Research and Development – Methodology reports from Statistics Sweden 2007:2

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

Carl-Erik Särndal
Sixten Lundström

Statistiska centralbyrån Statistics Sweden

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.

.

Research and Development – Methodology reports from Statistics Sweden 2007:2

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

Carl-Erik Särndal
Sixten Lundström

Statistiska centralbyrån
2007

Research and Development – Methodology reports from Statistics Sweden 2007:2

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

Statistics Sweden
2007

Previous publication	Quantifying the Quality of Macroeconomics Variables 2006:1 Stochastic population projections for Sweden 2006:2 A comparison of disclosure risk measures for tables 2007:1
Producer	Statistics Sweden, Research and Development Department SE-701 89 ÖREBRO + 46 19 17 60 00
Inquiries	Carl-Erik Särndal, +46 19 17 60 43 carl.sarndal@rogers.com Sixten Lundström, + 46 19 17 64 96 sixten.lundstrom@scb.se

Cover Ateljén, SCB

ISSN 1653-7149

URN:NBN:SE:SCB-2007-X103OP0702_pdf (pdf)

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

Preface

Nonresponse occurs in practically all sample surveys. Some decades ago, nonresponse rates were low, by today's standards; they were no major cause for concern. However, survey nonresponse is on the increase in many countries, including Sweden.

As is well known, high nonresponse has a negative impact on the quality of the statistics produced in a survey, unless powerful adjustment procedures can be brought to bear. In this regard, Statistics Sweden is in a comparatively favourable position, because the many administrative registers that are available provide a rich source of auxiliary information.

Statistics Sweden has devoted considerable resources to the study of the nonresponse and its consequences. For a long time, nonresponse rates have been carefully monitored in most of the agency's surveys. Over the past few decades several projects have focused on questions related to survey nonresponse.

The present article by Carl-Erik Särndal and Sixten Lundström, *Assessing Auxiliary Vectors for Control of Nonresponse Bias in the Calibration Estimator*, contributes further insight into the biasing effects of nonresponse. The indicator examined in the article is a useful tool in diagnosing nonresponse bias.

Statistics Sweden, June 2007

Folke Carlsson

Anna Björk

Disclaimer

The series Research and Development – Methodology reports from Statistics Sweden is published by Statistics Sweden and includes results on development work concerning methods and techniques for statistics production. Contents and conclusions in these reports are those of the author(s).

Contents

Abstract

This paper deals with calibration estimation for surveys with nonresponse. Efficient weighting adjustment for unit nonresponse requires powerful auxiliary information. The theory in the paper is inspired by the survey environment in Scandinavia (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 weights in the calibration estimator are computed on information about a specified auxiliary vector. Even with the "best possible" auxiliary vector, some bias remains in the estimator. A close approximation to the remaining bias is presented and analyzed. The relationship between the bias expression and the auxiliary vector in use is a focal point in the article.

The many potential auxiliary variables allow the statistician to compose a wide variety of possible auxiliary vectors. The need arises to compare these vectors to assess their effectiveness for bias reduction. To this end we define and examine an indicator useful for ranking alternative auxiliary vectors in regard to their ability to reduce the bias.

The indicator is computed on the auxiliary vector values for the sampled units, responding and nonresponding. An advantage is its independence of the study variables, of which there are many in a large survey. The properties of the indicator are examined in the theory sections of the paper. The indicator tends, with increasing sample size, to a population analogue, shown to be linked to the bias through an approximately linear relationship. The higher the value of indicator, the more likely it is that the bias will be low, for many study variables.

Empirical studies occupy the final sections of the paper. A synthetic population is constructed and potential auxiliary vectors are ranked with the aid of the indicator. Another empirical illustration illustrates how the indicator is used for selecting auxiliary variables in a large survey at Statistics Sweden.

Keywords: administrative registers, auxiliary information, auxiliary vector, bias indicator, calibration estimator, nonresponse bias, response distribution, response influence.

1. Introduction

When nonresponse occurs in a survey, a pressing objective is to "cleanse" the survey estimates of bias, through an efficient weighting scheme. This paper deals with calibration estimators for surveys with unit nonresponse. The calibrated weights are computed from information carried by an auxiliary vector, more or less powerful. A perfect auxiliary vector would be one that completely eliminates the bias. No such vector can be counted on in practice. 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 to be produced at all in the survey, one must ultimately settle for one auxiliary vector and use it in the computation of calibrated weights and survey estimates.

In practice, a pool of potential auxiliary variables is identified in a preliminary step. The search may involve a matching of different administrative registers. The Scandinavian countries, the Netherlands and several other countries in northern Europe are privileged, equipped as they are with many reliable administrative 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 potential auxiliary variables, a number of different auxiliary vectors can be formed. We need to compare these vectors to assess their effectiveness for bias reduction. Such a bias indicator was proposed on intuitive grounds by Särndal and Lundström (2005). This paper examines the properties of the indicator in further depth and shows its use as a tool for building the auxiliary vector, via, for example, a stepwise forward or a stepwise backward selection of variables, as Sections 9 and 11 illustrate.

It is known that desirable features of an auxiliary vector include the following: (i) it should explain the response pattern; (ii) it should well explain the study variable(s) in the survey, and (iii) it should identify the principal domains of interest in the survey. Särndal and

Lundström (2005) refer to (i) to (iii) as "principles for an auxiliary vector". The emphasis in this paper is on the aspect (i).

This paper is organized as follows: In Section 2 we specify the auxiliary information and the calibration estimator. A close approximation of its bias is given in Section 3. This expression, called nearbias, becomes the focus of attention in the following sections. It depends on (a) the known auxiliary vector values $\mathbf{x}_k$, (b) the unknown response probabilities θ_k, and (c) the unknown study variable values y_k.

If the response probabilities were known, nonresponse bias would cease to be a problem: The inverse response probabilities, $\phi_k = 1/\theta_k$, would provide the weights necessary for unbiased estimation. We call ϕ_k the *response influence* of population unit *k*. It is an unobservable quantity, a latent trait of unit *k*. We produce predicted response influences with the aid of the known auxiliary vector values. This is done in two ways: In Sections 6 and 7, the predicted influences are theoretical values, defined for all *N* population units. Their computable, sample-based counterparts follow in Section 8.

The proposed bias indicator, denoted $\hat{Q}$, is defined in Section 8 as the variance of the predicted influences of the responding units. An intuitive reason why such a variance can serve as an indicator of bias is that a variability in the predicted influences (which are surrogates for the true influences ϕ_k) is desirable to well reflect the unique features of 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 value of the population analogue of $\hat{Q}$ increases.

A computation of $\hat{Q}$ requires the values of the auxiliary vector for the sampled units, respondents as well as nonrespondents (but is independent of the study variable values). The composition of the auxiliary vector becomes critically important. The value of $\hat{Q}$ increases with the number of variables in the vector, and the prospects for reduced bias are improved. Section 9 discusses the uses of $\hat{Q}$ as a diagnostic tool in the search for the "best auxiliary vector", among those are possible in the survey.

A constructed population is used in Section 10 to confirm 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 $\hat{Q}$.

2. Auxiliary information for the calibration estimator

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), Bethlehem and Schouten (2004), Deville (2002), Folsom and Singh (2000), Fuller, Loughin and Baker (1994), Harms (2003), Lundström (1997), Rizzo, Kalton and Brick (1996), Thomsen et al (2006). Some of these authors focus on the calibration approach to estimation, notably Deville (2002), Harms (2003) and Lundström (1997), and so does this paper, where the basic premises are as in the book by Särndal and Lundström (2005).

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

The response set r results when the designated sample s is exposed to an unknown response distribution $q(r|s)$, such that unit k has an unknown response probability θ_k, assumed positive. Refusal, not-at-home or other types of nonresponse may lie behind a failure to record the value y_k of the study variable denoted y, which is allowed to be continuous or categorical. (As an example of the latter, $y_k = 1$ if k has a property of interest, such as "unemployed", and $y_k = 0$ otherwise.) There may be yet other causes for a failure to obtain the desired y-data. Although called 'response probability', θ_k may be viewed more generally as the probability that the value y_k becomes recorded for the unit $k \in s$. With probability $1 - \theta_k$ it goes missing, for whatever reason. Thus recorded data include the value y_k for $k \in r$ and the outcome of the response: $R_k = 1$ for

$k \in r$, $R_k = 0$ for $k \in s - r$. For any realized sample *s*, we assume $E_q(R_k|s) = \theta_k$, where q refers to the response phase.

The use of auxiliary information is essential. 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: Its value 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 $\sum_U \mathbf{x}_k^*$ is allowed to include population counts taken from demographic sources on age group by sex by region. The individual value $\mathbf{x}_k^*$ is assumed known for every $k \in s$ and consequently for every $k \in r$. (If $A \subseteq U$ is a set of units, we write $\sum_A$ for $\sum_{k \in A}$.)

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 of this is 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; the register information transcribed to the sample data file is then material for the $\mathbf{x}_k^\circ$ -vector. A difference between $\mathbf{x}_k^\circ$ and $\mathbf{x}_k^*$ in this paper is that $\sum_U \mathbf{x}_k^*$ is known while $\sum_U \mathbf{x}_k^\circ$ is unknown. Nevertheless, the computable unbiased estimate $\sum_s d_k \mathbf{x}_k^\circ$ is important information input for calibrated weight computation.

