An Introduction to Calibration Weighting for Establishment Surveys

Phillip S. Kott
RTI International, 6110 Executive Blvd., Suite 902, Rockville, MD 20852, U.S.A

Abstract

Calibration weighting is a general technique for adjusting probability-sampling weights to increase the precision of estimates, account for unit nonresponse or frame errors, or force internal estimates to be consistent with external measures. Special cases of calibration weighting include poststratification, weighting-class adjustments for nonresponse, raking (iterative proportional fitting), and separate ratio estimation. Although the literature has primarily focused on linear calibration techniques having close ties to the general regression or GREG estimator, attractive nonlinear calibration techniques have also been developed. The goal here is to provide a general introduction to calibration weighting for establishment surveys using as a motivating example a survey of hospital emergency departments.

Key Words: Survey weight; Ratio; Poststratification; Raking; Generalized exponential form; Variance.

1. Introduction

Establishment surveys based on randomly drawn samples come with survey weights allowing valid inferences about the target population. Given a probability sample with probability-based survey weights, calibration weighting involves a mild adjustment to those weights that forces the weighted totals for a set of calibration (or benchmark) variables to equal values determined using more complete information whether from the the target population itself or a larger sample. Many establishment surveys have access to frame data that can be used for this purpose.

There are three purposes of calibration weighting:

1. Decreasing the variance of many (nearly) unbiased estimators.
2. Removing or decreasing selection biases caused by unit nonresponse or frame errors.
3. Forcing estimates to be numerically consistent with external measures.

1.1 A Motivating Example

The following example of an establishment survey will motivate much of this discussion. Suppose we are interested in producing annual estimates of drug-related emergency-department visits to US hospitals. We have a list of all US hospital with emergency departments, but contacting each and collecting the information we want would be prohibitively expensive. Instead, we draw a random sample from the list and collect the information only from that sample. Moreover, it may be that not all the sample hospitals cooperate, and we able are to survey only a subset of them. For simplicity, we will assume that all cooperating hospitals supply all the information we request of them.

The list of all US hospitals with emergency departments is called a "sampling frame." Let F denote this frame, and k an emergency department on this frame. The frame comes with a fair amount of information about all the hospitals on it, such as their location, whether each is publicly or privately owned, and the annual number of emergency-department visits in a recent year (drug-related or not).

We draw a random sample S of n hospitals in F. Often, this sample is drawn by first dividing the population into mutually exclusive strata based on location, size and ownership (i.e., public or private), then selecting a simple random sample in each stratum. An example of just such a sample of hospital emergency departments can be found within the SUDAAN 11/WTADJX examples in http://www.rti.org/sudaan/. This is not a real sample. Rather, it is based on a public use file but some numbers have been adjusted and others have been created out of whole cloth for our illustrative purposes.

Let
y_{1k} be the number of annual drug-related visits to k collected on the survey,
d_k be the probability-sampling (or design) weight associated with k (the inverse of its probability of sample selection), and
q_k be the number of emergency-departments visits to k on the frame.

Assume that the frame contains the complete universe U of US hospitals with emergency departments and, for now, that there is no nonresponse.

1.2 Estimating Totals

Our primary goal is to estimate the universe total $T_{y_1} = T_1 = \sum_{k \in U} y_{1k} = \sum_U y_{1k}$. One simple way to do that is with the *expansion estimator* $t_1 = \sum_S d_k y_{1k}$, which is unbiased under probability theory (i.e., $E_{prob}(t_1) = T_1$). Under mild conditions, it also has small relative error, $(t_1 - T_1)/T_1$ when n is large (technically, $(t_1 - T_1)/T_1 = O_P(1/\sqrt{n})$).

