COMPARING ESTIMATES AND VARIANCES FOR A DATA SET WITH HOT DECK IMPUTATIONS

Núria Díaz-Tena and Frank Potter
Mathematica Policy Research, P.O. Box 2393, Princeton, New Jersey 08543-2393

KEY WORDS: **Item Nonresponse, Imputation, Multiple Imputation, Hot Deck Imputation, Supplemental Security Income**

1. Introduction

The National Survey of SSI Children and Families (NSCF) collected data on children with disabilities and their families who received or applied for Supplemental Security Income (SSI). The survey, sponsored by the Office of Research, Evaluation, and Statistics of the Social Security Administration, had two major objectives:

1. to provide information on the characteristics, experiences, and needs of current SSI child recipients and their families;
2. to evaluate the effects of the Personal Responsibility and Work Opportunity Act of 1996 (P.L. 104-193, otherwise known as the Welfare Reform Act) on SSI children.

The NSCF, administered in 2001-2002, was the first national survey of SSI children in more than 20 years. The survey was intended to fill a critical data need by providing current information on the health and well-being of SSI children and their families.

Mathematica Policy Research, Inc. developed the sampling design for the NSCF questionnaire (Potter et al. 2000); implemented the survey (including data collection); created the recoded variables; developed weights (Diaz-Tena and Potter 2003); computed variance estimation parameters for the survey; and imputed missing values. The recoded variables provided a wide range of information about the applicant's household; from the health of the applicant, the earned and unearned income of the household, to the household members' education levels.

This paper tries to reinforce the importance of imputation by checking the bias introduced by not imputing the missing values, and studies different possible ways of computing the variance of an estimate with imputed values: using Rao & Shao's variance estimation with balance repeated weights and using multiple imputation. Because if special variance estimation techniques are not used, the variance is erroneously underestimated.

We focus our variance estimation comparison to the value of the household car. Section 2 describes hot deck imputation; Section 3 explains the Rao & Shao replicate method to compute variance estimates in the presence of imputed values with a single imputation. Section 4 describes the Rubin multiple imputation procedure to compute variance estimation in the presence of imputed data. Section 5 compares the means and standard errors for the different imputation methods, shows the importance of imputation in the presence of missing data, and the increment in variance if the imputed data are not considered observed.

2. Hot Deck Imputations

In survey research, responding sampled units often fail to answer some survey items, and values are imputed for the missing responses to produce a complete data set. We imputed the value of the first car of the household to compare the different variance estimation methodologies in the presence of imputed data. This value of the car had approximately nine percent missing data.

The missing data were imputed using unweighted hot deck imputation (Ford, 1983). We checked which variables were correlated with the value of the car, and we formed imputation cells with them. The imputation cells were formed with:

- One or more cars in the household
- Value of any other car of the household
- Debt on the cars of the household
- Total household income
- Urban or rural area
- Census region (Northeast, South, Midwest and West).

We categorized these variables and grouped them to form the imputation cells. Three different schemes to form imputation cells were used by changing the categorization of these variables. And the different imputation cells resulted in different imputed values.

We checked the median number of donors, the median ratio of donors to receptors, and the variability of the means of the values of the cars among the donors in the different imputation cells. Results are shown in Table 1 for three different imputation schemes with different size cells. The preferable imputation scheme is Imputation 1 because it has more imputation cells than the other schemes and still enough donors to impute the missing values. In imputation 1, 82 imputation cells were defined, with a median of 18 donors per cell, a median ratio of 4 donors to receptors, and a standard

error of 5,203 across the means of the values of the car in the imputation cells (a large standard error across the imputation cells shows the diversity from cell to cell). Imputation 2 was defined collapsing some of the Imputation 1 cells. We ended up with fewer imputation cells (46), more donors per imputation cell (median of 51), a larger ratio of donors to receptors by imputation cell (median of 5), and less variability among the imputation cells (se of $4,377). And Imputation 3 collapsed more cells from Imputation 2 with the same consequences as before to the imputation cells: fewer imputation cells (25), more donors per imputation cell (median of 87), same ratio between donors and receptors, and less variability between the value of the car among the imputation cells ($1,414).

