or of the $\mathbf{x}_k^\circ$ kind. An auxiliary variable x_k is as efficient for reducing the nearbias when it belongs in $\mathbf{x}_k^\circ$ (carrying information to the sample level only) as when it qualifies for inclusion in $\mathbf{x}_k^*$ (carrying information up to the population level). One notes also that

$$\sum\nolimits_U \mathbf{x}_k = \begin{pmatrix} \sum_U \mathbf{x}_k^* \\ \sum_U \mathbf{x}_k^\circ \end{pmatrix}$$

in (5) differs from the information $\mathbf{X}$ used in computing the calibrated weights (4) as soon as $\mathbf{x}_k$ contains an $\mathbf{x}_k^\circ$-component. Noting that $\sum_U (y_k - \mathbf{x}_k' \mathbf{B}_U) = 0$ by (2), we can write (5) as

$$\text{nearbias}(\hat{Y}_W) = \sum\nolimits_U (\theta_k M_k - 1) y_k \tag{6}$$

where

$$M_k = \Big(\sum\nolimits_U \mathbf{x}_k \Big)' \Big(\sum\nolimits_U \theta_k \mathbf{x}_k \mathbf{x}_k' \Big)^{-1} \mathbf{x}_k \tag{7}$$

The scalar M_k is a derived variable, defined by (7) for all $k \in U$. This variable is a focal point in diagnosing the bias.

One objective is to compare alternative $\mathbf{x}_k$-vectors in regard to their capacity to control the bias. A suitable benchmark is then the "primitive auxiliary vector," $\mathbf{x}_k = 1$ for all $k \in U$. It gives $\hat{Y}_W = N\bar{y}_{r;d} = N(\sum_r d_k y_k)/(\sum_r d_k)$, $M_k = N/\sum_U \theta_k = 1/\bar{\theta}_U$ for all k, and

$$\text{nearbias}(N\bar{y}_{r;d}) = N(\bar{y}_{U;\theta} - \bar{y}_U) \tag{8}$$

where $\bar{y}_{U;\theta} = \sum_U \theta_k y_k / \sum_U \theta_k$ and $\bar{y}_U = \sum_U y_k / N$. The vector $\mathbf{x}_k = 1$ recognizes no differences among units and is therefore inefficient for nonresponse adjustment. The nearbias is large if the theta-weighted mean and the unweighted mean differ considerably, as when large y-value units respond with low probability.

We define the *nearbias ratio* as

$$P = \frac{\text{nearbias}(\hat{Y}_W)}{\text{nearbias}(N\bar{y}_{r;d})} = \frac{\sum_U (\theta_k M_k - 1) y_k}{N(\bar{y}_{U;\theta} - \bar{y}_U)} \tag{9}$$

It shows how well a specified vector $\mathbf{x}_k$ succeeds in controlling the bias, as compared with the primitive vector. A highly effective $\mathbf{x}_k$-vector brings a near-zero value of P.

4. Response Influence and Zero Nearbias

Under what conditions can the nearbias be zero? One aspect of this is seen by writing (6) as

$$\text{nearbias}(\hat{Y}_W) = \sum_U \theta_k M_k e_k \tag{10}$$

where $e_k = y_k - \mathbf{x}_k' \mathbf{B}_U$ is the ordinary least squares regression residual for unit k. That (6) and (10) are equal follows from $\sum_U \theta_k M_k \mathbf{x}_k' \mathbf{B}_U = \sum_U \mathbf{x}_k' \mathbf{B}_U = \sum_U y_k$. As (10) shows, $\text{nearbias}(\hat{Y}_W) = 0$ holds if $e_k = 0$ for all $k \in U$, that is, if $\mathbf{x}_k$ perfectly explains y_k. Most large surveys involve many y-variables. If $e_k = 0$ for all $k \in U$ *and* for every one of the perhaps numerous y-variables, then the survey gives unbiased estimation. This is a vain hope. If we focus instead on the response distribution, there are conditions under which the nearbias is zero or near zero *for every* y-variable.

It is convenient to work with the inverse of the response probability θ_k rather than with θ_k itself. We define the *response influence* of k as $\phi_k = 1/\theta_k$, assuming that $0 < \theta_k \leq 1$ for all k. The unknown value ϕ_k can be seen as a latent trait of unit k. A high influence ϕ_k accompanies a unit k with a low response probability θ_k, just as a high sampling weight $d_k = 1/\pi_k$ accompanies a unit with a low inclusion probability π_k. Unlike the y_k, the ϕ_k remain unknown even for responding units. If the ϕ_k were known, they would serve to compute the unbiased estimator $\sum_r d_k \phi_k y_k$ for $Y = \sum_U y_k$; nonresponse bias would cease to be a problem. We call ϕ_k "influence" to distinguish it from "weight," which is a known number that can be readily applied to an observed variable value. The unknown ϕ_k do not qualify for this purpose.

An ideal auxiliary vector $\mathbf{x}_k$ is one that perfectly explains the influence ϕ_k. It meets the following condition.

There exists a constant vector $\boldsymbol{\lambda}$ such that

$$\phi_k = 1/\theta_k = \boldsymbol{\lambda}' \mathbf{x}_k \quad \text{for all } k \in U \tag{11}$$

In survey practice, we cannot hope to find an ideal vector $\mathbf{x}_k$. But if one existed and could be used, the nearbias would be zero, as the following result shows.

Result 4.1. If $\mathbf{x}_k$ meets the condition (11), then $\text{nearbias}(\hat{Y}_W) = 0$.

Proof. When (11) holds, then $(\sum_U \mathbf{x}_k)'(\sum_U \theta_k \mathbf{x}_k \mathbf{x}_k')^{-1} = \boldsymbol{\lambda}'(\sum_U \theta_k \mathbf{x}_k \mathbf{x}_k')\times (\sum_U \theta_k \mathbf{x}_k \mathbf{x}_k')^{-1} = \boldsymbol{\lambda}'$, so $\theta_k M_k = \theta_k \boldsymbol{\lambda}' \mathbf{x}_k = 1$ for all $k \in U$, and, by (6), $\text{nearbias}(\hat{Y}_W) = 0$.

A simple application of Result 4.1 occurs for the classification vector, $\mathbf{x}_k = (\gamma_{1k}, \ldots, \gamma_{Jk})'$, where $\gamma_{jk} = 1$ if unit k belongs to the specified group j and

$\gamma_{jk} = 0$ if not; $j = 1, 2, \ldots, J$. Result 4.1 then states that the nearbias is zero if the response probability θ_k is constant for all units within one and the same group. This is in fact an assumption often made (explicitly or implicitly) in practice. It happens to be a convenient one. Few would believe it to "hold true," for example, in regard to a set of age/sex groups for a population of individuals. The remaining bias can be large. Better **x**-vectors are required to effectively combat the bias.

5. Least Squares Prediction of the Influence

The influences $\phi_k = 1/\theta_k$ are unknown and nonobservable. We can, however, obtain predicted influences, using the auxiliary data $\mathbf{x}_k$ for $k \in s$. These predicted values serve in Section 8 to construct the bias indicator. To motivate these sample-based predictions, we first derive their population-based counterparts. Let us determine the vector $\boldsymbol{\lambda}$ to minimize the weighted sum of squared differences $\text{WSS} = \sum_U \theta_k(\phi_k - \boldsymbol{\lambda}'\mathbf{x}_k)^2$. It is mathematically convenient to work with theta-weighted sums of squares; WSS stands for "weighted sum of squares." It is assumed that not all ϕ_k are equal. Differentiate WSS with respect to $\boldsymbol{\lambda}$, set the derivative equal to zero to obtain the estimating equation $\sum_U \theta_k(\phi_k - \boldsymbol{\lambda}'\mathbf{x}_k)\mathbf{x}_k' = \mathbf{0}'$, or equivalently,

$$\boldsymbol{\lambda}'\Big(\sum\nolimits_U \theta_k \mathbf{x}_k \mathbf{x}_k'\Big) = \Big(\sum\nolimits_U \mathbf{x}_k\Big)' \tag{12}$$

If the matrix on the left-hand side is nonsingular, the solution is $\boldsymbol{\lambda}' = \hat{\boldsymbol{\lambda}}_U'$, where

$$\hat{\boldsymbol{\lambda}}_U' = \Big(\sum\nolimits_U \mathbf{x}_k\Big)'\Big(\sum\nolimits_U \theta_k \mathbf{x}_k \mathbf{x}_k'\Big)^{-1} \tag{13}$$

The resulting predicted value of ϕ_k is

$$\hat{\phi}_{Uk} = \hat{\boldsymbol{\lambda}}_U'\mathbf{x}_k = \Big(\sum\nolimits_U \mathbf{x}_k\Big)'\Big(\sum\nolimits_U \theta_k \mathbf{x}_k \mathbf{x}_k'\Big)^{-1}\mathbf{x}_k = M_k \tag{14}$$

The quantities M_k reappear here. They were seen earlier to be important in the expression (6) for the nearbias. We need several properties of the M_k, beginning with the following result.