The expansion estimate can be used to estimate the universe total for any item given its values in the sample. In particular, $t_N = \sum_S d_k 1 = \sum_S d_k$ is an unbiased estimator for N, the size of U. Define u_{gk} to be 1 when k is in a group U_g (e.g., *public* hospitals), and 0 otherwise. Let $N_g = \sum_{U_g} 1 = \sum_{k \in U} u_{gk}$ be the size of U_g, which can be estimated by

$$t_{N_g} = \sum_{S_g} d_k = \sum_S d_k u_{gk}.$$

For public hospitals in our example data set, N_g is 1632, while its estimator, t_{N_g}, is 1647.

In general, the relative error of t_{N_g} will be small under mild conditions when n_g is large. This is some disagreement, however, about how large is "large." Opinions range from 6 to 20.

2. Sample Balance

When, regardless of which random sample is drawn,

$$\sum_S w_k = \sum_U 1 \equiv N, \tag{1}$$

$$\sum_S w_k q_k = \sum_U q_k \equiv T_q, \tag{2}$$

$$\text{or } \sum_S w_k u_{public,k} = \sum_U u_{public,k} \equiv \sum_{U_{metro}} 1 \equiv N_{public}, \tag{3}$$

and each $w_k = d_k$, the sample design is said to be *balanced* (or *design balanced*) on 1, the q_k, or the public-indicator ,respectively. The three equations t are called "balancing equations."

It is not hard to show that any simple random or stratified simple random sample is balanced on 1. Similarly, a stratified simple random sample where no strata contains both public and private hospitals is balanced on the public indicator. Finally, a sample drawn with probabilities proportional to q_k will be balanced on the q_k (as long as no q_k is larger than $\sum_U q_j/n$).

Sometimes, we would like an equation like $\sum_S w_k q_k = \sum_U q_k \equiv T_q$ to hold exactly – and perhaps additional balancing equations as well – even when the sample is not design balanced on the q_k. Calibration weighting creates a set of weights $\{w_k\}$ close the probability-sampling weights (the d_k) which satisfy balancing equation. In particular, $w_k = d_k \alpha_k$, and each α_k is close to 1 (technically, each $\alpha_k = 1 + O_P(1/\sqrt{n})$.)

Using probability-sampling weights assures us that balancing equations hold on average across all possible samples. Calibration weighting provides a mild adjustment that makes the equalities hold exactly given a particular sample. The balancing equations, which may or may hold exactly depending on the design, become *calibration equations* which our choices for the w_k force to hold.

A type of calibration equation that is rarely a balancing equation but is often used in practice has the form:

$$\sum_S w_k y_{2k} = \tilde{t}_2,$$

where y_{2k} is a variable collected on the survey, but $\tilde{t}_2$ is an estimate from an outside source deemed more reliable than $\sum_S d_k y_{2k}$. This type of calibration equation is used to make estimates from different sources numerically consistent. We will not discuss it further here.

The calibration estimator for $T_y = \sum_U y_k$, where y_k is any variable collected on the survey, is $t_{y,cal} = \sum_S w_k y_k$. Under mild conditions on the sample design and population, $t_{y,cal}$ is nearly unbiased in that $\text{Bias}(t_{y,cal})/\text{Var}(t_{y,cal}) \approx 0$.

Sometimes the benchmark totals for calibration equations come from a source other than the sampling frame. Often, however, the frame of an establishment survey will have some information on every element in the frame or nearly so. *Control variables* are

variables with known values for every element on the frame. In practice, some of these values may have to be imputed. *This should be done without using values collected on the survey itself.* Indeed, even if the survey provides more accurate information than the frame, survey values should not be used to update or correct errors perceived to be on the frame.

3. Simple Ratio Adjustments

Given a probability sample, the calibration equation equation in (2) will always hold if calibration weights are produced using the following adjustment:

$$w_k = \left(\frac{T_q}{\sum_S d_j q_j} \right) d_k,$$

since $\sum_S w_k q_k = \left(\frac{T_z}{\sum_S d_j q_j} \right) \sum_S d_k q_k = T_q$. In our example data set,

$$\alpha_k = \frac{T_z}{\sum_S d_j z_j} = \frac{132491601}{130612288} \approx 1$$

for all k in the sample.

