DISCUSSION

Susan M. Hinkins, Internal Revenue Service

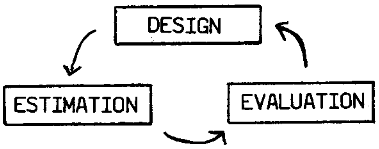
I would also like to thank Rod Little for valuable work in the area of nonresponse. A major contribution of his presentation is that it clearly puts in focus two basic approaches for dealing with missing data, and it gives additional emphasis to the use of the propensity score [5]. The general issues have been exposed. As a practitioner, it feels good to see a basic framework developing.

My orientation is that of an applied statistician - with the Internal Revenue Service (IRS). A significant amount of my experience there has been with nonresponse problems, estimating missing information. When the Internal Revenue Service is mentioned, the first words to cross one's mind may not be "sample surveys." But every April many of you take part in at least one of our surveys; the Individual Income Tax Returns comprise one of the populations that we sample. The IRS also surveys Corporate Returns, Partnerships, and many other more obscure subjects.

In what follows, I will make only a few comments - first on some general issues brought out in this presentation, and then on some problems of particular interest in the IRS sample survey environment.

1. THE SAMPLE SURVEY SITUATION

If the sample survey situation were simplified to just three steps, these might be:



The presentation we just heard was primarily concerned with estimation. To reiterate Rod's emphasis, when there is nonresponse, the estimation procedure requires some form of modelling - in terms of either propensity or prediction. Obviously, however, consideration of these two basic types of modelling carries throughout all three stages.

Design Stage -- One of the most important things that can be done here is to make every effort to minimize nonresponse. But it is also crucial in the design to address the modelling issues as well. It has been my experience, and I believe of others as well, that fully successful modelling for propensity or prediction is a major difficulty. Certain types of covariates may be needed to estimate propensity scores; other types of covariates may be needed to predict missing values. This basic framework of propensity and prediction makes the design problem simpler to approach - I have to ask whether I am looking for predictors of response or looking for predictors of missing variables.

Evaluation Stage -- When nonresponse is at all sizeable, the evaluation portion of the survey must also be enhanced, to include some form of sensitivity analysis. As Rubin has emphasized [6], the sensitivity of the final estimates to the models and to the imputation procedures must be evaluated. This can be done using multiple imputations. This is a non-trivial job, requiring a considerable committment of resources.

2. MODELLING ISSUES

Rod's emphasis on a model-based approach is valuable in general, and of course necessary when there is nonresponse. Recently, there has been considerable discussion of model-based approaches in sample surveys, especially regarding what conditional distribution should be used for inference. Some of the specific questions seem to be:

- How much time and effort to spend on modelling?
- What model is appropriate?
- What is the appropriate means of evaluating the result?

Regarding the first point, it is unfortunate that in practice one will often decide on a procedure first and only formalize the model later. This happens because a procedure is often selected primarily for its practicality: Because it works on the computer system, or because someone else used this method and it worked. Sometimes the model relies solely on undocumented subject matter expertise or knowledge. In the crunch to get a procedure working, I have "modelled" by simply asking a subject matter expert which types of returns will be similar to each other. In other words, in many applications, too few resources have been given to the problem of modelling, to designing an estimation procedure.

To go on to the next point, we are particularly interested in the multivariate modelling problem. In some of our larger surveys, there are several hundred items of interest, and there are likely to be different patterns of response for different variables. The data can be missing because of several different mechanisms:

- in some cases we simply cannot wait for the information, but must produce advance estimates.
- blocks of items are sometimes unavailable due to differences between the taxpayer's accounting practices and the IRS tax form. (The IRS does not collect this financial data primarily for our sample survey.)
- items can be randomly lost in transcription, in the process of recording the numbers from the return. (As noted in the presentation, for such a variable any procedure except simple weighting will increase the variance.)

There is a need for different types of models within one data set, and there may also be a need for different degrees of modelling even for one variable. For example, we have one particular nonresponse problem where the current imputation procedure is essentially equivalent to mean imputation within cross-classes [2]. This is an effective procedure for estimates of aggregates, but in our case it is very expensive. It may be possible to consider a simpler approach for some categories of returns, and to improve the estimation for other classes of records.

