Research and Development – Methodology reports from Statistics Sweden 2011:1

Statistiska centralbyrån Statistics Sweden

Three Factors to Signal Nonresponse Bias – With applications to Categorical Auxiliary Variables

Carl-Erik Särndal

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

Previous publication:

2006:1 Quantifying the quality of macroeconomic variables

2006:2 Stochastic population projections for Sweden

2007:1 Jämförelse av röjanderiskmått för tabeller

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

2007:3 Kartläggning av felkällor för bättre aktualitet

2008:1 Optimalt antal kontaktförsök i en telefonundersökning

2009:1 Design for estimation: Identifying auxiliary vectors to reduce nonresponse bias

2009:2 Demographic methods for the statistical office

Research and Development – Methodology reports from Statistics Sweden 2011:1

Three Factors to Signal Nonresponse Bias – With applications to Categorical Variables

Carl-Erik Särndal

Statistiska centralbyrån
2011

Research and Development – Methodology reports from Statistics Sweden 2011:1

Three Factors to Signal Nonresponse Bias – With applications to Categorical Variables

Statistics Sweden
2011

Producer

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

Enquiries

Thomas Laitila, +46 19 17 62 18
thomas.laitila@scb.se

Cover Ateljén, SCB

ISSN 1653-7149 (online)
URN:NBN:SE:SCB-2011-X103BR1101_pdf (pdf)

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

Preface

Nonresponse is an increasing problem threatening the validity and thrust in sample survey statistics. The problem has to be addressed in every part of a survey design in order to maintain and increase the usefulness of statistics. One important counteraction is improved utilization of available auxiliary information. Such information has traditionally been used in the design step of a survey, e.g. for stratification, and in the estimation step for reducing estimator variance. Using the calibration estimator, auxiliary information can also be a powerful tool for reducing the bias introduced in estimates due to nonresponse, which has been shown by the author in earlier contributions. One delicate problem for application of the calibration estimator is the choice of auxiliary information. This paper contributes with new results and insights on this problem and provides with new practical tools improving the ability to produce valid statistics under non-response.

Statistics Sweden, March 2011

Lilli Japec

Contents

Abstract

The methods for nonresponse bias adjustment weighting used in sample surveys at Statistics Sweden rely on two features: (1) the availability of many potential auxiliary variables derived from several administrative registers; (2) the use of a bias indicator to identify the auxiliary variables likely to be the most efficient ones. As a consequence of (1), many potential auxiliary vectors can be constructed. Every choice of vector defines a calibration estimator. Its remaining bias depends on the strength of the auxiliary vector. The theoretical basis of the bias indicator is explained. It serves to compare different auxiliary vectors, in order to settle on one that can be used in statistics production with good prospects of significant reduction of bias in most of the survey estimates. We express the indicator in linear algebra terms as a product of three factors, shown to reflect three familiar statistical concepts. We focus on the important case of categorical auxiliary variables, each defined in terms or two or more properties or traits, as when "Age" is defined by the traits "Young", "Middle aged" and "Elderly". Together, the available auxiliary variables represent a considerable number of predefined traits. An examination the bias indicator and its three factors brings the insight that the auxiliary vector should not necessarily contain all of the available traits; some may be insignificant or even harmful for the objective of bias reduction. One is led to a selection of influential traits, rather than to a selection of entire categorical variables. We illustrate this by numerical examples. We outline a stepwise forward selection procedure for the search for influential traits.

Key words: Calibration, nonresponse adjustment, nonresponse bias, auxiliary variables, administrative registers, bias indicator.

1 The calibration estimator for a survey with nonresponse

The literature is rich in contributions that examine different aspects of estimation in the presence of survey nonresponse. Examples during the last decade include Beaumont (2005); Crouse and Kott (2004), Deville (2002), Kott (2006, 2008). A central question is the reduction of the bias that the nonresponse causes in the estimates. Increased variance can also an issue; the balance between variance increase and bias reduction is considered for example in Little and Vartivarian (2005).

This paper is devoted to a study of nonresponse bias. We wish to reduce that bias as much as possible; the variance aspect is not considered. A justification is that the squared bias is often the dominant component of the Mean Squared Error. We assume that the sample size is quite large, as is typically the case in government surveys; consequently, variance is quite low.

The bias cannot be estimated. Instead we need methods capable of signaling when an effective reduction of the unknown bias has taken place, but without assurance that the bias is reduced to near-zero levels. This point of view is held in recent articles. The issue is one of selecting auxiliary variables likely to be effective for reducing bias. Särndal and Lundström (2005) propose two bias indicators; extensions of that work is reported in Särndal and Lundström (2008, 2010). An alternative bias indicator, with somewhat different motivation and derivation but with the same general purpose, is proposed in Schouten (2007).

It is traditionally argued that the chosen auxiliary vector should meet the objectives (i) to explain the nonresponse mechanism and more particularly its (unknown) response probabilities, and/or (ii) to explain the study variable y. These are ideals, never satisfied in practice. The best one can hope for is a partial fulfillment of one or the other objective. Moreover, the two objectives interact. An efficient nonresponse adjustment requires both objectives to be satisfied to a significant extent, not just one of them. This fact becomes clear from the analysis in this paper. We propose a

statistical indicator, computable on the data for respondents, helpful in showing that partial fulfillment is achieved.

We work in an environment where many alternative auxiliary vectors $\mathbf{x}_k$ can be constructed. The objective is to build the vector $\mathbf{x}_k$ from a supply of auxiliary variables so that prospects are good for small bias in the survey estimates. To guide the process we use the *bias indicator* Δ_A introduced in Section 3. The different forms of auxiliary information are discussed in Section 2. The indicator Δ_A has a telling expression in terms of three simple factors, as explained in Section 4. Sections 5, 6 and 7 prove the statistical properties claimed for each factor. Sections 8 to 11 focus on an application that is especially important in statistical agencies: The case of categorical auxiliary variables. It is emphasized that all the categories, or traits, that define an available categorical variable such as "Age" need not (and ordinarily should not) be retained. Hence we outline in Sections 8 to 11 a process that focuses on the selection of influential traits, rather than on a selection of complete categorical auxiliary variables.

We consider a finite population $U = \{1,2,...,k,...,N\}$. A probability sample s is drawn. Nonresponse occurs: a response set r is realized as a subset of s. We have $U \supset s \supset r$. The values y_k of the study variable *y* are observed only for the units $k \in r$. Those data on *y*, together with auxiliary data, form the material for estimating the population *y*-total $Y = \sum_U y_k$.

The sample s is drawn with a 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 (design-weighted) response rate is

$$P = \sum_r d_k / \sum_s d_k \qquad (1.1)$$

The auxiliary vector value $\mathbf{x}_k = (x_{1k},...,x_{jk},...,x_{Jk})'$ is available for $k \in s$, where x_{jk} is the value for unit *k* of the *j*:th auxiliary variable, x_j. We examine the calibration estimator given by

$$\tilde{Y}_{CAL} = \sum_r d_k m_k y_k \quad ; \quad m_k = (\sum_s d_k \mathbf{x}_k)'(\sum_r d_k \mathbf{x}_k \mathbf{x}_k')^{-1}\mathbf{x}_k \qquad (1.2)$$

The adjustment factor m_k is computable for $k \in s$, but used in the estimator only for $k \in r$. The weights are calibrated to fulfill

$$\sum_r d_k m_k \mathbf{x}_k = \sum_s d_k \mathbf{x}_k \tag{1.3}$$

Equivalently we can write $\tilde{\tilde{Y}}_{CAL} = (\sum_s d_k \mathbf{x}_k)' \mathbf{B}_\mathbf{x}$ with

$$\mathbf{B}_\mathbf{x} = (\sum_r d_k \mathbf{x}_k \mathbf{x}_k')^{-1} (\sum_r d_k \mathbf{x}_k y_k) \tag{1.4}$$

which can be seen as the result of a weighted least squares regression fit: The vector $\mathbf{B}$ that minimizes $\sum_r d_k (y_k - \mathbf{x}_k' \mathbf{B})^2$ is $\mathbf{B} = \mathbf{B}_\mathbf{x}$. Each vector specification $\mathbf{x}_k$ generates a different calibration estimator $\tilde{\tilde{Y}}_{CAL}$. However, $\tilde{\tilde{Y}}_{CAL}$ is not without bias, not even for the best among the available choices of $\mathbf{x}_k$.

2 The auxiliary information

The availability of potent auxiliary variables varies between countries. Surveys in the Scandinavian countries rely on rich sources of auxiliary data, derived from numerous administrative registers. Auxiliary information exists at two levels: at the population level, transmitted by a vector denoted $\mathbf{x}_k^*$, and/or at the sample level, transmitted by a vector $\mathbf{x}_k^\circ$. Both are known vector values for $k \in s$, that is, for respondents as well as for nonrespondents. The population total $\sum_U \mathbf{x}_k^*$ is known; by contrast, $\sum_U \mathbf{x}_k^\circ$ is unknown but estimated without bias by $\sum_s d_k \mathbf{x}_k^\circ$. The auxiliary vector is $\mathbf{x}_k = \begin{pmatrix} \mathbf{x}_k^* \\ \mathbf{x}_k^\circ \end{pmatrix}$. Behind the estimator $\tilde{\tilde{Y}}_{CAL}$ given by (1.2) lies the calibration equation $\sum_r d_k m_k \mathbf{x}_k = \begin{pmatrix} \sum_s d_k x_k^* \\ \sum_s d_k \mathbf{x}_k^\circ \end{pmatrix}$. In order to benefit from the potential for reduced variance when $\sum_U \mathbf{x}_k^*$ is a known population total, one can alternatively determine calibrated weights w_k to fulfill instead $\sum_r w_k \mathbf{x}_k = \mathbf{X} = \begin{pmatrix} \sum_U \mathbf{x}_k^* \\ \sum_s d_k \mathbf{x}_k^\circ \end{pmatrix}$. This yields the estimator $\hat{Y}_{CAL}$ $= \sum_r w_k y_k$ with weights $w_k = d_k \{\mathbf{X}'(\sum_r d_k \mathbf{x}_k \mathbf{x}_k')^{-1} \mathbf{x}_k\}$. They satisfy $\sum_r w_k \mathbf{x}_k^* = \sum_U \mathbf{x}_k^*$ (a known total) and $\sum_r w_k \mathbf{x}_k^\circ = \sum_s d_k \mathbf{x}_k^\circ$ (an unbiasedly estimated total).

