Component of Statistics Canada
Catalogue no. 12-001-X Business Survey Methods Division

Article

The calibration approach in survey theory and practice

by Carl-Erik Särndal

SURVEY
METHODOLOGY

A JOURNAL
PUBLISHED BY
STATISTICS CANADA

JUNE 2005
•
VOLUME 31
•
NUMBER 1

Canada

December, 2007

Statistics Canada Statistique Canada

Canada

The calibration approach in survey theory and practice

Carl-Erik Särndal [1]

Abstract

Calibration is the principal theme in many recent articles on estimation in survey sampling. Words such as “calibration approach” and “calibration estimators” are frequently used. As article authors like to point out, calibration provides a systematic way to incorporate auxiliary information in the procedure.

Calibration has established itself as an important methodological instrument in large-scale production of statistics. Several national statistical agencies have developed software designed to compute weights, usually calibrated to auxiliary information available in administrative registers and other accurate sources.

This paper presents a review of the calibration approach, with an emphasis on progress achieved in the past decade or so. The literature on calibration is growing rapidly; selected issues are discussed in this paper.

The paper starts with a definition of the calibration approach. Its important features are reviewed. The calibration approach is contrasted with (generalized) regression estimation, which is an alternative but conceptually different way to take auxiliary information into account. The computational aspects of calibration are discussed, including methods for avoiding extreme weights. In the early sections of the paper, simple applications of calibration are examined: The estimation of a population total in direct, single phase sampling. Generalization to more complex parameters and more complex sampling designs are then considered. A common feature of more complex designs (sampling in two or more phases or stages) is that the available auxiliary information may consist of several components or layers. The uses of calibration in such cases of composite information are reviewed. Later in the paper, examples are given to illustrate how the results of the calibration thinking may contrast with answers given by earlier established approaches. Finally, applications of calibration in the presence of nonsampling error are discussed, in particular methods for nonresponse bias adjustment.

Key Words: Auxiliary information; Weighting; Consistency; Design-based inference; Regression estimator; Models; Nonresponse; Complex sampling design.

1. Introduction

1.1 Calibration defined

It is useful in this paper to refer to a definition of the calibration approach. I propose the following formulation.

Definition. The *calibration approach* to estimation for finite populations consists of

(a) a computation of weights that incorporate specified *auxiliary information* and are restrained by *calibration equation*(*s*),

(b) the use of these weights to compute linearly weighted estimates of totals and other finite population parameters: weight times variable value, summed over a set of observed units,

(c) an objective to obtain nearly design unbiased estimates as long as nonresponse and other non-sampling errors are absent.

In the literature, “calibration” frequently refers to (a) alone; I shall often use the term for (a) to (c) together. Earlier definitions, although less extensive, agree essentially with mine. Ardilly (2006) defines calibration (or, more precisely, “calage généralisé”) as a method of re-weighting used when one has access to several variables, qualitative or quantitative, on which one wishes to carry out, jointly, an adjustment.

Kott (2006) defines calibration weights as a set of weights, for units in the sample, that satisfy a calibration to known population totals, and such that the resulting estimator is randomization consistent (design consistent), or, more rigorously, that the design bias is, under mild conditions, an asymptotically insignificant contribution to the estimator’s mean squared error. This is the property I call “nearly design unbiased”.

The Quality Guidelines (fourth edition) of Statistics Canada (2003) say: “Calibration is a procedure than can be used to incorporate auxiliary data. This procedure adjusts the sampling weights by multipliers known as calibration factors that make the estimates agree with known totals. The resulting weights are called calibration weights or final estimation weights. These calibration weights will generally result in estimates that are design consistent, and that have a smaller variance than the Horvitz-Thompson estimator.”

Part (c) of the definition merits a comment. Nothing prevents producing weights calibrated to given auxiliary information without requiring (c). But most published work on calibration is in the spirit of (c), so it makes good sense to include it. When non-sampling errors are present, bias in the estimates is unavoidable, whether they are made by calibration or by any other method. In line with (c), I

[1] Carl-Erik Särndal, 2115 Erinbrook Crescent, #44, Ottawa, Ontario, K1B 4J5, Canada. E-mail: carl.sarndal@rogers.com.

consider design-based inference to be the standard in this paper. The randomization-based variance of an estimator is thus important. However, the paper focuses on “motivations behind (point) estimation”; for reasons of space, the important question of variance estimation is not addressed.

1.2 Comments arising

The definition in Section 1.1 prompts some comments and references to earlier literature:

(1) *Calibration as a linear weighting method.* Calibration has an intimate link to practice. The fixation on weighting methods on the part of the leading national statistical agencies is a powerful driving force behind calibration. To assign an appropriate weight to an observed variable value, and to sum the weighted variable values to form appropriate aggregates, is firmly rooted procedure. It is used in statistical agencies for estimating various descriptive finite population parameters: totals, means, and functions of totals. Weighting is easy to explain to users and other stakeholders of the statistical agencies.

Weighting of units by the inverse of their inclusion probability found firm scientific backing long ago in papers such as Hansen and Hurwitz (1943), Horwitz and Thompson (1952). Weighting became widely accepted. Later, post-stratification weighting achieved the same status. Calibration weighting extends both of these ideas. Calibration weighting is outcome dependent; the weights depend on the observed sample.

Inverse inclusion probability weights are, by definition, greater than or equal to unity. A commonly heard interpretation is that “an observed unit represents itself and a number of others, not observed”. Calibrated weights, on the other hand, are not necessarily greater than or equal to unity, unless special care is taken in the computation to obtain this property.

Calibration is new as a term in survey sampling - about 15 years old - but not as a technique for producing weights. Those who maintain “I practiced calibration long before it was called calibration” have a point. The last 15 years widened the scope and the appeal of the technique. Weighting akin to calibration has long been used by private survey institutes, for example, in connection with quota sampling, a form of non-probability sampling outside the scope of this paper.

Weighting of observed variable values was an important topic before calibration became a popular term. Some authors derived the weights via the argument that they should differ as little as possible from the unbiased sampling design weights (the inverse of the inclusion probabilities). Others found the weights by recognizing that a linear regression estimator can be written as a linearly weighted sum of the observed study variable values. Terms such as “survey sample weighting” and “regression weighting” and “case weighting” are used. Among such “early papers” are Alexander (1987), Bankier, Rathwell and Majkowski (1992), Bethlehem and Keller (1987), Chambers (1996), Fuller, Loughin and Baker (1994), Kalton and Flores-Cervantes (1998), Lemaître and Dufour (1987), Särndal (1982) and Zieschang (1990). I comment later on the technique “repeated weighting”, promoted by the Dutch national statistical agency, CBS. The newer term “calibration” conveys a more specific message and a more definite direction than the older “weighting”.

(2) *Calibration as a systematic way to use auxiliary information*. Calibration provides a systematic way to take auxiliary information into account. As Rueda, Martínez, Martínez and Arcos (2007) point out, “in many standard settings, the calibration provides a simple and practical approach to incorporating auxiliary information into the estimation”.

Auxiliary information was used to improve the accuracy of survey estimates long before calibration became popular. Numerous papers were written with this goal in mind, for more or less specialized situations. Today, calibration does offer a systematic outlook on the uses of auxiliary information. For example, calibration can deal effectively with surveys where auxiliary information exists at different levels. In two-stage sampling information may exist for the first stage sampling units (the clusters), and other information for the second stage sampling units. In surveys with nonresponse (that is, essentially all surveys), information may exist “at the population level” (known population totals), and other information “at the sample level” (auxiliary variable values for all those sampled, responding and non-responding). Calibration with “composite information” is reviewed in Sections 8 and 9.

Regression estimation, or generalized regression (GREG) estimation, competes with calibration as a systematic way to incorporating auxiliary information. It is therefore important to contrast GREG estimation (described in Section 3) with calibration estimation (described in Section 4). The two approaches are different.

(3) *Calibration to achieve consistency*. Calibration is often described as “a way to get consistent estimates”. (Here “consistent” refers not to “randomization consistent” but to “consistent with known aggregates”.) The calibration equations impose consistency on the weight system, so that, when applied to the auxiliary variables, it will confirm (be consistent with) known aggregates for those same auxiliary variables. A desire to promote credibility in published statistics is an often cited reason for demanding consistency. Some users of statistics dislike finding the same population

quantity estimated by two or more numbers that do not agree.

The totals with which consistency is sought are sometimes called control totals. “Controlled weights” or “calibrated weights” suggest improved, more accurate estimation. The French term for calibration, “calage”, has a similar connotation of “stability”.

Consistency through calibration has a broader implication than just agreement with known population auxiliary totals. Consistency can, for example, be sought with appropriately estimated totals, arising in the current survey or in other surveys.

Consistency among tables estimated from different surveys is the motive behind *repeated weighting*, the technique developed at the Dutch national statistical agency CBS in several articles: Renssen and Nieuwenbroek (1997); Nieuwenbroek, Renssen, and Hofman (2000); Renssen, Kroese, and Willeboordse (2001); Knottnerus and van Duin (2006). The stated objective is to accommodate user demands to produce numerically consistent outputs. As the last mentioned paper points out, repeated weighting can be seen as an additional calibration step for a new adjustment of already calibrated weights. The final weights realize consistency with given margins.

Consistency with known or estimated totals may bring the extra benefit of improved accuracy (lower variance and/or reduced nonresponse bias). However, in some articles, especially those authored in statistical agencies, consistency for user satisfaction seems a more imperative motivation than the prospect of increased accuracy.

When the primary motivation for calibration is not so much an agreement with other statistics as rather to reduce variance and/or nonresponse bias, then “balanced weight system” is a more appropriate description than “consistent weight system”, because the objective is then to balance the weights to reflect the outcome of the sampling, the response to the survey, and the information available.

(4) *Calibration for convenience and transparency*. As Harms and Duchesne (2006) point out, “The calibration approach has gained popularity in real applications because the resulting estimates are easy to interpret and to motivate, relying, as they do, on design weights and natural calibration constraints.” Calibration on known totals strikes the typical user as transparent and natural. Users who understand sample weighting appreciate that calibration leaves the design weights “slightly modified only”, while respecting the controls. The unbiasedness is only negligibly disturbed. The simpler forms of calibration invoke no assumptions, only “natural constraints”. Yet another advantage is appreciated by users: In many applications, calibration gives a unique weighting system, applicable to all study variables, of which there are usually many in large government surveys.

(5) *Calibration in combination with other terms*. Some authors use the word “calibration” in combination with other terms, to describe various directions of thought. Examples of this proliferation of terms are: Model-calibration (Wu and Sitter 2001); *g*-calibration (Vanderhoeft, Waeytens and Museux 2000); Harmonized calibration (Webber, Latouche and Rancourt 2000), Higher level calibration (Singh, Horn and Yu 1998); Regression calibration (Demnati and Rao 2004); Non-linear calibration (Plikusas 2006); Super generalized calibration (Calage super généralisé; Ardilly 2006); Neural network model-calibration estimator and Local polynomial model-calibration estimator (Montanari and Ranalli 2003, 2005), Model-calibrated pseudo empirical maximum likelihood estimator (Wu 2003), and yet others. Also, calibration plays a significant role in the indirect sampling methods proposed in Lavalleé (2006). In a somewhat different spirit, not reviewed here, are concepts such as calibrated imputation (Beaumont 2005a), and bias calibration (Chambers, Dorfman and Wehrly (1993), Zheng and Little (2003)). The following review pages do not give justice to all the innovations within the sphere of calibration, but the names alone do suggest directions that have been explored.

(6) *Calibration as a new direction for thought*. If calibration represents “a new approach” with clear differences compared with predecessors, we must examine such questions as: Does calibration generalize earlier theories or approaches? Does calibration give better, more satisfactory answers on questions of importance, as compared with earlier recognized approaches? Sections 4.5 and 7.1 in this paper illustrate how the answers provided by calibration compare with, or contrast with, those obtained in earlier modes of reasoning.

The practice of survey sampling encounters “nuisances” such as nonresponse, frame deficiencies and measurement errors. It is true that imputation and reweighing for nonresponse are widely practiced, through a host of techniques. But they are somehow “separate issues”, still waiting to be more fully embedded into a comprehensive, more satisfactory theory of inference in sample surveys. Many theory papers deal with estimation for an imagined ideal survey, nonexistent in practice, where nonresponse and other non-sampling errors are absent. This is not a criticism of the many excellent but idealized theory papers. The foundations need to be explored, too.

Sections 9 and 10 indicate that calibration can provide a more systematic outlook on inference in surveys even in the presence of the various non-sampling errors. Future fruitful developments are expected in that regard.

2. Basic conditions for design-based estimation in sample surveys

This section sets the background for Sections 3 to 7. By "basic conditions" I will mean single phase probability sampling of elements and full response. In practice, survey conditions are not that simple and perfect, but many theory papers nevertheless address this situation.

A probability sample s is drawn from the finite population $U = \{1, 2, ..., k, ..., N\}$. The probability sampling design generates for element k a known inclusion probability, $\pi_k > 0,$ and a corresponding sampling design weight $d_k = 1/\pi_k$. The value y_k of the study variable y is recorded for all $k \in s$ (complete response). The objective is to estimate a population total $Y = \sum_U y_k$ with the use of auxiliary information. The study variable y may be continuous or, as in many government surveys, categorical. For example, if y is dichotomous with value $y_k = 0$ or $y_k = 1$ according as person k is employed or unemployed, then the parameter $Y = \sum_U y_k$ to be estimated is the population count of unemployed people. (If $A \subseteq U$ is a set of elements, I write $\sum_A$ for $\sum_{k \in A}$.) The basic design unbiased estimator of Y is $\hat{Y}_{\text{HT}} = \sum_s d_k y_k$, the Horwitz-Thompson estimator. It is, however, inefficient when powerful auxiliary information is available for use at the estimation phase.

The general notation for the auxiliary vector will be $\mathbf{x}_k$. In some countries, for some surveys, the sources of auxiliary data permit extensive vectors $\mathbf{x}_k$ to be built. But some examples of simple vectors are: (1) $\mathbf{x}_k = (1, x_k)'$, where x_k is the value for element k of a continuous auxiliary variable x; (2) the classification vector used to code membership in one of P mutually exclusive and exhaustive groups, $\mathbf{x}_k = \boldsymbol{\gamma}_k = (\gamma_{1k}, ..., \gamma_{pk}, ..., \gamma_{Pk})'$, so that, for $p = 1, 2, ..., P$, $\gamma_{pk} = 1$ if k belongs to group p, and $\gamma_{pk} = 0$ 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', the dimension of $\mathbf{x}_k$ being $P + Q - 1$, where P and Q are the respective number of categories, and the 'minus-one' is to avoid a singular matrix in the computation of weights calibrated "to the margins"; (5) the extension of (4) to more than two 'side-by-side' categorical classifications. Cases 4 and 5 are particularly important for production in national statistical agencies.

