ANALYZING IMPUTED SURVEY DATA SETS WITH MODEL ASSISTED ESTIMATORS

Robert E. Fay[1]
U.S. Bureau of the Census, Washington, DC 20233-9001

KEY WORDS: *Missing data, sample surveys, multiple imputation, fractionally weighted imputation.*

Abstract Multiple imputation has been previously applied for mass imputation, that is, the imputation from a subsample with complete information to a larger sample. In such applications, the missing data rates are often substantial, such as 80 percent or more, but valid inferences should, in principle, be within reach when probability sampling is used. Yet, limitations of multiple imputation can become severely amplified in this setting (Fay 1994).

This article combines an alternative strategy to multiple imputation, called *fractionally weighted imputation,* with a model assisted approach to estimation. This combination affords several advantages: 1) more efficient model based estimates when the model is true, 2) consistent estimates of the variances for the model based estimates, 3) resistance of the model assisted estimates to model failure, 4) consistence estimates of the variance of the model assisted estimates, 5) estimates of bias for the model based estimates, and 6) consistent estimates of the variance of the bias estimates. With this array of information, analysts will be in an improved position to analyze imputed data sets effectively.

1. INTRODUCTION

In surveys subject to nonresponse, imputation has often been used to complete the data set for purposes of subsequent estimation and other analysis. Rubin (1978, 1987) proposed multiple imputation (MI) as a means to assess the effect on uncertainty associated with estimates using imputed data. In addition, MI typically yields more efficient estimates than single imputation (Rubin 1987)

A second use of MI was suggested: as a strategy of estimation and inference for the problem of double sampling. In this setting, a probability sample or subsample is drawn from a population or larger original sample. Characteristics directly determined only for elements of the subsample are then imputed to the other elements of the original population or sample, generally using covariate information available for all of the elements. Because of the scale of imputation involved, such an approach has also been called *mass imputation.* For example, such an approach was used to impute 1980 industry and occupation codes to public use files from the 1970 decennial census on the basis of a doubly coded sample (Clogg *et. al.* 1991). This approach views the problem of inference to a larger sample or population as one of estimating the missing data for unobserved cases.

Fay (1991, 1992, 1994) showed simple examples in which the MI variance estimator is inconsistent for direct estimators that analysts are likely to use in the analysis of such data sets. In each of the examples, the imputer selects a correct but parsimonious predictive model for the imputation. The analyst attempts to investigate relationships that the imputer did not reflect in the imputation model and obtains from MI an overestimate of the variance. In fact, these problems with MI are particularly pronounced for high missing data rates, such as 80 percent, compared to the consequences for less extreme rates, such as 20 or 30 percent (Fay 1994). Consequently, these findings recommend particular attention to the implications for mass imputation.

In a recent discussion, Binder (1993) characterized the examples as a "blue eyes/brown eyes" problem, in which an imputer omits consideration of eye color in imputing other missing characteristics, and yet an analyst wishes to investigate effects of eye color with the imputed data set. As Fay (1994) illustrated by example, MI inference produces excessively wide confidence intervals for tests of an effect of eye color effect for mass imputation, if no color

effect is actually present. Under the same conditions, *fractionally weighted imputation* (FWI) (Fay 1994) and an extension of the variance estimator proposed by Rao and Shao (1992) provide consistent variance estimates. This combination does not provide a complete solution to the "blue eyes/brown eyes" problem, however, since if a real effect of eye color is in fact present in the population, the bias in the estimated effect would be an important concern.

Meng (1994) recently discussed the "blue eyes/brown eyes" problem from the MI perspective, making a distinction between "congenial" analyses in which the analyst's analysis is compatible with the imputer's assumptions and "uncongenial" analyses on which the counterexamples are based.

This article addresses the "blue eyes/ brown eyes" problem in the context of mass imputation. The solution stems not from an alternative imputation strategy but instead model assisted estimators that work with the imputed values. Basically, the customary analysis of imputed data sets has been model based, since the estimators treats the imputed values as if they were observed. The situation with MI is essentially similar, since the MI estimator averages estimates from multiple separate analyses, each of which employs the imputed values as if they were observed. Consequently, the estimators are potentially biased from model misspecification, even when the probability of inclusion, that is, the sampling and response mechanisms, are entirely known.

An alternative estimation strategy, based on model assisted estimation, represents a more complete solution to the "blue eyes/brown eyes" problem. The general reference for this approach, Särndal, Swensson, and Wretman (1992), has been preceded by a series of related work, including Cassel, Särndal, and Wretman, (1976, 1977), Särndal (1984), and Särndal and Hidiriglou (1989). This article combines the general features of the model assisted approach with extensions of results for the Rao-Shao variance estimator applied to FWI.

