Measuring and Reducing Inconsistency Among Questionnaire Items Through Imputation: An Application to the NSOPF

Kimberly L. Ault, Mansour Fahimi, Ruth E. Heuer
Kimberly L. Ault, RTI International, 6110 Executive Blvd. Suite 420, Rockville, MD 20852

Contributed Paper

Key Words: Imputation, Editing, Inconsistency
Abstract Word Limit: 200 or 1200 characters

For complex surveys the task of imputing a large number of variables is a major undertaking, since the resulting data must satisfy various consistency checks that are often intertwined. For instance, values of certain variables might have to add up to those of others, while in other cases certain variables might restrict the values other variables can take on. When an observed/imputed of an item is deemed inconsistent, oftentimes logically imputation is used to resolve (edit) the observed inconsistency. Alternatively, such values are set to missing and then imputed to avoid the preceding laborious process. This work provides an overview of a methodology for measuring that can be used for imputation of a large number of items and discusses a technique for measuring and reducing the number of inconsistent cases through imputation. Missing data are imputed using a weighted sequential hot-deck methodology, while ensuring that all values are consistent with respect to all known skip patterns and logical constraints. The research is based on the 2004 National Study of Postsecondary Faculty (NSOPF:04) survey data.
I. Introduction

For complex surveys the task of imputing a large number of variables is a major undertaking, since the resulting data must satisfy various consistency checks that are often intertwined. For instance, values of certain variables might have to add up to those of others, while in other cases certain variables might restrict the values other variables can take on. When an imputed value is deemed inconsistent with other variables, it is sometimes logically imputed (edited) to resolve the observed inconsistency. Alternatively, such values are set to missing and then imputed to avoid the preceding laborious process. This paper discusses a method for measuring and reducing the number of inconsistent cases through imputation, where inconsistency is measured as a function of the number of cases that are set to missing during the editing process. Missing data are imputed using a weighted sequential hot-deck methodology, which result in values that are consistent with respect to all known skip patterns and logical constraints. The research is based on the 2004 National Study of Postsecondary Faculty (NSOPF:04) survey data.

A. Imputation in Complex Datasets

Much of the research on imputation has concentrated on best methods for imputing for a single variate at a time. In large complex datasets, the situation is much harder, because the resulting data must satisfy multiple logical consistencies that are often intertwined. These relationships can take the form of one variable being the sum or ratio of others, or that while not being the exact sum (or ratio), it should approximate that relationship.

When developing an imputation plan, 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. 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.

B. Description of Item Nonresponse and Inconsistency

In surveys, the sampled units may not provide answers for all the survey items, or their responses may be inconsistent and then set to missing. In these cases values are imputed for the missing responses to produce a complete data set.

This paper addresses the issue of quality control or logical edit during or after the imputation for complex surveys. The inter-variable relationship is usually evaluated after imputation, most often manually, or simply ignored. When an imputed value is found to be inconsistent with other variables, it is sometimes manually edited to make it consistent with other variables, or it is set to missing and to be imputed again.

This paper outlines an edit and imputation system that automates the whole process of imputation and logical check and edit. The resulting imputed values are consistent with all known skip patterns and logical constraints. A hierarchical sequential hot deck method is used for the imputation and the program is coded in SAS. The edit and imputation system was applied
to the 2004 National Study of Postsecondary Faculty (NSOPF:04) survey data where approximately 150 survey items were imputed.

**C. Description of NSOPF data**

The 2004 National Study of Postsecondary Faculty (NSOPF:04), conducted by RTI International (RTI) and sponsored by the U.S. Department of Education’s National Center for Education Statistics (NCES), is a nationally representative study that collects data regarding the characteristics, workload, and career paths of full- and part-time postsecondary faculty and instructional staff at public and private not-for-profit 2- and 4-year institutions in the United States. Conducted previously in 1988, 1993, and 1999, it serves a continuing need for data on faculty and other instructional staff. The sample consisted of approximately 35,630 faculty and instructional staff selected from 980 sampled institutions in the 50 states and District of Columbia. The NSOPF:04 data were collected using a self-administered web-based questionnaire with Computer-Assisted Telephone Interview (CATI) nonresponse follow-up.

