Calibration versus other reweighting methods in surveys

Seppo Laaksonen
University of Helsinki, Finland, email: Seppo.Laaksonen@Helsinki.Fi

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. Over years, a number of methods have been proposed and used. An appropriate methodology requires necessarily auxiliary data. The more and predicatbe auxiliary data are available more options to reweight initial sampling weights exist. Basically, two types of such variables can be tried, either aggregate or micro. Typically, aggregate variables are margins of the target population frame. Consequently, such information gives opportunity to use calibration methods. On the other hand, if micro variables are available, response propensity methodology is possible, as well. This paper compares these two approaches, both theoretically and using simulated data.

Keywords: Reweighting, response-propensity methodology, macro vs micro auxiliary variables, empirical examples

1. Introduction

Nonresponse and coverage problems are common in surveys. Both problems are worsening rather than decreasing. Unless the fieldwork is successful or special data collection modes are found and used, post-survey adjustments are the only option to try for improving the data quality. We here concentrate on weighting adjustments. Without appropriate auxiliary data reweighting is useless. That is, we cannot do much. Auxiliary variables we test are of two types, (i) aggregate or macro and (ii) micro.

A number of methods have been proposed and used for reweighting. Since population statistics or other aggregate data are most often available, the methods that can get benefit on such data were first started to apply. Post-stratification was earlier the most common methodology to adjust for coverage and unit non-response errors (Holt and Smith 1979). This can also called as 'a basic calibration method.' This methodology was widened by Deville and Särndal (1992), in particular, leading to a more general approach so that several margins can be used to benchmark the reweights as precisely as the recent known population figures imply. When these two researchers together with Sautory (1993) invented a SAS macro Calmar, the prosperity of this methodology was ready to begin. Later, a second version of the SAS macro, Calmar 2, was published, giving also new options for calibration (Le Quennec and Sautory 2005). This macro gives opportunity to insert the margins of two levels even (e.g. households and household members).

Another approach to reweighting is to exploit micro level auxiliary variables as well as possible. The basic ideas of this methodology are mainly from the 1980's (e.g. Little 1986). Laaksonen (2007) applied this technique so that he first estimated a logistic regression model for predicting the response propensities to the individual respondents, not for the groups (often called homogeneity groups) of the respondents as he made much earlier (Ekholm and Laaksonen 1991). In the second stage, he divides the initial sampling weights with these propensities and in the end he benchmarks these preliminary weights to correspond to the known population of the explicit strata. This was done since the initial sampling design was the explicit stratification and these population figures were available in the beginning, before the fieldwork was started.

The success of this methodology much depends on the richness of the micro-level auxiliary variables.

This paper combines both approaches, that is, calibration methods and response propensity modeling weighting, respectively. On other hand, we compare these with each other using a simulated data set that is based on the extension of the real data set. This gives opportunity to see how well each method works. We also compare different variations of such calibration methods that are invented in the Calmar 2.

In Section 2, we describe our simulated data set. Section 3 presents the key principles of the Calmar 2 as well as our response propensity modeling. Section 4 summaries our empirical findings, and the final section presents some conclusions.

2. Data and simulation principles

The basis for the simulated data set is the Finnish Security Survey 2010 (Aromaa et al 2010). In this presentation we present results only on one variable, that is, yearly income of 15-74 years old people in Finland. Naturally, the data consists of many other variables, relating crime victimisation, in particular, and these also are interesting, but we do not present these in details. Some general views are given, nevertheless.

The sample data set was extended from about 3 000 respondents to the target population data set with 180 000 people. This extension was rather straightforward but some randomness was added to income values, among others. This does not lead to any critical things, but how to create missingness, it may be considered to be partially critical. We followed as well as possible the initial unit nonresponse and hence the response rate of our survey is about equal in it, that is, 49 per cent. The missingness itself is much randomized but it has a similar feature as in the initial survey. This was created by a person who was going to make simulations. There is thus a missingness indicator for each target population unit. The simulation strategy, naturally, follows the survey principles:

(i) Four explicit strata by four large regions was 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, dividing the target population by the number of the respondents (assuming ignorable unit non-response).

(iv) The three different calibrated weights were computed using Calmar 2. The margin variables were here applied: four strata, two genders, five age groups. These variables are quite easily available in Finland as well as in many countries. In principle, we could add more margins but this is not realistic since it is possible in few cases in practice only, and would require more resources.

(v) Response propensity based weights were respectively calculated. These weights include some calibration as well..

(vi) The mean estimates were calculated by these different weights.

(vii) The procedure from (ii) to (vi) was repeated enough many times and the output data set obtained.

(viii) The results between simulated results and true values were compared.