As is known (see for example Särndal and Lundström (2005)), $\tilde{\tilde{Y}}_{CAL}$ given by (1.2) and $\hat{Y}_{CAL}$ have the same bias to first order approximation. When the objective is a study of bias, as in this article, we are indifferent in the choice between $\tilde{\tilde{Y}}_{CAL}$ and $\hat{Y}_{CAL}$. We work with the former; the weights are then calibrated to the level of the sample.

We use vectors $\mathbf{x}_k$ such that, for some constant vector $\boldsymbol{\mu} \neq \mathbf{0}$,

$$\boldsymbol{\mu}'\mathbf{x}_k = 1 \text{ for all } k \in U \tag{2.1}$$

It is not a severe restriction. Most vectors $\mathbf{x}_k$ useful in practice are of this kind. For example, if $\mathbf{x}_k = (1, x_k)'$, where x_k is a continuous variable value, then take $\boldsymbol{\mu} = (1, 0)'$; if $\mathbf{x}_k = (0,...,1,...,0)'$, where the one and only "1" codes class membership of k, then take $\boldsymbol{\mu} = (1,...,1,...,1)'$.

We gauge $\tilde{Y}_{CAL}$ against two extremes: One is a "worst possible scenario", arising for the primitive **x**-vector, $\mathbf{x}_k = 1$, the other is the unrealized ideal of full response, admitting unbiased estimation.

The primitive vector, $\mathbf{x}_k = 1$ for all k, gives $m_k = 1/P$ for all k, so $\tilde{Y}_{CAL}$ becomes the expansion estimator

$$\tilde{Y}_{EXP} = (1/P)\sum_r d_k y_k = \hat{N}\,\bar{y}_{r;d} \tag{2.2}$$

where $\hat{N} = \sum_s d_k$, which is unbiased for N. The bias of $\tilde{Y}_{EXP}$ can be large, compared with an alternative $\tilde{Y}_{CAL}$ based on a much stronger auxiliary vector. (If the population size N were known and to replace $\hat{N}$ in (2.2), the asymptotic bias would be the same.)

A comment on the notation: When needed for clarity and emphasis, means and other quantities are given two indices separated by a semi-colon. The first shows the set of units in the summation(s), the second, following the semi-colon, shows the weighting used, as in $\bar{y}_{r;d} = \sum_r d_k y_k / \sum_r d_k$. If the weighting is uniform, the second index is dropped, as in $\bar{y}_U = \sum_U y_k / N$.

3 The bias indicator

We consider $\tilde{Y}_{CAL}$ and $\tilde{Y}_{EXP}$ given by (1.2) and (2.2) and form the statistic $\Delta_A = (\tilde{Y}_{EXP} - \tilde{Y}_{CAL}) / \hat{N}$, called the *bias indicator*. It is used as a tool to compare different possible vectors $\mathbf{x}_k$ for $\tilde{Y}_{CAL}$. Because $\tilde{Y}_{CAL} / \hat{N} = \tilde{Y}_{EXP} / \hat{N} - \Delta_A$, we interpret Δ_A as the distance travelled from the initial crude mean estimate, $\hat{\bar{y}}_{U,EXP} = \tilde{Y}_{EXP} / \hat{N}$, to a better alternative, $\hat{\bar{y}}_{U,CAL} = \tilde{Y}_{CAL} / \hat{N}$, likely to have less nonresponse bias because based on a more powerful auxiliary vector than the trivial $\mathbf{x}_k = 1$. Given the data $\mathbf{x}_k$ for $k \in s$ and y_k for $k \in r$, Δ_A can be routinely computed, along with other summary survey results, such as the nonresponse rate *P* given by (1.1) and the coefficient of determination for the regression of *y* on **x**, denoted later in the paper as $R^2_{y\mathbf{x}}$.

The case of full response (where y_k is available for all $k \in s$) represents an unrealized ideal that makes unbiased estimation possible. One possibility for full response is the unbiased Horvitz-Thompson estimator

$$\tilde{Y}_{FUL} = \sum_s d_k y_k = \hat{N}\, \bar{y}_{s;d} \tag{3.1}$$

In the presence of auxiliary information, a more variance efficient (and nearly unbiased) alternative for full response is $\tilde{Y}^*_{FUL} = \sum_s w_k y_k$, where $w_k = d_k \{\mathbf{X}'(\sum_s d_k \mathbf{x}_k \mathbf{x}_k')^{-1} \mathbf{x}_k\}$, with $\mathbf{x}_k = \mathbf{x}^*_k$ (such that $\boldsymbol{\mu}' \mathbf{x}^*_k = 1$ for all k) and a known population total $\mathbf{X} = \sum_U \mathbf{x}^*_k$. As long as $\tilde{Y}_{FUL}$ represents an essentially unbiased full response estimator, we need not further specify its form in this paper, where the main objective is the study of bias.

For an alternative perspective, we view the three estimates $\tilde{Y}_{CAL}$, $\tilde{Y}_{EXP}$ and $\tilde{Y}_{FUL}$, computed for a given outcome (s,r), as three points on a horizontal axis, and Δ_A as one out of three distances of interest: $\Delta_T = (\tilde{Y}_{EXP} - \tilde{Y}_{FUL})/\hat{N}$, which would be an estimate of bias under the worst possible scenario $\mathbf{x}_k = 1$, $\Delta_R = (\tilde{Y}_{CAL} - \tilde{Y}_{FUL})/\hat{N}$, which would be an estimate of the bias that still remains in $\tilde{Y}_{CAL}$ even with a better choice of $\mathbf{x}_k$, and $\Delta_A = (\tilde{Y}_{EXP} - \tilde{Y}_{CAL})/\hat{N}$, which is computable, in contrast to Δ_T and Δ_R. Then $\Delta_R = \Delta_T - \Delta_A$, where the index *T* suggests "total", *A* "accomplished" or "accounted for", and *R* "remainder". Given (s,r), Δ_T is an unknown (positive or negative) constant, not affected by the choice of $\mathbf{x}_k$. For a succession of improved $\mathbf{x}_k$-vectors, $\tilde{Y}_{CAL}$ will distance itself (in the positive or the negative direction) from $\tilde{Y}_{EXP}$; $|\Delta_A|$ increases. For a highly effective vector $\mathbf{x}_k$, $\tilde{Y}_{CAL}$ will come near $\tilde{Y}_{FUL}$, leaving a small remainder Δ_R. We note that $\Delta_R = 0$ for the (never existing) perfect relationship $y_k = \mathbf{x}_k'\boldsymbol{\beta}$ for all $k \in r$. There is, however, no guarantee that all choices of $\mathbf{x}_k$ will create an $\tilde{Y}_{CAL}$ lying between $\tilde{Y}_{EXP}$ and $\tilde{Y}_{FUL}$. Whether it does or not is unknown to the statistician/analyst, as is the size of Δ_R, so Δ_A is an indicator of bias, not a quantifier of bias.

The statistic Δ_A is suggested here as one possible tool in comparing potential **x**-vectors and for identifying efficient auxiliary variables for the vector. If $|\Delta_A|$ is greater for the vector $\mathbf{x}_k = \mathbf{x}_{1k}$ than for the alternative $\mathbf{x}_k = \mathbf{x}_{2k}$, it signals a preference for basing the calibration estimator $\tilde{Y}_{CAL}$ on $\mathbf{x}_{1k}$. More generally, we should choose $\mathbf{x}_k$ to make $|\Delta_A|$ large, because it is likely (but not guaranteed) that $\tilde{Y}_{CAL}$ based on this choice lies closer to the unbiased estimator.

4 Factoring the bias indicator

We express Δ_A and related quantities in matrix language. Define

$$\mathbf{D} = \bar{\mathbf{x}}_{r;d} - \bar{\mathbf{x}}_{s;d} \; ; \; \mathbf{C} = \left(\sum_r d_k (\mathbf{x}_k - \bar{\mathbf{x}}_{r;d})(y_k - \bar{y}_{r;d}\right) / (\sum_r d_k) \; ;$$
$$\mathbf{\Sigma} = \sum_r d_k \mathbf{x}_k \mathbf{x}_k' / \sum_r d_k \quad (4.1)$$

where the means are

$$\bar{y}_{r;d} = \sum_r d_k y_k / \sum_r d_k \; ; \; \bar{\mathbf{x}}_{r;d} = \sum_r d_k \mathbf{x}_k / \sum_r d_k \; ;$$
$$\bar{\mathbf{x}}_{s;d} = \sum_s d_k \mathbf{x}_k / \sum_s d_k$$

Here $\mathbf{D} = (D_1,...,D_j,...,D_J)'$, where $D_j = \bar{x}_{j|r;d} - \bar{x}_{j|s;d}$ measures what may be called a lack of balance, or a lack of representativity, in the variable x_j: When $\left|D_j\right|$ is large, the mean of the respondents, $\bar{x}_{j|r;d} = \sum_r d_k x_{jk} / \sum_r d_k$, is far from the mean of all those sampled, $\bar{x}_{j|s;d} = \sum_s d_k x_{jk} / \sum_s d_k$. The component C_j of $\mathbf{C} = (C_1,...,C_j,...,C_J)'$ is the (positive or negative) covariance between x_j and the study variable *y*,

$$C_j = Cov(x_j, y) = \sum_r d_k (x_{jk} - \bar{x}_{j|r;d})(y_k - \bar{y}_{r;d}) / \sum_r d_k \quad (4.2)$$