Result 5.1. The quantities M_k, defined by (7) for $k \in U$, satisfy the equations

$$\sum\nolimits_U \theta_k M_k \mathbf{x}_k' = \sum\nolimits_U \mathbf{x}_k'; \qquad \sum\nolimits_U \theta_k M_k = N; \qquad \sum\nolimits_U \theta_k M_k^2 = \sum\nolimits_U M_k \tag{15}$$

These equations follow by straightforward algebra, the second one with the aid of (2). The variation of the ϕ_k around their theta-weighted mean, $\bar{\phi}_{U;\theta} = \sum_U \theta_k \phi_k / \sum_U \theta_k = 1/\bar{\theta}_U$, is measured by $\sum_U \theta_k(\phi_k - 1/\bar{\theta}_U)^2$. A part of that variation is explained by the least squares predictions $\hat{\phi}_{Uk} = M_k$ based on a given vector $\mathbf{x}_k$. The equation "total variation = explained variation + residual variation" reads

$$\sum\nolimits_U \theta_k(\phi_k - 1/\bar{\theta}_U)^2 = \sum\nolimits_U \theta_k(M_k - 1/\bar{\theta}_U)^2 + \sum\nolimits_U \theta_k(\phi_k - M_k)^2 \tag{16}$$

The cross-product term is zero as a consequence of (15). We develop each of the three terms in (16), use (15), and divide through by N. The resulting equation is

$$\bar{\phi}_U - 1/\bar{\theta}_U = (\bar{M}_U - 1/\bar{\theta}_U) + (\bar{\phi}_U - \bar{M}_U) \tag{17}$$

where $\bar{\phi}_U - 1/\bar{\theta}_U$, $\bar{M}_U - 1/\bar{\theta}_U$ and $\bar{\phi}_U - \bar{M}_U$ are nonnegative terms with

$$\bar{\phi}_U = \sum_U \phi_k/N, \qquad \bar{M}_U = \sum_U M_k/N, \qquad \bar{\theta}_U = \sum_U \theta_k/N \tag{18}$$

In addition to the unweighted mean $\bar{M}_U = \sum_U M_k/N$, we need the theta-weighted mean $\bar{M}_{U;\theta} = \sum_U \theta_k M_k / \sum_U \theta_k$. It follows from (15) that

$$\bar{M}_{U;\theta} = N/\sum_U \theta_k = 1/\bar{\theta}_U \tag{19}$$

Thus $\bar{M}_{U;\theta}$ depends on the response distribution (through the mean response probability $\bar{\theta}_U$) but is independent of the auxiliary vector $\mathbf{x}_k$. By contrast, the unweighted mean $\bar{M}_U$ depends on $\mathbf{x}_k$. Key properties of $\bar{M}_U$ are shown in the following result.

Result 5.2. For any given auxiliary vector $\mathbf{x}_k$, $\bar{M}_U = \sum_U M_k/N$ satisfies

$$1/\bar{\theta}_U = \bar{M}_{U;\theta} \le \bar{M}_U \le \bar{\phi}_U \tag{20}$$

where the different means are defined by (18). The lower bound on $\bar{M}_U$, $1/\bar{\theta}_U$, occurs for the primitive vector, $\mathbf{x}_k = 1$ for all k. The upper bound on $\bar{M}_U$, $\bar{\phi}_U$, is attained for the ideal (in practice nonexistent) vector $\mathbf{x}_k$ that meets condition (11).

The inequalities in (20) follow from the nonnegativity of each of the two terms in parenthesis on the right-hand side of (17). The stated upper and lower bounds of $\bar{M}_U$ are easily verified.

By definition, the influences satisfy $\phi_k = 1/\theta_k > 1$ for all $k \in U$. Do the predictions satisfy $\hat{\phi}_{Uk} = M_k > 1$ for all $k \in U$? The answer is that while this is likely to hold for a majority of units, it may not necessarily hold for all units. By the nonnegativity of the first term on the right-hand side of (17), $\bar{M}_U \ge 1/\bar{\theta}_U > 1$, which does not exclude that a few M_k may fail to exceed unity. This is of no serious consequence for the rest of the article.

6. Other Moments of the Predicted Influences

To see the relation between the predictions $\hat{\phi}_{Uk} = M_k$ and nearbias($\hat{Y}_W$) we need further moments of the M_k, in addition to $\bar{M}_U$ and $\bar{M}_{U;\theta}$: (i) the theta-weighted variance, denoted Q^2, (ii) the theta-weighted coefficient of variation, denoted H, and (iii) the theta-weighted coefficient of correlation between M_k and ϕ_k, denoted $R_{M\phi}$. All of Q^2, H and $R_{M\phi}$ are unknown, theoretical quantities, dependent as they are on $(y_k, \mathbf{x}_k, \theta_k)$ for all $k \in U$. Sample-based, computable analogues of Q^2 and H are given in Section 8.

The theta-weighted variance of the predictions $\hat{\phi}_{Uk} = M_k$ for $k \in U$ is given by

$$Q^2 = \frac{1}{\sum_U \theta_k} \sum_U \theta_k (M_k - \bar{M}_{U;\theta})^2 \tag{21}$$

It forms the prototype for the computable bias indicator q^2 in Section 8. Expanding the square and arranging terms, using Result 5.1 we get:

$$Q^2 = \frac{\sum_U M_k}{\sum_U \theta_k} - \frac{N^2}{\left(\sum_U \theta_k\right)^2} = (1/\bar{\theta}_U)(\bar{M}_U - 1/\bar{\theta}_U) \tag{22}$$

Among the properties of Q^2 are: (a) for any given vector $\mathbf{x}_k$, $Q^2 \geq 0$, because Q^2 is a variance, hence nonnegative; (b) the minimum value, $Q^2 = 0$, occurs for the primitive vector, $\mathbf{x}_k = 1$ for all $k \in U$; (c) the upper bound on Q^2, denoted Q^2_{sup}, would be realized only for the ideal vector $\mathbf{x}_k$ that meets condition (11); by Result 5.2 we have

$$Q^2_{\text{sup}} = \frac{\sum_U \phi_k}{\sum_U \theta_k} - \frac{N^2}{\left(\sum_U \theta_k\right)^2} = (1/\bar{\theta}_U)(\bar{\phi}_U - 1/\bar{\theta}_U) \tag{23}$$

(d) extending the $\mathbf{x}_k$-vector by adding further x-variables to it will increase the value of Q^2 (or possibly leave it unchanged). The proof of (d) relies on the fact that the extended vector produces a value of the term "explained variation" in (16) which is at least as large as the value of that same term for the vector that excludes those additional variables.

Another useful quantity is the (theta-weighted) coefficient of variation of the M_k, defined as the standard deviation Q divided by the corresponding mean $\bar{M}_{U;\theta} = 1/\bar{\theta}_U$, so that

$$H = Q/\bar{M}_{U;\theta} = \sqrt{\bar{M}_U \bar{\theta}_U - 1} \tag{24}$$

The upper bound on H is $H_{\text{sup}} = \sqrt{\bar{\phi}_U \bar{\theta}_U - 1}$. The theta-weighted coefficient of correlation between M_k and ϕ_k is

$$R_{M\phi} = \frac{\sum_U \theta_k (M_k - \bar{M}_{U;\theta})(\phi_k - \bar{\phi}_{U;\theta})}{\left(\sum_U \theta_k (M_k - \bar{M}_{U;\theta})^2\right)^{1/2} \left(\sum_U \theta_k (\phi_k - \bar{\phi}_{U;\theta})^2\right)^{1/2}}$$

where $\bar{M}_{U;\theta} = \bar{\phi}_{U;\theta} = 1/\bar{\theta}_U$ by (19). Noting that $\sum_U \theta_k (\phi_k - \bar{\phi}_{U;\theta})^2 = N(\bar{\phi}_U - 1/\bar{\theta}_U)$, and using (21) and (22), we get

$$R_{M\phi} = \sqrt{\frac{\bar{M}_U - 1/\bar{\theta}_U}{\bar{\phi}_U - 1/\bar{\theta}_U}} \tag{25}$$

The quantities Q^2, H and $R_{M\phi}$ are related in the following way:

$$1 - R^2_{M\phi} = \frac{\bar{\phi}_U - \bar{M}_U}{\bar{\phi}_U - 1/\bar{\theta}_U} = 1 - \frac{Q^2}{Q^2_{\text{sup}}} = 1 - \frac{H^2}{H^2_{\text{sup}}} \tag{26}$$

7. Towards an Indicator of the Bias Remaining

The nearbias, given by (5) or (6), is expressed in the following result as the sum of a principal term that is linear in Q^2 (and in H^2) and a residual term, Δ, which is often small by comparison.