In calibration reasoning it is crucially important to specify exactly the *auxiliary information*. Under the basic conditions we need to distinguish two different cases relative to $\mathbf{x}_k$:

(i) $\mathbf{x}_k$ is a known vector value for every $k \in U$ (complete auxiliary information)

(ii) $\sum_U \mathbf{x}_k$ is known (imported) total, and $\mathbf{x}_k$ is known (observed) for every $k \in s$

It is often the survey environment that dictates whether (i) or (ii) prevails. Case (i), complete auxiliary information, occurs when $\mathbf{x}_k$ is specified in the sampling frame for every $k \in U$ (and thus known for every $k \in s$). This environment is typical of surveys on individuals and households in Scandinavia and other North European countries equipped with high quality administrative registers that can be matched with the frame to provide a large number of potential auxiliary variables. The population total $\sum_U \mathbf{x}_k$ is obtained simply by adding the $\mathbf{x}_k$.

Case (i) gives considerable freedom in structuring the auxiliary vector $\mathbf{x}_k$. For example, if x_k is a continuous variable value specified for every $k \in U$, then we are invited to consider x_k^2 and other functions of x_k for inclusion in $\mathbf{x}_k$, because totals such as $\sum_U x_k^2$ and $\sum_U \log x_k$ are readily computed. If the relationship to the study variable y is curved, it may be a serious omission not to take into account known totals such as the quadratic one or the logarithmic one.

Case (ii) prevails in surveys where (i) is not met, but where $\sum_U \mathbf{x}_k$ is imported from an outside source considered accurate enough, and the individual value $\mathbf{x}_k$ is available (observed in data collection) for every $k \in s$. Then $\sum_U \mathbf{x}_k$ is sometimes called an "independent control total", to mark its origin from outside the survey itself. Case (ii) is less flexible: If x_k is a variable with a total $\sum_U x_k$ imported from a reliable source, then $\sum_U x_k^2$ may be unavailable, barring x_k^2 from inclusion into $\mathbf{x}_k$.

3. Generalized regression estimation under the basic conditions

3.1 The GREG concept

Before examining calibration, let us consider *generalized regression* (GREG) *estimation* (or just *regression estimation*), for two good reasons: (1) GREG estimation can also be claimed to be a systematic way to take auxiliary information into account; (2) some (but not all) GREG estimators are calibration estimators, in that they can be expressed in terms of a calibrated linear weighting.

GREG estimators and calibration estimators have been extensively studied in the last two decades. The terms alone, "GREG estimation" and "calibration estimation", reflect a clear difference in thinking. Statisticians who work in the area are of two types: Those dedicated to "GREG thinking" and those dedicated to "calibration thinking". The distinction may not be completely clear-cut, but it helps structuring this review paper, so I will use it. I am not venturing to say that the latter thinking is more prevalent in national statistical agencies and the former more prevalent in the academic circles, but perhaps there is such a tendency.

The GREG estimator concept evolved gradually since the mid-1970's. The simple (linear) GREG is explained in Särndal, Swensson and Wretman (1992); a thorough review of regression estimation is given in Fuller (2002). The central idea is that predicted y-values $\hat{y}_k$ can be produced for all N population elements, via the fit of an *assisting model* and the use of the auxiliary vector values $\mathbf{x}_k$ known for all $k \in U$. The predicted values serve to build a nearly design unbiased estimator of the population total $Y = \sum_U y_k$ as

$$\begin{aligned}\hat{Y}_{\text{GREG}} &= \sum\nolimits_U \hat{y}_k + \sum\nolimits_s d_k (y_k - \hat{y}_k) \\ &= \sum\nolimits_s d_k y_k + \left(\sum\nolimits_U \hat{y}_k - \sum\nolimits_s d_k \hat{y}_k\right). \qquad (3.1)\end{aligned}$$

The obvious motivation behind this construction is the prospect of a highly accurate estimate $\hat{Y}_{\text{GREG}}$ through a close fitting assisting model that leaves small residuals $y_k - \hat{y}_k$. That modeling is the corner stone of GREG thinking. Some authors use the (also justifiable) name *general difference estimator* for the construction (3.1).

The great variety of possible assisting models generates a wide family of GREG estimators of the form (3.1). The assisting model, an imagined relationship between $\mathbf{x}$ and y, can have many forms: linear, non-linear, generalized linear, mixed (model with some fixed, some random effects), and so on. Whatever the choice, the model is "assisting only"; even though it may be short of "true", (3.1) is nearly deign unbiased under mild conditions on the assisting model and on the sampling design, so that $(\hat{Y}_{\text{GREG}} - Y)/N = O_p(n^{-1/2})$ and $(\hat{Y}_{\text{GREG}} - Y)/N = (\hat{Y}_{\text{GREG,lin}} - Y)/N + O_p(n^{-1})$, where the statistic $\hat{Y}_{\text{GREG,lin}}$, the result of linearizing $\hat{Y}_{\text{GREG}}$, is unbiased for Y.

3.2 Linear GREG

By linear GREG I mean one that is generated by a linear fixed effects assisting model. The predictions are $\hat{y}_k = \mathbf{x}_k' \mathbf{B}_{s;dq}$ with

$$\mathbf{B}_{s;dq} = \left(\sum\nolimits_s d_k q_k \mathbf{x}_k \mathbf{x}_k'\right)^{-1} \left(\sum\nolimits_s d_k q_k \mathbf{x}_k y_k\right)$$

so (3.1) becomes

$$\hat{Y}_{\text{GREG}} = \left(\sum\nolimits_U \mathbf{x}_k\right)' \mathbf{B}_{s;dq} + \sum\nolimits_s d_k (y_k - \mathbf{x}_k' \mathbf{B}_{s;dq}). \qquad (3.2)$$

The q_k are scale factors, chosen by the statistician. The standard choice is $q_k = 1$ for all k. The choice of the q_k has some (but often limited) impact on the accuracy of $\hat{Y}_{\text{GREG}}$; near-unbiasedness holds for any specification (barring outrageous choices) for the q_k. Although the model is simple, the linear GREG (3.2) contains many estimators, considering the many possible choices of the auxiliary vector $\mathbf{x}_k$ and the scale factors q_k. Under general conditions,

$$(\hat{Y}_{\text{GREG}} - Y)/N = \left(\sum\nolimits_s d_k E_k - \sum\nolimits_U E_k\right)\Big/N + O_p(n^{-1})$$

where $\sum_s d_k E_k$ is the Horvitz-Thompson estimator in the residuals $E_k = y_k - \mathbf{x}_k' \mathbf{B}_{U;q}$ with $\mathbf{B}_{U;q} = (\sum_U q_k \mathbf{x}_k \mathbf{x}_k')^{-1}$ $(\sum_U q_k \mathbf{x}_k y_k)$. Hence, the design-based properties $E(\hat{Y}_{\text{GREG}}) \approx Y$ and $\text{Var}(\hat{Y}_{\text{GREG}}) \approx \text{Var}(\sum_s d_k E_k)$. A close fitting *linear* regression of y on $\mathbf{x}$ holds the key to a small variance for $\hat{Y}_{\text{GREG}}$ (and this is very different from claiming that "a linear regression is the true regression").

The linear GREG in Särndal, Swensson and Wretman (1992) was motivated via the linear assisting model ξ stating that $E_\xi(y_k) = \boldsymbol{\beta}' \mathbf{x}_k$ and $V_\xi(y_k) = \sigma_k^2$. Generalized least squares fit gives the estimator (3.2) with $q_k = 1/\sigma_k^2$. In that context, an educated guess about the variation of the residuals $y_k - \boldsymbol{\beta}' \mathbf{x}_k$ determines the q_k. When the vector $\mathbf{x}_k$ is fixed, the modeling effort boils down to an opinion about the residual pattern. The choice $\sigma_k^2 = \sigma^2 x_k$ gives the classical ratio estimator. If $q_k = \boldsymbol{\mu}' \mathbf{x}_k$ for all $k \in U$ and a constant vector $\boldsymbol{\mu}$, then (3.2) reduces to "the cosmetic form" $(\sum_U \mathbf{x}_k)' \mathbf{B}_{s;dq}$.

As Beaumont and Alavi (2004) and others have pointed out, the linear GREG estimator is bias-robust (nearly unbiased although the assisting model falls short of "correct"), but it can be considerably less efficient (have larger mean squared error) than model dependent alternatives which, although biased, may have a considerably smaller variance. Thus one may claim that linear GREG is not variance robust; nevertheless, it is a basic concept in design-based survey theory.

The specification of $\mathbf{x}_k$ should include variables (with known population totals) that served already in defining the sampling design. Design stage information should not be relinquished at the estimation stage; instead, a "repeated usage" is recommended. For example, in stratified simple (STSI) random sampling, the vector $\mathbf{x}_k$ in estimator (3.2) should include, along with other available variables, the dummy coded stratum identifier, $\boldsymbol{\gamma}_k = (\gamma_{k1}, \gamma_{k2}, \ldots, \gamma_{kh}, \ldots, \gamma_{kH})'$, where $\gamma_{kh} = 1$ if element k belongs to stratum h, and $\gamma_{kh} = 0$ if not; $h = 1, \ldots, H$.

We can write the linear GREG (3.2) as a weighted sample sum, $\hat{Y}_{\text{GREG}} = \sum_s w_k y_k$, with

$$\begin{aligned} &w_k = d_k g_k;\ g_k = 1 + q_k \boldsymbol{\lambda}' \mathbf{x}_k; \\ &\boldsymbol{\lambda}' = \left(\sum\nolimits_U \mathbf{x}_k - \sum\nolimits_s d_k \mathbf{x}_k\right)' \left(\sum\nolimits_s d_k q_k \mathbf{x}_k \mathbf{x}_k'\right)^{-1}. \qquad (3.3)\end{aligned}$$

The weights w_k happen to be *calibrated* to (consistent with) the known population $\mathbf{x}$-total: $\sum_s w_k \mathbf{x}_k = \sum_U \mathbf{x}_k$. That $\hat{Y}_{\text{GREG}}$ is expressible as a linearly weighted sum with calibrated weights is a fortuitous by-product. It is not part of GREG thinking, whose central idea formulated in (3.1) is the fit of an assisting model. A few other GREG's than the simple linear one also have the calibration property, as will be noted later.

3.3 Non-linear GREG

Two features of the linear GREG (3.2) make it a favourite choice for routine production in statistical agencies: (i) the auxiliary population total $\sum_U \mathbf{x}_k$ becomes factored out, so the estimation can proceed as long as an accurate value for that total can be computed or imported, and (ii) when written as the linearly weighted sum $\hat{Y}_{\text{GREG}} = \sum_s w_k y_k$, the weight system (3.3) is independent of the y-variable and can thereby be applied to all y-variables in the survey. We need not know $\mathbf{x}_k$ individually for all $k \in U$; knowing $\sum_U \mathbf{x}_k$ suffices. Needless to say, if we do know all $\mathbf{x}_k$, more efficient (still nearly design unbiased) members of the GREG family (3.1) can be sought. This will also counter another criticism of the linear GREG, namely that a linear model is unrealistic for some types of data. For example, for a dichotomous y-variable, a logistic assisting model may be both more realistic and yield a more precise GREG estimator.

By a non-linear GREG estimator I mean one generated as in (3.1) by an assisting model of other type than "linear in $\mathbf{x}_k$ with fixed effects". Among the first to extend the GREG concept in this direction are Firth and Bennett (1998) and Lehtonen and Veijanen (1998); see also Chambers *et al.* (1993). In the last few years, several authors have studied model-assisted non-linear GREG's.

Non-linear GREG is a versatile idea; a variety of estimators become possible via assisting models ξ of the following type:

$$E_\xi(y_k|\mathbf{x}_k) = \mu_k \quad \text{for} \quad k \in U \qquad (3.4)$$

where the model mean μ_k and the model variance $V_\xi(y_k|\mathbf{x}_k)$ are given appropriate formulations.

One application of (3.4) is when $\mu_k = \mu(\mathbf{x}_k, \boldsymbol{\theta})$ is a specified non-linear function in $\mathbf{x}_k$. Having estimated $\boldsymbol{\theta}$ by $\hat{\boldsymbol{\theta}}$, the fitted values needed for $\hat{Y}_{\text{GREG}}$ in (3.1) are $\hat{y}_k = \mu(\mathbf{x}_k, \hat{\boldsymbol{\theta}})$ for $k \in U$. For example, if the modeler specifies $\log \mu_k = \alpha + \beta\, x_k$, the predictions for use in (3.1) are, following parameter estimation, $\hat{y}_k = \exp(\hat{\alpha} + \hat{\beta}\, x_k)$.

Other applications of (3.4) include generalized linear models such that $g(\mu_k) = \mathbf{x}_k'\boldsymbol{\theta}$, for a specified link function $g(\cdot)$, and $V_\xi(y_k|\mathbf{x}_k) = v(\mu_k)$ is given an appropriate structure. We estimate $\boldsymbol{\theta}$ by $\hat{\boldsymbol{\theta}}$, the fitted values needed for the non-linear GREG estimator (3.1) are $\hat{y}_k = \hat{\mu}_k = g^{-1}(\mathbf{x}_k'\hat{\boldsymbol{\theta}})$. For example, using a logistic assisting model, $\mathbf{x}_k'\boldsymbol{\theta} = \text{logit}(\mu_k) = \log(\mu_k/(1-\mu_k))$, and $\hat{y}_k = \hat{\mu}_k = \exp(\mathbf{x}_k'\hat{\boldsymbol{\theta}})/(1+\exp(\mathbf{x}_k'\hat{\boldsymbol{\theta}}))$.

Lehtonen and Veijanen (1998) examine the case of a categorical study variable with I classes, $i = 1, 2, ..., I$, $y_{ik} = 1$ if element k belongs to category i, and $y_{ik} = 0$ if not. For example, in a Labour Force Survey with I = 3 categories, "employed", "not employed" and "not in the labour force", an objective is to estimate the respective population counts $Y_i = \sum_U y_{ik}$, $i = 1, 2, 3$. These authors use the logistic assisting model

$$E_\xi(y_{ik}|\mathbf{x}_k) = \mu_{ik};\, \mu_{ik} = \exp(\mathbf{x}_k'\boldsymbol{\theta}_i) \Bigg/ \left(1 + \sum_{i=2}^{I} \exp(\mathbf{x}_k'\boldsymbol{\theta}_i)\right). \quad (3.5)$$

Estimates $\hat{\boldsymbol{\theta}}_i$ of the $\boldsymbol{\theta}_i$ are obtained by maximizing design-weighted log-likelihood. The resulting predictions $\hat{y}_{ik} = \hat{\mu}_{ik}$ are used to form $\hat{Y}_{i\,\text{GREG}} = \sum_U \hat{y}_{ik} + \sum_s d_k (y_{ik} - \hat{y}_{ik})$, for $i = 1, 2, ..., I$.