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

$$\mathbf{x}_k = \begin{pmatrix} \mathbf{x}_k^* \\ \mathbf{x}_k^\circ \end{pmatrix} \quad ; \quad \mathbf{X} = \begin{pmatrix} \sum_U \mathbf{x}_k^* \\ \sum_s d_k \mathbf{x}_k^\circ \end{pmatrix} \tag{2.1}$$

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

The target of the estimation is the y -total $Y = \sum_U y_k$. Särndal and Lundström (2005) examine the calibration estimator of Y based on the information $\mathbf{X}$ in (2.1). It is given by $\hat{Y}_W = \sum_r w_k y_k$ 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 two objectives: To reduce the nonresponse bias and to reduce the variance of $\hat{Y}_W$. The weights are calibrated to the given information: $\sum_r w_k \mathbf{x}_k = \mathbf{X}$. In this paper we consider vectors $\mathbf{x}_k$ with the following property: There exists a constant vector $\boldsymbol{\mu}$ such that

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

'Constant' means that $\boldsymbol{\mu}$ must not depend on *k*, nor on *s* or on *r*. Condition (2.2) 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) the classification vector used to code J mutually exclusive and exhaustive population groups, $\mathbf{x}_k = \boldsymbol{\gamma}_k = (\gamma_{1k}, ..., \gamma_{jk}, ..., \gamma_{Jk})'$, so that, for $j = 1, 2, ..., J$, $\gamma_{jk} = 1$ if k belongs to group j , and $\gamma_{jk} = 0$ if not;

(3) the combination of (1) and (2), $\mathbf{x}_k = (\boldsymbol{\gamma}'_k, x_k \boldsymbol{\gamma}'_k)'$;

(4) the vector $\mathbf{x}_k$ that codifies two classifications stringed out 'side-by-side', and the dimension of $\mathbf{x}_k$ is $J_1 + J_2 - 1$, where J_1 and J_2 are the respective number of categories, and the 'minus-one' is to avoid a singular matrix in the computation of weights;

(5) the extension of (4) to more than two 'side-by-side' classifications.

In view of (2.2), the calibration estimator is

$$\hat{Y}_W = \sum\nolimits_r w_k y_k = \sum\nolimits_r d_k v_k y_k \tag{2.3}$$

with $d_k = 1/\pi_k$ and, with $\mathbf{X}$ given by (2.1),

$$v_k = \mathbf{X}'(\sum\nolimits_r d_k \mathbf{x}_k \mathbf{x}'_k)^{-1} \mathbf{x}_k \tag{2.4}$$

Despite the "best possible" calibration, residual bias always remains in $\hat{Y}_W$. This 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.

3. Expressions for the remaining bias

The bias of $\hat{Y}_W$ is derived 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 bias of $\hat{Y}_W$, $\text{bias}(\hat{Y}_W) = E_p E_q(\hat{Y}_W|s) - Y$, is intractable, because $\hat{Y}_W$ is non-linear. We focus on the approximation obtained by Taylor expansion, denoted $\text{nearbias}(\hat{Y}_W)$. The approximation is close, even for rather modest sizes of the response set r, as simulations have shown. Although $\text{nearbias}(\hat{Y}_W)$ is unknown, because a function of the whole population, it forms the basis for designing methods to reduce the bias. Särndal and Lundström (2005), chapter 9, derive the following expression for the nearbias:

$$\text{nearbias}(\hat{Y}_W) = (\textstyle\sum_U \mathbf{x}_k)'(\mathbf{B}_{U;\theta} - \mathbf{B}_U) \tag{3.1}$$

where $\mathbf{x}_k = \begin{pmatrix} \mathbf{x}_k^* \\ \mathbf{x}_k^\circ \end{pmatrix}$, $\sum_U \mathbf{x}_k = \begin{pmatrix} \sum_U \mathbf{x}_k^* \\ \sum_U \mathbf{x}_k^\circ \end{pmatrix}$,

$\mathbf{B}_{U;\theta} = \left(\sum_U \theta_k \mathbf{x}_k \mathbf{x}_k'\right)^{-1} \sum_U \theta_k \mathbf{x}_k y_k$, and

$\mathbf{B}_U = (\sum_U \mathbf{x}_k \mathbf{x}_k')^{-1} (\sum_U \mathbf{x}_k y_k)$. Under mild conditions, $(1/\text{N})\left(\text{bias}(\hat{Y}_W) - \text{nearbias}(\hat{Y}_W)\right)$ is of order $n^{-1/2}$, where n is the sample size. The derivation of (3.1) need not be reproduced here. Similar expressions are also given in, or can be derived from, sources such as Bethlehem (1988) and Fuller (2002), although their conditions differ from ours.

To achieve $\text{nearbias}(\hat{Y}_W) = 0$ is a farfetched possibility. It would happen if all θ_k are equal, an unrealistic hope. As Result 4.1 will show, $\text{nearbias}(\hat{Y}_W) = 0$ holds under yet another condition, but it is also one that will almost certainly not hold in practice. No matter how good the auxiliary information, some bias remains; what we can do is to try to reduce it.

A comment on our notation: Several symbols are given two indices, separated by a semicolon. The principle is that 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 ('an unweighted formula'), the second index is suppressed, as in $\mathbf{B}_U$.

In studying the nearbias, we need not specify which variables in $\mathbf{x}_k$, if any, are of the $\mathbf{x}_k^*$ kind and which, if any, are of the $\mathbf{x}_k^\circ$ kind. An auxiliary variable x_k is equally 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, however, that $\sum_U \mathbf{x}_k = \begin{pmatrix} \sum_U \mathbf{x}_k^* \\ \sum_U \mathbf{x}_k^\circ \end{pmatrix}$ in (3.1) differs from the information $\mathbf{X}$ used in computing the weights (2.3), as soon as the joint vector $\mathbf{x}_k$ contains an $\mathbf{x}_k^\circ$-component.

That the nearbias (3.1) involves the difference between $\mathbf{B}_{U;\theta}$ and $\mathbf{B}_U$ emphasizes one of the predicaments with nonresponse: We end up estimating not the desired (unweighted) regression coefficient $\mathbf{B}_U$, but the "tainted" regression coefficient $\mathbf{B}_{U;\theta}$. The difference between the two causes a more or less pronounced bias in $\hat{Y}_W$. An equivalent expression for (3.1) follows easily:

$$\text{nearbias}(\hat{Y}_W) = \sum_U \theta_k M_k e_k \tag{3.2}$$

where $e_k = y_k - \mathbf{x}_k'\mathbf{B}_U$ is the ordinary least squares regression residual and

$$M_k = (\sum_U \mathbf{x}_k)'(\sum_U \theta_k \mathbf{x}_k \mathbf{x}_k')^{-1}\mathbf{x}_k \tag{3.3}$$

We have $\sum_U e_k = 0$ as a consequence of (2.2). In addition to (3.1) and (3.2), we need a third equivalent expression for the nearbias,

$$\text{nearbias}(\hat{Y}_W) = \sum_U (\theta_k M_k - 1) y_k \tag{3.4}$$

To see the equivalence with (3.2), note that

$$\text{nearbias}(\hat{Y}_W) = \sum_U \theta_k M_k y_k - \sum_U \theta_k M_k \mathbf{x}_k' \mathbf{B}_U$$

A development of the last term now leads to (3.4):

$$\sum_U \theta_k M_k \mathbf{x}_k' \mathbf{B}_U = (\sum_U \mathbf{x}_k)'(\sum_U \theta_k \mathbf{x}_k \mathbf{x}_k')^{-1}(\sum_U \theta_k \mathbf{x}_k \mathbf{x}_k' \mathbf{B}_U) =$$

$$\sum_U \mathbf{x}_k' \mathbf{B}_U = \sum_U y_k$$

The quantities M_k, defined by (3.3) for all $k \in U$, are important for diagnosing the bias. We can view M_k as a derived variable, contingent on the values $\mathbf{x}_k$ of the given auxiliary vector and on the (unknown) response probabilities θ_k.

We shall compare alternative $\mathbf{x}_k$-vectors in regard to their capacity to control the bias. As a benchmark we use the 'primitive auxiliary vector', $\mathbf{x}_k = 1$ for all $k \in U$, which gives $\hat{Y}_W = N\,\bar{y}_r = N \sum_r y_k / n_r$, where n_r is the size of the response set *r*. Then $M_k = N / \sum_U \theta_k = 1/\bar{\theta}_U$ for all *k*, and (3.4) reduces to the well-known expression

$$\text{nearbias}(N\,\bar{y}_r) = N(\bar{y}_{U;\theta} - \bar{y}_U) \tag{3.5}$$

where $\bar{y}_{U;\theta} = \sum_U \theta_k y_k / \sum_U \theta_k$ and $\bar{y}_U = \sum_U y_k / N$. When the theta-weighted mean and the unweighted mean differ considerably, $\hat{Y}_W = N\,\bar{y}_r$ has a large nearbias. For example, if large y-value units respond with low probability, there is a considerable negative nearbias. The vector $\mathbf{x}_k = 1$ recognizes no differences among units and is inefficient for nonresponse adjustment.

We use two measures of relative bias. Each depends on three factors: (i) the values $\mathbf{x}_k$ of the auxiliary vector used to compute

$\hat{Y}_W$, (ii) the response probabilities θ_k, and (iii) the values y_k of the study variable. The first measure sets the nearbias in relation to the value of the target of estimation, $Y = N\,\bar{y}_U$:

$$\text{relbias}(\hat{Y}_W) = \frac{\text{nearbias}(\hat{Y}_W)}{N\bar{y}_U} = \frac{\sum_U (\theta_k M_k - 1) y_k}{N\,\bar{y}_U} \tag{3.6}$$

The second measure shows how well a specified vector $\mathbf{x}_k$ succeeds in controlling the bias, when compared with the primitive vector:

$$P = \frac{\text{nearbias}(\hat{Y}_W)}{\text{nearbias}(N\bar{y}_r)} = \frac{\sum_U (\theta_k M_k - 1) y_k}{N(\bar{y}_{U;\theta} - \bar{y}_U)} \tag{3.7}$$

When several candidate $\mathbf{x}_k$ -vectors are compared, the more effective ones will bring comparatively smaller values of both $\text{relbias}(\hat{Y}_W)$ and P.

