Discussion
by
Charles H. Alexander, Bureau of the Census, Washington, D.C. 20233

This is an historic project. The authors are introducing "multiple revolutions":

- imputation not weighting for unit (as opposed to item) nonresponse
- model-based, not donor-based imputations
- multiple imputation
- proper (full Bayesian) imputation

Perhaps the first "revolution" doesn't count, since unit nonresponse for the medical examination may be regarded as item response in the context of the entire survey. However, the other three aspects of their work are genuinely revolutionary; as far as I know, this is the first real, live implementation of model-based, multiple, Bayesian imputation for one of the "mainstream establishment" Federal surveys.

There has been resistance to these "revolutionary" developments because they require some "paradigm shifts":

- acceptance of "made-up" data
- forcing data users to "do statistical inference", willingly or not
- Bayesian vs. "likelihood" inference
- getting mathematical statisticians involved in item nonresponse, traditionally the domain of subject-matter experts and the computer processing staffs

To this list would be added "model-based vs. design-based inference" if their method is viewed as replacing weighting for unit nonresponse.

The NHANES III Survey is fertile ground for this multiple imputation revolution. It is an important survey, with a high missingness rate for important variables. Good co-variates are often available, even for "total nonresponse" in the medical examination. In similar circumstances, even we counter-revolutionaries have considered model-based, multiple imputation for missing income on the Consumer Expenditure Survey. In fact, to show how attitudes have softened, we are even proposing to impute adjusted responses to "replace" **observed** data on labor force status for the American Community Survey, to force State estimates to agree with the less biased Current Population Survey estimates.

Let me return to reasons for resistance to the new paradigm. First, there is the reluctance to use "made-up" data. The model-based imputations generate values that were never observed for any real individual. This may be well and good for those who are interested in "a population scalar quantity of interest". But some data users have a much more intimate relationship with their data. They may study the details of graphical relationships. They may even pore over printouts of case-by-case observations, gaining valuable insights about the population. It is important to such users that the data be "real". A previous discussant referred to the dangers of "data dredging", but often this "true love" for real micro data is a legitimate and valuable passion. Perhaps the desire for "real data" is related to the "omitted variable problem" mentioned below, i.e., to the possible distortion of complex relationships, when imputing data.

My concern is:

> for graphical or case-by-case analysis, what is a user to do with multiply imputed date sets?

This concern is reinforced by a quote from Schafer, Khare and Ezzati-Rice (1993, p. 472): "This section presents some graphical displays and exploratory analyses of the imputed values....Rather than examining all ten sets of multiple imputations, which would have been very tedious, we focussed our attention on set MI_0...."

On the other hand, the mystique of imputing real values "borrowed" from a "donor" may be illusory. For item nonresponse, is it really possible to avoid "making up" relationships by hot deck or donor methods, except in the exceptional case where the donor matches the recipient on all observed variables? (I would have said "no", but the New Imputation Methodology (NIM) being developed by Statistics Canada gives me pause. I'm not familiar enough to endorse it one way or the other, but this method has already advanced the state-of-the-art as far as considering relationships of variables in donor-based methods).

A second source of resistance is the "omitted variable problem", or "Bob Fay's concern", which is mentioned by Schafer, et. al. (As I understand it, this is really a problem with the "modelling" rather than the "multiple".) Pragmatically, this is a problem for two reasons. First, the solution to this problem seems to require (for now, at least) that you need to call in an expert to decide whether your model is adequate for your intended use before you use the method. Secondly, it is not yet clear how widespread the problems are, even after you call in an expert.

It will be interesting to see how our view of this problem evolves over time. I see two possible futures. One is that serious "omitted variable" problems are regularly encountered and multiple imputation eventually will be replaced by some other method of accounting for the imputation variance: perhaps a modification of replication methods, perhaps something else. The other is that multiple imputation will be routinely used without extensive checking, much as design-based methods are regularly used without verifying the appropriate normality (or t-distribution) of estimates. One hopes that the latter outcome would be based on a collective experience that serious problems with the method are rare.

Time will tell. In the meantime, I do not see this problem as a sufficient reason to do nothing rather than do multiple imputation, which currently is the only method general enough to do what Schafer, et. al, have done for NHANES III. Nonetheless, we should not kid ourselves that users will never "....use the imputations for a purpose for which they were never intended."

The authors hint at third future, in which averaging U_1..., U_m is found to improve the stability of the sampling error so much that multiple-imputation-enhanced replication methods become the standard methods for estimating variances. This possibility is intriguing, but clearly speculative. It may be instructive to think more about how Ui and U_j are related, to try to understand the results.

Including multiple imputations on the public use file will force data users to confront the problems of inference in the presence of uncertainty. I think that, unlike design effects and sampling weights, the basic message of the multiple imputations will be easy to understand: when imputation is needed, the value for a given unit could be this value, or that value, or one of these other values. This will illustrate uncertainty more graphically than anything else we provide to the data user. I do not suggest that this will lead to a "proper inference" in which the coverage probability of a given interval appropriately influences the analyst's state of mind. My impression is that variances and confidence intervals typically are used only as a rough guide to how good the estimates are; given the various measurement biases not included in the calculations, this may be just as well.

It will be interesting to see how users react to the multiple data sets. Will they find a way around confronting the uncertainty, such as analyzing the first set of values and forgetting the rest? It might be worth trying a customer survey to ask whether and how people use the multiple imputations after the file is released.

An alternative, which may simplify (or oversimplify) things for the user would be to calculate the "relative increase in variance due to nonresponse" as described by Schafer, et al, and then "generalize" these increases, as is routinely done for "generalized variance functions". For a particular survey, if a group of estimates all have about the same relative increase (or if the relative increases for the group can be fit well by a simple function of the magnitude of the estimate), then this would be used to define the "generalized" value for each estimate. The data user would then operate on a singly imputed file, but would inflate the complete- data estimated variance by the generalized relative increase. This is similar in spirit to the variance multiplier used by Judkins and Winglee (1992), described in Schafer, Khare, and Ezzati-Rice (1993).

Another concern is the computations necessary to implement the proper or full Bayesian imputation, taking into account the uncertainty in the estimated parameter $\hat{\theta}$. The full Bayesian method apparently stretches the authors' computer time and storage, and thereby limits the number of variables, both the variables to be imputed and the explanatory variables. Improvements in computers may eventually solve this problem, but will not by themselves eliminate the necessity to monitor the convergence of the method. Right now, it requires an expert just to compute the estimates. The applicability of the method will be severely limited until the development of the method gets beyond this stage.

My question is how bad is it, for a large survey, to assume that $\hat{\theta}$ is known and impute accordingly? For this survey, the total effect of imputation on the confidence interval is not all that large and I would be very surprised if the portion of this due to uncertainty about $\hat{\theta}$ is worth worrying about. For the aforementioned project to impute missing income for the Consumer Expenditure Surveys, this issue will be investigated by a joint research project involving BLS, Census, and a team led by one of the authors of Schafer, et al.

Finally, Bayesian methods do present a perception problem for official statistics because of the notion of starting out with a "prior belief" about what the answer should be. However, with a noninformative prior, this problem should not be fatal.