Another development is the application of GREG reasoning to estimation for domains, as in Lehtonen, Särndal and Veijanen (2003, 2005) and Myrskylä (2007). Mixed models are used in the first two of these papers to assist the non-linear GREG. Let U_a be a domain, $U_a \subset U$, whose total $Y_{ia} = \sum_{U_a} y_{ik}$ we wish to estimate, $i = 1, 2, ..., I$. The 2005 paper derives the predictions for the non-linear GREG from the logistic mixed model stating that for $k \in U_a$

$$E_\xi(y_{ik}|\mathbf{x}_k; \mathbf{u}_{ia}) = \exp(\mathbf{x}_k'\boldsymbol{\theta}_{ia}) \Bigg/ \left(1 + \sum_{i=2}^{I} \exp(\mathbf{x}_k'\boldsymbol{\theta}_{ia})\right) \qquad (3.6)$$

with $\boldsymbol{\theta}_{ia} = \boldsymbol{\beta}_i + \mathbf{u}_{ia}$, where $\mathbf{u}_{ia}$ is a vector of domain specific random deviations from the fixed effects vector $\boldsymbol{\beta}_i$.

Non-linear GREG's assisted by models such as (3.5) and (3.6) require model fitting for every y-variable separately; there is no uniformly applicable weight system. However, the question arises: Are there examples of non-linear GREG's such that the practical advantages of linear GREG are preserved, that is, a linearly weighted form with calibrated weights independent the y-variable. The answer is in the affirmative. Two directions in recent literature are of interest in this regard:

Breidt and Opsomer (2000), Montanari and Ranalli (2005) consider model-assisted local polynomial GREG estimators, for the case of a single continuous auxiliary variable with values x_k known for all $k \in U$. Several choices have to be made in the process: (1) the order q of the local polynomial expression, (2) the specification of the kernel function, and (3) the value of the band width. The resulting estimator can be expressed in terms of weights calibrated with respect to population totals of the powers of x_k, so that $\sum_s w_k x_k^j = \sum_U x_k^j$ for $j = 0, 1, ..., q$.

Breidt, Claeskens and Opsomer (2005) develop a penalized spline GREG estimator for a single x-variable; the assisting model is $m(x; \boldsymbol{\beta}) = \beta_0 + \beta_1 x + ... + \beta_q x^q + \sum_{j=1}^{K} \beta_{q+j} (x - \kappa_j)_+^q$, where $(t)_+^q = t^q$ if $t > 0$ and 0 otherwise, q is the degree of the spline, and the κ_k are suitably spaced knots, for example, uniformly spaced sample quantiles of the x_k-values. After estimation of the β-parameters, they obtain the predictions $\hat{y}_k = m(x_k; \hat{\boldsymbol{\beta}})$ needed for the general GREG formula (3.1). The authors point out that the resulting GREG estimator is calibrated for the parametric portion of the model, that is, $\sum_s w_k x_k^j = \sum_U x_k^j$ for $j = 0, 1, ..., q$, and also for the truncated polynomial terms in the model as long a they are left unpenalized.

We can summarize GREG estimation as follows. The linear GREG has practical advantages for large scale statistics production: It can be expressed as a linearly weighted sum of y_k-values with weights calibrated to $\sum_U \mathbf{x}_k$, the weights are independent of the y_k-values and may be applied to all y-variables in the survey. It is sufficient to know a population auxiliary total $\sum_U \mathbf{x}_k$, imported from a reliable source. Non-linear GREG may give a considerably reduced variance, as a result of the more refined models that can be considered when there is complete auxiliary information (known $\mathbf{x}_k$ for all $k \in U$); near design unbiasedness is preserved. Certain non-linear GREG's can be written as linearly weighted sums.

In academic exercises with artificially created populations and relationships, one can provoke situations where a nonlinear GREG has a large variance advantage over a linear GREG. Such experiments are important for illustration. However, to meet the daily production needs in national statistical agencies; "farfetched" nonlinear GREG's seem to be of fairly remote interest at this point in time; the assisting models for GREG must meet requirements of robustness and practicality. The attraction of a minor reduction of the sampling variance is swept away by worries about other (non-sampling) errors and troubles in the daily production process.

The progression from linear to non-linear GREG creates opportunities and generates questions. What is the most appropriate formulation of the model expectation μ_k? How sensitive are the results to the specification of the variance part of the assisting model? To what extent is computational efficiency an issue? Further research will respond more fully to these questions.

4. The calibration approach to estimation

4.1 Calibration under basic conditions

A crucial step in the GREG approach reviewed in the previous section is to produce predicted values $\hat{y}_k$ through the fit of an assisting model. By contrast, the calibration approach, as defined in Section 1.1, does not refer explicitly to any model. It emphasizes instead the information on which one can calibrate. A key element of "calibration thinking" is the linear weighting of the observed y-values, with weights made to confirm computable aggregates. This conceptual difference will sometimes lead to different estimators in the two approaches.

The calibration approach has considerable generality; it can deal with a variety of conditions: complex sampling designs, adjustments for nonresponse and frame errors. This section, however, focuses on the basic conditions in Section 2: single phase sampling and full response. The notation remains as in Section 2. The material available for estimating the population total $Y = \sum_U y_k$ is: (i) the study variable values y_k observed for $k \in s$, (ii) the known design weights $d_k = 1/\pi_k$ for $k \in U$, and (iii) the known vector values $\mathbf{x}_k$ for $k \in U$ (or an imported total $\sum_U \mathbf{x}_k$). These simple conditions prevail in Deville and Särndal (1992) and Deville, Särndal and Sautory (1993), papers which gave the approach a name and inspired further work. Even though the background is simple, calibration raises several issues, some of them computational, as reviewed in Section 5.

The objective in Sections 4.2 and 4.3 is to determine weights w_k to satisfy the calibration equation $\sum_s w_k \mathbf{x}_k = \sum_U \mathbf{x}_k$, then use them to form the calibration estimator of Y as $\hat{Y}_{\text{CAL}} = \sum_s w_k y_k$, which we can confront with the unbiased Horvitz-Thompson estimator by writing $\hat{Y}_{\text{CAL}} = \hat{Y}_{\text{HT}} + \sum_s (w_k - d_k) y_k$. It follows that the bias of $\hat{Y}_{\text{CAL}}$ is $E(\hat{Y}_{\text{CAL}}) - Y = E(\sum_s (w_k - d_k) y_k)$. Meeting the objective of near design unbiasedness requires $E(\sum_s (w_k - d_k) y_k) \approx 0$, whatever the y-variable. Evidently, the calibration should strive for small deviations $w_k - d_k$.

The objective "calibration for consistency with known population auxiliary totals" can be realized in many ways. We can construct many sets of weights calibrated to the known $\sum_U \mathbf{x}_k$. This section examines this proliferation from two perspectives noted in the literature: the *minimum distance method* and the *instrumental vector method*. Yet another construction of a variety of calibrated weights is proposed in Demnati and Rao (2004).

4.2 The minimum distance method

In this method, the calibration sets out to modify the initial weights $d_k = 1/\pi_k$ into new weights w_k, determined to "be close to" the d_k. To this end, consider the distance function $G_k(w, d)$, defined for every $w > 0$, such that $G_k(w, d) \geq 0$, $G_k(d, d) = 0$, differentiable with respect to w, strictly convex, with continuous derivative $g_k(w, d) = \partial G_k(w, d) / \partial w$ such that $g_k(d, d) = 0$. Usually the distance function is chosen such that $g_k(w, d) = g(w/d)/q_k$, where the q_k are suitably chosen positive scale factors, $g(\cdot)$ is a function of a single argument, continuous, strictly increasing, with $g(1) = 0, g'(1) = 1$. Let $F(u) = g^{-1}(u)$ be the inverse function of $g(\cdot)$. Minimizing the total distance

$\sum_s G_k(w_k, d_k)$ subject to the calibration equation $\sum_s w_k \mathbf{x}_k = \sum_U \mathbf{x}_k$ leads to $w_k = d_k F(q_k \mathbf{x}_k' \boldsymbol{\lambda})$, where $\boldsymbol{\lambda}$ is obtained as the solution (assuming one exists) of

$$\sum\nolimits_s d_k \mathbf{x}_k F(q_k \mathbf{x}_k' \boldsymbol{\lambda}) = \sum\nolimits_U \mathbf{x}_k. \qquad (4.1)$$

The weights have an optimality property, because a duly specified objective function is minimized, but it is a "weak optimality" in the sense that there are many possible specifications of the distance function and the scale factors q_k.

Much attention has focused on the distance function $G_k(w_k, d_k) = (w_k - d_k)^2 / 2d_k q_k$. It gives $g_k(w_k, d_k) = (w_k / d_k - 1) / q_k$; $g(w/d) = w/d - 1$; $F(u) = g^{-1}(u) = 1 + u$. The term "the linear case" is thus appropriate. The task is then to minimize the "chi-square distance" $\sum_s (w_k - d_k)^2 / 2d_k q_k$, subject to $\sum_s w_k \mathbf{x}_k = \sum_U \mathbf{x}_k$. Equation (4.1) reads $\sum_s d_k \mathbf{x}_k (1 + q_k \mathbf{x}_k' \boldsymbol{\lambda}) = \sum_U \mathbf{x}_k$, which is easily solved for $\boldsymbol{\lambda}$. The resulting estimator of $Y = \sum_U y_k$ is $\hat{Y}_{\text{CAL}} = \sum_s w_k y_k$ with weights $w_k = d_k g_k$ given by (3.3). That is, $\hat{Y}_{\text{CAL}} = \hat{Y}_{\text{GREG}}$ as given by (3.2), and the residuals that determine the asymptotic variance are $E_k = y_k - \mathbf{x}_k' \mathbf{B}_{U;q}$ as given in Section 3.2. Some negative weights w_k may occur.

The linear GREG estimator implies weights that happen to be calibrated (to $\sum_U \mathbf{x}_k$), and the opposite side of the same coin says that the linear case for calibration (with chi-square distance) brings the linear GREG estimator. The tendency in some articles and applications to intertwine GREG thinking and calibration thinking stems from this fact. Many successful applications of the use of auxiliary information stem, in any case, from this linearity on both sides of the coin. The Canadian Labour Force Survey is an example, and an interesting recent development for that survey is the use of composite estimators, with part of the information coming from the survey results in previous months, as described in Fuller and Rao (2001).

The calibration equation is satisfied for any choice of the positive scale factors q_k in (4.1). A simple choice is $q_k = 1$ for all *k*. But it is not always the preferred choice. For example, if there is a single, always positive auxiliary variable, and $\mathbf{x}_k = x_k$, then many will intuitively expect $\hat{Y}_{\text{CAL}} = \sum_s w_k y_k$ to deliver the usual ratio estimator $\sum_U x_k (\sum_s d_k y_k) / (\sum_s d_k x_k)$, and it does, but by taking $q_k = x_k^{-1}$, not $q_k = 1$.

Another distance function of considerable interest is $G_k(w_k, d_k) = \{w_k \log(w_k / d_k) - w_k + d_k\} / q_k$. It leads to $F(u) = g^{-1}(u) = \exp(u)$, "the exponential case". Then (4.1) reads $\sum_s d_k \mathbf{x}_k \exp(q_k \mathbf{x}_k' \boldsymbol{\lambda}) = \sum_U \mathbf{x}_k$. Numeric methods are required to solve for $\boldsymbol{\lambda}$, to obtain the weights $w_k = d_k \exp(q_k \mathbf{x}_k' \boldsymbol{\lambda})$. No negative weights w_k will occur.

Deville and Särndal (1992) show that a variety of distance functions satisfying mild conditions will generate asymptotically equivalent calibration estimators. Alternative distance functions are compared in Deville, Särndal and Sautory (1993), Singh and Mohl (1996), Stukel, Hidiroglou and Särndal (1996). Some distance functions will guarantee weights falling within specified bounds, so as to rule out too large or too small (negative) weights. Changes in the distance function will often have minor effect only on the variance of the calibration estimator $\hat{Y}_{\text{CAL}} = \sum_s w_k y_k$, even if the sample size is rather small. Questions about the existence of a solution to the calibration equation are discussed in Théberge (2000).

4.3 The instrument vector method

An alternative to distance minimization is the instrumental vector method, considered in Deville (1998), Estevao and Särndal (2000, 2006) and Kott (2006). It can also generate many alternative sets of weights calibrated to the same information.

We can consider weights of the form $w_k = d_k F(\boldsymbol{\lambda}' \mathbf{z}_k)$, where $\mathbf{z}_k$ is a vector with values defined for $k \in s$ and sharing the dimension of the specified auxiliary vector $\mathbf{x}_k$, and the vector $\boldsymbol{\lambda}$ is determined from the calibration equation $\sum_s w_k \mathbf{x}_k = \sum_U \mathbf{x}_k$. The function $F(\cdot)$ plays the same role as in the distance minimization method; several choices $F(\cdot)$ are of interest, for example, $F(u) = 1 + u$ and $F(u) = \exp(u)$.

Opting for the linear function $F(u) = 1 + u$, we have $w_k = d_k (1 + \boldsymbol{\lambda}' \mathbf{z}_k)$. It is an easy exercise to determine $\boldsymbol{\lambda}$ to satisfy the calibration equation $\sum_s w_k \mathbf{x}_k = \sum_U \mathbf{x}_k$. The resulting calibration estimator is

$$\hat{Y}_{\text{CAL}} = \sum\nolimits_s w_k y_k;\ w_k = d_k (1 + \boldsymbol{\lambda}' \mathbf{z}_k),$$

$$\boldsymbol{\lambda}' = \left(\sum\nolimits_U \mathbf{x}_k - \sum\nolimits_s d_k \mathbf{x}_k\right)' \left(\sum\nolimits_s d_k \mathbf{z}_k \mathbf{x}_k'\right)^{-1}. \qquad (4.2)$$

Whatever the choice of $\mathbf{z}_k$, the weights $w_k = d_k (1 + \boldsymbol{\lambda}' \mathbf{z}_k)$ satisfy the calibration equation. The standard choice is $\mathbf{z}_k = \mathbf{x}_k$. In particular, setting $\mathbf{z}_k = q_k \mathbf{x}_k$, for specified q_k, gives the weights (3.3).

