IMPUTATION METHODS FOR LARGE COMPLEX DATASETS: AN APPLICATION TO THE NEHIS

Ibrahim S. Yansaneh, Leslie S. Wallace, and David A. Marker
Ibrahim S. Yansaneh, Westat Inc., 1650 Research Boulevard, Rockville, Maryland 20850

Key Words: Large datasets, Imputation, Regression, Hot Deck.

1. Introduction

Large complex datasets typically contain large numbers of variables measured on even larger numbers of respondents. Such datasets are the logical result of surveys that attempt to understand the relationships among characteristics of the population of inference and multiple outcome measures. Such surveys are frequently conducted by or for government agencies, covering topics such as health, welfare, education, and many others. These data are expensive to collect, but once collected, provide a wealth of analytic possibilities.

To improve those analytic capabilities, it is common to impute for item nonresponse, allowing more respondents to be incorporated in the analysis of complex multivariate relationships. Without imputation, one is restricted to analyzing the responses for the subset of cases, which responded to all of the questions being examined. This subset is often unrepresentative of the entire population, thus providing misleading analyses. The goal of imputation is to try and minimize the bias resulting from this nonresponse.

Much of the research on imputation has concentrated on best methods for imputing for a single variate at a time. See, for instance, Little and Rubin (1987) and Nordholt (1998). In large complex datasets, the situation is much harder, because the resulting data must satisfy multiple logical consistencies that are often intertwined. Also, in developing models to use in imputation, it is desirable to anticipate the main analyses that are planned for the imputed data and to try to avoid attenuating the variance among the variables whose relationships are being investigated. By their very nature, large complex datasets are analyzed by many users over many years. It is impossible to anticipate all of the significant analyses that will be conducted by the analysts. It is only possible to work with those who designed the original study, to try and anticipate which relationships are most important to accurately preserve during the imputation process.

This paper discusses the issues that must be addressed when trying to impute for large complex datasets. As an example, we will refer to the imputation for the National Employer Health Insurance Survey (NEHIS), which Westat has been conducting for three United States health agencies. NEHIS collected information on the health insurance plans offered by 40,000 private-sector establishments and governments (collectively referred to as establishments in this paper), and 50,000 health insurance plans offered by those private and public-sector respondents. More than 100 variables were collected for each establishment and each health plan. Fifty of these variables were selected for imputation. Since it was necessary to model these variables separately for public and private sectors, and for fully insured and self-funded health plans, this required almost 150 separate imputation models.

Imputation methods can be categorized as either deterministic or stochastic. Deterministic methods are generally simpler, but artificially reduce the variability of the data. Thus, deterministic methods, such as mean or modal imputation, or regression without a residual, tend to be used only when the percentage of missing data is quite small. There is a wide range of stochastic imputation methods, among the most common being Hot Deck and regression methods. Much recent research, for example, Schafer (1997), has examined the potential use of multiple imputation. Multiple imputation was not considered here for the complications mentioned above; that is, the tremendous number of variables being imputed and the lack of knowledge concerning the planned analyses. The former makes the inclusion of multiple imputed values for each variable extremely difficult, while the latter prevents the development of "proper" imputations, as required by Little and Rubin (1987).

For NEHIS imputation, each of the 50 variables had to be modeled separately for the public sector and private sector (most, but not all, were applicable for both sectors). For some variables it was also necessary to model establishments that self-funded their health plans separately from those that were fully insured. When examined in this way, item response rates varied from 99 percent to 29 percent; but in almost all cases the response rates were above 70 percent. Even though the government did not plan to publish estimates for the few variables with low response rates, it planned to use them in a variety of modeling efforts since no other source exists for this information. Imputed data based on low response rates were thought to be preferable to using the unimputed data for modeling purposes.

