PREDICTIVE MEAN NEIGHBORHOOD IMPUTATION WITH APPLICATION TO THE PERSON-PAIR DATA OF THE NATIONAL HOUSEHOLD SURVEY ON DRUG ABUSE

A.C. Singh, E.A. Grau, and R.E. Folsom, Jr.
Statistics Research Division, RTI International, RTP, NC 27709-2194

Key Words: Multivariate Imputation; Model-based Imputation; Nearest Neighbor Hot Deck; Predictive Mean; Multiplicity Factors; Pair Data Analysis

Introduction

In 1999, the instrument used to administer the National Household Survey on Drug Abuse (NHSDA) was changed from a paper and pencil format (PAPI) to a computer assisted format (CAI). In previous years, imputation of missing values for most of the drug use variables was accomplished with an unweighted sequential hot deck. For other variables, including person-pair data, no imputation was attempted at all. In the spirit of efforts to improve the quality of estimates from the redesigned NHSDA, and as a result of fundamental differences between PAPI and CAI, there was a need to change the way missing data were edited and imputed. Changes in the editing rules from PAPI to CAI put more of a burden on statistical imputation for resolving inconsistent values. These rules are referred to as "flag and impute", where ambiguous or inconsistent responses are flagged and replaced with consistent values in imputation. In addition, imputation was required for more variables in CAI. Finally, many of the variables in the NHSDA are closely related to each other, often in a hierarchical manner. These points all illustrate that the need for a method that was both rigorous, flexible, and preferably multivariate. This paper presents a new imputation method with these characteristics, termed Predictive Mean Neighborhoods (PMN), that was used to impute missing values in many variables in the NHSDA, including both drug use variables and variables derived from the person-pair data.

We give some background of the NHSDA in the following section, including the motivation for using a new methodology, both in general and for the person-pair data in particular. The PMN method is introduced next, followed by some details of the application of the methodology to the person-pair data. A comparison of the PMN method with other methods follows, with a final section providing some concluding remarks and suggestions for further work

Background of NHSDA Pair Data and Motivation for New Method

The NHSDA is a multistage stratified cluster sample survey of 66,706 respondents, which includes questions about drug use behavior and general mental health. From a particular household that is screened, either 0, 1, or 2 persons were selected to be interviewed, where a pair is defined by the selection of 2 persons. In addition to basic demographics, drug use variables, and other variables in the sample, each respondent provides a household composition roster, with a list of household members and their relationship to the respondent. By matching the household rosters of both respondents in a pair, it is possible to determine the relationship between pair members (mother-child, sibling-sibling, etc) in each respondent's household. In addition, for analysis of these pairs it is necessary to have the appropriate pair weights, defined according to pair domains of interest. These pair domains are listed below:

1. Parent-child aged 12-17, with focus on parent
2. Parent-child aged 12-17, with focus on child
3. Parent-child aged 12-14, with focus on parent
4. Parent-child aged 12-14, with focus on child
5. Parent-child aged 15-17, with focus on parent
6. Parent-child aged 15-17, with focus on child
7. Parent-child aged 12-20, with focus on parent
8. Sibling aged 12-14, sibling aged 15-17, with focus on sibling aged 15-17
9. Sibling aged 12-17, sibling aged 18-25, with focus on sibling aged 18-25

In order to appropriately adjust the pair weights (see Chromy and Singh, 2001), counts of these pair domains are required both at a pair level and at a household level. In particular, multiplicities are counts of the number of possible pair domains for a selected respondent who is the focus in a particular pair domain. Household-level pair domain counts are of the number of individuals in a household that could belong to each pair domain, given a particular respondent who is the focus in a domain was selected. In the CAI sample of the NHSDA, imputation was required for basic demographic variables (including income and insurance status), drug use variables, household composition variables (e.g., household size), pair relationship indicators, multiplicities (the pair domain counts at a pair level), and the household level pair domain counts.

In previous NHSDA's, an unweighted sequential hot deck was the tool of choice for almost all the imputation revised variables, which included both

demographic and drug use measures. Hot-deck imputation involves the replacement of a missing value with a valid code taken from another respondent who is "similar" and has complete data. The data set is first partitioned into imputation classes, which are defined according to variables that are very closely related to the variable with missing values to be imputed. Responding and non-responding units are then sorted together within these imputation classes by a variable or collection of variables (less) closely related to the variable of interest Y. For sequential hot-deck imputation (Little and Rubin, 1987, p. 62), a missing value of Y is replaced by the nearest responding value preceding it in the sequence. This method is mostly non-random, in that randomness can only be induced if more than one donor exists with identical levels of all other variables in the sort, and a final random number is used to determine the donor.

Although the unweighted sequential hot deck was simple and quick to implement, it was not acceptable to use a method where sample weights were not accounted for, and was not sufficiently flexible for the imputation of the person-pair variables. Although other imputation methods were available that accounted for sample weights, no method fulfilled all the requirements that we had for the 1999 NHSDA CAI sample. In particular, we required a method that had the following attributes:

- **Methods that incorporate sampling weights are preferred over unweighted methods.** Incorporating weights leads to asymptotically unbiased and consistent estimates of population means and totals

- **Model-assisted methods are preferred over methods that don't use models.** Using a model allows us to get a more precise estimate of the predictive mean, which in turn gives more nearly unbiased estimates of the overall means and totals. In addition, a larger number of covariates can be used to determine donors, and their relative importance can be determined objectively through model fitting.

