A SIMULATION STUDY OF TWO METHODS OF VARIANCE ESTIMATION WITH HOT DECK IMPUTATION

Michael E. Jones, J. Michael Brick, Graham Kalton, Westat; Richard Valliant, University of Michigan
Michael E. Jones, Westat, 1650 Research Boulevard, Rockville, Maryland 20850

KEY WORDS: Model-Assisted, Adjusted Jackknife

1. Introduction

In survey research, responding sampled units often fail to answer some survey items, and values are imputed for the missing responses to produce a completed data set. When the imputed values are treated as observed values, variance estimates understate the true variances of the survey estimates. The biases may be substantial, even when the rate of missingness is low. Several different methods have been proposed for computing variances that account for imputation.

This paper examines two variance estimation methods that are appropriate when a single value is imputed for each missing item using a hot deck imputation scheme. The two methods—a model-assisted method and an adjusted jackknife method—are described briefly in the next section. The third section describes the design of a simulation study that was conducted to evaluate the estimated variances and confidence intervals produced by the two methods. The imputations were performed using both an unweighted and a weighted hot deck scheme. The estimates are totals of the form $\hat{\theta} = \Sigma w_i y_i$, where w_i is the weight and y_i is the observed value of the characteristic for sampled unit i. To evaluate the variance estimation methods in a realistic setting, we use a real data set and a typical disproportionate stratified simple random sample design for that data set. The missing data and imputation procedures are simulated using three different structures. The imputation procedure is consistent with the missing data mechanism for one of the structures, but not for the other two.

The fourth section gives the results of the simulations for the two methods and shows how their variance estimates differ from those computed ignoring the fact that some of the missing values are imputed. The final section gives some concluding remarks on the methods and their applicability in this type of setting.

2. Description of the Methods

The first method of variance estimation examined in this paper uses the model-assisted approach introduced by Särndal (1992). The second method is the adjusted jackknife variance estimation procedure proposed by Rao and Shao (1992). Below, we briefly describe each of these methods and the procedures used to implement them.

2.1 Model-Assisted Variance Estimation

The model-assisted approach with hot deck imputation assumes that the data are missing at random (MAR) in the hot deck cells and that a model for the distribution of the y's holds. With hot deck imputation, an appropriate model is that the y_i's are independent and identically distributed (*iid*) random variables within the hot deck cells, e.g., for cell g, $y_{gi} \overset{iid}{} (\mu_g, \sigma_g^2)$. The difference between the imputed estimate, $\hat{\theta}_I$, and the finite population parameter, θ_N, can be decomposed into $\hat{\theta}_I - \theta_N = (\hat{\theta}_n - \theta_N) + (\hat{\theta}_I - \hat{\theta}_n)$, where $\hat{\theta}_n$ is the usual inverse probability weighted estimator of θ_N with complete response. Let the subscripts ξ, p, R, and I denote operations made with respect to the assumed cell mean superpopulation model (ξ), the probability sample design (p), the random response mechanism (R), and the imputation method (I). Under the cell mean model, both unweighted and weighted versions of the hot deck are valid, and we examine both versions. With the unweighted hot deck, a donor is selected with equal probability from responding units in the same cell; with the weighted hot deck, a donor is selected with probability proportional to its sampling weight. See Brick et al. (2002).

Särndal (1992) expresses the total variance for the imputed estimator as

$$\begin{aligned} V_{TOT} &= E_\xi E_p E_R E_I (\hat{\theta}_I - \theta_N)^2 \\ &= V_{SAM} + V_{IMP} + 2V_{MIX} \end{aligned} \qquad (1)$$

where the three components are

$$V_{SAM} = E_\xi E_p (\hat{\theta}_n - \theta_N)^2,$$
$$V_{IMP} = E_\xi E_p E_R E_I (\hat{\theta}_I - \hat{\theta}_n)^2, \quad \text{and}$$
$$V_{MIX} = E_\xi E_p E_R E_I [(\hat{\theta}_I - \hat{\theta}_n)(\hat{\theta}_n - \theta_N)].$$