3. Details of calibration and response propensity weighting

Calmar 2 is a bit old macro but working well. It offers the five calibration options (Le Quennec and Sautory 2005):

- Linear
- Raking-ratio
- Logistic method
- Truncated linear method
- Sinus hyperbolicus method.

In this paper we present results from the three methods, that is, linear, logistic and sinus hyperbolicus. It is well known that the linear method may give negative weights that are not acceptable. We organised our rules for sampling and calibration so that we avoided this nuisance. The same problem does not appear in other techniques applied. Naturally, the truncated linear method could help if negative weights applied, but what are the correct limits for the truncation, it is self-evident. A good solution for these limits could be to allow the minimum reduction from the basic weights as 50 per cent, and respectively for the maximum increase 50 per cent as well. These limits thus are symmetric, in relative measures.

The strategy for creating 'advanced sampling reweights' is as follows:

(i) We have the gross sample design weights that are the inverses of the inclusion probabilities. These inclusion probabilities vary by strata but are equal within each stratum.

(ii) We assume that the response mechanism within each stratum is ignorable, and hence compute the initial (basic) weights analogously to the weights (i). These are available only for the respondent k, and symbolised by w_k.

(iii) Next we take those initial weights and divide these by the estimated response probabilities (called also response propensities) of each respondent obtained from the probit (logit link gives quite similar results) model, and symbolised by p_k.

(iv) Before going forward, it is good to check that the probabilities p_k are realistic, that is, they are not too small, for instance. All probabilities are below 1, naturally.

(v) 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 initial weights in each stratum. This is made by multiplying the weights (iii) by the ratio $q_h = \frac{\sum_h w_k}{\sum_h w_k / p_k}$ (Laaksonen 2007). This is one option for the response propensity modelling weighting, called 'pure' in Table1. In this study, we however test the three other alternatives, that is, the same ones that are in calibration. Note that in this case, the initial weights before calibration are just the weights of the first option, not the basic weights.

(vi) It is good also to check these weights against basic statistics, such as the mean, the maximum, the minimum and the coefficient of variation. This was here made for the first sample and as soon as this was considered to give plausible results, the next repetitions were performed in the same way.

We used much effort in our initial study to gather as many auxiliary variables as possible. So, we had the same chances to use these in our simulations as well. So, our final response propensity model consists of the following explanatory variables: interaction of gender and age group, education level, stratum, partnership including different categories for the duration of the current marriage, and a category for second or more marriages, children or not at home, unemployed or not, mother tongue, number of rooms, if living in the municipality born or not. All these are not in all samples very significant but a good point in response propensity based adjustments is that an insignificancy does not violate the adjusted weights but its impact on an estimate is less remarkable in such case. It thus maybe always improves the estimates at all.

.

4. Summary of the results

We present our main results for income in Table 1.

Table 1. Estimates after 100 simulations by different weights

	Estimate	Bias	Standard error
TRUE	44919.8	0.0	0.0
Basic weight	45802.7	883.0	55.7
Calibration			
Linear	45656.8	737.0	55.9
Logistic	45667.0	747.3	55.9
Sinus hyperbolicus	45667.8	748.0	56.0
Response propensity			
Pure	45039.6	119.9	51.9
And Calibration			
Linear	44997.9	78.1	52.0
Logistic	44997.9	78.1	52.0
Sinus hyperbolicus	44997.9	78.2	52.0

The results are very clear. All weights with adjustments improve the estimates to some extent, but all are upward biased. The second conclusion is that our three calibrations give almost equal estimates. It is clear that a more complex case than ours here may give more variation. The third key finding is that all response propensity based methods are best, even so that the ‘pure’ case is better than any ordinary calibration. Respective, we find that these different calibrations give some advantage for response propensity methods. This has not been tested before as far as we know. The standard errors after 500 simulations are relatively high but these can be declined if more simulations are performed. The basic findings have been about similar even after 100 simulations.

5. Conclusion

Our study clearly shows that the combination of the response propensity weighting and calibration is a superior method to pure calibration, for instance. Of course, without a rich pattern of auxiliary variables, it is difficult to succeed. We tested several other estimates in the same way. Our overall outcome remains the same but there is less differences in some violence estimates that in this income estimate. We even found a case in which we cannot say which method is best. This means that it is hard to find auxiliary variables that are well correlated to such an estimate. This is not unusual.

References

Aromaa. K.. Heiskanen. M.. Laaksonen. S.. Nikula. J. & Ruuskanen. E. (2010): *Finnish results from the EU pilot survey*. Final report delivered to the European Commission on 22 February 2010. 71 pp. Not published. Available from the authors.

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

Deville. J-C.. Sarndal. 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

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

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

Laaksonen. S. (2008). Retrospective Two-Stage Cluster Sampling for Mortality in Iraq. *International Journal of Market Research* 50. 3. 403-417

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.