3. Variance Estimation With Single Imputation

There are different methods of variance estimation when missing data have been imputed as; the model assisted approach introduced by Särndal (1992); the adjusted jackknife variance procedure proposed by Rao and Shao (1992); and the modified balance repeated replication (BRR) by Rao and Shao (1999). What follows is the methodology of the modified balanced repeated replication, and we briefly describe this variant of the BRR method and the procedures used to implement it.

Balance repeated replication can run into problems in estimating a variance of a ratio from stratified sampling designs of totals for small domains. This problem is mainly caused by a sharp perturbation of the weights to construct replicate estimates. Robert Fay proposed a modified method obtained by perturbing the weights by +/- 100ε for the half sample and its complement, where $0 < \varepsilon < 1$ (Dippo, Fay and Morganstein, 1984). We denote this method by BRR(ε).

Judkins (1990) studied by simulation the empirical performance of BRR(ε). He found that BRR(0.5) performs well and is a compromise between the standard BRR and the jackknife. We used ε = 0.5, and Potter et al. (2003) describes all the details of the construction of the 72 balanced repeated replication weights ($w_{i,r}$) applied to the SSI Kids Survey, where *r* denotes the replicate weight *r* = (1,2,3…,72), *R* is the number of replicates weights (72) and *i* is the case. We wanted the number of BRR weights to be smaller than the number of Primary Sampling Units (PSUs) in the sampling design. The sampling design consisted in a stratified cluster design with 75 PSUs.

The following equations (1) to (4) from Rao and Shao, show how to construct estimates for the mean and variance of imputed values, where *k* identifies the imputation cells. Equation (1) defines how to compute R adjusted values $y_{i,k,r}^{adj}$ that depend on the initial value of y_i modified by the difference between the mean of the non imputed (NI) values in that imputation cell *k* weighted by the balance repeated weight *r* and the survey weight.

$$(1) \quad y_{i,k,r}^{adj} = y_i + \left(\bar{y}_{r,k}^{NI} - \bar{y}_k^{NI}\right)$$

Equation (2) averages the adjusted values obtained in equation (1) by imputation cell *k*, averaging all the cases (n_k) in the imputation cell *k* and then averaging all the different means in each one of the K imputation cells. And equation (3) computes the mean of the estimate by averaging all the adjusted values in (2) by each balanced replicate. This mean has the same value obtained as treating the imputed values as observed.

$$(2) \quad \bar{y}_r^{adj} = \sum_{k=1}^{K} \frac{1}{K}\left(\sum_{i=1}^{n_k} \frac{\bar{y}_{i,k,r}^{adj}}{n_k}\right)$$

$$(3) \quad \bar{y}^{adj} = \sum_{r=1}^{R} \frac{\bar{y}_r^{adj}}{R}$$

However, equation (4) provides us with the variance of the mean treating the imputed values as imputed, where η is the number of PSUs.

$$(4) \quad V(\bar{y}^{adj}) = \sum_{r=1}^{R}\left(\frac{\eta-1}{\eta}\right)\left(\frac{\bar{y}_r^{adj} - \bar{y}^{adj}}{R-1}\right)^2$$

Rao and Shao's estimation of the variance takes into account the nonimputed values in each imputation cell; the variance of the estimate depends on the size and the definition of the imputation cell, as shown in equation (5), and its value is larger than if the imputed values would have been treated as observed values.

$$(5) \quad V(y_{i,k,r}^{adj}) = V(y_i) + V\left(\bar{y}_{r,k}^{NI} - \bar{y}_k^{NI}\right)$$

