Imputation Variance Estimation in Schools and Staffing Survey

Fan Zhang[1], Mike Brick[2], Steve Kaufman[3], Elizabeth Walter[1]

Key Words: Imputation, Variance Estimation, Schools and Staffing Survey

1. Introduction

Missing data is a common problem in virtually all surveys. In cross-sectional surveys, missing data may mean no responses are obtained for a whole unit being surveyed (unit nonresponse), or that responses are obtained for some of the items for a unit but not for other items (item nonresponse). Unit and item nonresponse cause a variety of problems for survey analysts. Missing data can contribute to bias in the estimates and make the analyses harder to conduct and results harder to present.

The most frequently used method to compensate for item nonresponse in National Center for Education Statistics (NCES) surveys is imputation.

In practice, imputed values are often used as true values to estimate the population parameters. However, it is no longer appropriate to use the standard formulae to estimate the variance when there is imputed data. Treating imputed values as observed values can lead to underestimating variances if standard formulae are used. This underestimation may become more appreciable as the proportion of imputed items increases.

Analysts have developed a number of procedures to handle variance estimation of imputed survey data. In particular, Rubin (1987) proposed a multiple imputation procedure to estimate the variance due to imputation by replicating the process a number of times and estimating the between replicate variation. Särndal (1992) outlined a number of model-assisted estimators of variance, while Rao and Shao (1992) proposed a technique that adjusts the imputed values to correct the usual or naive jackknife variance estimator for hot deck imputation. Kaufman (1996) proposed a variance estimation method similar to Särndal's method that can be used with a nearest neighbor imputation approach. Shao and Sitter (1996) proposed to perform an imputation procedure on each bootstrap sub-sample to incorporate the imputation variability. This proposed bootstrap procedure is consistent irrespective of the sampling design, the imputation method, or the type of statistic used in inference. Shao and Sitter's method does not require any model or explicit variance formulae. Once the imputation procedure is programmed appropriately, Shao and Sitter's method is easy to implement. However, since B imputations should be performed for each item, extensive computation is required for large scale surveys. Maintaining the large amount of imputed data can be operationally difficult.

In this study, we applied Shao and Sitter's bootstrap method to the Schools and Staffing Survey (SASS) 1993-94 Public School Teacher Survey component to assess the magnitude of imputation variance.

2. 1993-94 Schools and Staffing Survey (SASS)

SASS 1993-94 Public School Teacher Survey has a two stage stratified sampling design. First, public schools are stratified. Within each stratum, schools are sorted and systematically selected using a probability proportionate to size algorithm. Then within each selected school, teachers are stratified. Within each school and teacher stratum, teachers are selected systematically with equal probability. The SASS 1993-94 Public School Teacher Survey data contains information on the 47,105 public school teachers who responded to the survey. The range of item response rates is 71-100%.

3. SASS 93/94 Imputation Procedure

Four types of imputation methods are used in SASS 1993-94. They are (paraphrasing from Abramson et al., 1996, page 80):

(1) Using data from other items of the same unit on the questionnaire;

[1] Synectics for Management Decisions., Inc.
[2] Westat, Inc.
[3] National Center for Education Statistics

(2) Extracting data from a related component of SASS;
(3) Extracting data from the frame file (the information about the sample case from the sampling frame);
(4) Extracting data from the record for a sample case with similar characteristics ("hot deck ").

Imputation methods (1) – (3) are deductive or logical imputation. Whenever it was possible, a item nonresponse was imputed by methods (1) – (3). If a missing item can not be imputed by methods (1) – (3), then imputation method (4) was used. Method (4) is a (sequential) hot deck method. The procedure started with the specification of imputation classes defined by certain relevant variables (matching variables). Then the records were sorted by STGROUP (Groups of states with similar schools) / STATE / TEALEVEL (Instructional level for teacher) / GRADELEV (Grade levels taught this year) / URB (Type of community where school located) / TEAFIELD (Teaching assignment field) / ENROLMNT (Number of students enrolled in the school). The records were then treated sequentially. A nonmissing y-variable was used as a starting point for the process. If a record had a response for the y-variable, that value replaced the value previously stored for its imputation class. If the record had a missing response, it was assigned the value currently stored for its imputation class. If there was no donor in the class, the class was collapsed with another class.

