DISCUSSION

Roderick J.A. Little, University of Michigan
Department of Biostatistics, University of Michigan, Ann Arbor MI 48109-2029

The authors of this session are to be commended for tackling important and difficult applied problems. Nonresponse on income and asset amounts in government surveys is an important obstacle to valid inference, because it is extensive, and nonrespondents tend to differ systematically from respondents. Imputation can be a useful tool for addressing the problem, but requires sound and thoughtful modeling. The imputation problems considered here are complicated by the size of the data sets, the multivariate nature of the outcomes, the multiple units of aggregation in the surveys -- job, individual, family, census block, and so on -- and skewed distributions with lumps of probability at zero representing people who do not have the income or asset in question.

Little (1988a) suggested that imputations should be (1) based on sound prediction models, either implicit (as in a hot deck) or explicit (as in a regression-based methods); (2) conditioned on observed variables for each case; (3) multivariate for multivariate outcomes, that is should reflect correlations between the variables that are jointly missing; (4) draws from a predictive distribution rather than means, to preserve distributions; (5) multiple as in Rubin (1987) rather than single, to allow imputation uncertainty to be reflected in the inference. New tools of statistical inference, such as Rubin's multiple imputation theory, improved imputation models, and powerful Bayesian simulation tools such as Gibbs' sampling, are making these principles more attainable in practice. See, for example, Kennickell (1991), or Khare et al. (1993).

The **Paulin and Sweet paper** (PS) concerns nonresponse in family Wages and Salary (WS) for the Consumer Expenditure Survey. Issues include (1) the level of aggregation; (2) choice of outcome and (3) choice of covariates.

(1) *Level of aggregation*. PS choose to model WS aggregated over the set of jobs within a family. While this approach yields a univariate regression problem, it is not well suited to deal with partial nonresponse within a family, where some members respond and others don't. It is also conceptually unappealing in that WS attaches to a job, not a family. An alternative approach is to impute WS for individuals (or better for jobs within individuals), and then aggregate the imputed WS amounts for each family. This approach is somewhat more work, but it handles partial nonresponse within a family, and is conceptually more appealing. For jobs within a family ordered in a sensible way, let W_j be the WS for job j, with associated covariates X_j. Then multivariate imputation can be based on the sequence of univariate regressions $(W_1|X_1), (W_2|W_1, X_2),..., (W_k|W_1,...W_{k-1}, X_k)$, where the conditioning is designed to preserve multivariate associations of WS values within families.

(2) *Choice of Outcome*. PS model the mean of W_i or $\log(W_i)$ as a linear combination of covariates. The log model is more natural for economists, and may be less heteroscedastic and interaction-prone. However, care is needed to avoid excessively high imputes of WS when unwinding the log transformation; see for example David et al. (1986). The regression of W_i can be expected to need interactions with T_i, the number of hours worked. If T_i is reasonably well measured, an alternative is to model W_i / T_i, a form of wage rate. If the variance is assumed proportional to T_i^{-1} we have an example of what I like to call an extended ratio model, which has some pleasing properties (Little 1988b).

Choice of Covariates. The PS paper places too much emphasis on strategies for reducing the number of covariates, in my view. I agree with Rubin's advice that, in the imputation context, the added variance from including unnecessary variables in the imputation model is less serious than the bias

from omitting variables that in fact have non-zero effects. Simply put,

Let those betas bounce!

This approach should be linked with *multiple* imputation of draws from the predictive distribution, to reduce imputation variance and allow easy assessment of the added variance from imputation.

Thus, in PS's problem, I would consider carefully if additional covariates should be added to the model. Although I am not sure what covariates are available for use in the survey, one set of obvious candidates are more detailed occupational codes, given the obvious relationship between occupation and WS. For example, David et al. (1986) use much more occupational detail in their model for WS in the Current Population Survey.

Steve Heeringa's paper describes an interesting problem concerning multivariate imputation of asset data in the National Institute of Aging's Health and Retirement Survey. Heeringa and his colleagues are to be commended for their ambitious plans to apply modern imputation technology based on Gibbs' sampling for the general location model in this complex setting. Kennickell's (1991) work was pioneering in this regard. Heeringa reports plans for analysis rather than results at this stage, and I shall make some comments on the proposed methods below.