Finally, $\mathbf{\Sigma}$ is a $J \times J$ weighting matrix, assumed non-singular. By (2.2) we have the properties

$$\bar{y}_{r;d} = \bar{\mathbf{x}}_{r;d}' \mathbf{B}_{\mathbf{x}} \quad ; \quad \bar{\mathbf{x}}_{r;d}' \mathbf{\Sigma}^{-1} \bar{\mathbf{x}}_{r;d} = \bar{\mathbf{x}}_{r;d}' \mathbf{\Sigma}^{-1} \bar{\mathbf{x}}_{s;d} = 1 \quad (4.3)$$

Result 4.1: We can express Δ_A as a bilinear form in the vectors $\mathbf{D}$ and $\mathbf{C}$:

$$\Delta_A = (\tilde{Y}_{EXP} - \tilde{Y}_{CAL}) / \hat{N} = \mathbf{D}' \mathbf{\Sigma}^{-1} \mathbf{C} \quad (4.4)$$

Proof. We have $\tilde{Y}_{EXP} / \hat{N} = \bar{y}_{r;d} = \bar{\mathbf{x}}'_{r;d}\mathbf{B}_{\mathbf{x}}$ by the first part of (4.3), $\tilde{Y}_{CAL} / \hat{N} = \bar{\mathbf{x}}'_{s;d}\mathbf{B}_{\mathbf{x}}$ by (2.1) and therefore $\Delta_A = \mathbf{D}'\mathbf{B}_{\mathbf{x}}$. By the second part of (4.3), $\mathbf{D}'\mathbf{B}_{\mathbf{x}} = \mathbf{D}'\boldsymbol{\Sigma}^{-1}\mathbf{C}$, proving (4.4). □

It is convenient to measure Δ_A in standard deviations. Let

$$S^2_{y|r;d} = \sum_r d_k (y_k - \bar{y}_{r;d})^2 / \sum_r d_k = S^2_y$$

The simpler notation S^2_y will be used. We write Δ_A / S_y as a product of three easily interpreted factors:

$$\frac{\Delta_A}{S_y} = \frac{\tilde{Y}_{EXP} - \tilde{Y}_{CAL}}{\hat{N} \times S_y} = \; cv_m \times R_{y\mathbf{x}} \times R_{DC} \tag{4.5}$$

where

$$cv_m = (\mathbf{D}'\boldsymbol{\Sigma}^{-1}\mathbf{D})^{1/2} \;\; ; \;\; R_{y\mathbf{x}} = \frac{(\mathbf{C}'\boldsymbol{\Sigma}^{-1}\mathbf{C})^{1/2}}{S_y} \;\; ;$$

$$R_{DC} = \frac{\mathbf{D}'\boldsymbol{\Sigma}^{-1}\mathbf{C}}{(\mathbf{D}'\boldsymbol{\Sigma}^{-1}\mathbf{D})^{1/2}(\mathbf{C}'\boldsymbol{\Sigma}^{-1}\mathbf{C})^{1/2}}$$

The notation is suggestive: $(\mathbf{D}'\boldsymbol{\Sigma}^{-1}\mathbf{D})^{1/2} = cv_m$ is the coefficient of variation of the weight adjustment factors m_k, $\mathbf{C}'\boldsymbol{\Sigma}^{-1}\mathbf{C} / S^2_y = R^2_{y\mathbf{x}}$ is the coefficient of determination (the proportion of variance explained) for the multiple regression fit of y_k on $\mathbf{x}_k$, $k \in r$, and $(\mathbf{D}'\boldsymbol{\Sigma}^{-1}\mathbf{C})^2 / (\mathbf{D}'\boldsymbol{\Sigma}^{-1}\mathbf{D})\;(\mathbf{C}'\boldsymbol{\Sigma}^{-1}\mathbf{C}) = R^2_{DC}$ is the coefficient of determination for the fit of a weighted regression through the origin of D_j on C_j, $j = 1, 2, ..., J$. These properties of the three factors are proved in the following sections. It follows that $0 \le R_{y\mathbf{x}} \le 1$, $-1 \le R_{DC} \le 1$, and a typical range for the first factor is $0.1 \le cv_m \le 0.8$. The factorization (4.5) was introduced in Särndal and Lundström (2010). The factors and their role are examined more completely and rigorously in this paper.

In constructing the **x**-vector, we need to examine the progression of the values of the three factors when the vector dimension J increases as a result of the inclusion of more and more *x*-variables. At the outset, $J = 1$, $\mathbf{x}_k = 1$, and $cv_m = R_{y\mathbf{x}} = \Delta_A = 0$, while R_{DC} is undefined. The case $J = 2$ requires a closer examination; we then have $\mathbf{x}_k = (1, x_k)'$ where the entering variable x_k is continuous or categorical (0-1). Then

$$cv_m = \left|\bar{x}_{r;d} - \bar{x}_{s;d}\right| / S_x \quad , \quad R_{y\mathbf{x}} = \left|R_{yx}\right| \quad , \quad R_{DC} = \pm 1$$

where $S_x^2 = (\sum_r d_k)^{-1} \sum_r d_k (x_k - \bar{x}_{r;d})^2$ and $R_{yx} = Cov(x, y) / S_x S_y$. The sign of R_{DC} depends on whether $\bar{x}_{r;d} - \bar{x}_{s;d}$ has the same sign or not as the product moment correlation coefficient R_{yx}. We have

$$\frac{\Delta_A}{S_y} = \frac{\bar{x}_{r;d} - \bar{x}_{s;d}}{S_x} \times R_{yx}$$

Consequently, the nonresponse adjustment brings

$$\tilde{\bar{Y}}_{CAL}/\hat{N} \quad = \quad \tilde{\bar{Y}}_{EXP}/\hat{N} \quad - \quad S_y \times \frac{\bar{x}_{r;d} - \bar{x}_{s;d}}{S_x} \times R_{yx} \tag{4.6}$$

Both $\left|\bar{x}_{r;d} - \bar{x}_{s;d}\right| / S_x$ and $\left|R_{yx}\right|$ need to be distinctly non-zero in order for the adjustment effect to be important. A high correlation $\left|R_{yx}\right| = 0.9$ is in itself of little interest if accompanied by a low value of the imbalance such as $\left|\bar{x}_{r;d} - \bar{x}_{s;d}\right| / S_x = 0.1$. This *x*-variable brings the adjustment $\Delta_A = 0.09\, S_y$, relatively modest compared with another *x*-variable for which both factors equal, say, 0.5, which would bring the more pronounced adjustment $\Delta_A = 0.25\, S_y$.

In practice the dimension of the **x**-vector is often quite large; not uncommonly $J > 30$, or greater. To illustrate (4.5) with fairly typical numbers, suppose $cv_m = 0.6$, $R_{y\mathbf{x}} = 0.5$ and $R_{DC} = 0.6$. Then $\Delta_A / S_y = 0.18$, and $\tilde{\bar{Y}}_{CAL}/\hat{N} = \tilde{\bar{Y}}_{EXP}/\hat{N} - 0.18\, S_y$. Hence the crude estimate $\tilde{\bar{Y}}_{EXP}/\hat{N}$ undergoes an downward adjustment of 0.18

standard deviations to arrive at $\tilde{\tilde{Y}}_{CAL}/\hat{N}$. Although that adjustment may appear modest, it should be gauged against a reference such as the standard deviation of an estimated *y*-mean in a typical large survey. For example, with m = 10,000 respondents, the standard deviation of the estimated mean under simple random sampling (and random nonresponse) is in the neighbourhood of $S_y/\sqrt{m} = 0.01\, S_y$ (or less, if efficient auxiliary information is used); by comparison, $\Delta_A = 0.18\, S_y$ is large, implying a mean squared error dominated totally by the squared bias component. A confidence interval centered on $\tilde{\tilde{Y}}_{EXP}/\hat{N}$ is completely invalid. But $\tilde{\tilde{Y}}_{CAL}/\hat{N}$ in (4.6) is likely to be considerably less biased. A step has been taken in the right direction. Whether the adjustment comes close to completely eliminating the bias remains unknown.

5 The first factor

Result 5.1. The first factor of the decomposition (4.5), $(\mathbf{D}'\boldsymbol{\Sigma}^{-1}\mathbf{D})^{1/2}$, is the coefficient of variation (standard deviation divided by mean), computed for $k \in r$, of the weight adjustment factors $m_k = (\sum_s d_k \mathbf{x}_k)'(\sum_r d_k \mathbf{x}_k \mathbf{x}_k')^{-1}\mathbf{x}_k$ that appear in $\tilde{Y}_{CAL} = \sum_r d_k m_k y_k$.

