DISCUSSION

Ralph E. Folsom, Research Triangle Institute
Research Triangle Park, N.C. 27709-2194

Before I begin discussing the individual papers, I want to congratulate our session chair and her employer, NCHS, for sponsoring this important work. As you have seen, the presenters and their collaborators took on a large scale, real world, test of their methods and achieved satisfying results. The multiple imputation application is particularly impressive. While I personally prefer solutions that have more of a nonparametric flavor, I am not aware of any alternative strategies that attempt as comprehensive a solution.

Turning to the papers presented today, I am going to concentrate my comments on the first three. The fourth paper by Clifford Johnson and his colleagues at NCHS begins to delve into the thorny issue of how to disseminate multiply imputed data sets. More work clearly needs to be done in that area.

Focusing on the Fahimi and Judkin paper, my reaction to their sequential approach is one of considerable sympathy. I would have probably taken a similar tack. Of the two alternative methods contrasted, I also favor their regression mean modeling and "hot deck" residual imputation approach. I prefer this scheme to the predictive mean matching method because the regression mean and residual imputation lends itself directly to linearization variance estimation. To facilitate this linearization approach to proper variance estimation, one requires a probability sampling scheme for selecting and assigning residuals. An algorithm like Brenda Cox's "weighted sequential hot-deck" can be adapted nicely for this purpose. Yes, in spite of claims to the contrary, randomized single imputations combined with proper linearizations can yield valid total variance estimators. Admittedly, this design-based approach is not as easy on data users as multiple imputations.

One thing I do find troubling about regressions mean and residual imputation approaches, is their potential for biasing tail probability estimates when the model R^2 is low. The serum cholesterol models reported here all suffer from this weakness. Faced with the task of estimating the fraction of various subpopulations that exceeded a particular high serum cholesterol value, I would prefer a two-stage imputation scheme that first predicted the likelihood of falling into a particular interval. For this purpose, one could use a polytemous logistic regression model given the other continuous and categorical predictors from the questionnaire and exam.

This also brings up a concern I have with the multiple imputation model. I was intrigued by the joint categorical and continuous data distribution used in the multiple imputation solution. Shafer and company fit the marginal distribution of the categorical data and the conditional distribution of the continuous data given the categories. This formulation achieves a relatively economical form for the categorical distributions given the continuous data. In practice though, even this relatively economical model quickly becomes impractical as the number of polytemous variables multiplies. I believe this difficulty constitutes a serious limitation to the simultaneous imputation approach when most of the important survey variables are polytemous. To overcome this difficulty, in their NHANES application, Shafer and colleagues stoop to treating categorical variables as continuous normal variates. If these categorical questionnaire variables were important survey outcomes, this fix would be less tolerable. The problem with too many categorical variables was exacerbated by including the PSU's or STANDS as fixed effects in the categorical model. I believe this approach to capturing design complexity in terms of fixed effects for the 44 primary sampling units is ill conceived. Primary sampling unit effects are inherently random effects. If one can deal with this source or variation by including STAND among the categorical predictors, what about the second stage sampling units? I believe NHANES uses census block groups as second stage sampling units. These units exert stronger neighborhood effects than the county-level primary units. There are clearly too many block groups in the NHANES sample to treat these second stage sampling units as fixed effects. Naturally, I believe that sample design-based variance estimation is the most practical way to incorporate these clustering effects into the analysis. Otherwise, I believe that pure model-based solutions will be forced ultimately to entertain hierarchical or random effects models. This is still a fairly onerous computational chore. An intermediate approach is to obtaining more matched census variables at the county and block group level. This matching also gives one more auxiliary variables that are useful for missing data prediction.

One final question I had regarding the multiple imputation paper has to do with the degrees-of-freedom formula. My question is, "How can the combined imputation and sampling variance have considerably more degrees-of-freedom than the dominant sampling variance contribution? "The SUDAAN sampling variance for NHANES is based, I believe, on a between PSU within stratum mean square that has at most 22 degrees-of-freedom. Am I correct in assuming that the ***df*** formula presented by Shafer and his coauthors views all design-based variance estimators as having infinite degrees-of-freedom. Given the design variance instabilities alluded to in the paper, I was surprised at the inference drawn regarding these degree-of-freedom values. The authors claim that their large ***df*** number is an indication of good quality estimation for the between imputation variance component. My interpretation of their big ***df*** values is that the erroneous assumption of infinite ***df*** for the sampling term combined with a small relative contribution by the between imputation component leads to a gross overestimate of total degrees-of-freedom.

This brings up the issue of the instability of NHANES design-based variance estimates or associated design effects. The authors acknowledge that this may be due to the small number of NHANES PSUs. This should not be taken as a general indictment of design-based variance estimators. There is a need for a robust degrees-of-freedom estimator for design-based variance approximations. When the degrees-of-freedom estimate is too small for acceptable inference, some combination of modeling design effects and smoothing over variables in needed. Often, in area household samples, I find that one can fall back to more stable variance estimates that account for second and subsequent stages of clustering without suffering much if any downward bias.