Even "deliberately awkward choices" for $\mathbf{z}_k$ give surprisingly good results. For example, let x_k be a single continuous auxiliary variable, and $\mathbf{z}_k = c_k x_k^{p-1}$. Suppose $p = 3$, and $c_k = 1$ for 4 elements only, chosen at random from $n = 100$ elements in a realized sample *s*, and $c_k = 0$ for the remaining 96. The near-unbiasedness of $\hat{Y}_{\text{CAL}} = \sum_s d_k (1 + \boldsymbol{\lambda}' \mathbf{z}_k) y_k$ is still present. Even with such a sparse **z**-vector, the increase in variance, relative to better choices of $\mathbf{z}_k$, may not be excessive.

When both sampling design and **x**-vector are fixed, Estevao and Särndal (2004) and Kott (2004) note that there is an asymptotically optimal **z**-vector given by

$$\mathbf{z}_k = \mathbf{z}_{0k} = d_k^{-1} \sum\nolimits_{\ell \in s} (d_k d_\ell - d_{k\ell}) \mathbf{x}_\ell$$

where $d_{k\ell}$ is the inverse of the second order inclusion probability $\pi_{k\ell} = P(k \,\&\, \ell \in s)$, assumed strictly positive. The resulting calibration estimator, $\hat{Y}_{\text{CAL}} = \sum_s d_k(1+\boldsymbol{\lambda}'\mathbf{z}_{0k})y_k$, is essentially the "randomization-optimal estimator" due originally to Montanari (1987) and discussed by many since then.

Andersson and Thorburn (2005) view the question from the opposite direction and ask: In the minimum distance method, can a distance function be specified such that its minimization will deliver the randomization-optimal estimator? They do find this distance; not entirely surprisingly, it is related to (but not identical to) the chi-square distance.

4.4 Does calibration need an explicitly stated model?

The calibration approach as presented in Sections 4.2 and 4.3 proceeds by simply computing the weights that reproduce the specified auxiliary totals. There is no explicit assisting model, unless one were to insist that picking certain variables for inclusion in the vector $\mathbf{x}_k$ amounts to a serious modeling effort. Instead, the weights are justified primarily by their consistency with the stated controls. Early contributions reflect this attitude, from Deming (1943), and continuing with Alexander (1987), Zieschang (1990) and others. This begs the question: Is it nevertheless important to motivate such "model-free calibration" with an explicit model statement? It is true that statisticians are trained to think in terms of models, and they feel more or less compelled to always have a statistical procedure accompanied by a model statement. It may indeed have some pedagogical merit, also in explaining calibration, to state the associated relationship of y to $\mathbf{x}$, even if it is as simple as a standard linear model.

But will a stated model help the users and practitioners better understand the calibration approach? To most of them the approach is perfectly clear and transparent anyway. They need no other justification than the consistency with stated controls. Will a search for "the true model with the true variance structure" bring significantly better accuracy for the bulk of the many estimates produced in a large government survey? It is unlikely.

The next section deals with model-calibration. For that variety, proposed by Wu and Sitter (2001), modeling has indeed an explicit and prominent role. These authors call the linear calibration estimator, $\hat{Y}_{\text{CAL}} = \sum_s w_k y_k$ with weights w_k given by (3.3), "a routine application without modeling". The description is appropriate in that all that is necessary is to identify the x-variables with their known population totals.

4.5 Model-calibration

The idea of model-calibration is proposed in Wu and Sitter (2001) and pursued further in Wu (2003) and Montanari and Ranalli (2003, 2005). The motivating factor is that complete auxiliary information allows a more effective use of the $\mathbf{x}_k$ known for every $k \in U$ than what is possible in model-free calibration, where a known total $\sum_U \mathbf{x}_k$ is sufficient. The weights are required to be consistent with the computable population total of the predictions $\hat{y}_k$, derived via an appropriate model formulation. Thus the weight system may not be consistent with the known population total of each auxiliary variable, unless there is special provision to retain this property. Model-calibration still satisfies all three parts, (a) to (c), of the definition of calibration in Section 1.1; in particular, the estimators are nearly design unbiased.

Consider a non-linear assisting model of the type (3.4). We estimate the unknown parameter $\boldsymbol{\theta}$ by $\hat{\boldsymbol{\theta}}$, leading to fitted values $\hat{y}_k = \hat{\mu}_k = \mu(\mathbf{x}_k, \hat{\boldsymbol{\theta}})$ computed with the aid of the $\mathbf{x}_k$ known for all $k \in U$. It follows that the population size N is known and should be brought to play a significant role in the calibration. If minimum chi-square distance is used, we find the weights of the model-calibration estimator $\hat{Y}_{\text{MCAL}} = \sum_s w_k y_k$ by minimizing $\sum_s (w_k - d_k)^2 / (2 d_k q_k)$, for specified q_k, and $d_k = 1/\pi_k$, subject to the calibration equations

$$\sum\nolimits_s w_k = N; \sum\nolimits_s w_k \hat{y}_k = \sum\nolimits_U \hat{y}_k. \qquad (4.3)$$

For simplicity, let us take $q_k = 1$ for all k; we derive the calibrated weights, rearrange terms and find that the model-calibration estimator can be written as

$$\hat{Y}_{\text{MCAL}} = N\{\bar{y}_{s;d} + (\bar{\hat{y}}_U - \bar{\hat{y}}_{s;d})\tilde{B}_{s;d}\} \qquad (4.4)$$

where $\bar{y}_{s;d} = \sum_s d_k y_k / \sum_s d_k$; $\bar{\hat{y}}_{s;d} = \sum_s d_k \hat{y}_k / \sum_s d_k$, and

$$\tilde{B}_{s;d} = \left(\sum\nolimits_s d_k(\hat{y}_k - \bar{\hat{y}}_{s;d})y_k\right) \Big/ \sum\nolimits_s d_k(\hat{y}_k - \bar{\hat{y}}_{s;d})^2.$$

The regression implied by $\tilde{B}_{s;d}$ is one of observed y-values on predicted y-values. The idea of this regression would hardly occur to the modeler is his/her attempts to structure the relation between y_k and $\mathbf{x}_k$, but it proves effective in building the calibration estimator. Wu and Sitter (2001) present evidence that

$$(\hat{Y}_{\text{MCAL}} - Y)/N = \left(\sum\nolimits_s d_k \tilde{E}_k - \sum\nolimits_U \tilde{E}_k\right) \Big/ N + O_p(n^{-1})$$

with $\tilde{E}_k = y_k - \bar{y}_U - (\mu_k - \bar{\mu}_U)\tilde{B}_U$, where $\tilde{B}_U = (\sum_U (\mu_k - \bar{\mu}_U) y_k) / \sum_U (\mu_k - \bar{\mu}_U)^2$, and $\bar{\mu}_U = \sum_U \mu_k / N$. The coefficient $\tilde{B}_U$ may not be near one even in large samples. It expresses a regression of y_k on its assisting model mean $\mu_k = \mu(\mathbf{x}_k, \boldsymbol{\beta})$. That is, $\hat{Y}_{\text{MCAL}}$ can be viewed as a regression estimator that uses the model expectation μ_k as the auxiliary variable, leaving $\tilde{E}_k$ as the residuals that determine the asymptotic variance of $\hat{Y}_{\text{MCAL}}$.

How does this asymptotic variance compare with that of the non-linear GREG construction (3.1) for the same non-linear assisting model and the same $\hat{y}_k = \hat{\mu}_k$? Formula (3.1) implies a slope equal to unity in the regression between y_k and $\hat{y}_k = \hat{\mu}_k$; viewed in that light, $\hat{Y}_{\text{GREG}}$ is a difference estimator rather than a regression estimator and hence less sensitive to the pattern in the data. The non-linear GREG $\hat{Y}_{\text{GREG}}$ is in general less efficient than $\hat{Y}_{\text{MCAL}}$. (It is of course possible to modify $\hat{Y}_{\text{GREG}}$ to also account for the information contained in the known population size N.)

On the other hand, compared with the linear (model-free) calibration estimator $\hat{Y}_{\text{CAL}} = \sum_s w_k y_k$ with weights as in (3.3), the model-calibration estimator $\hat{Y}_{\text{MCAL}}$ given by (4.4) may have a considerable variance advantage but implies a loss of the practical advantages of a consistency with the known population total $\sum_U \mathbf{x}_k$ and a multi-purpose weight system applicable to all y-variables. The y-values in (4.4) are linearly weighted, but the weights now also depend on the y-values. It is thus debatable if $\hat{Y}_{\text{MCAL}}$ is a bona fide calibration estimator.

In an empirical study, Wu and Sitter (2001) compare $\hat{Y}_{\text{MCAL}} = \sum_s w_k y_k$, calibrated according to (4.3), with the non-linear GREG, $\hat{Y}_{\text{GREG}} = \sum_U \hat{y}_k + \sum_s d_k (y_k - \hat{y}_k)$ given by (3.1), for the same non-linear assisting model and same $\hat{y}_k = \hat{\mu}_k$. The study confirms that $\hat{Y}_{\text{MCAL}}$ has a variance advantage over the non-linear $\hat{Y}_{\text{GREG}}$. They created a finite population U of size $N = 2{,}000$ with values (y_k, x_k), $k = 1, \ldots, 2{,}000$, such that $\log(y_k) = 1 + x_k + \varepsilon_k$; the 2,000 values x_k are realizations of the Gamma(1,1) random variable, and ε_k is a normally distributed error. The auxiliary information consists of the population size N and the known values x_k for $k = 1, \ldots, 2{,}000$. Repeated simple random samples of size n = 100 were taken; the assisting model for both estimators was the log-linear $E_\xi(y_k | x_k) = \mu_k$ with $\log(\mu_k) = \alpha + \beta x_k$. This model was fit for each sample, using pseudo-maximum quasi-likelihood estimation. The fitted values $\hat{y}_k = \exp(\hat{\alpha} + \hat{\beta} x_k)$ were used to form both $\hat{Y}_{\text{MCAL}}$ and $\hat{Y}_{\text{GREG}}$. The simulation variance was markedly lower for $\hat{Y}_{\text{MCAL}}$. (The linear GREG (3.2), identical to the model-free calibration estimator, was also included in the Wu and Sitter study; not surprisingly, it is even less efficient than the non-linear GREG, under the strongly non-linear relationship imposed in their experiment.)

Montanari and Ranalli (2005) provide further evidence, for several artificially created populations, on the comparison between $\hat{Y}_{\text{MCAL}}$ and the non-linear $\hat{Y}_{\text{GREG}}$. Their assisting model, $y_k = \mu_k + \varepsilon_k$, is fitted via nonparametric regression (local polynomial smoothing), yielding predictions $\hat{y}_k = \hat{\mu}_k$ for $k \in U$. With this type of model fit, the predictions $\hat{y}_k = \hat{\mu}_k$ are highly accurate. Not surprisingly, the model-calibration estimator $\hat{Y}_{\text{MCAL}}$ achieves only marginal improvement over the non-linear $\hat{Y}_{\text{GREG}}$.

We can summarize the calibration approach as follows: The estimator of $Y = \sum_U y_k$ has the linearly weighted form $\hat{Y} = \sum_s w_k y_k$. In linear (model-free) calibration, the calibration equation reads $\sum_s w_k \mathbf{x}_k = \sum_U \mathbf{x}_k$; a known population auxiliary total $\sum_U \mathbf{x}_k$ is required, but complete auxiliary information (known $\mathbf{x}_k$ for all $k \in U$) is not required; the same weights can be applied to all y-variables (multi-purpose weighting); the estimator is identical to the linear GREG estimator (but derived by different reasoning). In model-calibration, the assisting model mean μ_k is non-linear in $\mathbf{x}_k$; complete auxiliary information is usually required; the calibration constraints include the equation $\sum_s w_k \hat{y}_k = \sum_U \hat{y}_k$; the weights w_k depend on the y_k-values, implying a loss of the multi-purpose property.

5. Computational aspects, extreme weights and outliers

The computation of calibrated weights raises important practical issues, discussed in a number of papers. All computation must proceed smoothly and routinely in the large scale statistics production of a national statistical agency. Undesirable (or unduly variable) weights should be avoided. Many practitioners support the reasonable requirement that all weights be positive (even greater than unity) and that very large weights should be avoided.

A few of the weights computed according to (3.2) can turn out to be quite large or negative. Huang and Fuller (1978) and Park and Fuller (2005) proposed methods to avoid undesirable weights.

In the distance minimization method, the distance function can be formulated so that negative weights are excluded, while still satisfying the given calibration equations. The software CALMAR (Deville, Särndal and Sautory 1993) allows several distance functions of this kind. An expended version, CALMAR2, is described in LeGuennec and Sautory (2002). Other statistical agencies have developed their own software for weight computation. Among those are GES (Statistics Canada), CLAN97 (Statistics Sweden), Bascula 4.0 (Central Bureau of Statistics, The Netherlands), g-CALIB-S (Statistics Belgium). These strive, in different ways, to resolve the computational issues arising. The user needs to consult the users' guide in each particular case to see exactly how the computational issues, including an avoidance of undesirable weights, are handled.

GES uses mathematical programming to minimize the chi-square distance, subject to the calibration constraints as well as to individual bounds on the weights, so that they will satisfy $A_k \le w_k \le B_k$ for specified A_k, B_k. Bascula 4.0 is described in Nieuwenbroek and Boonstra (2002). The software g-CALIB-S, described in Vanderhoeft, Waeytens and Museux (2001), Vanderhoeft (2001), uses generalized inverse (the Moore–Penrose) for the weight computation; consequently one need not be concerned about a possible redundancy in the auxiliary information.

In Bankier, Houle and Luc (1997) the objective is two-fold: to keep the computed weights within desirable bounds, and to drop some x-variables to remove near-linear dependencies. Isaki, Tsay and Fuller (2004) consider quadratic programming to obtain both household weights and person weights that lie within specified bounds.

An intervention with the weights (so as to get rid of undesirable weight values) raises the question how far one can deviate from the design weights d_k without compromising the desirable feature of nearly design unbiased estimation. An idea that has been tried is to modify the set of constraints so that tolerances are respected for the difference between the estimator for the auxiliary variables and the corresponding known population totals. Hence, Chambers (1996) minimizes a "cost-ridged loss function".

