Item Imputation Made Easy

Andrea Piesse, David Judkins, and Zizhong Fan
Westat

Keywords: Hot deck, Iterative imputation, Predictive mean matching, Semi-parametric method

1. Introduction

This paper presents the item imputation approach used for the Youth Media Campaign Longitudinal Survey (YMCLS). This survey is the primary vehicle for assessing the impact of a national multimedia campaign to encourage 9- to 13-year-olds to be physically active every day. The YMCLS produces estimates of campaign effects on attitudes and behavior related to physical activity, as well as estimates of overall trends. As with any survey, respondents do not always provide answers to all items. We have developed a highly automated imputation approach to handle all items where missing data need to be filled in. The method draws primarily on the concept of iterative imputation via cyclic n-partition hot decks (Judkins, 1997); it may be viewed as a semi-parametric equivalent of Bayesian Monte Carlo Markov chain methods, such as the Gibbs sampler (see Marker, Judkins, and Winglee, 2001). The approach described here is a two-step process involving predictive mean matching. Central to the automation is the development of a master index file that describes the attributes of all items in the input dataset. The end result is a complete dataset of analytically relevant items.

2. YMCLS Background

In June 2002, the Centers for Disease Control and Prevention (CDC) launched a national campaign to encourage 9- to 13-year-olds to be physically active, using the brand and tagline, "VERB™ *It's What You Do*." The YMCLS was designed to assess the impact of the campaign and uses a stratified, list-assisted, random-digit dialed sample of households, from which a parent and at most two youth aged 9 to 13 years were selected. The survey has been conducted four times. The baseline survey was conducted in the middle of 2002, prior to any VERB advertising, and repeated midway through 2003, 2004, and 2005. Initially, two stratified random samples were drawn: a community and a national sample (subsequently referred to as Panel 1). In 2004, a new national sample (subsequently referred to as Panel 2) was selected independently of the existing national sample. The Panel 2 sample was drawn for three main reasons: 1) a considerable proportion of the original sample of 9- to 13-year-olds had aged out of the campaign's target population; 2) the size of the original sample had decreased through attrition; and 3) there was a desire to assess whether any "time-in-sample" bias existed in the original sample.

This paper primarily describes the automated item imputation approach used on the Panel 1 sample in 2004. Some comparisons with the performance of the previously developed item imputation approach, used on the Panel 2 sample in 2004, are given in Section 5.

3. Imputation Requirements

Imputed data were required for approximately 160 parent and youth items from each panel in 2004. Given the large number of items, we sought to develop an imputation methodology that could be largely automated, yet improve the quality of the imputed data. The objectives were to impute all items in a single software run, to adhere to "skip patterns", to preserve distributions, and to preserve covariance structures within waves of data (including between parent and youth items) as well as between waves of data. Skip patterns refer to the situation where a respondent's answer to one or more previous items determines whether or not they are asked a subsequent item. For example, a youth who reported not having seen any of the campaign advertising would not then be asked about the amount of advertising seen.

4. The Automated Approach

Our approach consisted of four main steps:

1. Create a master index file describing the attributes of all items on the input dataset;
2. Impute data for items using a preliminary hot deck;
3. Re-impute data for items using nearest neighbor matching based on predicted means; and
4. Review the summary output.

4.1 Master Index File

In order to automate the item imputation process, there is a clear need for a mechanism through which controlling parameters can be passed to the software to be used. For example, the items to be imputed must be distinguished from those for which imputation is not required. In what order should these items be imputed? One of the objectives for the automated methodology was to preserve skip patterns in the data. Since skip patterns are almost always governed in part by the order in which the items occur in

the survey instrument, another field on the master index file stores the questionnaire order of the items. This order then determines the order of imputation in the automated process. Another factor relevant to the preservation of skip patterns is the knowledge of the "skip controllers" for each item. In other words, which of the previous items determine whether or not a given item will be asked of the respondent? A field for skip controllers is also on the master index file. It is important to note that both the questionnaire order and the skip controllers for each item can be determined as soon as the survey instrument is finalized, generally well in advance of the need for item imputation. In addition, the completion of these fields on the master index file can be (and is perhaps better) handled by subject area staff, presenting an opportunity for closer liaison with the statistician during the imputation process.