For imputation method (1), the imputed values are from other observed items of the same unit and in method (3) the imputed values are from the sampling frame file (PSS or CCD). For imputation method (2), the LEA's (Local Education Agency – another component of SASS) missing item is imputed through information from the sampled school which belongs to that LEA. According to Abramson et al. (1996), this type of imputation was performed only to the one-school LEAs. Therefore, the imputed values by methods (1), (2), or (3) are independent of the sample and the sample design. Assume the simplest response mechanism: respondents always respond and nonrespondents never respond. Then if the population is $\{y_1, y_2, ..., y_N\}$, the imputed values can be assumed to be $\{z_1, z_2, ..., z_N\}$. Here if y_k is actually observed, then $z_k = y_k$, otherwise z_k equals the value imputed by any method of (1), (2), or (3). Let $t_y = \sum_{k=1}^{N} y_k$ be the population total of y, $t_z = \sum_{k=1}^{N} z_k$ be the population total of z, and $\hat{t}_z = \sum_s z_k / \pi_k$ be the Horvitz-Thompson estimator of t_z (here π_k is the inclusion probability of unit k). We have the following decomposition

$$MSE(\hat{t}_z) = V(\hat{t}_z) + (t_z - t_y)^2 .$$

The first part, $V(\hat{t}_z)$, can be estimated by treating the imputed values as observed values while the second part is the bias of the imputation and assessing this bias is out of the scope of this study. If the imputation bias is small, then treating the values imputed by any method of (1), (2), or (3) as observed values and using a standard variance estimation formula will not underestimate the variance.

For method (4)—the hot deck imputation, however, the imputed data can not be treated as observed data. Actually every imputed value is a function of the sample, therefore the imputed values cannot be represented as a set of fixed values as $\{z_1, z_2, ..., z_N\}$. Therefore in this study, we investigated the imputation variance of method (4) – the hot deck method.

4. Imputation Variance Estimation Procedure

SASS surveys are designed to produce reliable state estimates, and samples are selected systematically without replacement with large sampling rates within strata. To reflect the increase in precision due to large sampling rates, a without replacement bootstrap variance estimator procedure has been implemented for the 1993-94 SASS. Instead of drawing a simple random sample with replacement from the original sample, the bootstrap is done systematically without replacement with probability proportional to size as the original sampling was performed (Abramson et al., 1996).

In SASS 1993-94 components, 48 replicate weights were created to estimate variance using the bootstrap method. These replicate weights were subjected to various adjustments, including a sampling adjustment, a noninterview adjustment, and a ratio adjustment. In order to reflect these adjustments, these replicate weights should be used in the variance estimation. To this end, we used the Shao and Sitter's method in the following manner:

(1) For each set of replicate weights $\{w_{ik}\}_{k=1,2,...,n}$ (i = 1, 2, ..., 48), cases with $w_{ik} = 0$ are dropped. Denote the remaining cases, which make up a bootstrap sub-sample, as $Y_{1i} = \{y_k : k \in A_{Ri}, \eta_k : k \in A_{Mi}\}$ (i = 1, 2, ..., 48). Here A_{Ri} is the set of observed values and A_{Mi} is the set of missing values.
(2) Apply the same imputation method as was used to create the full sample imputation values and use $\{y_k : k \in A_{Ri}\}$ to impute $\{\eta_{ik}^* : k \in A_{Mi}\}$ (i = 1, 2, ...,

48). This re-imputed bootstrap sub-sample is denoted as s_i. That is

$s_i = \{ y_k : k \in A_{Ri} \} \cup \{ \eta_{ik}^* : k \in A_{Mi} \}$,

here η_{ik}^* is imputed value. The missing values in the full sample are also imputed using the nonmissing values in the full sample. This set of imputed values is denoted as

$s_0 = \{ y_k : k \in A_R \} \cup \{ \eta_k^* : k \in A_M \}$.

Thus, 48 sets of imputed bootstrap sub-samples and 1 set of imputed full sample are obtained.