Proof. The factor m_k is computable for $k \in s$. Two weighted means are of interest:

$\overline{m}_{r;d} = (\sum_r d_k m_k)/(\sum_r d_k)$ and $\overline{m}_{s;d} = (\sum_s d_k m_k)/(\sum_s d_k)$. A development making use of (2.1) shows that

$$\overline{m}_{r;d} = 1/P \quad ; \quad \overline{m}_{s;d} = (1/P)\,\overline{\mathbf{x}}_{s;d}'\boldsymbol{\Sigma}^{-1}\overline{\mathbf{x}}_{s;d} \tag{5.1}$$

where P is the response rate (1.1). The weighted variance of m_k over the response set is

$$S^2_{m|r;d} = \frac{1}{\sum_r d_k}\sum_r d_k (m_k - \overline{m}_{r;d})^2 = S^2_m \tag{5.2}$$

The simpler notation S^2_m will be used. Using that $\sum_r d_k m_k^2 = \sum_s d_k m_k$ we get $S^2_m = \overline{m}_{r;d}(\overline{m}_{s;d} - \overline{m}_{r;d})$. It follows that $\overline{m}_{s;d} \geq \overline{m}_{r;d}$ for any outcome (s,r). The coefficient of variation of m_k for $k \in r$ is

$$cv_m = \frac{S_m}{\overline{m}_{r;d}} = \sqrt{\frac{\overline{m}_{s;d}}{\overline{m}_{r;d}} - 1} \tag{5.3}$$

Using (5.1) and the second part of (4.3) we complete the proof by noting that

$$cv_m^2 = \overline{\mathbf{x}}_{s;d}'\boldsymbol{\Sigma}^{-1}\overline{\mathbf{x}}_{s;d} - 1 = (\overline{\mathbf{x}}_{r;d} - \overline{\mathbf{x}}_{s;d})'\boldsymbol{\Sigma}^{-1}(\overline{\mathbf{x}}_{r;d} - \overline{\mathbf{x}}_{s;d}) = \mathbf{D}'\boldsymbol{\Sigma}^{-1}\mathbf{D} \quad \square \tag{5.4}$$

Comment 5.1. Result 5.1 expresses $cv_m^2 = \mathbf{D}'\boldsymbol{\Sigma}^{-1}\mathbf{D}$ as a quadratic form in the vector $\mathbf{D} = (D_1,...,D_j,...,D_J)'$, where $D_j = x_{j|r;d} - x_{j|s;d}$ measures the lack of balance in the variable x_j. A large D_j may or may not be harmful in the sense "causing large bias". The issue depends, as seen more clearly later, on the size of the matching component C_j of $\mathbf{C} = (C_1,...,C_j,...,C_J)'$. The value of $cv_m^2 = \mathbf{D}'\boldsymbol{\Sigma}^{-1}\mathbf{D}$ increases (or possibly stays the same) when further *x*-variables are added to the auxiliary vector $\mathbf{x}_k$.

Comment 5.2. Can one claim that it is desirable to choose the **x**-vector so that the variance $S_m^2 = cv_m^2 / P^2$ of the weight factors m_k is large? An argument in the affirmative is linked to the degree to which $\mathbf{x}_k$ explains the response mechanism, more specifically the inverse of the response probability denoted θ_k. The theoretical weight factors $\phi_k = 1/\theta_k$ would, if known, make $\sum_r d_k \phi_k y_k$ unbiased for $Y = \sum_U y_k$. Predictors of the ϕ_k can be derived by seeking $\boldsymbol{\lambda}$ to minimize the sum of squares $\sum_U \theta_k(\phi_k - \boldsymbol{\lambda}'\mathbf{x}_k)^2$. The optimal predictions are $\hat{\phi}_k = \hat{\boldsymbol{\lambda}}'\mathbf{x}_k = (\sum_U \mathbf{x}_k)'(\sum_U \theta_k \mathbf{x}_k \mathbf{x}_k')^{-1}\mathbf{x}_k = M_k$, say. The total variance of the ϕ_k, $S^2_{\phi|U;\theta} = \sum_U \theta_k(\phi_k - \bar{\phi}_{U;\theta})^2 / \sum_U \theta_k$, is decomposed as $S^2_{\phi|U;\theta} = S^2_{M|U;\theta} + S^2_{res|U;\theta}$, where the component "variance explained by $\mathbf{x}_k$" is $S^2_{M|U;\theta} = \sum_U \theta_k(M_k - \bar{M}_{U;\theta})^2 / \sum_U \theta_k$, and $S^2_{res|U;\theta} = \sum_U \theta_k(\phi_k - M_k)^2 / \sum_U \theta_k$. A desirable choice of $\mathbf{x}_k$ is one that yields a large component of variance explained, $S^2_{M|U;\theta}$. The latter contains unknown population quantities; we replace those by their analogues computed on data for respondents: M_k becomes m_k, and $S^2_{M|U;\theta}$ becomes S_m^2, and the case can be made that it is desirable to choose $\mathbf{x}_k$ to make the weight factor variance $S_m^2 = \mathbf{D}'\boldsymbol{\Sigma}^{-1}\mathbf{D} / P^2$ large.

6 The second factor

Result 6.1. The factor $(\mathbf{C}'\boldsymbol{\Sigma}^{-1}\mathbf{C})^{1/2}/S_y$ in (4.5) is the coefficient of multiple correlation between y_k and $\mathbf{x}_k$. Hence, $\mathbf{C}'\boldsymbol{\Sigma}^{-1}\mathbf{C}/S_y^2 = R_{y\mathbf{x}}^2$ is the coefficient of determination for the multiple regression fit of y_k on $\mathbf{x}_k$ for $k \in r$.

Proof. The regression fit of y on $\mathbf{x}$ gives for unit k the predicted value $\hat{y}_k = \mathbf{x}_k'\mathbf{B}_\mathbf{x}$, where $\mathbf{B}_\mathbf{x}$ is given by (1.4). Out of the total variance S_y^2 of y, the component of variance explained is $\sum_r d_k(\hat{y}_k - \bar{y}_{r;d})^2 / \sum_r d_k = (\sum_r d_k y_k \hat{y}_k)/(\sum_r d_k) - \bar{y}_{r;d}^2 = \mathbf{C}'\boldsymbol{\Sigma}^{-1}\mathbf{C}$, where we have used (2.1) and (4.3). Hence the ratio of explained-to-total variance (the coefficient of determination) is

$$\sum_r d_k(\hat{y}_k - \bar{y}_{r;d})^2 / \sum_r d_k(y_k - \bar{y}_{r;d})^2 = (\mathbf{C}'\boldsymbol{\Sigma}^{-1}\mathbf{C})/S_y^2 = R_{y\mathbf{x}}^2 \quad \square \quad (6.1)$$

Comment 6.1: Result 6.1 shows $\mathbf{C}'\boldsymbol{\Sigma}^{-1}\mathbf{C} = R_{y\mathbf{x}}^2 \times S_y^2$ as a quadratic form in the covariance vector $\mathbf{C} = (C_1,...,C_j,...,C_J)'$, where C_j is given by (4.2). A value of, say, $R_{y\mathbf{x}}^2 = 0.9$ indicates a strong regression relationship between **x** and *y*, but is in itself insufficient to bring about a large value $|\Delta_A|$, because $R_{y\mathbf{x}}^2$ is only one of three contributing factors. This begs the question: What is the value and the importance of "an improved fit" of $y_k = \mathbf{x}_k'\boldsymbol{\beta} + \varepsilon_k$? The question has two aspects: (i) the values y_k are viewed as fixed constants (as they are for a given finite population), while the **x**-vector expands, through an inclusion of additional *x*-variables, and (ii) the composition of the **x**-vector is fixed, but the y_k-values change in a direction of smaller regression residuals. It is fitting in both cases to use the term "improved fit" to mean that the residual variance $S_e^2 = (\sum_r d_k e_k^2)/(\sum_r d_k)$ decreases, where $e_k = y_k - \hat{y}_k = y_k - \mathbf{x}_k'\mathbf{B}_\mathbf{x}$

with $\mathbf{B}_{\mathbf{x}}$ given by (1.4). Then $S_y^2 = S_{\hat{y}}^2 + S_e^2$, where $S_{\hat{y}}^2 = \sum_r d_k (\hat{y}_k - \bar{\hat{y}}_{r;d})^2 / \sum_r d_k = \mathbf{C}'\mathbf{\Sigma}^{-1}\mathbf{C}$.

In the aspect (i), we have $R_{y,\mathbf{x}}^2 = 1 - S_e^2 / S_y^2$, where S_y^2 is a constant since the y_k-values do not change. The enlargement of the **x**-vector brings a decrease in S_e^2 and an increase in $R_{y,\mathbf{x}}^2$. It will also increase $cv_m^2 = \mathbf{D}'\mathbf{\Sigma}^{-1}\mathbf{D}$. If many efficient variables are admitted into the **x**-vector, $R_{y,\mathbf{x}}^2$ may come near unity. Thus in the aspect (i), "improved fit" entails an increase both in the first factor, cv_m^2, and in the second factor, $R_{y,\mathbf{x}}^2$. However, the effect on the third factor, R_{DC}^2, is unpredictable; it can change in either direction.

Consider now the aspect (ii). The **x**-vector is made up of a fixed set of *x*-variables with fixed values. The *y*-values undergo change so that the residuals $y_k - \hat{y}_k = e_k$ become progressively smaller. We have $R_{y,\mathbf{x}}^2 = 1 - [S_e^2 / (\mathbf{C}'\mathbf{\Sigma}^{-1}\mathbf{C} + S_e^2)]$. The improved fit (the reduced S_e^2) leaves $\mathbf{C}'\mathbf{\Sigma}^{-1}\mathbf{C}$ unchanged, so $R_{y,\mathbf{x}}^2$ increases. The factors $cv_m^2 = \mathbf{D}'\mathbf{\Sigma}^{-1}\mathbf{D}$ and R_{DC}^2 are also unchanged. Thus in the aspect (ii), "improved fit" entails an increase in the factor $R_{y,\mathbf{x}}^2$ but a status quo in the factors $cv_m^2 = \mathbf{D}'\mathbf{\Sigma}^{-1}\mathbf{D}$ and R_{DC}^2. In the unlikely event of "perfect fit", then $e_k = 0$ for all $k \in r$, so $R_{y,\mathbf{x}}^2 = 1$, but despite this, $|\Delta_A| / S_y = cv_m \times R_{y\mathbf{x}} \times |R_{DC}|$ may not reach a particularly large value.

7 The third factor

Result 7.1. The square of the third factor in (4.5), $(\mathbf{D}'\boldsymbol{\Sigma}^{-1}\mathbf{C})^2/(\mathbf{D}'\boldsymbol{\Sigma}^{-1}\mathbf{D})(\mathbf{C}'\boldsymbol{\Sigma}^{-1}\mathbf{C})$, is the coefficient of determination (the proportion of variance explained) in the fit of a (weighted) regression through the origin of D_j on C_j, $j = 1, 2, ..., J$, which are components of $\mathbf{D} = (D_1, ..., D_j, ..., D_J)'$ and $\mathbf{C} = (C_1, ..., C_j, ..., C_J)'$, respectively.