4. Variance Estimation with Multiple Imputation

The multiple imputation technique creates multiple complete data sets (M data sets). The imputations are model based (Rubin, 1987) and we chose to maximize the log-likelihood of the multivariate normal distribution of the variables used to create the imputation cells and the value of the car. We created three complete data sets and estimated the mean of the first car in each data set ($\bar{y}_m$) as shown in equation (6), where $y_{i,m}$ is the value of the car for observation *i* in the m^{th} data set. The grand mean is the mean among the M complete data sets as shown in equation (7).

$$(6) \quad \bar{y}_m = \sum_i \frac{y_{i,m}}{n}$$

(7) $\bar{y} = \sum_m \frac{y_m}{M}$

The variance computation using multiple imputation has the intuitive idea of computing the mean of the variances and the variances of the means (as expressed in equation (8)), or computing the within and between variances (as noted in equation (9)).

(8) $Var(\bar{y}) = \sum_{m=1}^{M} \frac{Var(\bar{y}_m)}{M} + \left(1 + \frac{1}{M}\right) \sum_{m=1}^{M} \frac{(\bar{y}_m - \bar{y})^2}{M-1}$

(9) $Var(\bar{y}) = Var_{Within}(\bar{y}) + Var_{Between}(\bar{y})$

5. Results

We compared differences among the means and standard errors, in Table 2, for the different methodologies using imputed values. We imputed three times using hot deck imputation with the imputation cell described in Imputation 1 (these three imputations, using the same imputations cells are different because we used a different sorting variable to choose the donor, and we denoted those imputations as Hot Deck 1A, Hot Deck 1B and Hot Deck 1C). We imputed one time with Imputation 2 (Hot Deck 2), one time with the Imputation 3 (Hot Deck 3), and we have the three multiple imputations (Multiple 1, Multiple 2, and Multiple 3) obtained using a different random seed.

Table 2 shows: the means of the value of the car, the increment of the mean with imputed values to the mean with the incomplete data set, the standard error computed as if all the observations were observed, the standard error using Rao and Shao variance estimation, and the increment in standard error between the last two standard errors. The data sets with imputed values estimate a mean 4 percent points larger than the mean with missing data. This is an important result indicating why missing data should be imputed to avoid bias.

The variances computed as if the missing values were observed are smaller than the variances computed with the Rao and Shao method. The Rao and Shao's method takes into account the variance of the nonimputed values in the imputation cells used for hot deck imputation. We can observe that the increase in variance is smaller in the Hot Deck 1 imputations (an increase in the standard error using Rao and Shao around 3.7 percent more than if the imputed values are treated as observed), The increase in variance gets larger for Hot Deck 2 (an increase in standard error of 4.48 percent) and even larger for Hot Deck 3 (an increase in standard error of 10.03 percent).

Besides comparing the increase in variances using Rao and Shao methodology and how the increment depends on the selection of the imputation cells, we wanted to see how multiple imputation increases the between-variance term. Table 3 shows: the means (means of the M data sets), the within-variance (the mean of the variances of the M data sets), the total variance using multiple imputation variance estimation (the addition of the within- and the between-variance), and the increase in variance due to the addition of the between variance to the within-variance.

The increase in variance using multiple imputation is small (the increase provided by the between-variance). The increase of the standard error for Hot Deck 1A, 1B and 1C was approximately 3.7 percent using Rao and Shao's variance estimation (Table 2), but it is only a 0.73 percent increment using multiple imputation (table 3). Multiple imputation adds variance with the between variance (the variances of the means of the different complete data sets).

6. Conclusions

Missing data should be imputed to avoid bias.

Using Rao and Shao's variance estimation with balanced repeated weights or using multiple imputation to estimate the variance, increases the variance of the estimate because some values have been imputed. However the increment is computed as: (1) the variation of the nonimputed values in the defined imputation cells created for the hot deck imputation (in Rao and Shao), (2) the difference between the different complete data set means (in Rubin's multiple imputation).