In this study the values of the three components are evaluated using the conditional method suggested by Brick et al. (2002). The estimator for V_{SAM} is calculated as the standard variance estimator using the imputed values as if they were reported values. We call this the naïve variance estimator. Brick et al. (2002) show that under the cell mean model and hot deck imputation, the bias of the naïve variance estimator as an estimator for V_{SAM} is small if no

donor is used too often. With stratified simple random sampling, $\hat{V}_{SAM} = \Sigma n_h^{-1} N_h^2 (1-f_h)\tilde{s}_h^2$, where $f_h = n_h / N_h$, $\tilde{s}_h^2 = \Sigma(\tilde{y}_{hi} - \bar{\tilde{y}}_h)^2 (n_h - 1)^{-1}$, $\bar{\tilde{y}}_h = \Sigma \tilde{y}_{hi} / n_h$, and $\tilde{y}_{hi}$ is the observed value if unit *i* responds or the imputed value if the unit does not respond to the item. An unbiased estimator (under the model) for V_{IMP} is

$$\hat{V}_{IMP} = 2 \sum_{g=1}^{G} [\sum_{i \in A_{M_g}} w_i^2 + \sum \sum_{\substack{i<j \\ i,j \in A_{M_g}}} w_i w_j \gamma_{ij}] \hat{\sigma}_g^2 ,$$

where *g* denotes the hot deck cell, $\hat{\sigma}_g^2$ denotes an estimate of the unit variance in cell *g*, A_{M_g} is the set of nonrespondents in cell *g*, and $\gamma_{ij} = 1$ if nonrespondents *i* and *j* have the same donor and $\gamma_{ij} = 0$ otherwise. By definition, $\gamma_{ii} = 1$. An unbiased estimator for V_{MIX} is

$$\hat{V}_{MIX} = \sum_{g=1}^{G} [\sum_{i \in A_{R_g}} \sum_{j \in A_{M_g}} w_i w_j d_{ij} - \sum_{j \in A_{R_g}} w_j^2] \hat{\sigma}_g^2 ,$$

where A_{R_g} is the set of respondents, and $d_{ij} = 1$ if respondent *i* is the donor for nonrespondent *j* and $d_{ij} = 0$ otherwise. Putting the components together, the model-assisted variance estimator is

$$\hat{V}_{MA} = \hat{V}_{SAM} + \hat{V}_{IMP} + 2\hat{V}_{MIX} .$$

Either weighted or unweighted estimates of σ_g^2 could be used in the model-assisted variance estimate. We computed variance estimates with both estimators of σ_g^2 for the weighted hot deck imputation. The weighted estimate is

$$\hat{\sigma}_g^2 = n_{Rg} (n_{Rg} - 1)^{-1} \sum_{i \in A_{Rg}} w_i (y_i - \bar{y}_{Rg})^2 / \sum_{i \in A_{Rg}} w_i ,$$

where $\bar{y}_{Rg}$ is the weighted mean for respondents in cell *g*. The unweighted estimate of σ_g^2 used the same formula with w_i set equal to 1. Since the simulations showed that the weighted and unweighted estimates produced nearly identical results, we present only the results using the weighted estimate of σ_g^2.

2.2 Adjusted Jackknife Variance Estimation

The adjusted jackknife method was developed using a model that makes assumptions about the response mechanism without the distributional assumptions of the model-assisted method. The adjusted jackknife model assumes a uniform response probability model within each hot deck cell. Under this model, the theory for the adjusted jackknife requires the use of a weighted hot deck to provide unbiased estimates. After describing the method we discuss an extension to a superpopulation model.

The standard jackknife variance estimator for a complete sample is

$$\hat{V}(\hat{\theta}_n) = \sum_{k=1}^{K} c_k (\hat{\theta}_n^{(k)} - \hat{\theta}_n)^2 , \qquad (2)$$

where $\hat{\theta}_n^{(k)}$ is the estimate of θ_N based on the observations included in the *k*-th replicate, *K* is the number of replicates and c_k is a factor associated with replicate *k*. In a stratified design with *L* strata, equation (2) can be re-written as

$$\hat{V}(\hat{\theta}_n) = \sum_{h=1}^{L} \sum_{k=1}^{n_h} (1-f_h)(\frac{n_h - 1}{n_h})(\hat{\theta}_{h_n}^{(k)} - \hat{\theta}_n)^2 , \qquad (3)$$

where n_h is the number of sample units in variance stratum *h*, $f_h = n_h / N_h$ and $\hat{\theta}_{h_n}^{(k)}$ is the estimate for replicate *k* in stratum *h* when unit *k* is deleted and the other units are given replicate weights of $w_i^{(k)} = n_h (n_h - 1)^{-1} w_i$.