Proof. Consider the quadratic form $\mathbf{D}'\boldsymbol{\Sigma}^{-1}\mathbf{D}$ as a measure of total variability of the D_j, to be decomposed as a sum of an explained component and a residual component. To the data points (D_j, C_j), $j = 1, 2, ..., J$, fit a weighted simple regression through the origin,

$$D_j = K \times C_j + E_j$$

where the slope K is to be determined. Let $\mathbf{E} = (E_1, ..., E_j, ..., E_J)'$. A weighted least squares minimization of $\mathbf{E}'\boldsymbol{\Sigma}^{-1}\mathbf{E} = (\mathbf{D} - K\,\mathbf{C})'\boldsymbol{\Sigma}^{-1}(\mathbf{D} - K\,\mathbf{C})$ gives $K = K_o$, where

$$K_o = (\mathbf{D}'\boldsymbol{\Sigma}^{-1}\mathbf{C})/(\mathbf{C}'\boldsymbol{\Sigma}^{-1}\mathbf{C})$$

The resulting minimum value of the residual component is

$$(\mathbf{D} - K_o\mathbf{C})'\boldsymbol{\Sigma}^{-1}(\mathbf{D} - K_o\mathbf{C}) = \mathbf{D}'\boldsymbol{\Sigma}^{-1}\mathbf{D} - K_o^2(\mathbf{C}'\boldsymbol{\Sigma}^{-1}\mathbf{C}) \qquad (7.1)$$

where $K_o^2(\mathbf{C}'\boldsymbol{\Sigma}^{-1}\mathbf{C})$ is the explained component. Thus the ratio of explained variability to total variability (the coefficient of determination) is

$$K_o^2(\mathbf{C}'\boldsymbol{\Sigma}^{-1}\mathbf{C})/(\mathbf{D}'\boldsymbol{\Sigma}^{-1}\mathbf{D}) = (\mathbf{D}'\boldsymbol{\Sigma}^{-1}\mathbf{C})^2/(\mathbf{D}'\boldsymbol{\Sigma}^{-1}\mathbf{D})(\mathbf{C}'\boldsymbol{\Sigma}^{-1}\mathbf{C}) = R_{DC}^2 \quad (7.2)$$

Comment 7.1: Although both $R_{y\mathbf{x}}$ and R_{DC} are correlation coefficients, their interpretations are quite different. The former is a measure, *for the responding units* $k \in r$, of the degree of relationship between the study variable y_k and the auxiliary vector $\mathbf{x}_k$. By contrast, R_{DC} is a measure, *for the participating auxiliary variables* $j = 1, ..., J$, of the degree of relationship between the lack of balance D_j and the covariance $C_j = Cov(x_j, y)$. While $R_{y\mathbf{x}}$ increases as more *x*-variables enter into $\mathbf{x}_k$, $|R_{DC}|$ may not have this property when *J* increases; it may decrease.

The factor $R_{DC} = (\mathbf{D}'\boldsymbol{\Sigma}^{-1}\mathbf{C})/\{(\mathbf{D}'\boldsymbol{\Sigma}^{-1}\mathbf{D})^{1/2}(\mathbf{C}'\boldsymbol{\Sigma}^{-1}\mathbf{C})^{1/2}\}$ can be interpreted as a measure of proportionality between the lack of balance $D_j = \bar{x}_{j|r;d} - \bar{x}_{j|s;d}$ and the covariance C_j $= Cov(y, x_j) = R_{y,x_j} \times S_{x_j} \times S_y$. For a fixed *y*-variable, and a fixed dimension *J* of $\mathbf{x}_k$, the maximum value, $R_{DC}^2 = 1$, would be attained if the proportionality $\mathbf{D} = A \times \mathbf{C}$ holds for some constant A. Let $A_0 = A \times S_y$; then $R_{DC}^2 = 1$ holds if the standardized lack of balance for the variable x_j is proportional to that auxiliary variable's correlation with the study variable y:

$$\frac{\bar{x}_{j|r;d} - \bar{x}_{j|s;d}}{S_{x_j}} = A_0 \times R_{y,x_j} \;;\; j = 1, \ldots, J \qquad (7.3)$$

We can also interpret $R_{DC} = (\mathbf{D}'\boldsymbol{\Sigma}^{-1}\mathbf{C})/\{(\mathbf{D}'\boldsymbol{\Sigma}^{-1}\mathbf{D})^{1/2}(\mathbf{C}'\boldsymbol{\Sigma}^{-1}\mathbf{C})^{1/2}\}$ as the cosine of the angle (in *J*-dimensional space) between the vectors $\mathbf{D}$ and $\mathbf{C}$. When the dimension J increases, the angle normally grows wider and $|R_{DC}|$ decreases. If the perfect proportionality, $\mathbf{D} = A \times \mathbf{C}$, holds for some constant A, then the angle is zero, and $R_{DC}^2 = 1$.

At the heart of the matter of achieving an important adjustment $|\Delta_A|$ lies the progression in the value of R_{DC}^2 when more and more variables are allowed into $\mathbf{x}_k$. For $J = 2$ we have $\mathbf{x}_k = (1, x_k)'$ and $R_{DC}^2 = 1$, whatever the entering variable x_k. In practice, the dimension J of $\mathbf{x}_k$ may be considerable, say $J > 30$ or more. We may then be far from attaining the proportionality (7.3). A higher dimension J will normally cause a lower value of R_{DC}^2. As the vector $\mathbf{x}_k$ expands, R_{DC}^2 has a tendency to decrease, as a result of an increased scatter of the J points (D_j, C_j), $j = 1, 2, ..., J$. We prefer a set of auxiliary variables that produces a small scatter around the fitted line through the origin.

In summary, extending the dimension J of the **x**-vector by the addition of further *x*-variables affects the value of $|\Delta_A| / S_y = cv_m \times R_{y\mathbf{x}} \times |R_{DC}|$ in the following manner: cv_m and $R_{y\mathbf{x}}$ will increase, but $|R_{DC}|$ is likely to decrease from its maximum $|R_{DC}| = 1$ when $J = 2$. The increase in the first two may more than offset the decrease in $|R_{DC}|$, so that $|\Delta_A| / S_y$ increases. However, this is not always so; $|\Delta_A| / S_y$ may start to decrease because of a pronounced drop in $|R_{DC}|$. Thus a *reversal* may occur in the value of $|\Delta_A|$, when the increases in cv_m and $R_{y\mathbf{x}}$ are insufficient to offset a decrease in $|R_{DC}|$, with a decrease in $|\Delta_A| / S_y$ as a result.

8 A single categorical auxiliary variable

In the preceding theory, the auxiliary variables can be continuous or categorical. We now examine the important case of a single categorical auxiliary variable defined by a set of mutually exclusive and exhaustive traits or properties. We shall use the former term. For example, the variable Age may be defined by three traits: Young, Middle-aged and Elderly. A set of units sharing the same trait is called a trait group, or simply a group.

Suppose the auxiliary vector is defined initially by a total of J_{tot} traits. (In the final analysis we may not keep all.) The auxiliary vector $\mathbf{x}_k = (\gamma_{1k},...,\gamma_{jk},...,\gamma_{J_{tot}k})'$ codes the trait of unit k, where $\gamma_{jk} = 1$ if k has the trait j, and $\gamma_{jk} = 0$ otherwise, $j = 1,..., J_{tot}$. That is, the vector is of the form $\mathbf{x}_k = (0,...,1,...,0)'$ with a single entry "1". Let s_j be the subset of the sample s consisting of the units *k* with the trait *j*. Its size is random, unless that trait group was designated as a stratum in the probability sampling design. The responding subset of s_j is denoted r_j. For trait *j*, define also the following weighted quantities: $\hat{N}_j = \sum_{s_j} d_k$ (an unbiased estimator of the population group count N_j), $Q_{sj} = \hat{N}_j / \hat{N}$ (the trait group size as a proportion of the full sample), $Q_{rj} = \sum_{r_j} d_k / \sum_r d_k$ (the proportion of the whole response set), $P_j = \sum_{r_j} d_k / \hat{N}_j$ (the response rate), and $\bar{y}_{r_j;d} = \sum_{r_j} d_k y_k / \sum_{r_j} d_k$ (the study variable mean for respondents). Here $\hat{N}$ and $\bar{y}_{r;d}$ are defined as before by (2.2). Then the adjustment Δ_A that we seek to make large in absolute value can be written as

$$\Delta_A = \mathbf{D}'\boldsymbol{\Sigma}^{-1}\mathbf{C} = \sum_{j=1}^{J_{tot}} (Q_{rj} - Q_{sj})(\bar{y}_{r_j;d} - \bar{y}_{r;d}) = \bar{y}_{r;d} \sum_{j=1}^{J_{tot}} H_{Aj} \qquad (8.1)$$

where

$$H_{Aj} = Q_{sj} \times \frac{P_j - P}{P} \times \frac{\bar{y}_{r_j;d} - \bar{y}_{r;d}}{\bar{y}_{r;d}} \tag{8.2}$$

Starting from the crude estimator $\hat{\bar{y}}_{U,EXP} = \tilde{Y}_{EXP} / \hat{N} = \bar{y}_{r;d}$ of $\bar{y}_U = \sum_U y_k / N$, the adjustment brings the improved estimator $\hat{\bar{y}}_{U,CAL} = \hat{\bar{y}}_{U,EXP} - \Delta_A$, that is,

$$\hat{\bar{y}}_{U,CAL} = \bar{y}_{r;d}(1 - \sum_{j=1}^{J_{tot}} H_{Aj}) \tag{8.3}$$

which also has a more familiar expression, commonly referred to in the literature as the Weighting Class estimator,

$$\hat{\bar{y}}_{U,CAL} = \hat{\bar{y}}_{U,WC} = \hat{N}^{-1} \sum_{j=1}^{J_{tot}} \hat{N}_j \, \bar{y}_{r_j;d} \tag{8.4}$$

Although a very simple application of the general procedure, it has drawn much attention in the literature. It is usually taken for granted that the categorical auxiliary variable be used "as is", with all of its J_{tot} predefined traits. The question whether all traits are worth keeping is seldom if ever raised. But the form (8.3) prompts the question, because it puts the emphasis on the contribution of each particular trait to a desired departure from the crude estimate $\bar{y}_{r;d}$. As (8.2) shows, three factors contribute to $|H_{Aj}|$, called the importance of the j:th trait (coded by γ_{jk}), namely, *the relative trait group size* Q_{sj}, *the response relative* $(P_j - P)/P$, *the respondent mean relative* $(\bar{y}_{r_j;d} - \bar{y}_{r;d})/\bar{y}_{r;d}$. All three factors need to reach significant levels in order to make the j:th trait important. Some of the J_{tot} traits may not contribute enough to $|H_A| = \left| \sum_{j=1}^{J_{tot}} H_{Aj} \right|$ to be worth keeping. Others may be outright counterproductive.