Outlying values in the auxiliary variables may be a cause of extreme weights. Calibration in the presence of outliers is discussed in Duchesne (1999). His technique of "robust calibration" may introduce a certain bias in the estimates; it may, however, be more than offset by a reduction in variance.

When the set of constraints is extended to make the weights restricted to specified intervals, a solution to the optimization problem is not guaranteed. The existence of a solution is considered in Théberge (2000), who also proposes methods for dealing with outliers.

6. Calibration estimation for more complex parameters

The calibration approach adapts itself to the estimation of more complex parameters than a population total. Examples are reviewed in this section. Single phase sampling and full response continue to be assumed; the notation remains as in Section 2. One example is the estimation of population quantiles (Section 6.1), another is the estimation of functions of totals (Section 6.2). Other examples in this category, not reviewed here, are Théberge (1999), for the estimation of bilinear parameters, and Tracy, Singh and Arnab (2003), for calibration with respect to second order moments.

6.1 Calibration for estimation of quantiles

The median and other quantiles of the finite population are important descriptive measures, especially in economic surveys. To estimate quantiles, the finite population distribution function must first be estimated. Before calibration became popular, several papers considered the estimation of quantiles, with or without the use of auxiliary information. More recent articles have turned to the calibration approach for the same purpose, including Kovaĉević (1997), Wu and Sitter (2001), Ren (2002), Tillé (2002), Harms (2003), Harms and Duchesne (2006) and Rueda *et al.* (2007). As these papers illustrate, there is more than one way to implement the calibration approach. The non-smooth character of the finite population distribution function causes certain complexities; these are resolved by different authors in different ways.

Let $\Delta(\cdot)$ denote the Heaviside function, defined for all real z so that $\Delta(z)=1$ if $z \geq 0$ and $\Delta(z)=0$ if $z<0.$ The unknown distribution function of the study variable y is

$$F_y(t) = \frac{1}{N}\sum\nolimits_U \Delta(t-y_k). \qquad (6.1)$$

The α-quantile of the finite population is defined as $Q_{y\alpha} = \inf\{t \mid F_y(t) \geq \alpha\}$. The auxiliary variable x_j, taking values x_{jk}, has the distribution function $F_{x_j}(t) = (1/N)\sum_U \Delta(t-x_{jk})$ with α-quantile denoted $Q_{x_j\alpha}$, $j=1,2,\ldots,J.$ A natural estimator of $F_y(t)$ based on the design weights $d_k = 1/\pi_k$ is

$$\hat{F}_y(t) = \frac{1}{\sum_s d_k}\sum\nolimits_s d_k \Delta(t-y_k).$$

A calibration estimator $F_y(t)$ of takes the form

$$\hat{F}_{y\mathrm{CAL}}(t) = \frac{1}{\sum_s w_k}\sum\nolimits_s w_k \Delta(t-y_k) \qquad (6.2)$$

where the weights w_k are suitably calibrated to a specified auxiliary information; then from $\hat{F}_{y\mathrm{CAL}}(t)$ we obtain the α-quantile estimator as $\hat{Q}_{y\alpha} = \inf\{t \mid \hat{F}_{y\mathrm{CAL}}(t) \geq \alpha\}$. A formula analogous to (6.2) holds for $\hat{F}_{x_j\mathrm{CAL}}(t)$.

Without explicit reference to any model, Harms and Duchesne (2006) specify the information available for calibration as a known population size, N, and known population quantiles $Q_{x_j\alpha}$ for $j=1,2,\ldots,J.$ The complete auxiliary information, with values $\mathbf{x}_k = (x_{k1},\ldots,x_{kJ})'$ known for $k \in U$, is not required. (But in practice, the complete information would usually be necessary, because accurate quantiles of several x-variables are not likely to be importable from outside sources.) They determine the w_k to minimize the chi-square distance $\sum_s (w_k - d_k)^2 / 2d_k q_k$, for specified q_k, subject to the calibration equations

$$\sum\nolimits_s w_k = N; \hat{Q}_{x_j\mathrm{CAL},\alpha} = Q_{x_j\alpha},\ j=1,2,\ldots,J$$

for suitably defined estimates $\hat{Q}_{x_j\mathrm{CAL},\alpha}$. Now, if we were to specify $\hat{Q}_{x_j\mathrm{CAL},\alpha} = \inf\{t \mid \hat{F}_{x_j\mathrm{CAL}}(t) \geq \alpha\}$, then it is in general not possible to find an exact solution of the calibration problem as stated. Instead, Harms and Duchesne substitute smoothed estimators, called "interpolated distribution estimators", of the distribution functions $F_{x_j}(t), j = 1, 2, ..., J.$ They replace $\Delta(\cdot)$ by a slightly modified function. Weights w_k can now be obtained, as well as a corresponding estimated distribution function $\hat{F}_{y\text{CAL}}(t)$; finally, $Q_{y\alpha}$ is estimated as $\hat{Q}_{y\alpha} = \hat{F}_{y\text{CAL}}^{-1}(\alpha)$.

The resulting calibrated weights w_k allow us to retrieve the known population quantiles of the auxiliary variables. This is reassuring; one would expect such weights to produce reasonable estimators for the quantiles of the study variable y. Moreover, in the case of a single scalar auxiliary variable x, the resulting calibration estimator delivers exact population quantiles for y when the relationship between y and x is exactly linear, that is, when $y_k = \beta x_k$ for all $k \in U$. An idea involving smoothed distribution functions is also used in Tillé (2002).

The computationally simpler method of Rueda *et al.* (2007) is an application of model-calibration, in that they calibrate with respect to a population total of *predicted y*-values. Complete auxiliary information is required. Using the known $\mathbf{x}_k$, compute first the linear predictions $\hat{y}_k = \hat{\boldsymbol{\beta}}'\mathbf{x}_k$ for $k \in U$, with $\hat{\boldsymbol{\beta}} = (\sum_s d_k q_k \mathbf{x}_k \mathbf{x}_k')^{-1} (\sum_s d_k q_k \mathbf{x}_k y_k)$, where $d_k = 1/\pi_k$ and the q_k are specified scale factors. The weights w_k are obtained by minimizing the chi-square distance subject to calibration equations stated in terms of the predictions, so as to have consistency at J arbitrarily chosen points $t_j, j = 1, ..., J$:

$$\frac{1}{N}\sum\nolimits_s w_k \Delta(t_j - \hat{y}_k) = F_{\hat{y}}(t_j), j = 1, ..., J$$

where $F_{\hat{y}}(t_j)$ is the finite population distribution function of the predictions $\hat{y}_k$, evaluated at t_j. It is suggested that a fairly small number of arbitrarily selected points t_j may suffice, say less than 10. Once the w_k are determined, the α-quantile estimate is obtained from $\hat{F}_{y\text{CAL}}(t) = (1/N)\sum_s w_k \Delta(t - y_k)$.

Quantile estimation provides a good illustration that the calibration approach can be carried out in more than one way when somewhat more complex parameters are being estimated. Both methods mentioned give nearly design unbiased estimation. The Harms and Duchesne (2006) weights are multi-purpose, independent of the y-variable; by contrast, the method of Rueda *et al.* (2007) requires a new set of weights for every new y-variable. Empirical evidence, by simulation, suggests that both methods compare favourably with the earlier quantile estimation methods, not based explicitly on calibration thinking (but on the same auxiliary information).

An extension of the calibration approach to the estimation of other complex parameters, such as the Gini coefficient, is sketched in Harms and Duchesne (2006).

6.2 Calibration for other complex parameters

Plikusas (2006), and Krapavickaitė and Plikusas (2005) examine calibration estimation of certain functions of population totals. (Their term "non-linear calibration" signifies "non-linear function of totals"; I do not use it here.) A simple example is the estimation of a ratio of two totals, $R = \sum_U y_{1k} / \sum_U y_{2k}$, where y_{1k} and y_{2k} are the values for element k of the variables y_1 and y_2, respectively. (The distribution function (6.1) is in effect also of ratio type, with $y_{2k} = 1$, and $N = \sum_U 1$ as the denominator total.) These authors examine the calibration estimator $\hat{R}_{\text{CAL}} = \sum_s w_k y_{1k} / \sum_s w_k y_{2k}$. Its weights w_k, common to the numerator and the denominator, are determined by calibration to auxiliary information stated as follows: There is one auxiliary variable, x_{1k}, for y_{1k}, and another, x_{2k}, for y_{2k}; the ratio of totals $R_0 = \sum_U x_{1k} / \sum_U x_{2k}$ is a known value, by a complete enumeration at a previous occasion or from some other accurate source. The proposed calibration equation is $\sum_s w_k e_k = 0$, where $e_k = x_{1k} - R_0 x_{2k}$. Because $\sum_U e_k = 0$, the weights, by minimum chi-square distance, are

$$w_k = d_k \left\{1 - \left(\sum\nolimits_s d_k e_k\right)\left(\sum\nolimits_s d_k e_k^2\right)^{-1} e_k\right\}.$$

These weights correctly retrieve the known ratio value R_0; setting $y_{1k} = x_{1k}$ and $y_{2k} = x_{2k}$ in $\hat{R}_{\text{CAL}}$, we have

$$\frac{\sum_s w_k x_{1k}}{\sum_s w_k x_{2k}} - R_0 = \frac{\sum_s w_k e_k}{\sum_s w_k x_{2k}} = 0.$$

The empirical evidence in Plikusas (2006), and Krapavickaitė and Plikusas (2005) suggests that their calibration estimator compares favourably (lower variance, while maintaining near design unbiasedness) with other estimators, derived through other arguments than calibration, while relying on the same auxiliary information.

7. Calibration contrasted with other approaches

As many have noted, users view calibration as a simple and convincing way to incorporating auxiliary information, for simple parameters (Section 4), as for more complex parameters such as quantiles, ratios and others (Section 6). Simplicity and practicality are undeniable advantages, but aside from that, is calibration also "theoretically superior"? Are there instances where calibration can be shown to give more accurate and/or more satisfactory answers on questions of importance, when contrasted with other design-based approaches?

Section 4.5 gave one indication that calibration thinking may have an advantage over GREG thinking, in that model-calibration may give more precise estimates than the non-linear GREG, for the same assisting model. The following Section 7.1 gives another example where calibration reasoning and GREG reasoning give diverging answers, with an advantage for the calibration method.

7.1 An example in domain estimation

The example in this section, from Estevao and Särndal (2004), shows, for a simple practical situation, a conflict between the results of GREG thinking and calibration thinking. The context is the estimation of the y-total for a sub-population (a domain).

A probability sample s is drawn from $U = \{1, 2, ..., k, ..., N\}$; the known design weights are $d_k = 1/\pi_k$. Let U_a be a domain; $U_a \subset U$. The domain indicator is δ_{ak} with value $\delta_{ak} = 1$ if $k \in U_a$ and $\delta_{ak} = 0$ if not. The target of estimation is the domain total $Y_a = \sum_U y_{ak}$, where $y_{ak} = \delta_{ak} y_k$, and y_k is observed for $k \in s$. The Horvitz-Thompson estimator $\hat{Y}_{\text{HT}} = \sum_s d_k y_{ak}$, although design unbiased, has low precision, especially if the domain is small; the use of auxiliary information will bring improvement. An auxiliary vector value $\mathbf{x}_k$ is specified for every $k \in U$.

As is frequently the case in practice, the elements belonging to a domain of interest are not identified in the sampling frame. (If they are, some very powerful information is available from the start, but frequently real world conditions are not that favourable.) But suppose elements in a larger group U_C are identifiable; $U_a \subset U_C \subset U$. For example, suppose y is "income" and U_C a professional group specified for the persons listed in the frame, while U_a is a professional sub-group not identified in the frame. We can identify the sample subsets $s_C = s \cap U_C$ and $s_a = s \cap U_a$, and we can benefit from knowing the total $\sum_U \mathbf{x}_{Ck}$, estimable without bias by $\sum_s d_k \mathbf{x}_{Ck}$, where $\mathbf{x}_{Ck} = \delta_{Ck} \mathbf{x}_k$, and δ_{Ck} is the information group indicator: $\delta_{Ck} = 1$ if $k \in U_C$ and $\delta_{Ck} = 0$ if not. The domain auxiliary total $\sum_U \mathbf{x}_{ak}$ is unavailable, because U_a is not identified. Calibration to satisfy $\sum_s w_k \mathbf{x}_{Ck} = \sum_U \mathbf{x}_{Ck}$ gives the nearly design unbiased estimator $\hat{Y}_{a\text{CAL}} = \sum_U w_k y_{ak}$, where $w_k = d_k(1 + \boldsymbol{\lambda}'\mathbf{z}_k)$, with $\boldsymbol{\lambda}' = (\sum_U \mathbf{x}_{Ck} - \sum_s d_k \mathbf{x}_{Ck})' (\sum_s d_k \mathbf{z}_k \mathbf{x}'_{Ck})^{-1}$. The asymptotically optimal instrument for the given vector $\mathbf{x}_k$ is (see Section 4.3) $\mathbf{z}_k = \mathbf{z}_{0Ck} = d_k^{-1} \sum_{\ell \in s} (d_k d_\ell - d_{k\ell}) \mathbf{x}_{C\ell}$.

By contrast, regression thinking for the same auxiliary information leads to $\hat{Y}_{a\text{GREG}} = \sum_s d_k y_{ak} + (\sum_U \mathbf{x}_{Ck} - \sum_s d_k \mathbf{x}_{Ck})' \mathbf{B}_{\tilde{s};d}$, also nearly design unbiased, where the regression coefficient $\mathbf{B}_{\tilde{s};d} = (\sum_{\tilde{s}} d_k \mathbf{x}_k \mathbf{x}'_k)^{-1} \sum_{\tilde{s}} d_k \mathbf{x}_k y_k$ is the result of a weighted least squares fit at a suitable level, using all (when $\tilde{s} = s$) or part (when $\tilde{s} \subset s$) of the data points $(y_k, \mathbf{x}_k)$ available for $k \in s$.

For example, the modeller may opt for a regression fit "extending beyond the domain" (so that $\tilde{s} \supset s_a = s \cap U_a$), in an attempt to borrow strength for $\hat{Y}_{a\text{GREG}}$ by letting it depend also on y-data from outside the domain. By contrast, $\hat{Y}_{q\text{CAL}}$ relies exclusively on y-data in the domain, and this is in effect better. Estevao and Särndal (2004) show that $\hat{Y}_{a\text{CAL}}$ with $\mathbf{z}_k = \mathbf{z}_{0Ck}$ has smaller (asymptotic) variance than $\hat{Y}_{a\text{GREG}}$, no matter how $\tilde{s}$ is chosen. Bringing in y-data from the outside does not help; calibration thinking and regression thinking do not agree.