4. Response influence and zero nearbias

As (3.2) shows, $\text{nearbias}(\hat{Y}_W) = 0$ holds if the residuals $e_k = y_k - \mathbf{x}'_k\mathbf{B}_U$ are zero for all $k \in U$, that is, if $\mathbf{x}_k$ predicts y_k without error, for every population unit. Most large surveys involve many y-variables. To achieve a zero bias for every one of those would require the residuals e_k to be zero for all units as well as for all y-variables. To achieve this is a vain hope. However, if we focus instead on the response distribution, there are conditions under which the nearbias is zero *for every* y-variable.

To see this, 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. Prior to data collection, both ϕ_k and y_k are unknown characteristics of unit $k \in U$. But unlike the y_k, the ϕ_k remain unknown even for observed sample units. If the ϕ_k were known, they would serve as weights for unbiased estimation. For example, $\sum_r d_k\phi_k y_k$ would be unbiased for $Y = \sum_U y_k$, and 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. Still, they are important in the following. Their relation to the M_k is explained in section 6.

An ideal auxiliary vector is one that perfectly explains the influence ϕ_k. More precisely, an ideal vector $\mathbf{x}_k$ is one that meets the following condition:

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

$$\phi_k = 1/\theta_k = \boldsymbol{\lambda}'\mathbf{x}_k \quad \text{for all } \; k \in U \tag{4.1}$$

In a survey, we cannot hope to find an ideal vector $\mathbf{x}_k$. But if one existed and could be used, the nearbias would be completely eliminated. This is the message of the following result given in Särndal and Lundström (2005).

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

Proof
When (4.1) holds, then

$$(\sum\nolimits_U \mathbf{x}_k)'(\sum\nolimits_U \theta_k \mathbf{x}_k \mathbf{x}'_k)^{-1} = \boldsymbol{\lambda}'(\sum\nolimits_U \theta_k \mathbf{x}_k \mathbf{x}'_k)(\sum\nolimits_U \theta_k \mathbf{x}_k \mathbf{x}'_k)^{-1} = \boldsymbol{\lambda}'$$

so $\theta_k M_k = 1$ for all $k \in U$. Hence, by (3.4), nearbias($\hat{Y}_W$) = 0. □

To illustrate, suppose the available information allows a classification of the population units or the sampled units into J mutually exclusive groups. Then $\mathbf{x}_k = (\gamma_{1k},...,\gamma_{Jk})'$, where $\gamma_{jk} = 1$ if k belongs to group j and $\gamma_{jk} = 0$ if not. By Result 4.1, the nearbias is zero for this $\mathbf{x}$-vector if the θ_k are constant within groups, for example, a set of age/sex groups for a population of individuals. Such an assumption is often made in practice. It is a convenient one, but few would believe it to hold true. The remaining bias can be large; better $\mathbf{x}$-vectors are sought.

5. Least squares prediction of the influence

Consider a fixed auxiliary vector $\mathbf{x}_k$, as given in (2.1). The influences $\phi_k = 1/\theta_k$ are unknown and non-observable, even for the units observed in a sample. We can, however, use the auxiliary data $\mathbf{x}_k$ for $k \in s$ to compute predicted influences, which will then serve to obtain the bias indicator. This is done in section 8. To motivate these sample-based predictions, we first examine the population-based predictions of the ϕ_k. Let us determine the vector $\boldsymbol{\lambda}$ so as to minimize a (weighted) sum of the squared differences $\phi_k - \boldsymbol{\lambda}'\mathbf{x}_k$. Theta-weighted sums of squares are convenient here. We determine $\boldsymbol{\lambda}$ to minimize $WSS = \sum_U \theta_k (\phi_k - \boldsymbol{\lambda}'\mathbf{x}_k)^2$, where *WSS* stands for '<u>w</u>eighted <u>s</u>um of <u>s</u>quares'. (It is assumed that not all ϕ_k are equal.) We differentiate *WSS* with respect to $\boldsymbol{\lambda}$ and set the derivative equal to zero to get the estimating equation

$$\sum_U \theta_k (\phi_k - \boldsymbol{\lambda}'\mathbf{x}_k)\mathbf{x}_k' = \mathbf{0}' \tag{5.1}$$

or, equivalently,

$$\boldsymbol{\lambda}'(\sum_U \theta_k \mathbf{x}_k \mathbf{x}_k') = (\sum_U \mathbf{x}_k)' \tag{5.2}$$

If the matrix on the left hand side is non-singular, the solution is

$$\boldsymbol{\lambda}' = \hat{\boldsymbol{\lambda}}_U' = (\sum_U \mathbf{x}_k)'(\sum_U \theta_k \mathbf{x}_k \mathbf{x}_k')^{-1} \tag{5.3}$$

The resulting predicted value of ϕ_k is

$$\hat{\phi}_{Uk} = \hat{\boldsymbol{\lambda}}_U' \mathbf{x}_k = (\sum_U \mathbf{x}_k)'(\sum_U \theta_k \mathbf{x}_k \mathbf{x}_k')^{-1} \mathbf{x}_k = M_k \tag{5.4}$$

The quantities M_k defined by (3.3) reappear here. They were seen earlier to play an important role in the expressions (3.2) and (3.4) for the nearbias.

The minimum value of the criterion *WSS* is

$$WSS_{\min} = \sum_U \theta_k (\phi_k - \hat{\boldsymbol{\lambda}}'_U \mathbf{x}_k)^2 = \sum_U \phi_k - \sum_U M_k \tag{5.5}$$

Since $\boldsymbol{\lambda} = \hat{\boldsymbol{\lambda}}_U$ satisfies the estimating equation (5.2), we have

$$\sum_U (\theta_k M_k - 1)\mathbf{x}'_k = \mathbf{0}' \tag{5.6}$$

Post-multiplying by the constant $\boldsymbol{\mu}$ and using (2.2), we get $\sum_U (\theta_k M_k - 1) = 0$, or

$$\sum_U \theta_k M_k = N \tag{5.7}$$

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 $\sum_U \theta_k (\phi_k - 1/\bar{\theta}_U)^2$. This sum of squares has an orthogonal components decomposition: The vector $\mathbf{x}_k$ yields the predictions $\hat{\phi}_{Uk} = M_k$ given by (3.3), and the equation "total variation = explained variation + residual variation" reads

$$\sum_U \theta_k (\phi_k - 1/\bar{\theta}_U)^2 = \sum_U \theta_k (M_k - 1/\bar{\theta}_U)^2 + \sum_U \theta_k (\phi_k - M_k)^2 \tag{5.8}$$

or, expanding and dividing through by *N*,

$$\bar{\phi}_U - 1/\bar{\theta}_U = (\bar{M}_U - 1/\bar{\theta}_U) \; + \; (\bar{\phi}_U - \bar{M}_U) \tag{5.9}$$

where

$$\bar{\phi}_U = \sum_U \phi_k / N \quad , \quad \bar{M}_U = \sum_U M_k / N \;, \quad \bar{\theta}_U = \sum_U \theta_k / N \tag{5.10}$$

In obtaining (5.9) we have used that $\sum_U \theta_k M_k^2 = \sum_U M_k$.

By definition, the influences satisfy $\phi_k = 1/\theta_k > 1$ for all *k*. This begs the question: Do the predictions satisfy $\hat{\phi}_{Uk} = M_k > 1$ for all *k*? The answer is: Yes, for a great majority but not necessarily all *k*. By the non-negativity of the first term on the right hand side of (5.9), their mean satisfies $\bar{M}_U \geq 1/\bar{\theta}_U > 1$, which does not exclude that a few M_k may fail to exceed unity. This is no major drawback.

6. Moments of the predicted influences

To see how the predictions $\hat{\phi}_{Uk} = M_k$ are related to the nearbias of $\hat{Y}_W = \sum_r w_k y_k$, we need several moments of the M_k: (i) the unweighted mean, $\overline{M}_U = \sum_U M_k / N$, (ii) the theta-weighted mean, $\overline{M}_{U;\theta} = \sum_U \theta_k M_k / \sum_U \theta_k$, (iii) the theta-weighted variance, denoted Q, (iv) the theta-weighted coefficient of variation, denoted H, and (v) the theta-weighted coefficient of correlation between M_k and ϕ_k, denoted $r_{M\phi}$. The quantities Q, H and $r_{M\phi}$ are unknown, because they depend on the unknown response distribution. Sample-based, computable analogues are given in section 8.

Consider first the theta-weighted mean $\overline{M}_{U;\theta}$. By (5.7),

$$\overline{M}_{U;\theta} = \sum_U \theta_k M_k / \sum_U \theta_k = N / \sum_U \theta_k = 1/\overline{\theta}_U \tag{6.1}$$

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

Result 6.1
For any given auxiliary vector $\mathbf{x}_k$, $\overline{M}_U = \sum_U M_k / N$ satisfies

$$\overline{\phi}_U \geq \overline{M}_U \geq \overline{M}_{U;\theta} = 1/\overline{\theta}_U \tag{6.2}$$

where the different means are defined by (5.10). The lower bound on $\overline{M}_U$, $1/\overline{\theta}_U$, occurs for the primitive vector, $\mathbf{x}_k = 1$ for all k. The upper bound on $\overline{M}_U$, $\overline{\phi}_U$, would be attained only for an ideal vector $\mathbf{x}_k$ that meets condition (4.1).

Proof

The part $\overline{M}_U \leq \overline{\phi}_U$ follows from (5.5), because $WSS_{\min} \geq 0$. The part $\overline{M}_{U;\theta} = 1/\overline{\theta}_U \leq \overline{M}_U$ follows by the non-negativity of the first term on the right hand side of (5.8), whereby $0 \leq \sum_U \theta_k (M_k - 1/\overline{\theta}_U)^2 = N(\overline{M}_U - 1/\overline{\theta}_U)$. That the upper and lower bounds hold for the specified $\mathbf{x}_k$-vectors is easily verified.