Notice that this calibration weight adjustment produces the simple ratio estimator:

$$t_{y,cal} = \sum_S w_k y_k = T_q \left(\frac{\sum_S d_k y_k}{\sum_S d_k q_k} \right)$$

no matter what the survey variable y_k. For example, it would be the estimator even if y_k were $y_{1k} u_{pulic,k}$, which is the number of drug-related emergency department visits to hospital k when k is publicly owned and zero otherwise. Recall that q_k is positive for all k, so that $t_{y,cal.}$ is no longer the traditional ratio estimator for the total number of drug-related emergency department visits to public hospitals.

Now suppose U were divided into G mutually exclusive groups each of size N_g. The G calibration equations $\sum_S w_k u_{gk} = N_g$ are satisfied by the weight adjustment:

$$w_k = \left(\frac{N_g}{\sum_S d_j u_{jg}} \right) d_k \equiv \left(\frac{N_g}{\sum_{S_g} d_j} \right) d_k$$

for $k \in S_g$. These calibration weight adjustments produce the *poststratified estimator*:

$$t_{y,cal} = \sum_S w_k y_k = \sum_{g=1}^{G} \sum_{S_g} w_k y_k = \sum_{g=1}^{G} N_g \frac{\sum_{S_g} d_k y_k}{\sum_{S_g} d_k}.$$

4. Reducing Variances: The Working Prediction Model

Calibration weighting will tend to decrease the variance of an estimated total like $T_y = \sum_U y_k$ when y_k behaves like as a linear function of the calibration variables. For example, consider the *working prediction* (*or outcome) model*:

$$y_k = \beta q_k + \varepsilon_k, \tag{4}$$

which is often useful in establishment surveys. Observe that we can write:

$$\begin{aligned} t_y &= \sum_S d_k y_k = \beta \sum_S d_k z_k + \sum_S d_k \varepsilon_k = \beta t_z + \sum_S d_k \varepsilon_k, \\ t_{y,cal} &= \sum_S w_k y_k = \beta \sum_S w_k z_k + \sum_S w_k \varepsilon_k = \beta T_z + \sum_S w_k \varepsilon_k, \text{ and} \\ T_y &= \sum_U y_k = \beta \sum_U z_k + \sum_U \varepsilon_k = \beta T_z + \sum_U \varepsilon_k \end{aligned}$$

From which it is easy to see that while the variance or t_y as an estimator of T_y has two components: the difference between βt_q and βT_q and the difference between $\sum_S d_k \varepsilon_k$ and $\sum_U \varepsilon_k$, the variance of $t_{y,cal}$ as an estimator of T_y has only one: the difference between $\sum_S w_k \varepsilon_k$ and $\sum_U \varepsilon_k$.

The ε_k in equation (4) need not be normally distributed or even uncorrelated. In fact, the prediction model in equation (4) may not even hold in the traditional sense that $E_{model}(\varepsilon_k|q_k) = 0$ whether or not k is in the sample. Still, there can be a reduction in variance from calibration weighting.

An analogous variance-reduction rationale works for a vector of calibration variables $\mathbf{z}_k$, with $\boldsymbol{\beta}^T z_k$ replacing βq_k. Examples include:

A *group-mean model* like $m_1 u_{public,k} + m_2 u_{private,k}$, where m_1 is the public y-mean and m_2 the private mean. This is the model for the *poststratified estimator.*

A *group ratio model* like $b_1 (q_k u_{public,k}) + b_2 (q_k u_{private,k})$. This is the model for *the separate ratio estimator*.

A *simple regression model* is $a + bq_k$. This is the model for the *simple regression estimator*. It requires more than ratio adjustments to find calibration weights.

