DISCUSSION

Donald B. Rubin, Educational Testing Service

It gives me great pleasure to discuss these papers on the important applied problem of handling nonresponse in surveys. I am especially pleased to see in them the use of multiple imputations because I firmly believe that multiple imputations will become the standard, accepted technique for handling item nonresponse in the future.

The technique of multiple imputations replaces each missing datum by a vector of possible imputations, as depicted in Figure 1. The values in the vector in general represent: (a) possible values under one model for nonresponse; and (b) possible values under different models for nonresponse.

For example, if the vector of multiple imputations had four components, the first two might represent two possible values under model 1 and the second two might represent two possible values under model 2. By model, I don't necessarily mean a formal, explicit statistical model, like a normal linear regression model, but include hot-deck and other implicit models. Different regression models might differ because they include different independent variables, and different hot-deck models might differ because they use different classification or matching variables.

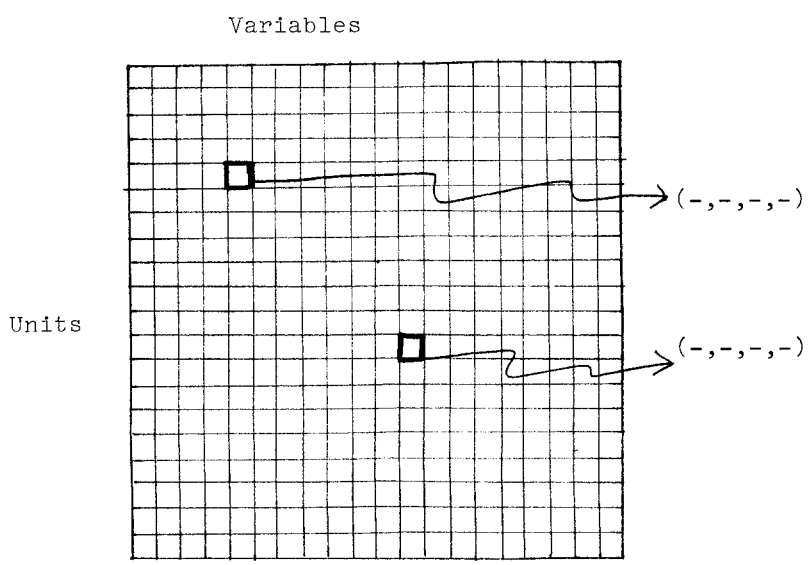
Although at first it might appear difficult to analyze a multiply-imputed data set, in fact, it is quite easy. For example, with four values in each vector, create four complete-data sets, and analyze each by standard complete-data methods just as if there were no nonresponse. The results under one model can be combined to produce one resultant answer as the Herzog and Lancaster papers illustrate. Answers from different models should not be explicitly combined.

Multiple Imputations Under One Model

Multiple imputations under one model are used to represent the best guess about the distribution of possible values for the missing items. Formally, they are intended to simulate the posterior distribution of the missing items. Although randomly drawing from possible values is one way to represent this posterior distribution, in practice it might be more efficient to systematically choose a few values.

There are two advantages to multiple imputations under one model. As Herzog and Oh-Scheuren illustrate, multiple imputations reduce the real variance of estimation. (This effect, however, is not likely to be large.) The second, and usually more important, effect (also illustrated by Herzog and Oh-Scheuren) is that multiple imputation allows the simple calculation of valid standard errors for estimates.

Figure 1: Pictorial Description of a Multiply Imputed Data Set with Four Imputations for Each of Two Missing Values



Multiple Imputations from Different Models

The unique function of multiple imputations is to display sensitivity of inference to different imputation models, that is, to different assumptions about how nonrespondents might differ from respondents. The great advantage of such sensitivity analyses is that they expose the potential sizes of hidden residual biases.

The residual bias can be important in practice, and so displaying possible directions and sizes for it is essential to a well-grounded applied enterprise. The work presented here using administrative data suggests that in some contexts we may be lucky and have small residual bias after carefully controlling for important survey variables. In the Current Population Survey (CPS), missing social security benefit data seem to fall roughly into this category (Herzog). On the other hand there may be large residual bias as appears to be the case with missing CPS property income (Oh-Scheuren) or wages (Greenlees, Reece and Zieschang, 1980).