□

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

$$Q = \frac{1}{\sum_U \theta_k} \sum_U \theta_k (M_k - \overline{M}_{U;\theta})^2 \tag{6.3}$$

The quantity Q is important as the prototype for the bias indicator $\hat{Q}$ in Section 8. Expanding the square and arranging terms we get using (6.1)

$$Q = \frac{\sum_U M_k}{\sum_U \theta_k} - \frac{N^2}{(\sum_U \theta_k)^2} \quad = (1/\overline{\theta}_U)(\overline{M}_U - 1/\overline{\theta}_U) \tag{6.4}$$

Noteworthy properties of Q are: (a) for any given vector $\mathbf{x}_k$, $Q \geq 0$, because Q is a variance, hence non-negative; (b) the minimum value, $Q = 0$, occurs for the primitive vector, $\mathbf{x}_k = 1$ for all $k \in U$; (c) the upper bound on Q, denoted Q_{sup}, would be realized only for a vector $\mathbf{x}_k$ that meets the perfect fit condition (4.1); by Result 6.1 we have

$$Q_{\text{sup}} = \frac{\sum_U \phi_k}{\sum_U \theta_k} - \frac{N^2}{(\sum_U \theta_k)^2} = (1/\overline{\theta}_U)(\overline{\phi}_U - 1/\overline{\theta}_U) \tag{6.5}$$

(d) extending the $\mathbf{x}_k$-vector by adding further x-variables to it will increase the value of Q (or possibly leave it unchanged). The proof of this property relies on the fact that the extended vector produces at least as small a value of the term "explained variation" in (5.8), as compared to the value of that same term for the shorter vector that excludes those additional variables.

Another useful quantity is the coefficient of variation of the M_k, for $k \in U$, given by

$$H = \bar{\theta}_U \sqrt{Q} = \sqrt{\bar{M}_U \bar{\theta}_U - 1} \tag{6.6}$$

That $\bar{\theta}_U \sqrt{Q}$ is a coefficient of variation (standard deviation divided by corresponding mean), follows from the fact that Q is the theta-weighted variance of the M_k, and $\bar{M}_{U;\theta} = 1/\bar{\theta}_U$ is the theta-weighted mean. The upper bound on H is $H_{\sup} = \sqrt{\bar{\phi}_U \bar{\theta}_U - 1}$.

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 (6.1). Noting that $\sum_U \theta_k (\phi_k - \bar{\phi}_{U;\theta})^2 = \bar{\phi}_U - 1/\bar{\theta}_U$, and using (6.3) and (6.4), we get

$$r_{M\phi} = \sqrt{\frac{\bar{M}_U - 1/\bar{\theta}_U}{\bar{\phi}_U - 1/\bar{\theta}_U}} \tag{6.7}$$

The coefficient of non-determination $1 - r_{M\phi}^2$, which satisfies $0 \le 1 - r_{M\phi}^2 \le 1$, has several useful expressions:

$$1 - r_{M\phi}^2 = \frac{\bar{\phi}_U - \bar{M}_U}{\bar{\phi}_U - 1/\bar{\theta}_U} = 1 - \frac{Q}{Q_{\sup}} = 1 - \frac{H^2}{H_{\sup}^2} \tag{6.8}$$

7. Towards a computable indicator of the bias

The nearbias, given by any one of (3.1), (3.2) or (3.4), is expressed in the following result as the sum of a principal term that is linear in Q (and in H^2) and an error term, R, 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

$$\text{nearbias}(\hat{Y}_W) \;=\; N(\bar{y}_{U;\theta} - \bar{y}_U)(1 - r^2_{M\phi}) + R \tag{7.1}$$

where $1 - r^2_{M\phi}$ is given by (6.8) and $R = \sum_U \theta_k M_k E_k$ with

$$E_k = y_k - \bar{y}_U - (\phi_k - \bar{\phi}_U)\frac{\bar{y}_U - \bar{y}_{U;\theta}}{\bar{\phi}_U - 1/\bar{\theta}_U} \tag{7.2}$$

Proof
To unit k belongs the values y_k and ϕ_k. Let α and β be specified constants. Then we can also associate with unit k the unique value $E_k = y_k - \alpha - \beta\phi_k$. Let us fix α and β to be the values that minimize the theta-weighted sum of squares $\sum_U \theta_k(y_k - \alpha - \beta\phi_k)^2$, namely, $\beta = B = \dfrac{\bar{y}_U - \bar{y}_{U;\theta}}{\bar{\phi}_U - 1/\bar{\theta}_U}$ and $\alpha = A = \bar{y}_U - B\bar{\phi}_U$. Now insert $y_k = A + B\phi_k + E_k$ in (3.4) and simplify to get

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

But $\sum_U(\theta_k M_k - 1) = 0$ by (5.7), and $\sum_U E_k = 0$. Then (7.1) follows from (6.8).

□

In a survey we need to compare different possible $\mathbf{x}_k$ -vectors to assess their effectiveness for bias reduction. As a benchmark we can use the primitive auxiliary vector, $\mathbf{x}_k = 1$ for all k, which yields $\hat{Y}_W = N\,\bar{y}_r$ and $\text{nearbias}(N\,\bar{y}_r) = N(\bar{y}_{U;\theta} - \bar{y}_U)$. For any other, more effective vector $\mathbf{x}_k$, Result 7.1 states that the principal term of $\text{nearbias}(\hat{Y}_W)$ 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 (4.1), $\overline{M}_U$ increases toward its upper bound $\bar{\phi}_U$, the fraction $1 - r^2_{M\phi}$ tends to zero, and $\text{nearbias}(\hat{Y}_W)$ approaches zero. Thus the bias may be reduced considerably if steps are taken to strengthen the $\mathbf{x}_k$ -vector.

The remainder term R in (7.1) is not in general zero, but it is indeed zero under any one of several conditions stated in the following result.

Result 7.2
Consider a fixed auxiliary vector $\mathbf{x}_k$. The remainder term $R = \sum_U \theta_k M_k E_k$ in (7.1) 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 the perfect fit condition (4.1); (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 (7.2); (iv) for some constants c_0 and c_1 , $y_k = c_0 + c_1\phi_k$ for $k \in U$.

Here, condition (iii) states that $\mathbf{x}_k$ explains perfectly the variation remaining in y_k after a removal of the dependence on ϕ_k . Condition (iv) states that the variation in y_k is perfectly explained by the influence ϕ_k , a case of "purely non-ignorable nonresponse."

Proof
In case (i), the result follows by noting that $M_k = 1/\bar{\theta}_U$ for all *k*. In case (ii), $\text{nearbias}(\hat{Y}_W)$ is zero by Result 4.1, and the proportion

$1-r^2_{M\phi}$ in (7.1) is also equal to zero, because $\overline{M}_U = \overline{\phi}_U$ by Result 6.1. Hence, $R = 0$. When case (iii) holds, then it follows from (5.6) and (5.7) that $R = 0$. When case (iv) holds, simple algebra shows that $\text{nearbias}(\hat{Y}_W) = N(\overline{y}_{U;\theta} - \overline{y}_U)\,(1-r^2_{M\phi}) = N\,c_1(\overline{M}_U - \overline{\phi}_U)$, hence $R =$ 0.

□

The following Result 7.3 is an immediate consequence of (7.1).

Result 7.3
If the term R in (7.1) is small in comparison with the first term on the right hand side of (7.1) then

$$P = \frac{\text{nearbias}(\hat{Y}_W)}{\text{nearbias}(N\overline{y}_r)} = \frac{\sum_U (\theta_k M_k - 1) y_k}{N(\overline{y}_{U;\theta} - \overline{y}_U)} \approx 1 - r^2_{M\phi} \qquad (7.3)$$

where the coefficient of non-determination $1-r^2_{M\phi}$ has the alternative forms shown in (6.8).

The ratio $P = \text{nearbias}(\hat{Y}_W)/\text{nearbias}(N\overline{y}_r)$ measures how well the given vector $\mathbf{x}_k$ succeeds in controlling the bias, when compared to the primitive vector. This ratio depends on three factors: (i) the values of the $\mathbf{x}_k$-vector, (ii) the response probabilities θ_k, and (iii) the study variable values y_k. The ratio is approximated in (7.3) by $1-r^2_{M\phi}$, which depends on the first two factors but is independent of the y-variable. Thus $1-r^2_{M\phi}$ represents "the part of the nearbias ratio P that is independent of the study variable".

When different $\mathbf{x}_k$-vectors are at our disposal in a survey, we seek one that is likely to be best for controlling the bias of *all* study variables y in the survey. If R is small, formula (7.3) suggests that we should seek an $\mathbf{x}_k$-vector with a large value of Q or of H. If we concentrate on Q, formula (7.3) shows the nearbias to be roughly a linear function of Q, $\text{nearbias}(\hat{Y}_W) \approx C_0 - C_1 Q$, where $C_0 = N(\overline{y}_{U;\theta} - \overline{y}_U)$ and $C_1 = C_0 / Q_{\text{sup}}$ do not depend on $\mathbf{x}_k$. When a certain $\mathbf{x}_k$-vector is replaced by "an improved one", with an accompanying increase in the value of Q, then we expect the absolute value of $\text{nearbias}(\hat{Y}_W)$ to be reduced in a roughly linear manner. Ideally, the chosen vector $\mathbf{x}_k$ should bring about a value of Q near to its upper bound $Q_{\sup}$, because then the nearbias would be near zero for all y-variables in the survey.

Formula (6.3) defines Q as the variance of the predicted influences $\hat{\phi}_{Uk} = M_k$. We conclude that the greater the variance of these predictions, the better the chances that the bias will be small. This rhymes with the intuition that the more the predictions $\hat{\phi}_{Uk}$ can reflect the unique features of the respondents, the better the chances for a small bias.