(3) Calculate the $\hat{\theta}_i$ of interest from s_i, weighted by replicate weights $\{w_{ik}\}$ $(i = 1,...48)$, and the $\hat{\theta}$ from full sample s_0, weighted by the full sample weight $\{w_k\}$. The variance of $\hat{\theta}$ is estimated by

$v(\hat{\theta}) = \frac{1}{48} \sum_{i=1}^{48} (\hat{\theta}_i - \hat{\theta})^2$.

Another difference between the variance estimator we used above and Shao-Sitter's estimator is that in our formula the deviation is around the full sample estimate $\hat{\theta}$ whereas in Shao-Sitter's formula the deviation is around the average of the bootstrap estimates $\bar{\theta}^*$. The balanced repeated replication method (BRR) is implemented in WesVar PC, but the bootstrap method is not. Abramson et al. (1996) suggests that with any BRR software package, the BRR option should be specified for 1993-94 SASS data analysis. The formulae used in WesVar PC for the BRR option is the formula we used above. In general,

$$\sum_{i=1}^{B} (\hat{\theta}_i - \bar{\theta}^*)^2 \le \sum_{i=1}^{B} (\hat{\theta}_i - \hat{\theta})^2 = \frac{1}{B} \sum_{i=1}^{B} (\hat{\theta}_i - \bar{\theta}^*)^2 + (\bar{\theta}^* - \hat{\theta})^2$$

here $\bar{\theta}^* = B^{-1} \sum_{i=1}^{B} \hat{\theta}_i$. Notice

$$E(\bar{\theta}^* - \hat{\theta})^2 = E_P E_B (\bar{\theta}^* - \hat{\theta})^2 .$$

Here E_P is with respect to sample design, E_B is with respect to bootstrap subsampling, and typically $E_B(\bar{\theta}^*) = \hat{\theta}$. Therefore $E_B(\bar{\theta}^* - \hat{\theta})^2 = Var_B(\bar{\theta}^*)$. An unbiased estimator of $Var_B(\bar{\theta}^*)$ is

$$\hat{V}_B(\bar{\theta}^*) = \frac{1}{B} \frac{1}{B-1} \sum_{i=1}^{B} (\hat{\theta}_i - \bar{\theta}^*)^2 .$$

Therefore

$$\frac{1}{B} \sum_{i=1}^{B} (\hat{\theta}_i - \hat{\theta})^2 \approx \left(1 + \frac{1}{B-1}\right) \frac{1}{B} \sum_{i=1}^{B} (\hat{\theta}_i - \bar{\theta}^*)^2 .$$

When B is large the bias in variance estimation is small and can be easily corrected by factor $(B-1)/B$. In our study, we compare standard error estimates instead of variance estimates and $B = 48$, so the adjustment factor is $\sqrt{47/48} \approx 0.99$. We do not apply this adjustment because it is close to 1. In addition, we use the same formula to calculate both the standard error estimates cooperating imputation variance and the standard error estimates without cooperating imputation variance. And the ratio of these two types of standard error estimates is used as the measurement of the difference. Therefore, the adjustment factor has no effect on this ratio.

The variables used for this study include 6 categorical variables and 7 continuous variables. Their stage 2 imputation—method (4), rates range from 2 percent to 25 percent (see table 1).

Most of the variables used for sorting or matching the records are not included in the data file; they had to be reconstructed by using other variables in the data file. This caused a discrepancy between the data imputed for this study and the original imputed data in the file. To prevent confounding the imputation difference with imputation variance, we imputed the full sample with our sorting and matching variables and denote this imputed full sample as s_0. This is the sample used in the variance estimation (see imputation procedure step 3 above).

5. Imputation Variance Estimates

From Table 2 to Table 4, we compare standard errors which do not take the imputation variance into account ($ste(\hat{\theta})$) with the standard errors incorporated with imputation variance ($ste_I(\hat{\theta})$). It is important to emphasize that both $ste_I(\hat{\theta})$ and $ste(\hat{\theta})$ are estimates of standard errors instead of true standard errors and therefore both of them are also subjected to sampling errors.