An example of the *general linear model is* $\boldsymbol{\beta}^T \mathbf{z}_k$ (e.g., $\beta_0 + \beta_1 u_{urban,k} + \beta_2 u_{public,k} + \beta_3 q_k$). More on this later.

5. Unit Nonresponse

Suppose the unit response mechanism can be described with the following *selection model*: Each $k \in U$ is in one of G mutually exclusive groups (poststrata), and the probability of $k \in U_g$ responds if sampled is $\pi_{(g)}$, which can vary *across* groups.

Denote the set of respondents within each group by R_g. Suppose, at first, that N_g is known (e.g., the number of public hospitals in a region of a certain size). In principle, $t_{N_g}^{(2)} = \sum_{R_g} d_k (1/\pi_{(g)})$ is an unbiased estimator for N_g with a small relative error when r_g is large.

By using the calibration equation,

$$\sum_{R_g} d_k \alpha_k = \sum_{R_g} d_k (1/\pi_{(g)}) = N_g$$

(which should approximately hold), to adjust the original weights, we implicitly estimate $\pi_{(g)}$. The solution comes from solving the calibration equation:

$$\alpha_k = \frac{1}{\hat{\pi}_{(g)}} = \left(\frac{N_g}{\sum_{R_g} d_j} \right) \Rightarrow w_k = \left(\frac{N_g}{\sum_{R_g} d_j} \right) d_k \quad \text{for } k \in R_g .$$

Note that α_k need not be near 1.

When N_g is not known, it can be replaced by its full-sample estimate *but only if that is known*, in this case because we can presumably determine ownership in the full sample but not in the population:

$$\sum_{R_g} d_k \alpha_k = \sum_{R_g} d_k (1/\hat{\pi}_{(g)}) = \sum_{S_g} d_k = \hat{N}_g ,$$

$$\text{So} \quad \alpha_k = \frac{1}{\hat{\pi}_{(g)}} = \left(\frac{\sum_{S_g} d_j}{\sum_{R_g} d_j} \right) \Rightarrow w_k = d_k \left(\frac{\sum_{S_g} d_j}{\sum_{R_g} d_j} \right) \text{ for } k \in R_g .$$

Since calibration weighting reduces variances, this choice for the α_k should be better than the unweighted maximum-likelihood solution, n_g/r_g for $k \in R_g$, *for estimating* T_y (but *not* for estimating the response rate within S_g).

6. Linear Calibration

Ratio adjustment doesn't work for the combination of these two calibration equations: $\sum_S w_k = N$ and $\sum_S w_k q_k = T_q$. In general, we can write a system of P calibration equations as a matrix equation:

$$\sum_S w_k \mathbf{z}_k = T_{\mathbf{z}}.$$

For example, with $P = 2$,

$$\mathbf{z}_k = \begin{pmatrix} 1 \\ q_k \end{pmatrix}, \quad T_{\mathbf{z}} = \begin{pmatrix} N \\ T_q \end{pmatrix}.$$

With *linear calibration weighting:*

$$w_k = d_k \left[1 + \mathbf{h}^T \mathbf{z}_k \right],$$

where **h** is chosen so that the calibration equation holds. This produces the the general regression (GREG) estimator:

$$\begin{aligned} \sum_S w_k y_k &= \sum_S d_k \left(1 + \mathbf{h}^T \mathbf{z}_k \right) y_k \\ &= \sum_S d_k \left(1 + \left(T_{\mathbf{z}} - t_{\mathbf{z}}\right)^T \left(\sum_S d_j \mathbf{z}_j \mathbf{z}_j{}^T \right)^{-1} \mathbf{z}_k \right) y_k \\ &= \sum_S d_k y_k + \left(T_{\mathbf{z}} - t_{\mathbf{z}}\right)^T \left(\sum_S d_j \mathbf{z}_j \mathbf{z}_j{}^T \right)^{-1} \sum_S d_k \mathbf{z}_k y_k \\ &= \sum_S d_k y_k + \left(T_{\mathbf{z}} - t_{\mathbf{z}}\right)^T \mathbf{b} \end{aligned}$$

