AN ASSESSMENT OF ALTERNATIVE DATA REPLACEMENT TECHNIQUES

Nancy A. Mathiowetz, Agency for Health Care Policy and Research
2101 E. Jefferson Street, Rockville, MD 20852

Keywords: imputation, exact matching

Introduction

Missing data is a problem that pervades most large data sets and most certainly is a problem for data based on survey responses. Missing data may be due to a selected sample element refusing to participate in the survey or the interviewer's inability to locate the selected unit (unit nonresponse) or among cooperative respondents, due to a respondent's inability or refusal to provide the information of interest (item nonresponse). In addition, item nonresponse may be a product of post-data collection editing. While both unit and item-based sources of missing data are problematic for the analyst of survey data, the research presented here will be limited to the problem of item nonresponse.

In most statistical agencies responsible for releasing large survey data bases, there is a tendency to provide complete data, that is, data files in which rates of missing data have been eliminated or minimized. Such files are advantageous in that they permit the use of standard statistical software based on methods of analysis requiring complete data. In addition, by having the data collector provide complete data, information that is only available to the collector, for instance, confidential information or information concerning the nuances of the data collection effort can be used in the imputation algorithm. Depending upon the approach used for replacing missing data, imputation may also serve to reduce nonresponse bias.

As noted by Kalton and Kasprzyk (1986), a number of different strategies exist for replacing missing data. These methodologies include deductive imputation, overall mean imputation, class mean imputation, random imputation within classes, sequential hot deck imputation, and regression imputation. In addition, replacement of missing data can be facilitated by the use of either exact or statistical matching to an external data file, where an exact match is defined as a match in which the linkage of data for the same person from different files is sought and does not indicate that the matches are made without error (Statistical Policy Working Paper 5, 1980).

NMES Estimation Strategy

The Household Component of the 1987 National Medical Expenditure Survey (NMES), was a one-year multi-wave study which collected information on health status, utilization of health care services and associated expenditures for the civilian noninstitutionalized household population. Respondents were interviewed five times between February 1987 and July 1988 concerning use of and charges for health care, health insurance coverage, employment, income, and related characteristics of survey participants for the reference period, calendar year 1987. The overall response rate was 80 percent. Details concerning the sample design can be found in Cohen, et al. (1991).

One of the primary analytic measures from NMES are the total charges associated with health care utilization. Previous experience in conducting the 1977 National Medical Care Expenditure Survey suggested that for many types of health care utilization, household respondents are unable to provide information on the total charge for the visit. As a means to reduce the potential bias in medical expenditure estimates derived solely from data collected from household respondents, the 1987 NMES included a Medical Provider Survey (MPS). The MPS was designed to follow up all medical providers who provided care to NMES respondents in a hospital, in a clinic setting, all providers of home health care, and all providers reported by persons in households with at least one member eligible for Medicaid. In addition, the design included all medical providers associated with persons in a nationally

representative 25 percent sample of households, including providers associated with ambulatory office-based events which were otherwise excluded from the MPS. This 25 percent sample was designed to provide a data base for modeling the differences in charges reported by household respondents and medical providers and to increase the pool of potential donors for imputation purposes.

Data from the MPS were linked to the HS reported events via a probabilistic matching strategy adopted for NMES. The probabilistic approach to matching or record linkage was outlined by Felligi and Sunter (1969). The Felligi and Sunter probabilistic approach was developed specifically to account for the possibility of errors in the data. The approach depends upon matching or comparison rules constructed by the user, but permits the error structure of the reported data to guide the determination of the likelihood of the match. The algorithm compares two events and the probability that they represent the same event is estimated. These probabilities or likelihoods can then be used in a decision making process to determine which events match. The probabilities of matching across the set of all possible comparisons can also be used to approximate the chances of making false matches or not making matches when the two events do match. The CANLINK software package (Statistics Canada, 1985), developed by Statistics Canada was used for the probabilistic match.

When MPS data were available for a given event, they were used to construct the expenditures for that event. In the absence of MPS data, respondent-reported data were used. The remaining missing data were imputed using a weighted sequential hot-deck procedures, where donors consisted of both household reported and MPS data that had been linked to a household event.