We note that large group sizes Q_{sj} are preferable to small group sizes. For the most important few traits, suppose H_{Aj} is positive, that is, $(P_j - P)/P$ and $(\bar{y}_{r_j;d} - \bar{y}_{r;d})/\bar{y}_{r;d}$ have the same sign. Then less important traits such that these two factors are of opposite sign will "pull in the wrong direction" and should be relegated to a "rest group", comprising all others. Therefore, under the objective to realize a high value of $|\Delta_A| = |\mathbf{D}'\boldsymbol{\Sigma}^{-1}\mathbf{C}|$, one may be led to keep only a subset consisting of J traits, where $J \le J_{tot}$. Then, (8.3) and (8.4) become

$$\hat{\bar{y}}_{U,CAL} = \bar{y}_{r;d}(1 - \sum_{j=1}^{J} H_{Aj}) = \hat{N}^{-1}\sum_{j=1}^{J} \hat{N}_j \bar{y}_{r_j;d} = \hat{\bar{y}}_{U,selWC} \qquad (8.5)$$

which may be called the *Selective Weighting Class estimator*.

We can code a *J*-category classification as $\mathbf{x}_k = (1, \gamma_{1k}, \gamma_{2k}, ..., \gamma_{J-1,k})'$ as an alternative to $\mathbf{x}_k = (\gamma_{1k}, ..., \gamma_{jk}, ..., \gamma_{Jk})'$. This substitution leaves $\mathbf{D}'\boldsymbol{\Sigma}^{-1}\mathbf{D}$, $\mathbf{C}'\boldsymbol{\Sigma}^{-1}\mathbf{C}$ and $\mathbf{D}'\boldsymbol{\Sigma}^{-1}\mathbf{C}$ unchanged, and, as a consequence, cv_m, $R_{y\mathbf{x}}$, R_{DC} and $\Delta_A / S_y = cv_m \times R_{y\mathbf{x}} \times R_{DC}$ are also unchanged. To illustrate, a higher value of $|\Delta_A|$ may be realized by keeping only $J = 2$ out of a total of $J_{tot} = 5$ traits, say, those coded by γ_{1k} and γ_{2k}, so that the vector is $\mathbf{x}_k = (1, \gamma_{1k}, \gamma_{2k})'$. The other three traits, those coded by γ_{3k}, γ_{4k} and γ_{5k}, form the rest group. An example of "keeping fewer than all" is seen in the empirical section 10.

The important special case $J = 2$ implies a dichotomy of the units: Those *with* the specified trait, defined by the sets s_1 and r_1, and those *without* it, defined by the complement sets comprising all other traits, $\bar{s}_1 = s - s_1$ and $\bar{r}_1 = r - r_1$. Then $\mathbf{x}_k = (1, \gamma_{1k})'$, where $\gamma_{1k} = 1$ if k has the given trait and $\gamma_{1k} = 0$ otherwise. (For identical results, we can code the vector as $\mathbf{x}_k = (\gamma_{1k}, \gamma_{2k})'$, where $\gamma_{2k} = 1 - \gamma_{1k}$ indicates the complement set of "all others".) Then (8.1) becomes

$$\Delta_A = (Q_{r1} - Q_{s1})(\bar{y}_{r_1;d} - \bar{y}_{\bar{r}_1;d})$$

where $\bar{y}_{\bar{r}_1;d}$ is the design-weighted *y*-mean for the complement set $\bar{r}_1$. Two features of the specified trait interact to create a high value of $|\Delta_A|$: (i) the proportion of respondents, Q_{r1}, differs markedly from the proportion of sampled units, Q_{s1}, *and* (ii) the mean $\bar{y}_{r_1;d}$ for respondents with the trait differs markedly from the mean $\bar{y}_{\bar{r}_1;d}$ for respondents without the trait. Both differences, not just one of them, need to be pronounced in order to generate a high $|\Delta_A|$.

Remark. If known, the population count N_j replaces $\hat{N}_j$ in (8.4). The estimator is then known as the Population Weighting Adjustment estimator**.** The difference in bias compared with (8.4) is inconsequential, but the variance may be significantly smaller.

9 Empirical illustration I

We present an example in two parts, representing two different fictitious (but not unrealistic) data sets, part I in this section, part II in section 10. Suppose the data set I shown in Table 9.1 is the result of probability sampling from a large population whose size, N, may be of the order of several million, as in sampling from the National Population Register of Sweden.

Table 9.1. Data set I, for a simple random sample of size n = 4,000. Frequencies by trait group and overall, shown for responding, non-responding and whole sample. Within parenthesis, frequency of units with the trait Υ. Asterisk indicates unobservable frequency

	Trait group			All traits (thereof Υ)
	$j = 1$ (thereof Υ)	$j = 2$ (thereof Υ)	$j = 3$ (thereof Υ)	
Response	400 (200)	500 (200)	1500 (400)	2400 (800)
Non-response	1100 ($700^{\times}$)	400 ($50^{\times}$)	100 ($50^{\times}$)	1600 ($800^{\times}$)
Total sample	1500 ($900^{\times}$)	900 ($250^{\times}$)	1600 ($450^{\times}$)	4000 ($1600^{\times}$)

To fix ideas, we may think of $J_{tot} = 3$ traits of the variable Age: Young (j = 1), Middle-aged (j = 2) and Elderly (j = 3). The objective is to estimate the proportion in the population with the attribute denoted Υ, say the use of a certain drug, expected to be more prevalent among the young. (To mark the distinction, we use "attribute" for a *dichotomous study variable*, and "trait" for a *dichotomous trait indicator variable*.)

Table 9.1 shows the frequency count of persons, by trait and overall, broken down further into responding, non-responding and total sample. Shown in parenthesis is the frequency, out of those in a given cell, with the targeted attribute Υ. The entries marked with

superscript $\times$ are unknown to the statistician/analyst. Any conclusions must be based on the other numbers.

The dichotomous study variable *y* has the value $y_k = 1$ if person *k* has the attribute Υ, and $y_k = 0$ otherwise. We wish to estimate the proportion (the prevalence) of persons with the attribute Υ, $\bar{y}_U = \sum_U y_k / N$. The probability sample *s* of size $n = 4{,}000$ is drawn by simple random sampling, so the design weighting is uniform: $d_k = N / n$ for all *k*, and $\hat{N} = N$.

High (but not unrealistically high) nonresponse occurs, at a rate clearly more pronounced among the young, which also have a higher prevalence of Υ. The primitive estimate based on the response set r of size m is $\hat{\bar{y}}_{U,EXP} = \tilde{\tilde{Y}}_{EXP} / N = \bar{y}_r = \sum_r y_k / m =$ 800/2400= 0.3333. An adjusted estimate will be $\hat{\bar{y}}_{U,CAL}$ given by (8.5) with $J = 2$ or $J = J_{tot} = 3$, and $\hat{N} = N$.

Some features of the data are shown in Table 9.2, with notation as defined at the beginning of Section 8. The analyst routinely computes the numbers in rows one to four as part of a *nonresponse analysis.* Considerable group differences exist, both in the response rate P_j and in the prevalence $\bar{y}_{r_j}$ of the attribute Υ. There is strong incentive to adjust by age group, and to publish the Weighting Class estimate (8.4), $\hat{\bar{y}}_{U,CAL} = \tilde{\tilde{Y}}_{CAL} / N = (1/N) \sum_{j=1}^{3} \hat{N}_j \, \bar{y}_{r_j} = 0.3842$.

Table 9.2. Data set I. Entries in the first four rows are computed by the analyst; those of the bottom row, marked ×, are unavailable

	Trait group			All traits
	$j = 1$	$j = 2$	$j = 3$	
Proportion of sample, Q_{sj}	0.3750	0.2250	0.4000	1
Proportion of response, Q_{rj}	0.1667	0.2083	0.6250	1
Response rate, P_j	0.2667	0.5556	0.9375	$P = 0.6000$
Prevalence of γ in the response, $\bar{y}_{r_j}$	0.5000	0.4000	0.2667	$\bar{y}_r = 0.3333$
Prevalence of γ in the sample, $\bar{y}_{s_j}$	$0.6000^{\times}$	$0.2778^{\times}$	$0.2813^{\times}$	$\bar{y}_s = 0.4000^{\times}$

We turn now to some facts that are beyond the reach of the analyst, because of the nonresponse. They involve the entries in the bottom line of Table 9.2, marked with superscript ×:

(i) If the whole sample had responded, the unbiased mean estimate (the prevalence of γ) would be $\hat{\bar{y}}_{U,FUL} = \tilde{Y}_{FUL} / N = \bar{y}_s = \sum_s y_k / n =$ 1600/4000= 0.4000, considerably higher than the primitive estimate computed on respondents, $\hat{\bar{y}}_{U,EXP} = \tilde{Y}_{EXP} / N = \bar{y}_r = 0.3333$, which is a severe underestimation.