Section 2 reviews the Rao-Shao variance estimator for simple random sampling, emphasizing its ability to capture the variances and covariances for the separate components of FWI. Section 3 extends the results to an additional component, imputations for the observed cases, and discusses the class of model assisted estimators that result. Section 4 presents simple Monte Carlo illustrations of the performance of the model assisted estimators compared to model based versions. Section 5 concludes with a summary of the potential implications of this approach and future extensions.

2. RAO AND SHAO: JACKKNIFE VARIANCE ESTIMATION FOR THE HOT DECK

Rao and Shao (1992) modified the standard jackknife variance to account for the variability introduced through use of the hot deck. For simplicity, this section describes the case of a hot deck with a single imputation cell, although the results extend to an arbitrary number of cells. The results also extend to complex multi-stage samples when the sampling variance for estimators applied to complete data are consistently estimated from the standard variance estimators for sampling with replacement at the first stage.

Suppose a simple random sample, y_j , j = 1,...,n, of size n is drawn from an infinite (or extremely large) population. Suppose further that the values of y_j are observed only for a subset of r respondents, $j \in A_r$. The hot deck provides an imputed value, y_j^*, for each nonrespondent $j \in A_m$. Suppose further that the data are missing at random (e.g., Rubin 1978).

The "hot deck" must conform to specified conditions (Rao and Shao 1992). For example, in the simplest case, a single imputation class and a simple random sampling design, imputations are made through simple random sampling with replacement from the donors. For multi-stage stratified sampling, which may lead to differential probabilities of selection and associated weights, the authors consider the selection of "donors" with probabilities proportional to their respective survey weights within the imputation class. Estimates are produced from the singly imputed data set in the normal manner, that is, by using the imputed values as if they were observed for purposes of

estimation. The analysis is modified at the point of variance estimation to reflect the uncertainty due to missing data.

The estimator of the mean may be written:

$$\bar{y}_{(HD)} = \left(\frac{r}{n}\right)\bar{y}_r + \left(1 - \left(\frac{r}{n}\right)\right)\bar{y}^*_{nr} \quad (2.1)$$

where $\bar{y}_r$ is the respondent mean and $\bar{y}^*_{nr}$ is the mean of the imputed values.

The standard jackknife variance formula is:

$$v_{J(1)} = \frac{n-1}{n}\sum_{j=1}^{n}(\bar{y}_{(HD)}(-j) - \bar{y}_{(HD)})^2 \quad (2.2)$$

where

$$\bar{y}_{(HD)}(-j) = \begin{cases} \frac{1}{(n-1)}(n\bar{y}_{(HD)} - y_j) & \text{if } j \in A_r \\ \frac{1}{(n-1)}(n\bar{y}_{(HD)} - y^*_j) & \text{if } j \in A_{nr} \end{cases} \quad (2.3)$$

represents the mean of y computed by omitting observation j. Thus, (2.3) treats imputed values as if they were observed, and may appropriately be called "naive" for doing so. Rao and Shao modify (2.2) and (2.3) by:

$$v_J = \frac{n-1}{n}\sum_{j=1}^{n}(\bar{y}^a_{(HD)}(-j) - \bar{y}_{(HD)})^2 \quad (2.4)$$

where

$$\bar{y}^a_{(HD)}(-j) = \begin{cases} \frac{1}{n-1}[r\bar{y}_r - y_j + \sum_{i \in A_{nr}}(y^*_i + \bar{y}_r(-j) - \bar{y}_r)] & \text{if } j \in A_r \\ \frac{1}{n-1}[r\bar{y}_r + (n-r)\bar{y}^*_{nr} - y^*_j] & \text{if } j \in A_{nr} \end{cases} \quad (2.5)$$

and where $\bar{y}_r(-j) = (r\bar{y}_r - y_j)/(r-1)$. In other words, if $j \in A_{nr}$, then (2.5) is computed in the same way as (2.3), by omitting the imputed value for j. If $j \in A_r$, then y_j is omitted and the imputed values are adjusted to reflect y_j's influence on the mean of the imputed values. Rao and Shao establish the consistency of this variance estimator, both for the single imputation class, as shown, and for multiple classes.

