STRATEGIES FOR THE ANALYSIS OF IMPUTED DATA IN A SAMPLE SURVEY

James M. Lepkowski, Sharon A. Stehouwer, and J. Richard Landis
The University of Michigan

The National Medical Care Utilization and Expenditure Survey (NMCUES) was designed to collect data about the U.S. civilian noninstitutionalized population during 1980. Information was obtained on health, access to and use of medical services, associated charges and sources of payment, and health insurance coverage. The survey used a panel of approximately 6,200 sample households which were interviewed four or five times over a one year period in 1980 to 1981. A more complete description of the survey design is available in Bonham, 1983, and Research Triangle Institute, 1983. The NMCUES was co-sponsored by the National Center for Health Statistics and the Health Care Financing Administration. Data collection was provided under contract by the Research Triangle Institute, and its subcontractors, National Opinion Research Center and SysteMetrics, Inc.

The complexity of the survey requires an analyst to be familiar with a range of design features during the analysis, both to determine appropriate analytic methods and to investigate the impact that the design may have on estimation and inference. In this paper we discuss the impact that imputation to compensate for missing data can have on survey estimates. The focus is on the effect on means, sampling errors for those means, and relationships between variables not controlled in the imputation process.

Survey Nonresponse

Nonresponse in panel surveys such as the NMCUES occurs when sample individuals refuse to participate in the survey (total nonresponse), when initially participating individuals drop out of the survey (attrition nonresponse), or when data for specific items on the questionnaire are not collected (item nonresponse). In general, response rates for reporting units (RU's) and persons in the NMCUES were high, with approximately 90 percent of the sample RU's agreeing to participate in the survey and approximately 94 percent of the individuals in the participating RU's supplying complete information. However, for key survey items, the amount of missing data for responding persons is substantial because of item nonresponse. For example, total charges for more than one-half (51 percent) of the hospital outpatient department visits were missing. Survey estimates of means and proportions may be biased if nonrespondents tend to have different health care experiences than respondents or if there is a substantial response rate differential across subgroups of the target population. Furthermore, annual totals will tend to be underestimated unless allowance is made for the loss of data due to nonresponse.

Two methods commonly used to compensate for survey nonresponse are data imputation and the adjustment of sampling weights. For the NMCUES, data imputation was used to compensate for attrition and item nonresponse and weighting adjustments were used to compensate for total nonresponse.

Item Nonresponse and Imputation

Item nonresponse was a problem in the NMCUES for health care expenditures, income, and other sensitive topics. The extent of missing data varied by question, and imputation to replace missing data with nonmissing values from other respondents for all items in the data file would have been expensive. Imputations were made for missing data on important demographic, economic, and expenditure items; imputation rates are shown in Table 1 for selected measures from several of the files available from the NMCUES public use data tapes.

Demographic items such as age, sex, and education had the lowest nonresponse rates. Income items had higher rates of nonresponse, and for total personal income, which is a cumulation of earned income and 11 sources of unearned income, nearly one-third of the persons required imputation for at least one component of income. The disability items (bed days, work loss days, and cut down days) have rates of nonresponse that are intermediate to the demographic and income items.

The highest rates of nonresponse and imputation occurred for the important charge items on the various medical event records associated with each respondent. Total charges for medical visits, hospital stays, and prescribed medicine and other medical expenses records were missing for 25.9, 36.3, and 19.4 percent of the events reported, respectively.[1] Among the source of payment data, missing data rates for the source of payment item were small, but for the amount paid by the first source of payment item, the rate was generally higher.

The methods used to impute data for missing values were diverse and tailored to the variable requiring imputation (Cox, 1982). Three types of imputation predominated: logical, sequential hot deck, and weighted hot deck imputation procedures. The logical imputations were used to eliminate missing data that could be determined readily from other data items that provided overlapping information. The sequential hot deck was used primarily for small numbers of imputations for the demographic items, while the weighted sequential hot deck was used more extensively for the remaining item imputations.