(Here $\mathbf{h} = (\sum_S d_j \mathbf{z}_j \mathbf{z}_j{}^T)^{-1} \left(T_{\mathbf{z}} - t_{\mathbf{z}}\right).$)

Using linear calibration to treat unit nonresponse, we have the calibration equation:

$$\sum_R w_k \mathbf{z}_k = \sum_R d_k (1 + \mathbf{h}^T \mathbf{z}_k)\, \mathbf{z}_k = \sum_S d_k \mathbf{z}_k,$$

which leads to

$$\frac{1}{\hat{\pi}_k} = 1 + \mathbf{h}^T \mathbf{z}_k \quad \Rightarrow \quad \pi_k = \frac{1}{1 + \boldsymbol{\gamma}^T \mathbf{z}_k},$$

where **h** estimates **γ.** But this estimate can be less than 0 or greater than 1. Even when there is no nonresponse, a linear calibration weight can be less than 0.

Truncated linear calibration is an iterative method that successively fixes wayward α_k at their truncated values, removes the associated k from the population and sample, and performs linear calibration again (and again) on the remainder of the sample and population.

7. Raking

Suppose we have many calibration equations like $\sum_S d_k u_{gk} = t_{N_g}$ but the groups are *not* all mutually exclusive. For example, urban/non-urban, four regions, public/private. If we cross-classified the population, some groups could have few sampled or responding units. Some may even be empty.

We can ratio adjust first for urban and non-urban, then ratio adjust the result for the four regions, and then ratio adjust that result for public and private. We then can repeat the process as many times as needed until all calibration equations are effectively satisfied (which may never happen). This type of calibration weighting in called "iterative proportional fitting" or "raking."

Raking is equivalent to adjusting weights like so:

$$\alpha(\mathbf{h}^T\mathbf{z}_k) = \exp(\mathbf{h}^T\mathbf{z}_k) = [\exp(u_{urban,k})]^{h_1}[\exp(u_{non\text{-}urban,k})]^{h_2}[\exp(u_{east,k})]^{h_3}\cdots$$

where $\mathbf{z}_k = (u_{urban,k} \;\; u_{non\text{-}urban,k} \;\; u_{east,k} \;\; \ldots\ldots)^T$.

When there are *no* nonresponse or coverage errors: $\exp(\mathbf{h}^T\mathbf{z}_k) \approx 1 + \mathbf{h}^T\mathbf{z}_k$. When used *for* nonresponse or coverage adjustment, by contrast,

$$\alpha_k = \frac{1}{\hat{\pi}_k} = \exp\left(\mathbf{h}^T\mathbf{z}_k\right) \Rightarrow \pi_k = \exp(-\boldsymbol{\gamma}^T\mathbf{z}_k),$$

which need *not* be $\approx 1 - \boldsymbol{\gamma}'\mathbf{z}_k$. These estimated probabilities of response are always nonnegative but can exceed 1.

Finding an $\mathbf{h}$ such that $w_k = d_k\exp(\mathbf{h}^T\mathbf{z}_k)$ and $\sum_S w_k\mathbf{z}_k = T_{\mathbf{z}}$ (or $\sum_R w_k\mathbf{z}_k = \sum_S w_k\mathbf{z}_k$) can often be done *even when components of* $\mathbf{z}_k$ *are continuous.*

