Measuring and Reducing Inconsistency among Questionnaire Items through Imputation: An Application to the NSOPF

Kimberly L. Ault, Mansour Fahimi, and Ruth E. Heuer
RTI International
6110 Executive Blvd., Suite 902, Rockville, MD 20852

I. Introduction

Most of the research work on imputation has concentrated on improving methods of imputing missing values a single variable at a time. For large complex surveys a more elaborate imputation process is needed, because the resulting data must satisfy various constraints and dependencies that are often intertwined. These constraints can take the form of one variable being the sum or ratio of others, one variable dictating the possible values of other variables, or some other functional relationship. Missing data are among the inevitable facts of survey research. Oftentimes, sampled units do not respond to certain survey items, or fail to complete different sections of a questionnaire. Also, there are instances where respondents provide responses that are inconsistent with others. Such responses are typically set to missing and then imputed along with the originally missing values.

The huge task of imputing several hundred questionnaire items can seem insurmountable because of the complex relationships between questionnaire items. When statisticians develop imputation plans, they must do so by relying on the guidance provided by questionnaire developers and the data editors to ensure that the resulting plan can accommodate all questionnaire logics and edit rules. This paper discusses a procedure for developing an imputation plan that incorporates both the complex questionnaire skip patterns and the numerous edit specifications. The ultimate goal is to develop a simple, repeatable produce where the imputed values are consistent with respect to all known skip patterns and logical constraints. The employed imputation technique is based on a sequential hot-deck method whereby missing values of some 140 variables from the 2004 National Study of Postsecondary Faculty (NSOPF:04) survey are imputed. The success of the methodology is evaluated by measuring the extent of inconsistency and examining bias in estimated means and percentages before and after the imputation.

The NSOPF:04 was conducted by RTI International and sponsored by the U.S. Department of Education's National Center for Education Statistics. It 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. The sample consisted of approximately 35,630 faculty and instructional staff selected from about 1,000 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 nonresponse follow-up.

II. Imputation Methodology

Sequential hot-deck imputation involves defining imputation classes, which generally consist of a cross-classification of covariates related to the imputation item, and then replacing missing values sequentially from a single pass through the survey data within the imputation classes. When this form of imputation is performed using the sampling weights in selection of suitable donors, the procedure is called weighted sequential hot-deck imputation. 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 (Cox, 1980).

In order to implement the weighted sequential hot-deck procedure, imputation classes and sorting variables that are relevant (strong predictor) for each item being imputed need to be 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 is 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 may be 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 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. Imputation Process

While developing the imputation plan for this survey, simple frequencies, correlation coefficients, and different patterns of missing data were reviewed to determine the general order of imputation. Implementation of this step also relied on knowledge of the questionnaire skip patterns and edit rules. The goal was to impute data that met the constraints imposed on reported data. The resulting imputation plan involved the following key steps

1. Calculation of percent missing, the numerator for which included number of inconsistent values set to missing during the edit process and number of missing values due to nonresponse;
2. Determination of the order of imputation;
3. Imputation of questionnaire items;
4. Assessment of inconsistency after imputation;
5. Examination of the imputed data.

Calculation of Percent Missing Data

For each variable that was selected for imputation the total percent missing was calculated. For this purpose the numerator consisted of the number of inconsistent values that were set to missing during the edit process and those resulting from direct nonresponse, while the denominator consisted of the respondent base for each questionnaire item. For this study, an inconsistent value was defined as any value that was in conflict with at least one observed value. Such values were changed to missing and appropriate flags were created to indicate that these values were set to missing due to observed inconsistencies.

An example inconstant data would be when the respondent reported that their year of birth was 1960 and the year they completed their PhD was 1970. For these cases, both data values were set to missing due to inconsistent values. Often, only one data value was set to missing when inconsistent data was reported. For example, the respondent reported having another job at another institution and then reported $0 in compensation for employment. The edit rule assumed the income amount was wrong (because it is a sensitive item) and accepted the response to the other item.

Over 90 percent of the questionnaire items had less than 5 percent missing data due only to nonresponse. Only 21 percent of the items had inconsistent data values that were set to missing. After these missing value percentages were computed for all variables, they were used in developing the imputation groups as described next.