The logical imputation was used in instances where the choice of a plausible value could be made from other available data. For example, race was not recorded during the survey for children under 14 years of age. Instead, a logical imputation was made during the processing of the

[1] All estimates in this paper include attrition imputation whether or not the imputed records had real or imputed data. Attrition imputation rates were quite small for all record types (usually less than one percent of records of a given type were imputed in the attrition imputation process). Removing attrition imputations from the analysis is not likely to change results.

data that assigned the race of the head of the reporting unit to the child. Similarly, extensive editing was performed for the charge data before any imputations were made. For example, if the first source of payment was available, only one source of payment was indicated, and total charge was missing, the value of the first source of payment amount was assigned to the total charge item.

In the sequential hot deck procedure, the data were grouped within imputation classes and then, within those classes, sorted by variables that were correlated with the item for which imputations were to be made. An initial value was assigned as a "cold deck" value, such as the mean of the nonmissing cases for the item within the imputation classes. The first record in the imputation class was then examined. If it was missing, the "cold deck" value replaced the missing data code; if real, the real value replaced the "cold deck" value as a "hot deck" value. Then the next record was examined. Again, if missing, the "hot deck" value was used to replace missing data, and, if real, the "hot deck" value was replaced. The process continued sequentially through the imputation classes.

The weighted hot deck was the most frequently used imputation procedure applied to the NMCUES. It was a modification of the sequential hot deck which uses the sampling weights assigned to each record to determine which real values were used to impute for a particular record which needed an imputation. Records were again classed and sorted by measures expected to be correlated with items requiring imputations, and the procedure was applied to several items simultaneously to reduce the amount of processing required.

Imputations for the important charge items involved a combination of logical imputations or edits followed by the weighted hot deck procedure. For example, for medical visit total charges an extensive edit was performed to eliminate as many inconsistencies as possible between the source of payment data and total charge items. The medical visit records were then separated into three types: emergency room, hospital outpatient department, and doctor visits. Within each type, the records were classed and sorted by different variables prior to a weighted hot deck imputation. For instance, for doctor visits the records were classified by the reason for visit, the type of doctor seen, whether work was done by a physician, and the age of the individual. Within the groups formed by these classing variables, the records were further sorted by type of insurance coverage and the month of visit. The weighted hot deck procedure was used with the classed and sorted data file to impute simultaneously for missing values of total charge, sources of payment, and sources of payment amounts.

Since extensive imputations were made for missing values for a large number of the key items in the NMCUES, they can be expected to influence estimates made from the survey in several ways. Although the weighted hot deck is expected to preserve the means of the nonmissing observations when those means are for the total sample or classes within which imputations were made (see Cox, 1980), this will not be the case for sampling variances. Sampling variances can be substantially underestimated when imputed values from an imputation process are used in the estimation process (see Kalton and Kasprzyk, 1982). For example, sampling variances computed using all data, real as well as imputed, for a variable with one-quarter of its values imputed will be based on one-third more values than were actually collected in the survey for the given item. The variance would be underestimated by a factor of at least one-third (Kalton, 1982). In addition, relationships between variables can be attenuated by imputation. Santos (1981) demonstrates that the attenuation of correlations can be substantial. In the next section, we present empirical findings which illustrate the effect imputations in the NMCUES can have on survey results.

Table 1
Percent of data imputed for selected survey items in four of the NMCUES Public Use Data Files

Tape location	Description	Percent imputed
	Person File (n=17,123)	
P54	Age	0.1
P57	Race	20.0[1]
P59	Sex	0.1
P62	Highest grade attended	0.1
P67	Perceived health status	0.8
P592	Functional limitation score	3.2
P125	Number of bed disability days	7.9
P128	Number of work loss days	8.9
P135	Number of cut down days	8.2
P399	Wages, salary, business income	9.7
P434	Pension income	3.5
P445	Interest income	21.6
P462	Total personal income	30.4[2]
	Medical Visit File (n=86,594)	
M117	Total charge	25.9
M123	First source of payment	1.8
M125	First source of payment amount	11.6
	Hospital Stay File (n=2,946)	
H252	Nights hospitalized	3.1
H124	Total charge	36.3
H130	First source of payment	2.2
H132	First source of payment amount	17.6
	Medical Expenses File (n=58,544)	
E117	Total charge	19.4
E123	First source of payment	2.8
E125	First source of payment amount	10.0

[1] Race for children under 14 imputed from race of head.

[2] Cumulative across 12 types of income.