Further complicating the imputation, the data were subject to numerous logical consistency requirements. These range from situations where employer and employee contributions must add up to the total premium, to much more complex arrangements involving combinations of single and family-coverage enrollments and contributions and total plan costs. This required frequent passes through imputation-edit-reimputation cycles to achieve an imputed dataset that matched the cleanliness of the reported data. Many of these consistency requirements could only be evaluated when the last of the variables involved was imputed; so it was very important to determine an order of imputation that would maximize the available covariates at each step, and simultaneously allow for checking logical edits as soon as they could possibly be checked. To accomplish all of these tasks, the variables were broken up into chunks of related variables, and the chunks grouped together so they could be imputed simultaneously, since a variable in one chunk would not be needed to check the imputation of a variable in another chunk in the same group. The groups were then ordered in a logical sequence to provide for the maximum available covariates at each step.

Section 2 describes the methods of imputation appropriate for large complex datasets. Section 3 discusses the issue of selecting an order of imputation. Imputation models and the process of implementing them are discussed in Section 4. Section 5 provides a summary and conclusions. Each section includes examples from the NEHIS imputation.

2. Methods of Imputation

Item nonresponse is an unavoidable feature of sample surveys in general. It occurs when some responses are missing from an otherwise cooperating sampled unit. This type of nonresponse frequently arises from (i) refusal or insufficient knowledge on the part of the respondent; (ii) inability to find a suitable respondent; (iii) invalid data being discarded as a result of edit checks; and (iv) missing data because questions are not asked, generally due to the incorrect application of the skip patterns. Some of these sources of item nonresponse (for example, the first) may be correlated with response and hence may induce bias in the survey estimates, while others (for example, the third) may lead to data being missing at random.

Two broad classes of imputation methods are frequently employed to compensate for item nonresponse in surveys. These are regression methods and imputation class methods. The imputation class methods involve (i) partitioning the sample into a number of imputation classes based on categories of variables known to be strongly correlated with the analysis variable of interest; and (ii) assigning a value from a record with a response on the item in question (donor) to a record with a missing value on that item (recipient). In the regression imputation procedure, the imputation variable is regressed on the covariates for all cases with non-missing values of the imputation variable. The missing values may then be imputed either by using the predicted value from the model given the values of the covariates for the record with missing item response (deterministic regression), or by using this predicted value plus a randomly chosen residual (stochastic regression). If the imputation variable is categorical, then log-linear or logistic models may be used. See, for example, Little and Rubin (1987), Marker et al. (1997), and Nordholt (1998) for extensive discussions of the standard methods of imputation.

In order to preserve the multivariate relationships among blocks of variables, it is preferable to impute blocks of missing data concurrently, rather than on an item-by-item basis. This is referred to as *Block Imputation*. Imputation as a block should be considered only when the imputation variables in the block are very highly correlated and have similar covariates (that are not among the variables in the block). When a set of imputation variables in a block has very strong covariates but the covariates are not the same, imputation should be implemented independently for each variable in the set.

In general, regression methods are preferable to Hot Deck methods when the covariates are predominantly continuous and highly correlated. Regression imputation should be considered when enough covariates (which may be continuous or categorical) are available for respondents as well as nonrespondents that can be used to model the response for the imputation variable. The Hot Deck imputation procedure is most appropriate when dealing with categorical variables, when covariates are weakly correlated with the imputation variable, or when the level of missing data is substantial.

Choosing an imputation method requires a thorough understanding of the nature of missing data, that is, the extent and patterns of the observed nonresponse. Such knowledge is very useful for the partitioning of the sample into appropriate imputation cells (in the case of imputation class methods); and the identification of covariates that are important with respect to modeling nonresponse (in the case of regression methods). The following guidelines may be taken into consideration when deciding on an imputation strategy for a large complex dataset:

(i) Covariates cannot be missing when the imputation variable is missing. Thus, some variables that appear to be highly correlated may be inappropriate as covariates;

