Calibration after response propensity weighting

Laaksonen, Seppo
University of Helsinki, Finland,
Correspondence: Seppo.Laaksonen@Helsinki.Fi
17 August 2015

Keywords: Reweighting, simulation study, macro vs micro auxiliary variables, empirical examples

1. The main target

Reweighting of survey data of the respondents is required for several reasons, especially due to problems in frame coverage and frame data quality on one hand, and due to selective unit non-response, on the other.

This paper tries to give a good strategy to solve both problems. In principle, this strategy can work well given that appropriate macro and micro auxiliary data are available and used exhaustively. Unfortunately, this is not usually the case but it is unfortunate that survey experts do not pay enough attention to get such data.

I understand well that appropriate micro/macro auxiliary variables are very difficult to get in many cases. Some examples:

- Happiness results in all surveys are too positive since unhappier people have not been contacted or they do not reply.
- We know that politically, socially etc. active people are well participating but how to adjust the weights it is not solved although some type of proxy variables are attempted.
- Poverty indicators using objective incomes are not believable since poorest people are much missing even from the sampling frame or they are not responding. Item non-response in subjective income is much lower but unit non-response is still there.

Our reweighting approach consists of the three steps.

- First, the basic sampling weights are computed using the sampling design of the survey and assuming that the response mechanism is ignorable.
- Then the response propensity weights are constructed and
- Finally, these weights are used as initial weights in calibration.

Thus both micro and macro auxiliary variables may be fully exploited. We compare the different methods with each other using a simulated data set that is based on a modification of a real data set. This gives opportunity to see how well each method works. We also compare variations of such calibration methods that are invented in the Calmar 2.

2. Calibration, response propensity weighting and their joint method

A success story of calibration methods begun early 1990's when the article of Deville and Särndal (1992) was published and especially when the first Calmar SAS macro was coded by the French statistical office INSEE in 1993 (Sautory 2003). The new version, Calmar 2, developed in 2003, offers the user new resources for performing calibrations and implements the generalized calibration method of handling non-response.

Calmar 2 thus is a macro into which a user can choose the three types of options:

(i) the initial weight d_k desired to calibrate,
(ii) the calibration margins that are expected to be as true population totals as possible,
(iii) the calibration methods with alternative distance functions.

The original version of Calmar offered four calibration methods, corresponding to four different distance functions. These methods are characterized by the form of function as follows:
• the *linear* method: the calibrated estimator is the generalized regression estimator:
• the *exponential* method: where all the calibration variables are qualitative, this is the raking ratio method;
• the *logistic* method: this method provides lower limits L and upper limits U on the weight ratios w_k/d_k *in which w* refers to the calibrated weight;
• the *truncated linear* method, very similar to the logit method.
• the sinus hyperbolicus method that is exponential and does not to give negative or other implausible weights like the exponential or raking ratio method.

We present results from the three methods, that is, linear, logistic and sinus hyperbolicus; this done both concerning pure calibration and joint method.

The strategy for creating 'response propensity weights' is as follows:

(i) We have the basic weights d_k for each respondent *k*.
(ii) The response propensities p_k using logistic (or probit) regression are estimated and their quality is checked.
(iii) We take the basic weights and divide these by the estimated response probabilities of each respondent getting the preliminary weights.
(iv) Since the sum of the weights (iii) does not match to the known population statistics by strata *h*, they should be calibrated so that the sums are equal to the sums of the basic weights in each stratum. This is made by multiplying the weights (iii) by the ratio $q_h = \frac{\sum_h d_k}{\sum_h d_k / p_k}$. This is one option for the response propensity modelling weighting, called 'pure' in Figure 1.

2. Data and simulation principles

The data for simulations is created from the 2010 Finnish Security Survey (Laaksonen & Heiskanen 2014) so that its three independent data sets (face-to-face, phone and web) of the respondents are first pooled together and then extracted to a population of 179985 persons. This extraction started from about 3 000 respondents. It was rather straightforward but some randomness was added. The creation of missingness is more demanding. We followed as well as possible the initial unit nonresponse and hence the response rate of our survey is about equal with it, that is, 49 per cent. The simulation data set thus consists of a missingness indicator for each target population unit, and it is respectively copied for each simulation sample.

The simulation strategy, naturally, follows the survey principles:

(i) Four explicit strata by four large regions were formed.
(ii) Simple random sample with unequal allocation was drawn from each stratum, altogether 2000 individuals
(iii) Basic sample weights were computed for the respondents as usually, assuming ignorable unit non-response.
(iv) The three different calibrated weights were computed using Calmar 2. The margin variables are here: four strata, two genders, five age groups.
(v) Response propensity based weights were respectively calculated. These weights include some calibration as well. We included basically all available micro auxiliary variables in the respective model (interaction of gender and age group, stratum, education, number of rooms, married, citizen, mother tongue and children). The mean estimates were calculated by these different weights.
(vi) The procedure from (ii) to (v) was repeated enough many times and the output data set obtained.
(vii) The results between simulated results and true values were compared.

