Calibration Weighting for Nonresponse with a Flawed but Survey-Corrected Frame Variable

Phillip S. Kott[1]

Abstract
Sometimes in survey sampling we have access to a frame variable that is imperfectly measured. For example, the frame may contain an imperfect indicator of whether a housing unit is owned or rented. Although (we will assume) the error in this variable can be corrected on the survey itself, using a corrected-frame values as a calibration variable will generally bias the resulting estimates. We will show how to avoid that source of bias when adjusting for unit nonresponse through calibration weighting. This can be done by treating the flawed-frame variable as a shadow variable to the corrected-frame variable in the weight-adjustment function. In other words, by calibrating on the flawed version of the variable while assuming, more reasonably, that whether or not a sample unit responds is a function of the corrected version. Since only the respondents are reweighted, we only needed to have corrected versions of the respondents' values in the weight-adjustment function. Some simple simulations will show the effectiveness this weighting approach.

Keywords: Model variable, Shadow variable, Weight-adjustment factor, Bias.

1. Introduction

Sometimes in survey sampling we have access to a frame variable that is imperfectly measured. For example, the frame may contain an indicator of whether a housing unit is owned or rented that is not always correct. Although the error in this variable can be corrected on the survey itself (which we assume here is always provides correct values)), using a corrected-frame values as a calibration variable will almost always bias the resulting estimates because only population units responding to the survey can have their values corrected, not units outside the respondent sample.

Now suppose unit response is a function of the true value of a variable whose sometimes flawed counterpart is on the frame. Traditional calibration weighing using either the frame-variable values as they are or the partial-corrected frame values (i.e., corrected by the survey for respondents) will result in biased estimates.

It is possible, however, to include the true values of the variable recorded in the survey among the weight-adjustment model variables while the frame values of the variable are treated as shadow variables, that is, calibration variables that are not part of the response model. This can be done with the SUDAAN 11 calibration-weighting procedure WTADJX (RTI 2012) as we shall see.

In Section 2, we provide the mathematics behind this procedure. Section 3 demonstrates its use with a simplistic but enlightening example. Section 4 contains some concluding remarks.

2. The Mathematics

Let k denote an element of the population U, and R the respondent sample. We want to estimate a total $T_y = \sum_U y_k$ using a vector of calibration variables $\mathbf{z}_k$, with known population totals: $\boldsymbol{T}_\mathbf{z} = \sum_U \mathbf{z}_k$. Given sampling weights d_k and probabilities of unit response of the assumed form:

[1] RTI International, 6110 Executive Blvd #902, Rockville, MD 20852

$p_k = 1/(1 + \exp(\mathbf{x}_k{}^T\boldsymbol{\gamma}))$,

where $\mathbf{x}_k$ is a vector of covariates of the same dimension as $\mathbf{z}_k$ and $\boldsymbol{\gamma}$ is unknown, then a consistent estimator for T_y is

$$t_y = \sum_R d_k[1 + \exp(\mathbf{x}_k^T\mathbf{g})]y_k,$$

where $\mathbf{g}$ is an estimate for $\boldsymbol{\gamma}$ found by solving the calibration equation**:**

$$\boldsymbol{T}_\mathbf{z} = \sum_R d_k(1 + \exp(\mathbf{x}_k^T\mathbf{g}))\mathbf{z}_k.$$

Such a solution often exists because there are as many unknowns (in $\mathbf{g}$) as calibration equations (components of $\boldsymbol{T}_\mathbf{z}$). Note that the components of $\mathbf{x}_k$ need only to known for respondents. Extension to $\mathbf{x}_k$ to vectors with less components than $\mathbf{z}_k$ are possible. See Kott (2014) for more details.

3. A Toy Example

We demonstrate how this can work with a simplistic example, in which we will estimate a population mean rather than a total. When unity is contained in both the x_k and z_k vectors (or the equivalent; *i.e*., some linear combination of the components of each vector is 1) estimating a population mean is analogous to estimating a population total.

We start with a simple random sample (with replacement) of size 1000. Each unit (k = 1, …, 1,000) is assigned a random variable r between 0 to 1. When $k > 500$ then the true value $x_k = 0$; otherwise $x_k = 1$. When $k > 300$ then the frame values $z_k = 0$; otherwise $z_k = 1$. As a result, there is some, but not perfect correlation between z and x.

If $x_k = 1$ and $r_k \leq .6$ then unit k responds; otherwise it doesn't. If $x_k = 1$ and $r_k \leq .9$ then unit k responds; otherwise it doesn't. We can only observe x_k when unit k responds (in fact, we known what all the x_k are; we only pretend some units don't respond and their x-values are not observed). Clearly, whether or not x_k is observed depends on its value. Nonresponse is *not* missing at random. Nevertheless, we want to estimate the population mean of x_k from the respondent ("observed") sample.

We will compare three estimation techniques based on one generated respondent sample. The SAS-callable SUDAAN code we use for this appears in the appendix. In the first method, we reweight the respondent sample so that the mean of the z_k in the respondent sample equals that in the whole sample. Although we use SAS-callable SUDAAN, this is a simple reweighting exercise that could be done by hand. In particular, the weight-adjustment factor for a responding unit j is the number of units in the whole sample with the same z-value as unit j divided by the number of units with the same z-value in the respondent sample. The weight-adjustment factor is multiplied by the design weight (here N/n) to produce the calibration (adjusted) weight.