Result 7.1. Consider a given auxiliary vector $\mathbf{x}_k$ for the calibration estimator $\hat{Y}_W$. Then

$$\text{nearbias}(\hat{Y}_W) = N(\bar{y}_{U;\theta} - \bar{y}_U)(1 - R^2_{M\phi}) + \Delta \tag{27}$$

where $1 - R^2_{M\phi}$ has the alternative expressions given by (26) and $\Delta = \sum_U \theta_k M_k E_k$ with

$$E_k = y_k - \bar{y}_U - (\phi_k - \bar{\phi}_U)\frac{\bar{y}_U - \bar{y}_{U;\theta}}{\bar{\phi}_U - 1/\bar{\theta}_U} \tag{28}$$

Proof. The values y_k and ϕ_k are associated with unit k. For any given constants α and β, we can also associate with unit k the unique value $y_k - \alpha - \beta\phi_k$. Let us determine α and β to minimize $\sum_U \theta_k(y_k - \alpha - \beta\phi_k)^2$. We get $\beta = B = (\bar{y}_U - \bar{y}_{U;\theta})/(\bar{\phi}_U - 1/\bar{\theta}_U)$ and $\alpha = A = \bar{y}_U - B\bar{\phi}_U$. Now insert $y_k = A + B\phi_k + E_k$ in (6) and simplify to get

$$\text{nearbias}(\hat{Y}_W) = A\sum\nolimits_U(\theta_k M_k - 1) + B\,N(\bar{M}_U - \bar{\phi}_U) + \sum\nolimits_U(\theta_k M_k - 1)E_k$$

But $\sum_U(\theta_k M_k - 1) = 0$ by (15), and $\sum_U E_k = 0$. Then (27) follows from (26).

The magnitude of the reminder term Δ is discussed at the end of this section. Result 7.1 states that the principal term of nearbias($\hat{Y}_W$), based on a given auxiliary vector $\mathbf{x}_k$, equals a proportion, $1 - R^2_{M\phi}$, of its value, $N(\bar{y}_{U;\theta} - \bar{y}_U)$, for the primitive vector, for which $R^2_{M\phi} = 0$. As the auxiliary vector $\mathbf{x}_k$ improves and approaches its ideal form (11), $\bar{M}_U$ increases towards its upper bound $\bar{\phi}_U$, the proportion $1 - R^2_{M\phi}$ tends to zero, and nearbias($\hat{Y}_W$) approaches zero if Δ is small. The bias may be considerably reduced if steps are taken to strengthen the $\mathbf{x}_k$-vector. We note the following consequence of (27).

Result 7.2. If the remainder term Δ is small in comparison with the first term on the right-hand side of (27) then

$$P = \frac{\text{nearbias}(\hat{Y}_W)}{\text{nearbias}(N\bar{y}_r)} = \frac{\sum_U(\theta_k M_k - 1)y_k}{N(\bar{y}_{U;\theta} - \bar{y}_U)} \approx 1 - R^2_{M\phi} = 1 - \frac{Q^2}{Q^2_{\text{sup}}} \tag{29}$$

The nearbias ratio P, which resembles a proportion, depends on $(y_k, \mathbf{x}_k, \theta_k)$ for $k \in U$. It measures how well the given vector $\mathbf{x}_k$ succeeds in controlling the bias, as compared to the primitive vector. In (29), P is approximated by $1 - R^2_{M\phi}$, which depends on $(\mathbf{x}_k, \theta_k)$ for $k \in U$, but is independent of the y-variable. Thus $1 - R^2_{M\phi}$ represents "the proportion of the nearbias ratio P that is independent of the study variable."

The remainder term Δ in (27) is not in general zero, but it is indeed zero under any one of the several conditions in the following result.

Result 7.3. Consider a fixed auxiliary vector $\mathbf{x}_k$. The remainder term $\Delta = \sum_U \theta_k M_k E_k$ in (27) is equal to zero under any one of the following four conditions: (i) $\mathbf{x}_k$ is the primitive vector $\mathbf{x}_k = 1$ for all k; (ii) $\mathbf{x}_k$ satisfies condition (11); (iii) for some constant vector $\boldsymbol{\mu}$,

$E_k = \boldsymbol{\mu}'(\mathbf{x}_k - \bar{\mathbf{x}}_U)$ for $k \in U$, where E_k is given by (28); (iv) for some constants c_0 and c_1, $y_k = c_0 + c_1\phi_k$ for $k \in U$.

Here, case (iii) states that $\mathbf{x}_k$ explains perfectly the variation remaining in y_k after a removal of the dependence on ϕ_k. Case (iv), stating that y_k is perfectly explained by the influence ϕ_k, may be described as "purely nonignorable nonresponse."

Proof. In case (i), the result follows by noting that $M_k = 1/\bar{\theta}_U$ for all k. In case (ii), nearbias($\hat{Y}_W$) is zero by Result 4.1, and the proportion $1 - R^2_{M\phi}$ in (27) is also equal to zero, because $\bar{M}_U = \bar{\phi}_U$ by Result 5.2. Hence, $\Delta = 0$. In case (iii), $\Delta = 0$ follows from (15). Finally, in case (iv), simple algebra and the use of (26) show that nearbias($\hat{Y}_W$) $= N(\bar{y}_{U;\theta} - \bar{y}_U)(1 - R^2_{M\phi}) = Nc_1(\bar{M}_U - \bar{\phi}_U)$, hence $\Delta = 0$.

Remark 7.1. When different $\mathbf{x}_k$-vectors are at our disposal in a survey, we want to identify one that is likely to effectively control the bias of *all* study variables y. Formula (29) shows that the nearbias ratio is roughly a linearly decreasing function of $Q^2 = Q^2(\mathbf{x}_k)$, the constant Q^2_{sup} being independent of $\mathbf{x}_k$. Hence we should seek an $\mathbf{x}_k$ for which Q^2 is large. When a certain $\mathbf{x}_k$-vector is replaced by "an improved one," with an accompanying increase in the value of Q^2, we expect the nearbias ratio P to drop in a roughly linear manner. Ideally, the chosen vector $\mathbf{x}_k$ should bring a value of Q^2 close to the upper bound Q^2_{sup}, guaranteeing near-zero bias for all y-variables. If Q^2 and H were computable in a survey, either one would serve as an indicator of bias remaining. But both depend on the whole population with its unknown response probabilities. In experiments, such as those in Section 10, we can study how Q^2 tracks the nearbias for different vectors $\mathbf{x}_k$. A computable, sample-based counterpart of Q^2, denoted q^2, is given in Section 8. The use of q^2 as a diagnostic tool is discussed in Section 9.

Remark 7.2. Formula (21) defines Q^2 as the variance of the predicted influences $\hat{\phi}_{Uk} = M_k$. By Result 7.2, the larger the variance Q^2, the better the chances that the bias will be small. This is in line with the intuition that the more the predictions $\hat{\phi}_{Uk}$ can reflect the individual features of the respondents, the better the chances of a small bias.

8. Sample-based Counterparts

The population quantities M_k, Q^2 and H have sample-based counterparts, m_k, q^2 and h, given in this section. They are computed from two kinds of input: (i) the vector values

$$\mathbf{x}_k = \begin{pmatrix} \mathbf{x}^*_k \\ \mathbf{x}^\circ_k \end{pmatrix}$$

known for $k \in s$, and (ii) the outcome of the response phase, $I_k = 1$ for $k \in r$ and $I_k = 0$ for $k \in s - r$. They do not depend on the y-values.

Formula (14) gave the predicted influences for $k \in U$ as $\hat{\phi}_{Uk} = \hat{\boldsymbol{\lambda}}'_U \mathbf{x}_k = M_k$, where $\hat{\boldsymbol{\lambda}}'_U$ is the solution of the population-based estimating equation (12). The corresponding sample-based estimating equation is obtained by substituting the unbiased estimates $\sum_s d_k \mathbf{x}_k$ and $\sum_r d_k \mathbf{x}_k \mathbf{x}'_k$ for the unknown population sums in (12). It should be noted

that $E_p(E_q(\sum_r d_k \mathbf{x}_k \mathbf{x}_k' | s)) = E_p(\sum_s d_k \theta_k \mathbf{x}_k \mathbf{x}_k') = \sum_U \theta_k \mathbf{x}_k \mathbf{x}_k'$. The estimating equation is $\boldsymbol{\lambda}'(\sum_r d_k \mathbf{x}_k \mathbf{x}_k') = (\sum_s d_k \mathbf{x}_k)'$; its solution is $\boldsymbol{\lambda}' = \hat{\boldsymbol{\lambda}}_s' = (\sum_s d_k \mathbf{x}_k)'(\sum_r d_k \mathbf{x}_k \mathbf{x}_k')^{-1}$, supposing the matrix can be inverted. The sample-based prediction of ϕ_k, computable for $k \in s$, is $\hat{\phi}_{sk} = \hat{\boldsymbol{\lambda}}_s' \mathbf{x}_k = m_k$, where

$$m_k = \Big(\sum_s d_k \mathbf{x}_k\Big)' \Big(\sum_r d_k \mathbf{x}_k \mathbf{x}_k'\Big)^{-1} \mathbf{x}_k \tag{30}$$

The scalar values m_k are close to (but not in general equal to) the weight factors v_k used to compute the estimator $\hat{Y}_W = \sum_r d_k v_k y_k$ given by (3) and (4). We do have $m_k = v_k$ when the auxiliary information is exclusively at the sample level, so that $\mathbf{x}_k = \mathbf{x}_k^\circ$. Otherwise, m_k and v_k differ by a usually small amount. The equations in (15) have the counterparts

$$\sum_r d_k m_k \mathbf{x}_k' = \sum_s d_k \mathbf{x}_k'; \quad \sum_r d_k m_k = \sum_s d_k; \quad \sum_r d_k m_k^2 = \sum_s d_k m_k \tag{31}$$

