Discussion

Roderick J.A. Little
University of Michigan

INTRODUCTION

These fine papers take contrasting approaches to the problem of Nonresponse Follow-up (NRFU) analysis for the Census. The paper of Fay and Town (FT) falls within the tradition of design-based inference, where the aim is to make minimal modeling assumptions. To the extent that models are used they are "nonparametric", or perhaps "multiparametric", since the models implied by this approach tend to be saturated with many parameters. Zanutto and Zaslavsky (ZZ) work within a more model-based framework, smoothing via parametric assumptions. Both sets of authors are sophisticated practitioners within these traditions, and are aware of strengths of competing approaches. However, the contrast provides fertile ground for simulation comparisons of the differing approaches, which I hope to see in the future. In general, I would say that the aim should be to be more parametric and model-based for small areas, where data are sparse and smoothing is needed, and more nonparametric at higher levels of aggregation, where the data should be allowed to dominate over modeling assumptions. ZZ attempt to do this in their model by allowing a richer interaction structure at higher levels of aggregation.

DESIGN ISSUES

The introduction in FT provides a fascinating snapshot of the latest Bureau position on NRFU sample design. Last year the thinking seemed to focus on an initial Census mailing, followed by NRFU sampling of nonmail returns. The major design issues concerned the choice of sampling fraction for the second stage, and whether the sample design would sample blocks -- a "block" sample -- or sample individual households with blocks -- a "unit" sample. The block sample would be less costly because of more clustered fieldwork, but riskier for analysis, in that nonrespondents for entire blocks not included in NRFU would require imputation. Current plans call for a more complete "truncated census" to achieve a 90% response rate, followed by unit NRFU sampling at a 10% rate for the remaining 10% of nonrespondents. This seems much closer to a full Census than the earlier design, and given this it will be interesting to see if significant cost savings can be achieved over the 1990 Census design. In their presentation, FT showed data from the Census Test that suggested substantial loss of efficiency in moving from the unit to the block sample; it remains to be seen whether this finding holds up with further analysis, and whether it is reasonable to extrapolate findings from the Census Test areas to the country as a whole.

In work in progress with Ron Kessler and Steve Lewitsky at Michigan, I have been studying the possible gains in efficiency from subsampling call-backs for the National Comorbidity Study, a large national sample of mental health. The idea is that a small minority of cases in the sample consume disproportionate resources due to the large number of call-backs required to reach them, and a design that subsamples units after a certain number of callbacks may be more efficient; for example D(2,.5) = a design that follows 50% of those who do not respond at the second callback. Preliminary simulations indicate that the unit costs might be reduced by this form of subsampling, but not by a whole lot. The problem has obvious analogies with the NRFU problem in the Census, although the cost structure of a Census may be sufficiently different to make comparisons hazardous.

TOP-DOWN ANALYSIS APPROACHES

Turning to analysis issues, ZZ and FT favor a "top-down" strategy for filling in missing households, where the known count of nonrespondent households in an area are allocated across household types, either by hot-deck matching to sampled nonrespondents (FT), or by estimating probabilities of each household type from a loglinear model for the counts (ZZ). The alternative "bottom-up" approach builds a model for the joint distribution of household types as a sequence of conditional models, and then grows new households by drawing their characteristics according to this model (Schafer 1995). I like Schafer's comprehensive approach, but it is ambitious and a lot of work. Top-down seems more realistic given current capabilities.

The thrust of ZZ's paper is on the use of administrative records, but initially I focus on estimation from the NRFU and Census mailback data alone (Zaslavsky and Zanutto 1995). Figure 1 presents a stylized diagram of the missing-data problem, when a block NRFU sample is taken. Each row represents a block, the first b of which appear in the NRFU sample. The key task is to distribute counts for nonrespondents in blocks not in the NRFU sample -- the block with a question mark in Figure 1 -- into K household types. Two main sources of information are (A) distributions of respondents by household type in the same block, and (B) distributions of nonrespondents by household type in the sampled blocks. Figure 1 has somewhat the structure of a partially-classified contingency table, to which the methods described by Fuchs (1982) or Little and Rubin (1987, chapter 9) might be applied. However an important difference is that blocks are nested within S, not crossed, because of the NRFU design. Keys in the analysis are (a) to form block covariates, based on the respondent information or external data, that remove block heterogeneity, and (b) to use the NRFU sample to characterize how respondents and nonrespondents differ.