8. Two Distinct Types of Models

Although $\alpha_k = \alpha(\mathbf{h}^T\mathbf{z}_k) = \exp(\mathbf{h}^T\mathbf{z}_k)$ and the $\pi_k = \exp(-\boldsymbol{\gamma}^T\mathbf{z}_k)$ in the *selection* model are *not* linear functions of the component of $\mathbf{z}_k$, the working *prediction* (outcome) model for a *y*-variable remains linear; e.g., drug-related visits is a linear function of frame visits and other covariates. Formally, $E_{model}(y_k - \mathbf{z}_k^T\boldsymbol{\beta}|\mathbf{z}_k) = 0$ for some vector $\boldsymbol{\beta}$ whether or not *k* responds when sampled (i.e., the response mechanism is assumed to be ignorable; the sampling mechanism may be ignorable as well, but the use of the d_k in determining the calibration weights makes that additional assumption superfluous). A key difference between the two types of models is that it is possible for the prediction model to hold for some survey variables but not others. By contrast, either the selection model holds or it doesn't regardless of the survey variables.

If *either* the selection model or the prediction model holds, then the resulting estimator for T_y is nearly unbiased in some sense: there is *double protection* against nonresponse bias.

9. Bounding and Centering Parameters

Selection and prediction models can also be used to address coverage errors when population totals for the **z**-variables are known from an outside source but there are duplications or missing elements on the frame (so that $F \neq U$),

When there is *no* nonresponse or coverage error (or adjustments for them have already been made), there are many possible versions of $\alpha_k = \alpha(\mathbf{h}^T\mathbf{z}_k) \approx 1 + \mathbf{h}^T\mathbf{z}_k$ that are nearly identical with $\alpha(0) = \alpha'(0) = 1$. *There is no reason to set up a loss function to choose among them.*

One attractive version is *generalized raking*

$$\alpha_{GR}\left(\mathbf{h}^T\mathbf{z}_k;\, \ell, u\right) = \frac{\ell(u-1) + u(1-\ell)\exp(A\mathbf{h}^T\mathbf{z}_k)}{(u-1) + (1-\ell)\exp(A\mathbf{h}^T\mathbf{z}_k)},$$

where $0 \le \ell < 1 < u \le \infty$ (and A defined so that $\alpha_{GR}'(0;\, \ell,\, u) = 1$), which constrains α_k between the *bounding parameters* ℓ and u. With conventional raking, $\ell = 0$ and $u = \infty$. Newton's method is used to find **h**. As a result the components of $\mathbf{z}_k$ need not be 0/1.

Even more general is the *general exponential form* (*of weight adjustment*):

$$\alpha_{GEF}\left(\mathbf{h}^T\mathbf{z}_k;\, \ell_k, u_k, c_k\right) = \frac{\ell_k(u_k - c_k) + u_k(c_k - \ell_k)\exp(A_k\mathbf{h}^T\mathbf{z}_k)}{(u_k - c) + (c_k - \ell_k)\exp(A_k\mathbf{h}^T\mathbf{z}_k)},$$

where $0 \le \ell_k < c_k < u_k \le \infty$, (and A_k is defined so that $\alpha_{GEF}'(0;\, \ell_k,\, u_k,\, c_k) = 1$). It allows the bounds (and the center) to vary across k. Yet when all $c_k = 1$, and there is no nonresponse or coverage errors, it is asymptotically indistinguishable from linear calibration.

If response (or coverage) is a function of the components of $\mathbf{z}_k$, then *the choice of* $\alpha_k = \alpha(\mathbf{h}^T\mathbf{z}_k)$ *matters,* since $\alpha(0)$ may differ from 1. In particular, the general exponential form can be $\alpha_{GEF}(\mathbf{h}^T\mathbf{z}_k;\, \ell,\, u,\, c)$, where $\alpha_{GEF}(\mathbf{0};\, \ell,\, u,\, c) = c$. A special case, $\alpha_{GEF}(\mathbf{h}^T\mathbf{z}_k;\, 1,\, \infty,\, c > 1)$ corresponds to a logistic response model.