Hence $\sum_r d_k m_k$ is unbiased for the population size N, because $\sum_s d_k$ has this property. To the means $\bar{M}_U$ and $\bar{M}_{U;\theta}$ correspond

$$\bar{m}_{s;d} = \frac{\sum_s d_k m_k}{\sum_s d_k}; \qquad \bar{m}_{r;d} = \frac{\sum_r d_k m_k}{\sum_r d_k} = \frac{\sum_s d_k}{\sum_r d_k} \tag{32}$$

Hence $1/\bar{m}_{r;d}$ is a measure of the survey response rate, independently of $\mathbf{x}_k$. The quantity Q^2 was defined by (21) as the (theta-weighted) variance of the predicted influences $\hat{\phi}_{Uk} = M_k$ for $k \in U$. By the same logic, q^2 is now defined to be the (design-weighted) variance of the sample-based predictions $\hat{\phi}_{sk} = m_k$ for $k \in r$:

$$q^2 = \frac{1}{\sum_r d_k} \sum_r d_k (m_k - \bar{m}_{r;d})^2 \tag{33}$$

A simple development and the use of (31) gives

$$q^2 = \frac{\sum_s d_k m_k}{\sum_r d_k} - \frac{\Big(\sum_s d_k\Big)^2}{\Big(\sum_r d_k\Big)^2} = \bar{m}_{r;d}(\bar{m}_{s;d} - \bar{m}_{r;d}) \tag{34}$$

Since $q^2 \geq 0$ and $\bar{m}_{r;d} \geq 1$, it follows that $\bar{m}_{s;d} \geq \bar{m}_{r;d}$. It is useful to remember the interpretation of q^2 as the variance of the sample-based predicted influences. But a familiar line of reasoning brings an alternative interpretation of q^2: replace each population sum in the expression (22) for Q^2 by the corresponding unbiased estimate. That is, the sums $\sum_U \theta_k \mathbf{x}_k \mathbf{x}_k'$, $\sum_U \theta_k$, $\sum_U \mathbf{x}_k$ and N in Q^2, are replaced by the respective unbiased estimates, $\sum_r d_k \mathbf{x}_k \mathbf{x}_k'$, $\sum_r d_k$, $\sum_s d_k \mathbf{x}_k$ and $\sum_s d_k$. In particular, $\sum_U M_k = (\sum_U \mathbf{x}_k)'(\sum_U \theta_k \mathbf{x}_k \mathbf{x}_k')^{-1} (\sum_U \mathbf{x}_k)$ in (22) is replaced by $(\sum_s d_k \mathbf{x}_k)'(\sum_r d_k \mathbf{x}_k \mathbf{x}_k')^{-1}(\sum_s d_k \mathbf{x}_k) = \sum_s d_k m_k$. We arrive at (34).

Some properties of q^2 are: (a) for any given auxiliary vector $\mathbf{x}_k$, $q^2 \geq 0$, because q^2 is a variance; (b) $q^2 = 0$ for the primitive vector, $\mathbf{x}_k = 1$ for all k; (c) $q^2 = 0$ for complete response, $r = s$; (d) $q^2 = 0$ if $\bar{\mathbf{x}}_{s;d} = \bar{\mathbf{x}}_{r;d}$; (e) in contrast to Q^2, q^2 does not have a finite upper bound; (f) for a given $\mathbf{x}_k$, q^2 converges in probability, under mild conditions, to Q^2,

because each population sum in Q^2 is replaced in q^2 by a corresponding design-unbiased estimate.

The convergence of q^2 to Q^2 may be slow, and the sample-to-sample variability of q^2 may be considerable, unless both s and r are rather large sets. For best results, q^2 should be used with the large sample sizes, often one thousand or more, that are typical of government surveys.

The indicator q^2 was proposed and used in Särndal and Lundström (2005), under the notation *IND1*, for the purpose of comparing different vectors $\mathbf{x}_k$ for bias reduction. This role of q^2 is further developed in the following sections. A different tool for the selection of x-variables is proposed in Bethlehem and Schouten (2004) and Schouten (2007). It depends both on the response outcome and on the response variable y. It seeks to combine two aspects of an auxiliary vector $\mathbf{x}_k$: on the one hand how well it explains the response pattern, on the other hand how well it explains the study variable y. It is thus a y-dependent indicator. By contrast, q^2 in this article has an objective to serve as a guide no matter how many y-variables the survey may contain. The question of choice among available auxiliary variables for bias reduction appears earlier in the literature. For example, Rizzo, Kalton, and Brick (1996) view the choice of auxiliary variables as a somewhat more important question than the choice among alternative algorithms for computing the weights given a set of variables.

The population-based coefficient of variation H given by (26) has the sample-based analogue

$$h = q/\bar{m}_{r;d} = \sqrt{\frac{\bar{m}_{s;d}}{\bar{m}_{r;d}} - 1}$$

A reason to prefer h to q^2 in empirical work is that it mitigates the tendency in q^2 to increase with increasing rates of nonresponse. The population correlation coefficient $R_{M\phi}$ does not have a sample-based counterpart.

9. A Diagnostic Tool for Assessing the Bias Reduction Potential of an Auxiliary Vector

When a survey encounters a sizeable nonresponse, the onus is on the survey producer to adjust the estimates. A rich source of auxiliary data is a necessary prerequisite. Such an environment is found in a number of North European countries, where reliable registers of the total population provide extensive auxiliary data for surveys on individuals and households. These databases contain many potential auxiliary variables. In a preliminary inventory, a pool of potential x-variables is identified. A range of possible auxiliary vectors $\mathbf{x}_k$ can be considered. Both types of information, $\mathbf{x}_k^*$ and $\mathbf{x}_k^\circ$, may be present in $\mathbf{x}_k$. We want to compare those $\mathbf{x}_k$-vectors in regard to their capacity to reduce the bias remaining in the calibration estimator $\hat{Y}_W = \sum_r w_k y_k$. In practice, one vector will ultimately be chosen for computing the weights $w_k = d_k v_k$.

In practice, how do we compare the various candidate vectors $\mathbf{x}_k$? By (29), an increase in $Q^2 = Q^2(\mathbf{x}_k)$ is accompanied by a roughly linear decrease in the relative nearbias P. The empirical evidence in Section 10 supports this contention. In practice, $Q^2(\mathbf{x}_k)$ is

replaced by the computable indicator $q^2 = q^2(\mathbf{x}_k)$ given by (33) or (34). What assurance do we have that $q^2(\mathbf{x}_k)$ will guide us correctly to the preferred $\mathbf{x}_k$-vector? Suppose we compare two candidate $\mathbf{x}$-vectors, $\mathbf{x}_{1k}$ and $\mathbf{x}_{2k}$, related hierarchically so that $\mathbf{x}_{2k}$ is made up of $\mathbf{x}_{1k}$ and an additional vector $\mathbf{x}_{+k}$: $\mathbf{x}_{2k} = (\mathbf{x}_{1k}^{'}, \mathbf{x}_{+k}^{'})'$. Then $Q^2(\mathbf{x}_{2k}) \geq Q^2(\mathbf{x}_{1k})$ by property (d) in Section 6, which says that adding further variables to $\mathbf{x}_{1k}$ increases Q^2 (or possibly leaves it unchanged). The same ordering holds for the sample-based counterparts: $q^2(\mathbf{x}_{2k}) \geq q^2(\mathbf{x}_{1k})$, for any realized sample s and response set r. That is, if Q^2 indicates that $\mathbf{x}_{2k}$ is preferred to $\mathbf{x}_{1k}$, then q^2 will agree with this order of preference, whatever the realization (s, r). Still there is no guarantee that the bias is smaller for $\mathbf{x}_{2k}$ than for $\mathbf{x}_{1k}$, but (29) suggests that this is so.

The situation is different if the compared vectors $\mathbf{x}_{2k}$ and $\mathbf{x}_{1k}$ are not related hierarchically, that is, when $\mathbf{x}_{2k}$ is not the result of adding more variables to $\mathbf{x}_{1k}$. Then $q^2(\mathbf{x}_{2k}) \geq q^2(\mathbf{x}_{1k})$ may hold for some realizations (s, r), but not necessarily all, as illustrated at the end of the next section.

The indicator $q^2 = q^2(\mathbf{x}_k)$ provides a tool for a stepwise selection of x-variables from a pool of J potentially interesting x-variables, continuous or categorical. In Step 1 of a stepwise forward selection, compute $q^2(\mathbf{x}_k)$ for each single x-variable; retain the one that yields the largest value of $q^2(\mathbf{x}_k)$. In Step 2, compute $q^2(\mathbf{x}_k)$ for each of the $J - 1$ vectors $\mathbf{x}_k$ composed of the variable from Step 1 and one additional x-variable; of those vectors, retain the one that yields the highest increase in $q^2(\mathbf{x}_k)$, and so on, if further steps are needed. Typically, the successive increases in $q^2(\mathbf{x}_k)$ taper off, as exemplified by the study in Section 11. It is assumed that all $\mathbf{x}_k$-vectors considered in the stepwise procedure satisfy the condition (2). If x is a one-dimensional continuous variable, the vector under consideration in Step 1 is $\mathbf{x}_k = (1, x_k)'$.