The stratified jackknife procedure in (3) requires the computation of Σn_h replicate estimates $\hat{\theta}_{h_n}^{(k)}$. A commonly used strategy to reduce the number of computations is to combine units into computational or variance strata. Let h^* denote a combined variance stratum and *r* a group of sample units within the combined stratum. Then, the grouped jackknife variance estimator is

$$\hat{V}(\hat{\theta}_n) = \sum_{h^*} \sum_{r} (1-f_{h^*}) \frac{n_{h^*(r)}}{n_{h^*}} (\hat{\theta}_{h^*}^{(r)} - \hat{\theta}_n)^2 , \qquad (4)$$

where

n_{h^*} = the number of sample units in combined variance stratum h^*;

$n_{h^*(r)}$ = the number of units retained in combined variance stratum h^* when units in group *r* are deleted;

f_{h^*} = n_{h^*} / N_{h^*}, where N_{h^*} is the number of population units in h^*; and

$\hat{\theta}_{h^*}^{(r)}$ = the estimate for the full population when units in group *r* in combined variance stratum h^* are deleted.

The retained units are given replicate weights of $w_i^{(r)} = n_{h^*}(n_{h^*(r)})^{-1} w_i$. As described later, a grouped jackknife variance estimator was used in the simulations. We verified that this grouped jackknife variance procedure gave essentially the same average estimates as the ungrouped jackknife in the case of complete response.

Rao and Shao (1992) proposed the adjusted jackknife variance estimator as a modification of (2) to handle the situation in which some of the data have been imputed. This adjusted estimator is

$$\hat{V}_{RS} = \sum_{k=1}^{K} c_k (\hat{\theta}_I^{(k)} - \hat{\theta}_I)^2 ,$$

where

$$\hat{\theta}_I^{(k)} = \sum_{g=1}^{G} [\sum_{i \in A_{Rg}} w_i^{(k)} y_i + \sum_{j \in A_{Mg}} w_j^{(k)} (\hat{y}_j^{(k)} + \hat{e}_j^*)] ,$$

and G is the number of hot deck cells. The residual $\hat{e}_j^*$ is described in Kim et al. (2002) and is a result of writing an imputed value as a predicted value plus a residual term. This variance estimator can easily be computed using equation (4) rather than equation (2).

In developing the theory of the adjusted jackknife, Rao and Shao (1992) assume the finite population factor (*fpc*) is ignorable. In the simulations, we modified their variance estimator to include the *fpc* term shown in equation (3).

3. Sample Design for Simulation Study

To describe the simulation study, we begin by presenting the sampling frame and the design used to select the sample of units. The methods for creating missing values for the sampled units and imputing for the missing values are then described.

The sampling frame for the simulations is a subset of the file of 14,571 public school districts extracted from the 1999-2000 Common Core of Data (CCD), a file of all the school districts in the 50 states and the District of Columbia. Districts that had missing data, that were not "regular" districts (e.g., administrative districts), and that had no enrolled children were excluded from the sampling frame. We also excluded 25 districts with the largest enrollments because in most studies these districts would be sampled with certainty. The final frame consists of 11,941 districts.

The sample design used in the simulation is a stratified simple random sample of 1,020 school districts. Twelve strata were created using number of students (district size), and percentage of students at or below the poverty level (poverty status) as stratification factors. The 12 strata were created by the cross classification of four district size categories (1 to 4 in increasing size) and three poverty status categories (1 to 3 in increasing percentage in poverty). The strata and number of districts in the frame are given in Table 1. The table also gives the strata sample sizes and sampling rates used in the simulation. The sampling rates increase as district sizes increase, which is typical of many school district sample designs. The sampling rates in some strata are over 10 percent, and at this rate finite population correction factors begin to be important in estimating variances. Table 1 also contains the strata means and variances for the number of administrators in the district, *y*, which is the variable of interest in the simulations.