Neither the variance Q nor the coefficient of variation H are computable in a survey, because they depend on the whole population with its unknown response probabilities. In empirical experiments, such as those reported in Section 10, we can, however, study the relationship between Q and the nearbias. The sample-based counterpart of Q, denoted $\hat{Q}$, is given in Section 8, and Section 9 shows how $\hat{Q}$ can be used as a diagnostic tool for indicating the bias.

8. Sample-based counterparts

The population-based quantities M_k, Q and H have sample-based analogues, m_k, $\hat{Q}$ and $\hat{H}$, which we now define. They can be computed from two sources of input: (i) the vector values $\mathbf{x}_k = \begin{pmatrix} \mathbf{x}_k^* \\ \mathbf{x}_k^\circ \end{pmatrix}$, known for $k \in s$, and (ii) the outcome of the response phase, that is, $R_k = 1$ for $k \in r$ and $R_k = 0$ for $k \in s - r$. They do not depend on the y-values.

Formula (5.4) gives 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 (5.2). The corresponding sample-based estimating equation for $\boldsymbol{\lambda}$ is obtained by substituting unbiased estimates for the population sums in (5.2). These estimates are $\sum_s d_k \mathbf{x}_k$ and $\sum_r d_k \mathbf{x}_k \mathbf{x}'_k$, noting that $E_p\left(E_q(\sum_r d_k \mathbf{x}_k \mathbf{x}'_k | s)\right) = E_p\left(\sum_s d_k \theta_k \mathbf{x}_k \mathbf{x}'_k)\right) = \sum_U \theta_k \mathbf{x}_k \mathbf{x}'_k$. The estimating equation is $\boldsymbol{\lambda}' \left(\sum_r d_k \mathbf{x}_k \mathbf{x}'_k\right) = \left(\sum_s d_k \mathbf{x}_k\right)'$; its solution is $\boldsymbol{\lambda}' = \hat{\boldsymbol{\lambda}}'_s = \left(\sum_s d_k \mathbf{x}_k\right)' \left(\sum_r d_k \mathbf{x}_k \mathbf{x}'_k\right)^{-1}$, supposing the matrix can be inverted. The sample-based prediction of ϕ_k, computable for $k \in s$, is

$$\hat{\phi}_{sk} = \hat{\boldsymbol{\lambda}}'_s \mathbf{x}_k = m_k = \left(\sum_s d_k \mathbf{x}_k\right)' \left(\sum_r d_k \mathbf{x}_k \mathbf{x}'_k\right)^{-1} \mathbf{x}_k \qquad (8.1)$$

We have $\sum_r d_k m_k = \sum_s d_k$, which implies that $\sum_r d_k m_k$ is an unbiased estimate of the population size N, because $\sum_s d_k$ has this property. The quantities m_k are related to (but not in general equal to) the weight factors v_k in the calibration estimator $\hat{Y}_W = \sum_r d_k v_k y_k$ given by (2.3). 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. Two different weighted means of the m_k now become important:

$$\overline{m}_{r;d} = \frac{\sum_r d_k m_k}{\sum_r d_k} = \frac{\sum_s d_k}{\sum_r d_k} \quad ; \quad \overline{m}_{s;d} = \frac{\sum_s d_k m_k}{\sum_s d_k} \tag{8.2}$$

The quantity Q was defined by (6.3) and (6.4) as the (theta-weighted) variance of the predicted influences $\hat{\phi}_{Uk} = M_k$ for $k \in U$. By the same logic, $\hat{Q}$ is defined to be the (design-weighted) variance of the predictions $\hat{\phi}_{sk} = m_k$ for $k \in r$:

$$\hat{Q} = \frac{1}{\sum_r d_k} \sum_r d_k (m_k - \overline{m}_{r;d})^2 \tag{8.3}$$

A simple development gives

$$\hat{Q} \; = \; \frac{W}{\sum_r d_k} - \frac{(\sum_s d_k)^2}{(\sum_r d_k)^2} \; = \overline{m}_{r;d} (\overline{m}_{s;d} - \overline{m}_{r;d}) \tag{8.4}$$

where

$W = \sum_r d_k m_k^2 = \sum_s d_k m_k = (\sum_s d_k \mathbf{x}_k)'(\sum_r d_k \mathbf{x}_k \mathbf{x}_k')^{-1}(\sum_s d_k \mathbf{x}_k)$.

Since $\hat{Q} \geq 0$ and $\overline{m}_{r;d} \geq 1$, it follows that $\overline{m}_{s;d} \geq \overline{m}_{r;d}$. It is useful to remember the interpretation of $\hat{Q}$ as the variance of the sample-based predicted influences. But a familiar line of reasoning allows an alternative interpretation of $\hat{Q}$: Replace each population sum in Q given by (6.4) by its corresponding unbiased estimate: In place of 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, substitute their 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 $\sum_s d_k$. Then $\sum_U M_k = (\sum_U \mathbf{x}_k)'(\sum_U \theta_k \mathbf{x}_k \mathbf{x}_k')^{-1}(\sum_U \mathbf{x}_k)$ in (6.4) becomes replaced by $W = \sum_s d_k m_k$, and we have arrived at (8.4).

Some properties of $\hat{Q}$ are: (a) For any given auxiliary vector $\mathbf{x}_k$, $\hat{Q} \geq 0$, because $\hat{Q}$ is a variance; (b) $\hat{Q} = 0$ for the primitive vector,

$\mathbf{x}_k = 1$ for all k; (c) $\hat{Q} = 0$ when $r = s$, that is, when the response is complete; (d) $\hat{Q} = 0$ if $\overline{\mathbf{x}}_{s;d} = \overline{\mathbf{x}}_{r;d}$; (e) unlike Q given by (6.4), $\hat{Q}$ does not have a specifiable upper bound; (f) for a given $\mathbf{x}_k$, $\hat{Q}$ converges in probability, under mild conditions, to the value Q, because to each population sum in Q corresponds a design unbiased estimate in $\hat{Q}$; (e) the convergence of $\hat{Q}$ to Q may be rather slow, and the sample-to-sample variability of $\hat{Q}$ may be considerable, unless both s and r are rather large sets. For best results, $\hat{Q}$ should be used with the large sample sizes, often more than one thousand units, that are typical of government surveys

The population-based coefficient of variation *H* given by (6.6) has a sample-based analogue, computable on the values m_k for $k \in r$, namely,

$$\hat{H} = \frac{1}{\overline{m}_{r;d}} \sqrt{\hat{Q}} = \sqrt{\frac{\overline{m}_{s;d}}{\overline{m}_{r;d}} - 1}$$

Both $\hat{H}$ and $\hat{Q}$ give the same ranking of $\mathbf{x}_k$-vectors, but one may prefer $\hat{H}$ since it mitigates the tendency present in $\hat{Q}$ to increase with increasing rates of survey nonresponse.

The indicator $\hat{Q}$ was introduced on intuitive grounds in Särndal and Lundström (2005), under the notation *IND1*, and used there to compare different candidate vectors $\mathbf{x}_k$ in regard to their potential for bias reduction. This role of $\hat{Q}$ is further developed in the following sections. A different tool for the selection of *x*-variables, among the many that may be at hand, is developed and illustrated in Bethlehem and Schouten (2004). It has no direct resemblance to $\hat{Q}$, although both attempt to meet the same goal. Rizzo, Kalton and Brick (1996) agree with the two just cited sources in viewing the choice of auxiliary variables as a comparatively more important question than the choice among alternative algorithms for computing the weights once a set of such variables has been fixed.

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 notably in a number of North European countries, where reliable registers of total population provide extensive auxiliary data for surveys on individuals and households. These data bases contain many potential auxiliary variables. Following a preliminary inventory aimed at identifying a set of potential x-variables, a range of possible auxiliary vectors $\mathbf{x}_k$ can be considered. We want to compare those 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 $w_k = d_k v_k$.

How do we compare the various candidate vectors $\mathbf{x}_k$ to assess their capacity to reduce as the bias of $\hat{Y}_W$? (Both types of information, $\mathbf{x}_k^*$ and $\mathbf{x}_k^\circ$, may be present in $\mathbf{x}_k$.) The approximation (7.3) suggests that an increase in $Q = Q(\mathbf{x}_k)$ is accompanied by a roughly linear decrease in the nearbias. The empirical evidence in Section 10 supports this contention. But $Q(\mathbf{x}_k)$ depends on the whole population and must be replaced in practice by $\hat{Q} = \hat{Q}(\mathbf{x}_k)$ given by (8.3) or (8.4).

What assurance do we have that $\hat{Q}(\mathbf{x}_k)$ will guide us correctly to the preferred $\mathbf{x}_k$ -vector? Suppose we compare two possible $\mathbf{x}$-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 $Q(\mathbf{x}_{2k}) \geq Q(\mathbf{x}_{1k})$ by property (d) in Section 6. That is, adding further variables to $\mathbf{x}_{1k}$ increases the value of Q. The same holds for the sample-based counterparts: $\hat{Q}(\mathbf{x}_{2k}) \geq \hat{Q}(\mathbf{x}_{1k})$, for any realized sample s and response set *r*. This still does not guarantee that the bias is smaller for $\mathbf{x}_{2k}$ than for $\mathbf{x}_{1k}$, but (7.3) suggests that this is so. In other words, if *Q* indicates that $\mathbf{x}_{2k}$ is preferred to $\mathbf{x}_{1k}$, then $\hat{Q}$ will agree with this indication, for any realization (*s*, *r*).

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 obtainable by adding further auxiliary variables to $\mathbf{x}_{1k}$. Then $\hat{Q}(\mathbf{x}_{2k}) \geq \hat{Q}(\mathbf{x}_{1k})$ may hold for some realizations (*s*, *r*), but not necessarily for all.