Figure 2 shows the data based on an NRFU unit sample. The structure is closer to that of a standard partially-classified contingency table, since blocks are crossed with S. By random selection, the distribution of HH types within blocks is the same for nonrespondents not in the NRFU sample (S = 0) as for nonrespondents in the NRFU sample (S = 1), suggesting that this information could be used directly for imputation with minimal modeling assumptions. However the S = 1 block is sparse! The ZZ model borrows strength from the (much more extensive) respondent distributions, by excluding interactions in their multinomial logistic model. The model is more detailed, and hence more realistic, than the models that underpin the ratio estimation methods of Fuller, Isaki and Tsay (1994), at the expense of greater complexity. The approach is, of course, subject to model misspecification, although ZZ try hard to confine the potential biases to lower levels of aggregation by building in interactions at the higher levels of aggregation. In contrast the estimation methods in the FT paper are comparatively simple, involving the imputation of real households by a nearest neighbor hot-deck. These methods involve relatively weak modeling assumptions, and are protected from bias by the random NRFU sampling. On the other hand the methods are relatively inefficient, particularly since no attempt is made to "borrow strength" from nonrespondents in similar blocks or respondents in same or similar blocks. Clearly, simulations that explore the bias-variance-simplicity trade-offs in these approaches would be of great interest.

The main focus of the ZZ paper is in the use of administrative records (ARs) to enhance estimation under their top-down model. Clearly there is a very high potential pay-off in the use of ARs, and research on incorporating them into the Census process is a major priority. There are clearly formidable problems to be overcome, logistics, matching and confidentiality to name a few. The analysis involves difficult issues

of response error, matching error and missing data. ZZ have taken an important step by their formulation of the problem, and by suggesting how administrative records can be imbedded within their analysis framework. Their ingenious simulation using the CPS-IRS match file highlights the potential impact of ignoring biases in the ARs, although of course it will be important to conduct similar simulations on datasets closer to the Census setting.

The issue of bias is best conceptualized by considering the joint distribution of HH type based on Census and ARs, forming a KxK matrix if there are K HH types. If there was a perfect correspondence between the classifications, then all cases fall on the diagonal where the two HH types are the equal, and simple substitution of the AR type for the Census type is the correct imputation. It is more likely that a given observed AR type maps into a set of Census types in this matrix; the appropriate imputation is then a draw from an estimate of the distribution over this set. Random AR error implies that the distribution of Census types is centered at the diagonal value, and bias occurs if the Census types are not centered at the diagonal value. The serious challenge is to estimate these joint distributions in homogeneous subgroups, given the available information.

VARIANCE ESTIMATION

Let me close by making a few remarks on FT's proposed variance methodology. FT propose a clever modification of the Rao and Shao method of variance estimation with incomplete data, based on a modified jackknife. The method provides an approach to variance estimation for nearest-neighbor matching imputation in the unit NRFU sample. The formulae involve variation between the nearest neighbor and second-nearest neighbor that would be matched when the nearest is deleted by the Jackknife. The method is quite simple to implement, and as in many matching methods, parametric assumptions are minimal. However, as FT acknowledge, the method ignores error in matching, that is, it assumes homogeneity of neighbors. This assumption may be questionable when the fraction of missing data on nonrespondents is large, as when the sampling rate of nonrespondents in the NRFU sample is small, and sampled nonrespondents are spread out geographically so that neighbors are not that close. The preliminary empirical evidence presented by FT suggests that the lack of homogeneity does indeed result in an overestimation of the variance. Also the lack of modeling assumptions at the low level of aggregation may lead to standard errors that are excessively variable, and require some smoothing. These problems might be eased by extensions of the method that make better use of household characteristics in forming matches.

REFERENCES

Fuchs, C. (1982). Maximum Likelihood Estimation and Model Selection in Contingency Tables with Missing Data. *Journal of the American Statistical Association* 77, 270-278.

Fuller, W.A., Isaki, C.T. and Tsay, J.H. (1994). Design and Estimation for Samples of Census Nonresponse. *Proceedings of the 1994 Bureau of the Census Annual Research Conference*, 289-305.

Little, R.J.A. and Rubin, D.B. (1987). *Statistical Analysis with Missing Data.* New York: John Wiley.

Schafer, J. L. (1995). Model-Based Imputation of Census Short Forms. *Proceedings of the 1995 Bureau of the Census Annual Research Conference*, 267-299.

Zaslavsky, A. M. and Zanutto, E. (1995). Models for Imputing Nonsample Households with Sampled Nonresponse Follow-Up. *Proceedings of the 1995 Bureau of the Census Annual Research Conference*, 673-686.

(This page needs to be scanned)

Figure 1. Data for "Top-Down" Estimation, NRFU Block Sample

BLOCK	BLOCK COVARS	MAILING RESPONDENTS (R=1) HH TYPE					In NRFU Sample? (S)	MAILING NONRESPONDENTS (R=0) HH TYPE				
		1	2	.....	*K*	Total		1	2	.....	*K*	Total

Figure 2. Data for "Top-Down" Estimation, NRFU Unit Sample

BLOCK	BLOCK COVARS	MAILING RESPONDENTS (R=1)					MAILING NONRESPONDENTS (R=0)									
		In NRFU Sample (S=1) HH TYPE					Not in NRFU Sample (S=0) HH TYPE									
		1	2	.....	*K*	Total	1	2	.....	*K*	Total	1	2	.....	*K*	Total