Ignorable vs. Nonignorable Models

An important distinction between models for nonresponse is the difference between ignorable and nonignorable models. Ignorable models assume that if we could properly control for all "background" variables recorded in the survey and used in the model (e.g., regression independent variables or hot-deck classification variables) there would be no residual bias. The Welniak-Coder paper addresses the problem of biases due to ignorable models: their deletion (nonresponse) mechanism used to create missing values, although using realistic rates on nonresponse, relied on classifications based on background variables (but not the values of the variables being deleted). That is, the resultant missing data are "missing at random" (Rubin, 1976, Biometrika).

Nonignorable models assume that the nonresponse is due not only to these observed background variables, but also to unobserved variables possibly correlated with missing values that are to be imputed. The Herzog-Lancaster and Oh-Scheuren papers admit the possibility of such variables (and thus the possibility of nonignorable models) because they study the problem of biases due to the actual mechanism which creates nonresponse.

The study of a number of alternative ignorable and nonignorable models is very important and is easily implemented by including imputations to represent a variety of models of both types.

Explicit vs. Implicit Models

Although the tradition among mathematical statisticians has been to employ explicit probability models (like the two-stage linear/log-linear model in Herzog), the tradition among applied statisticians, especially in the survey area when dealing with nonresponse, has been to employ implicit models, such as the Census hot-deck. At least in the near future, I do not anticipate a strong movement among applied statisticians to use explicit models.

Although trained, in some sense, to be distrustful of them, I continue to be impressed with the power of hot-deck-like methods. The techniques are simple, and the resultant imputations for the CPS seem to be quite good. Nevertheless, as Scheuren points out (Scheuren, 1978), in a new survey context where years of experience and hot-deck refinements do not exist, implicit models may not do very well unless based on preliminary explicit modelling efforts aimed at discovering relationships among variables.

The distant future belongs to explicit models, but current explicit models appear relatively expensive and inflexible when applied to large data bases such as the CPS. There just does not exist a large collection of models between the normal linear regression model and the log-linear model that allows constraints to be handled and an appropriate number of parameters to be used: the normal models typically have too few parameters for the size and complexity of these data bases, whereas log-linear models typically have too many when data are not artificially grouped. It is always dangerous to use models that are contradicted by the observed data. In smaller surveys, commonly used explicit models are more competitive because they are not as easily contradicted.

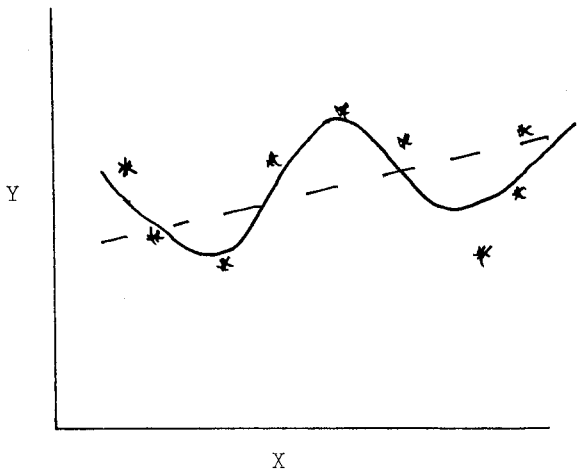
A critically important activity when using models, implicit or explicit, is to fine-tune them to be as appropriate as possible for the data. Welniak and Coder show one method for studying how to tune hot-deck models. Ideally, of course, many features of the distribution of variables should be checked, not just the means. Also, ideally, to guard against nonignorable nonresponse "real" responses from nonrespondents would be obtained and the models fine-tuned to these real responses. In any case, the Welniak-Coder paper shows how implicit models can be adjusted to be locally accurate. (Performing such fine-tuning within a formal, explicit modelling framework could be quite difficult.)