(ii) For continuous covariates, regression imputation is used if a highly predictive regression model (one with a high R-squared value) can be obtained. Remember that a high R-squared value is a necessary, but not sufficient, condition for the choice of regression as a method of imputation; and

(iii) If the imputation variable is categorical, and is weakly correlated with its covariates, the choice of imputation method is between Hot Deck and deterministic imputation (mean or modal). If the item nonresponse rate is high, then Hot Deck imputation should be used. If the item nonresponse rate is extremely low and there are no highly correlated covariates, mean or modal imputation may be used, subject to any constraints imposed on the variable. Mean or modal imputation should not be used if the imputation variable is evenly spread across many categories.

To understand the extent and patterns of missingness in the NEHIS dataset, frequency distributions of all imputation variables and covariates were constructed and examined. For each set of variables, the most appropriate imputation strategy was selected. Alternative approaches were evaluated in terms of the quality of the imputations and the associated costs. Some approaches, which may be sub-optimal, were chosen because they kept the number of passes through the data to a minimum, thereby drastically reducing data processing costs while producing results that are essentially comparable to those produced by optimal but very expensive and time-consuming approaches. Other approaches were not considered for implementation in NEHIS for a variety of reasons. For instance, regression imputation was not used primarily because of the pervasive problem of missing data in the most highly correlated covariates, but also because of time and cost considerations. The cold deck method was not used because of the lack of comparable past data on the same population. Logistic regression imputation was not used because of the relatively small number and the relatively low nonresponse rates of binary imputation variables in the NEHIS dataset.

A modification of the Hot Deck, which we will refer to as the Hot-Deck-Variant (HDV) method, was implemented in situations where there was only one significant continuous covariate for a given imputation variable and this covariate turned out to be a count variable (for example, number of enrollees in a health insurance plan or number of employees at an establishment). In this procedure, the covariate itself (rather than categories of it), was used as the soft boundary. This procedure has the advantage of easy implementation and its results are comparable to those obtained from regression imputation (Aigner et al., 1975).

The Hot Deck and HDV procedures were the most frequently used imputation methods in NEHIS primarily because the NEHIS dataset contains a large number of imputation variables with moderate response rates and weak covariates and also for computational convenience. For all imputation class methods, the imputation classes were based on percentiles of the distributions of the covariates. Of the approximately 150 imputation models implemented in NEHIS, 10% used deterministic imputation, 30% used HDV, and the rest used Hot Deck.

3. Order of Imputation

In implementing an imputation task for a large complex dataset consisting of a large number of variables which are interrelated, possibly measured at different levels (for example, establishments and health insurance plans within establishments, schools and students within schools, hospitals and patients within hospitals, etc.), and subject to many constraints, the order in which the variables are imputed is of paramount importance. The following factors should be taken into consideration in deciding on the order of imputation for a large complex dataset:

(i) If one variable is used in the construction of a second variable, then the first variable should be imputed before the second;

(ii) The imputations should follow the logical sequence (if any) suggested by the patterns of missingness of the imputation variables, that is, the joint frequencies that identify sets of imputation variables that are missing together. For instance, if the first variable happens to be a strong covariate of the second, and is present in most cases where the second variable is missing, then the first variable should be imputed first;

(iii) Within groups, variables using deterministic imputation should be imputed before variables requiring stochastic imputation;

(iv) Within groups, decisions about the order of imputation need only be made for imputation variables that are very highly correlated with one another. The order of imputation is not crucial for variables that are not highly correlated; and

(v) If the best covariates are the same for a set of imputation variables within a group, and those variables are frequently all missing for the same

cases, then those imputation variables should be imputed as a block.

The chunks into which the NEHIS imputation variables were partitioned covered the following data areas: fully insured premiums; premium equivalents (self-funded plans); plan enrollments; plan costs; deductibles and co-payments; additional plan-level variables; and additional establishment-level variables.