**II. Imputation Method**

**A. Description of Weighted Sequential Hot Deck Imputation**

The method that was used in this study was a weighted sequential hot-deck imputation. This procedure involves defining imputation classes, which generally consist of a cross-classification of covariates, and then replacing missing values sequentially from a single pass through the survey data within the imputation classes. This procedure takes into account the unequal probabilities of selection in the original sample to specify the expected number of times a particular respondent’s answer will be used as a donor. These expected selection frequencies are specified so that, over repeated applications of the algorithm, the weighted distribution of the all values—imputed and observed—will resemble that of the target universe in expectation. Under this methodology, while each respondent record has a chance to be selected for use as a hot-deck donor, the number of times a respondent record can be used for imputation will be controlled.

To implement the weighted sequential hot-deck procedure, imputation classes and sorting variables that are relevant (strong predictor) for each item being imputed were defined. For this study, imputation classes were developed by using a Chi-squared Automatic Interaction Detection (CHAID) analysis. The CHAID segmentation process divides the data into groups based on the most significant predictor of the item being imputed. Subsequently, this procedure will be repeated using the remaining predictor variables to split each of the emerging groups into smaller subgroups. In this process, a number of subgroups created during a previous iteration might get merged back to form new subgroups. This splitting and merging process continues until no more statistically significant predictors are found, at which point imputation classes are defined from the resulting segments. When dealing with categorical variables, the CHAID process may merge certain categories of such variables that are found not to be significantly different. Similarly, continuous variables are categorized to create the strongest categorical predictors of the item in question.
III. Defining and Measuring Inconsistency

A. Definition of Missing and Inconsistency

For this study, an inconsistent value was defined as a value that conflicts with one or more values reported on the respondent’s questionnaire. Any values in the dataset that were changed were flagged and based on these flags, we could determine which data values were missing due to nonresponse and which values were missing due to inconsistency. To measure the amount of inconsistency in the data before and after imputation, we needed to calculate the following values:

- Let \( M \) = Number of missing (due to nonresponse) before imputation and consistent after imputation
- Let \( M' \) = Number of missing (due to nonresponse) before imputation and inconsistent after imputation
- Let \( I \) = Number of inconsistent (missing due to editing rules) before imputation and consistent after imputation
- Let \( I' \) = Number of inconsistent (missing due to editing rules) before imputation and inconsistent after imputation
- Let \( O \) = Number of observed (or responses) before imputation
- Let \( N \) = Total number of cases (or \( M + M' + I + I' + O \))

To determine the number of values that were consistent and inconsistent after imputation, the imputed data was edited (just as the original data was edited, including performing logical imputations) and was set back to missing if an inconsistency is found.

B. Criteria for Successful Imputation

We defined the imputation procedure to be successful at reducing inconsistency if the following criteria were met:

1) if \( M' < M \) and \( I' < I \).
2) if \( I' \) is substantially less than \( I \). If \( I' \) is only slightly less then the imputation process has not done a very good job of correcting inconsistencies.
3) if \( M' \) is very small. If \( M' \) is only slightly less than \( M \) then the imputation process has generated a very large number of inconsistencies.
4) if the error consistency rate, defined by \( I'/I \), is less than 25%. This rate measures the proportion of the original problem cases that remain problems after imputation.
5) if the nonresponse imputation error rate, defined by \( M'/M \) is less than 25%. This rate measures the proportion of the values imputed for nonresponse that turn out to be inconsistent.