(ii) In a group-by-group comparison of sampled units with responding units, the *y*-mean (the prevalence of γ) shows the following contrasts: 0.6000 vs. 0.5000 for $j = 1$; 0.2778 vs. 0.4000 for $j = 2$, and 0.2813 vs. 0.2667 for $j = 3$. The difference is considerable for $j = 1$ and $j = 2$, small for $j = 3$. A nearly complete elimination of the nonresponse error would have required the difference to be near zero for each trait. This is not achieved by conditioning on age group. Unaware of this, the analyst may venture *the assumption* of MAR (missing at random), conditional on age group. Under that

assumption, $\hat{\bar{y}}_{U,CAL} = \tilde{\bar{Y}}_{CAL} / N = (1/N)\sum_{j=1}^{3} \hat{N}_j \, \bar{y}_{r_j;d} = 0.3842$, is deemed an unbiased estimate. But the assumption is invalid; 0.3842 remains far from the unbiased estimate 0.4000. This does not imply that Age is an inefficient auxiliary variable. On the contrary, it achieves an important partial although incomplete adjustment for bias.

The analyst can in effect do more than the traditional nonresponse analysis in the first four lines of Table 9.2. He/she should start from the expression (8.3) for $\hat{\bar{y}}_{U,CAL}$, compute for $j = 1, 2, 3$ the quantities H_{Aj} given by (8.2) and thereby see the importance of each individual trait. The result is shown in Table 9.3.

Table 9.3. Data set I. Analysis of trait influence. The bottom line entry H_{Aj} is the product of the preceding three entries

	Trait group			Sum
	$j = 1$	$j = 2$	$j = 3$	
Q_{sj}	0.3750	0.2250	0.4000	1
$(P_j - P)/P$	-0.5556	-0.0741	0.5625	
$(\bar{y}_{r_j;d} - \bar{y}_{r;d})/\bar{y}_{r;d}$	0.5000	0.2000	-0.2000	
H_{Aj}	-0.1042	-0.0033	-0.0450	H_A = -0.1525

Rows 2 and 3 in Table 9.3 have weighted means equal to zero: $\sum_{j=1}^{3} Q_{sj}(P_j - P) = 0$; $\sum_{j=1}^{3} Q_{rj}(\bar{y}_{r_j} - y_r) = 0$. The table shows that the trait $j = 1$ (Young) brings high values on both $|P_j - P|/P$ (due to a low response rate) and $|\bar{y}_{r_j;d} - \bar{y}_{r;d}|/\bar{y}_{r;d}$ (due to a high prevalence of the attribute of interest y). The trait $j = 3$ (Elderly) is also important: A high response rate is paired with a rather low prevalence of y. Together those two traits account for nearly all of

$H_A = \sum_{j=1}^{3} H_{Aj}$ and therefore for nearly all of the adjustment $\Delta_A = \bar{y}_{r;d} H_A$; Thus Table 9.3 informs the analyst that $j = 1$ and $j = 3$ are the critical traits, primarily accountable for the underestimation in the primitive estimate $\hat{\bar{y}}_{U,EXP} = 0.3333$. The trait $j = 2$ (Middle aged) is unimportant. The adjusted estimate by (8.3) is $\hat{\bar{y}}_{U,EXP}(1 - H_A) = 0.3333(1 + 0.1525) = 0.3842$, confirming the result of the Weighting Class formula (8.4).

The next question becomes: Are all three traits necessary? Should the number of traits be reduced? Table 9.4 throws some light on these questions. It illustrates the progression of a stepwise forward selection of traits. Consider the stepwise algorithm such that, in a given step, we enter the trait for which $|\Delta_A| = \bar{y}_{r;d} |H_A|$ has its highest value, where $H_A = \sum_{j=1}^{J} H_{Aj}$, and J is the dimension of the **x**-vector, $J \le J_{tot} = 3$.

At Step 0 we have the primitive vector $\mathbf{x}_k = 1$. At step 1, with $J = 2$, the choice is between the three vectors $\mathbf{x}_k = (1, \gamma_{jk})'$, the entering variable γ_{jk} being the indicator of trait j; j = 1 or 2 or 3. If j is admitted, the union of the other two traits forms the "rest group". The symbol $\cup$ denotes a merger of trait groups; for example, $2 \cup 3$ denotes the merger of j = 2 and j = 3. The vector formulations $\mathbf{x}_k = (1, \gamma_{jk})'$ and $\mathbf{x}_k = (\gamma_{jk}, \gamma_{\bar{j}k})'$ (where $\bar{j}$ denotes "*not-j*") are equivalent; they leave $\mathbf{D}'\boldsymbol{\Sigma}^{-1}\mathbf{D}$, $\mathbf{D}'\boldsymbol{\Sigma}^{-1}\mathbf{C}$ and $\mathbf{C}'\boldsymbol{\Sigma}^{-1}\mathbf{C}$ invariant. Thus cv_m, $R_{y\mathbf{x}}$ and R_{DC} are also invariant.

Table 9.4. Data set I. Values of cv_m, $R_{y\mathbf{x}}$ and R_{DC}, of their product Δ_A / S_y, and of Δ_A, shown for each of three steps. n.d. stands for "not defined". Standard deviation $S_y = 0.4714$

	Step 0	Step 1 Two trait groups			Step 2 Three trait groups
	1∪2∪3	1 , 2∪3	2 , 1∪3	3 , 1∪2	1, 2, 3
cv_m	0	0.5590	0.0410	0.4648	0.5855
$R_{y\mathbf{x}}$	0	0.1581	0.0725	0.1826	0.1937
R_{DC}	n.d.	- 1	- 1	- 1	-0.9512
Δ_A / S_y	0	-0.0884	-0.0030	-0.0849	-0.1078
Δ_A	0	-0.0417	-0.0014	-0.0400	-0.0508

In Step 1, the largest value of $|\Delta_A|$ is 0.0417, realized by entering j = 1 (but j = 3 is a close second). The Selective Weighting estimate after step 1 is thus $\hat{\bar{y}}_{U,CAL} = \tilde{Y}_{CAL} / N = 0.3333 + 0.0417 = 0.3750$, thereby reducing the underestimation considerably, from - 6.7% (= 0.3333 – 0.4000) to -2.5% (= 0.3750 – 0.4000). Table 9.4 reinforces the impression from Table 9.3 that j = 2 (Middle aged) is the least important trait.

Step 2 entails, in general, a selection of the trait corresponding to the largest value of $|\Delta_A|$, given the Step 1 selection. Here, there are no more than $J_{tot} = 3$ traits, so Step 2 implies that all three become taken into account. We see that cv_m and $R_{y\mathbf{x}}$ increase, as they must, whereas $|R_{DC}|$ recedes from 1 to 0.9512. Nevertheless, $|\Delta_A|$ increases from 0.0417 to 0.0508. The resulting estimate after Step 2 is therefore $\hat{\bar{y}}_{U,CAL} = 0.3333 + 0.0508 = 0.3842$, which reduces the underestimation to -1.7% (= 0.3842 – 0.4000). For these data, $|\Delta_A|$ increases in each step. This increasing pattern cannot be taken for granted, as the next section will illustrate.

10 Empirical illustration II

Table 10.1 presents a variation on the data in Table 9.1. The two data sets have some similarities. The data are the same for "All traits", for the distribution of the total sample on the three traits, and for the response in trait group $j = 1$. The crude estimate remains $\hat{\bar{y}}_{U,EXP} = 0.3333$, the unbiased (unavailable) full response estimate remains $\hat{\bar{y}}_{U,FUL} = 0.4000$. But differences in other respects will make the conclusions differ considerably from those in Section 9.

Table 10.1. Data set II, for a simple random sample of size *n* = 4,000. Frequencies by trait group and overall, shown for responding, non-responding and whole sample

	Trait group			
	$j = 1$ (thereof $\mathcal{Y}$)	$j = 2$ (thereof $\mathcal{Y}$)	$j = 3$ (thereof $\mathcal{Y}$)	All traits (thereof $\mathcal{Y}$)
Response	400 (200)	1200 (330)	800 (270)	2400 (800)
Non-response	1100 ($500^{\times}$)	400 ($220^{\times}$)	100 ($80^{\times}$)	1600 ($800^{\times}$)
Total sample	1500 ($700^{\times}$)	1600 ($550^{\times}$)	900 ($350^{\times}$)	4000 ($1600^{\times}$)

Considerable trait group differences prevail in Table 10.2, for both P_j and $\bar{y}_{r_j}$. This will again lead the analyst to adjust by age group and to publish the Weighting Class estimate $\hat{\bar{y}}_{U,CAL}$ $= (1/N)\sum_{j=1}^{3} \hat{N}_j \, \bar{y}_{r_j;d} = 0.3734$. This is done without assessing the importance of each individual trait, something which is instead accomplished in Table 10.3.

Table 10.2. Data set II. Entries in the first four row are computed by the analyst; those of the bottom row, marked ×, are unavailable

	Trait group			All traits
	$j = 1$	$j = 2$	$j = 3$	
Proportion of sample, Q_{sj}	0.3750	0.4000	0.2250	1
Proportion of response, Q_{rj}	0.1667	0.5000	0.3333	1
Response rate, P_j	0.2667	0.7500	0.8889	$P = 0.6000$
Prevalence of $\mathcal{Y}$ in the response, $\bar{y}_{r_j}$	0.5000	0.2750	0.3375	$\bar{y}_r = 0.3333$
Prevalence of $\mathcal{Y}$ in the sample, $\bar{y}_{s_j}$	$0.4667^{\times}$	$0.3438^{\times}$	$0.3889^{\times}$	$\bar{y}_s = 0.4000^{\times}$

Table 10.3 suggests that the trait $j = 1$ is important (the value $|H_{A1}| = 0.1042$ is large by comparison) and that both $j = 2$ and $j = 3$ are unimportant. The numbers in the $j = 1$ column coincide with those seen before in Table 9.3. But for $j = 2$ and especially for $j = 3$, the situation has radically changed.

