A SIMULATION STUDY TO EVALUATE THE ROBUSTNESS OF RECENT METHODS FOR PREPARING VARIANCE ESTIMATES IN THE PRESENCE OF HOT DECK IMPUTATION

Michael Sinclair, Mathematica Policy Research
Nuria Diaz-Tena, Mathematica Policy Research
Lap-Ming Wun, the Agency for Healthcare Research and Quality[1]

Michael Sinclair, Mathematica Policy Research, 600 Alexander Park, Princeton, NJ 08540

KEY WORDS: Variance Estimation, Hot Deck Imputation, Simulation

I. Introduction

Many large-scale surveys currently use a variety of single imputation methods–as discussed by Chapman (1976), Cox (1980; and Kalton and Kasprzyk (1986)—to handle item nonresponse. Since the use of such imputation increases the underlying variation in the survey results, methods are needed to assess the impact. Until fairly recently, methods to assess the impact of the imputation on the variance have not been available. Rao and Shao (1992) and Shao (2002), presented a method to measure the variance of an estimate due to the combined effect of the sample design and the use of imputation to compensate for item nonresponse. This method discussed in section II, is based on the use of a replication method of variance estimation combined with specific adjustments to the imputed values. Our research sought to explore the use of this method on the expenditure data collected in the Medical Expenditure Panel Survey (MEPS), sponsored by the Agency for Healthcare Research and Quality (AHRQ). Since the MEPS utilizes a single hot deck-imputation method, and the sample design and the data files were structured to facilitate the use of a replicate variance estimation method, the survey met the basic requirements to apply Shao's method. This paper presents results from a simulation study conducted to evaluate the feasibility of applying Shao's procedure to MEPS data[1]. We will discuss, that the MEPS imputation procedures do not meet all of the methodological assumptions given by Shao. In particular, Shao's method assumes the covariates used in the imputation process are fully reported and that variance estimates are needed only for univariate statistics. The goal of our simulations was to quantify the biases in estimates of variance when these assumptions were violated.

[1] The views expressed in this paper are those of the authors and no official endorsement by the Department of Health and Human Services or the Agency for Healthcare Research and Quality is intended or should be inferred. There is no analysis of any MEPS data presented in this paper.

In the MEPS as well as other large-scale surveys, a multivariate missing data pattern exists. This problem is handled by imputing one questionnaire item at a time or in groups in a specified order. Often a variable imputed in a prior step is used to impute another variable in a subsequent step. As a result, the covariates or predictor variables used to impute an item are not necessarily fully observed. Hence, the first goal of the simulation was to evaluate the impact on Shao variance estimates when the imputation process included a covariate previously imputed. We also wanted to evaluate if the bias would vary depending on whether the covariate was imputed in a prior step using either a weak or a strong predictor.

The second goal of the simulation was to quantify the bias of variance estimates when Shao's method is applied on variables developed as the sum of individual components rather than component parts. This is important because MEPS primary analytical expenditure variables are often formulated from the sum of various components. In addition, since Shao's method was designed for univariate statistics, we wanted to determine if when the method was applied to component parts whether a bias would result for a sum or a ratio statistic. Simulation methods are discussed in section III; the simulation findings are presented in section IV, and closing remarks are given in section V.

II. Shao's Method of Variance Estimation Due to Sampling and Imputation

As the starting setup, we consider that a set of survey responses for a single characteristic y is divided into two groups, those that are not imputed, for which we will superscript the y values by NI, and the imputed values by I, as outlined in (1). We will assume that a cell-based imputation method, such as a hot deck imputation scheme, has been applied in a prior step to impute the missing values using k cells reflecting a specified group of Cat_v categories for each variable v used in the imputations, $v = 1..,V$. To apply the method, the procedure begins by computing the overall mean for y for each of the k-imputation

cells limited to the non-imputed values as given in (1).

$$(1)\quad \bar{y}_k^{\,NI} = \sum_{i=1:i\in k}^{n_k^{NI}} w_i \times y_i \Big/ \sum_{i=1}^{n_k^{NI}} w_i$$

The method then assumes that a replication-based/resampling variance estimation procedure is used, such as the jackknife (see Rao, Wu, and Yue 1992), to compute the sampling error associated with the estimated value of the mean or the total population characteristic associated with *y*.

