The effect of model quality on model-assisted and model-dependent estimators of totals and class frequencies for domains

Ari Veijanen[1], Risto Lehtonen[2] and Carl-Erik Särndal[3]

We compare the effects of model quality on model-assisted and model-dependent estimators of totals and class frequencies for population subgroups or domains. As an example of a model-assisted method, we consider the generalized regression (GREG) estimator. The GREG estimator is known to retain good properties even when the model is incorrect; it is always nearly design-unbiased, for example. We compare GREG with a synthetic estimator, defined as the sum of fitted values over each domain. A synthetic estimator is known to have small variance, but it may suffer from considerable design bias. If the bias is large, a confidence interval is misleadingly narrow and does not cover the true value with the desired degree of confidence. Synthetic estimators can be expected to depend heavily on model quality. The paper draws on results in Lehtonen, Särndal and Veijanen (2003, 2004) and on more recent research (Lehtonen, Veijanen and Särndal 2005).

We study four aspects of model quality: (1) The mathematical form of the model. This aspect is likely to be particularly important for binary variables, for which logistic models are preferred to linear ones. (2) The kind of auxiliary information included in the model. We assume that the auxiliary variables values are known for all population elements and that domain membership is known for all population units. We can expect that inclusion of domain indicators in the model is important. (3) Should we formulate a fixed domain effects model or a mixed model with random effects for each domain? (4) How sensitive are GREG and synthetic estimators to outlying domains or outliers in the data?

Our simulation experiments concerning quality aspects (1)-(3) are based on repeated sampling from a fixed population. Our study of aspect (4) involves simulation of populations from a superpopulation model. The sampling weights were constant throughout the population, so as not to create an unfair advantage for the GREG methods (considering that in synthetic and other model-dependent methods, the sampling weights are usually ignored).

In the experiments, model improvement has a distinct impact on the accuracy of synthetic estimators, especially in large domains. For the synthetic estimators, the inclusion of domain indicators in the model was important. Without them, the synthetic estimators are highly inaccurate. A mixed model was found preferable to a model with fixed domain indicators. Model improvement was a much less important factor for the GREG estimators. As expected, they were nearly unbiased regardless of the model, whereas the bias of the synthetic estimators was sometimes large.

Nevertheless, the synthetic estimators usually had smaller mean squared error than GREG estimators. An exception to this was found in a robustness study with a single outlier domain.

[1] Statistics Finland
[2] University of Jyväskylä
[3] Canada

The presence of the outlier domain reduced the benefits of synthetic estimation. For a distinctly deviating domain, the GREG estimator assisted by mixed model was clearly better than corresponding synthetic estimator. The synthetic estimator is affected by the problem of estimated random effects that are biased towards zero, whereas the bias correction in the GREG estimator provides robustness. A general conclusion is that GREG estimators are little affected by the quality of model and they may in many cases be preferable to synthetic estimators, especially when the underlying model is of questionable quality.

References

Lehtonen R., Särndal C.-E. and Veijanen A. (2003). The effect of model choice in estimation for domains, including small domains. *Survey Methodology*, 24, 51-55.

Lehtonen R., Särndal C.-E. and Veijanen A. (2004). Does the model matter? A comparison of the accuracy of model-assisted and model-dependent estimators of domain totals. (Under review for Survey Methodology).

Lehtonen R., Särndal C.-E. and Veijanen A. (2005). Robustness of GREG estimators and synthetic estimators of domain totals to outlying domains. (Draft manuscript).