Embedding Logical Check and Edit in an Automated Hot-Deck Imputation of Survey Data

Rong Huang, Wei Yen
UCLA Center for Health Policy Research, 10911 Weyburn Ave., Suite 300, Los Angeles, CA 90024

KEY WORDS: Logical Check, Automated Edit and Imputation, Quality Control, Hot Deck Imputation, Survey

1. Introduction

In surveys, the sampled units may not provide answers for all the survey items, or their responses may be inconsistent. In these cases values are imputed for the missing responses to produce a complete data set. In recent years there has been a growing amount of theoretical and empirical work on imputation, but little work addresses the quality control or logical edit during or after the imputation, especially 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 article introduces an imputation system that automates the whole process of imputation, 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 is applied to the California Health Interview Survey (CHIS) 2003 data file.

2. The Problem

The California Health Interview Survey (CHIS) is the largest state health survey with the aim of providing timely information on priority health issues affecting the state's population. The survey addresses the health status of individuals and covers some major health conditions, for example, asthma, diabetes, cancer, and heart disease. The survey also asks about many health-related behaviors important for disease prevention.

As in the cases with any household survey containing questions about sensitive information, such as income, insurance, and health conditions, the CHIS encountered significant levels of item nonresponses to some questions. For most questions, the item nonresponse rates were very low, often less than 1 percent. But the nonreponse rate for income is around 25 percent.

Nonresponses in survey data come from three ways:

1. The respondent may not know or refuse to give the information.
2. The interview gets interrupted and the information is lost.
3. The interviewer forgets to ask a question or record a response.

Other than those nonresponses, information is lost when errors occur. They are caused by three reasons:

1. The respondent gets confused and gives inaccurate information.
2. The interviewer may misunderstand the answer or miscode the response.
3. The item values for one item contradict with other item(s).

The nonresponses and erroneous values constitute missing values in survey data. In CHIS 2001 survey data, all the items with missing value were not imputed except the key variables used for weights calibration, such as ethnicity, age, and strata. Only a few very important construct variables, which were derived from several survey item variables, were imputed. This approach was chosen because of its efficiency – only a few variables need to be imputed. But it has turned out this approach is not desirable in the long run. (1) The imputation at the construct variable level had resulted in potential inconsistencies among construct variables. This was mainly due to different interpretation and treatment of item missing values. The risk of inter-construct-variable inconsistency increased with the number of imputed construct variables. (2) Most item nonresponses were not imputed. Therefore, the statistician and programmers had spent a lot of time to communicate with researchers to get the appropriate algorithms for different projects. Despite the tremendous effort, there were still similar but not the exact same variables constructed for the same concept in different research projects. This wasted time and introduced confusion. (3) Observations with missing value for variables of interest are by default excluded from analysis for simplicity. For example, the online

query system AskCHIS (www.AskCHIS.com), which provides interactive access to CHIS 2001 results, offers the estimates based on the complete cases. The total population in one query of asthma by gender is different from the total population in the other query of heart disease by gender. This is because the missing cases of one variable asthma are different from those of the variable heart disease. This will confuse the users. Furthermore, the estimates based on complete cases may introduce bias if the missing pattern is not MCAR (missing completely at random).

The other option to deal with missing data is to impute most missing data, except some sensitive variables where those nonresponses have their own meaning, such as the question about sexual orientation. However, most current imputation methods ignore the logical constraint and complicated skip patterns during the imputation. The imputed value might violate the logical constraint and thus need to be re-imputed or edited. The quality checks and edits after imputation are usually practiced manually. And it consumes large amount of time when the time constraints to release the survey data are severe. Granquist (1997) estimated that the data cleaning could use up to 40 percent of the total resources spent on a survey. In addition, some logical inconsistency may not be detected on time due to the non-systematical nature of the process.

The other challenge for skip check is related to multiple imputation. Hot deck imputation is the most practiced procedure where recorded units in the sample are used to substitute missing values. Although the imputed data set is easy to use, there will be inference bias if the imputed data set is analyzed in the same way as a complete data set. Rubin (1978) proposed to use multiple imputation as a way of getting valid inference. Instead of filling in a single value for each missing value, Rubin's multiple imputation strategy replaces each missing value with two or more plausible values that represent the uncertainty about the impute value. Each of the two or more resulting complete data sets is then analyzed using standard complete-data methods. These analyses are combined to reflect both within-imputation variability and between-imputation variability. The logical check, edit, and re-imputation for each imputed data set will consume tremendous effort and time if the whole process cannot be automated. This is one great obstacle for multiple imputation of large-scale survey data in real world.