For the jackknife, the sample is divided into *R* mutually exclusive groups by removing groups one at a time from the full sample and forming replicate samples, each containing one group short of the full sample. In the case of stratified samples, groups usually are formed within strata. In stratified clustered designs like the MEPS, the groups can be formed by combining all the observations within one or more PSUs in each PSU strata. We index the replicates by the pair {g,j} where g indexes the stratum *(g = 1,…L),* from which the *j*th group of observations (for example, all those in PSU *j*) were deleted. The variance of the survey estimate of a function of the *y* values, denoted by Θ, such as a mean or sum, is given in (2), where η_g is the number of groups created in stratum *g,* and $\Theta_{g,j}$ is the estimate of Θ from the *(g,j)* replicate after reweighting the survey results (using adjusted survey weights for respondent *i*, $rw_{(g,j),i}$) to account for the omission of the *j*th group from stratum *g.*

$$(2)\quad V(\hat{\Theta}) = \sum_{g=1}^{L} \frac{\eta_g - 1}{\eta_g} \sum_{j=1}^{\eta_g} \left(\hat{\Theta}_{(g,j)} - \hat{\Theta}\right)^2$$

With the replication method or resampling method in place, we compute, similar to (1), the mean value for each *k* cell and replicate combination as given in (3) .

$$(3)\quad \bar{y}_{(g,j),k}^{\,NI} = \sum_{i=1:i\in k}^{n_{(g,j)k}^{NI}} rw_{(g,j),i} \times y_i \Big/ \sum_{i=1}^{n_{(g,j)k}^{NI}} rw_{(g,j),i}$$

To prepare the adjusted variance estimates accounting for the imputation, a set of adjusted imputed values is prepared for each replicate using (4) which adds to each value the difference between the overall mean of the non-imputed values and the similar mean for the replicate for the *k*th cell used to impute the value. The adjusted, imputed values are used in place of the original imputed value to compute the variance from (2) due to both imputation and sampling. The difference in the adjusted variance and the "naïve" variance (considering only the sampling error) reflects the added variation due to the imputation process.

$$(4)\quad \tilde{y}^I_{(g,j),i:i\in k} = y^I_{(g,j),i:i\in k} + \left[\bar{y}^{NI}_{(g,j),k} - \bar{y}^{NI}_k\right]$$

III. A Limited Simulation Study

The first task of our simulation procedures was to generate a population of values that would resemble to some degree the structure of the MEPS health expenditure variables. For this, we created a mock population containing 400,000 person-based records with four expenditure like outcomes that were highly correlated with four person characteristics and to some degree with each other. The variables are described briefly in Table III.1. Table III.2 presents the correlation coefficients among them.

As indicated in Tables III.1 and III.2, we created a total of four survey expenditure-type responses which would be known for the entire population, including Y1, the reported total physician costs per year; Y3, the cost of each physician's visit; Y5, the total dental costs for the year; and Y9, yearly prescription costs. As we structured the artificial data, reported physician costs were dependent on Y3 and X1, the number of physician visits. The variable Y5 was dependent on X4, the number of dental visits. The cost of each physician visit, Y3, is heavily affected by X2, whether the plan is an HMO and, to a much lesser degree, by X3, whether the plan is an employer-sponsored plan. For prescription costs, we made the variable dependent on X1, the number of office visits, and X2, whether the plan was an HMO. We note that we did not attempt to simulate the distribution of actual data values from the MEPS, but merely to generate some data for exploratory purposes that would have an intuitive distribution and relationships for the defined characteristics.

The data and the relationships generated above provided us with a basis for evaluating the desired properties of Shao's variance estimation procedure. The steps that followed included the creation of missing data patterns among the four expenditure items, conducting sampling of the population, and performing various types of hot deck imputations on each of data items in the samples generated To create the missing data patterns we created four new variables each initially equal to the four original variables, and then assigned missing values at random to these variables based on the patterns in the covariates, and in the case of Y1, the missing pattern was also dependent Y3. In this manner, the conditional distribution of the missing responses was missing at random (MAR) given the known covariate values. The missing rates are also presented in Table III.2. In general, we attempted to create a range in the missing rates from a low of 14.7 percent for Y2

(based on Y1) total physician costs to a high of 32.6 percent for Y4 (based on Y3), the cost of each physician's visit.