If Hot Deck imputation is used, the Rao and Shao methodology is appropriate to compute the increment in variance. And the definition of imputation cells dictates the increase in variance. How the imputed cells are defined impacts how much the variance increases.

If the data follow some model assumptions, multiple imputation is easy to implement and obtain results for your analysis without being restricted to using balance repeated replication weights. And the variance will increase depending of the variability among the different imputations.

7. References

Diaz-Tena, N., Potter, F. (2003). "Nonresponse Adjustments for a Survey of Children with Disabilities Using Information of a Responsible Adult". *Proceedings of the Section on Survey Research Methods* [CD-ROM]. Alexandria, VA: American Statistical Association, 2003.

Dippo, C. S., Fay, R. E. and Morganstein, D. H. (1984). "Computing variances from complex samples with replicate weights". *Proceedings Survey Research Section of the American Statistical Association* pp 489-94.

Ford, B. L., Madow, W. G, Olkin, I., and Rubin, D. B. (1983). "An Overview of Hot-Deck Procedures in Incomplete Data in Survey Samples. American Press, New York pp 185 – 207.

Judkins, D. R. (1990). "Fay's Method of Variance Estimation". Journal of Official Statistics, vol. 6, no. 3, 223-39.

Potter, F., Jang, d., Friedman, E., Diaz-Tena, N. Gosh, B. (2003). "Comparison of Procedures to Account for Certainty Primary Sampling Units". *Proceedings of the Section on Survey Research Methods* [CD-ROM]. Alexandria, VA: American Statistical Association, 2003.

Potter, F. Mitchell, S. (2000). "Report on Sampling Design and Estimated Survey Cost. Evaluation of the Effects of the 1996 Welfare Reform Legislation on Children with Disabilities: Survey Design and OMB Clearance Package". Mathematica Policy Research, Inc. MPR Reference NO.: 8535-610.

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

Rao, J. N. K., and Shao, J. (1999). Modified balanced repeated replication for complex survey data. *Biometrika*, 86, 403-415.

Rubin, D.B. (1987). "Multiple Imputation for Nonresponse in Surveys". John Wiley & Sons.

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

Table 1. Three Different Imputation Cells to Group Donors and Receptors to use Hot Deck Imputation.

Imputation Cells	Imputation 1	Imputation 2	Imputation 3
Number of Imputation Cells	82	46	25
Median Number of Donors per Imputation Cell	18	51	87
Median Ratio of Donors to Receptors per Imputation Cell	4	5	5
Between SE of Means of the Value of the cars across Imputation Cells	5,203	4,377	1,414

Table 2. Mean and Standard Error for all Data Sets, Standard Errors of the Complete Data Sets Treating the Imputed Values as Observed Values, and Using the Rao and Shao Methodology for the Hot Deck Imputed Data Sets.

	Mean	Increment of Means Among Imputed Versus Incomplete Data Sets	SE (Observed)	SE (Rao and Shao)	SE Increment Between Observed and Rao and Shao
Incomplete Data Set	2,213		91.0		
Hot Deck 1A	2,290	3.47	92.0	95.5	3.76%
Hot Deck 1B	2,309	4.32	92.4	95.8	3.72%
Hot Deck 1C	2,302	4.02	90.5	94.0	3.89%
Hot Deck 2	2,297	3.81	88.3	92.3	4.48%
Hot Deck 3	2,321	4.88	88.7	97.5	10.03%
Multiple 1	2,320	4.83	86.9		
Multiple 2	2,327	5.13	93.6		
Multiple 3	2,313	4.50	87.8		

Table 3. Mean and Standard Error for the Complete Data Sets Treating the Imputed Values as Observed Values (within), and Using Multiple Imputation for Variance Estimation.

	Mean	SE (within)	SE (Rubin)	SE Increase
Multiple 1, 2, 3	2,320	89.5	89.8	0.41
Hot Deck 1A, 1B, 1C	2,300	91.6	92.3	0.73
Hot Deck 1A, 2, 3	2,303	89.7	91.6	2.18