Table 1. Strata definitions, sampling rates, and strata population statistics

Stratum	District size	Poverty status	N_h	n_h	Sampling rate	Number of administrators	
						Mean	Variance
1	1	1	615	32	0.0520	1.02	0.46
2	1	2	1,147	59	0.0514	1.05	0.48
3	1	3	1,292	66	0.0511	0.99	0.53
4	2	1	1,720	111	0.0645	2.59	3.10
5	2	2	2,305	149	0.0646	2.35	2.60
6	2	3	1,893	122	0.0644	2.63	4.33
7	3	1	692	75	0.1084	5.34	16.12
8	3	2	579	63	0.1088	5.54	11.45
9	3	3	527	57	0.1082	5.93	13.95
10	4	1	342	83	0.2427	12.05	175.73
11	4	2	449	110	0.2450	13.27	332.71
12	4	3	380	93	0.2447	13.64	370.26
Total			11,941	1,020		3.62	44.78

By construction, information on the number of administrators is available for all districts in the sampling frame. To create missing values, we assigned missing values to sampled units within what we term "response cells". In some cases the response cells are the sampling strata, termed STR cells, whereas in other cases they are what are termed HD cells, as described below. Within a given response cell, units were assigned at random to missing or nonmissing at a specified rate. For each type of response cell, five schemes for assigning rates of missingness were chosen. In three of the schemes, the rates of missingness varied across the response cells, whereas in the other two schemes the rates were constant across the cells.

The HD cells were defined by the cross classification of four geographic regions and a fourfold categorization of the number of full time equivalent teachers in the district. The HD cells are somewhat correlated with the sampling strata, but each cell contains units from more than one stratum. The 16 HD cells are displayed in Table 2, along with the population mean and variance of the number of administrators per district in each cell.

Table 2. Definition of the HD cells, together with the population mean and variance of the number of administrators per district for each cell

HD cell	Region	Categorized number of teachers	Number of administrators	
			Mean	Variance
1	NE	1	1.63	0.62
2	SE	1	1.82	1.70
3	C	1	1.36	0.75
4	W	1	1.05	0.76
5	NE	2	2.47	1.14
6	SE	2	4.25	5.62
7	C	2	2.81	3.45
8	W	2	2.09	2.15
9	NE	3	4.33	7.37
10	SE	3	7.50	16.05
11	C	3	6.34	20.85
12	W	3	3.79	6.37
13	NE	4	14.48	215.73
14	SE	4	20.99	435.91
15	C	4	20.90	712.97
16	W	4	9.92	272.19

The simulations were conducted by first drawing a stratified simple random sample as described above. Once the sample was selected, response status (respondent/nonrespondent) was randomly assigned to each sampled unit according to the given response scheme. A standard hot deck imputation procedure was then used to assign donor values to any missing values (unweighted or weighted for the model-assisted approach, and only weighted for the adjusted jackknife approach). The estimated total number of administrators was computed for this simulated sample with imputed values, and variance estimates were computed for this estimated total using both the model-assisted and the adjusted jackknife variance estimation methods. The adjusted jackknife method was based on three combined variance strata and 40 groups of units per stratum for a total of 120 replicates. This process was repeated 10,000 times for each response scheme.

The cells for the hot deck imputation were defined in two alternative ways. One way defined them to be the strata (STR) given in Table 1 and the other define them to be the HD cells given in Table 2. In the simulation we examined three combinations of response mechanism (STR or HD cells) and hot deck cell formation (STR or HD cells) We refer to these combinations as STR/STR, HD/HD, and STR/HD, where the first set of letters identifies the response mechanism and the second set identifies the type of hot deck cell. The first two columns of Table 3 show the three combinations with the five sets of response rates, which together generate 15 separate simulation studies. The third column of the table gives the empirical mean response rate from the 10,000 simulation runs.

As indicated earlier, the key assumptions of the model-assisted method are that the data are MAR and the units within a hot deck cell are independent and identically distributed. To evaluate the applicability of the model assumptions, the estimation strategy must be considered. With a stratified sample, the strata means and variances are estimated and then aggregated over the strata. Thus, to satisfy the model assumptions, the value imputed for a missing unit in hot deck cell g in stratum h must have mean $\mu_g = \mu_h$ and variance $\sigma_g^2 = \sigma_h^2$.

Clearly, the STR/STR combination satisfies that condition. On the other hand, the HD/HD combination does not satisfy these conditions. The donors may be from a different stratum than the missing unit. It is worth noting that the HD/HD combination is consistent with a uniform response probability model as assessed with the adjusted jackknife method. Furthermore, the weighted hot deck ensures that the estimated total over all possible imputations is unbiased under this response probability model. Finally, the STR/HD combination does not satisfy the MAR assumption (the response probability depends upon the stratum and this is not included in the imputation procedure). For this combination, the assumptions for neither the cell mean model nor the uniform response probability model are satisfied.