The paper provides a very interesting and apparently successful application of the idea of seeking interval data on asset amounts when an exact amount is withheld. The apparent reduction in the level of nonresponse from this strategy is impressive. More research would be useful to determine which of the two methods for obtaining interval data -- the sequence of questions or the range card -- is preferable. The range card seems less cumbersome if it can produce data of comparable quality.

The interval data create an interesting problem for analysis. If the ranges are narrow enough -- how narrow is another interesting research topic -- a simple imputation method would multiply impute random draws of amounts within the interval, (see for example Little 1992). The Heeringa paper proposes a much more sophisticated model-based strategy. While the general approach is appealing, modifications may be needed to make the approach workable.

Suppose there are Q assets to be imputed, let Y_j be the jth asset amount $(j = 1,\ldots,Q)$, and let W_j denote the interval indicator for Y_j, with I_j ordered categories. Heeringa proposes to fit the general location model to the incomplete data on these Q continuous and Q categorical variables. A serious practical issue is that the categorical variables form an $(I_1 \times I_2 \times \ldots \times I_Q)$ way contingency table. With $Q = 11$ asset amounts and five intervals for each amount, this yields a table with $5^{11} \cong 49$ million cells! The proposal to restrict the contingency table model to 2-way associations helps, but the number of parameters remains large and problems with sparse cells can be expected. The problems with the size of the basic model for the W's and Y's are compounded when covariates are included, as seems essential for imputing cases where asset amounts are not bounded within an interval. Another problem with the general location model is that it implies a constant covariance matrix for the Y's within each cell. A transformation to the log scale may help, but the assumption still seems unlikely to be met. Also, some provision is needed to deal with the $W_j = 0, Y_j = 0$ cells, which do not fit the assumptions of the model.

One way of reducing the dimensionality of the contingency table is to replace the W_j's by binary variables for presence or absence of the amount, and treat the non-zero interval data as interval-censored. When applying the Gibbs' algorithm, draws for interval-censored values are simply constrained to lie in the observed intervals by rejection sampling. While many draws are rejected and computing time may be lengthy, the gain in model parsimony seems very worthwhile.

The problem of how to deal with the zero cells remains. Little and Su (1987) propose an approach where the zero's are all treated as missing data, and then the imputed amounts set to zero at the end. That paper concerned maximum likelihood estimation, but the approach might be applied to the Gibbs' methods proposed by Heeringa. An even simpler approach is to eliminate the categorical W_j variables altogether, and treat the zero amounts as interval-censored in the range $(-\infty, 0)$, as in a Tobit model. For non-recipients, negative draws are accepted by the Gibbs' algorithm, and then at the end are replaced by zeros.

Speed of convergence of these algorithms may be a problem, and imputes in the right tail of the asset distribution need care. I like Heeringa's idea of using historical data here, but would rather use the data as a basis for a prior distribution, rather than using values directly in a cold deck. In short, many practical issues arise, but the power of the general approach makes the effort well worth pursuing.

References

David, M., Little, R.J.A., Samuhel, M.E. and Triest, R.K. (1986). Alternative methods for CPS income imputation. *Journal of the American Statistical Association,* 81, 29-41.

Kennickell, A.B. (1991). Imputation of the 1989 Survey of Consumer Finances: Stochastic Relaxation and Multiple Imputation. *Proceedings of the Survey Research Methods Section, American Statistical Association 1991*, 1-9 .

Khare, M., Little, R.J.A., Rubin, D.B. and Schafer, J.L. (1993). Multiple Imputation of HNANES III. To appear in *Proceedings of the Survey Research Methods Section, American Statistical Association 1993.*

Little, R.J.A. (1988a). Missing data in large surveys. *Journal of Business and Economic Statistics,* 6, 287-301 (with discussion).

Little, R.J.A. (1988b). Statistics at the World Fertility Survey. *The American Statistician,* 42, 31-36.

Little, R.J.A. (1992). Incomplete Data in Event History Analysis. Pp. 209-230 in *Demographic Applications of Event History Analysis*, J. Trussell, R. Hankinson and J. Tilton, eds. Oxford: Clarendon Press.

Little, R.J.A. and Su, H.L. (1987). Missing-data adjustments for partially-scaled variables. *Proceedings of the Survey Research Methods Section, American Statistical Association 1987,* 644-649.

Rubin, D.B. (1987), *Multiple Imputation for Nonresponse in Surveys*, New York: John Wiley.