The complex nature of the NEHIS dataset had a tremendous impact on the order in which the variables were imputed. For instance, health insurance plan costs are a function of plan enrollments and premiums. Therefore, enrollments and premiums were imputed before plan costs. Also, within the chunk consisting of premiums, the employer and employee contributions to the premiums for single coverage were found to be the most highly correlated covariates for the corresponding contributions and premiums for family coverage. Therefore, the single-coverage contributions and premiums were imputed first, and then used in the imputation of their family-coverage counterparts. Several variables within some chunks consisting of premiums and enrollments were imputed simultaneously as a block. Examples of such blocks of variables are the number of retirees under 65 and the number of retirees over 65 for all plans; the employer contributions to premiums and the premiums for single coverage for fully-insured plans in the public sector; and the number of enrollees and the number of enrollees with family coverage for private sector plans.

The imputation process for cost variables for fully insured plans provides an illustration of the importance of a thorough understanding of the structure of the dataset to the formulation of an imputation strategy. For fully insured plans, the plan cost variable of interest is the total annual premium. The total annual premium is defined as the sum of the annual premium estimated from monthly premiums and enrollments and a noise factor. Adding the noise factor was deemed appropriate because the enrollment figures are for a point in time (end of plan year), while the total annual premium is for the entire year. The distribution of the noise factor is expected to be highly skewed and to contain extreme values due primarily to retiree-only plans with large numbers of retirees and due to plans with extremely large enrollments. Therefore, the imputation process started with an exploratory data analysis on the noise factor and various transformations of the factor.

Two types of outliers were identified: *unreasonable outliers*, which arise from the imputation of values that are clearly inconsistent with the reported data; and *reasonable outliers*, which have values that are consistent with the rest of the data but have other characteristics that render them undesirable as donors for imputation. In the case of unreasonable outliers, the imputed values, and other values derived from them, were deleted from the database prior to imputation. No modification of the data was made in the case of reasonable outliers, but the associated plans were excluded from the donor pool during imputation. The imputation process was further complicated by the fact that cost variables such as total annual premium are likely to be highly correlated with the total number of enrollees (active and retiree) and this relationship depends on whether or not a plan has enrollees, and whether or not it is a retiree-only plan. Therefore the imputation of total annual premium was done separately for various subgroups of plans: single service plans, major medical plans with no enrollees, retiree-only major medical plans, and all other major medical plans. Depending on the quality of their covariates, the total annual premium was either imputed directly or constructed by addition after the noise factor is imputed.

4. The Imputation Model and Process

When refining the imputation process, certain summary output for the imputation variables and potential covariates are useful. These include basic frequencies, pairwise correlations, and patterns of missingness. Knowledge about the questionnaire design and subject matter expertise are also valuable. Each of these factors is discussed below.

Basic frequencies for each imputation variable should be reviewed for several reasons. First, they help verify that the correct file is being used, a non-trivial step in complex imputation tasks with several iterations and levels of processing. Second, the nonresponse rates can be used to help choose the imputation method. For example, in NEHIS, modal or mean imputation was used for variables that had less than a two- percent nonresponse rate. Note that care must be taken to remove inapplicable cases from the computation of the nonresponse rate. Third, basic frequencies help identify values that are not missing, but require special handling. For example, inapplicable values could be identified and excluded from the donor pool for Hot Deck imputation. Also, cases with extreme values could be identified and excluded from correlation computations, regression models, and the donor pool in the case of Hot Deck imputation. In some cases, transformations of variables are imputed, and then the variables themselves are derived.

Pairwise correlations also help to refine the imputation process. Two kinds of correlations may be helpful: correlations between an imputation variable and a potential covariate, and correlations between two potential covariates. The correlations between imputation variables and potential covariates may be

used to choose covariates, or to identify variables that can be imputed as a block. Variables being imputed together in a block (i.e., from the same donor) should be highly correlated with common covariates. It may be helpful to review joint frequencies of each pair of variables with a large correlation. Continuous variables may be collapsed into meaningful categories such as "inapplicable", "missing", and "greater than or equal to zero". These cross-tabs are used to refine the list of covariates and the order of imputation, as discussed in Section 3.