This approach can be extended to *fractionally weighted imputation* (Fay 1994) for $m > 1$. For each $j \in A_{nr}$, m independent selections from the hot deck may be made, giving m imputed values $y^*_{j\ell}$, $\ell = 1,...,m$. Each imputation is given weight $1/m$ times the original survey weight. The weighted observations may then be used in the calculation of (2.1), giving the estimator $\bar{y}_{(FWI)}$. Expressions (2.4) and (2.5) provide a consistent estimator of the variance under the same conditions as the hot deck. For any given m, fractionally weighted imputations generally provide estimates with lower variance than multiple imputation (Fay 1994).

3. MODEL ASSISTED ANALYSIS OF IMPUTED DATA SETS

Consider a domain d for which an analyst requires an estimate of the mean. The usual approach is to compute (2.1) from observed y_j and imputed y^*_j, $j \in A_d$, the observed count in the domain, n_d, and number of respondents, r_d. The Rao-Shao variance estimator (2.4) and (2.5) may be used for the resulting estimator, $\bar{y}_{d(HD)}$, but generally the estimator will be biased unless the means within imputation cells within the domain are the same as the corresponding imputation cell means over all domains. If the choice of imputation cells is highly effective (the implicit assumption of MI as well), the resulting variance estimator is satisfactory for inferences about the domain mean.

Using the same hot deck to compute fractionally weighted imputations, $y^*_{j\ell}$, $\ell = 1,...,m$, for $j \in A_r$, enables application of model assisted estimation. Differences within domain d between the imputed values and the observed values for $j \in A_r$ may be used as an estimate of the bias of the model based imputations. Using the bias estimates either in the form of a difference estimator:

$$\hat{Y}_{ma} = \hat{Y}_{mb} + \left(\frac{1-p}{p}\right)(\hat{Y}_r - \hat{Y}_r^*)$$

or ratio estimator:

$$\bar{y}_{ma} = \bar{y}_{mb} + \left(\frac{n_{nr}}{n}\right)(\bar{y}_r - \bar{y}_r^*)$$

may be employed, where $\bar{y}_{mb}$ represents a model based estimator, such as (2.1) or its FWI extension, and $\bar{y}_{ma}$ represents a model assisted estimator based on it.

4. A COMPARISON OF IMPUTATION PROCEDURES AND VARIANCE ESTIMATORS

Following the examples in Fay (1994), we consider imputation for a problem with two imputation classes, *s* and *t*, and domains, *a* and *b*, cutting across the imputation classes. In the Monte Carlo study, the imputation class variable and tabulation variable were independently drawn from Bernoulli with *prob* = .5, so that they divide the population into 4 cells with equal expected sample sizes. The probability of response, *p*, was set at .2, corresponding to mass imputation.

To illustrate the consequences of bias in imputation, cell means were assigned to vary more substantially with the imputation classes *s* and *t*, than domains, *a* and *b*. The complete assignment of means was 1.5 to the cell (*s, a*), .5 to (*s, b*), -.5 to (*t, a*), and -1.5 to (*t, b*).

Table 1 shows the results, including the average length of confidence intervals, for the grand mean. Mass imputation produces an estimate with considerably lower variance than analysis of the observed data only, and inferences are satisfactory for all alternatives. Table 1 also includes the properties of the estimated bias from the model assisted approach. The confidence intervals are quite short.

Table 2 presents results for an imputation cell. In this case, mass imputation does not improve on the information present in the observed data. Table 3 shows the results for a domain, *a*, whose estimates are affected by bias in the imputation. Only the model assisted estimators and analysis of the observed data only behave acceptably under such conditions -- the coverage of confidence intervals for other estimators is so poor because of bias. Obviously, the results make an extremely strong case for the model assisted perspective under these or similar circumstances.

The last entry in Table 3, showing the performance of the bias estimator, suggests that this may be an effective tool for assessing potential bias in domain estimates.

5. CONCLUDING REMARKS

In applications to mass imputation, model based estimation carries with it the potential effects of model misspecification. The model assisted approach offers the analyst tools to assess the potential effect of error. If the model assisted analysis indicates that the model based estimates have negligible bias, then the simplicity of the model based estimates may favor their continued use. In the presence of significant model bias, however, the model assisted estimates become attractive alternatives.

[1] The author is Senior Mathematical Statistician at the U.S. Bureau of the Census, Washington, DC 20233. This article reports results of research undertaken by a staff member of the Census Bureau. The views expressed are attributable to the author and do not necessarily reflect those of the Census Bureau. The author thanks Gregg Diffendal and Philip Gbur for helpful comments.

