Estimation in the Presence of Nonresponse and Other Survey Imperfections – A Handbook at Statistics Sweden

Sixten Lundström
Statistics Sweden
Klostergatan 23
Örebro, Sweden

This paper describes a document that is one in a series of Current Best Methods (CBM) manuals produced in recent years at Statistics Sweden. Their objective is to present in easily accessible form those techniques that are viewed as "best" for a given aspect of the statistical production process. They are intended as guides for survey statisticians in survey design, redesign and maintenance. Although produced mainly for statisticians at Statistics Sweden, they can also provide useful information for many other readers. The objective of this CBM, Lundström and Särndal (2001), is to give an up-to-date account of methods of estimation for use when data collection has been "disturbed" by nonresponse and frame imperfections.

Nonresponse adjustment is not treated as an isolated issue. We integrate nonresponse adjustment into the broader context of estimation. The issue is how to make the best possible estimates based on the data collected from those who respond to the survey, and on any relevant auxiliary information that may be available about the population and its elements, whether respondents or nonrespondents. Frame imperfections are also discussed, though we do not provide an exhaustive treatment of this difficult problem on which the literature has remained comparatively silent.

The principal methods for nonresponse adjustment are *reweighting* and *imputation*. In this CBM, reweighting is treated by a general approach - the *calibration approach* - which has the favourable property of incorporating most "standard" methods found in different places in the literature. Imputation entails replacing missing values by proxy values. Different imputation methods, and the "imputed estimators" that they lead to, are discussed.

All methods discussed rely on the use of auxiliary information. The "stronger" the auxiliary information is, the smaller the errors in the estimates will be. The procedure for selecting such information is based on a general expression for the nonresponse bias of the calibration estimator, results from simulation studies, the literature, etc. The CBM states that the following general recommendations should be followed when selecting auxiliary information.

(I) the auxiliary information should, as far as possible, explain the variation of the response probabilities;

(ii) the auxiliary information should, as far as possible, explain the variation in the most important study variables; and

(iii) the auxiliary information should identify as closely as possible the (principal) domains of interest.

When principle (i) is fulfilled the nonresponse bias is reduced for all estimates. However, if only principle (ii) is fulfilled the nonresponse bias is reduced only for the estimates based on the main study variables. In the latter case the variance will also decrease. When principle (iii) is fulfilled the effect is mainly a reduction of the variance for the domain estimates.

The CBM also discusses how to estimate the variance of the calibration estimator and the imputed estimator. In particular, we comment on the use of CLAN97, a program constructed at Statistics Sweden (Andersson and Nordberg, 1998) and designed to compute point and standard error estimates in sample surveys. It can be adapted to most designs in current use at Statistics Sweden, and to complex parameters, constructed as certain types of functions of totals.

In the case of nonresponse and when using the calibration estimator, we rely on an analogy with the estimator for two-phase sampling (treated in CLAN97), where the desired sample is first selected from the population, and a set of respondents is thereafter realised as a subset of the sample. We suggest that proxies derived by a calibration procedure replace the unknown response probabilities. Thus, it becomes straightforward to estimate the variance for the calibration estimator. The variance estimator is based on the assumption that the conditional nonresponse (given the realised sample) is negligible.

Variance estimation in the presence of imputation is a complex statistical problem. In recent years, considerable research has been put into "correct" variance computation when using imputation. Developments seem to be far from a conclusion, and one can expect the next few years to bring new results. The recommendations in this CBM are therefore preliminary.

The variance is a sum of two components, the sampling variance and the nonresponse variance. We give directions for how to estimate these components, when using the imputed estimator. A guiding principle is the desire to benefit as far as possible from CLAN97. The sampling variance estimate is calculated on the completed data set for some imputation methods, but for other methods residuals are added to the values. The dependence of the nonresponse variance on the imputation method is an inconvenience in that a new formula must be worked out for each imputation method. The formula also depends on the sampling design in use. These inconveniences are particularly pronounced when more than one imputation method is used in the same survey. The CBM gives expressions for the nonresponse variance in some special cases.

REFERENCES

Andersson, C. and Nordberg, L. (1998). CLAN97 – a SAS-program for computation of point- and standard error estimates in sample surveys. Statistics Sweden.

Lundström, S. and Särndal, C. E. (2001). Estimation in the Presence of Nonresponse and Other Survey Imperfections. Statistics Sweden.

RESUME

SCB, l'agence nationale de statistique de la Suède, vient de produire un manuel d'une centaine de pages résumant les "meilleures techniques disponibles" concernant l'estimation dans les enquêtes touchées par la non-réponse et les erreurs de couverture. Afin de réduire autant que possible la variance et plus particulièrement les biais causés par la non-réponse et les erreurs de couverture, l'utilisation judicieuse d'information auxiliaire joue un rôle primordial, et ceci pour les deux formes d'estimation qui entrent en considération: l'estimation par calage et l'estimation faisant appel à l'imputation. Le manuel présente également un compte rendu des procédures d'estimation de la variance, à l'aide des logiciels déjà disponibles, pour chacun des deux formes d'estimation.