As Section 11 illustrates, $\hat{Q}(\mathbf{x}_k)$ provides a tool for a stepwise selection of x-variables from a pool of J potentially interesting *x*-variables, categorical or continuous. In the first step of a stepwise forward procedure, compute $\hat{Q}(\mathbf{x}_k)$ for each single *x*-variable; retain the one that yields the largest value of $\hat{Q}(\mathbf{x}_k)$. In the second step, compute $\hat{Q}(\mathbf{x}_k)$ for each of the $J-1$ vectors $\mathbf{x}_k$ composed of the variable from step one and one additional *x*-variable; of those, retain the one that yields the highest increase in $\hat{Q}(\mathbf{x}_k)$, and so on, if further steps are deemed necessary. Typically, the consecutive increases in $\hat{Q}(\mathbf{x}_k)$ will taper off, since the best variables enter first. The empirical evidence in Section 11 confirms this pattern.

An alternative is to use $\hat{Q}(\mathbf{x}_k)$ for a stepwise backward deletion of *x*-variables, one at a time: Start with the vector $\mathbf{x}_k$ comprising all J *x*-variables deemed of interest. There are two reasons why one may not wish to retain all variables in that vector: (i) some of the *x*-variables contribute little to the objective of reducing bias, and (ii) inspection may reveal some undesirably large or small weights. Compute first the value of $\hat{Q}(\mathbf{x}_k)$ for the full vector; then compute $\hat{Q}(\mathbf{x}_k)$ for every one of the $J-1$ different vectors with one *x*-variable deleted. The vector that causes the least reduction in $\hat{Q}(\mathbf{x}_k)$ is retained, provided a renewed inspection of the weights is satisfactory. If desired, the stepwise deletion continues: If at any step deletion causes a significant drop in $\hat{Q}(\mathbf{x}_k)$, it is a sign that the *x*-variable in question is important for bias reduction and should not be sacrificed.

10. Empirical study of the relation between the nearbias and the bias indicator

The objective is to study the behaviour of $Q(\mathbf{x}_k)$ and $\hat{Q}(\mathbf{x}_k)$. First, let us examine how well $Q(\mathbf{x}_k)$ succeeds in tracking the value of $\text{nearbias}(\hat{Y}_W(\mathbf{x}_k))$. We compose a number of auxiliary vectors $\mathbf{x}_k$, we compute both $\text{nearbias}(\hat{Y}_W(\mathbf{x}_k))$ and $Q(\mathbf{x}_k)$ for each vector, and observe how these two quantities move together when $\mathbf{x}_k$ changes. By Result 7.1, we expect $Q(\mathbf{x}_k)$ 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. We expect a comparison of two possible vectors $\mathbf{x}_{1k}$ and $\mathbf{x}_{2k}$ to show that when $Q(\mathbf{x}_{2k}) > Q(\mathbf{x}_{1k})$, then $\text{nearbias}(\hat{Y}_W(\mathbf{x}_{2k})) < \text{nearbias}(\hat{Y}_W(\mathbf{x}_{1k}))$, and this regardless of the response distribution, which is unknown in practice.

A study of this kind requires values y_k, $\mathbf{x}_k$ and θ_k specified for $k = 1, 2, ..., N$. We experimented with several constructed populations. The main conclusions were similar. Results are reported here for one population. We used 16 different vectors $\mathbf{x}_k$. To get some representation of different response conditions, we used four different response distributions with specified response probabilities θ_k, $k = 1, 2, ..., N$. For each response distribution, we compute $\text{nearbias}(\hat{Y}_W(\mathbf{x}_k))$ and $Q(\mathbf{x}_k)$ for the 16 different $\mathbf{x}_k$-vectors. We also compute the nearbias ratio $P(\mathbf{x}_k)$ and the coefficient of non-determination $T(\mathbf{x}_k)$, defined respectively as

$$P(\mathbf{x}_k) = \frac{\text{nearbias}(\hat{Y}_{\text{W}}(\mathbf{x}_k))}{N(\bar{y}_{U;\theta} - \bar{y}_U)} = \frac{\sum_U (\theta_k M_k - 1) y_k}{N(\bar{y}_{U;\theta} - \bar{y}_U)} \quad ;$$

$$T(\mathbf{x}_k) = 1 - \frac{Q(\mathbf{x}_k)}{Q_{\text{sup}}} = 1 - r^2_{M\phi}$$

We plot the 16 points $\left(P(\mathbf{x}_k), T(\mathbf{x}_k)\right)$. The primitive vector $\mathbf{x}_k = 1$ gives the point (1,1). The other 15 points lie inside the unit square. If the term R is small, Result 7.3 suggests that the points will align themselves, apart from some scatter, around the diagonal of the unit square, and that a decrease in $T(\mathbf{x}_k)$ is accompanied by a linear decrease in $P(\mathbf{x}_k)$. The results shown later confirm these expectations. (Note that $P(\mathbf{x}_k) = T(\mathbf{x}_k) = 0$ would occur for an $\mathbf{x}_k$-vector that satisfies condition (4.1). Our study has no such ideal vector, but $P(\mathbf{x}_k)$ and $T(\mathbf{x}_k)$ come close to zero for some of our vectors $\mathbf{x}_k$. Note also that if the numerator in the ratio $P(\mathbf{x}_k)$ is near zero, as it is for a very powerful $\mathbf{x}_k$-vector, then $P(\mathbf{x}_k)$ may have a small negative value. This did not occur in the experiment reported here.)

We constructed a population with values $(y_k, x_{1k}, x_{2k}, \theta_k)$ specified for $k = 1, 2, ..., N = 6{,}000$, where x_{1k} and x_{2k} are the values for unit *k* of two continuous auxiliary variables, x_1 and x_2, y_k is the value of the continuous study variable, y, and θ_k is the response probability of k. The 6,000 values (y_k, x_{1k}, x_{2k}) were obtained by the following steps:

Step 1

<u>The continuous auxiliary variable</u> x_1. The values x_{1k}, $k = 1, 2, ..., N = 6000$, 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

<u>The continuous auxiliary variable</u> 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

$$\mu_{x_{2k}|x_{1k}} = \alpha + \beta x_{1k} + K\,h(x_{1k}) \quad ; \quad \sigma^2_{x_{2k}|x_{1k}} = \sigma^2 x_{1k} \qquad (10.1)$$

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 non-linear appearance to the plotted points (x_{2k}, x_{1k}). This was done on purpose, to avoid an 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

$$\mu_{y_k|x_{1k},x_{2k}} = c_0 + m_k \quad ; \quad \sigma^2_{y_k|x_{1k},x_{2k}} = \sigma^2_M\, m_k \quad ; \quad m_k = c_1 x_{1k} + c_2\, x_{2k} \quad (10.2)$$

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 $\sigma^2_M = 2$. (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.

The 16 alternative auxiliary vectors $\mathbf{x}_k$ were constructed by equally many different uses of x_{1k} and x_{2k}. We transformed each variable into a categorical, size-grouped variable, as is often done in practice to gain stability. The 6,000 variable values x_{1k} were size ordered, and eight groups were formed. The first group consists of the units with the 750 largest values x_{1k}, the second group consists of the next 750 units of the size ordering, and so on. The same procedure was used to group the 6,000 values x_{2k}. Each variable was used in four different *group modes*:

- Mode 8G: Used as categorical with the eight size groups, numbered 1 to 8;
- Mode 4G: Used as categorical with four merged size groups;
- Mode 2G: Used as categorical with two merged size groups;
- Mode N: Not used (a merger of all size groups into one).

In mode 8G of variable x_1, unit *k* is assigned the vector value $\boldsymbol{\gamma}_{(x_1;8)k}$, a vector of dimension eight consisting of seven entries "0" and a single entry "1", marking the size group to which *k* belongs. For example, the vector value $\boldsymbol{\gamma}_{(x_1;8)k} = (0,0,0,0,1,0,0,0)'$ states that unit *k* belongs to size group five of the x_1-variable. In the successive mergers that follow, two adjoining groups will define a new group, doubling the group size. We get mode 4G by merging groups 1 and 2, putting the units with the 1,500 largest x_{1k}-values into a first new group, the merger of groups 3 and 4 forms a second new group, and so on. This assigns to unit k the vector value $\boldsymbol{\gamma}_{(x_1;4)k}$. For mode 2G, unit *k* has the vector value $\boldsymbol{\gamma}_{(x_1;2)k}$; for example, $\boldsymbol{\gamma}_{(x_1;2)k} = (1,0)'$ states that *k* belongs to the first of the two new groups, each of size 3,000. Merging causes a successive loss of information. In the ultimate mode N, all groups are merged together, all x_1-information is relinquished, and $\boldsymbol{\gamma}_{(x_1;1)k} = 1$ for all *k*.

The variable x_2 is put to use in the same four modes; 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 $\boldsymbol{\gamma}_{(x_2;1)k} = 1$. Now 4 × 4 = 16 different auxiliary vectors $\mathbf{x}_k$ are formed by using the group information as shown in the following display.

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

The "+" indicates that the $\mathbf{x}_k$ -vector is formed by placing the $\boldsymbol{\gamma}$ -vectors "side by side", the effect being a calibration on the margins. That is, the case 8G+8G 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 N+N disregards all the information and gives 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+N has $\mathbf{x}_k = (\boldsymbol{\gamma}'_{(x_1;4)k}, 1)'_{(-1)} = \boldsymbol{\gamma}_{(x_1;4)k}$, and so on.

We report results for four different response distributions:

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

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

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

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

The constant c was chosen in each option to deliver a mean response probability of $\overline{\theta}_U = \sum_U \theta_k / N = 0.70$. 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. The preceding theory suggests that the linear relationship between $Q(\mathbf{x}_k)$ and the nearbias will prevail for most response distribution, but "unusual" response patterns can always interfere with this property. Options (ii) and (iv), where the response is entirely *y*-variable dependent, might be called "purely non-ignorable". Many other choices could be considered in an experimental study; we expect the principal conclusions to be similar.