Impact of Imputation on Estimates

Estimated means and sampling errors from the NMCUES for bed disability days, work loss days, work loss days in bed, cut down days, and restricted activity days are presented in Table 2. For each survey measure, separate estimates were computed using all data (i.e., both real and imputed) and using only the real data. The unweighted and weighted mean, unweighted and weighted simple random sampling (SRS) standard error of the mean, and the weighted complex standard error which accounts for the stratified, multistage nature of the design are presented.

For each measure, the weighted means computed using all the data and using only the real data are quite similar. This similarity is not unexpected given that the weighted hot deck imputation procedure is designed to preserve the weighted mean for overall sample estimates. The SRS standard errors, however, are smaller when all data are used simply because the SRS variance is inversely related to the sample size. For the complex standard error, three of the five measures have smaller standard errors when all data are used, and the other two measures show the opposite. One may conclude from Table 2 that imputation for the disability measures has little or no effect on estimated means or their standard errors for the total population, primarily because the amount of missing data for these measures is small (approximately seven or eight percent).

In contrast, for other measures that have larger amounts of missing data, imputation has larger effects. For example, consider the means and standard errors for total charge for a hospital outpatient department visit shown in Table 3. There were 9,529 hospital outpatient department visits (real visit records plus those generated from the attrition imputation process), and 4,841 of these have a total charge that was imputed from one of the other hospital outpatient department visit records. Thus, more than one-half of the total charges were missing for this particular medical event. Despite the large amount of missing data, the weighted means using all the data and using only real values are quite similar. However, sampling errors are changed substantially when imputed values are added to real values to form an estimate. The weighted and unweighted SRS standard errors are markedly smaller for all data than for the real data.

To investigate whether this decrease in sampling error is due to changes in sample size, changes in the element variance, or both, the element variances were computed by multiplying the weighted simple random sampling variances by the sample sizes. Since the element variances are quite similar using all data and real data, the difference in standard error can be mostly attributed to the loss in sample size when going from all data to real data.

Not all of the real data were used as donors for imputation, and some of the real values were used as donors several times. Table 3 also suggests that those real values not used as donors have a lower mean total charge than those used as donors, but values used as donors more than twice tend to have even smaller mean total charges. These means reflect the use of imputation classes within which the mean total charge and the amount of missing data varied.

The difference in complex standard errors between all data and the real data in Table 3 demonstrate large effects from imputation. However, neither of these complex standard errors is the actual standard error of the weighted mean estimated using all the data. The mean computed using all data includes 4,841 values that were actually subsampled with replacement from the 4,688 real values. In addition, the imputations were made across the primary sampling units and strata used in the variance estimation procedure. The assumption that the observations were selected independently between primary sampling units and strata is incorrect. Hence, the complex standard error for all data shown in Table 3 fails to account for two sources of variability present in estimates based on all data: the double sampling used to select values for imputation and correlation between primary sampling units and strata induced by imputation. At the same time, the complex standard error for the weighted mean computed using only the real data is an incorrect estimate of the standard error of the mean based on all the data. The actual sampling error of the weighted mean for all the data is probably larger than that shown for the mean estimated using all the data in Table 3; it may even be larger than the sampling error computed using only the real data.

Since it is not clear how to estimate the actual sampling error for the weighted mean estimated using all the data, an alternative estimation strategy was developed to provide a mean for which item nonresponse is compensated and sampling errors can be estimated. Rather than using an imputation strategy, an adjustment to the sampling weights was used to compensate for missing data. In particular, the imputation classes for hospital outpatient department charges were created. Within each class, the sum of weights for recipients and for donors and the sum of the number of donations were made within imputation classes. The sum of the weights for imputed records was then divided by the number of donations, and this average weight value was used to increase the weights of donors proportionate to the number of times they were reported as donors. The adjusted weights for donors within imputation classes will sum to the sum of weights for imputed values and donors combined. Estimates of means using these adjusted weights for only the real data should be similar to means obtained from all the data. In addition, sampling errors for this adjusted mean can be computed using the real data and the adjusted weights.

The estimated mean and its standard error under this adjusted weighting procedure is also shown in Table 3. The mean is virtually identical to that obtained using all the data, and the standard error is quite close to that obtained from the real data using the unadjusted weights. The differences between sampling standard errors for the weighted mean using only the real data and for the mean using adjusted weights are due to the effects of increased variability of weights in the adjustment process.