Before presenting the results of the simulations, we define the statistics used to evaluate the two variance estimation methods, and also the naïve variance estimator. The imputed estimate of the total number of administrators in the population is $\hat{\theta}_I = \Sigma N_h \bar{y}_{hI}$, where $\bar{y}_{hI} = n_h^{-1} \Sigma \tilde{y}_{hi}$. The bias of $\hat{\theta}_I$ is estimated by $bias(\hat{\theta}_I) = \Sigma_s (\hat{\theta}_{Is} - \theta_N)/10{,}000$, where $\hat{\theta}_{Is}$ is the estimate

from sample s and θ_N is the known population total from the frame. The relative bias, relative to the population total, is estimated by $relbias(\hat{\theta}_I) = bias(\hat{\theta}_I)/\theta_N$.

The mean square error of $\hat{\theta}_I$ is estimated by $MSE(\hat{\theta}_I) = \Sigma_s(\hat{\theta}_{Is} - \theta_N)^2/10{,}000$. If the cell mean model holds, then $relbias(\hat{\theta}_I)$ should be close to zero, and the $MSE(\hat{\theta}_I)$ should be nearly equal to the variance of $\hat{\theta}_I$. Since a major use of variance estimates is for producing confidence intervals, we also calculated nominal 95 percent confidence intervals around the estimated total for each of the 10,000 simulations for each response mechanism and we evaluated the coverage properties of these intervals.

4. Simulation Results

Table 3 shows the main simulation results when missing values are assigned using the weighted hot deck imputation scheme that is assumed in the theory of the adjusted jackknife. Based on the 10,000 draws for each simulation, the table gives the mean response rate, the relative bias of the imputed estimator, the estimated *MSE*, the averages of the three variance estimates (naïve, model-assisted and adjusted jackknife) as percentages of the *MSE*, and the confidence interval (CI) coverage rates for the estimates. The results are given for each response mechanism and simulation scheme described above.

We begin by examining the relative bias of the imputed estimates in Table 3. The imputed estimates are unbiased estimates under the STR/STR and the HD/HD schemes because the response mechanism assignments and imputation cell assignments are the same and the probability of selecting donors is proportional to their weights. The estimates are also unbiased when the response rates are constant across strata in the STR/HD scheme for the same reason. However, the imputed estimates are biased when the response rates vary by stratum and the imputation cells are defined by the hot deck cells. The relative bias is not large, but biases affect confidence interval coverage rates and the relationship between variance and *MSE*. Since all the theories of variance estimation with imputed data assume unbiased estimates, the properties of the variance estimates in this situation are of particular interest.

Table 3. Relative bias, variance estimates and mean square error of imputed estimates, the percent of alternative variance estimates of the mean square error and confident interval coverage with weighted hot deck imputed estimates, by simulation scheme

Scheme	Mean response rate	Rel-bias (%)	*MSE* (×10^6)	Variance estimate as percent of *MSE*			Confidence interval coverage (%)		
				$\hat{V}_{SAM}$	$\hat{V}_{MA}$	$\hat{V}_{RS}$	$\hat{V}_{SAM}$	$\hat{V}_{MA}$	$\hat{V}_{RS}$
STR/STR									
0.2 to 0.6	0.39	0.1	9.03	17.1	96.4	85.3	56.5	90.6	89.2
0.4	0.40	0.0	5.88	26.6	96.9	86.4	67.0	92.4	91.3
0.4 to 0.8	0.59	0.0	5.07	30.9	96.0	84.4	70.8	92.6	91.2
0.7	0.70	0.1	3.06	52.3	99.8	91.5	83.5	94.2	92.8
0.6 to 0.9	0.74	0.0	3.28	48.4	101.0	91.2	81.8	93.9	92.3
HD/HD									
0.2 to 0.6	0.40	0.1	8.37	21.1	97.6	89.1	62.1	91.9	90.1
0.4	0.40	0.1	5.53	31.7	100.1	87.9	72.1	93.4	91.7
0.4 to 0.8	0.60	0.1	4.69	36.5	100.7	88.3	75.5	93.4	91.3
0.7	0.70	0.0	2.94	57.0	102.5	92.2	85.0	94.2	93.0
0.6 to 0.9	0.75	0.0	3.18	52.6	102.6	91.2	83.8	94.1	92.4
STR/HD									
0.2 to 0.6	0.39	-2.0	8.72	19.1	85.6	81.0	56.2	85.8	84.8
0.4	0.40	0.1	5.53	31.7	100.1	87.9	72.1	93.4	91.7
0.4 to 0.8	0.59	-1.3	4.79	34.8	95.1	82.9	71.7	89.6	88.1
0.7	0.70	0.0	2.94	57.0	102.5	92.2	85.0	94.2	93.0
0.6 to 0.9	0.74	-0.8	3.23	50.9	98.5	87.3	81.7	91.8	90.1