An alternative is to use $q^2(\mathbf{x}_k)$ for a stepwise backward deletion of x-variables, one at a time, beginning with the full vector $\mathbf{x}_k$, composed of all J x-variables deemed to be of interest. For either of two reasons one may not wish to retain all the variables in that vector: (i) some of the x-variables may contribute little to the objective of reducing bias, or (ii) inspection of the set of weights w_k produced by $\mathbf{x}_k$ may reveal some undesirably large or small values. The following procedure may be followed. In Step 1, compute $q^2(\mathbf{x}_k)$ for the full vector. In Step 2, compute $q^2(\mathbf{x}_k)$ for each of the J different vectors $\mathbf{x}_k$ with one x-variable deleted; consider retaining the vector causing the least reduction in $q^2(\mathbf{x}_k)$. Additional steps follow the same routine. A significant drop in $q^2(\mathbf{x}_k)$ is a sign that the deleted x-variable is important for bias reduction. The procedure stops if at a certain step both of properties A and B hold, where A = "the drop in $q^2(\mathbf{x}_k)$ by deleting the next x-variable is numerically important," and B = "inspection of the set of weights is satisfactory."

10. Empirical Study of the Relation Between the Nearbias and the Bias Indicator

The first objective in this section is to study empirically how well $Q^2 = Q^2(\mathbf{x}_k)$ succeeds in tracking the value of nearbias$(\hat{Y}_W(\mathbf{x}_k))$. For a given constructed y-variable, we compose a number of auxiliary vectors $\mathbf{x}_k$, we compute both nearbias$(\hat{Y}_W(\mathbf{x}_k))$ and $Q^2(\mathbf{x}_k)$ for each vector, and we observe how these two quantities move together when $\mathbf{x}_k$ changes. By Result 7.2, we expect $Q^2(\mathbf{x}_k)$ to be able to rank the vectors $\mathbf{x}_k$ in regard to their ability

to reduce the bias, if not perfectly, so at least with a high rate of success. A comparison of two vectors $\mathbf{x}_{1k}$ and $\mathbf{x}_{2k}$ is expected to show, for any response distribution, that when $Q^2(\mathbf{x}_{2k}) > Q^2(\mathbf{x}_{1k})$, then nearbias$(\hat{Y}_W(\mathbf{x}_{2k})) <$ nearbias$(\hat{Y}_W(\mathbf{x}_{1k}))$. Our empirical results confirm this pattern.

An empirical study of this kind requires values $(y_k, \mathbf{x}_k, \theta_k)$ specified for $k = 1, 2, \ldots, N$. We experimented with several such constructed populations. The conclusions were similar. We report here results for one study variable with value y_k specified for $k = 1, 2, \ldots, N = 6,000$, together with 16 different specifications of $\mathbf{x}_k$, under each of four different response distributions with θ_k specified for all k. For each response distribution, we computed nearbias$(\hat{Y}_W(\mathbf{x}_k))$ and $Q(\mathbf{x}_k)$ for the 16 different $\mathbf{x}_k$-vectors, as well as the nearbias ratio and the coefficient of nondetermination, computed in the image of (29) as

$$P(\mathbf{x}_k) = \frac{\text{nearbias}(\hat{Y}_W(\mathbf{x}_k))}{N(\bar{y}_{U;\theta} - \bar{y}_U)} = \frac{\sum_U (\theta_k M_k(\mathbf{x}_k) - 1) y_k}{N(\bar{y}_{U;\theta} - \bar{y}_U)}; \qquad T(\mathbf{x}_k) = 1 - \frac{Q^2(\mathbf{x}_k)}{Q^2_{\text{sup}}}$$

We plotted the 16 points $(|P(\mathbf{x}_k)|, T(\mathbf{x}_k))$. The primitive vector $\mathbf{x}_k = 1$ gives the point (1,1). The other 15 points lie inside the unit square. If the remainder term Δ is small, Result 7.3 suggests that the points will, apart from some scatter, align themselves around the diagonal of the unit square, and that a decrease in $T(\mathbf{x}_k)$ is accompanied by a linear decrease in $P(\mathbf{x}_k)$. We have $P(\mathbf{x}_k) = T(\mathbf{x}_k) = 0$ for an $\mathbf{x}_k$-vector that satisfies condition (11). Our study has no such ideal vector, but both $P(\mathbf{x}_k)$ and $T(\mathbf{x}_k)$ come near zero for some of the more powerful vectors $\mathbf{x}_k$. (Although this did not occur in the experiment reported here, a powerful $\mathbf{x}_k$-vector may yield a small negative value of $P(\mathbf{x}_k)$.)

The 16 vectors $\mathbf{x}_k$ were created by different uses of the values x_{1k} and x_{2k} of two continuous auxiliary variables, x_1 and x_2. We created (y_k, x_{1k}, x_{2k}) for $k = 1, 2, \ldots, 6,000$ 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 $a = 2$, $b = 5$. This theoretical mean is $\mu_{x_1} = ab = 10$; the theoretical variance is $\sigma^2_{x_1} = ab^2 = 50$. The mean of the 6,000 realized values x_{1k} was 10.0 and the variance was 49.9.

Step 2: The continuous auxiliary variable x_2. For unit k, with the value x_{1k} fixed in 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 $B_k = \sigma^2_{x_{2k}|x_{1k}} / \mu_{x_{2k}|x_{1k}}$, where

$$\mu_{x_{2k}|x_{1k}} = \alpha + \beta x_{1k} + K\, h(x_{1k}) \text{ and } \sigma^2_{x_{2k}|x_{1k}} = \sigma^2 x_{1k} \tag{35}$$

with $h(x_{1k}) = x_{1k}(x_{1k} - \mu_{x_1})(x_{1k} - 3\mu_{x_1})$. Suitable values were assigned to the constants α, β, K and σ^2. The conditional expectation of x_{2k} given x_{1k} is the sum of the linear term $\alpha + \beta x_{1k}$ and the polynomial term $K\, h(x_{1k})$, which gives a somewhat nonlinear appearance to the plotted points (x_{2k}, x_{1k}). This was done on purpose, to avoid the argument that some simulation results may happen just because of a linear relationship. We used the values $\alpha = 1$, $\beta = 1$, $K = 0.001$, $\mu_{x_1} = 10$ 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 $b_k = \sigma^2_{y_k|x_{1k},x_{2k}}/\mu_{y_k|x_{1k},x_{2k}}$, where

$$\mu_{y_k|x_{1k},x_{2k}} = c_0 + c_1 x_{1\,k} + c_2 x_{2k} \text{ and } \sigma^2_{y_k|x_{1k},x_{2k}} = \sigma_0^2(c_1 x_{1k} + c_2 x_{2k}) \tag{36}$$

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 the values $c_0 = 1$, $c_1 = 0.7$, $c_2 = 0.3$ 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.

Each of the two continuous variables x_1 and x_2 was employed in the experiment in four different group modes, denoted 8G, 4G, 2G, and NG, to obtain $4 \times 4 = 16$ different auxiliary vectors $\mathbf{x}_k$. The procedure for the variable x_1 was the following. 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. This defines mode 8G of variable x_1; unit k is assigned the group indicator vector $\boldsymbol{\gamma}_{(x_1;8)k}$, of dimension eight with seven entries "0" and a single entry "1" to identify the group membership of k. For example, $\boldsymbol{\gamma}_{(x_1;8)k} = (0,0,0,0,1,0,0,0)'$ implies that k is one of the 750 units in Group 5 of the x_1-variable. Next, successive group mergers are carried out; two adjoining groups always defining a new group, doubling the group size and causing a progressive loss of information. 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 $\boldsymbol{\gamma}_{(x_1;4)k}$ is associated with unit k. In mode 2G, unit k has the indicator vector $\boldsymbol{\gamma}_{(x_1;2)k}$ such that $\boldsymbol{\gamma}_{(x_1;2)k} = (1,0)'$ 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 form a single group, all x_1-information is relinquished, and $\boldsymbol{\gamma}_{(x_1;1)k} = 1$ for all k.

The same procedure was used to transform the 6,000 values x_{2k} into the group modes 8G, 4G, 2G, and NG. Correspondingly, the group information for 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, $4 \times 4 = 16$ different auxiliary vectors $\mathbf{x}_k$ are formed by combining the group information as shown in the following display.

	Use made of x_{2k}			
Use made of x_{1k}	Eight size groups	Four size groups	Two size groups	Not used
Eight size groups	8G + 8G	8G + 4G	8G + 2G	8G + NG
Four size groups	4G + 8G	4G + 4G	4G + 2G	4G + NG
Two size groups	2G + 8G	2G + 4G	2G + 2G	2G + NG
Not used	NG + 8G	NG + 4G	NG + 2G	NG + NG

The "+" indicates that the $\mathbf{x}_k$-vector is formed by placing the two $\boldsymbol{\gamma}$-vectors "side by side," the effect being a calibration on the margins. For case 8G + 8G, unit k has the auxiliary vector $\mathbf{x}_k = (\boldsymbol{\gamma}'_{(x_1;8)k}, \boldsymbol{\gamma}'_{(x_2;8)k})'_{(-1)}$, 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 case 8G + 8G makes the most complete use of the group information. At the other extreme, the case NG + NG disregards all the information and gives rise to the primitive auxiliary vector $\mathbf{x}_k = 1$ for all k. There are 14 intermediate cases. For example, the case 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 case 4G + NG has $\mathbf{x}_k = (\boldsymbol{\gamma}'_{(x_1;4)k}, 1)'_{(-1)} = \boldsymbol{\gamma}_{(x_1;4)k}$.

