VARIANCE ESTIMATION WITH MISSING BEST VALUES IN THE NIPRCS

Mary Gessley Nixon, Graham Kalton, J. Michael Brick, Westat, Inc.,
Trena Ezzati-Rice, NCHS, Elizabeth R. Zell, CDC
Mary Nixon, Westat, Inc., 1650 Research Blvd., Rockville, MD 20850-3129

Key Words: Hot Deck Imputation, Weighting Adjustments, Replication Variance Estimation

1. Introduction

Weighting adjustments and imputation are widely used in survey research to compensate for missing data. Since these procedures affect the precision of survey estimates, their use needs to be taken into account in variance estimation. In particular, the common practice of treating imputed responses as observed values in variance estimation clearly leads to an overestimation of the precision of the survey estimates. Rao and Shao (1992) and Shao (1993) have recently developed replication variance estimation techniques for computing standard errors for sample means and estimates of population totals when the missing data have been imputed using hot deck methods. Some empirical findings on the use of these and other techniques are presented below.

The objective of this study is to obtain variance estimates for estimates of vaccination levels obtained from the National Immunization Provider Record Check Study (NIPRCS). The NIPRCS determines a sampled child's vaccination status for each of five vaccinations (DTP, Polio, MMR, Hib, and Hep B) based on a combination of the data obtained from household reports and vaccination provider reports. First, vaccination data are collected from household respondents for all children aged 19 to 35 months old in households sampled for the National Health Interview Survey (NHIS). These data are obtained through the Immunization Supplement, which collects data on the number of doses of each of the five vaccinations that the child has received (Peak and Cadell, 1996). The vaccination providers of the children are then contacted in the record check component of the study to obtain provider reports of the numbers of doses received. By comparing and reconciling the responses from these two sources, the 'best values" of the child's vaccination status for the five vaccinations are determined (Ezzati-Rice *et al.*, 1996). Since provider reports are required for determining best values, the best values for children with no provider data are missing.

Both imputation and weighting adjustments are applied to compensate for missing best values, as described in the next section. Section 3 then describes the replication methods used to obtain variance estimates for vaccination coverage estimates from the NIPRCS taking into account both the stratified, multi-stage NHIS sample design and the compensations made for missing data. Section 4 presents the results of this study and the last section presents some summary remarks.

2. Compensating for Missing Best Values

This analysis is based on preliminary data relating to the 1,230 children aged 19 to 35 months old for whom the NHIS Immunization Supplement was completed in Quarters 1 and 2 of 1994. Best values of the number of doses of each of the five vaccinations that a sampled child had received were produced from a combination of household and provider reports for 852 children. Best values for all five vaccinations are missing for 378 children (31 percent) because no provider data were obtained.

The analysis focuses on the proportions of children who have received the recommended numbers of doses for each of the vaccinations separately and in certain combinations. A child who has received the recommended doses is called up-to-date. To be up-to-date, a child needs to have at least 4 doses of DTP, 3 of Polio, 1 of MMR, 3 of Hib, and 3 of Hep B. Combinations of vaccinations that are of particular interest are: at least 4 doses of DTP, 3 of Polio and 1 of MMR, termed 431 up-to-date; 431 up-to-date and at least 3 doses of Hib, termed 4313 up-to-date; and 4313 up-to-date and at least 3 doses of Hep B, termed 43133 up-to-date.

Three alternative methods of dealing with the missing best values in estimating the proportions of children up-to-date for the individual vaccinations and their combinations are applied in the following analyses. One is simply to drop the cases with missing best values from the analysis. Since analyses of NIPRCS data have shown that children with and without best values differ in some systematic ways, the missing best values are not missing completely at random (MCAR) and hence the survey estimates produced by this method are biased. Nevertheless, we apply this method to provide estimates for comparative purposes.

The other two methods that have been used to deal with the missing best values are imputation and weighting adjustments, both of which are widely-used general-purpose methods of compensating for missing survey data. Imputation is commonly used to assign values for item nonresponses, whereas weighting adjustments are generally used to compensate for total nonresponse. Weighting adjustments are rarely appropriate for handling item nonresponses because the varying response patterns across items would lead to a multitude of weights, a feature that causes problems for analyses that involve the relationships between items. In the case of the NIPRCS, however, the best values for the five vaccinations are either all present or all missing. Thus a single weighting adjustment can be applied to compensate for the missing best values for all five vaccinations.