Dangers with Implicit Models

Even though there are compelling practical reasons to use implicit imputation models rather than explicit imputation models in large surveys, there are inherent dangers with implicit models. My experience suggests that current implicit imputation models suffer two problems. They (1) do not smooth (borrow strength) enough, and, on the other hand, (2) they underestimate actual variability. The figure below illustrates these.

Suppose the points in Figure 2 represent the actual data, (X,Y) where X is the background variable used for imputation and Y is the variable to be imputed. The solid line represents the values of Y predicted by an implicit model, and the dashed line represents the values of Y predicted by an explicit (linear) model. The solid line follows the points closely just as a hot-deck tends to use the observed values of Y corresponding to the closest value of X. Relative to the dashed line, the solid line predicted values of Y are less regular (not smooth enough) and the residual variability of the predicted value is less (too small). In other words, implicit models tend to incorporate noise into predictions. This is not to say that implicit models should be avoided, rather that theoretical considerations underlying explicit models should be kept in mind when using implicit models. For example, in hot-deck cells with one or two donors, borrow stength from adjacent cells.

Figure 2: Illustration of Dangers of Implicit Model (solid line) Relative to Explicit Model (dashed line)



Summary

In conclusion, I find these papers to be an important collection, that in combination, form a package that represents exactly the kind of work needed to handle the nonresponse problem.

A final comment is that I hope resources can be set aside to provide multiple imputations, because without these, it is often very difficult to assess the extra variability in inferences due to nonresponse. Setting aside funds for multiple imputations in a survey system is, in a sense, similar to setting aside funds for evaluation of a new social program. In order to know how successful the primary procedure (imputation system/new program) is, some resources have to be allocated to secondary procedures (multiple imputations/evaluation experiment). There is always an argument that all resources should be devoted to the primary procedure, but if this procedure must ever be defended, sensible secondary procedures are necessary.

References

[1] Greenlees, J.S., Reece, W.S. and Zieschang, K.D. (1980), "Imputation of Missing Values When the Probability of Response Depends upon the Variable Being Imputed," submitted to the Journal of the American Statistical Association.

[2] Herzog, T.N. and Lancaster, C. (1980), "Multiple Imputation Modeling for Individual Social Security Benefit Amounts--Part I," American Statistical Association, Proceedings of the Section on Survey Research Methods.

[3] Herzog, T.N. (1980), "Multiple Imputation Modeling for Individual Social Security Benefit Amounts--Part II," American Statistical Association, Proceedings of the Section on Survey Research Methods.

[4] Oh, H.L. and Scheuren, F.J. (1980a), "Estimating the Variance Impact of Missing CPS Income Data," American Statistical Association, Proceedings of the Section on Survey Research Methods.

[5] Oh, H.L. and Scheuren, F.J. (1980b), "Differential Bias Impacts of Alternative Census Bureau Hot-Deck Procedures for Imputing Missing CPS Income Data," American Statistical Association, Proceedings of the Section on Survey Research Methods.

[6] Rubin, D. (1976), "Inference and Missing Data," Biometrika, Vol. 63, No. 3, pp. 581-92.

[7] Rubin, D. (1980), "Using Multiple Imputations in Sample Surveys--A Phenomenological Bayesian Approach to Nonresponse," American Statistical Association, Proceedings of the Section on Survey Research Methods.

[8] Rubin, D. (1980), "Using Multiple Imputations to Handle Nonresponse," Incomplete Data in Sample Survey: The Theory of Current Practice, National Academy of Sciences, Panel on Incomplete Data.

[9] Scheuren, F.J. (1978), "Discussion of the Paper Entitled 'Income Data Collection and Processing for the March Income Supplement to the Current Population Survey,' by J. Coder," Proceedings of the Data Processing Workshop: Survey of Income and Program Participation, U.S. Department of Health, Education, and Welfare.

[10] Welniak, E.J. and Coder, F. (1980), "A Measure of the Bias in the March CPS Earnings Imputation System and Results of a Sample Bias Adjustment Procedure," American Statistical Association, Proceedings of the Section on Survey Research Methods.