Sub-Sample, Non-Response and Weighting

Seppo Laaksonen
Statistics Finland
00022- Statistics Finland, Finland
Seppo.Laaksonen@Stat.Fi

1. Introduction

The auxiliary variables of a survey may be derived from various sources (see e.g. Laaksonen 1999), but for weighting purposes these are usually taken from registers, other administrative sources and surveys. This sort of auxiliary variables may be called *external*, if we want to distinguish these of *internal* auxiliary variables which are derived from the same survey.

Internal auxiliary variables are particularly used for imputations when some item values are missing. These also are much used in panel surveys if a certain respondent responds in one wave, but not in the other. In this panel survey case, internal auxiliary information may be used both for weighting adjustments and for imputations.

This paper does not deal with the standard survey as described above. We have the two special characteristics:
(i) The survey consists of the two steps. In the first example, the second step is directed to the *non-respondents* of the initial survey so that *a sub-sample* of these has been picked up, and these have been interviewed using a questionnaire consisting of a couple of the crucial questions. In the second example, respectively, a *sub-sample* consists of such *respondents* of the initial survey who are willing to contribute to a more detailed survey.
(ii) In the both examples, when making attempts for post-survey adjustments, we may exploit both external and internal auxiliary variables.

2. Methods

Example 1 is from an innovation survey with an ordinary sampling design, that is, pre-stratification by type of industry and size is used, and the sampled units are drawn randomly within pre-strata. Because the unit non-response rate was high, a sub-sample of non-respondents was taken. These replied excellently because the only one crucial question was asked: '*have you done any innovations or have you used money for innovative activities during the last 2 years*?'

We first created the standard design weights for the whole sample. These weights were used as the *first basic weights* for the sub-set of the respondents. For the sub-sample of the non-respondents we created the analogous sample weights, called *sampling weights for non-respondents*, so that the sums of these weights tallied in each pre-stratum with the sums of the design weights. When estimating figures from the sample covering the units answered either to the full or limited questionnaire, we thus used both of these weights, called *sampling weights for extended sample*. At contrast, when estimating the figures from the sample answering to the full questionnaire we needed the *second basic sampling weights* which were based on the similar formats but given that the response mechanism is ignorable within each pre-stratum.

If, however, there is a tendency that the response mechanism is not ignorable, for which some evidence may be found from the sub-sample, we have to look forward to developing the adjusted weights. In this case, we tested the two techniques, (a) calibration, and (b) the technique which first exploits logistic regression so that the data cover the extended sample. The response variable is such which = 1 if a unit responded either to the full survey or the limited survey, and = 0, otherwise. For this model all external crucial auxiliary variables may be exploited as explanatory variables in a best way, including the same ones as for calibration. In addition, the internal auxiliary variables may and should be used, in this case, thus picking up from the sub-sample. This is not enough, it is necessary to *use the sampling weights of the extended sample* in this modelling. They gave the estimated response propensities by which the second basic sampling weights were divided. The last operation was to scale these weights by calibration at pre-stratum level.

Example 2 is from a special survey done for Finnish citizens older than 15 years. The topic of the survey is concerned their leisure time activities (in the nature and other outdoor activities). First, a CAPI survey was conducted, covering various leisure time and hobby questions. In the end of this survey was asked the willingness of each respondent to receive a special postal survey questionnaire in which more detailed questions would be presented. This example is a bit more complex than the previous one, because it consists of the three steps:
(i) First, the well-designed *post-stratified sampling weights* for the first set of respondents were created. The post-strata were based on region, gender, age group and participation season.
(ii) Then, we used the set of volunteers from the first group, and built a *logistic regression* model using the weights from step (i). We were able to use both external and internal auxiliary variables. The internal ones were here taken from the first questionnaire. The response propensities derived from this model were next included in the *second basic weights* by multiplying with their inverses, and then the weights were calibrated in the same manner as in Example 1, at post-stratum level.
(iii) The final step was similar to step (ii) but now a logistic regression model was built to the sub-set of volunteers, thus the response variable is = 1 if a person replied to this second survey, and = 0 if did not. The weights used here were those created in the previous step, step (ii).

As the result, we have the *adjusted sampling weights* for the persons who replied to the both surveys. These weights may thus be used when estimating the maximum number of variables, including the variables of the first survey. On the other hand, when analysing the first survey only, we can use the basic (post-stratified) weights. Our target is however, that some and rather many figures would be reasonably similar obtained from each of these two surveys. At least, a user would be happy with such a situation, since the results would not be contradictory. We go to empirical results and look forward how well we have succeeded.

3. Some results

Example 1

The predicted response probabilities have been used when constructing the new weights and then estimating results of the innovation survey. The results show how selective non-response is, and consequently, that the non-adjusted estimates are very biased. The both relevant adjustments exploiting the sub-sample of the non-respondents give the slightly too low estimates. However, the estimate based on the weighted response propensity modelling lies within the 95% confidence interval, whereas this based on the calibration method lies outside this interval. This suggests to prefer the previous method. On the other hand, the results clearly show an overestimate when using the unweighted response propensity modelling.

Example 2

There are found many interesting results from these three successive behaviour logistic regression models. We exclude these details from this report. The models have been exploited in our estimations, and the results clearly show how biased some figures may be in a sub-group if any adjustments exploiting a higher level survey have not been done. Results will be improved while more internal auxiliary variables are used. We however cannot say so far what restrictions will be occurred if ‘ too many such variables have been used, or these have not been used correctly.’

REFERENCE

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

Laaksonen, S. (1999). Weighting and Auxiliary Variables in Sample Surveys. In: G. Brossier and A-M. Dussaix (eds). 'Enquêtes et Sondages. Méthodes, modèles, applications, nouvelles approches," 168-180. Dunod. Paris.

RESUME

Nous présentons deux cas différents où un sous-échantillon a été exploité pour ajuster les poids, un sous-échantillon est tiré soit (i) des non-répondants soit (ii) des répondants.