REFERENCES

Cassel, C.M., Särndal, C.-E. and Wretman, J.H. (1976), "Some Results on Generalized Difference Estimation and Generalized Regression Estimation for Finite Populations," *Biometrika*, 63, 615-620.

_____ (1977) , *Foundations of Inference in Survey Sampling*, New York: John Wiley.

Clogg, C. C., Rubin, D.B., Schenker, N., Shultz, B., and Weidman, L. (1991), "Multiple Imputation of Industry and Occupation Codes in Census Public-use Samples Using Bayesian Logistic Regression," *Journal of the American Statistical Association*, 86, 68-78.

Fay, R. E. (1990), "VPLX: Variance Estimation for Complex Samples," *Proceedings of the*

for Complex Samples," *Proceedings of the Section on Survey Research Methods,* American Statistical Association, Alexandria, VA, pp. 266-271.

_____ (1991), "A Design-Based Perspective on Missing Data Variance," *Proceedings of the 1991 Annual Research Conference*, Washington, DC: U.S. Bureau of the Census, 429-440.

_____ (1992), "When Are Inferences from Multiple Imputation Valid?" *Proceedings of the Section on Survey Research Methods,* American Statistical Association, Alexandria, VA, pp. 227-232.

_____ (1994), "Valid Inferences from Imputed Survey Data," draft manuscript submitted to the *Journal of the American Statistical Association.*

Meng, X.-L. (1994), "Multiple-Imputation with Uncongenial Sources of Input," to appear in *Statistical Science.*

Rao, J.N.K. and Shao, J. (1992), "Jackknife Variance Estimation with Survey Data Under Hot Deck Imputation," *Biometrika*, **79**, 811-822.

Rubin, D. B. (1978), "Multiple Imputations in Sample Surveys: A Phenomenological Bayesian Approach to Nonresponse," *Proceedings of the Survey Research Methods Section*, Washington, DC: American Statistical Association, pp. 20-34.

_____ (1987), *Multiple Imputation for Nonresponse in Surveys*, New York: John Wiley & Sons.

Särndal, C.-E. (1984), "Design-consistent vs. Model-dependent Estimators for Small Domains", *Journal of the American Statistical Association*, **79**, 624-631.

Särndal, C.-E. and Hidiriglou, M.A. (1989), "Small Domain Estimation: A Conditional Analysis," *Journal of the American Statistical Association*, **84**, 266-275.

Särndal, C.-E., Swensson, B., and Wretman, J. (1992), *Model Assisted Survey Sampling,* New York, Springer-Verlag.

Table 1 Performance of Estimators and Estimators of Their Variances, for the Overall Mean, $p = .2$, $n = 400$, $m = 5$

Estimator	Act. Var.	Est. Var.	% Cov	CI Len
Observed only	.0436	.0446	94.8	.413
FWI	.0221	.0224	95.1	.292
Hot deck	.0242	.0243	94.9	.305
FWI_{ma} (diff. est.)	.0243	.0243	94.9	.305
HD_{ma} (diff. est.)	.0265	.0263	94.9	.317
MI, MI var	.0240	.0233	94.0	.331
Bias (diff. est.)	.0020	.0020	94.9	.086

Table 2 Performance of Estimators and Estimators of Their Variances, for the Mean of an Imputation Class, $p = .2$, $n = 400$, $m = 5$

Estimator	Act. Var.	Est. Var.	% Cov	CI Len
Observed only	.0320	.0326	94.7	.350
FWI	.0329	.0336	94.7	.355
Hot deck	.0370	.0374	94.6	.376
FWI_{ma} (diff. est.)	.0373	.0374	94.0	.375
HD_{ma} (diff. est.)	.0413	.0413	94.2	.395
MI, MI var	.0367	.0357	93.8	.441
Bias (diff. est.)	.0040	.0039	95.3	.121

Table 3 Performance of Estimators and Estimators of Their Variances, for the Mean of a Cross-Class not Used in Imputation, $p = .2$, $n = 400$, $m = 5$

Estimator	Act. Var.	Est. Var.	% Cov	CI Len
Observed only	.0828	.0852	94.6	.569
FWI	.0291	.0290	34.9	.333
Hot deck	.0331	.0329	39.5	.355
FWI_{ma} (diff. est.)	.0470	.0473	94.9	.425
HD_{ma} (diff. est.)	.0509	.0512	95.0	.442
MI, MI var	.0313	.0350	47.3	.392
Bias (diff. est.)	.0139	.0142	94.6	.232