My final comments relate to the Ezzati-Rice, et.al recommendation regarding alternative weight adjustments for questionnaire nonresponse. Expanding on this recommendation, I think further consideration should be given to making response propensity weight adjustments for complete examination (unit) nonresponse. Most of the missing exam data is completely missing and therefore the exam data imputations seldom benefit from any within exam predictors. Inverse response propensity weights utilizing all the pre-exam questionnaire variables should achieve most of the bias and variance reduction benefits of imputing selected MEC variables. Simultaneous imputation of missing exam and questionnaire items could still be beneficial.

I have recently been working on generalized raking solutions for response propensities in the context of unit and item nonresponse adjustments. This approach has interesting nonresponse bias reduction features when the data imputation model is faulty. I have also discovered some simple linearization variance estimators for response propensity weighted statistics. In short, I think that unit nonresponse adjustment for completely missing exams deserves further consideration.

In the time remaining, I will outline some of my generalized raking results.

A Generalized Raking Solution for the Questionnaire (Unit) Response Propensity

Definitions:

$s \equiv$ A sample of n units with sampling weights W_i

$X_i \equiv$ $(1, x_i)$ with x_i a p element vector of questionnaire response predictors

$r_i \equiv$ A one-zero questionnaire response indicator

Y_i ≡ A vector of questionnaire outcomes "observed" when $r_i = 1$.

ρ_i ≡ Prob $[r_i = 1 \mid X_i , Y_i]$

= $[1 + \exp (-X_i \beta)]^{-1}$, assumed independent of Y_i given X_i.

Generalized Raking Solution for β

The β vector solution equations have the generalized raking form

$$U_{S1} \equiv \sum_{i \epsilon s} W_i X_i^T [(r_i \div \hat\rho_i) - 1] = \underset{\sim p+1}{\phi} \quad (1)$$

$$= \sum_{i \epsilon s} W_i X_i^T [r_i \hat\alpha_i - (1 - r_i)] = \underset{\sim p+1}{\phi}$$

with

$$\hat\alpha_i \equiv [(1 - \hat\rho_i) \div \hat\rho_i] = exp(-X_i \hat\beta)$$

The associated mean estimator for y_{ti}, an element of Y_i is

$$\bar{y}_{r+\rho} \equiv [\sum_{i \epsilon s} W_i (r_i \div \hat\rho_i) y_{ti}] \div \hat N$$

with

$$\hat N \equiv \sum_{i \epsilon s} W_i = [\sum_{i \epsilon s} W_i (r_i \div \hat\rho_i)]$$

The Associated Imputation Estimator

$$\bar{y}_{r+L} \equiv \sum_{i \epsilon s} W_i [r_i y_{ti} + (1 - r_i) X_i \hat\gamma_{r\alpha}] \div \hat N$$

where the $\hat\gamma_{r\alpha}$ coefficients satisfy

$$U_{S2} \equiv \sum_{i \epsilon s} W_i (r_i \hat\alpha_i) X_i^T (y_{ti} - X_i \hat\gamma_{r\alpha}) = \underset{\sim p+1}{\phi} \quad (2)$$

I have shown (Folsom, 1991) that

$$\bar{y}_{r+\rho} = \bar{y}_{r+L} \quad (3)$$

The Delta Linearization for $\bar{y}_{r+\rho} = \bar{y}_{r+L}$

If we define

$$\tilde{y}_{ti} \equiv [r_i y_{ti} + (1 - r_i) X_i \hat\gamma_{r\alpha}]$$

and

$$\hat L_{ti} \equiv X_i \hat\gamma_{r\alpha}$$

then the linearized variate is

$$Z_{ti} \equiv [(\hat L_{ti} - \bar{y}_{r+\rho}) + (r_i \div \hat\rho_i) \hat e_{ti}] \div \hat N$$

$$= [(\tilde{y}_{ti} - \bar{y}_{r+L}) + (r_i \hat\alpha_i) \hat e_{ti}] \div \hat N$$

where

$$\hat e_{ti} \equiv (y_{ti} - X_i \hat\gamma_{r\alpha})$$

The associate linearization variance estimator has, for simple random sampling, a form reminiscent of the double sampling regression variance estimator. When the squared coefficient of variation in the $(1 \div \hat\rho_i)$ weight adjustments is small, $\bar{y}_{r+\rho}$ achieves variance reduction roughly equivelant to the SRS double sampling regression estimator. Deville and Särndal (1992) present similar results in the context of post-stratification type generalized raking adjustments for sampling error reduction.

REFERENCES

Folsom, R.E. (1991). "Exponential and Logistic Weight Adjustments for Sampling and Nonresponse Error Reduction," Proceedings of the Social Statistics Section of the American Statititistical Association. 197-202, 1991.

Deville, J.C., and Särndal, C.E. (1992). "Calibration Estimation in Survey Sampling," Journal of the American Statistical Association, 87, 376-382.