Tables 10.1 to 10.4 show $\text{relbias}(\hat{Y}_W(\mathbf{x}_k)) = \text{nearbias}(\hat{Y}_W(\mathbf{x}_k))/(N\,\bar{y}_U)$, and (in parenthesis) the value of $Q(\mathbf{x}_k)$ (both quantities multiplied by 100) for the 16 $\mathbf{x}_k$-vectors. In each table, the case N+N gives $Q(\mathbf{x}_k)$ = 0, and $\text{relbias}(\hat{Y}_W(\mathbf{x}_k))$ is at its highest level. At the other extreme, the case 8G+8G gives the highest value of $Q(\mathbf{x}_k)$ and the lowest value of $\text{relbias}(\hat{Y}_W(\mathbf{x}_k))$. Other cases are intermediate. An increase in the value of $Q(\mathbf{x}_k)$ is in a vast majority of all cases accompanied by a decrease in $\text{relbias}(\hat{Y}_W(\mathbf{x}_k))$, as the preceding theory suggests. To each table corresponds one of the Figures 10.1 to 10.4, showing the plot of the points $(P(\mathbf{x}_k), T(\mathbf{x}_k))$ for the 16 auxiliary vectors $\mathbf{x}_k$.

Table 10.1. Relbias $(\hat{Y}_W(\mathbf{x}_k)$ in % and value of $Q(\mathbf{x}_k)$ 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)

Figure 10.1. 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)

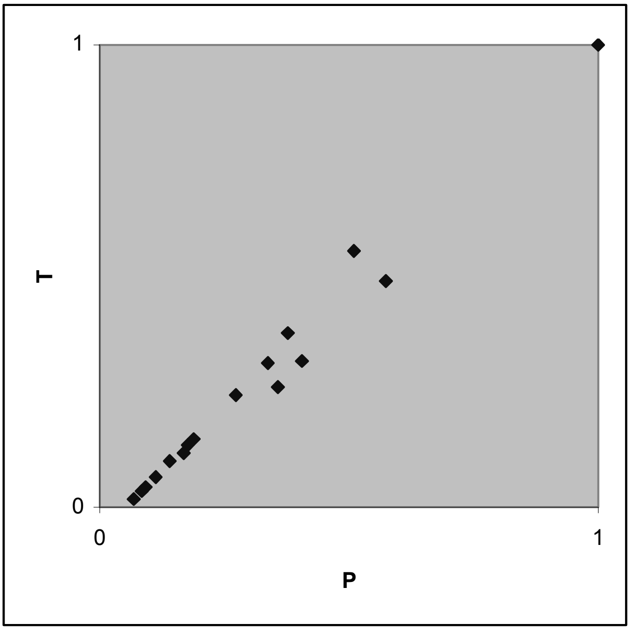


Table 10.2. Relbias $(\hat{Y}_W(\mathbf{x}_k)$ in % and value of $Q(\mathbf{x}_k)$ in % (within parenthesis) for 16 auxiliary vectors $\mathbf{x}_k$. Response distribution IncExp(10+ y)

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)

Figure 10.2. Plot of $(P(\mathbf{x}_k), T(\mathbf{x}_k))$ for 16 auxiliary vectors $\mathbf{x}_k$. Response distribution IncExp(10+ y)

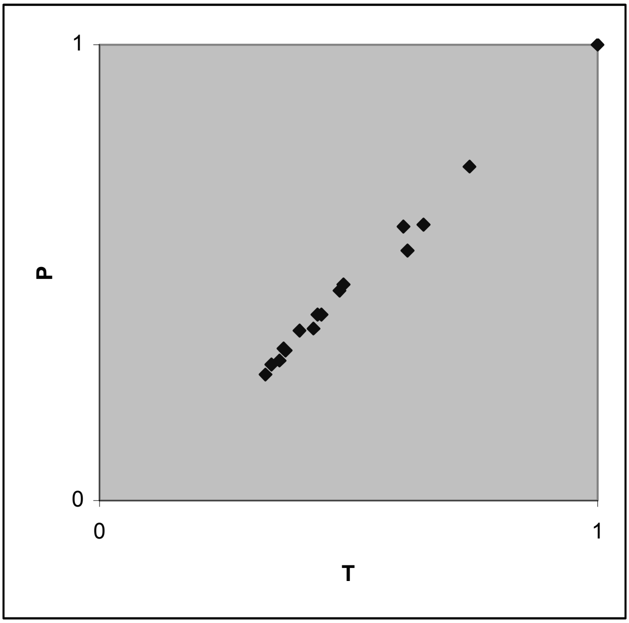


Table 10.3. Relbias $(\hat{Y}_W(\mathbf{x}_k)$ in % and value of $Q(\mathbf{x}_k)$ in % (within parenthesis) for 16 auxiliary vectors $\mathbf{x}_k$. Response distribution DecExp($x_1 + x_2$)

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)

Figure 10.3. Plot of $(P(\mathbf{x}_k), T(\mathbf{x}_k))$ for 16 auxiliary vectors $\mathbf{x}_k$. Response distribution DecExp($x_1 + x_2$)

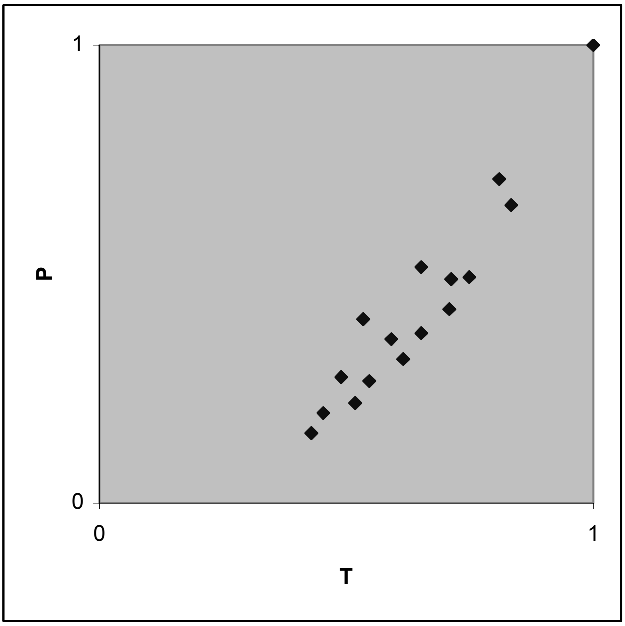
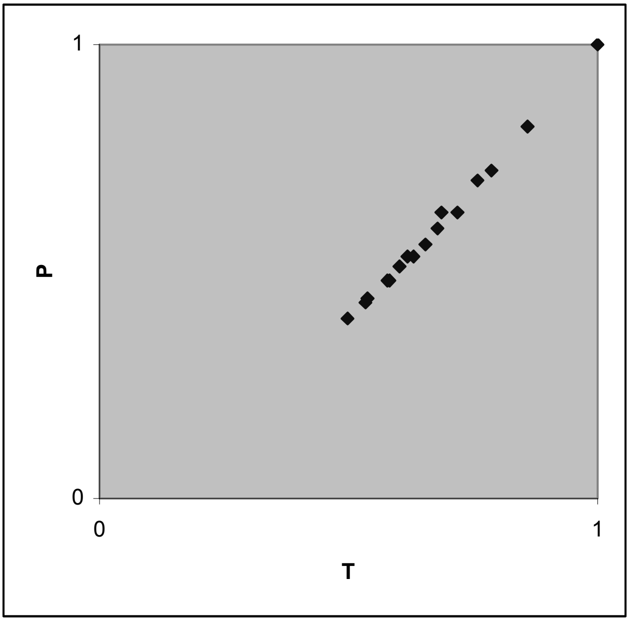


Table 10.4. Relbias $(\hat{Y}_W(\mathbf{x}_k)$ in % and value of $Q(\mathbf{x}_k)$ in % (within parenthesis) for 16 auxiliary vectors $\mathbf{x}_k$. Response distribution DecExp(y)

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)

Figure 10.4. Plot of $(P(\mathbf{x}_k), T(\mathbf{x}_k))$ for 16 auxiliary vectors $\mathbf{x}_k$. Response distribution DecExp(y)



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 yield near-zero nearbias for the x-dependent response distributions. For example, Table 10.1 for IncExp(10+ $x_1 + x_2$) shows a nearbias (in %) decreasing from 13.2 (case N+N) to 0.2 (case 8G+8G). Although the decreasing pattern holds also for the y-dependent response distributions, a difference is that the nearbias does not come close to zero for the best vectors. For example, in Table 2 for IncExp(y) the decrease progresses from 13.1 (case N+N) to 3.6 (case 8G+8G). To use a powerful $\mathbf{x}_k$-vector is important also for non-ignorable nonresponse.
2) **Linear relationship between $T(\mathbf{x}_k)$ and $P(\mathbf{x}_k)$.** The visual impression in all of Figures 10.1 to 10.4 is one of strong linearity. Results 7.1 and 7.3 lead us to expect that as the auxiliary vector $\mathbf{x}_k$ improves, $T(\mathbf{x}_k)$ and $P(\mathbf{x}_k)$ decrease together in a nearly linear fashion. To measure the linearity, we computed the product-moment correlation coefficient, denoted r_{TP}, based on the 16 points $\left(P(\mathbf{x}_k), T(\mathbf{x}_k)\right)$. Table 10.5 shows values of r_{TP} near one for all four response distributions, indicating near perfect linearity. We also computed the Spearman rank correlation coefficient, denoted R_{TP}, based on the 16 points $\left(P(\mathbf{x}_k), T(\mathbf{x}_k)\right)$. Table 10.5 shows that R_{TP} is also near one for all four response distributions, so for this population, $T(\mathbf{x}_k)$ gives an almost perfect ranking the $\mathbf{x}_k$-vectors. It follows that $Q(\mathbf{x}_k)$ has the same favourable property, because $Q(\mathbf{x}_k)$ is a linear function of $T(\mathbf{x}_k)$.
3) **The size of the remainder term** R. Result 7.3 leads us to expect the points $\left(P(\mathbf{x}_k), T(\mathbf{x}_k)\right)$ aligned, except for some scatter, around the diagonal of the unit square. This assumes that the term R in 7.3 is comparatively small. The diagonal pattern is strong in Figures 10.1, 10.2 and 10.4, but is somewhat less pronounced in Figure 10.3 for DecExp($x_1 + x_2$), although the linear relationship remains strong. Figure 10.3 exemplifies a case where the remainder term R in (7.1) is not negligible, compared to the principal term.

