Journal of Official Statistics, Vol. 28, No. 3, 2012, pp. 359–362

Discussion

Paul A. Smith[1]

1. Introduction

The debate over whether design-based inference is the right approach for official statistics has a long history, as attested by the many references in Prof. Little's article, and shows little sign of early resolution. But his article gives persuasive arguments about the internal consistency and simplicity of having one approach across a wide range of statistics, which looks attractive. I'd like to explore some of the implementation details; as so often, I think that some of the details need to be worked through before we can evaluate the overall approach.

First let me admit that my training has been very largely in design-based inference; I have little experience of using model-based methods in practice, which is doubtless reflected in the questions I pose below.

Prof. Little says many sensible things in his discussion, which should be completely uncontroversial, but which seem not always to be considered in practical applications, and I would just like to agree wholeheartedly with these before moving on to the main matters:

- randomisation sampling should be the standard for all estimation methods (Section 2.2);
- sampling information is important in model-building (if only as a model check, Section 4.2);
- the need for good training in model-building and -testing (Section 6).

2. Where Model-Based Methods are Most Challenging

The development and many of the references quoted in Prof. Little's article consider relatively large populations with small sampling fractions (most similar to social surveys of one sort or another). Business surveys have characteristics which mean that different elements of the methods have more importance (Rivière 2002) – for example, sampling is practically never ignorable, so design information is essential to fit a suitable model; there is often a lot of auxiliary information; sampling fractions can be high; and distributions of variables can be very skewed. To take one particular example, small area estimation for business surveys cannot follow the same shrinkage-type approach as for social surveys, as many small areas contain no businesses at all, and some contain one enormous business, and any type of smoothing is too easily shown not to reflect the actual distribution. It is interesting that Hansen et al. (1983, Section 2.1) use a business survey example for where

[1]Office for National Statistics, Cardiff Road, Newport, NP10 8XG, UK. Email: paul.smith@ons.gov.uk
Acknowledgment: The views in this discussion are the author's, and do not necessarily represent those of the ONS.

apparently satisfactory model fits can give misleading inferences. For a paradigm shift of the magnitude suggested by Prof. Little, we need more information on the efficacy of model-based methods for business surveys.

The arguments in Prof. Little's article also seem targeted mostly at multi-variable, annual surveys. Will the approach work well in practice where there is a monthly survey with half a dozen variables and a short turnround time?

3. Internal Consistency

The potential big winner in Prof. Little's approach is the internal consistency of different estimates from the same survey (that is, consistency of estimates of different variables and at different levels, and distinct from "design consistency"). The same estimation approach can be used for the estimate for a whole country, a region and a borough. The design-model compromise (DMC) is avoided, and presumably all the estimates can be produced from a single software application.

What is less clear is whether there is a single *model* which will work well for a whole country, a region and a borough (or any other output domains). One of the big advantages of the design-based approach (where sample sizes are sufficiently large) is that a single set of weights can be calculated which allows estimates for any domain to be constructed so that they form an internally consistent system – lower levels of a hierarchy sum to higher levels; linear combinations of variables are preserved (for example, different breakdowns of the same total all sum to that total). An ideal calibrated Bayes (CB) approach would extend these properties to small domain estimates to produce an internally consistent system of estimates at all levels.

I said *potential* big winner because if such an internally consistent approach is not possible, then we risk replacing the design-model compromise with a model-model compromise. We would need to decide when one model or another should be used (a new "point of inferential schizophrenia"?), within the same survey, and would need to explain why the results are not internally consistent even though the underlying approach is the same. There may nevertheless be a good case for standardising on a single approach which can be used everywhere.

4. Sensitivity

Prof. Little points out that good frequentist properties provide a degree of protection against model misspecification. But design-based (model-assisted) estimation can be quite sensitive to the choice of model (see for example Hedlin et al. 2001, who also advocate careful model selection and evaluation for model-assisted estimation). Therefore the robustness of model-based estimation (including CB) is an important topic, for which a body of evidence needs to be gathered together. Only a review of how this works in a range of situations will have the credibility to provide a foundation to change the estimation paradigm for official statistics.

There are several components of sensitivity. First, how sensitive are the results to the model choice? Often there will be two models with similar properties and similar fits – do the outputs change substantially if the model changes? Second, how sensitive is the model choice to change in the data – this is the classical model robustness, which we would

expect to deal with by using robust models. And third, how sensitive are outputs to the choice of prior? If noninformative priors are used, presumably there will be little sensitivity, but some evaluation of this will be important.

An important element of model robustness for official statistics is the stability of models over time. Do models remain valid and sufficiently robust to changes? In particular, are the models able to cope with sudden changes (for example from the effects of natural disasters) and to operate satisfactorily around turning points in series? These are often the times when the signal in the data is most uncertain, and paradoxically when the most attention is being focused on the statistical outputs. At other times, model reviews may be analogous to periodic survey redesigns and therefore less contentious.

The last element of sensitivity of models concerns their objectivity. One of the concerns when model-based small area estimates were first produced in the Office for National Statistics was whether users would challenge the models on which they are based, since judgement is required in the choice of model; this has not been a problem in practice. However, where statistical outputs are used for resource allocation, users are often very vocal. We can expect that at some stage someone will challenge whether a different model (which gives them more money!) would not be more appropriate. Careful model choice and model checking will be an important part of demonstrating objectivity here – and perhaps should also have greater prominence for model-assisted methods.

5. Evaluation in Policy Decision Contexts

The best way to encourage users to agree to a change is to demonstrate that the new methods are better – for example, that resource allocations are more efficient, or that decisions can be made earlier, or that the same outputs can be produced for less. Prof. Little's Example 6, of language assistance at polls in areas where a single language minority with limited proficiency in English is above a given threshold, is interesting in this respect. It deals with some statistical properties, particularly reduced variance, but a classification relative to a threshold will still have some classification error – some areas that would be seen to be above the threshold if we could do a full enumeration will not get assistance based on the estimate, and vice versa. So it would be advantageous to work through to the *outcomes* – what would be the effect on the total costs of the assistance programme? How many people would miss out on or gain assistance relative to the design-based rule? Are the calibrated Bayesian estimates giving more efficient use of public funds (people with help available per $ spent)? These are difficult questions to answer, but the availability of one or two case studies would go a long way to making the case for change.

6. Conclusions

So in summary I find the idea of a single approach which can be applied in all situations attractive, but there is considerable work still to do to demonstrate the properties of calibrated Bayes estimators in a range of official survey contexts. We will need to have this backup to be able to sell a change to all users (not just a subset that works with models as in Section 4.1), some of whom are very conservative. If some improved properties (internal consistency, mean squared error or total survey error, cost) can be built in, the additional benefits will make the task of getting agreement to a change easier.

7. References

Hansen, M.H., Madow, W.G., and Tepping, B.J. (1983). An Evaluation of Model-Dependent and Probability-Sampling Inferences in Sample Surveys (with discussion). Journal of the American Statistical Association, 78, 776–793.

Hedlin, D., Falvey, H., Chambers, R., and Kokic, P. (2001). Does the Model Matter for GREG Estimation? A Business Survey Example. Journal of Official Statistics, 17, 527–544.

Rivière, P. (2002). What Makes Business Statistics Special? International Statistical Review, 70, 145–159.

Received June 2012