The second method is the same as the first, except that the z-values are replaced by partially-corrected c-values, defined as equal to z_k when unit k in not a respondent observed and as x_k otherwise. That is to say, we calibrate on our best guess at the x-value: x_k when it is known, z_k otherwise. This is well known to produce biased estimates, which we will confirm.

In the third method, we again reweight so the mean of the z_k in respondent sample equals that in the whole sample, but we do it in such a way that that units with the same x-value in the respondent sample get the same weight adjustment factor. This can be done with matrix algebra in this simple example, but we use the WTADJX procedure in SUDAAN 11 employing the code in the appendix.

As the three tables below show, the first two method produce badly biased estimates in that their errors are larger than twice their estimated standard errors computed (by the WTADJUST routine in SUDAAN 11) under the erroneous assumption that the weighting procedure is unbiased.

By contrast, the error in the estimate using the third method is well within one standard error. It appears to be asymptotically unbiased, as theory predicts. The realized response rate in the simulation for the 1,000-unit sample was (approximately) 91.8% when $x_k = 0$ and 59.0% when $x_k = 1$. The weight-adjustment factor for the third method was 1.077 when $x_k = 0$ and 1.741 when $x_k = 1$. Ideal would have been 1/.918 = 1.089 and 1/.590 = 1.695. Even though these factors do not appear that close to ideal, they produce an asymptotically unbiased estimator for the mean of the *x*-values.

Table 1. Calibrating and Response Modeling on the Frame Values (*z*)

Variable		
x	Mean	0.4451
	SE Mean	0.0165

Table 2. Calibrating and Modeling on the Partially-Corrected Frame Values (*c*)

Variable		
x	Mean	0.4200
	SE Mean	0.0156

Table 3. Calibrating on the frame values (*z*)**, but modeling on the survey values** (*x*)

Variable		
x	Mean	0.5057
	SE Mean	0.0223

4. Concluding Remarks

- If we weaken the correlations between *z* and *x*, the standard errors (as estimated by WTADJX) increase.
- This toy example demonstrates the usefulness of letting a variables in the weight-adjustment model differ from related calibration variable.
- In actual practice, the survey-corrected frame variable will be only one of many variables in the model and calibration equations, and it need not be binary.
- Moreover, the target of estimation will not likely be the total for the corrected frame value, although the correct frame variable may be a predictor of the target variable.
- If the corrected frame value is not correlated with the target variable, then the nonresponse depending on it will not be a potential source of bias.

- When the components are x_k are 0/1 membership indicators of mutually exclusive groups, the functional form the response function $p(.)$ doesn't matter except when restrictions on the range of the function prevents calibration for some forms but not others (e.g., $p(\mathbf{x}_k^T\boldsymbol{\gamma}) = 1/(1 + \exp(\mathbf{x}_k^T\boldsymbol{\gamma}))$ cannot be less than 1 but $p(\mathbf{x}_k^T\boldsymbol{\gamma}) = 1/(1 + \mathbf{x}_k^T\boldsymbol{\gamma})$ can). See Kott (2014).
- Often we do not know if nonresponse is ignorable or nonignorable, but with survey-correctible frame variables we do know. It depends on the true variable values in x rather than their proxy values in z.
- There are procedures in R, such as 'Sampling' (Tille and Matei 2013), that can be used in place of WTADJX when more than matrix algebra is needed.

References

Kott, P. S. (2014). Calibration weighting when model and calibration variables can differ. In *Contribution to Sampling Statistics: ITACOSM 2013 Selected Papers* (pp. 1–18). Heidelberg, Germany: Springer.

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

Tille, Y. and Matei, A., (2013), *Package* '*Sampling.*' A software routine available online at http://cran.r-project.org/web/packages/sampling/sampling.pdf.

Appendix: SAS-callable SUDAAN Code for the Three Methods

```
* The data set is named D.

/* Calibrating on the frame values (z) */

PROC WTADJUST  DATA = D ADJUST = NONRESPONSE DESIGN = SRS;
VAR X;
MODEL RESPONSE = Z; * an intercept term is implicit;
PRINT MEAN SE_MEAN  ;
OUTPUT ADJFACTOR/FILENAME =OUT1 REPLACE;

/*  Calibrating on the partially-corrected frame values (c) */

PROC WTADJUST DATA = D ADJUST = NONRESPONSE DESIGN = SRS;
VAR X;
MODEL RESPONSE = C; * an intercept term is implicit;
PRINT MEAN SE_MEAN  ;
OUTPUT ADJFACTOR/FILENAME =OUT2 REPLACE;

/* Calibrating on the frame values (z), but modeling on the survey values (x) */

PROC WTADJX  DATA = D ADJUST = NONRESPONSE DESIGN = SRS;
VAR X;
MODEL RESPONSE = X; *  an intercept term is implicit;
CALVARS Z;
PRINT MEAN SE_MEAN ;
OUTPUT ADJFACTOR/FILENAME =OUT3 REPLACE;
```