The second step in the automated approach developed for the YMCLS is the imputation of items using a preliminary hot deck. In this step, a non-missing response from a "donor" is used as the imputed value for an individual with a missing response for that item. The donor is chosen so as to resemble the recipient on a particular set of characteristics. For some characteristics, it is deemed important that the donor and recipient are an exact match; these characteristics are said to constitute a "hard boundary" in the hot deck imputation. For other characteristics, it may be sufficient that the donor and recipient are as similar as possible, without requiring an exact match. These characteristics are said to constitute a "soft boundary." The automated imputation process requires two more fields on the master index file to identify the auxiliary variables to be used as the hard and soft boundaries in the preliminary hot deck. There is an additional requirement that the set of auxiliary variables specified is the same for each item to be imputed. Note that these auxiliary variables must be complete, i.e. non-missing, in the input dataset.

Other fields on the master index file specify the set of missing data codes for each item requiring imputation, the measurement level (nominal, binary, ordinal, or continuous) and variable type (character or numeric) for every variable, and an indicator of whether or not each variable is an eligible predictor in the modeling phase (step two of the iterative imputation procedure). Other parameters that are not item specific, such as those controlling the model selection criteria, can be passed directly to the automated imputation software and need not be part of the master index file.

4.2 Preliminary Hot Deck

Step one of the imputation approach uses a preliminary hot deck to produce a complete dataset for use in step two. As described in Section 4.1, the user must specify a common set of auxiliary variables to be used as the hard and soft boundaries for each item imputed by the preliminary hot deck. These auxiliary variables must be non-missing in the input dataset and so may come from the survey frame, if necessary. In addition to the user-specified auxiliary variables, any skip controllers for an individual item are included as part of the hard boundary. Treating the skip controllers as hard boundaries, along with the order of imputation specified in the master index file, ensures that the preliminary imputed values adhere to the skip patterns in the data.

4.3 Predictive Mean Matching

Hot deck imputation has long been used to handle missing data (see Kalton and Kasprzyk, 1986). The method is fairly simple and relatively low in cost. The sets of variables used to determine the hard and soft boundaries are often chosen a priori. However, when the number of items to be imputed is large, it is unlikely that the user will be able or have the time to choose optimal hot deck boundaries for each item. Here "optimal" refers to a choice of boundary variables that are strongly associated with both nonresponse propensity and the item to be imputed itself. Such a choice will result in reduced nonresponse bias in estimates involving the imputed item.

An alternative to the simple hot deck that attempts to tailor the imputation process towards each specific item is predictive mean matching. This method involves modeling the item to be imputed in terms of a set of eligible predictors. Based on the model, predicted means are then calculated for both records where the item is missing and records where it is non-missing. Donors for those records requiring imputation are selected by matching on the predicted means, according to some specified distance metric. The imputed value is then the value of the item on the donor record. One advantage of this approach is that a large number of eligible predictors can be considered in the modeling step. The use of this extra information should result in improved quality of the imputed data. Since the final imputed value comes from a matched donor record, the method protects against model mis-specification (e.g., error distribution and homoscedasticity assumptions) and prevents the imputation of impossible values.

Step two of our automated imputation approach exploits the fully imputed dataset created by the preliminary hot deck. Again, items are imputed in the order specified in the master index file. Binary, ordinal, and continuous items are individually modeled using stepwise linear regression with the eligible predictors identified in the master index file. Predicted means are then calculated for each item. The actual matching was implemented via a hot deck where the predicted mean constituted the soft boundary, and any skip controllers for the item constituted the hard boundary. The use of this hard boundary, along with the specified order of imputation, ensures that the imputed values adhere to the skip patterns in the data. The soft boundary specification attempts to match donors and imputation recipients on their

predicted mean values from the model, but accepts a "nearest neighbor" when no exact match exists. Linear regression models are used for binary and ordinal items despite their mis-specification in order to improve the run time of the automated approach. The models are restricted to main effects for the same reason.

Items of other data types, i.e. unordered categorical items with three or more levels, are re-imputed in step two using the preliminary hot deck. These items require re-imputation because of the possibility that the values of some or all of the item's skip controllers may have changed when those skip controllers were themselves re-imputed earlier in step two of the process.

4.4 Review of Imputation Output

Although the imputation approach is highly automated, and perhaps particularly for that reason, there is still a need for careful review of the output dataset. Whatever software might be used to implement the method, some basic comparisons of the data items before and after imputation are required. In particular, the summary should report any skip pattern violations in the input and/or output dataset. The report on the input dataset serves as an additional quality control measure to the edit checking step that should already have taken place.