Imputation

The imputation method used for assigning missing best values in this study was a form of hot deck imputation (see, for example, Brick and Kalton, 1996). The NIPRCS sample was divided into imputation classes using auxiliary variables known for both children with and children without best values. Each child without best values was then assigned the full set of best values from a child with best values in the same imputation class.

The imputation classes were formed based on the results of an exploratory analysis to identify auxiliary variables that were good predictors either of the propensity of a child to have best values (response propensity) or of being up-to-date on the 431 combination. A total of 20 imputation classes were formed by cross-classifying the following variables: whether or not the child was up-to-date for the 431 combination according to the household report (up-to-date, not up-to-date, unknown); whether or not immunization records were consulted for the household report; the age of the child (under 2 years old, 2 years old); and the highest level of education of adult household members (12 years or less, more than 12 years).

Within each imputation class, a sample of children with best values (respondents) was selected to serve as donors of best values to children without best values (nonrespondents). If the number of nonrespondents exceeded the number of respondents in an imputation class, all the respondents were chosen once as donors, and a sample of them was selected to donate their best values a second time. The number of times a respondent could serve as a donor was limited to two. If the number of nonrespondents was more than twice as large as the number of respondents in an imputation class, then the search for donors was extended to adjacent imputation classes. The search extended to classes that differed in terms of the highest education of the adult household members. Selection of donors from imputation classes that differed in terms of the other auxiliary variables was not permitted.

Weighting

With the weighting approach, the 378 nonrespondents are dropped from the analysis and the weights of the 852 respondents are increased in compensation. The procedure is carried out within weighting classes, which were defined to be the same 20 classes as used for imputation. Within each class, the weights of the respondents were inflated by the ratio of the sum of the weights for all children in the class (respondents and nonrespondents) to the sum of the weights for the respondents.

The use of weighting adjustments for nonresponse typically results in increased variation in the weights and lower precision of the survey estimates. A useful index of the loss of precision from weighting is $(1+CV^2)$, where CV is the coefficient of variation of the weights (Kish, 1992). This index is reported in the results section.

3. Variance Estimation

The NHIS is a stratified three-stage sample of households (Massey *et al.*, 1989). As a result, methods of variance estimation that are applicable for complex sample designs need to be employed in assessing the precision of the survey estimates. Balanced repeated replications (BRR) and jackknife repeated replications are employed here for this purpose (Wolter, 1985).

For the NIPRCS data, a simplification of the NHIS sample design that treats the design as two PSUs sampled with replacement from 62 "pseudo" strata was used to form the replicates (Parsons, 1990). Two sets of replicates were formed from the pseudo-strata and collapsed PSUs: one set of 64 replicates using BRR, another set of 62 replicates using a jackknife methodology.

The 64 replicate weights for the BRR approach were created by multiplying the full sample weights in a PSU by either 0 or 2 depending on the entries in a 64×64 Hadamard matrix. The BRR estimated variance of the full sample estimate $\hat{y}$ is

$$\mathrm{var}_{brr}(\hat{y}) = \frac{1}{64}\sum_{r=1}^{64}(\hat{y}_{(r)} - \hat{y})^2 , \qquad (3.1)$$

where $\hat{y}_{(r)}$ is the estimate obtained using the BRR replicate weights for replicate r.

With the jackknife approach, the replicate weights are the same as the full sample weights except for the sampled children in the stratum in which the replicate number is equal to the stratum number. In that stratum, the replicate weights for all the children within a PSU are either 0 or 2 times the full sample weights, depending on the random selection of the PSU to retain in the replicate. The jackknife variance estimate of $\hat{y}$ is

$$\mathrm{var}_{jk}(\hat{y}) = \sum_{r=1}^{62} (\hat{y}_{(r)} - \hat{y})^2. \tag{3.2}$$