10. Instrumental Variables

With linear calibration:

$$w_k = d_k\left[1 + \mathbf{h}^T\mathbf{z}_k\right] = d_k\left[1 + \mathbf{z}_k{}^T\mathbf{h}\right] = d_k\left[1 + \mathbf{z}_k{}^T\left(\sum\nolimits_S d_j\mathbf{z}_j\mathbf{z}_j{}^T\right)^{-1}\left(T_\mathbf{z} - t_\mathbf{z}\right)\right] \Rightarrow$$

$$\sum\nolimits_S w_k\mathbf{z}_k = \sum\nolimits_S d_k\mathbf{z}_k + \sum\nolimits_S d_k\mathbf{z}_k\mathbf{z}_k{}^T\left(\sum\nolimits_S d_j\mathbf{z}_j\mathbf{z}_j{}^T\right)^{-1}\left(T_\mathbf{z} - t_\mathbf{z}\right) = T_\mathbf{z}$$

Similarly for a vector $\mathbf{x}_k$ with the same number of components as $\mathbf{z}_k$,

$$w_k = d_k\left[1+\mathbf{h}^T\mathbf{x}_k\right] = d_k\left[1+\mathbf{x}_k{}^T\left(\sum_S d_j\mathbf{z}_j\mathbf{x}_j{}^T\right)^{-1}\left(T_\mathbf{z}-t_\mathbf{z}\right)\right] \;\Rightarrow$$

$$\sum_S w_k\mathbf{z}_k = \sum_S d_k\mathbf{z}_k + \sum_S d_k\mathbf{z}_k\mathbf{x}_k{}^T\left(\sum_S d_j\mathbf{z}_j\mathbf{x}_j{}^T\right)^{-1}\left(T_\mathbf{z}-t_\mathbf{z}\right) = T_\mathbf{z}$$

Components of $\mathbf{x}_k$ that are not in $\mathbf{z}_k$ are called *instrumental variables.*

In the simple ratio, q_j is the calibration vector and 1 is the instrumental-variable vector:

$$w_k = d_k\left[1+\mathbf{1}^T\left(\sum_S d_j\mathbf{q}_j\mathbf{1}^T\right)^{-1}\left(T_\mathbf{q}-t_\mathbf{q}\right)\right]$$

$$= d_k\left[1+\left(\sum_S d_j q_j\right)^{-1}\left(T_q-t_q\right)\right] = d_k\frac{T_q}{\sum_S d_j q_j}.$$

Pseudo-optimal calibration sets all $\mathbf{x}_k = (d_k - 1)\mathbf{z}_k$ (including an intercept if there is one). This will often reduce the variance when there is *no* nonresponse or coverage error and y_k is *not* strictly a linear function of the components of **z**. This is because **b** in the GREG is replaced by an estimate of $[\mathbf{Var}_{design}(t_\mathbf{z})]^{-1}\mathbf{Cov}_{design}(t_\mathbf{z}, t_y)$ under Poisson sampling or *stsrs* with large stratum sample sizes.

When there is *no* nonresponse or coverage error, linear calibration with instruments is nearly equal to setting

$$\alpha_{GR}\left(\mathbf{h}^T\mathbf{x}_k;\ell,u\right) = \frac{\ell(u-1)+u(1-\ell)\exp(A\mathbf{h}^T\mathbf{x}_k)}{(u-1)+(1-\ell)\exp(A\mathbf{h}^T\mathbf{x}_k)}.$$

If we add *k*-indices to the upper and lower bounds (as can be done in SUDAAN 10 and 11), then the weights themselves, not just the weight adjustments, can be bound. For example, we can restrict weights to be no smaller than 1 or we can bound the size of $w_k q_k$.

When there *is* nonresponse, fitting $\alpha_{GEF}(\mathbf{h}^T\mathbf{x}_k;\ell\geq 1,u,c>\ell)$ bounds the weight adjustment between ℓ and u, and thus the estimated probability of response between $1/u$ and $1/\ell$ since $\hat{\pi}_k = 1/\alpha_k$. It also allows some of the components in the $\mathbf{x}_k$ vector governing the response model to be *y*-variables known only for respondents. Thus, treating unit nonresponse that is *not* missing at random is possible (and available with WTADJX in SUDAAN 11). The components of $\mathbf{x}_k$ are then called "response-model variables."

