”Optimal” Calibration Weights Under Nonresponse in Survey Sampling

Per Gösta Andersson, Stockholm University

1 Introduction

High nonresponse is a very common problem in sample surveys today. In statistical terms we are worried about increased bias and variance of estimators for population quantities such as totals or means. Different methods have been suggested in order to compensate for this phenomenon. We can roughly divide them into imputation and calibration and it is the latter approach we will focus on here. A wide spectrum of possibilities is included in the class of calibration estimators.
We need to distinguish between different levels of availability of variable values: The population level, the sample level and the response level. The sample is drawn by probability sampling from the population and our calibration approach is design-based. The response is the subset of the sample for which the study variable values are individually observed. Auxiliary variables are essential. For the calibration technique studied in this paper, an auxiliary variable must contain information at a higher level than the response and its value must be known individually for all units in the response.

Notation and setup

We will start with a population U of size N from which we take a probability sample s of size n_s with inclusion probabilities $\pi_1, \ldots, \pi_N$. Nonresponse means that we only observe the response set r of size n_r. Our aim is to estimate the study variable total $t_y = \sum_U y_k$. We assume access to an auxiliary variable vector x of dimension J, where either $(x_k)_{k \in U}$ (the population level) or $(x_k)_{k \in s}$ (the sample level) are known or possibly a mixture.

2 Calibration estimation

2.1 Calibration estimators under full response

Starting with the full response situation ($r = s$) and following the procedure as established by Deville and Särndal (1992), the calibration estimator is defined as

$$\hat{t}_{y\,cal} = \sum_s w_{ks} y_k,$$

where the sample dependent weights w_{ks} are chosen so that

$$\sum_s w_{ks} x_k = t_x, \text{ (the calibration equation)} \tag{1}$$

while also minimizing the quadratic distance measure

$$(w_s - w_{0s})' R (w_s - w_{0s}),$$

where $w_s = (w_{ks})_{k \in s}$, $w_{0s} = (1/\pi_k)_{k \in s} = (d_k)_{k \in s}$ and R is diagonal.
In other words, given the constraint (1) the w_{ks} should be "as close as possible" to the design weights d_k.
The resulting weights are

$$w_s = w_{0s} + R^{-1} x' (X R^{-1} X')^{-1} (t_x - \hat{t}_x)$$

It turns out that the model assisted GREG estimator $\hat{t}_{y\,r}$ (Särndal, Swensson and Wretman (1992)) is a calibration estimator for which

$$R = \left(w_{0s} I_{n_s}\right)^{-1},$$

where I_{n_s} is the unit diagonal matrix of size n_s.
Another calibration estimator is the optimal regression estimator $\hat{t}_{y\,opt}$ (see e.g. Rao (1994) and Montanari (1998)), for which

$$R = (\frac{\pi_{kl} - \pi_k \pi_l}{\pi_{kl} \pi_k \pi_l})^{-1}_{k,l \in s},$$

as shown by Andersson and Thorburn (2005).
Asymptotically, this estimator has (in a design-based sense) minimum variance among linear regression type estimators.

2.2 Calibration estimators under nonresponse

In the nonresponse case, a possible calibration estimator is

$$\hat{t}_{y\,cal} = \sum_r w_{kr} y_k,$$

where it should hold that

$$\sum_r w_{kr} x_k = X, \tag{2}$$

where $X = \sum_U x_k$, if the auxiliary information is known up to the population level. Otherwise, $X = \sum_s d_k x_k$, the unbiased estimator of t_x. (We can also combine the two types of information in the constraint X.)

Weights fulfilling the requirement (2) are presented by Särndal and Lundström (2005). Using the direct approach, where all information is used in one single calibration, we get

$$w_{kr} = d_k(1 + x_k'(\sum_r d_k x_k x_k')^{-1}(X - \sum_r d_k x_k) \tag{3}$$

The resulting estimator will henceforth be denoted $\hat{t}_{y\,cal}$. Other approaches, including two-step procedures, are presented and investigated by Särndal and Lundström (2005) and Andersson and Särndal (2016).

A natural question to ask is: What is the underlying distance measure generating these weights? Särndal and Lundström (2005) do not comment on this particular issue, but according to Lundström and Särndal (1999), we should choose "w_k 'as close as possible' to the d_k", which does not seem quite adequate under nonresponse. Going back to Lundström (1997) we will find that the corresponding distance measure is actually

$$(w_r - w_{0r})'(w_{0r}I_{n_r})^{-1}(w_r - w_{0r}),$$

where $w_r = (w_{kr})_{k\in r}$ and $w_{0r} = (d_k)_{k\in r}$.