Determination of the Order of Imputation

Developing the order of imputation was the next step in the imputation process. The questionnaire items were segmented according to skip pattern and conditionality. For example, if one variable was to be used in the construction of a second variable, then the first variable was imputed before the second. Also, if the imputed value of one variable determined the value of another variable, then the deterministic imputation was done before the stochastic imputation of that variable.

Initially, variables were separated into two broad groups: unconditional and conditional variables. The unconditional group consisted of variables that applied to all respondents, while the conditional group 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. The unconditional group was divided into two subgroups based on the percent missing: less than one percent versus greater than one percent missing. The conditional variables were divided into three subgroups based on the level of conditionality where this level was essentially determined by the structure of the questionnaire and the edit specifications. Thus, the final number of imputation groups was five, the distribution of which is as follows:

- Group 1 (unconditional less than one percent missing);
- Group 2 (unconditional more than one percent missing;
- Group 3 (conditional level 1;
- Group 4 (conditional level 2); and
- Group 5 (conditional level 3).

Imputation of Questionnaire Items

After the variables were segmented into the above five groups and the order of imputation was determined in each group, the actual imputation began. Prior to execution of the weighted hot-deck imputation, however, the majority of the missing values of the three key demographic variables of gender, race, and ethnicity were retrieved through cold-deck imputation using information from the sampling frame and administrative records. Any remaining missing values for these three variables were logically imputed using respondent names. These variables were imputed prior to any other variables since they were used as key predictors for all other variables.

The imputation method used for all five groups was a weighted sequential hot deck method. All Group 1 variables (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.

After Group 1 variables were imputed, these imputed variables were used in a CHAID analysis to create the imputation classes for the four remaining groups. To find a set of predictor variables for each imputation variable, a CHAID analysis would typically be performed on all potentially correlated variables on the data file. Because the large number of variables in this data file, the time and effort required to find an optimal set of predictor variables for each variable would outweigh any significant difference in the final distributions when the percentage of missing values is low. By using this subset of variables, the CHAID process of defining imputation classes was quickly completed and avoided imputation classes with high ratios of missing to observed cases. After each group was imputed and before the next group was imputed, any logical imputations and/or editing rules were applied.

Assessment of Inconsistency after Imputation

After all missing values were imputed, the resulting data were edited (just as the original data were edited and logically imputed). Subsequently, all newly emerging inconsistent values due to imputation were set back to missing in order to measure the amount of inconsistency that resulted from the imputation process. Using original edit flags (before imputation) and the updated edit flags (after imputation), the following counts were calculated to measure inconsistency before and after imputation:

- MC: Number of missing values (due to nonresponse) that were imputed consistently;
- MI: Number of missing values (due to nonresponse) that were imputed inconsistently;
- IC: Number of inconsistent values (set to missing) and imputed consistently; and
- II: Number of inconsistent values (set to missing) and imputed inconsistently.

Of the 31 variables that had inconsistent values before imputation, 18 variables had inconsistent values after imputation. The variables that had inconsistency after imputation were only variables that had inconsistency before imputation. That is, no new inconsistencies were introduced into the data due to imputation. Table 1 displays the average percent inconsistent before imputation and the average percent inconsistent after imputation. In most cases, the percent of inconsistent values after imputation was less than the percent of inconsistent before imputation. The percent of inconsistent values after imputation ranged from less than 1 percent to about 4 percent. Of the 18 variables with remaining inconsistency, 15 had less than 1 percent of the values being inconsistent. The remaining three variables that had greater than one percent remaining inconsistency were income-related variables and were the variables with the highest percent missing due to nonresponse. Overall, the average percent of inconsistent data before imputation was 1.16 percent and the average percent of inconsistent data after imputation was 0.5 percent

To better quantify these comparisons, the following indicators of relative error were calculated:

$$EI = \frac{II}{IC + II} \qquad EM = \frac{MI}{MC + MI}$$

In the above equations, EM represents the relative error in imputation of missing values due to nonresponse, where EI represents the corresponding error for inconsistent values set to missing during the initial edit phase. Table 2 shows these relative errors for the variables with remaining inconsistency. In most cases, the relative error in imputation due to nonresponse was less than the relative error due to inconsistency which supported the goal of the imputation process. Overall, the average relative error due to nonresponse was about 5 percent and the average relative error due to inconsistency was 23.6 percent.