4. Summary of the results

Table 1 shows the averages of all indicators in the population. In simulations, we thus try to get the estimates that are as close these values as possible. It is not necessarily easy for many reasons. One special reason is that there are a fairly small number of observations for some indicators. This is indicated in the right column. This missingness is not only due to nonresponse. it is mainly due to the topic itself. For example, if a person never had a partner, the answer is empty. The reason for some missingness is not exactly known, such as violence by stranger recently or harassment ever.

Table 1. Statistics about the population of the 15-79 years old 179 985 persons in simulations

Indicator in simulations	Average in population	Response rate, %
Income	44905€	100
Worry (about crime)	28%	100
Harassment recently	43%	99
Harassment ever	74%	24
Violence by stranger recently	33%	24
Violence by stranger ever	87%	41
Violence by partner	16%	74
Violence by ex-partner	30%	45

We drew 150 samples from this simulation population using maybe the most common design, stratified random sampling. The sample allocation had not any specific target and hence it is not far from proportionality, the average sampling fractions being 6.2%, 3.7%, 5.3% and 7.8%.

The relative deviation from the true value is the most illustrative way to compare results since it is not needed to look at the indicator values themselves. These comparisons are presented in the eight graphs of Figure 1.

Figure 1. The relative bias in per cent for each of 8 indicators by eight weighting methods

The differences between the three calibration estimates are minor. This is concerned the pure calibration methods CAL!, CAL3 and CAL5 on one hand, and the same methods after the response propensity weighting on the other.

Almost all weights with adjustments improve the estimates to some extent. The study also shows that the combination of the response propensity weighting and calibration is a superior method to pure calibration. Nevertheless, it is not best in each case. It is even so that the basic weights are best in one case (violence by ex-partner). The two reasons behind this are obvious: a small number of respondents and non-good auxiliary variables. Calibration is best only in one case (violence by stranger ever). Surprisingly, the results are for this indicator worsening when calibrating after response propensity weights. We thus see that any weighting method does not work ideally in each case.

References

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

Deville. J-C., Särndal. C-E. & Sautory. O (1993). Generalized Raking Procedures in Survey Sampling. *Journal of the American Statistical Association*. 1013-1020.

Ekholm. A. & Laaksonen. S. (1991). Weighting via Response Modelling in the Finnish Household Budget Survey. *Journal of Official Statistics*. 7.2. 325-337

ESS _ European Social Survey (2014). Documentation of ESS Post-Stratification Weights. http://www.europeansocialsurvey.org/docs/methodology/ESS_post_stratification_weights_documentation.pdf

Holt. D. & Smith. T.M.F. (1979). Post-Stratification. Journal of the Royal Statistical Society. Series A (General). Vol. 142. 33-46.

Kott, P.S. (2006). Using Calibration Weighting to Adjust for Nonresponse and Coverage Errors. *Survey Methodology* 32, 2, 133-142

Lumley, T. (2013). Survey: Analysis of Complex Survey Samples. R package version 3.29. URL http://CRAN.R-project.org/package=survey

Lumley, T., Shaw, P. & Dai, J.Y. (2011). Connections between Survey Calibration Estimators and Semiparametric Models for Incomplete Data. *International Statistical Review* 79, 2, 200–220, doi:10.1111/j.1751-5823.2011.00138.x

Laaksonen. S. (2007). Weighting for Two-Phase Surveyed Data. *Survey Methodology.* December Vol. 33. No. 2. pp. 121-130. Statistics Canada.

Laaksonen, S. (2015). Sampling Design Data File. *Survey Statistician* 72, 61-66.

Laaksonen, S. & Heiskanen, M. (2014). Comparison of Three modes for a Crime Victimization Survey. Journal of Survey Statistics and Methodology, *2 (4): 459-483 doi:10.1093/jssam/smu018*

Laaksonen, S., Kemppainen, T., Stjernberg, M., Kortteinen, M., Vaattovaara, M. and Lönnqvist, H. (2015). Tackling City-Regional Dynamics in a Survey Using Grid Sampling. *Survey. Research Methods by the European Survey Research Association, Vol 9, No 1, 55-65.* *www.surveyresearch.org*

Le Guenne. J. & Sautory. O. (2005). CALMAR 2 : Une Nouvelle Version de la Macro Calma de Redressment D'Échantillion Par Calage.
http://vserver-insee.nexen.net/jms2005/site/files/documents/2005/327_1-JMS2002_SESSION1_LE-GUENNEC-SAUTORY_CALMAR-2_ACTES.PDF

Little. R.J.A. (1986). Survey Nonresponse Adjustments for Estimates of Means. *International Statistical Review.* 54. 139-157.

Robins, J.M., Rotnitzky, A. & Zhao, L.-P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association* 89, 846–866.

Sautory, O. (2003). CALMAR 2: A NEW VERSION OF THE CALMAR CALIBRATION ADJUSTMENT PROGRAM. Proceedings of Statistics Canada's Symposium: Challenges in Survey Taking for the Next Decade

Särndal, C-E. (2007). Calibration Approach in Survey Theory and Practice*. Survey Methodology,* 33, No. 2, pp. 99-119