8. Calibration estimation in the presence of composite information

As the preceding sections have shown, many papers choose to study estimation for direct, single phase sampling of elements, without any nonresponse. The information available for calibration is simple; the k:th element of the finite population $U = \{1, 2, ..., k, ..., N\}$ has an associated auxiliary vector value $\mathbf{x}_k$.

However, in an important category of situations, the auxiliary information has *composite structure*. The complexity of the information increases with that of the sampling design. In designs with two or more phases, or in two or more stages, the information is typically composed of more than one component, reflecting the features of the design. The information is stated in terms of more than one auxiliary vector. For example, in two-stage sampling, some information may be available about the first stage sampling units (the clusters), other information about the second stage units (the elements).

Consequently, estimation by calibration (or by any alternative method) must take the composite structure of the information systematically into account. The total information has several pieces; the calibration can be done in more than one way. All relevant pieces should be taken into account, for best possible accuracy in the estimates. To accomplish this in a general or "optimal" way is not a trivial task. Calibration reasoning offers one way.

Regression reasoning, with a duly formulated assisting model, is an alternative way, but it will strike some users as more roundabout. Hence, surveys that allow composite auxiliary information bring further perspectives on the contrast between calibration thinking and GREG thinking.

Two-phase sampling and two-stage sampling are discussed in this section. Another example of composite information occurs for nonresponse bias adjustment, as discussed in Section 9.

Another aspect of composite information occurs when the objective is to combine information from several surveys. This, too, can be a way to add strength and improve accuracy of the estimates. It is a motivating factor (in addition to the user oriented motive to achieve consistency among surveys) in the previously mentioned repeated weighting methodology of the Dutch statistical agency. Combined auxiliary information for GREG estimation is considered in Merkouris (2004).

8.1 Composite information for two-phase sampling designs

Double sampling refers to designs involving two probability samples, s_1 and s_2, from the same population $U = \{1, ..., k, ..., N\}$. Auxiliary data may be recorded for both U and s_1, the study variable values y_k are recorded only for $k \in s_2$ with an objective to estimate $Y = \sum_U y_k$. Hidiroglou (2001) distinguishes several kinds of double sampling: In the *nested case* (traditional two phase sampling), the first phase sample s_1 is drawn from U, the second phase sample s_2 is a sub-sample from s_1, so that $U \supset s_1 \supset s_2$. Two *non-nested cases* can be distinguished: In the first of these, s_1 is drawn from the frame U_1; s_2 from the frame U_2, where U_1 and U_2 cover the same population U; the sampling units may be defined differently for the two frames. In the second non-nested case, s_1 and s_2 are drawn independently from U.

To illustrate how composite information intervenes in the estimation, consider the nested case. The design weights are $d_{1k} = 1/\pi_{1k}$ (s_1 sampled from U); $d_{2k} = 1/\pi_{2k} (\pi_{2k} = \pi_{k|s_1}$ in sub-sampling s_2 from s_1). The combined design weight is $d_k = d_{1k} d_{2k}$. The basic unbiased estimator $\hat{Y} = \sum_{s_2} d_k y_k$ can be improved by a use of auxiliary information, specified here at two levels:

Population level: The vector value $\mathbf{x}_{1k}$ is known (given in the frame) for every $k \in U$, thus known for every $k \in s_1$ and for every $k \in s_2$; $\sum_U \mathbf{x}_{1k}$ is a known population vector total;

First sample level: The vector value $\mathbf{x}_{2k}$ is known (observed) for every $k \in s_1$, and thereby known for every $k \in s_2$; the unknown total $\sum_U \mathbf{x}_{2k}$ is estimated without bias by $\sum_{s_1} d_{1k} \mathbf{x}_{2k}$.

How do we best take this composite information into account? In an adaptation of GREG thinking, Särndal and Swensson (1987) formulated two linear assisting models, the first one stated in terms of the $\mathbf{x}_{1k}$-vector, the other one also brings in the $\mathbf{x}_{2k}$-vector. The two models are fitted; the resulting predictions, of two kinds, are used to create an appropriate GREG estimator $\hat{Y}_{\text{GREG}}$ of $Y = \sum_U y_k$.

Dupont (1995) makes the important point that the given composite information invites “two different natural approaches”: Besides the GREG approach, there is a calibration approach that will deliver final weights w_k for a calibration estimator $\hat{Y}_{\text{CAL}} = \sum_{s_2} w_k y_k$. It is of interest to compare the results of the two approaches. Both of them allow more than one option: In the GREG approach, there are alterative ways of formulating the linear assisting models with their respective variance structures. In the calibration approach, alternative formulations of the calibration equations are possible.

For example, a *two-step calibration* option is as follows: First find intermediate weights w_{1k} to satisfy $\sum_{s_1} w_{1k} \mathbf{x}_{1k} = \sum_U \mathbf{x}_{1k}$; then use the weights w_{1k} in the second step to compute the final weights w_k to satisfy

$$\sum\nolimits_{s_2} w_k \mathbf{x}_k = \sum\nolimits_{s_1} w_{1k} \mathbf{x}_k = \begin{pmatrix} \sum_U \mathbf{x}_{1k} \\ \sum_{s_1} w_{1k} \mathbf{x}_{2k} \end{pmatrix}$$

where $\mathbf{x}_k$ is the combined auxiliary vector

$$\mathbf{x}_k = \begin{pmatrix} \mathbf{x}_{1k} \\ \mathbf{x}_{2k} \end{pmatrix}.$$

Alternatively, in a *single step* option, we determine the w_k directly to satisfy

$$\sum\nolimits_{s_2} w_k \mathbf{x}_k = \begin{pmatrix} \sum_U \mathbf{x}_{1k} \\ \sum_{s_1} d_{1k} \mathbf{x}_{2k} \end{pmatrix}.$$

The final weights w_k are in general not identical in the two options. Suppose that $\sum_U \mathbf{x}_{1k}$ is an imported $\mathbf{x}_1$-total. At closer look, the two-step option requires more extensive information, because individually known values $\mathbf{x}_{1k}$ are required for $k \in s_1$, whereas it is sufficient in the single step option that they be available for $k \in s_2$. Some variance advantage may thus be expected from the two-step option, since $\sum_{s_1} w_{1k} \mathbf{x}_{2k}$ is often more accurate (as an estimator of $\sum_U \mathbf{x}_{2k}$) than $\sum_{s_1} d_{1k} \mathbf{x}_{2k}$ in the single step procedure. Nevertheless, this anticipation is not always confirmed; the single step method can be better, as when $\mathbf{x}_1$ and $\mathbf{x}_2$ are weakly correlated.

Dupont (1995) and Hidiroglou and Särndal (1998) examine links that exist, not surprisingly, between the two approaches. A GREG estimator, derived from assisting models with specific variance structures, may be identical to calibration estimator, if the weights of the latter are calibrated in a certain way. In other cases, differences may be small.

The efficiency of different options depends in rather subtle ways on the pattern of correlation among y_k, $\mathbf{x}_{1k}$ and $\mathbf{x}_{2k}$. For example, to what extent do $\mathbf{x}_1$ and $\mathbf{x}_2$ complement each other, to what extent are they substitutes for one another? In the GREG approach, it is difficult or even futile to pinpoint a variance structure that truly captures a “reality” behind the data. The calibration approach is more direct. Some of its possibilities are explored in Estevao and Särndal (2002, 2006).

8.2 Composite information in two-stage sampling designs

The traditional two-stage sampling set-up (clusters sampled at stage one, elements sub-sampled within selected clusters in stage two) has in common with two-phase sampling that the total information may have more than one component. There may exist (a) information at the cluster level (about the clusters); (b) information at the element level for all clusters; (c) information at the element level for the selected clusters only. Here again, authors are of two different orientations: some exploit the information via calibration thinking, others follow the GREG thinking route.

Estevao and Särndal (2006) develop calibration estimation for the traditional two-stage set-up, with composite information specified as follows: (i) for the cluster population U_I, there is a known total $\sum_{U_\mathrm{I}} \mathbf{x}_{(c)i}$, where $\mathbf{x}_{(c)i}$ is a vector value associated with the cluster U_i, for $i \in U_\mathrm{I}$; (ii) for the population of elements $U = \bigcup_{i\in U_\mathrm{I}} U_i$, there is a known total $\sum_U \mathbf{x}_k$, where the vector value $\mathbf{x}_k$ is associated with the element $k \in U$. Suppose both cluster statistics and element statistics are to be produced in the survey: Both the cluster population total $Y_\mathrm{I} = \sum_{U_\mathrm{I}} y_{(c)i}$ and the element population total $Y = \sum_U y_k$ are to be estimated.

If no relation is imposed between cluster weights $w_{\mathrm{I}i}$ and element weights w_k, the former are calibrated to satisfy $\sum_{s_\mathrm{I}} w_{\mathrm{I}i}\mathbf{x}_{(c)i} = \sum_{U_\mathrm{I}} \mathbf{x}_{(c)i}$, the latter to satisfy $\sum_s w_k \mathbf{x}_k = \sum_U \mathbf{x}_k$. (Here, s_I is the sample of clusters from U_I; s_i is the sample of elements from the cluster U_i; $s = \bigcup_{i\in s_\mathrm{I}} s_i$ is the entire sample of elements.) Then $\hat{Y}_{\mathrm{ICAL}} = \sum_{s_\mathrm{I}} w_{\mathrm{I}i} y_{(c)i}$ estimates the cluster population total Y_I, and $\hat{Y}_{\mathrm{CAL}} = \sum_s w_k y_k$ estimates the element population total Y.

Integrated weighting is often used in practice: A convenient relationship is imposed between the cluster weight $w_{\mathrm{I}i}$ and the weights w_k for the elements within the selected cluster. Two forms of integrated weighting are discussed in Estevao and Särndal (2006).

One of these is to impose $w_k = d_{k|i} w_{\mathrm{I}i}$, where $d_{k|i}$ is the inverse of the probability of selecting element k within cluster i. (For example, in single stage cluster sampling, when all elements k in a sampled cluster are selected, then $d_{k|i} = 1$. Consequently $w_k = w_{\mathrm{I}i}$ is imposed, and all elements in the cluster receive the same weight for computing element statistics, and that same weight is also used for computing cluster statistics.) The calibration equation $\sum_s w_k \mathbf{x}_k = \sum_U \mathbf{x}_k$ then reads $\sum_{s_\mathrm{I}} w_{\mathrm{I}i} \sum_{s_i} d_{k|i} \mathbf{x}_k = \sum_U \mathbf{x}_k$. The cluster weights $w_{\mathrm{I}i}$ are now derived by minimizing $\sum_{s_\mathrm{I}} (w_{\mathrm{I}i} - d_{\mathrm{I}i})^2 / d_{\mathrm{I}i}$ subject to the calibration equation that takes both kinds of information into account:

$$\begin{pmatrix} \sum_{s_\mathrm{I}} w_{\mathrm{I}i} \mathbf{x}_{(c)i} \\ \sum_{s_\mathrm{I}} w_{\mathrm{I}i} \sum_{s_i} d_{k|i} \mathbf{x}_k \end{pmatrix} = \begin{pmatrix} \sum_{U_\mathrm{I}} \mathbf{x}_{(c)i} \\ \sum_U \mathbf{x}_k \end{pmatrix}. \qquad (8.1)$$

Once the $w_{\mathrm{I}i}$ are determined, the element weights $w_k = d_{k|i} w_{\mathrm{I}i}$ follow.

Another reasonable integrated weighting is to impose $\sum_{s_i} w_k = N_i w_{\mathrm{I}i}$. For example, for single stage cluster sampling it implies that the cluster weight $w_{\mathrm{I}i}$ is the average of the element weights w_k in that cluster.

Two-stage sampling is also the topic in Kim, Breidt and Opsomer (2005). They assume auxiliary information for clusters, via a single quantitative cluster variable $x_{(c)i}$, but none for elements. They develop and examine a GREG type estimator of the element total $Y = \sum_U y_k$, $\hat{Y} = \sum_{i\in U_\mathrm{I}} \hat{\mu}_i + \sum_{i\in s_\mathrm{I}} d_i(\hat{t}_i - \hat{\mu}_i)$, where $\hat{t}_i$ is design unbiased for the cluster total $t_i = \sum_{U_i} y_k$, and $\hat{\mu}_i$ is obtained by local polynomial regression fit. The estimator can be expressed on the linearly weighted form, with weights that turn out to be calibrated to the population totals of powers of the cluster variable $x_{(c)i}$.

8.3 Household weighting and person weighting

Some important social surveys set the objective to produce both household estimates and person estimates; some study variables are household (cluster) variables, others are person (element) variables. Consequently, a number of papers have addressed the situation with single stage cluster sampling ($d_{k|i} = 1$) and the integrated weighting that gives all members of a selected household equal weight, a weight also used for producing household statistics. A general solution for this weighting problem, when both household information and person information are specified, is to obtain the household weights $w_{\mathrm{I}i}$ calibrated as in equation (8.1) with $d_{k|i} = 1$, then take $w_k = w_{\mathrm{I}i}$.

Several articles focus on auxiliary vector values $\mathbf{x}_k$ attributed to persons. Alexander (1987) derives weights by minimizing chi-square distance, whereas Lemaître and Dufour (1987) and Niewenbrook (1993) derive the integrated weights via a GREG estimator. The Lemaître and Dufour technique proceeds by an indirect construction of an "equal shares auxiliary vector value" for all persons in a selected household; their result is derivable from the direct procedure in Section 8.2.

The household-weighting/person-weighting question is revisited in more recent papers. Some authors display calibration thinking, others GREG thinking. Isaki, Tsay and Fuller (2004) formulate the problem as one of calibrated weighting; their weights respect both household controls and person controls; no explicit assisting models are formulated. By contrast, Steel and Clark (2007) proceed by the GREG approach, with linear assisting model statements and accompanying variance structures.

9. Calibration for nonresponse adjustment

9.1 Traditional adjustment for nonresponse

The context of many good theory articles is the simple one of Section 2, which includes total absence of nonresponse. It is good theory for conditions that seldom or never occur. (As an author of papers in that stream, I am not without guilt.) Practically all surveys encounter nonresponse; although undesirable, it is a natural feature, and theory should incorporate it, from the outset, via a perspective of selection in two phases.

In many surveys, nonresponse rates are extremely high today, compared with what they were 40 years ago, that is, so low that one could essentially ignore the problem. Today, survey sampling theory needs more and more to address the damaging consequences of nonresponse. In particular, one pressing objective is to examine the bias and to try to reduce it as far as possible.