As a final illustration of the effects that imputation can have on survey results, Figure 1 presents estimated mean charges per hospital outpatient department visit for four family income

Table 2

Sample Size, Rate of Imputation, Mean, Standard Errors, and Square Root of Design Effect for Five Disability Measures by Data Type: NMCUES,United States, 1980

Data Type by Disability Measure	Sample Size	Rate of Imput-ation	Unweighted Estimates		Weighted Estimates			Square Root of Design Effect
			Mean	SRS Std Error	Mean	SRS Std Error	Complex Std Error	
Bed Disability Days								
· All data	17,123	0.08	5.303	.1279	5.268	.1269	.1540	1.21
· Real data	15,777		5.253	.1326	5.228	.1319	.1599	1.21
Workloss Days								
· All data	13,069	0.12	3.614	.1221	3.696	.1220	.1629	1.34
· Real data	11,537		3.510	.1284	3.574	.1277	.1716	1.34
Work Loss Days in Bed								
· All data	13,069	0.16	1.516	.0508	1.568	.0518	.0592	1.14
· Real	10,970		1.530	.0556	1.578	.0568	.0652	1.15
Cutdown Days								
· All	17,123	0.08	6.831	.1681	6.881	.1697	.3343	1.97
· Real	15,724		6.609	.1721	6.639	.1735	.3322	1.91
Restricted Activities Days								
· All	17,123	0.18	13.746	.2559	13.805	.2573	.4716	1.83
· Real	14,049		13.036	.2732	13.064	.2742	.4658	1.70

Table 3

Sample Size, Means, Standard Errors, Square Root of Design Effect, and Element Variance for Total Charge for a **Hospital Outpatient Department(OPD)** Visit by Data Type: NMCUES, United States, 1980

Data Type†	Sample Size	Unweighted Estimates		Weighted Estimates			Square Root of Design Effect	Element Variance ($x 10^{-3}$)
		Mean	SRS Std Error	Mean	SRS Std Error	Complex Std Error		
All	9,529	51.86	1.030	51.61	1.018	1.914	1.88	9.87
Real only	4,688	52.28	1.436	52.27	1.430	2.936	2.05	9.59
Real(ReWt)	4,688	---	---	51.80	1.470	3.000	2.04	10.14
Imputed	4,841	51.45	1.476	50.98	1.447	2.323	1.60	10.14
Real:								
Not donor	929	47.83	2.108	48.53	2.117	3.935	1.86	4.17
Donor once	2,798	55.85	2.016	55.76	1.982	3.386	1.71	11.00
Donor twice	841	48.61	3.525	49.37	3.579	4.879	1.36	10.78
Donor 3-5 times	120	29.45	7.340	28.97	7.987	11.64	1.46	7.66

†ReWt denotes reweighting of real data within imputation subclasses

groups computed using all the data and using only the real data. For the real data, the mean charge per visit increases in a linear fashion as the family income increases. However, when all the data are used to estimate the mean charge per visit, the mean charge does not increase as rapidly with increasing family income. The strong relationship between family income and mean charge per hospital outpatient department visit in the real data has been attenuated by the imputed values.

The reason for this attenuation is shown in Figure 2. Sixteen imputation classes were formed for the imputation of total charges for hospital outpatient department visits. Figure 2 shows mean charge for real data for the total sample and the subgroup with family incomes less than $5,000 in 1980. The low income group has lower mean charges than the total sample. Since family income was not one of the variables used to form imputation classes, low family income persons within an imputation class with missing hospital