The Problem

Using a mix of household and medical provider expenditure data has been seen as a relatively cost effective means for collecting high quality expenditure data. Although using MPS data adds both time and costs to the NMES study, for some types of medical events (e.g. hospitalizations) and for some types of respondents (e.g. those on Medicaid), household respondents are for the most part, unable to provide accurate expenditure information. However, for other types of events, for example ambulatory office visits, household respondents may provide as accurate expenditure data as may be available from medical providers, at a lower cost and in a more timely manner (see Cohen and Carlson, in press).

The focus of this investigation is to determine whether for office-based physician visits, not including visits made by respondents on Medicaid, more timely and less costly means for replacing expenditure information can be found.

MPS:HS Replacement Strategy

As noted above, in 1987 a mixed approach was taken to replace missing expenditure data. For those cases where a link could be made between the household and medical provider reported event and for which expenditure data were reported by the provider, the MPS expenditure data were used. For nonmatches, household reported data were used when available, and replacement for the remaining cases was done using a sequential weighted hot deck procedure (Cox, 1980) where donors were based a pooled group of household and provider reported expenditures. Of the 91,274 office-based physician visits reported in NMES (not including those where Medicaid is cited as a source of payment), the source of the expenditure data is as follows: 53 percent based on household data, 23 percent based on MPS data, and the remaining 24 percent imputed.

The costs of collecting and matching MPS data are significant. In 1987, MPS data were collected on 56,388 ambulatory office events, representing contacts with more than 10,000 providers. Of these 56,388 events, 31,213 (or 55 percent) were matched to household reported events and used in the creation of public use files.[1] Therefore, information for over 25,000 events was never used in the construction of public use files. In addition to the cost of collecting the MPS data, the time needed to

contact providers, process the data, and then match the files added approximately two years to the creation of public use files.

Alternative Approaches

Two alternative approaches for replacing expenditure data for office based physician visits are evaluated in this research. The first method evaluates the use of the MPS as a cold deck imputation source. In doing so, all person and provider identification numbers were eliminated as potential variables for matching. The MPS data represent an "ideal" cold deck data base, since the actual "exact" matches do exist in the data. It therefore provides a best case scenario for evaluating a cold deck approach. Although the reader may question the merit of such an approach, given the duel goals of reducing the length of time and the costs associated with exact matching, the use of a cold deck data source would eliminate the need to contact household respondents' providers (e.g. one could create a database from insurance usual and customary charges). Cold-deck data were used to replace expenditure data for all events in the public use file for which household charge information was not available (that is all cases for which the public use file indicated that the total charge variable was based on either MPS or imputed data).

The second approach simply expands the hot deck approach used in 1987 (when the mix of household and MPS matched data were used as donors) to all events missing an expenditure estimate from the household. Donors were defined as those cases in the public use file which indicated that total charge was based on the household data and recipients were defined as those cases in which total charge information was based on either MPS or imputed data.

Findings

Tables 1 through 4 present the findings. Tables 1 and 2 examine the resulting event and person level distributions for the various approaches, the PUF estimate, the estimate based on the hot deck procedure, and the estimate based on the cold deck matching. The distributions are similar at both the event and person levels. At the event level, the 50th percentile charge is the same ($32), regardless of the data replacement method used. Looking at the three distributions, it appears that for events with charges over $500, both the hot deck and cold deck approaches produce significantly fewer than the approach used to generate the PUF files(1.4 percent and 1.6 percent vs. 2.1 percent, respectively). At the person level, we see a similar pattern--no difference in the charges at the 50 percentile, but lower annual mean charges per person for the hot deck and cold deck approaches as compared to the PUF file.

Tables 3 and 4 provide the distribution of the absolute value of the difference between the PUF estimate and the alternative estimates, at the event and the person level, respectively. The findings in Table 3 suggest that the hot deck approach may be the best alternative to the current PUF estimate. Looking at the event level data, we see that the mean absolute value of the difference between PUF and the hot deck approach is $26 as compared to $54 for the difference between the PUF and cold deck approaches. The 50th percentile values are $0 and $20, respectively. At the person level, however, the findings would suggest that the cold deck approach more closely resembles the person level PUF estimate. Although the 50 percentile absolute difference at the person level for the hot deck estimate represents slightly more than 2 percent of the mean charge estimate, the $77 mean absolute difference for the hot deck estimate is disturbing and appears to be driven by the 2.9 percent of the cases in which the PUF estimate and the hot deck estimate differed by more than $500. In almost 16 percent of the cases, the absolute difference (at the person level) between the PUF estimate and the hot deck approach was more than $100.