This relates to the third point, our population of interest is very skewed - a small number of records has a dominant effect on our estimates. A relatively few large units (for example very large corporations) dominate the estimate. With such a skewed distribution, the simple mean square error, averaging over aggregates, is probably not an adequate measure of the effectiveness of the procedure. This is an area that, from our point of view, needs further study.

3. PARTICULAR PROBLEMS OF INTEREST

As just mentioned, we are particularly interested in the problem of estimation and imputation when the population is highly skewed. Another dimension to consider in modelling for missing data in the IRS environment, is that most of the sample surveys are longitudinal. Therefore, observed information from a previous year's record could be used to predict missing information on this year's record. While the presentation today is primarily concerned with cross-sectional surveys, I would be interested in any comments Dr. Little - or others - might care to make on longitudinal samples.

The next point is that some of our data sets are made available on public-use files, which implies certain additional problems. Survey data are used for several objectives. While the primary users may be estimating population and subpopulation totals, there is an increasing interest in microdata, in small subsets of the sample. These different objectives may imply different techniques for handling nonresponse. On public-use data files, one must decide to what extent the imputation procedures will try to meet various needs of the users, and to what extent the secondary users may want to use their own models to redo the imputation. This should always be possible if high standards for documentation are maintained. (This sounds easy, but anyone who has documented files knows that it is a difficult and time-consuming chore.)

Finally, there are two areas where there is a need to design procedures that miss, or lose, data intentionally. The first is prompted because of disclosure issues on a public-use file. We can only release data in such a manner that it is essentially impossible for anyone to identify individual records. This is an area of particular concern for government practitioners as producers of such files, but it needs more attention generally from data users as well [4] [7].

The second is prompted by the need to find strategies to lower cost without compromising aggregate data and microdata. For example, in one survey we collect over 600 items from a record. To collect this information from one record may take as little as an hour or as long as a week [1]. Matrix sampling (a multivariate subsampling procedure) has been introduced into the design [3]. The obvious advantage to this purposeful omission of data is that one is able to feel reasonably confident that the model is correct.

These last topics are areas of particular importance in my work and I would hope that some of you might find them interesting and important areas for future research.

Again, I would like to thank Rod Little for a valuable paper, of particular interest to the applied statistician.

REFERENCES

[1] Cys, K., Hinkins, S., and Rehula, V. (1982). "Automatic and Manual Edits for Corporation Tax Returns," American Statistical Association, Proceedings of the Section on Survey Research Methods, 1982.

[2] Hinkins S. (1982). "Imputation of Missing Items on Corporate Balance Sheets," American Statistical Association, Proceedings of the Section on Survey Research Methods, 1982.

[3] Hinkins, S. (1984). "Matrix Sampling and the Effects of Using Hot Deck Imputation," American Statistical Association, Proceedings of the Section on Survey Research Methods, 1984.

[4] Oh, H.L. and Scheuren, F.J. (1984). "Statistical Disclosure Avoidance," presented at the Washington Statistical Society, May 24, 1984.

[5] Rosenbaum, P.R. and Rubin, D. (1983). "The Central Role of the Propensity Score in Observational Studies for Causal Effects," Biometrika, 70.

[6] Rubin, D. (1980). Handling Nonresponse in Sample Surveys by Multiple Imputations. U.S. Dept. of Commerce, Bureau of the Census Monograph.

[7] Spruill, N. (1983). "The Confidentiality and Analytic Usefulness of Masked Business Microdata," American Statistical Association, Proceedings of the Section on Survey Research Methods, 1983.

For further information see also:

[8] Little, R.J.A. (1982). "Models for Nonresponse in Sample Surveys." Journal of the American Statistical Association.

[9] Rubin, D. (1977). Formalizing Subjective Notions About the Effect of Nonresponse in Sample Surveys," Journal of the American Statistical Association.

[10] Sande, I. (1982). "Imputation in Surveys: Coping with Reality," The American Statistician.