Table 10.3. Data set II. Analysis of trait influence. The bottom line entry H_{Aj} is the product of the preceding three entries

	Trait group			Sum
	$j = 1$	$j = 2$	$j = 3$	
Q_{sj}	0.3750	0.4000	0.2250	1
$(P_j - P)/P$	-0.5556	0.2500	0.4815	
$(\bar{y}_{r_j;d} - \bar{y}_{r;d})/\bar{y}_{r;d}$	0.5000	-0.1750	0.0125	
H_{Aj}	-0.1042	-0.0175	0.0014	$H_A = -0.1203$

Table 10.4 reinforces the message in Table 10.3 that the only important trait is $j = 1$. The stepwise procedure will select this trait in Step 1; the **x**-vector is then $\mathbf{x}_k = (1, \gamma_{1k})'$, implying that $2 \cup 3$ forms the rest group. The alternatives $2, 1 \cup 3$ and $3, 1 \cup 2$ are inferior for these data.

Table 10.4. Data set II. Values of cv_m, $R_{y\mathbf{x}}$ and R_{DC}, of their product Δ_A / S_y, and of Δ_A, shown for each of three steps. n.d. stands for "not defined". Standard deviation $S_y = 0.4714$

	Step 0	Step 1 Two trait groups			Step 2 Three trait groups
	$1 \cup 2 \cup 3$	$1, 2 \cup 3$	$2, 1 \cup 3$	$3, 1 \cup 2$	$1, 2, 3$
cv_m	0	0.5590	0.2000	0.2298	0.5618
$R_{y\mathbf{x}}$	0	0.1581	0.1237	0.0063	0.1688
R_{DC}	n.d.	- 1	- 1	+1	-0.8969
Δ_A / S_y	0	-0.0884	-0.0247	0.0014	-0.0851
Δ_A	0	-0.0417	-0.0117	0.0007	-0.0401

Step 2 causes $|R_{DC}|$ to drop significantly from 1 to 0.8969. The modest increases in cv_m and $R_{y\mathbf{x}}$ are not enough to compensate, so $|\Delta_A|$ drops from 0.0417 to 0.0401. Hence, these data provide an example of a reversal in $|\Delta_A|$: The use of all $J_{tot} = 3$ traits does not yield the highest value of $|\Delta_A|$. When we act by the principle to end when $|\Delta_A|$ is at its highest value, $\mathbf{x}_k = (1, \gamma_{1k})'$ remains ultimate **x**-vector, and the finally published Selective Weighting Class estimate is $\hat{\bar{y}}_{U,CAL} = 0.3333 + 0.0417 = 0.3750$. This is closer to the (unavailable) unbiased estimate of 0.4000, and thus preferred to the complete Weighting Class estimate 0.3734 based on all three traits (although the difference for these data is small).

11 Selecting influential traits in the presence of several categorical auxiliary variables

The word influential is to be understood in the sense "important for bias reduction". Many government surveys involve categorical study variables, as when one needs to estimate the number of persons or households with attributes defined by, for example, employment status, or health condition, or drug usage, or choice of post-secondary education.

The auxiliary variables are often also categorical. They include "traditional ones" such as Sex and Age group. In Scandinavia, a variety of others enter into consideration: Income class, Education level, Presence of children, Urban versus rural dwelling, Marital status, Unemployment pattern, Pattern of paid sick leave, Level of debt, Country of birth, and many others.

To build an effective auxiliary vector, we outline here a possible procedure of stepwise forward selection. We expand gradually the dimension of the **x**-vector by adding one trait at a time (rather than one complete auxiliary variable at a time). The procedure derives from the conclusions in Sections 8 to 10 where we considered one single categorical auxiliary variable and the selection of its most influential traits. Here we select from among all those defined by a collection of categorical auxiliary variables.

Consider a set of $I \geq 2$ available categorical auxiliary variables, Age, Income class, Educational level, and so on, where the *i*:th variable has J_i mutually exclusive and exhaustive predefined traits, $i = 1,2,...,I$. Usually, not all $\sum_{i=1}^{I} J_i$ traits are critically important. Some may be of marginal value, others counterproductive from the perspective of achieving a numerically important adjustment $|\Delta_A|$. That is, we may prefer to use the *i*:th variable not in its complete form, but only through a selected few of its J_i traits. That is, out of

the total number $J_{tot} = \sum_{i=1}^{I} J_i$ of available traits, only influential ones will be retained for the **x**-vector, keeping in mind the decomposition $\Delta_A / S_y = \mathbf{D}'\boldsymbol{\Sigma}^{-1}\mathbf{C} / S_y = cv_m \times R_{y\mathbf{x}} \times R_{DC}$.

Which traits should be retained in a stepwise construction of the **x**-vector? As pointed out earlier, the value of $|\Delta_A|$ may decrease, rather than increase, by adding "unnecessary" traits. Both cv_m and $R_{y\mathbf{x}}$ increase by adding more traits, but this does not necessarily hold for $|R_{DC}|$ and therefore not for $|\Delta_A|$. The presence in unit (person) k of the trait j is coded $\gamma_{jk} = 1$; the absence is coded $\gamma_{jk} = 0$, $j = 1, 2, ..., J_{tot}$. Each dichotomous trait indicator γ_{jk} is now viewed as a potential auxiliary variable.

Traits are added one by one to the **x**-vector. A maximum number of $1 + J_{tot} - I$ can be used. The reason is that the *i*:th variable is exhausted when $J_i - 1$ of its J_i traits have been admitted, to avoid a singular matrix in the weight computation. The effective number of available traits is thus $1 + J_{tot} - I$. Consider the following stepwise forward selection algorithm: Include, in a given step, the trait identified by the highest value of $|\Delta_A| = |\mathbf{D}'\boldsymbol{\Sigma}^{-1}\mathbf{C}|$. At Step 0, the auxiliary vector is the primitive $\mathbf{x}_k = 1$, which gives $D_j = C_j = |\Delta_A| = 0$.

In Step 1, $\Delta_A = \mathbf{D}'\boldsymbol{\Sigma}^{-1}\mathbf{C}$ is computed for all J_{tot} traits available, that is, for all **x**-vectors $\mathbf{x}_k = (1, \gamma_{jk})'$, $j = 1, 2, ..., J_{tot}$. The trait for which $|\Delta_A|$ has its largest value is selected, say the trait $j = \mathrm{S1}$, coded by the dichotomous trait indicator γ_{S1k}.

In Step 2, Δ_A is computed for all $J_{tot} - 1$ traits not yet selected, that is, for all three dimensional **x**-vectors $\mathbf{x}_k = (1, \gamma_{S1k}, \gamma_{jk},)'$, $j = 1, 2, ..., J_{tot}$, $j \neq S1$. The trait that gives $|\Delta_A|$ its largest value is selected, and so on, in the succeeding steps.

The critical value of $|\Delta_A|$ (that is, the value of $|\Delta_A|$ that triggers the inclusion of the next trait), will be increasing in a number of steps (for a number of selected traits), until a decline in the critical value is likely to set in. A stopping rule that is recommended (although alternatives could be considered) is to end the procedure at the step where the decline in $|\Delta_A|$ sets in, that is, when $|\Delta_A|$ is no longer increased by the inclusion of yet another trait. In practice, one is then often led to use less than all the J_i traits of the *i*:th categorical variable, but only the most influential ones, up to a maximum number of $J_i - 1$.

12 Concluding comments

This paper begins on the somewhat pessimistic but nevertheless realistic note that the biasing effects of survey nonresponse can never be totally eliminated. It is recognized that : (a) although the bias that nonresponse causes in survey estimates cannot be estimated or quantified, it can be decreased by basing the estimation on an efficient auxiliary vector, and (b) the auxiliary vector is to be built by a judicious selection of efficient auxiliary variables. In environments such as Statistics Sweden, many auxiliary variables become available through the access to many administrative registers. The choice of "the best" among these is the responsibility of the statistician/analyst. Contrary to what one might initially expect, a use of all available auxiliary variables may not be the best action. Tools are needed for the selection of auxiliary variables. To this end we have examined the bias indicator $\Delta_A = \mathbf{D}'\boldsymbol{\Sigma}^{-1}\mathbf{C}$, factorized in formula (4.5) as $\Delta_A / S_y = cv_m \times R_{y\mathbf{x}} \times R_{DC}$. The behavior of the factor R_{DC} is critical. Its value tends to drop when further *x*-variables are added to the vector, and it may drop more than what is compensated for by the increases in cv_m and $R_{y\mathbf{x}}$. In particular, we studied the case where the auxiliary variables are categorical, each defined in terms of a given set of traits or properties. The construction of the auxiliary vector then presents itself as a selection of the most potent traits among those represented by the whole set of categorical auxiliary variables. We outlined one stepwise forward selection procedure of traits. Alternative procedures may have advantages; this is a topic for future investigation.

Referenser

Beaumont, J.F. (2005). On the use of data collection process information for the treatment of unit nonresponse through weight adjustment. *Survey Methodology*, **31**, 227-231.

Crouse, C. and Kott, P.S. (2004). Evaluation alternative calibration schemes for an economic survey with large nonresponse. Proc. Survey Research Methods Section, American Statistical Association.

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

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

Kott, P.S. (2008). Some research on calibration estimators at the National Agricultural Statistics Service. *Pakistan Journal of Statistics*, **24**, 145-161.

Little, R.J. and Vartivarian, S. (2005). Does weighting for nonresponse increase the variance of survey means? *Survey Methodology*, **31**, 161-168.

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

Särndal, C.E. and Lundström, S. (2008). Assessing auxiliary vectors for control of nonresponse bias in the calibration estimator. *Journal of Official Statistics*, **24**, 251-260.

Särndal, C.E. and Lundström, S. (2010). Design for estimation: Identifying auxiliary vectors to reduce nonresponse bias. To appear, *Survey Methodology*.

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

ISSN 1653-7149 (online)

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

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