Three alternative forms of replication variance estimation are employed to take account of the effects of the hot deck imputation, one using a jackknife methodology and the other two using BRR. The method using the jackknife methodology is due to Rao and Shao (1992). With this method, each imputed value in each replicate is adjusted by an amount equal to the difference between the respondent mean for its imputation class in the replicate and in the full sample. The adjustment for each imputed value in imputation class j in replicate r is given by

$$A_{(r)j} = \frac{\sum^{k} w_{(r)jk}\,\delta(jk) y_{jk}}{\sum^{k} w_{(r)jk}\,\delta(jk)} - \frac{\sum^{k} w_{(0)jk}\,\delta(jk) y_{jk}}{\sum^{k} w_{(0)jk}\,\delta(jk)} \tag{3.3}$$

where $w_{(r)jk}$ and $w_{(0)jk}$ are the weights of child k in imputation class j in replicate r and in the full sample, respectively, $\delta(jk) = 1$ if child (jk) is a respondent and $\delta(yk) = 0$ if not, and $y_{(r)jk}$ is the value of the y variable for child (jk). The adjusted value for child jk in replicate r is

$$z_{(r)jk} = y^{*}_{jk} + (1 - \delta(jk))A_{(r)j}. \tag{3.4}$$

where y^{*}_{jk} is the full sample value for child (jk) (either actual or imputed). The replicate estimates $\hat{y}_{(r)}$ are computed using these adjusted values, and the variance of $\hat{y}$ is estimated using equation (3.2). This will be called the Rao-Shao or JK-RS method.

One of the BRR approaches for variance estimation in the presence of imputed data is very similar to the Rao-Shao method outlined above. With this method, replicate imputed values are adjusted using $A_{(r)j}$ in equation (3.3), but using BRR weights. The replicate values are computed using equation (3.4) and employed in the BRR variance formula, equation (3.1). This method, which is proposed by Shao (1993), will be referred to as the BRR-Shao or BRR-S method.

In an alternative BRR approach suggested by Rao (1996), the imputation process is carried out separately for each of the 64 replicates using the same procedures as were employed in the full sample but applied only to children who are in the replicate. The estimated variance is given by equation (3.1), but with $\hat{y}_{(r)}$ denoting the replicate estimate computed with the imputations conducted within the replicate. This method is termed the BRR-Rao or BRR-R method.

Sparseness of the data in some imputation classes within the replicates may cause a problem with some of the methods, resulting in donors being used more frequently and the need to select more donors from adjacent rather than the same imputation classes. This is an issue that should be considered in the design of the imputation classes if these methods of variance estimation are contemplated.

Replication methods of variance estimation are also well suited to reflect the effect of nonresponse weighting class adjustments on the precision of survey estimates. The nonresponse adjustments made to the full sample weights can be reflected in the replicate weights by applying the adjustment process to each replicate separately. As described above, the weighting approach for compensating for missing best values consisted of increasing the weights of respondents to represent the nonrespondents within 20 weighting classes. The same procedure is applied separately within each replicate, using the same 20 weighting classes. The adjusted weights were then used to compute the replicate estimate $\hat{y}_{(r)}$ from the sample of children with best values in the replicate. The BRR variance estimate in equation (3.1) and the jackknife variance estimate in equation (3.2) are then used to measure the precision of the survey estimates. These methods are termed the BRR-W and JK-W methods, respectively.

One further method of variance estimation has also been employed. This method is based on the fact that for univariate analyses a hot deck imputation procedure gives the same estimates as a weighting adjustment procedure that adds the weight of each record with a missing value to its donor record, and drops the records with missing values from the data set (Kalton, 1983). The method involves converting the imputed data set into a reduced data set of respondents, with increased weights from the nonrespondents to whom they served as donors, and then computing the BRR and jackknife variance estimates from this reduced data set. Although the resultant variance estimate is likely to be an underestimate, it may capture much of the effect of imputation on the survey estimates, and hence be a considerable improvement on the common practice of treating the imputed values as if they were reported values. The main attraction of this *ad hoc* procedure, termed here the "donor weighted" method, is the simplicity of its implementation.

4. Results

The methods described in the preceding sections are now applied to the preliminary 1994 Quarter 1 and 2 NIPRCS data. In assessing the results, it should be noted that all of the estimates and variance estimates are subject to appreciable sampling error.