The next columns of the table give the variance estimates as percentages of the *MSE*. The naïve variance estimates, $\hat{V}_{SAM}$, seriously underestimate the *MSE* in all simulation schemes. The naïve variance estimates are relatively constant across all schemes, and hence underestimate the variance and *MSE* more seriously the larger the nonresponse rate.

Now consider the two estimators designed for handling imputed data. The variance estimates using the model-assisted method are uniformly larger than the adjusted jackknife variance estimates and are closer to the actual *MSE*s. In the STR/STR schemes, the model-assisted variance estimates are within four percentage points of the *MSE*, whereas the adjusted jackknife variance estimates are 10 to 15 percent underestimates. As noted earlier, the assumptions of both methods are satisfied in these schemes. In the HD/HD schemes, the model-assisted variance estimators still closely track the *MSE*, and they are on average still closer to the *MSE* than the adjusted jackknife variance estimates. This result is somewhat surprising since the superpopulation model needed for the model-assisted method does not hold in the HD/HD schemes, while the response probability model holds–and the theoretical justification for the adjusted jackknife–holds. In the STR/HD schemes, the two methods still do reasonably well, even when the imputed estimates are biased and the response rates are low. Table 3 shows the model-assisted variance estimate is 86 percent of the *MSE* (the adjusted jackknife is 81%) due to the biased estimator. If the variance estimates were presented as percentages of the variance rather than the *MSE*, the estimates for the STR/HD schemes would be very similar to those given for the other schemes in which the estimates are essentially unbiased.

One reason for the underestimation of the adjusted jackknife variance estimates may be a technical issue relating to the *fpc*. As noted earlier, the theory for the adjusted jackknife method was developed assuming that the *fpc* is negligible, but in the simulations the *fpc*'s were sizable in some strata. Stratum sampling rates ranged from about 0.05 to 0.24. The variance estimator used in the simulations was modified to include stratum-specific *fpc's* based on the number of initial sampled units (see equation 4). These *fpc's* were thus applied to the overall variance estimator. In contrast, the *fpc* with model-assisted variance estimator applies only to the sampling variance component.

To investigate the effect of the *fpc*, we ran 10,000 simulations using the STR/STR/0.2 to 0.6 scheme with same sample sizes as in Table 1, but with a population that was five times larger than the CCD school district population. In those runs, the stratum *fpc*'s were negligible and the average $\hat{V}_{MA}$ and $\hat{V}_{RS}$ were 97.4 percent and 101.5 percent of the empirical *MSE*s. These results support the idea that the *fpc*'s were an important source of the underestimation of the variance using the adjusted jackknife method in the main simulation.

Now we examine the confidence interval (CI) coverage rates of the methods shown in the last three columns of Table 3. As expected, both the model-assisted and adjusted jackknife variance estimators convincingly outperform the naïve estimator. With a nominal 95 percent coverage, the naïve CI coverage percentages range from 56 to 85 percent, whereas the model-assisted and adjusted jackknife CI coverage percentages range from 86 to 94 percent and 85 to 93 percent, respectively. The model-assisted coverage rates are closer to the nominal level than the adjusted jackknife estimates, which is a consequence of the variance estimates being larger and closer to the *MSE*. Once again, the model-assisted method has slightly better properties than the adjusted jackknife even in the HD/HD schemes. In the schemes with biased imputed estimates, the CI coverages of the methods are about 85 percent when the responses rates are lowest and 88 to 92 percent when the response rates are somewhat higher. Again, the performances of both methods are reasonably good considering the misspecification of the model and are far better than naïvely assuming that the imputed data are actual observations.

