Statistical Science
2011, Vol. 26, No. 2, 262–265
DOI: 10.1214/11-STS346C
Main article DOI: 10.1214/10-STS346

Discussion of "Impact of Frequentist and Bayesian Methods on Survey Sampling Practice: A Selective Appraisal" by J. N. K. Rao

Eric Slud

I would like to congratulate Professor Rao on having produced an overview of survey methodology which is at the same time a broad-ranging prospectus of current research and also an impressive retrospective from a modern viewpoint of the early historical developments. He shows us in broad terms where the various approaches to survey methodology have been successful and where they cannot quite be relied upon without further development.

Most of the paper is not specifically directed at contrasting the Bayesian and frequentist viewpoints. The most important distinctions for Rao seem to be between model-dependent and design-based methods, and Bayes methods are faulted in Rao's chosen terrain of "the large-scale production of official statistics from complex surveys" primarily for using models where models are not absolutely necessary. He takes for granted that models will be used in adjusting for nonresponse, in his formulation largely through calibration, and in small area estimation. The faults he finds with unnecessarily model-dependent survey estimation methods are:

- design-inconsistency (of model-based BLUP under misspecified models, and in other examples, in Section 3.2);
- requiring different sets of predictor variables for different attributes of interest (in Section 3.3);

and in Section 4.2, in relation to the nonparametric Bayesian and pseudo-Bayesian methods relying heavily on exchangeability, for their

- lack of generalizability to complex survey designs with clustering and unequal probability weighting.

Like many authors in survey sampling, Rao faults model-based analyses because of possible model misspecification. This discussion highlights aspects and consequences of model misspecification under the headings of Rao's paper.

1. MODEL MISSPECIFICATION IN LINEAR REGRESSION AND CALIBRATION

In Section 3.1 of his paper, Rao considers the behavior of a calibration estimator (of a population total) when the calibration constraints involve some but not all of the predictor variables entering a true superpopulation model. The context is a superpopulation in which the regression model

$$Y_i = \beta' X_i + \gamma' Z_i + \varepsilon_i \tag{1}$$

holds for all units i in the frame $\mathcal{U}$, with auxiliary variables X_i, Z_i known for all population units, and where it is desired to estimate the total $t_Y = \sum_{i \in \mathcal{U}} Y_i$ based on a probability sample of units $i \in \mathcal{S}$ with first-order inclusion weights $d_i = 1/\pi_i$. [In Rao's example, the weights d_i are all equal, $X_i = (1, x_i)'$, and $Z_i = x_i^2$, for a scalar auxiliary variable x_i.] A calibration estimator of t_Y might be based on the variables X_i alone, that is, on $\sum_{i \in \mathcal{S}} w_i Y_i$ where the modified weights w_i are determined by minimizing $\sum_{i \in \mathcal{S}} (w_i - d_i)^2 / d_i$ subject to the constraints $\sum_{i \in \mathcal{S}} w_i X_i = \sum_{i \in \mathcal{U}} X_i$. As described by Rao, it turns out that this calibration estimator is equivalent to the generalized regression (GREG) estimator based on the weights d_i and the predictor variable X_i. In the setting with constant d_i, this estimator would be the unweighted model-based regression estimator based on predictor X_i.

As Rao suggests, calibration might be based on a subset of the appropriate predictor variables when the same universal calibration constraints are used over many different choices of response variables. In the context (1) above, there are three ways in which this

---

Eric Slud is Professor, Statistics Program, Department of Mathematics, University of Maryland, College Park, Maryland 20742, USA (e-mail: evs@math.umd.edu).