With the missing data values in place we selected a total of 10,000 stratified random samples of a size of 800 persons each from the population. The samples were stratified based on eight combinations of the values for X1, number of physician office visits and X2, whether the health plan was an HMO. We allocated the sample across each stratum proportional to the stratum's population size; however given each stratum was created to be of equal size in this simulation, this produced a sample size of 100 from each stratum.

For each sample selected, we conducted hot deck imputation on each of the four expenditure items (Y2, Y4, Y6, and Y10) to evaluate the two properties of Shao's variance estimation methodology considered. The imputation methods conducted are presented in Table III.3. For each variable imputation, we used a random hot deck procedure that selected donors (people with non-missing data for the item) to provide replacement values for the missing data persons (the recipients) on a cell basis. In this approach, the cells are based on the missing and non-missing person case's values for the covariates (reported or imputed).

In Table III.3, the imputation methods #1, #2, #3, #6, and #7 all meet the required assumptions of Shao's approach in that the covariates used to form the cells for the imputation process were non-missing for both respondents and nonrespondents. In #4 and #5, since the imputation of Y2 was based on X1 and Y4, which was imputed in #2 and #3, this violated the first assumption under consideration. Finally, we applied the hot deck methodology to Y6, and Y10 to use with Y2 to evaluate the properties of applying the method to the sum of two imputed variables, both on a component basis (applied to Y2 and Y6 separately) and to the sum as a whole (Y2 + Y6, treating the sum as imputed if either component was imputed—reflecting a 32.7 percent missing rate, compared to component missing rates of 14.7 and 26.3 percent, respectively). We also examined, when applied on a component basis, the sum of Y2 and Y10 and the ratios of Y6 to Y2 and Y10 to Y2.

To evaluate the properties of Shao's method, we computed, for each of the 10,000 samples selected, the mean or total estimate for each imputed version of the four variables and their associated, naïve estimate of their sampling variance using a standard stratified textbook variance estimator as available in Proc Survey means under SAS version 8. Likewise, we computed the mean value among the 10,000 samples for the estimates and the variance of these estimates, to provide a comparative set of values that reflected an estimate of the actual variance due to the sampling and imputation procedures. For the complete data variables, the results showed these two sets of values to be within a relative 2 percent of each other for all these variables except Y3, estimated dental costs, which showed a relative difference of about 5 percent. Hence, we felt that the 10,000 simulations were sufficient to detect any noticeable differences from a violation in these simulations. Finally, for all 10,000 samples, we also prepared a standard jackknife estimate and its estimated variance, as well as an adjusted variance estimate using Shao's methodology.

IV. Results

We begin by discussing the properties of the Shao's method when one of the covariates used to impute the item was also imputed in a prior step (and the donors used are allowed to have imputed values for the covariate but not the variable in question). Table IV.1 presents, in rows 1 to 5, a comparison of the mean Shao estimated variance for Y2, Y4, Y6, and Y10 to the simulated values and the mean naïve textbook estimate. For Y4, we also studied the imputation process using both a "good" classing variable (X2, which was correlated with Y4 at .941) and a weak classing variable (X3, which was correlated with Y4 at only .169). The results indicated that the Shao estimator reproduced the simulated variance due to sampling and imputation within 16 percent of simulated values with the ratio of the Shao estimator to the simulated variance ranging from .847 to 1.001.[2] In rows 6 and 7, we compare the mean Shao's estimate of the variance to the simulated variance when Y2 is imputed, using Y4 as the covariate under the two imputation methods for Y4. With the use of imputed covariates, the Shao estimates are within 1 percent of the simulated variance with ratios of .997 (using X2 to impute Y4) and 1.008 (using X4 to impute Y4). Hence, these findings while based on a limited study, indicate that the method has the potential to be quite robust to this violation.

Next, we explored the application of Shao's method to the sum of variables Y2 and Y6, using X1 and Y3 to impute Y2, and X4 for Y6, so that the assumption that the covariates were known for all cases was met. As indicated previously, we examined the application of Shao's methodology using two approaches. First, we applied the methodology to each variable separately, then used the jackknife replication method to determine the

[2]At this time, the authors are uncertain as to why the Shao estimated variance tends to show an underestimate, compared to the simulated values, by up to 16 percent for some of the variables.

variance in the sum. Second, we applied the procedure to the sum of the two variables directly, where the sum was considered imputed if either of the two components was imputed. The results are presented in Table IV.2.