3. Framework of Logical Correction during Sequential Imputation

In CHIS 2003 data processing, data quality control is executed before the imputation for two reasons. The first reason is to identify internally inconsistent items and set some items as missing for imputation. The second reason is to make sure all the values from donor items are valid. Although the collaborating survey company, Westat, implements the complex skip pattern in the Computer Assisted Interview (CATI) system, it is still well worth the effort to apply the logical check and the skip pattern check program to the survey data as quality assurance. Since logical error detection are performed in pre-imputation, during-imputation, and after-imputation stages, the SAS program for logical error detection and correction are coded in the way that the program can be shared in those three stages. It not only improves the efficiency and consistency across the three stages, but also scrutinizes the rules and the implementation of those rules rigorously across the three stages.

Automatic logical correction within the imputation is composed of four steps. The first step is to create auxiliary variables that will be used for error identification and range check. For example, the poverty threshold variable is constructed based on the reported household income and household size. That variable will be used to determine whether certain questions about government-provided insurance programs will be asked or skipped. Another example is that the range for household income is calculated based on the bracket answer to income if the exact number is not provided and will be imputed.

The second step is to define each logical check site, which can be composed of one or more survey items, depending on whether those variables share the same logical conditions. A set of specific rules for each logical check site are defined based on skip conditions documented in the questionnaire as well as derived constraints from other response items. Each logical check site can test several sequential survey items. There are five elements involved in each logical check: (1) the condition statement for "should have response" which requires a response other than "refusal", "don't known", "not ascertained", or "skipped"; (2) the range constraint for the value of that item; (3) the condition statement for "should be skipped" which demands "skip" as the item value; (4) the variables that share the same condition for "should have response"; (5) and the variables that share the same condition for "should be skipped".

The third step is to define edit rules after error identification. If the survey item "should have response" but it does not, then it will be assigned "not ascertained" and will be imputed later. If the survey item "should be skipped" but it does not, then it will be assigned as "skipped". If its value is out of range, then it

is will be assigned "not ascertained" and will be imputed within the range constraint.

The rules of logical check at the second step and rules of edit at the third step are thoroughly evaluated on each check site at the pre-imputation stage. They can be programmed based on the questionnaire before the raw data is available. The logical check and edit serves four distinctive purposes: (1) to communicate with the survey agency, Westat, on the implementation and modification of the designed questionnaire and skip patterns; (2) to improve the accuracy of the data and the questionnaire by send back the errors detected by our logical check program; (3) to set up legitimate universe and range constraint for each survey item for imputation at the next stage; (4) to obtain information for future improvement to the data collection process. The rules for logical checks are defined based on the questionnaire structure, designed skip pattern, and field expertise. Although Fellegi (1976) and Bankier (1999) proposed algorithms that make data to satisfy all edits by changing the fewest (weighted) number of variables, those algorithms are not appropriate for complex survey data like CHIS because they are not sensitive to the complicated skip conditions. And the imputed value for the nonresponses will be biased toward nonmissing values that share the same or similar skip conditions as those nonresponses. For example, question AK1 ("which of the following were you doing last week?") is asked in CHIS 2003 adult survey. If the answer for AK1 is "1. working at a job", "-7. refused", or "-8. don't know", the next questions AK2 ("what is the main reason you did not work last week?") and AL22 ("are you receiving social security disability insurance?") will be skipped. If the answer for AK1 is "3. looking for a job", then the next AK2, AL22 and AK4 are all skipped. If the answer for AK1 is "2. with a job but not at work" or "4.not working at a job", then AK2 is the next question. And the following question AL22 will be skipped if the answer for AK2 is "2. vacation or leave" or "8. on layoff or strike". If we adopt Fellegi (1976) or Bankier (1999)'s algorithm to impute the nonresponses ("-7. refused" or "-8. don't know") for AK1, the imputed value will be biased toward "1. working at a job" because this imputed value will not modify any other variables in the survey data. Similarly, the imputation of nonresponses for AK2 will be biased toward value 2 and 8. Our procedure not only avoids that kind of bias, but also displays the error type, the variables involved in that error, and the recommended edit. Therefore, it is convenient for review and validation before any imputation is executed.

After that, the forth step is to compile all the individual checks together so that the whole survey data can be checked and edited simultaneously. The structure of the check and edit procedure can accommodate as many logical check conditions as possible, and rules for check and edit can be easily evaluated, modified, added or deleted for each check site. And the continuous as well as categorical variables can be examined at the same time.

The logical check and edit are carried out before and after the imputation for each survey items as shown in Figure 1. The raw data is first checked and edited by the logical check program. This will ensure the validity of all the donor values. Then one survey variable is imputed with the help of auxiliary variable to define the range of imputed value for each observation. After imputation, the whole data file is checked to identify whether the other survey variables are inconsistent with the imputed value of the current variable. If there is any inconsistency, then the logical check program will output the edit program and modify the data accordingly. The most frequently identified inconsistencies are related to the variables' universe, and the edits will either assign "-9" to those items for imputation later, or assign "-1" as legitimate skip. The logical check and edit are iterated until all variables are consistent. Then the imputation, check, and edit are repeated for the next variable.