We report results for four different response distributions:

(i) IncExp$(10 + x_1 + x_2)$, defined by $\theta_k = 1 - \mathrm{e}^{-c(10+x_{1k}+x_{2k})}$ with $c = 0.04599$
(ii) IncExp$(10 + y)$, defined by $\theta_k = 1 - \mathrm{e}^{-c(10+y_k)}$ with $c = 0.06217$
(iii) DecExp$(x_1 + x_2)$, defined by $\theta_k = \mathrm{e}^{-c(x_{1k}+x_{2k})}$ with $c = 0.01937$
(iv) DecExp(y), defined by $\theta_k = \mathrm{e}^{-cy_k}$ with $c = 0.03534$

The constant c was chosen in each option to deliver a mean response probability of $\bar{\theta}_U = \sum_U \theta_k / N = 0.70$. The value 10 (rather than 0) is used in options (i) and (ii) to avoid a high incidence of very small response probabilities θ_k. The four options represent contrasting features of the response probabilities: decreasing as opposed to increasing, dependent on x-values only as opposed to dependent on y-values only. Options (ii) and (iv), where the response is entirely y-variable dependent, might be called "purely nonignorable."

Many other response distributions could be considered in this experimental setting. The preceding theory suggests that the approximate linear relationship between $Q^2(\mathbf{x}_k)$ and the nearbias will prevail for any response distribution. It is nevertheless clear that one can provoke situations with relationships among y_k, $\mathbf{x}_k$ and θ_k such that the approximate linear relationship is significantly perturbed. This has not been an objective in this study.

Tables 1 to 4 show the value (in percent) of relbias$(\hat{Y}_W(\mathbf{x}_k)) =$ nearbias$(\hat{Y}_W(\mathbf{x}_k))/(N\bar{y}_U)$, and (in parenthesis) the value of $Q^2(\mathbf{x}_k)$ for the 16 $\mathbf{x}_k$-vectors. In each table, the case NG + NG gives $Q^2(\mathbf{x}_k) = 0$, and relbias$(\hat{Y}_W(\mathbf{x}_k))$ is at its highest level. At the other extreme, the case 8G + 8G gives the highest value of $Q^2(\mathbf{x}_k)$ and the lowest value of relbias$(\hat{Y}_W(\mathbf{x}_k))$. Other cases are intermediate. The tables confirm property (d) of Section 6, namely that the value $Q^2(\mathbf{x}_k)$ increases by moving upwards within every column and by moving from right to left within every row. All four tables show that the absolute values of relbias also follow a monotonic (but decreasing) pattern, something which however is not guaranteed by the preceding theory. To each table corresponds one of the Figures 1 to 4, showing the plotted points $(P(\mathbf{x}_k), T(\mathbf{x}_k))$ for the 16 auxiliary vectors $\mathbf{x}_k$. The tables and the figures prompt several comments:

1. *Comparing x-dependent response distributions with y-dependent response distributions*. The best of the auxiliary vectors (those for Case 8G + 8G), yield near-zero nearbias for the x-dependent response distributions. For example, Table 1 for IncExp$(10 + x_1 + x_2)$ shows relbias (in %) decreasing from 13.2 (Case NG + NG) to 0.2 (Case 8G + 8G). The decreasing pattern holds also for the y-dependent response distributions, IncExp$(10 + y)$ and DecExp(y), with the

Table 1. Relbias $\hat{Y}_W(\mathbf{x}_k)$ in % and value of $Q^2(\mathbf{x}_k)$ in % (within parenthesis) for 16 auxiliary vectors $\mathbf{x}_k$. Response distribution IncExp$(10 + x_1 + x_2)$

Use made of x_{1k}	Use made of x_{2k}			
	Eight size groups	Four size groups	Two size groups	Not used
Eight size groups	0.2 (9.5)	0.4 (9.3)	1.3 (8.7)	3.4 (6.5)
Four size groups	0.5 (9.2)	0.8 (9.0)	1.8 (8.4)	4.1 (6.0)
Two size groups	1.5 (8.5)	1.9 (8.2)	3.2 (7.3)	6.5 (4.3)
Not used	4.1 (6.7)	5.0 (6.3)	7.3 (5.0)	13.2 (0.0)

difference that the relbias does not come close to zero for the best vectors. In Table 2 for IncExp(y) the relbias (in %) decreases from 13.1 (Case NG + NG) to 3.6 (Case 8G + 8G). This emphasizes the importance of a powerful $\mathbf{x}_k$-vector also for nonignorable nonresponse.

2. *Linear relationship between $T(\mathbf{x}_k)$ and $P(\mathbf{x}_k)$.* The visual impression in all of Figures 1 to 4 is one of strong linearity. Results 7.1 and 7.2 lead us to expect that as the auxiliary vector $\mathbf{x}_k$ improves, $T(\mathbf{x}_k)$ and $P(\mathbf{x}_k)$ will decrease together in a nearly linear fashion. To measure this tendency, we computed the product–moment correlation coefficient, denoted r_{TP}, based on the 16 points $(P(\mathbf{x}_k), T(\mathbf{x}_k))$. Table 5 shows values of r_{TP} near one for all four response distributions, indicating near perfect linear relationship. We also computed the Spearman rank correlation coefficient, denoted R_{TP}, based on the 16 points $(P(\mathbf{x}_k), T(\mathbf{x}_k))$. Table 5 shows that R_{TP} is also near one in all four cases, so for this population, $T(\mathbf{x}_k)$ gives an almost perfect ranking of the 16 $\mathbf{x}_k$-vectors. The same attractive property applies to $Q^2(\mathbf{x}_k)$ since it is linearly related to $T(\mathbf{x}_k)$.
3. *The effect of the remainder term* Δ. Result 7.3 leads us to expect the points $(P(\mathbf{x}_k), T(\mathbf{x}_k))$ to be aligned, except for some scatter, around the diagonal of the unit

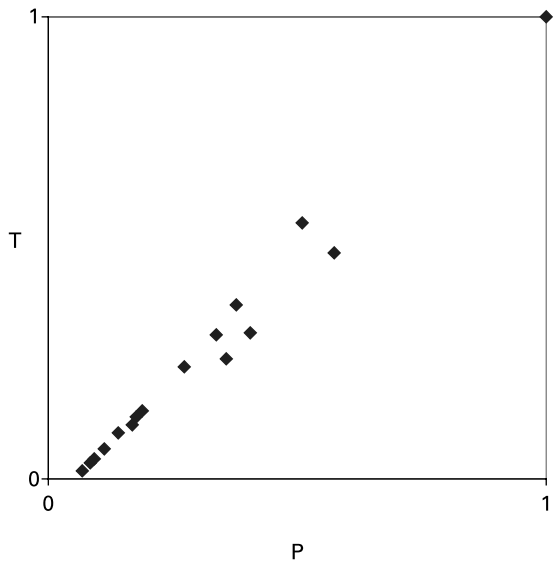

Fig. 1. Plot of $(P(\mathbf{x}_k), T(\mathbf{x}_k))$ for 16 auxiliary vectors $\mathbf{x}_k$. Response distribution IncExp$(10 + x_1 + x_2)$

Table 2. Relbias $\hat{Y}_W(\mathbf{x}_k)$ in % and value of $Q^2(\mathbf{x}_k)$ in % (within parenthesis) for 16 auxiliary vectors $\mathbf{x}_k$. Response distribution IncExp$(10 + y)$

Use made of x_{1k}	Use made of x_{2k}			
	Eight size groups	Four size groups	Two size groups	Not used
Eight size groups	3.6 (4.3)	3.8 (4.2)	4.3 (4.0)	5.3 (3.6)
Four size groups	4.0 (4.1)	4.3 (4.0)	4.9 (3.8)	6.0 (3.3)
Two size groups	4.9 (3.6)	5.3 (3.5)	6.2 (3.3)	7.9 (2.5)
Not used	7.1 (2.4)	7.9 (2.2)	9.6 (1.6)	13.1 (0.0)

square. This assumes that the term Δ in (27) is small by comparison. The diagonal pattern is most clearly pronounced in Figures 1, 2, and 4, but somewhat less prominent in Figure 3 for DecExp$(x_1 + x_2)$, although the linear relationship remains strong there also. Figure 3 suggests that in that case, Δ may not be negligible, compared to the principal term. However, the ranking of the $\mathbf{x}$-vectors remains excellent, with $R_{TP} = 0.92$.

4. *Interactions*. There is nonnegligible interaction between x_1 and x_2 in the population constructed for this experiment. We found that a cross-classification, such as 2G $\times$ 2G, gave smaller values of nearbias (and correspondingly lower values of $Q^2(\mathbf{x}_k)$) than a corresponding "side by side" arrangement, such as 2G + 2G, which disregards interactions.