A probability sample s is drawn from $U = \{1, 2, ..., k, ..., N\}$; the known design weight of element k is $d_k = 1/\pi_k$. Nonresponse occurs, leaving a response set r, a subset of s; the study variable value y_k is observed for $k \in r$ only. The unknown response probability of element k is $\Pr(k \in r|s) = \theta_k$. The unbiased estimator $\hat{Y} = \sum_r d_k \phi_k y_k$ is ruled out because $\phi_k = 1/\theta_k$ is unknown. To keep the idea of a linearly weighted sum, how do we then construct the weights? Unit nonresponse adjustment by weighting, based on "nonresponse modeling", has a long history. Calibration offers a newer perspective.

In what we may call "the traditional procedure", the probability design weights $d_k = 1/\pi_k$ are first adjusted for nonresponse and possibly for other imperfections such as outliers. The information used for this step is often a grouping of the sampled elements. Finally, if reliable population totals are accessible, the adjusted design weights are subjected to a calibration with respect to those totals.

The methodology of the Labour Force Survey of Canada, described in Statistics Canada (1998), exemplifies this widespread practice. A (modified) design weight is first computed for a given household, as the product of three factors. The product of the design weight and a nonresponse adjustment factor is called the sub-weight. The sub-weights are subjected in the final step to a calibration with respect to postcensal, highly accurate estimates of population by age group, sex and sub-provincial regions. The final weights meet the desirable objective of consistency, in regions within a province, with the postcensal estimates. The nonresponse bias remaining in the resulting estimates is unknown but believed to be modest.

The traditional procedure is embodied in the estimator type $\hat{Y} = \sum_r d_k (1/\hat{\theta}_k) y_k$, where θ_k has been estimated by $\hat{\theta}_k$ in a preliminary step, using response (propensity) modeling. What theory demands of the statistician is not an easy task, namely, to formulate "the true response model", capable of providing accurate, non-biasing values $\hat{\theta}_k$. But the factors $1/\hat{\theta}_k$ are applied in many surveys in an uncritical and mechanical fashion, for example, by straight expansion within the strata already used for sample selection.

The traditional procedure is apparent for example in Ekholm and Laaksonen (1991) and in Rizzo, Kalton and Brick (1996).

Practitioners often act as if the resulting $\hat{Y} = \sum_r d_k (1/\hat{\theta}_k) y_k$ (following a more or less probing response modeling trying to get the $\hat{\theta}_k$) is essentially unbiased, something which it is not (unless the ideal model happens to be specified); one acts (for purposes of variance estimation, for example) as if $\pi_k \hat{\theta}_k$ is the true selection probability of element k in a single step of selection, something which it is definitely not. This practice, with roots in the idyllic past, becomes more and more vulnerable as nonresponse rates continue their surreptitious climb.

An unavoidable bias results from the replacement of θ_k by $\hat{\theta}_k$. Decades ago, when the typical nonresponse was but a few per cent, it was defendable to ignore this bias, but with today's galloping nonresponse rates, the practice becomes untenable. By first principles, unbiased estimation is the goal, not an estimation where the squared bias is a dominating (and unknown) contributor to the Mean Squared Error. We must resolve to limit the bias as much as possible. Calibration reasoning can help in constructing an auxiliary vector that meets this objective.

9.2 Calibration for nonresponse bias adjustment

More or less contrasting with the traditional procedure are a number of recent papers that emphasize calibration reasoning to achieve the nonresponse adjustment. Recent references are Deville (1998, 2002), Ardilly (2006), chapter 3, Skinner (1998), Folsom and Singh (2000), Fuller (2002), Lundström and Särndal (1999), Särndal and Lundström (2005) and Kott (2006).

Calibration reasoning starts by assessing the total available auxiliary information: information at the sample level (auxiliary variable values observed for respondents and for nonrespondents), information at the population level (known population auxiliary totals). The objective is to make the best of the two sources combined, so as to reduce both bias and variance. The design weights are modified, in one or two calibration steps, to make them reflect (i) the outcome of the response phase, (ii) the individual characteristics of the respondents, and (iii) the specified auxiliary information. The information can be summarized as follows:

Population level: The vector value $\mathbf{x}_k^*$ is known (specified in the frame) for every $k \in U,$ thus known for every $k \in s$ and for every $k \in r$; $\sum_U \mathbf{x}_k^*$ is a known population total.

Sample level: The vector value $\mathbf{x}_k^\circ$ is known (observed) for every $k \in s,$ and thereby known for every $k \in r$; the unknown total $\sum_U \mathbf{x}_k^\circ$ is estimated without bias by $\sum_s d_k \mathbf{x}_k^\circ$.

Calibration on this composite information can be done in two steps (intermediate weights computed first, then used in the second step to produce final weights) or directly in one single step. Modest differences only are expected in bias and variance of the estimates. In the single step option, the combined auxiliary vector and the corresponding information are

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

Using an extension of the instrument vector method in Section 4.3, we seek calibrated weights $w_k = d_k v_k,$ where $v_k = F(\boldsymbol{\lambda}'\mathbf{z}_k)$ is the nonresponse adjustment factor, with a vector $\boldsymbol{\lambda}$ determined through the calibration equation $\sum_r w_k \mathbf{x}_k = \mathbf{X}$; the resulting calibration estimator is $\hat{Y}_{\text{CAL}} = \sum_r w_k y_k$. It is enough to specify the instrument vector value $\mathbf{z}_k$ for respondents only; $\mathbf{z}_k$ is allowed to differ from $\mathbf{x}_k$. The function $F(\cdot)$ has the same role as in Sections 4.2 and 4.3. Here, $F(\boldsymbol{\lambda}'\mathbf{z}_k)$ implicitly estimates the inverse response probability, $\phi_k = 1/\theta_k$ as Deville (2002), Dupont (1995), Kott (2006) have noted. In the linear case, $F(u) = 1 + u,$ and $v_k = 1 + \boldsymbol{\lambda}'\mathbf{z}_k,$ with $\boldsymbol{\lambda}' = (\sum_U \mathbf{x}_k - \sum_r d_k \mathbf{x}_k)'(\sum_r d_k \mathbf{z}_k \mathbf{x}_k')^{-1}.$

The variables that make up the vector $\mathbf{x}_k^\circ,$ although observed for sampled elements only, can be crucially important for the reduction of nonresponse bias (although less important than the $\mathbf{x}_k^*$ for the reduction of variance). For example, Beaumont (2005b) discusses data collection process variables can be used in building the $\mathbf{x}_k^\circ$ vector component.

9.3 Building the auxiliary vector

In some surveys, there are many potential auxiliary variables, as pointed out for example by Rizzo, Kalton and Brick (1996), and Särndal and Lundström (2005). For example, for surveys on households and individuals in Scandinavia, a supply of potential auxiliary variables can be derived from a matching of existing high quality administrative registers. A decision then has to be made which of these variables should be selected for inclusion in the auxiliary vector $\mathbf{x}_k$ to make it as effective as possible, for bias reduction in particular. As Rizzo, Kalton and Brick (1996) point out, "the choice of auxiliary variables is … probably more important than the choice of the weighting methodology."

Let us examine the bias, when $\mathbf{z}_k = \mathbf{x}_k$. We need to compare alternative $\mathbf{x}_k$ -vectors in order to finally settle one likely to yield the smallest bias. (I assume $\mathbf{x}_k$ to be such that $\boldsymbol{\mu}'\mathbf{x}_k = 1$ for all k and some constant vector $\boldsymbol{\mu},$ as is the case for many $\mathbf{x}_k$ -vectors, including the examples 1 to 5 at the beginning of Section 2.) A close approximation to the bias of $\hat{Y}_{\text{CAL}}$ is obtained by Taylor linearization as $nearbias(\hat{Y}_{\text{CAL}}) = (\sum_U \mathbf{x}_k)'(\mathbf{B}_{U;\theta} - \mathbf{B}_U),$ which involves a difference between the weighted regression coefficient $\mathbf{B}_{U;\theta} = \left(\sum_U \theta_k \mathbf{x}_k \mathbf{x}_k'\right)^{-1} \sum_U \theta_k \mathbf{x}_k y_k$ and the unweighted one, $\mathbf{B}_U = (\sum_U \mathbf{x}_k \mathbf{x}_k')^{-1}(\sum_U \mathbf{x}_k y_k)$. Unless all θ_k are equal, the bias caused by the difference in the two regression vectors may be substantial, even though $\mathbf{x}_k$ is a seemingly "good auxiliary vector". This expression for nearbias is given in Särndal and Lundström (2005); related bias expressions, under different conditions, are found in Bethlehem (1988) and Fuller *et al.* (1994). We can write alternatively $nearbias(\hat{Y}_{\text{CAL}}) = \sum_U (\theta_k M_k - 1) y_k,$ where $M_k = (\sum_U \mathbf{x}_k)'$ $(\sum_U \theta_k \mathbf{x}_k \mathbf{x}_k')^{-1} \mathbf{x}_k$. In comparing possible alternatives $\mathbf{x}_k,$ a convenient benchmark is the "primitive auxiliary vector", $\mathbf{x}_k = 1$ for all $k \in U,$ which gives $\hat{Y}_{\text{CAL}} = N\,\bar{y}_r = N \sum_r y_k / n_r,$ where n_r is the number or respondents, with $nearbias(N\,\bar{y}_r) = N(\bar{y}_{U;\theta} - \bar{y}_U),$ where $\bar{y}_{U;\theta} = \sum_U \theta_k y_k / \sum_U \theta_k$ and $\bar{y}_U = \sum_U y_k / N.$ The ratio

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

measures how well a candidate vector $\mathbf{x}_k$ succeeds in controlling the bias, when compared with the primitive vector. We seek an $\mathbf{x}_k$ that will give a small bias. But $relbias(\hat{Y}_{\text{CAL}})$ is not a computable bias indicator; it depends on unobserved y_k and on unobservable θ_k. We need a computable indicator that approximates $relbias(\hat{Y}_{\text{CAL}})$ and depends on the $\mathbf{x}$ -vector but not on the y-variables, of which the survey may have many.

It is easy to see that $relbias(\hat{Y}_{\text{CAL}}) = 0$ if an ideal (probably non-existent) $\mathbf{x}$-vector could be constructed such that $\phi_k = 1/\theta_k = \boldsymbol{\lambda}'\mathbf{x}_k$ for all $k \in U$ and some constant vector $\boldsymbol{\lambda}$.

For an $\mathbf{x}$-vector that can actually be formed in the survey, we can at least obtain predictions of the ϕ_k: Determine $\boldsymbol{\lambda}$ to minimize $\sum_U \theta_k (\phi_k - \boldsymbol{\lambda}'\mathbf{x}_k)^2$; we find $\boldsymbol{\lambda} = \hat{\boldsymbol{\lambda}}_U,$ where $\hat{\boldsymbol{\lambda}}_U' = (\sum_U \mathbf{x}_k)'(\sum_U \theta_k \mathbf{x}_k \mathbf{x}_k')^{-1}$; the predicted value of ϕ_k is $\hat{\phi}_{kU} = \hat{\boldsymbol{\lambda}}_U' \mathbf{x}_k = M_k$. The (theta-weighted) first and second moment of the predictions $\hat{\phi}_{kU} = M_k$ are, respectively, $\bar{M}_{U;\theta} = \sum_U \theta_k M_k / \sum_U \theta_k = N / \sum_U \theta_k = \ 1/\bar{\theta}_U$ and

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

where $\bar{M}_U = \sum_U M_k / N$. Särndal and Lundström (2007) show that $relbias(\hat{Y}_{\text{CAL}})$ and Q have under certain conditions an approximately linear relationship,

$$relbias(\hat{Y}_{\text{CAL}}) \approx 1 - \frac{Q}{Q_0}$$

where $\bar{\phi}_U = \sum_U \phi_k / N$ and $Q_0 = (1/\bar{\theta}_U)(\bar{\phi}_U - 1/\bar{\theta}_U)$ is the maximum value of Q. Thus if Q were computable, it could serve as an indicator for comparing the different candidate $\mathbf{x}_k$-vectors. A computable analogue $\hat{Q}$ of Q is instead obtained as the variance of the corresponding sample-based predictions $\hat{\phi}_{ks} = \hat{\boldsymbol{\lambda}}'_s \mathbf{x}_k = (\sum_s d_k \mathbf{x}_k)' (\sum_r d_k \mathbf{x}_k \mathbf{x}'_k)^{-1} \mathbf{x}_k = m_k$, so that

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

where

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

We expect *relbias* to decrease in a roughly linear fashion as $\hat{Q}$ increases; thus, independently of the y-variables, $\hat{Q}$ may be used as a tool for ranking different **x**-vectors in regard to their capacity of reduce the bias.

We can use $\hat{Q}$ as a tool to select x-variables for inclusion in the $\mathbf{x}_k$-vector, for example, by stepwise forward selection, so that variables are added to $\mathbf{x}_k$ one at a time, the variable to enter in a given step being the one that gives the largest increment in $\hat{Q}$. The method is described in Särndal and Lundström (2007).

10. Calibration to account for other non-sampling error

Nonresponse errors are critical determinants of the quality of published statistics. When we examine how the calibration approach may intervene in the treatment other sources of non-sampling error than the nonresponse, the literature to date is not surprisingly much less extensive. However, several authors sketch a calibration reasoning to also incorporate frame errors, measurement errors, and outliers. Calibration has a potential to provide a more general theory for estimation in surveys, encompassing the various non-sampling errors.

As Deville (2004) points out (my translation from the French): "The concept of calibration lends itself to be applied with ease and efficiency to a great variety of problems in survey sampling. Its scope goes beyond that of regression estimation, an idea to which some seem to wish to reduce the calibration approach". He provides a brief sketch of how a treatment of several of the nonresponse errors may be accomplished under the caption of calibration thinking.

Folsom and Singh (2000) present a weight calibration method using what they call the generalized exponential model (GEM). It deals with three aspects: extreme value treatment, nonresponse adjustment and calibration through post-stratification. The method provides built-in control for extreme values. Calibration to treat both coverage errors (under- or over-coverage of the frame) and nonresponse is discussed in Särndal and Lundström (2005) and Kott (2006). Skinner (1998) discusses uses of calibration in the presence of nonresponse and measurement error. He notes something which remains a challenge almost ten years later: "More research is needed to investigate the properties of calibration estimates in the presence of non-sampling errors".

11. Conclusion

If I am to select one issue for a concluding reflection on the contents of this paper, let me focus on the concept of auxiliary information. It is the pivotal concept in the paper. If there is not auxiliary information, there is no calibration approach; there is nothing to calibrate on. I noted on the other hand that regression (GREG) estimation is an alternative but different thought process for putting auxiliary information to work in the estimation.