Table 2 compares standard errors for the total estimator and the average estimator of continuous variables. The output shows the imputation does not inflate the variance for the total very much. For variable T0985, the standard error increases only 7 percent even though the imputation rate is as high as 27 percent. For the average per person estimators of continuous variables, the underlying estimator is actually a nonlinear estimator. When the imputation rate is high, inflation to the variance can be very high, too. For example, variable T0985 now shows $ste_I(\hat{\theta})$ is 41 percent higher than $ste(\hat{\theta})$. So if the imputed data are treated as true values, the underestimation can be severe.

Table 3 compares standard errors for the ratio estimators of continuous variables. Variable BASIC is the ratio of

teacher's basic salary to teacher's total income. Variable INSCH is the ratio of teacher's total income at school to teacher's total income. OUTSCH is the ratio of teacher's total income from outside of school to teacher's total income. ADITION is teacher's other income from school (total income inside school minus base salary) to teacher's total income. IN_OUT is teacher's total income inside school to teacher's total income outside school. Although some variables used for the ratios have high imputation rates (T1440, for example, has a 21.3% imputation rate) the increase in standard errors are very small. Again, for continuous variables, we observed smaller inflation in standard error.

Table 4 compares standard errors for the total estimator and percentage estimator of categorical variables. Here the total estimates are estimated total counts in each category and the percentage is the estimated percent of units in each category. Notice the inflation in variance is larger than the continuous variables. This might be due to the fact that the sample sizes of the categorical variables are smaller (there is more legitimate skipping for these items). It also shows that when imputation rates get higher, the increase in standard errors also gets larger. Now variable T0040 shows the biggest inflation: 2.04.

6. Summary

The techniques developed so far for the variance estimation of imputed data are not yet easy to implement or operationally convenient. Shao and Sitter's method is appealing but requires repeated imputations, so for large scale surveys the data files become too large.

For the deductive imputation methods (1) – (3), the imputed value can be treated as observed value and the use of standard formula should not cause variance underestimation.

Our empirical study shows that using the hot deck imputation method in the 1993-94 SASS can seriously affect the standard error especially for the discrete variables with small sample size.

But notice that the majority of items have very low hot deck imputation rates. For the SASS 1993-94 Public School Teacher component, only 11 out of 249 items had hot deck imputation rates above 10 percent (see Gruber, Rohr, and Fondelier, 1996, figure VIII-24, pp. 231-235). We used six of those items for this study. And, when the imputation rate is low, the inflation in variance is not severe, especially for continuous type variables with large sample size, no matter it is a linear or ratio estimator.

References

Abramson, R., Cole, C., Fondelier, S., Jackson, B., Parmer, R., and Kaufman, S. 1996. *1993-94 Schools and Staffing: Survey Sample Design and Estimation.* (NCES 96-089). Washington, D.C.: U.S. Department of Education, Office of Educational Research and Improvement. National Center for Education Statistics.

Gruber, K., Rohr, C., and Fondelier, S. 1996. *1993-94 Schools and Staffing Survey: Data File User's Manual, Volume I: Survey Documentation.* (NCES 96-142). Washington, D.C.: U.S. Department of Education, Office of Educational Research and Improvement. National Center for Education Statistics.

Kaufman, S. 1996. Estimating the variance in the presence of imputation using a residual. In *1996 Proceedings of the Section on Survey Research Methods* (pp. 423-428). Alexandria, VA: American Statistical Association.

Rao, J. N. K. and Shao, J. 1992. Jackknife variance estimation with survey data under hot deck imputation. *Biometrika* 79(4): 811-822.

Rubin, D. B. 1987. *Multiple Imputation for Nonresponse in Surveys.* New York: J. Wiley & Sons, Inc.

Särndal, C. E. 1992. Methods for estimating the precision of survey estimates when imputation has been used, *Survey Methodology,* 18(2): 241-252.

Shao, J. and Sitter, R. R. 1996. Bootstrap for imputed survey data, *Journal of the American Statistical Association,* 91: 1278-1288.