By using a component-based approach, the ratio of the Shao estimate of the variance compared to the simulated variance was similar to the previous results showing a ratio of .950. Similar results hold for the sum of Y2 and Y10 with a ratio of .915. On the other hand, by applying the method directly to the sum, the ratio changes to 3.25. This suggests that applying this method directly to the sum of the two variables could potentially inflate the estimated variance, due to sampling and imputation, and that care should be exercised in attempting to reduce the computational effort by applying the method to aggregate statistics. The results in Table IV.2 also indicate that the impact of the imputation process on the sum is not additive[3] and that the increase in the variance for the sum due to the imputation process is somewhere between that for each of the components. For ratios, the results when applied to the components separately also showed Shao's method to be quite robust with ratios of .935 and .927.

V. Conclusions

Shao's method for estimating the variance due to imputation and sampling has been shown to be a invaluable approach when a single imputation procedure is used, which is common in many large-scale surveys. This study, while limited, shows that the procedure appears to be fairly robust when the assumption that covariates are known for all cases is violated. The results also show that the method yields accurate results when the method is applied to the sum or ratio of two variables, providing the method is applied to the components separately.

References

Chapman, D.W. (1976) "A Survey of Nonresponse Imputation Procedures." Proceedings of the Social Statistics Section of the American Statistical Association, pp. 245-251.

Cox, B.G. (1980) "The Weighted Sequential Hot Deck Imputation Procedure." Proceedings of the Survey Research Methods Section of the American Statistical Association, pp. 721-726.

Kalton, G., and D. Kaspryzk. (1986) "The Treatment of Missing Data," Survey Methodology, pp. 1-16.

Shao, J. (2002) "Replication Methods for Variance Estimation in Complex Surveys with Imputed Data." In Survey Nonresponse, edited by Groves, Dillman, Eltinge, and Little. New York: Wiley, pp. 303-328.

Rao, J.N.K., C.F.J. Wu, and K. Yue. (1992) "Some Recent Work on Resampling Methods for Complex Surveys," Survey Methodology, vol 18, No 2., pp. 209-217.

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

[3]In Table IV.1, we show that the inflation in the variance due to imputation is estimated at 1.31 and at 1.79 for variables Y2 and Y6, respectively. In contrast, the impact of imputation on the sum of Y2 and Y6 is 1.60, which is between the two component effects.

TABLE III.1 MOCK POPULATION: VARIABLES GENERATED

Variable	Description	Generation Method	Values and Frequencies
X1	Number of Physician Visits	Sequential Assignment	4 Values (1-4), 25 Percent Each
X2	Health Plan is HMO	Sequential Assignment within X1	Yes=1, No=2, 50% Yes
X3	Employer Sponsored Plan	Bernoulli Based on X2 X2=1 (HMO) p=.85 X2=2 p=.70	Yes=1, No=2: 22.5 Percent No
X4	Number of Dental Visits	Random Uniform Based on X1 X1=<2: 75% X4=1, 15% X4=2, 10% X4=3 X1>2: 50% X4=1, 25%, X4=2, 25% X4=3	1=62.4 Percent 2=20 Percent 3=17.6 Percent
Y3	Cost of Each Physician Visit	Random Uniform Based on X2 (HMO status) X2=1: 25% Y3=$0, 50% Y3=$5, 25% Y3=$10 X2=2: 33% Y3=$30, 33% Y3=$40, 33% Y3=$50	Mean=$22.51
Y1	Total Physician Visit Costs For Year	Random Exponential Variable with Mean on X1 and Y3	Mean=$112.4 Sum=$ 44,966,580
Y5	Total Dental Costs	Random Exponential with Mean on X4 times $40	Mean=$62.90 Sum=$25,185,670
Y9	Total Prescription Costs For Year	Based on X1 times an initial cost per prescription based on X2 added to random exponential with mean $1.00	Mean=$136.22 Sum=$54,487,268

TABLE III.2 MOCK POPULATION: CORRELATIONS AND MISSING RATES

	Variable (Complete Data)/Variable with Missing Data	Y1/Y2	Y3/Y4	Y5/Y6	Y9/Y10
	Description	Physician Costs	Office Visit Cost	Dental Costs	Total Prescrip.Costs
	Missing Rate	32.6% (High)	14.7% (Low)	26.3% (Moderate)	20.5% (Moderate)
	Correlations				
X1	Number of Physician Office Visits	0.362	0.001	0.097	.371
X2	Health Plan is HMO	0.630	0.941	-0.001	.418
X3	Employer Plan Related to X2	0.114	0.169	0.001	.076
X4	Number of Dental Visits	0.082	-0.001	0.432	.085
Y1	Total Physician Costs	1.000	.669	.035	.452
Y3	Cost of Each Physician Visit	0.669	1.000	-0.001	.394
Y5	Total Dental Costs	0.035	-0.001	1.000	.0377