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

Table 10.5. Product-moment correlation coefficient r_{TP}, and Spearman rank correlation coefficient R_{TP}, for four response distributions; both correlations computed on 16 points $(P(\mathbf{x}_k), T(\mathbf{x}_k))$, corresponding to 16 vectors $\mathbf{x}_k$

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

Tables 10.1 to 10.4 and Figures 10.1 to 10.4 show, for the population used here, that, if computable, $Q(\mathbf{x}_k)$ would be a good instrument for ranking the possible $\mathbf{x}_k$-vectors. In an actual survey, we must rely on the sample-based analogue $\hat{Q}(\mathbf{x}_k)$. This begs the question: How well does $\hat{Q}(\mathbf{x}_k)$ succeed in ranking the $\mathbf{x}_k$-vectors?

For row-wise and for column-wise comparisons in Tables 10.1 to 10.4, the $\mathbf{x}_k$-vectors are in a hierarchical relationship, in the sense of Section 9. We know that if the vectors $\mathbf{x}_{1k}$ and $\mathbf{x}_{2k}$ belong in the same table row or in the same table column, and $Q(\mathbf{x}_{2k}) \geq Q(\mathbf{x}_{1k})$, then $\hat{Q}(\mathbf{x}_{2k}) \geq \hat{Q}(\mathbf{x}_{1k})$ 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 would show that $\hat{Q}(\mathbf{x}_{2k}) \geq \hat{Q}(\mathbf{x}_{1k})$ for any (*s*, *r*), confirming correctly that 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. Then they are not hierarchically related, and $\hat{Q}(\mathbf{x}_{2k}) \geq \hat{Q}(\mathbf{x}_{1k})$ will hold for some but most likely not for all outcomes (*s*, *r*). Especially if the difference $\text{nearbias}(\hat{Y}_W(\mathbf{x}_{1k})) - \text{nearbias}(\hat{Y}_W(\mathbf{x}_{2k}))$ is considerable (in absolute value), we would like to see that $\hat{Q}(\mathbf{x}_{2k}) \geq \hat{Q}(\mathbf{x}_{1k})$ holds in a high proportion of all outcomes (*s*, *r*), because it means $\hat{Q}$ will with high probability lead to the correct decision to base the estimator 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". For several pairs of vectors $(\mathbf{x}_{1k}, \mathbf{x}_{2k})$, we computed the proportion of the outcomes (*s*, *r*) for which $\hat{Q}(\mathbf{x}_{2k}) - \hat{Q}(\mathbf{x}_{1k})$ has the desired sign. It is of particular interest to compare vectors whose nearbias values are fairly close, so that $\hat{Q}$ is put to a difficult test. This is the case the following examples (i), (ii) and (iii), where the compared vectors are non-hierarchical.

(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 10.1, $\text{nearbias}(\hat{Y}_W(\mathbf{x}_{2k})) = 1.5 < 1.8 = \text{nearbias}(\hat{Y}_W(\mathbf{x}_{1k}))$ and, correspondingly, $Q(\mathbf{x}_{2k}) = 8.5 > 8.4 = Q(\mathbf{x}_{1k})$. Thus $\mathbf{x}_{2k}$ is the slightly better vector in terms of nearbias. The *Q*-values confirm this order of preference. For the $\hat{Q}$-values, the desired ordering, $\hat{Q}(\mathbf{x}_{2k}) > \hat{Q}(\mathbf{x}_{1k})$, occurred in a substantial proportion of all outcomes (*s*, *r*), namely, 70.7%.

(ii) Comparison of 2G+2G (vector $\mathbf{x}_{1k}$) with 4G+N (vector $\mathbf{x}_{2k}$) for DecExp(*y*). By Table 10.4, nearbias($\hat{Y}_W(\mathbf{x}_{2k})$) = -12.2 and nearbias($\hat{Y}_W(\mathbf{x}_{1k})$) = -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*-values: $Q(\mathbf{x}_{2k}) = 8.0 > 7.0 = Q(\mathbf{x}_{1k})$. The desired ordering $\hat{Q}(\mathbf{x}_{2k}) > \hat{Q}(\mathbf{x}_{1k})$ occurred in 78.1% of the 5,000 outcomes (*s*, *r*).

(iii) Comparison of N+2G (vector $\mathbf{x}_{1k}$) with 2G+N (vector $\mathbf{x}_{2k}$) for DecExp($x_1 + x_2$). By Table 10.3, nearbias($\hat{Y}_W(\mathbf{x}_{2k})$) = -11.5 and nearbias($\hat{Y}_W(\mathbf{x}_{1k})$) = -12.6. Hence, $\mathbf{x}_{2k}$ is the somewhat better vector, judging by the absolute value of nearbias. But this is one of the rare comparisons where the *Q*-values do not confirm the nearbias order: We have $Q(\mathbf{x}_{2k}) = 5.8 < 6.7 = Q(\mathbf{x}_{1k})$. Not surprisingly then, the desired ordering $\hat{Q}(\mathbf{x}_{2k}) > \hat{Q}(\mathbf{x}_{1k})$ 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 Total Population (RTP). The survey objective was to measure trends in certain types of crimes, in particular crimes against the person. It will 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 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 data base, LISA. All auxiliary variables are categorical. Groups were formed for a variable which is by nature continuous. Variables obtained from LISA were transcribed only to the sample data base, so they are of the $\mathbf{x}_k^{\circ}$ type defined in Section 2.

With this survey as a background, we illustrate the use of $\hat{Q}$ for a stepwise selection of variables, in the manner explained in Section 9.

In each step, the auxiliary vector $\mathbf{x}_k$ expands by one additional categorical variable, the one that yields the largest increase in $\hat{Q}$ at that point. A new variable joins already entered variables in the "side-by-side" (or "+") manner. Table 11.1 shows the variables entered into $\mathbf{x}_k$ in the first ten forward selection steps. Country of birth, entered in step one, is the dichotomous variable indicating Scandinavian born or not. Age group and sex, adjustment variables "by routine" in many surveys, do qualify for inclusion here, in steps 3 and 4. The pool of potential auxiliary variables included a number of others, not shown in the table.

Table 11.1 also shows the number of groups for each categorical variable, and the successive values of $1000 \times \hat{Q}$. Not unexpectedly, the increases in $\hat{Q}$ taper off after a few steps. This suggests that there would be little point, for bias reduction, to use more than the first six x-variables, and perhaps the first four would suffice.

In the survey, estimates were produced for many categorical study variables, as totals or as proportions. In this context 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 crime in some form. We thus have $Y = \sum_U y_k$, where $y_k = 1$ if person k has the specific property and $y_k = 0$ if not. The bias remaining in the finally produced count estimates remains unknown. But we can follow the stepwise evolution of the estimates. For a selected set of study variables we computed the estimated count at each step in Table 11.1. That is, we computed $\hat{Y}_W = \sum_r w_k y_k$ 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 at step 0 was computed without any $\mathbf{x}_k$-vector by direct expansion within strata, $\hat{Y} = \sum_{h=1}^{H} N_h \bar{y}_{r_h}$, where $\bar{y}_{r_h}$ is the mean response in stratum h.

Some count estimates changed by two or more percentage points in the progression from step 0 to step 6. This must be considered a large change for this survey; nonresponse has a strong impact. We have no way to guarantee that the count estimate in step 6 is more accurate (less biased) than the one in step 0, but theory leads to expect so. A typical pattern was that the greatest change in the estimate occurred in passing 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 development over the successive steps of the value $\hat{Q}$, as shown in Table 11.1. For count variables not affected much by nonresponse, the changes were small in all steps.

Table 11.1. National Crime Victim and Security Study; stepwise forward selection of variables for the auxiliary vector

Step	Auxiliary variable entering	Number of groups	Value of $1000\times\hat{Q}$
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	Martial 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

12. Concluding comment

This paper suggests to use the indicator $\hat{Q}$ as a tool for building the auxiliary vector for the final calibrated weights. The bias in the final estimates remains unknown. We do not resolve age-old questions such as: What is the size of the bias? How large is the squared bias component of the Mean Squared Error? Definite answers are impossible, because the response distribution is unknown. An important step that can be taken, and is taken in this paper, is to rank different auxiliary vectors for their potential to reduce the bias.

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.

Folsom, R.E. and Singh, A.C. (2000). *The generalized exponential model for sampling weight calibration for extreme values, nonresponse and poststratification*. American Statistical Association, Proceedings 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 noresponse 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 Journal, 22, 43-53.

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

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

ISSN 1653-7149

Publication services:
E-mail: publ@scb.se, phone: +46 19 17 68 00, fax: +46 19 17 64 44. Address: SE-701 89 Örebro.
Information and Library: E-mail: information@scb.se, phone: +46 8 506 948 01, fax: +46 8 506 948 99.
Over-the-counter sales: Statistics Sweden, Library, Karlavägen 100, Stockholm, Sweden.

www.scb.se