Since the model-assisted variance estimates are formed by summing estimates of three components, we compared the component estimates to their theoretical values computed from the entire population. In all cases we found that $2\hat{V}_{MIX}$ accounted for less than one percent of $\hat{V}_{MA}$, an ignorable amount. In line with expectation, as the average response rate decreases, $\hat{V}_{IMP}$ accounts for a greater percentage of $\hat{V}_{MA}$. However, even when the average response rate is relatively high, e.g., 70 percent, $\hat{V}_{IMP}$ still accounts for about 50 percent of the total variance estimate. This highlights the problem of ignoring the component of variance due to imputation, even with relatively high response rates. Comparisons of the averages of $\hat{V}_{SAM}$ and $\hat{V}_{IMP}$ with their simulated theoretical values show that they are both nearly unbiased in most cases The greatest biases occur in the STR/HD schemes with variable response rates across strata, when $\hat{V}_{IMP}$ underestimates V_{IMP}. For the scheme in which response rates vary from 0.2 to 0.6, the underestimation is nearly 20 percent.

Since an unweighted hot deck is justified with the model-assisted method if the superpopulation model and MAR assumption hold, we also simulated the performance of $\hat{V}_{MA}$ with unweighted hot deck imputation. The results are displayed in Table 4. The STR/STR schemes are not included in the table because in this case the unweighted and weighted hot decks are identical (the weights within each stratum are constant).

Table 4. Relative bias, variance estimates, confidence interval coverage rates and percent of confidence interval misses by side for the model-assisted method with unweighted hot deck imputed estimates and confidence interval misses by side for the model-assisted method with weighted hot deck imputed estimates, by simulation scheme

Scheme	Rel-bias (%)	V_{MA} / MSE	Confidence interval coverage (%)	Unweighted hot deck		Weighted hot deck	
			$\hat{V}_{MA}$	Low(%)	High(%)	Low(%)	High(%)
HD/HD							
0.2 to 0.6	2.7	81.3	94.1	3.0	2.9	7.4	0.7
0.4	2.7	80.5	94.3	1.8	4.0	5.5	1.1
0.4 to 0.8	1.9	87.8	95.0	2.7	2.3	5.9	0.7
0.7	1.4	93.9	95.1	2.2	2.8	4.8	1.0
0.6 to 0.9	1.2	89.6	94.7	2.9	2.4	5.0	0.9
STR/HD							
0.2 to 0.6	0.5	92.3	91.8	7.1	1.1	13.8	0.4
0.4	2.7	80.5	94.3	1.8	4.0	5.5	1.1
0.4 to 0.8	0.4	98.5	93.8	5.1	1.1	10.1	0.3
0.7	1.4	93.9	95.1	2.2	2.8	4.8	1.0
0.6 to 0.9	0.3	99.0	94.0	4.8	1.2	7.7	0.5

All of the imputed estimates shown in Table 4 using the unweighted hot deck are biased. This situation arises often in practice because it is impossible to determine cells such that the response probability is constant within the imputation cells. Most of the biases for the unweighted hot deck in Table 4 are relatively small, with no relative bias exceeding three percent of the estimate. Note, however, that the biases are often larger than the corresponding biases for the weighted hot deck given in Table 3.

The confidence interval coverages in Table 4 for the unweighted hot deck are remarkably close to the nominal level, despite the fact that the imputed estimates are biased. In fact, the coverage percentages for the unweighted hot deck are superior to those of the weighted hot deck, even when the weighted hot deck estimates are unbiased and the unweighted hot deck estimates are biased. Note that the improvement is not the result of increasing the variance estimates, since the model-assisted variance estimates for the weighted and unweighted hot deck are very similar.

To investigate the reasons for the improvement of the coverage percentages for the unweighted hot deck, we decomposed the two tails of the coverage intervals derived from the model-assisted variance estimator for both the weighted and unweighted hot deck imputation schemes. Table 4 shows the percentages of the confidence intervals that did not cover the population total divided between those that missed on the low side and those that missed on the high side. The weighted hot deck model-assisted coverages are asymmetric, with much higher likelihood of missing on the low side. On the other hand, the unweighted hot deck coverages are more symmetric. To investigate these findings further, we computed the correlation between the imputed estimates and their model-assisted standard error estimates. The high correlation (0.66 to 0.73) between the estimate and the standard error is largely responsible for the asymmetry of the weighted hot deck coverage intervals. When the estimate is low, the standard error is likely to be underestimated, thus causing the confidence interval to fail to cover the population value. The estimates from the unweighted hot deck have essentially the same correlations, but the positive bias of the imputed estimate re-centers the distribution and gives more symmetric coverages.