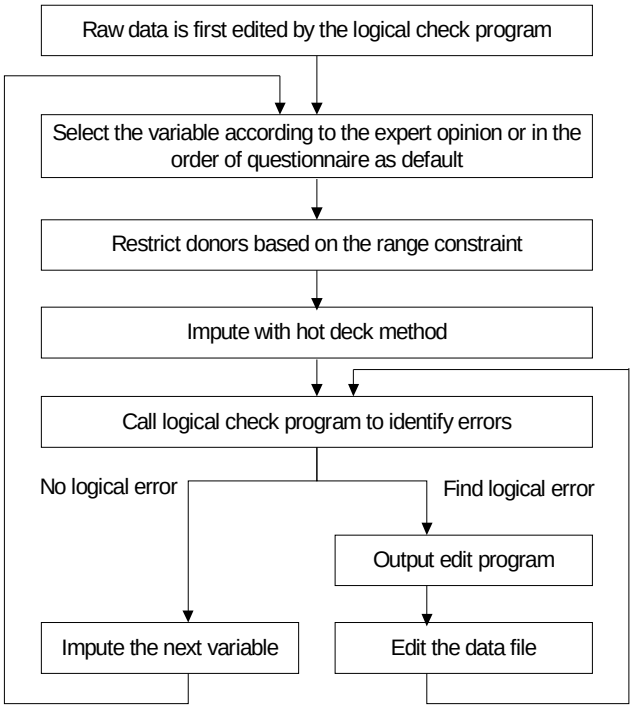

Fig 1. Automated logic check/edit and imputation

Table 1 shows that the number of logical inconsistencies are positively correlated with the number of variables in the questionnaire, which also associates with the complexity of the survey structure. The most complicated part for CHIS 2003 is the survey for the adult. Although Westat provides us a high-quality data file for adult survey, there are still logical inconsistencies detected in 0.18% of the total cells in the data file. Out of the total 613 variables in the adult data file that need to be imputed, there are logical inconsistencies detected for 286 variables at the pre-imputation stage. Those edits are reviewed and validated before imputation. There are totally 15,207 observations and 47,102 cells need editing. During the imputation, there are 53492 cells (0.21% of the total cells) get edited, and the edits affect 424 variables and 8379 observations. For each edit, the type of the inconsistency, which round of imputation causes that inconsistency, the variables involved in the edit, and the values before edit are all saved for future review. Those editing metadata is not only important for data processing programmers, but also for the data users in the future. The editing process is automatically and fully documented with the reasons for changes.

Table 1: Automatic error detection and edit before and during imputation

	Data File	Number of variables *	Number of observations *	Number of cells *
Before Imputation	Adult	286 (46.7%)	15207 (36.2%)	47102 (0.18%)
	Adolescent	10 (6.7%)	572 (14.3%)	670 (0.11%)
	Child	80 (49.1%)	1430 (16.8%)	2963 (0.21%)
During Imputation	Adult	424 (69.2%)	8379 (20.0%)	53492 (0.21%)
	Adolescent	57 (38.3%)	448 (11.2%)	1699 (0.28%)
	Child	76 (46.6%)	257 (3.0%)	545 (0.04%)

* The number in the parenthesis shows the percentage of affected variables, observations or cells from the respective totals in the survey data.

4. Discussion

A major challenge in imputation is the consistency among imputed and non-imputed values. Most imputation methods address simple abstract versions of the real problem, but they ignore the complicated and essentially unstructured logical relationship 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. The article presents a systematic detection and edit of logical inconsistency during the process of imputation.

Although the system is embedded with hot deck method for sequential single imputation, it can be extended to accommodate multiple imputation if different random numbers are used to impute different data sets. It works well with methods that impute each variable sequentially. And it can also be integrated with methods that impute a set of variables simultaneously, but the imputed data set will need to be checked, edited, and re-imputed iteratively until all survey variables are consistent. Whatever the imputation method is, the author recommends the system for logical check and edit before and after imputation.

5. References

Bankier, M., Lachance, M. and Poirier, P. (1999). "A generic implementation of the New Imputation Methodology," *Proceedings of the Survey Research Methods Section, American Statistical Association*, 548-553.

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

Granquist, L. (1997), "The New View on Editing," *International Statistical Review*, 65, 381-387.

Granquist, L. (1997), "An Overview of Methods of Evaluating Data Editing Procedures," *Statistical editing, Vol.2, Methods and Techniques. Statistical Standards and Studies*, 48, 112-122.

Rubin, D.B. (1978), "Multiple Imputations in Sample Surveys – A Phenomenological Bayesian Approach to Nonresponse," *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 20-28.