TABLE III.3 IMPUTATION METHODS

Variable	Steps and Imputation Method Used	Comments
Y2 Total Physician Visit Costs For Year	1. Hot Deck: Using X1 and Y3	Covariate Fully Reported; Ideal for Shao's Method
	4. Hot Deck: Using X1 and Y4, with prior imputation of Y4 based on X2	Violates Assumption that Classing Variables Are Reported (uses imputed version of Y4 instead of Y3)
	5. Hot Deck: Using X1 and Y4, with prior imputation of Y4 based on X3	Same as Above but Uses Weaker Predictor for Y4
Y4 Number of Dental Visits	2. Hot Deck: Using X2 (Strong)	To Evaluate Effect of Violation in Imputation of Y2 in # 4 and #5 Above
	3. Hot Deck: Using X3 (Weak)	
Y6 Total Dental Costs	6: Hot Deck Using X4	To Evaluate Application to Sums and Ratio (Y2 and Y6 and Y2 and Y10)
Y10 Total Prescription Costs	7. Hot Deck Using X1 and X2	To Evaluate Properties of Shao's Method When Applied to Sum (Y2 and Y10) on Component Basis. Also Explored Ratio of Y10/Y2

TABLE IV.1 SIMULATION RESULTS: IMPACT OF IMPUTED COVARIATES

	Variable	Missing Rate (Percent)	Mean Estimate Across 10,000 Samples	Mean of Textbook Estimate of Variance	Mean of Shao's Variance Estimate	Variance Across Simulated Samples	Ratio Actual To Naïve	Ratio Shao to Simulated
Assumptions Meet	1. Y2 Total Physician Visit Costs Per Year (Using X1 and Y3)	14.7	44,937,063	1.506E+12	1.975E+12	1.973E+12	1.31	1.001
	2. Y4 Mean Number of Dental Visits (Using X2 – Good)	32.6	22.52	.049	.098	.108	2.21	.902
	3. Y4 Number of Dental Visits (Using X3 –Fair)		20.66	.249	.383	.452	1.82	.847
	4. Y6 Estimated Total Dental Costs (Using X4)	26.3	25,240,044	1.135E+12	2.031E+12	2.227E12	1.79	.912
	5. Y10, Estimated Total Prescription Costs (Using X1 and X2)	20.5	54,538,566	3.452E+12	6.748E+12	7.548E+12	2.19	.894
Covariates Imputed	6. Y2 Total Physician Visit Costs Per Year (Using X1, Y4 based on X2)	14.7	44,882,221	1.501E+12	2.006E+12	2.012E+12	1.34	.997
	7. Y2 Total Physician Visit Costs Per Year (Using X1, Y4 based on X3)	32.6	43,953,852	1.686E+12	2.135E+12	2.117E+12	1.26	1.008

TABLE IV.2 SIMULATION RESULTS: APPLICATION TO SUM AND RATIO

Variable	Missing Rate Percent)	Mean Estimate Across 5,500 Samples	Mean of Textbook Estimate of Variance	Mean of Shao's Variance Estimate	Variance Across Simulated Samples	Ratio Actual To Naïve	Ratio Shao to Simulated
Y2+Y6 Total Physician Visit and Dental Costs Per Year(Applied Separately)	14.7/23.6	70,177,107	2.640E+12	1.371E+13	4.219E+12	1.60	3.25
Y2+Y6 Total Physician and Dental Costs Per Year(Applied to Sum Directly)	32.7 Jointly			4.006E+12			.950
Y2 + Y10 (Applied Separately)	14.7 and 20.5	99,475,629	4.975E+12	8.758E+12	9.568E+12	1.92	.915
Ratio Y6 to Y2	14.7 and 23.6	.562	.0008	.0013	.0014	1.76	.935
Ratio Y10 to Y2	14.7/20.5	1.214	.0028	.0048	.0052	1.84	.927