We also examined the correlation in the case of complete item response and found the estimate and its standard error had a correlation of 0.71 which is very similar to the correlations when the missing data are imputed. We hope to further investigate the positive bias of the estimates and the positive correlation of estimates and their standard errors to determine if they are a function solely of the simulation design or may be a more general phenomenon.

The correlations between the imputed estimates and their estimated standard errors were also found to be high when the standard errors were computed using the adjusted jackknife. Across the 15 schemes these correlations ranged from 0.64 to 0.71, values that are similar to the model-assisted correlations. As in the model assisted case, the correlation causes the coverage with the adjusted jackknife variance estimates to be asymmetrical. This fact, in combination with the underestimation of the variance, leads to undercoverage.

5. Conclusions

We have examined three methods of variance estimation when missing items have been imputed using hot deck imputation. The simplest of these, and the poorest, is to treat the imputed values as if they were reported values. As is well-known, this leads to severe underestimates of variance and confidence intervals that

provide far less than the nominal coverage levels. In our simulations, 95 percent confidence intervals based on this naïve approach had empirical coverage rates ranging from about 56 percent to 85 percent depending on the amount of imputation.

The two alternative variance estimation methods that we examined in the simulations are the model-assisted and adjusted jackknife methods. In the former, a superpopulation model is used to derive estimates of variance components associated with sampling and imputation. For hot deck imputation, cells should be formed such that units in each cell can be approximately modeled as having a common mean and variance and also having the same response probability. The adjusted jackknife method assumes that units within hot deck cells have the same response probability.

Through simulation we studied the performance of the two alternatives under several schemes for response mechanisms with weighted hot deck imputation. Both the model-assisted and the adjusted jackknife variance estimators were substantial improvements over the naïve variance estimator. The model-assisted estimator performed somewhat better than the adjusted jackknife in almost all combinations of simulation parameters. The model-assisted variance estimates were more nearly unbiased and had consistently better confidence interval coverage. This was true even when the response rate was constant within cells, a situation that is favorable to the adjusted jackknife. The model-assisted method also performed better when the model misspecification was more severe and unfavorable to both the model-assisted and adjusted jackknife estimators. We suspect that the poorer performance of the adjusted jackknife method in the simulations may be largely due to the relatively large *fpc*'s in some of the strata. The adjusted jackknife estimator was developed assuming the *fpc* is ignorable. We modified the estimator for the simulations, but the modification may not correctly account for the *fpc*. An *fpc* based on the count of nonmissing units may be more appropriate, but further development is needed here. A limited simulation was conducted with a larger population and negligible *fpc*'s. In these simulations the adjusted jackknife variance estimates did not underestimate the true variances.

We also examined the sampling and imputation components of the model-assisted variance estimator and found that they tracked the theoretical values closely, even when the superpopulation model did not hold. Surprisingly, when we examined the use of the model-assisted variance estimator with the unweighted hot deck in constructing confidence intervals we found that it actually led to better confidence interval coverage rates with greater symmetry than the weighted hot deck variance estimator. The explanation for this finding is that the bias of the imputed estimates with unweighted hot deck imputation offsets the correlation between the estimates and their standard error estimates.

When there was no item nonresponse, the correlation between the estimates and their standard errors was equally high at 0.71. This high correlation caused coverage rates for the estimates with full response to be asymmetric and fail to cover the population value at the nominal level. Note that the overall sample size was large (1,020 school districts). This finding is informative about some of the normality assumptions that are routinely made in surveys with large sample sizes.

6. References

Brick, J.M., Kalton, G., Kim, J.K. (2002). *Variance estimation with hot deck imputation using a model*. Unpublished manuscript, Westat.

Kim, J.K., Brick, J.M., Kalton, G., and Fuller, W. (2002). *Some theory of variance estimation with imputed data.* Unpublished report for National Center for Education Statistics.

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

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