Another tool that helps to structure the imputation process is patterns of missingness. For NEHIS, patterns were reviewed for cost variables, enrollment variables, and other variables thought to be potential candidates for block imputation. The patterns of missingness were reviewed to verify that the skip patterns and data editing rules were followed. It is important to identify variables that are inapplicable, if any, in each set since this may help shape the imputation plan. For example, cases with one variable inapplicable may be fundamentally different from other cases, so that different regression models should be developed or different donors should be used in Hot Deck imputation. In NEHIS, inapplicable variables may have indicated establishments that did not offer health insurance or plans that were self-funded, for instance.

In addition to the review of diagnostic output, knowledge of the questionnaire design, survey implementation, and subject matter are essential to developing a good imputation plan. Therefore the statistical staff primarily responsible for conducting the imputation should work closely with key field staff, data processing staff, and subject matter experts. Variables that are fundamental to the survey design or important research or reporting variables may be used as covariates for non-technical reasons (such as face validity).

Decisions were made regarding how consistent the imputed data should be. In NEHIS, the goal was to make the dataset after imputation at least as good as the one before imputation in terms of variable ranges and multivariate relationships between variables. For example, care was taken not to impute data values that were out of range, and to be sure that algebraic relationships among variables were preserved (such as one variable being the sum of three others). These consistency requirements frequently necessitated both an edit-impute cycle and an edit-construct cycle. The dataset before imputation was edited, then missing values were imputed, and then the imputed data were edited again. Any values that failed edits and were set to missing were then imputed. Similarly, during the course of imputation, an impute-construct cycle was implemented for sets of variables with algebraic relationships that needed to be maintained.

To illustrate the most important features of our imputation strategy for complex datasets, we discuss in detail the imputation process for selected cost variables for self-funded plans. Four plan cost variables were imputed for self-funded plans: total plan costs, benefits paid, stop-loss premium, and administrative costs. The imputation was done using Hot Deck, and was subject to the following constraints. First, total plan costs should equal the sum of benefits paid, stop-loss premium, and administrative costs. Second, the stop-loss premium must be nonnegative in general, and greater than zero for plans with enrollees at the end of the plan year. Administrative costs must be greater than zero, and benefits paid must be nonnegative. Third, the total plan cost per enrollee cannot be too small or too large. The limits on the range of total plan cost per enrollee were determined from the reported data, and the lower limit varied by plan type. Fourth, the stop-loss premium per enrollee cannot be too large. The limits were determined from the reported data and varied by the number of employees at the establishment.

The simplest way for the imputed data to meet the first constraint is to impute benefits paid, stop-loss premium, and administrative costs, and then compute total plan costs as the sum of the other three costs. However, stop-loss premium had very weak covariates, and total plan costs and benefits paid had similar strong covariates and were nearly perfectly correlated with each other. Therefore, the general approach was to impute total plan costs and benefits paid as a block and then impute administrative costs and stop-loss premium as follows:

(i) Impute benefits paid for cases with benefits paid and total plan costs missing, using HDV and total enrollment (sum of active enrollees, retirees, and cobra enrollees) as the continuous covariate;

(ii) For the cases from part (i) that now have only total plan costs missing, calculate total plan costs as the sum of the other three costs;

(iii) For the rest of the cases from part (i), impute total plan costs using the same donor that was used in part (i);

(iv) Impute total plan costs for cases with total plan costs missing and benefits paid present, using HDV and benefits paid as the continuous covariate;

(v) Impute benefits paid for cases with benefits paid missing and total plan costs present, using HDV and total plan costs as the continuous covariate;

(vi) Calculate administrative costs (or stop-loss premium) by subtraction where it is the only cost variable still missing; and

(vii) For the cases still needing imputation, which have both stop-loss premium and administrative costs missing, impute administrative costs as a percentage of administrative costs plus stop-loss premium. Then derive administrative costs and stop-loss premium from the imputed percentage.