After this evaluation, most of these inconsistencies were fixed manually through a series of conditional statements. Additional checks were performed to ensure that these manual changes to the data were accurate.

Examination of Imputed Data

Another measure of success of imputation has to do with how much bias is reduced in as a result of imputation, as one of the goals of imputation is to reduce the bias of survey estimates. This goal is achieved to the extent that systematic patterns of item nonresponse are correctly identified and modeled. For continuous variables, the estimated bias was calculated as the mean before imputation minus the mean after imputation. For categorical variables, the estimated bias was computed for each category as the percentage of faculty members in that category before imputation minus the corresponding percentage after imputation. The estimated bias was then tested (adjusting for multiple comparisons) to determine if the bias was significant at the 5 percent level. A categorical variable was deemed significantly biased if the bias for any of its categories was significant. The variables selected for this analysis were ones that had less than 85 percent response rate. For most variables examined, the bias after imputation is not significant. However, a few variables continue to have significant bias after imputation.

IV. Summary

The process outlined in the paper did not use a repetitive imputation and edit cycle between each imputation group where values that were inconsistently imputed would have been set back to missing and then re-imputed. This was not done since the possibility existed that there would be an infinite cycle of edit-impute. Additionally, the covariates used in the CHAID analysis were limited to a selected group. If other covariates had been used, the more precise imputation classes could have been developed and may have also helped with the level of inconsistency. However, this method of using a limited set of predictor variables proved to be a quick, reliable, and efficient method for imputing a large number of variables with low levels of missing values.

This imputation plan provides an example of how large-scale imputations can be performed and how inconsistency can be measured. Most imputation methods address simple abstract versions of the real problem, but they ignore the complicated and essentially unstructured logical relationships among survey items. A major challenge in devising general solutions for editing and imputation is to implement data checking and validation during the imputation process.

Table 1. Percent Inconsistent Before Imputation		
Less Than 1%	Number of Items	18
	Average Percent Inconsistent Before Imputation	0.18%
	Average Percent Inconsistent After Imputation	0.15%
Between 1% and 5%	Number of Items	12
	Average Percent Inconsistent Before Imputation	2.28%
	Average Percent Inconsistent After Imputation	0.74%
Total Average Percent Inconsistent Before Imputation		1.16%
Total Average Percent Inconsistent After Imputation		0.50%

Table 2. Percent Inconsistent Before Imputation		
Less Than 1%	Number of Items	18
	Average Relative Error - Nonresponse	5.50%
	Average Relative Error - Inconsistency	24.75%
Between 1% and 5%	Number of Items	12
	Average Relative Error - Nonresponse	4.21%
	Average Relative Error - Inconsistency	19.48%
Total Average Relative Error - Nonresponse		5.40%
Total Average Relative Error - Inconsistency		23.61%

References

Chromy, James R. (1979). Sequential Sample Selection Methods. Proceedings of the section on Survey Research Methods, American Statistical Association, pp.401-406.

Cox, Brenda. (1980). The Weighted Sequential Hot Deck Imputation Procedure. Proceedings of the section on Survey Research Methods, American Statistical Association, pp.721-726.

Fellegi, I.P. and Holt, D. (1976). "A Systematic Approach to Automatic Edit and Imputation". Journal of the American Statistical Association, 71, pp.17-35.

Huang, Rong and Yen, Wei. (2004) Embedding Logical Check and Edit in an Automated Hot-Deck Imputation of Survey Data. Proceedings of the section on Survey Research Methods, American Statistical Association, pp.3685-3688.

Iannacchione, Vince. (1982). Weighted Sequential Hot Deck Imputation Macros. Presented at the Seventh Annual SAS User's Group International Conference, San Francisco, CA.

U.S. Department of Education, National Center for Education Statistics. Imputation Methodology for the National Postsecondary Student Aid Study: 2004. NCES 2003–20, by Kimberly Ault, Stephen Black, Jim Chromy, Mansour Fahimi, Peter Siegel, Lily Trofimovich, Roy Whitmore, and Lutz Berkner. Project Officer: James Griffith. Washington, DC: 2003

Yansaneh, Ibrahim, Wallace, Leslie, and Marker, David. (1998). Imputation Methods for Large Complex Datasets: An Application to the NEHIS. Proceedings of the section on Survey Research Methods, American Statistical Association, pp.314-319.