IV. Statement of Problem

A. Method used impute interdependent data items

The questionnaire items were grouped according to skip pattern or conditionality and initially separated into two groups: conditional and unconditional variables. The first group
(unconditional) consisted of variables that applied to all respondents, while the second group (conditional) consisted of variables that applied to only a subset of the respondents. That is, conditional variables were subject to “gate” questions. After this initial grouping, these groups were divided into finer subgroups as detailed next. The unconditional group was divided into two subgroups based on the percent of missing values: less than 1 percent versus greater than 1 percent missing. The conditional variables were divided into three subgroups based on the level of conditionality where this level was essentially determined by the sequence of the questionnaire. For variables in the conditional group, the questionnaire skip patterns were reviewed and variables were grouped according to which variables determine the values of other variables. After these subgroups were constructed, missing values of the variables were imputed in order from lowest percent missing to highest percent missing within each subgroup, first for the unconditional variables and then for the conditional variables in an ascending level of their conditionality.

All unconditional variables that had less than one percent missing were imputed using imputation classes defined by a combination of gender, race, and ethnicity. Moreover, institution type, institution size, and faculty type were used as sort variables to place like records in closer proximity to improve the donor selection process. The imputation classes for the remaining unconditional variables (that had more than one percent missing) and all conditional variables were determined by a CHAID analysis based on key demographic variables and all imputed variables that had less than one percent missing. After all variables were imputed, the imputed data was consistency checks were applied to the entire faculty data file to ensure that the imputed values did not conflict with other questionnaire items, observed or imputed. This process involved reviewing all of the logical imputation and editing rules as well.

To determine the order of imputation, the following factors were taken into consideration:

- If one variable is used in the construction of a second variable, then the first variable should be imputed before the second.
- 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.
- Within groups, variables using deterministic imputation should be imputed before variables requiring stochastic imputation.
- 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.
- 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.
B. Results of imputation- Before and After Imputation

Inconsistency Counts

In planning any imputation process it is important to decide how consistent the imputed data should be. In the NSOPF:04, the goal was to make the data set after imputation at least as good as the one before imputation in terms of allowable 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 data set 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 re-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; once a value was imputed, others could be logically constructed using that value. These constructed variables were then subject to their own edits.

V. Summary and Conclusions

A. Problems encountered/Areas for improvement

This method did not use a repeated edit and imputation cycle that would have been preferable. This method would have edited the data after each group of variables had been imputed. Due to time constraints, the method described previously was the approach taken. If the edit and imputation cycle could have been applied then the number of inconsistencies would have decreased more the approach taken.

Additionally, when the imputation plan was developed the final edit plan was not fully in place. If the imputation plan could have taken into consideration the full edit plan, then again the number of inconsistent cases after imputation could have been decreased.

When future edit and imputation systems are being developed, it is important for the statistician who is performing the imputations and the data editor to work closely together to allow for the two systems to be compatible. What would be most useful is that when the data is flagged as being inconsistent that an additional flag should be set to indicate which edit the value failed on. This would allow the imputation plan to be able to define appropriate donors for each of the imputation variables.

Alternative methods identify strong covariates for use in hot deck or regression imputations. These covariates can be identified through examination of correlation matrices and patterns of missingness. Multivariate relationships can also be maintained by restricting donors to those with responses to an entire set of variables. Using a common donor for a set of variables can assure satisfaction of multivariate constraints across re-imputed variables provided that the set are either jointly missing/reported or that partially reported vectors are overwritten. Logical or numeric constraints can also be retained through repeated imputation-edit cycles. Cyclical hot-deck imputation methods can also be used, which use initial imputations of one variable to improve the imputations of others.
B. Other items related to imputation – effects on variance

The goal of any imputation should be to provide a database containing complete cases allowing for easy, consistent analyses. The resulting database should minimize bias from nonresponse in univariate analyses and attenuation of key multivariate relationships. The greater the resources (statistician skill, knowledge of potential uses, budget, time, etc.) available, the better one will be able to achieve these goals. From the examples in this paper, it is hoped that one will better understand how to address this with finite resources, while still trying to come as close as possible to these ideals.
References