For the population in this experiment, Tables 1 to 4 and Figures 1 to 4 support the idea that, if computable, $Q^2(\mathbf{x}_k)$ would be a good instrument for ranking the possible $\mathbf{x}_k$-vectors. In practice, the computable sample-based analogue $q^2(\mathbf{x}_k)$ must be used. How well does $q^2(\mathbf{x}_k)$ succeed in ranking the $\mathbf{x}_k$-vectors? For row-wise and for column-wise comparisons in Tables 1 to 4, the $\mathbf{x}_k$-vectors are in a hierarchical relationship, in the sense of Section 9. When the vectors $\mathbf{x}_{1k}$ and $\mathbf{x}_{2k}$ belong in the same table row or in the same table column, and

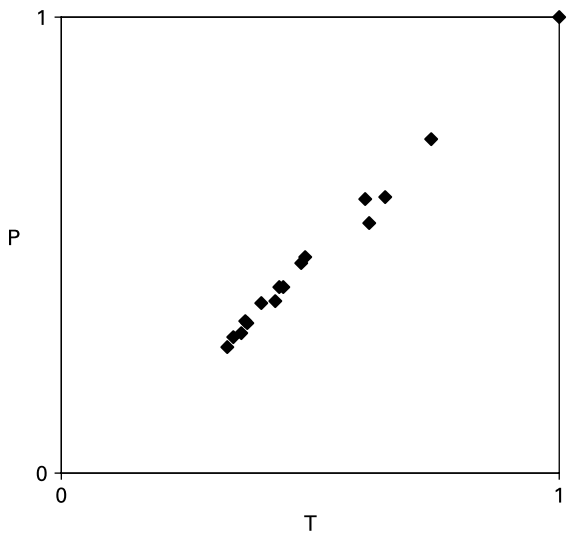

Fig. 2. Plot of $(P(\mathbf{x}_k), T(\mathbf{x}_k))$ for 16 auxiliary vectors $\mathbf{x}_k$. Response distribution IncExp$(10 + y)$

Table 3. Relbias $\hat{Y}_W(\mathbf{x}_k)$ in % and value of $Q^2(\mathbf{x}_k)$ in % (within parenthesis) for 16 auxiliary vectors $\mathbf{x}_k$. Response distribution DecExp$(x_1 + x_2)$

Use made of x_{1k}	Use made of x_{2k}			
	Eight size groups	Four size groups	Two size groups	Not used
Eight size groups	−2.8 (20.1)	−3.9 (17.0)	−5.6 (13.6)	−7.6 (10.3)
Four size groups	−3.5 (19.3)	−4.8 (16.0)	−6.6 (12.3)	−8.8 (8.8)
Two size groups	−4.9 (18.0)	−6.4 (14.4)	−8.7 (10.1)	−11.5 (5.8)
Not used	−7.2 (16.4)	−9.1 (12.3)	−12.6 (6.7)	−17.7 (0.0)

$Q^2(\mathbf{x}_{2k}) \geq Q^2(\mathbf{x}_{1k})$, then $q^2(\mathbf{x}_{2k}) \geq q^2(\mathbf{x}_{1k})$ follows for any outcome (s, r). For example, if $\mathbf{x}_{1k}$ is the vector for 4G + 2G, and $\mathbf{x}_{2k}$ is the one for 8G + 2G, then computation will necessarily show that $q^2(\mathbf{x}_{2k}) \geq q^2(\mathbf{x}_{1k})$ for any (s, r), confirming that the nearbias is smaller (in absolute value) for 8G + 2G than for 4G + 2G.

The situation is different if $\mathbf{x}_{1k}$ and $\mathbf{x}_{2k}$ do not belong to the same row or the same column. Over repeated outcomes (s, r), we may then find $q^2(\mathbf{x}_{2k}) \geq q^2(\mathbf{x}_{1k})$ for some but not all outcomes. Especially if the difference nearbias$(\hat{Y}_W(\mathbf{x}_{1k}))$ − nearbias$(\hat{Y}_W(\mathbf{x}_{2k}))$ is considerable (in absolute value), we would like to see that $q^2(\mathbf{x}_{2k}) \geq q^2(\mathbf{x}_{1k})$ holds in a vast majority of all outcomes (s, r), because then the indicator q^2 leads with large probability to the correct decision to base the estimator $\hat{Y}_W$ on $\mathbf{x}_{2k}$ rather than on $\mathbf{x}_{1k}$.

We shed further light on this question by Monte Carlo experiments, in which 5,000 outcomes (s, r) were realized. Repeated simple random samples s of size 1,000 were drawn, and, for every given s, r was realized according to each of the four response distributions. That is, unit k is declared "responding" if a Bernoulli experiment with the specified θ_k gives "success." We then computed the proportion of outcomes (s, r) in which the correct ordering is achieved. It is of particular interest to compare cases where the

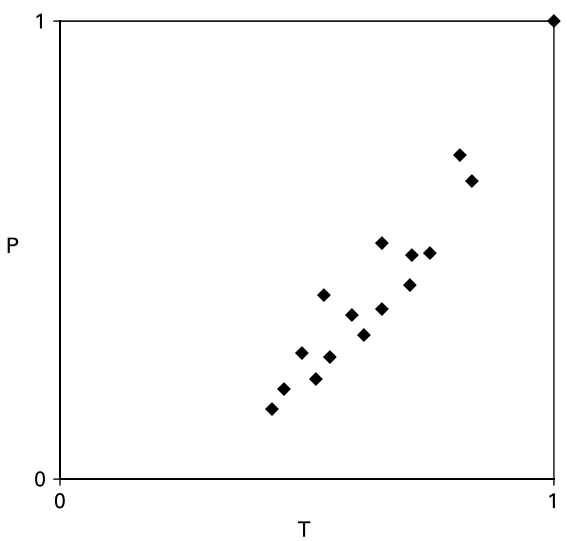

Fig. 3. Plot of $(P(\mathbf{x}_k), T(\mathbf{x}_k))$ for 16 auxiliary vectors $\mathbf{x}_k$. Response distribution DecExp$(x_1 + x_2)$

Table 4. Relbias $\hat{Y}_W(\mathbf{x}_k)$ in % and value of $Q^2(\mathbf{x}_k)$ in % (within parenthesis) for 16 auxiliary vectors $\mathbf{x}_k$. Response distribution DecExp(y)

Use made of x_{1k}	Use made of x_{2k}			
	Eight size groups	Four size groups	Two size groups	Not used
Eight size groups	−8.2 (12.6)	−8.9 (11.7)	−9.8 (10.6)	−11.0 (9.5)
Four size groups	−9.0 (11.6)	−9.8 (10.5)	−10.9 (9.3)	−12.2 (8.0)
Two size groups	−10.5 (10.0)	−11.5 (8.7)	−12.9 (7.0)	−14.8 (5.3)
Not used	−12.9 (7.8)	−14.4 (6.1)	−16.8 (3.5)	−20.5 (0.0)

nearbias values are close, so that a correct decision is hard to obtain. Some examples of these comparisons:

(i) Comparison of 4G + 2G (with vector denoted $\mathbf{x}_{1k}$) with 2G + 8G (with vector denoted $\mathbf{x}_{2k}$) for IncExp($10 + x_1 + x_2$). By Table 1, relbias($\hat{Y}_W(\mathbf{x}_{2k})$) $= 1.5 < 1.8 =$ relbias($\hat{Y}_W(\mathbf{x}_{1k})$) and, correspondingly, $Q^2(\mathbf{x}_{2k}) = 8.5 > 8.4 = Q^2(\mathbf{x}_{1k})$. Thus the vector $\mathbf{x}_{2k}$ is slightly better for reducing nearbias, an order of preference confirmed by Q^2. The correct ordering, $q^2(\mathbf{x}_{2k}) > q^2(\mathbf{x}_{1k})$, occurred here for a high proportion, 70.7%, of the 5,000 outcomes (s, r).

(ii) Comparison of 2G + 2G (vector $\mathbf{x}_{1k}$) with 4G + NG (vector $\mathbf{x}_{2k}$) for DecExp(y). By Table 4, relbias($\hat{Y}_W(\mathbf{x}_{2k})$) $= -12.2$ and relbias($\hat{Y}_W(\mathbf{x}_{1k})$) $= -12.9$. Thus $\mathbf{x}_{2k}$ is the slightly better vector, by the absolute value of nearbias. This order of preference is confirmed by the Q^2-values: $Q^2(\mathbf{x}_{2k}) = 8.0 > 7.0 = Q^2(\mathbf{x}_{1k})$. The correct ordering $\hat{Q}^2(\mathbf{x}_{2k}) > \hat{Q}^2(\mathbf{x}_{1k})$ was realized in 78.1% of the 5,000 outcomes (s, r).