Figure 1

Mean Charge for Hospital Outpatient Department Visits by Income Groups

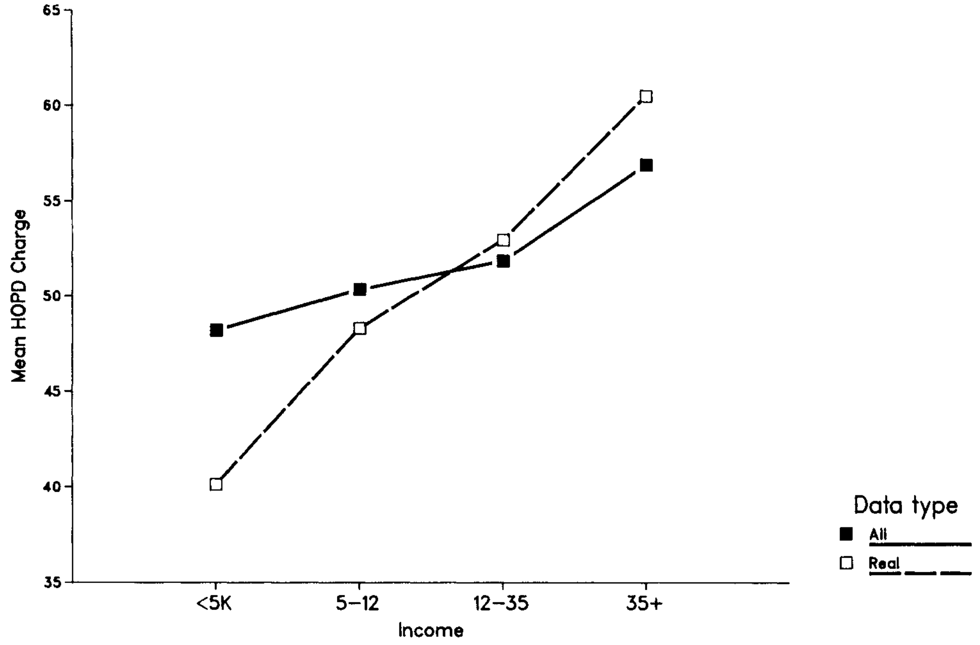
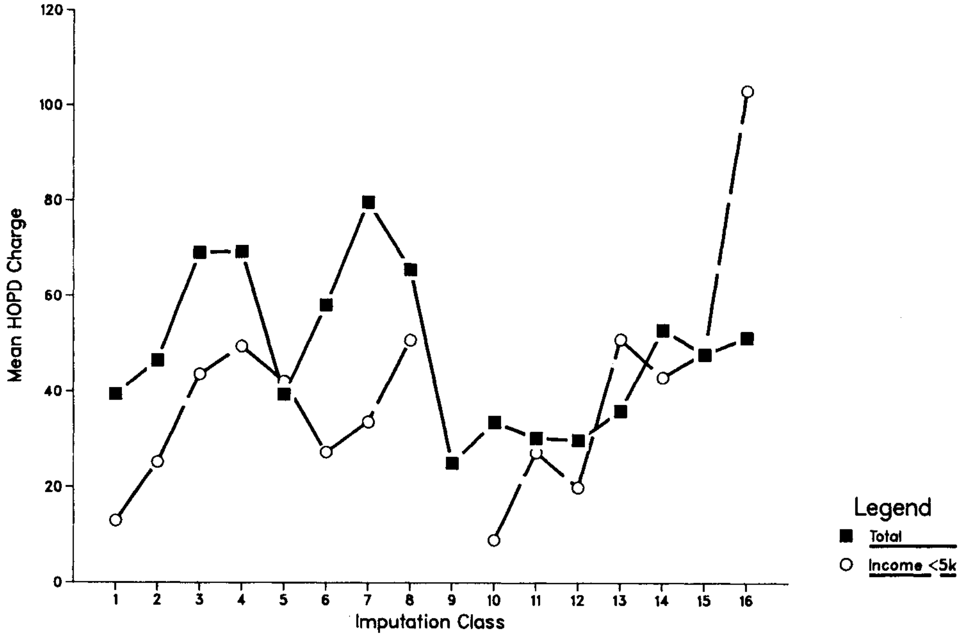


Figure 2

Mean Charge for Hospital Outpatient Department Visits by Imputation Class



outpatient department visit total charges were imputed a charge that was, on average, higher than the mean charge for low income persons with real data. This occurs in almost every imputation class. When the real and imputed data are combined for persons with family incomes less than $5,000, the effect of imputation is to increase the mean charge for this subgroup. Conversely, for persons with family incomes of $35,000 or more, total hospital outpatient department visit charges for persons with real data tend to be larger than values imputed to persons with missing charges. The overall impact of the imputation process on the relationship between charges for hospital outpatient department visits and family income is a regression toward the mean charge for real data for low and high income subgroups.