5. Automated Imputation for YMCLS

The automated imputation approach was implemented for the first time in 2004, by applying a SAS macro to data from Panel 1 of the YMCLS. Approximately 160 parent and youth items required imputation, with a median item response rate for imputed items from the Panel 1 parent and child interviews of 99.8 percent. Items with the highest missing rates were parent and youth responses to, "What is the name of the message or advertising?" (74% and 89%, respectively). Subject area staff created the master index file and populated the questionnaire order, skip controllers, and missing data code fields. Statisticians then entered the hard and soft boundaries, eligible predictor indicators, data types, etc. Since the Panel 1 parents and youth had been in the survey since 2002, two years' worth of longitudinal data was available for use in the imputation, as well as the cross-sectional data. The automated approach allowed the richness of these data to be exploited without delaying the project time frame. The availability of complete data from previous years also presented a number of choices of auxiliary variables for use in the preliminary hot deck.

Although one of the objectives was to develop a procedure that could impute all items in a single software run, we imputed parent-specific and youth-specific items separately. This was done for logistical reasons—we did not want to impute different values for items about the household or the parent when there were two records in the input dataset corresponding to the parent's two sampled youth. Following the order of instrument administration, the parent-specific items were imputed first. The eligible predictors included both parent and youth data from previous years of the survey, as well as 2004 parent and youth items. These results were then used to update the input dataset for the imputation of youth-specific items. Again, eligible predictors included both parent and youth longitudinal data, as well as 2004 parent and youth items.

Review of the output dataset and summary imputation reports identified a very small number of skip pattern failures that were later traced to last minute recoding of "other specify" type responses in the input dataset. These edits were corrected and the imputation re-run. The final output dataset, containing the results of parent and youth imputation, adhered to all skip patterns in the data and passed all other quality control checks.

In contrast, imputations for largely the same set of parent and youth items for Panel 2 of the YMCLS were conducted using previously developed hot deck routines, with some modifications to account for new or deleted data items, etc. These imputations were also implemented using SAS, but due to the large number of imputed items, the implementation was spread across approximately ten separate programs. In 2004, the median item response rate for imputed items from the Panel 2 parent and child interviews was 99.7 percent. Items with the highest missing rates were household income (88%) and parent and youth responses to, "What is the name of the message or advertising?" (66% and 79% respectively).

Despite initial challenges in setting up the master index file for the Panel 1 imputation, the overall integration facilitated by the automated approach was a great advantage compared to the Panel 2 method. In particular, review of the Panel 1 output required far less time than the quality control of the individual programs developed for Panel 2. Ensuring the adherence of the imputed Panel 2 data to skip patterns was also more difficult to manage. Although most of the hot deck routines for Panel 2 had been developed in previous years, imputation for Panel 1 was completed in a shorter time frame. The real gains from the automated approach are achieved when the master index file is largely prepared in advance.

6. Summary and Future Work

The imputation method described here can be applied to a wide range of surveys—cross-sectional, panel, in-person interview, RDD, etc.

The method is especially useful when a large number of items require imputation. In such situations, the amount of time and effort involved in creating a master index file and automating the imputation process is considerably less than that required to devise and execute an individually tailored imputation strategy for each item. In addition, the quality of

the imputed data resulting from the item-specific modeling approach is likely to be superior to what could be achieved in the same time frame using more traditional approaches. The improvement in quality may be particularly strong for panel surveys where there is a wealth of data available. The opportunity to involve subject area staff in the creation of the master index file also encourages closer collaboration among members of the project team.

Improvements to the automated imputation approach described here are already in development, but the basic foundation remains the same. An obvious next step is to increase the number of iterations of modeling and re-imputation, until some convergence criteria is satisfied. As the number of iterations increases the final imputed values should become less dependent on the preliminary imputation. Alternative modeling options could be explored, such as tree-based algorithms, or the introduction of models for unordered categorical items, searches for interaction terms, etc. A common donor option could be introduced to accommodate situations where it is required that the same donor be used for a given set of items. Matching on some coarsened version of the predicted means might result in reduced variances due to larger donor pools, and could perhaps pave the way for the incorporation of a multiple imputation option (see Rubin, 1987).

References

Judkins, D. (1997). Imputing for Swiss Cheese Patterns of Missing Data. *Proceedings of Statistics Canada Symposium '97*, pp. 143-148.

Kalton, G. and Kasprzyk, D. (1986). The Treatment of Missing Survey Data. *Survey Methodology,* Vol. 12, No. 1: 1-16. Kovar and Whitridge in Cox et al., 1995.

Marker, D. A., Judkins, D., and Winglee, M. (2001). *Large-scale Imputation for Complex Surveys*. In Groves, R. M., Dillman, D. A., Eltinge, J. L., and Little, R. J. A. (Eds) Survey Nonresponse. Wiley and Sons, N.Y.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*, Wiley, NY.