Table 1 presents the estimated percentages of children who are up-to-date for each of the five vaccinations and three combinations using the three different methods of handling the missing best values. Apart from Hep B, the estimates produced by the imputation and weighting adjustment methods are very similar, as expected. The estimates produced by ignoring the children with missing best values (i.e., dropping them from the analysis) are slightly higher than the adjusted estimates. As noted earlier, since the missing best values are not MCAR, the estimates produced by ignoring the children with missing best values are biased. The imputation and weighting methods should reduce the nonresponse bias because they are based on the more reasonable MAR assumption.

Table 2 gives the estimated variances of the imputed estimates using the alternative method of variance estimation described in Section 3. The first column of variance estimates is for the estimates obtained by ignoring the children with missing best values, and is presented for reference purposes. The variance estimates in the second column are obtained by the common practice of treating the imputed values as reported values, a practice that underestimates the variances. The last four columns give the variance estimates obtained by using the alternative methods of accounting for the imputations.

To facilitate the comparison of the variance estimates presented in Table 2, the variance estimates obtained by treating the imputed values as observed values are taken as bases, and the ratios of the variance estimates obtained by other methods to these bases are computed. The results are presented, in percentages, in Table 3. As can be seen from this table, the variance estimates obtained using the JK-RS, BRR-S and BRR-R methods are broadly similar, and are around twice as large as the observed estimates. The donor weighted method produces variance estimates that are around 80 percent larger than the observed estimate on average, but the method may not capture all of the effect of imputation on the precision of the survey estimates.

The estimated variances of the weighted estimates using the standard jackknife and BRR variance estimation procedures are given in Table 4; the variance estimates obtained by ignoring the children with missing best values are repeated for comparison purposes. The JK-W and BRR-W methods result in very similar variance estimates, with little to choose between them. The variances of the weighted estimates are only slightly larger than those obtained by ignoring the children with missing best values.

Finally, we compare the variances of the weighted and imputed estimates. Since the JK-W and BRR-W variance estimates are so similar, only the JK-W variance estimates are considered. These variance estimates are reproduced in Table 5 together with the variance estimates of the imputed estimates obtained using the JK-RS variance estimation procedure, the procedure most consistent with the JK-W procedure. The last column gives the ratios of the JK-RS to the JK-W variance estimates. These ratios show that the variance estimates for the imputed estimates are between 3 percent and 20 percent larger than those for the weighted estimates, with an average of 15 percent larger. The larger variances for the imputed estimates may be explained by the imputation variance resulting from the random selection of donors within imputation classes with hot deck imputation; with weighting adjustments, the increase in weights is spread evenly across all the respondents in the weighting class (Kalton and Kish, 1984).

5. Conclusions

When a relatively large fraction of the observations is missing, as is the case for best values in the NIPRCS, compensation for the missing data may be very important in reducing the effect of nonresponse bias. However, when imputation or weighting adjustments are employed to address the concern about bias, the standard approaches to variance estimation need to be modified to take account of the compensation made. In this study, treating the imputed best values as actual values resulted in the variance estimates of the imputed survey estimates being underestimated by a factor of two, or standard errors underestimated by 40 percent or more.

The methods used for variance estimation of imputed estimates are relatively straightforward but, apart from the donor weighted method, they require substantial computation. The donor weighted method is simple to apply, and it appears to be a marked improvement over the practice of treating imputed values as actual values for variance estimation. However, it is an *ad hoc* procedure that in general will not produce unbiased variance estimates. The Rao-Shao and BRR-Shao methods are currently limited to estimates that can be expressed as estimates of totals and means. The BRR-Rao method can also be applied to other estimates.

The replication variance approach for estimating variances of weighted estimates, with weighting adjustments to compensate for missing data, is straightforward and the variance estimates can be computed using existing replication variance estimation software. Weighting adjustments have the advantage over hot deck imputation of producing more precise survey estimates. Weighting adjustments for missing best values in NIPRCS are readily applied because best values for the five vaccinations are either all present or all missing. When there is a variety of patterns of missingness across the survey variables, imputation is generally a more appropriate strategy. When imputation is used, the imputation classes can be based on auxiliary variables chosen specifically for the variable being imputed. When weighting is used, the weighting classes need to be chosen for all the survey variables taken together.

References

Brick, J.M., and Kalton, G., (1996). Handling missing data in survey research. *Statistical Methods in Medical Research*, **5**, 215-238.