If we have a random mechanism generating the response set r from the sample s with probabilities θ_k of inclusion, we can view the nonresponse situation as a two-phase design. Then we should minimize the distance between w_{kr} and $d_k \cdot (1/\theta_k)$. Using some modelling θ_k can be estimated by $\hat{\theta}_k$, to be put to use for the distance minimization. However, we will not go in the direction of model-based inference. In order to reduce the bias effect under nonresponse one could instead in the distance measure think of comparing w_{kr} not with d_k, but with $d_k^* = d_k \cdot c$, where c is a constant larger than 1, aiming to compensate for the "average" nonresponse effect.

However, Lundström (1997) shows that for many interesting cases, namely when one can find a vector μ for which $\mu' x_k = 1$, for all k, the multiplicative increase in d_k^* implies the same resulting calibration weights w_{kr}. This follows from the result that if $\mu' x_k = 1$, for all k, we can simplify the expression of w_{kr} as

$$w_{kr} = d_k x_k'(\sum_r d_k x_k x_k')^{-1}X$$

Thus, we have an invariance property for the weights.

With this as a background we propose to use alternative "optimal" weights resulting from the distance measure

$$(w_r - w_{0r})'(\frac{\pi_{kl} - \pi_k\pi_l}{\pi_{kl}\pi_k\pi_l})^{-1}_{k,l\in r}(w_r - w_{0r}),$$

leading to $\hat{t}_{y\,opt}$.

As for the full response situation, there are cases for which the "optimal" weights are identical to (3), as e.g. under simple random sampling.
Using quotation marks around *optimal* is indeed deliberate, but under full response *optimal* has a very clear meaning. As mentioned earlier, the optimal regression estimator has asymptotically minimum variance among linear regression estimators. Adding nonresponse where the nonresponse mechanism is at least partially unknown, makes it difficult (impossible?) to define optimality criteria in a proper way.
In the following we will focus on a sampling design where generally $\hat{t}_{y\,cal} \neq \hat{t}_{y\,opt}$, namely Poisson sampling. The independence of drawings simplifies the "optimal" distance measure:

$$\sum_r \frac{\pi_k^2}{1-\pi_k}(w_{kr}-d_k)^2 = \sum_r \frac{(w_{kr}-d_k)^2}{d_k(d_k-1)}$$

For this measure it might be fruitful to replace d_k with d_k^*. where we include in d_k^* the reciprocal of an estimate of the average response probability $\bar{\theta} = \sum_U \theta_k/N$. One simple candidate is $\hat{\bar{\theta}} = n_r/n_s$, thus yielding $d_k^* = d_k \cdot (n_s/n_r)$. Another natural choice is $\hat{\bar{\theta}} = \sum_r d_k / \sum_s d_k$, since $E(\sum_s d_k) = N$ and $E(\sum_r d_k) = \sum_U \theta_k = N\bar{\theta}$, which lead to $E(\sum_r d_k / \sum_s d_k) \approx \bar{\theta}$.
In a following simulation study we will examine these two examples of alternative weightings, where the possible effects on the bias of the resulting calibration estimators are of special interest.

References

Andersson, P.G. and Thorburn, D. (2005). An optimal calibration distance leading to the optimal regression estimator. *Survey Methodology*, 31(1), 95-99.

Andersson, P.G. and Särndal, C-E. (2016). Calibration for nonresponse treatment: In one or two steps? *Statistical Journal of the IAOS*, 32, 375-381.

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

Lundström, S. (1997). *Calibration as a standard method for treatment of nonresponse.* Ph.D. thesis, Department of Statistics, Stockholm University.

Lundström, S. and Särndal, C-E. (1999). Calibration as a standard method for treatment of nonresponse. *Journal of Official Statistics*, 13, 305-327.

Montanari, G.E. (1998). On regression estimation of finite population means. *Survey Methodology*, 23, 69-77.

Rao, J.N.K. (1994). Estimating totals and distribution functions using auxiliary information at the estimation stage. *Journal of Official Statistics*, 10, 153-165.

Särndal, C-E, and Lundström, S. (2005). *Estimation in surveys with nonresponse.* Chichester, UK: Wiley.