Note that in place of the last two steps, consideration was given to imputing administrative costs directly and then obtaining stop-loss premium by subtraction. However, it was difficult to impute values for administrative costs that would pass the first edit constraint (that total plan costs equals the sum of the other three costs).

The imputation was done separately by plan type (major medical, other). Splitting the major medical plans by type was not possible in the public sector due to small sample sizes and insufficient donors for some plan types. After each of steps (ii), (iii), and (iv), total plan cost per enrollee was checked and cases with out-of-range values had any imputed cost variables set to missing. Similarly, after steps (vi) and (vii), stop-loss premium per enrollee was checked and cases with out-of-range values had any imputed cost variables set to missing. Imputation procedures were revised to minimize the number of cases with imputed cost variables set to missing. These values were then re-imputed in subsequent rounds of imputation.

5. Summary and Conclusions

This paper discusses the important components of an imputation strategy for large complex survey datasets. The general approach is to review frequencies, correlations, and patterns of missingness to determine the method and order of imputation, to choose covariates, and to decide whether or not block imputation is appropriate. Survey operations staff, data processing staff, and subject matter experts are an integral part of the imputation team. Knowledge of questionnaire skip patterns, the levels at which data are collected, variable construction and edits, and subject matter information should be incorporated into the imputation plan. Imputed data should meet the constraints imposed on reported data. This necessitates the inclusion of an impute-edit cycle in the process. The interdependence of imputation variables necessitates the inclusion of an impute-construct cycle in the process. Both of these cycles represent efforts to maintain consistency in the data.

Analyses on imputed data are preferable to analyses using only completed cases. Using the imputed dataset for analyses offers a reduction in variance over using only the completed cases due to the increase in the number of cases available for analysis in the imputed dataset. However, standard variance estimation methods treat the imputed data as reported data, which underestimates the variance. Variance estimates using such standard methods can be substantially downwardly biased, even for estimators involving variables with relatively low nonresponse rates. Thus, care must be taken to include the variance due to imputation in the variance estimates. Several methods are available for including the variance due to imputation in variance estimates. See Montaquila and Jernigan (1997), and references cited therein, for details.

Distributions on respondent characteristics before and after imputation can provide insight into the effect of imputation. To the extent that the respondent characteristics chosen are related to survey responses, changes in these distributions reflect the effectiveness of imputation.

6. References

Aigner, D.J., Goldberger, A.S., and Kalton, G. (1975). *On the Explanatory Power of Dummy Variable Regressions.* International Economic Review, 16, 2, 503-510.

Little, R.J.A and Rubin, D.B. (1987). *Statistical Analysis with Missing Data*, John Wiley.

Marker, D., Yansaneh, I.S., and Croos, J. (1997). *National Employer Health Insurance Survey (NEHIS) Proposed Methods for Imputing Variables.* Prepared under contract to the National Center for Health Statistics.

Montaquila, J. and Jernigan, R. (1997). *Variance Estimation in the Presence of Imputed Data.* Proceedings of the section on Survey Research Methods, American Statistical Association, 273-278.

Nordholt, E.S. (1998). *Imputation: Methods, Simulation Experiments, and Practical Examples.* International Statistical Review, 66, 2, 157-180.

Schafer. J.L. (1997). *Analysis of Incomplete Multivariate Data.* Chapman & Hall, London, UK.

Acknowledgments

Much of this work was performed under National Center for Health Statistics (NCHS) Contract #200-94-7003 for the Health Care Finance Administration (HCFA), the Agency for Health Care Policy and Research (AHCPR), and the NCHS. The opinions are those of the authors, and do not necessarily reflect those of Westat or the sponsoring agencies.