Ezzati-Rice, T., Zell, E., Massey, J., and Nixon, M. (1996). Improving the assessment of vaccination coverage rates with the use of both household and medical provider data. *Proceedings of the Section on Survey Research Methods, American Statistical Association.*

Kalton, G. (1983). *Compensating for missing survey data.* Ann Arbor: Survey Research Center, University of Michigan.

Kalton, G. and Kish, L. (1984). Some efficient random imputation methods. *Communications in Statistics*, **13**, 1919-1939.

Kish, L. (1992). Weighting for unequal P_i. *Journal of Official Statistics*, **8**, 183-200.

Massey, J.T., Moore, T.F., Parsons, V.L., and Tadros, W. (1989). *Design and estimation for the National Health Interview Survey, 1985-1994.* National Center for Health Statistics, Vital Health Stat **2** (110).

Parsons, V.L. (1990). Analytic limitations to current national health surveys. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 736-741.

Peak, R., and Cadell, D. (1996). Overview of the National Immunization Provider Record Check Study. *Proceedings of the Section on Survey Research Methods, American Statistical Association.*

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. (1996). Personal communication.

Shao, J. (1993). Balanced repeated replication. *Proceedings of the Section on Survey Research Methods, American Statistical Association,* 544-549.

Wolter, K.M. (1985). *Introduction to Variance Estimation.* New York: Springer-Verlag.

Table 1. Estimated percent of children up-to-date using imputation and weighting methods

Vaccination	Adjustment method		
	Ignore Missing	Imputation	Weighting
DTP	77.3%	76.3%	76.1%
Polio	84.8	83.6	83.9
MMR	89.4	88.6	88.8
Hib	91.1	90.4	90.6
Hep B	19.2	18.1	18.7
431	75.7	74.8	74.5
4313	74.2	73.2	73.1
43133	15.8	14.8	15.3
n	852	1,230	852

Table 2. Estimated variances of imputed estimates

Vaccination	Method of variance estimation					
	Ignore	Observed	JK-RS	BRR-S	BRR-R	Donor Weighted
DTP	2.02	1.30	2.22	2.66	2.59	2.43
Polio	1.61	1.04	1.93	2.43	1.72	1.99
MMR	1.51	0.66	1.69	2.19	2.00	1.64
Hib	1.51	0.96	1.74	1.88	1.90	1.30
Hep B	2.25	1.61	3.03	3.03	3.08	2.19
431	1.99	1.21	2.19	2.62	2.70	2.31
4313	2.13	1.21	2.28	2.69	2.75	2.31
43133	2.34	1.37	2.66	2.66	3.09	2.04
$1+cv^2$	1.15	1.14	1.14	1.14	1.14	1.21
n	852	1,230	1,230	1,230	1,230	852

Table 3. Estimated variances relative to imputed as observed

Vaccination	Observed	JK-RS	BRR-S	BRR-R	Donor Weighted
DTP	100	171	204	199	187
Polio	100	186	234	165	191
MMR	100	258	334	304	250
Hib	100	181	195	198	135
Hep B	100	188	188	191	136
431	100	181	217	223	191
4313	100	188	222	227	191
43133	100	194	194	225	149
Avg. %	100	193	224	217	179

Table 4. Estimated variances of weighted estimates

Vaccination	Method of variance estimation		
	Ignore	JK-W	BRR-W
DTP	2.02	1.93	2.07
Polio	1.61	1.66	1.72
MMR	1.51	1.64	1.74
Hib	1.51	1.51	1.56
Hep B	2.25	2.53	2.50
431	1.99	1.88	2.02
4313	2.13	1.93	2.04
43133	2.34	2.40	2.43
Mean	100	101	105
$1+cv^2$	1.15	1.17	1.17

Table 5. Estimated variances of weighted and imputed estimates

Vaccination	Weighting JK-W (A)	Imputation JK-RS (B)	Ratio (A)/(B)
DTP	1.93	2.22	1.15
Polio	1.66	1.93	1.16
MMR	1.64	1.69	1.03
Hib	1.51	1.74	1.15
Hep B	2.53	3.03	1.20
431	1.88	2.19	1.17
4313	1.93	2.28	1.18
43133	2.40	2.66	1.11