An objective in this paper has been to give a portrait of the two types of reasoning, and I made a point of noting how the thinking differs. I gave examples where essentially the same estimation objective is tackled by some authors through calibration reasoning, by others through GREG reasoning (or at least *primarily* by one or the other type). The respective estimators that they end up recommending may or may not agree. Whether or not the difference has significant consequence (for variance, for bias, for practical matters such as consistency and transparency) depends on the situation. This paper may help contributing an awareness of the separation existing between two thought processes that have guided researchers survey sampling.

References

Alexander, C.H. (1987). A class of methods for using person controls in household weighting. *Survey Methodology*, 13, 183-198.

Andersson, P.G., and Thorburn, D. (2005). An optimal calibration distance leading to the optimal regression estimator. *Survey Methodology*, 31, 95-99.

Ardilly, P. (2006). *Les techniques de sondage*. Paris: Editions Technip.

Bankier, M.D., Rathwell, S. and Majkowski, M. (1992). Two step generalized least squares estimation in the 1991 Canadian Census. Working paper, Census Operations Section, Social Surveys Methods Division, Statistics Canada.

Bankier, M., Houle, A.M. and Luc, M. (1997). Calibration estimation in the 1991 and 1996 Canadian censuses. *Proceedings*, *Section on Survey Research Methods*, American Statistical Association, 66-75.

Beaumont, J.-F. (2005a). Calibrated imputation in surveys under a quasi-model-assisted approach. *Journal of the Royal Statistical Society* B, 67, 445-458.

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

Beaumont, J.-F., and Alavi, A. (2004). Robust generalized regression estimation. *Survey Methodology*, 30, 195-208.

Bethlehem, J.G. (1988). Reduction of nonresponse bias through regression estimation. *Journal of Official Statistics*, 4, 251-260.

Bethlehem, J.G., and Keller, W.J. (1987). Linear weighting of sample survey data. *Journal of Official Statistics*, 3, 141-153.

Breidt, F.J., and Opsomer, J.D. (2000). Local polynomial regression estimators in survey sampling. *The Annals of Statistics*, 28, 1026-1053.

Breidt, F.J., Claeskens, G. and Opsomer, J.D. (2005). Model-assisted estimation for complex surveys using penalized splines. *Biometrika*, 92, 831-846.

Chambers, R.L. (1996). Robust case-weighting for multipurpose establishment surveys. *Journal of Official Statistics*, 12, 3-32.

Chambers, R.L., Dorfman, A.H. and Wehrly, T.E. (1993). Bias robust estimation in finite populations nonparametric calibration. *Journal of the American Statistical Association*, 88, 268-277.

Deming, W.E. (1943). *Statistical Adjustment of Data*. New York: John Wiley & Sons, Inc.

Demnati, A., and Rao, J.N.K. (2004). Linearization variance estimators for survey data. *Survey Methodology*, 30, 17-26.

Deville, J.C. (1998). La correction de la nonréponse par calage ou par échantillonnage équilibré. Paper presented at the Congrès de l'ACFAS, Sherbrooke, Québec.

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

Deville, J.C. (2004). Calage, calage généralisé et hypercalage. Internal document, INSEE, Paris.

Deville, J.C., and Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87, 376-382.

Deville, J.C., Särndal, C.-E. and Sautory, O. (1993). Generalized raking procedures in survey sampling. *Journal of the American Statistical Association*, 88, 1013-1020.

Duchesne, P. (1999). Robust calibration estimators. *Survey Methodology*, 25, 43-56.

Dupont, F. (1995). Alternative adjustments where there are several levels of auxiliary information. *Survey Methodology*, 21, 125-135.

Ekholm, A., and Laaksonen, S. (1991). Weighting via response modeling in the Finish Household Budget Survey. *Journal of Official Statistics*, 3, 325-337.

Estevao, V.M., and Särndal, C.-E. (2000). A functional form approach to calibration. *Journal of Official Statistics*, 16, 379-399.

Estevao, V.M., and Särndal, C.-E. (2002). The ten cases of auxiliary information for calibration in two-phase sampling. *Journal of Official Statistics*, 18, 233-255.

Estevao, V.M., and Särndal, C.-E. (2004). Borrowing strength is not the best technique within a wide class of design-consistent domain estimators. *Journal of Official Statistics*, 20, 645-660.

Estevao, V.M., and Särndal, C.-E. (2006). Survey estimates by calibration on complex auxiliary information. *International Statistical Review*, 74, 127-147.

Firth, D., and Bennett, K.E. (1998). Robust models in probability sampling. *Journal of the Royal Statistical Society*, 60, 3-21.

Folsom, R.E., and Singh, A.C. (2000). The generalized exponential model for design weight calibration for extreme values, nonresponse and poststratification. *Proceedings*, *Section on Survey Research Methods*, American Statistical Association, 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.

Fuller, W.A., and Rao, J.N.K. (2001). A regression composite estimator with application to the Canadian Labour Force Survey. *Survey Methodology*, 27, 45-51.

Hansen, M.H., and Hurwitz, W.N. (1943). On the theory of sampling from finite populations. *Annals of Mathematical Statistics*, 14, 333-362.

Harms, T. (2003). Extensions of the calibration approach: Calibration of distribution functions and its link to small area estimators. Chintex working paper no. 13, Federal Statistical Office, Germany.

Harms, T., and Duchesne, P. (2006). On calibration estimation for quantiles. *Survey Methodology*, 32, 37-52.

Hidiroglou, M.A. (2001). Double sampling. *Survey Methodology*, 27, 143–154.

Hidiroglou, M.A., and Särndal, C.-E. (1998). Use of auxiliary information for two-phase sampling. *Survey Methodology*, 24, 11-20.

Horvitz, D.G., and Thompson, D.J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47, 663-685.

Huang, E.T., and Fuller, W.A. (1978). Nonnegative regression estimation for sample survey data. *Proceedings*, *Social Statistics Section*, American Statistical Association, 300-305.

Isaki, C.T., Tsay, J.H. and Fuller, W.A. (2004). Weighting sample data subject to independent controls. *Survey Methodology*, 30, 35-44.

Kalton, G., and Flores-Cervantes, I. (1998). Weighting methods. In *New Methods for Survey Research* (Eds. A.Westlake, J. Martin, M. Rigg and C. Skinner),. Berkeley, U.K.: Association for Survey Computing.

Kim, J., Breidt, F.J. and Opsomer, J.D. (2005). Nonparametric regression estimation of finite population totals under two-stage sampling. Unpublished manuscript.

Knottnerus, P., and van Duin, C. (2006). Variances in repeated weighting with an application to the Dutch Labour Force Survey. *Journal of Official Statistics*, 22, 565-584.

Kott, P.S. (2004). Comment on Demnati and Rao: Linearization variance estimators for survey data. *Survey Methodology*, 30, 27-28.

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

Kovaĉević, M.S. (1997). Calibration estimation of cumulative distribution and quantile functions from survey data. *Proceedings of the Survey Methods Section*, Statistical Society of Canada, 139-144.

Krapavickaitė, D., and Plikusas, A. (2005). Estimation of a ratio in the finite population. *Informatica*, 16, 347-364.

Lavallée, P. (2007). *Indirect Sampling*. New York: Springer-Verlag.

LeGuennec, J., and Sautory, O. (2002). CALMAR2: une nouvelle version de la macro CALMAR de redressement d'échantillon par calage. *Actes des Journées de Méthodologie*, INSEE, Paris.

Lehtonen, R., and Veijanen, A. (1998). Logistic generalized regression estimators. *Survey Methodology*, 24, 51-55.

Lehtonen, R., Särndal, C.-E. and Veijanen, A. (2003). The effect of model choice in estimation for domains, including small domains. *Survey Methodology*, 29, 33-44.

Lehtonen, R., Särndal, C.-E. and Veijanen, A. (2005). Does the model matter? Comparing model-assisted and model-dependent estimators of class frequencies for domains. *Statistics in Transition*, 7, 649-674.

Lemaître, G., and Dufour, J. (1987). An integrated method for weighting persons and families. *Survey Methodology*, 13, 199-207.

Lundström, S., and Särndal, C.-E. (1999). Calibration as a standard method for treatment of nonresponse. *Journal of Official Statistics*, 15, 305-327.

Merkouris, T. (2004). Combining independent regression estimators from multiple surveys. *Journal of the American Statistical Association*, 99, 1131-1139.

Montanari, G.E. (1987). Post-sampling efficient prediction in large-scale surveys. *International Statistical Review*, 55, 191-202.

Montanari, G.E., and Ranalli, M.G. (2003). On calibration methods for design-based finite population inferences. Bulletin of the International Statistical Institute, 54th session, volume LX, contributed papers, book 2, 81-82.

Montanari, G.E., and Ranalli, M.G. (2005). Nonparametric model-calibration estimation in survey sampling. *Journal of the American Statistical Association*, 100, 1429-1442.

Myrskylä, M. (2007). Generalised regression estimation for domain class frequencies. Statistics Finland Research Reports 247.

Nieuwenbroek, N.J. (1993). An integrated method for weighting characteristics of persons and households using the linear regression estimator. Report, Central Bureau of Statistics, The Netherlands.

Nieuwenbroek, N.J., and Boonstra, H.J. (2002). Bascula 4.0 for weighting sample survey data with estimation of variances. The Survey Statistician, Software Reviews, July 2002.

Nieuwenbroek, N.J., Renssen, R.H. and Hofman, L. (2000). Towards a generalized weighting system. In *Proceedings*, *Second International Conference on Establishment Surveys*, American Statistical Association, Alexandria VA.

Park, M. and Fuller, W.A. (2005). Towards nonnegative regression weights for survey samples. *Survey Methodology*, 31, 85-93.

Plikusas, A. (2006). Non-linear calibration. Proceedings, Workshop on Survey Sampling, Venspils, Latvia. Riga: Central Statistical Bureau of Latvia.

Ren, R. (2002). Estimation de la fonction de répartition et des fractiles d'une population finie. *Actes des journées de méthodologie statistique*, INSEE Méthodes, tome 1, 100, 263-289.

Renssen, R.H., and Nieuwenbroek, N.J. (1997). Aligning estimates for common variables in two or more sample surveys. *Journal of the American Statistical Association*, 92, 368-374.

Renssen, R.H., Kroese, A.H. and Willeboordse, A.J. (2001). Aligning estimates by repeated weighting. Report, Central Bureau of Statistics, The Netherlands.

Rizzo, L., Kalton, G. and Brick, J.M. (1996). A comparison of some weighting adjustment methods for panel nonresponse. *Survey Methodology*, 22, 43-53.

Rueda, M., Martínez, S., Martínez, H. and Arcos, A. (2007). Estimation of the distribution function with calibration methods. *Journal of Statistical Planning and Inference*, 137, 435-448.

Särndal, C.-E. (1982). Implications of survey design for generalized regression estimation of linear functions. *Journal of Statistical Planning and Inference*, 7, 155-170.

Särndal, C.-E., and Swensson, B. (1987). A general view of estimation for two phases of selection with application to two-phase sampling and nonresponse. *International Statistical Review*, 55, 279-294.

Särndal, C.E., Swensson, B. and Wretman, J. (1992). *Model-assisted Survey Sampling*. New York: Springer-Verlag.

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

Särndal, C.-E., and Lundström, S. (2007). Assessing auxiliary vectors for control of nonresponse bias in the calibration estimator. Statistics Sweden: Research and development - methodology report 2007:2, to appear, *Journal of Official Statistics*.

Singh, A.C., and Mohl, C.A. (1996). Understanding calibration estimators in survey sampling. *Survey Methodology*, 22, 107-115.

Singh, S., Horn. S. and Yu, F. (1998). Estimation of variance of general regression estimator: Higher level calibration approach. *Survey Methodology*, 24, 41-50.

Skinner, C. (1998). Calibration weighting and non-sampling errors. *Proceedings International Seminar on New Techniques for Statistics,* Sorrento, November 4-6, 1998, 55-62.

Statistics Canada (1998). Methodology of the Canadian Labour Force Survey. Statistics Canada, Household Survey Methods Division. Ottawa: Minister of Industry, catalogue no. 71-526-XPB.

Statistics Canada (2003). Quality Guidelines (fourth edition). Ottawa: Minister of Industry, Catalogue no. 12-539-XIE.

Steel, D.G., and Clark, R.G. (2007). Person-level and household-level regression estimation in household surveys. *Survey Methodology*, 33, 51-60.

Stukel, D.M., Hidiroglou, M.A. and Särndal, C.-E. (1996). Variance estimation for calibration estimators: A comparison of jackknifing versus Taylor linearization. *Survey Methodology*, 22, 117-125.

Théberge, A. (1999). Extension of calibration estimators in survey sampling. *Journal of the American Statistical Association*, 94, 635-644.

Théberge, A. (2000). Calibration and restricted weights. *Survey Methodology*, 26, 99-107.

Tillé, Y. (2002). Unbiased estimation by calibration on distribution in simple sampling designs without replacement. *Survey Methodology*, 28, 77-85.

Tracy, D.S., Singh, S. and Arnab, R. (2003). Note on calibration in stratified and double sampling. *Survey Methodology*, 29, 99-104.

Vanderhoeft, C. (2001). Generalised calibration at Statistics Belgium. SPSS Module g-CALIB-S and current practices. Statistics Belgium Working Paper no. 3. Available at: www.statbel.fgov.be/studies/paper03_en.asp.

Vanderhoeft, C., Waeytens, E. and Museux, J.M. (2001). Generalised calibration with SPSS 9.0 for Windows baser. In *Enquêtes, Modèles et Applications* (Eds. J.J. Droesbeke and L. Lebart), Paris: Dunod

Webber, M., Latouche, M. and Rancourt, E. (2000). Harmonised calibration of income statistics. Statistics Canada, internal document, April 2000.

Wu, C. (2003). Optimal calibration estimators in survey sampling. *Biometrika*, 90, 937-951.

Wu, C., and Sitter, R.R. (2001). A model-calibration approach to using complete auxiliary information from survey data. *Journal of the American Statistical Association*, 96, 185-193.

Zheng, H., and Little, R.J.A. (2003). Penalized spline model-based estimation of the finite population total from probability-proportional-to-size samples. *Journal of Official Statistics*, 19, 99-117.

Zieschang, K.D. (1990). Sample weighting methods and estimation of totals in the Consumer Expenditure Survey. *Journal of the American Statistical Association*, 85, 986-1001.