Discussion

The results in Tables 2 and 3 and Figure 1 demonstrate the effect that imputation can have on estimated means, on estimated sampling errors, and on relationships between variables. The analyst of survey data which has imputations for important survey items is faced with selecting a strategy for handling imputation in estimation. The results in this paper permit comparisons among three different strategies for handling imputed data: use all the data, create weights for each item which adjust for item nonresponse and use only real data, and use only real data and the unadjusted sampling weights (i.e., ignore the effects of item nonresponse). It is useful to examine the advantages and disadvantages of each of these strategies by reviewing likely effects on four estimates of interest: sample means for the total sample, estimates of totals or aggregates, estimated sampling errors for means, and relationships between an imputed measure and a measure not controlled for in the imputation process.

In the instance of the NMCUES and the weighted sequential hot deck imputation procedure, the first strategy, using all the data (including imputed values), would provide mean values for the total sample similar to those obtained by using only the real data. Estimates of aggregates or totals (e.g., total charges for all hospital outpatient department visits) would be automatically adjusted for the failure to obtain responses from some persons. However, sampling errors estimated using all the data and standard variance estimation techniques will tend to underestimate the actual variance. In addition, relationships between imputed measures and variables not used as control variables in the imputation process will be attenuated by the imputations. In multivariable analyses, the effects of imputation may be difficult to anticipate, because variables which were and were not controlled for in the imputation process may appear in a model and may be effected in unexpected ways.

The second strategy is to adjust sampling weights separately for each estimate to account for item nonresponse, thus creating item nonresponse adjustment weights for every item requiring compensation for nonresponse. The advantages to this adjusted weight strategy is that sampling errors could be estimated for estimated means using standard sampling error estimation techniques, weighted means for real data for the total sample would essentially be preserved for estimates based on all the data, and estimated totals would not be subject to underestimation because of item nonresponse. On the other hand, it would be a sizeable task to create such weights for each item, and it is not clear what weight should be used for examining a relationship between two or more variables, where some of the variables have nonresponse adjusted weights and others do not.

The third strategy suggested here is to ignore the imputations altogether and simply analyze the real data. Weighted means for the total sample computed using real data will not differ substantially from those estimated using all the data. Sampling errors may be estimated for the real data means, and relationships in the data will not be attenuated by the imputation process. However, estimated totals for items with substantial amounts of missing data will be severe underestimates if only real data are used.

A final adaptive strategy may also be considered. Estimation of means and sampling errors of means and the analysis of relationships among survey variables may be done using only the real data. On the other hand, estimates of totals could be computed using all the data to avoid severe underestimates for survey measures with large item nonresponse rates. One still would be faced with the unresolved task of estimating sampling errors of totals which used imputed values, since a suitable sampling error estimation strategy is not readily available for surveys such as the NMCUES which have imputed data.

References

Bonham, Gordon S. "Procedures and Questionnaires of the National Medical Care Utilization and Expenditure Survey," National Medical Care Utilization and Expenditure Survey. Series A, Methodology Report No. 1. DHHS Publication No. 83-20001. Public Health Service. Washington. U.S. Government Printing Office, March, 1983.

Research Triangle Institute. Public Use Data Tape Documentation, National Medical Care Utilization and Expenditure Survey, 1980. National Center for Health Statistics and Health Care Financing Administration, DHHS. Washington, DC. U. S. Government Printing Office, August, 1983.

Cox, Brenda G. "The Weighted Sequential Hot Deck Imputation Procedure," Proceedings of the Section on Survey Research Methods, American Statistical Association, 1980, pp. 721-6.

Cox, Brenda G., Parker, A. Elaine, Sweetland, Scott S., and Wheeless, Sara C. "The Imputation of Missing Item Data for the National Medical Care Expenditure Survey," Research Triangle Park, NC: Research Triangle Institute, July, 1983.

Kalton, Graham and Kasprzyk, Daniel. "Imputing for Missing Survey Responses," Proceedings of the Section on Survey Research Methods, American Statistical Association, 1982.

Kalton, Graham. Compensating for Missing Survey Data. Ann Arbor, Michigan: Survey Research Center, 1983.

Santos, Robert. "Effects of Imputation on Regression Coefficients," Proceedings of the Section on Survey Research Methods, American Statistical Association, 1981, pp. 140-145.