Table 1: Variables used in this study

Name	Label	Hot Deck imputation rate (%)	Type
T0030	2 Full/Part-time teacher at this school	11.8	5 Categories
T0035	3A Have other assignment at this sch	9.8	Dichotomous
T0040	3B What is other assignment at this sch	24.0	6 Categories
T0140	11D Consecutive yrs teaching since break	5.2	Continuous
T0435	28A Any mathematics courses taken	5.7	Dichotomous
T0645	32B Programs changed views on teaching	2.0	5 Categories
T0860	40B(4) Number of students in the class	13.6	Continuous
T0985	41C Number of separate classes taught	27.0	Continuous
T1420	53B(1) Academic yr base tchng salary	8.3	Continuous
T1430	53B(2) Additional compensation earned	4.0	Continuous
T1440	53B(3) Earning from job outside sch sys	21.3	Continuous
T1455	53B(5) Income earned from other source	5.9	Continuous
T1520	55 Total income of all HHD family member	25.0	12 Categories

Source: Abramson et al. (1996).

Table 2: Standard error comparison for total estimates and average estimates of continuous variables

Name	Hot Deck imputation rate (%)	Total Estimate	$ste_1(\hat{\theta})/ste(\hat{\theta})$	Average Estimate*	$ste_1(\hat{\theta})/ste(\hat{\theta})$
T0140	5.2	8985367	0.99	11.01	0.96
T0860	13.6	24958128	1.01	22.79	1.10
T0985	27.0	2107888	1.07	12.79	1.41
T1420	8.3	86349560396	1.00	33713.26	1.01
T1430	4.0	1865774738	1.03	2093.88	1.05
T1440	21.3	2179435663	1.03	4384.44	1.05
T1455	5.9	588847739	1.01	1676.05	1.03

- These estimates are average per teacher.

Table 3: Standard error comparison for ratio estimates of continuous variables

Name	Hot Deck Imputation rate (%)	Estimate	$ste_1(\hat{\theta})/ste(\hat{\theta})$
Basic	--	0.94907	1.01
Insch	--	0.96957	1.03
Outsch	--	0.02395	1.02
Addition	--	0.02051	1.05
In_out	--	31.87	1.03

Basic = T1420/(T1420 + T1430 + T1440 + T1455)
Insch = (T1420 + T1430)/(T1420 + T1430 + T1440 + T1455)
Outsch=T1440/(T1420 + T1430 + T1440 + T1455)
Addition=T1430/(T1420 + T1430 + T1440 + T1455)
In_out=(T1420 + T1430)/(T1440 + T1455)

Table 4: Standard error comparison for total estimates and percentage estimates of discrete variables

Name	Hot Deck imputation rate (%)	Categories	Total Estimate	$ste_I(\hat{\theta})/\ ste(\hat{\theta})$	Percentage Estimate (%)	$ste_I(\hat{\theta})/\ ste(\hat{\theta})$
T0030	11.8					
		1	12994	1.10	5.61	1.10
		2	31489	1.14	13.60	1.18
		3	97607	1.12	42.15	1.19
		4	52767	1.11	22.79	1.13
		5	36706	1.38	15.85	1.35
T0035	9.8					
		1	54006	1.08	24.45	1.09
		2	166845	1.00	75.55	1.09
T0040	24.0					
		1	9613	1.44	13.49	1.54
		2	11737	2.04	16.47	2.09
		3	5093	1.26	7.15	1.29
		4	12311	1.73	17.28	1.66
		5	26962	1.27	37.84	1.52
		6	5543	1.62	7.78	1.71
T0435	5.7					
		1	2001004	0.99	78.12	0.98
		2	560289	1.00	21.88	0.98
T0645	2.0					
		1	122310	0.99	5.42	0.98
		2	822249	1.01	36.41	1.01
		3	498908	1.00	22.09	1.03
		4	711355	1.01	31.50	1.01
		5	103472	0.98	4.58	0.97
T1520	25.0					
		1	173	1.45	0.01	1.60
		2	863	1.63	0.03	1.68
		3	8850	1.03	0.35	1.04
		4	72952	1.18	2.85	1.15
		5	123771	1.19	4.83	1.22
		6	154036	1.10	6.01	1.12
		7	174850	1.18	6.83	1.16
		8	404821	1.18	15.81	1.30
		9	434259	1.08	16.95	1.14
		10	523142	1.27	20.42	1.26
		11	438739	1.12	17.13	1.19
		12	224836	1.22	8.78	1.21