It is also possible to fit a response model when the **x**-vector has fewer components than the calibration vector $\mathbf{z}_k$ (but we won't).

11. Some Comments on Variance Estimation

In general,

$$t_{y,cal} = \sum_S w_k y_k = \beta \sum_S w_k \mathbf{z}_k + \sum_S w_k \varepsilon_k$$

Traditional methods of linearization variance estimation applied to $t_{y,cal}$ capture the variability of the w_k in $\sum_S w_k \varepsilon_k$, but not the reduction in variance from $\sum_S w_k \mathbf{z}_k$ being calibrated.

Replication methods, by contrast, capture both (the former slightly more than is warranted). When calibration fails in a replicate (i.e., a calibration equation is not satisfied) and the replicate isn't dropped from the variance estimator, that estimator usually will be biased *upward*.

The linearization routine in SUDAAN 11 will captures the variance reduction from one calibration step assuming either $T_{\mathbf{z}}$ or, in the case of unit nonresponse, $\sum_S d_k \mathbf{z}_k$ is known. If there really is only one calibration step, it may underestimate slightly because the $(y_k - \mathbf{z}_k{}^T \mathbf{b})^2$, where $\mathbf{b}$ is a sample estimate will tend to be smaller than $(y_k - \mathbf{z}_k{}^T \mathbf{b}^*)^2$, where $\mathbf{b}^*$ is the probability limit of $\mathbf{b}$.

Otherwise, it may tend to *overestimate* the variance because it fails to capture variance reduction from previous steps.

Useful References and Additional Reading

Bankier, M. (2002). Regression Estimators for the 2001 Canadian Census, presented at the International Conference in Recent Advances in Survey Sampling.

Chang, T. and P. S. Kott (2008). Using Calibration Weighting to adjust for Nonresponse Under a Plausible Model, *Biometrika*, 95, 557-571.

Deville, J.-C. (2000). Generalized Calibration and Application to Weighting for Non-response, *Compstat: Proceedings in Computational Statistics; 14th Symposium Held in Utrecht, The Netherlands* (available online).

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

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

Fuller, W.A., Loughin, M.M., and Baker, H.D. (1994). Regression Weighting for the 1987-88 National Food Consumption Survey, *Survey Methodology*, 20, 75-85.

Kim, J. K. and Park, M. (2010). Calibration Estimation in Survey Sampling, International Statistical Review 78, 21-39.

Kott, P.S. (2011). A Nearly Pseudo-optimal Method for Keeping Calibration Weights from Falling Below Unity in the Absence of Nonresponse or Frame Errors, *Pakistan Journal of Statistics*, 27, 391-396.

Kott, P.S. (2009). Calibration Weighting: Combining Probability Samples and Linear Prediction Models, in *Handbook of Statistics_29B: Sample Surveys: Inference and Analysis*, D. Pfeffermann, & C.R. Rao (Eds.), New York: Elsevier.

Kott, P.S. and Day, C.D. (2012). Developing Calibration Weights and Standard-Error Estimates for a Survey of Drug-Related Emergency-Room Visits, Proceedings of the Third International Conference of Establishment Surveys, June 18-21, 2007, Montréal, Canada [CD-ROM]: forthcoming.

Kott, P.S. and Liao, D. (2012). Providing Double Protection for Unit Nonresponse with a Nonlinear Calibration-weighting Routine, *Survey Research Methods*, 6, 2, 105-11.

Rao, J.N.K. (1994). Estimating Totals and Distribution Functions Using Auxiliary Information at the Estimations Stage, *Journal of Official Statistics* 25, 1-21.

RTI (2012). *SUDAAN Language Manual, Release 11.0.* Research Triangle Park, NC: RTI International.