(iii) Comparison of NG + 2G (vector $\mathbf{x}_{1k}$) with 2G + NG (vector $\mathbf{x}_{2k}$) for DecExp($x_1 + x_2$). By Table 3, relbias($\hat{Y}_W(\mathbf{x}_{2k})$) $= -11.5$ and relbias($\hat{Y}_W(\mathbf{x}_{1k})$) $= -12.6$. Here $\mathbf{x}_{2k}$ is the somewhat better vector by the absolute value of relbias, but

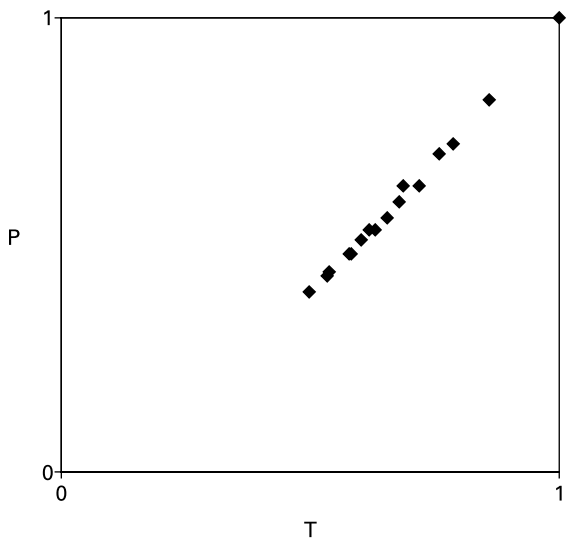

Fig. 4. Plot of $(P(\mathbf{x}_k), T(\mathbf{x}_k))$ for 16 auxiliary vectors $\mathbf{x}_k$. Response distribution DecExp(y)

Table 5. Product–moment correlation coefficient r_{TP}, and Spearman rank correlation coefficient R_{TP}, computed on 16 points $(P(\mathbf{x}_k), T(\mathbf{x}_k))$, for four response distributions

Response distribution	r_{TP}	R_{TP}
IncExp$(10 + x_1 + x_2)$	0.99	0.99
IncExp$(10 + y)$	1.00	0.99
DecExp$(x_1 + x_2)$	0.95	0.92
DecExp(y)	1.00	0.99

this is one of the rare instances where the order of preference is not confirmed by the Q^2-values: we have $Q^2(\mathbf{x}_{2k}) = 5.8 < 6.7 = Q^2(\mathbf{x}_{1k})$. Not surprisingly, the correct ordering $\hat{Q}^2(\mathbf{x}_{2k}) > \hat{Q}^2(\mathbf{x}_{1k})$ occurred in less than a majority of the 5,000 outcomes (s, r), namely, 31.2%.

11. Use of the Bias Indicator in the Swedish National Crime Victim and Security Study

In 2006, the Swedish National Council for Crime Prevention (*Brottsförebyggande Rådet*, acronym *BRÅ*) conducted a National Crime Victim and Security Study. As part of the study, Statistics Sweden carried out a survey in which 10,000 persons were sampled from the Swedish Register of the Total Population (RTP). The survey objective was to measure trends in certain types of crimes, in particular crimes against the person. It would provide an opportunity to assess levels of insecurity, and how these levels vary with respect to various groups in Swedish society.

A stratified simple random sample s of 10,000 persons was drawn from the RTP. The strata were defined by the cross classification of region of residence by age group. The regions are the 21 Swedish administrative areas known as "*län*." The three age groups were defined by the brackets 16–29, 30–74, and 75–79. This design reflects an objective to get accurate results for each of the 21 *län* as well as for each of the three age groups. The allocation of the sample to strata was roughly proportional to the population size in the stratum, with minor modifications to reflect the goal of sufficient accuracy for the domains of particular interest, the *län* and the age groups. The overall response rate was 77.8%. The nonresponse, more or less pronounced in the different domains of interest, interferes to some degree with the accuracy objective.

The pool of potential auxiliary variables consisted of those in the RTP and a subset of those in another Statistics Sweden's database, LISA. All auxiliary variables are categorical. Groups were formed for the variables that are continuous by nature. Variables obtained from LISA were transcribed only to the sample database, so they are of the $\mathbf{x}_k^\circ$ type defined in Section 2.

For this survey, we illustrate the use of q^2 as a tool for stepwise forward selection of variables, as explained in Section 9. In each step, the auxiliary vector $\mathbf{x}_k$ expands by addition of the categorical variable that yields the largest increase in q^2 at that point. A new variable joins already entered variables in the "side-by-side" (or "+") manner. Table 6 shows the variable entered into $\mathbf{x}_k$ in the first ten forward selection steps. Country of birth, entered in Step 1, is the dichotomous variable indicating Scandinavian-born or not. Age group and sex, adjustment variables "by routine" in many surveys, do qualify for

Table 6. National Crime Victim and Security Study; stepwise forward selection of variables for the auxiliary vector

Step	Auxiliary variable entering	Number of groups	Value of $1{,}000 \times q^2$
0	–	–	0
1	Country of birth	2	20.0
2	Income group	3	27.6
3	Age group	6	31.3
4	Gender	2	35.1
5	Marital status	2	38.6
6	Region	21	40.7
7	Family size group	5	41.4
8	Days unemployed	6	41.9
9	Urban centre dweller	2	42.3
10	Occupation	10	42.7

inclusion here, in Steps 3 and 4. The pool of potential auxiliary variables included a number of others, not shown in the table.

Table 6 also shows the number of groups for each categorical variable, and the successive values of $1,000 \times q^2$. Not unexpectedly, the increases in q^2 taper off after a few steps. This suggests that there would be little point, for bias reduction, in using more than the first six x-variables, and perhaps the first four would suffice.

Estimates were produced in the survey for many categorical study variables, as totals or as proportions. The typical targeted population total Y is a population count, the number of persons with a specific property, relating, say, to insecurity and/or fear of becoming a victim of some form of crime. Thus $Y = \sum_U y_k$, where $y_k = 1$ if person k has the property and $y_k = 0$ if not. The bias remaining in the final count estimates remains unknown. But we can follow the stepwise evolution of the estimates. For some study variables we computed the estimated count at each step in Table 6, $\hat{Y}_W = \sum_r w_k y_k$ with weights w_k based on the $\mathbf{x}_k$-vector with the variables selected up until and including the step in question. The estimate used in Step 0 was computed by straight expansion within strata as $\hat{Y}_W = \sum_{h=1}^{H} N_h \bar{y}_{r_h}$, where $\bar{y}_{r_h}$ is the mean response in stratum h.

Some estimated counts changed by two or more percentage points in the progression from Step 0 to Step 6. This is a large change; nonresponse has considerable effect. Still, there is no guarantee that the estimate in Step 6 is any more accurate (less biased) than the one in Step 0, but theory leads one to expect so. A typical pattern was that the greatest change in the estimate occurred in the transition from Step 0 to Step 1, that the change was quite noticeable also in Steps 2, 3, and 4, and that the change then subsided. This pattern agrees with the progression of q^2 shown in Table 6. Some variables appear to be little affected by the nonresponse, the change in the estimates being small in all steps.

12. Concluding Comment

This article suggests using the indicator q^2 as a tool for building the auxiliary vector for the final calibrated weights. The bias in the final estimates still remains unknown. We do

not resolve questions such as: How large is the squared bias component of the mean squared error? To what extent does the bias invalidate the inferences? Precise answers are not available, because the response distribution is unknown. Nevertheless, an important step is to rank different auxiliary vectors for their potential to reduce the bias. The indicator q^2 serves this purpose.

13. References

Bethlehem, J.G. (1988). Reduction of Nonresponse Bias through Regression Estimation. Journal of Official Statistics, 4, 251–260.

Bethlehem, J.G. and Schouten, B. (2004). Nonresponse Adjustment in Household Surveys. Discussion Paper 04007. Voorburg: Statistics Netherlands.

Deville, J.C. (2002). La correction de la non-réponse par calage généralisé. Actes des Journeés de Méthodologie, I.N.S.E.E., Paris [In French].

Folsom, R.E. and Singh, A.C. (2000). The Generalized Exponential Model for Sampling Weight Calibration for Extreme Values, Nonresponse and Poststratification. Proceedings of the American Statistical Association, Survey Research Methods Section, 598–603.

Fuller, W.A. (2002). Regression Estimation for Survey Samples. Survey Methodology, 28, 5–23.

Fuller, W.A., Loughin, M.M., and Baker, H.D. (1994). Regression Weighting in the Presence of Nonresponse with Application to the 1987–1988 Nationwide Food Consumption Survey. Survey Methodology, 20, 75–85.

Harms, T. (2003). Calibration Estimators for Prediction of Dynamics in Panels. Using Longitudinal Patterns to Improve Calibration Estimates about Developments in Panels. Chintex Working Paper no. 14, Federal Statistical Office, Germany.

Kersten, H.M.P. and Bethlehem, J.G. (1984). Exploring and Reducing the Nonresponse Bias by Asking the Basic Question. Statistical Journal of the United Nations, ECE 2, 369–380.

Lundström, S. (1997). Calibration as a Standard Method for Treatment of Nonresponse. Ph.D. Thesis, Stockholm University.

Rizzo, L., Kalton, G., and Brick, J.M. (1996). A Comparison of Some Weighting Adjustment Methods for Panel Nonresponse. Survey Methodology, 22, 43–53.

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

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 Decreasing Response Rates in Statistics Norway. Recommended Practice for Reducing the Effect of Nonresponse. Reports 2006/29. Oslo: Statistics Norway.

Received February 2007
Revised January 2008