Conclusions

In concluding we should note that the PUF estimate does not represent a gold standard, rather it represents a standard that we believe to be a reasonably good approach, given relatively high levels of resources (to obtain the

MPS data) and time (to complete MPS data collection and matching work). Even though Table 4 suggests that there are problems with the hot deck approach at the person level of estimation, the indications from Table 3 are that the hot deck approach (using only household reported data as donors) offers a potentially cost-efficient means for replacing total charges for office-based physician visits. However, it appears that special attention needs to be given to the determinents of high cost office visits and these determinents need to be added to either the class or sort specifications for the hot deck imputation models. Further research will attempt to improve the models for these types of events.

[1] The estimates of the percentage of ambulatory events that matched include visits where Medicaid is reported as a source of payment.

References

Cohen, S.B.,DiGaetano, R.G. and Waksberg, J. (1991) *National Medical Expenditure Survey: Sample Design of the 1987 Household Survey.* AHCPR Pub. No. 91-0037, U.S. Department of Health and Human Services, Rockville, MD.

Cohen, S.B. and Carlson, B.L. (In Press) A Comparison of Household and Medical Provider Reported Expenditures in the 1987 NMES. *Journal of Official Statistics.*

Cox, B.G. (1980). The weighted sequential hot deck imputation procedure. *Proceedings of the Section on Survey Research Methods, American Statistical Association,* 721-726.

Fellegi, I.P. and Sunter, A.B.(1969). A Theory of Record Linkage. *Journal of the American Statistical Association*, 64:1183-1210.

Kalton, G. and Kasprzyk, D. (1986). The Treatment of Missing Survey Data. *Survey Methodology*, 12, 1-16.

Statistics Canada (1985). *Generalized Iterative Record Linkage System.* Research and General Systems, Informatic Services and Development Division, Ottawa, Ontario.

Statistical Policy Working Paper 5 (1980). *Report on Exact and Statistical Matching Techniques.* U.S. Department of Commerce, Office of Federal Statistical Policy and Standards, Washington, D.C.

Table 1. Alternative Imputation Techniques: Event Level Charges

Charge/Visit	PUF Estimate	Hot Deck	Cold Deck
$0-$9	9.2%	8.5%	9.4%
10-19	8.1	8.2	8.2
20-29	25.4	25.6	25.1
30-39	17.8	18.2	17.7
40-49	9.8	10.1	10.0
50-99	18.1	19.0	19.0
100-499	10.3	9.8	9.9
500+	0.9	0.6	0.6
Mean	$55	$53	$54
50th percentile	$32	$32	$31
Number of Visits	91,274	91,274	91,274

Table 2. Alternative Imputation Techniques: Person Level Annual Charges

Charge/Visit	PUF Estimate	Hot Deck	Cold Deck
$0-$9	3.9%	4.1%	4.0
10-19	2.2	2.5	2.3
20-29	7.7	8.1	7.8
30-39	6.8	6.9	6.8
40-49	6.2	5.7	6.2
50-99	20.9	20.6	20.5
100-499	41.9	41.5	41.9
500+	19.3	19.5	19.2
Mean	$237	$227	$234
50 percentile	$107	$105	$107
Number of Persons	20,693	20,693	20,693

Table 3. Alternative Imputation Techniques:
Absolute Value of the Difference (PUF-Alternative) Estimate
Event Level Charges

ABS (DIFF)	PUF-HOT DECK	PUF-COLD DECK
$0-9	64.2%	27.7%
10-19	9.4	19.1
20-29	7.3	14.1
30-39	4.3	9.1
40-49	2.8	5.1
50-99	6.4	12.8
100-499	5.4	10.7
500+	0.5	1.2
Mean ABS Diff	$26	$54
50% ABS DIFF	$0	$20

Table 4. Alternative Imputation Techniques:
Absolute Value of the Difference (PUF-Alternative) Estimate
Person Level Annual

ABS (DIFF)	PUF-HOT DECK	PUF-COLD DECK
$0-9	54.7%	86.6%
10-19	7.2	2.0
20-29	5.7	1.6
30-39	4.1	1.4
40-49	3.5	0.9
50-99	9.4	3.3
100-499	12.7	3.8
500+	2.9	0.8
Mean ABS Diff	$77	$24
50% ABS DIFF	$5	$0