- **Methods with a live donor set are preferred over other methods**. This allows us the flexibility to subset donors according to internal consistency and other constraints.

- **Methods using distance functions to find donors "close" to recipients are preferred over methods using sort variables**. Donors can be chosen that are arbitrarily close to the recipient, and the sparseness of the donor set is less of an issue.

- **Methods that are potentially multivariate are preferred over methods that are necessarily univariate**. Multivariate imputations preserve the correlation between the imputed variables. It is natural for closely related variables to be imputed together.

- **Methods that deal with both discrete and continuous data are preferable.** If a distance function is used, we need a clear and objective way to define the distance function for both discrete and continuous variables. In a multivariate imputation, a method is needed that can deal with discrete and continuous variables simultaneously.

The New Imputation Method

To address the deficiencies associated with the unweighted sequential hot deck and to incorporate all the requirements listed above, a new method of imputation called Predictive Mean Neighborhoods (PMN) was developed for the NHSDA. PMN is a combination of modeling with a random nearest neighbor hot deck. This method can be applied to one variable at a time or to several variables at once. The technique incorporates predictive means from models, and the assigns values to item nonrespondents using neighborhoods determined by those predictive means.

The problem of univariate imputation arises when values of a single continuous variable, such as age at first use of marijuana, or a dichotomous discrete variable, such as lifetime use of marijuana, are missing for a respondent, and these need to be imputed independently of the imputation of other variables. The problem of multivariate imputation arises when values of two or more variables are missing for a single respondent. The case of a single polytomous variable such as marijuana recency with missing values can be viewed as a multivariate imputation problem.

A commonly used imputation method is random nearest neighbor hot deck (Little and Rubin, 1987, p. 65). With this method, donors and recipients are distinguished by the completeness of their records with regard to the variable(s) of interest (the donor has complete data, the recipient does not). A donor is selected at random from a set of complete records deemed close to the recipient with respect to a number of covariates. For the NHSDA, the set of covariates typically would include demographic variables as well as some other nonmissing drug use variables. To further ensure that a donor matches the recipient as closely as possible, discrete variables (or discrete categories of continuous variables) strongly correlated with drug use, such as age categories, can be used as classing variables. A difficulty arises when there is an insufficient number of donors to closely match a recipient on a number of important discrete covariates. What is defined as "close" would have to depend upon a sorting mechanism as described with the unweighted sequential hot deck, the disadvantages of which have already been made clear.

To remedy the problem of sparseness of the imputation class, the idea of predictive mean modeling is used. Using respondents with complete data, the mean of the outcome variable is modeled as a function of the covariates; thus the mean gives a one number summary of the effects of covariates (or predictors) on the outcome variable. In other words, instead of matching the values of the covariates through a sorting or classing mechanism, the covariates' effect is transformed into a single (continuous) number given by the predictive mean which, in turn, can be easily used to define the donor neighborhood. The predictive mean can be then used to define the imputation class (the "neighborhood") in the nearest neighbor hot deck. Assuming that the predicted mean for a randomly selected donor from this neighborhood (an item respondent) and the recipient (an item non-respondent) are approximately equal, the residual defined by the difference between this predicted mean and the observed values of the donor should approximate the residual that would have been obtained if it had been drawn from a known error distribution. This technique is called the univariate predicted mean neighborhood (UPMN) technique, since it is defined from a single predictive mean.

When the outcome variables are multivariate, the predictive mean vector is modeled, and as in the UPMN, donors in the neighborhood of the mean vector are used to choose a record for hot deck imputation. For a single categorical but polytomous variable, one can use a polytomous logit model to get the predictive mean vector (actually a vector of probabilities). An appealing way to determine this predictive mean vector would be to model all the response variables at once, including both discrete and continuous variables. This multivariate superpopulation model can be used to characterize the finite population from which an error distribution could be estimated. The model parameters could then be estimated from the complete sample data after taking account of the response mechanism and the sample design.

Although the above idea of multivariate modeling with an arbitrary set of outcome variables (including both discrete and continuous) is preferred, it is likely to be tedious in practice because of computational problems due to the sheer number of model parameters, and the difficulty in specifying a suitable covariance structure. Little and Rubin (1987) have proposed a joint model for discrete and continuous variables. While this has been implemented by Schafer (1998), his solution doesn't take account of survey design effects. As an alternative, a series of univariate (including the polytomous case) models are fit sequentially such that variables modeled earlier on in the hierarchy have a chance to be included in the covariate set for subsequent models in the hierarchy. This idea gives rise to the method of multivariate predictive mean neighborhood (MPMN) imputation. The vector of predictive means so obtained can then be used to compute a distance (Mahalanobis) to define a neighborhood for each record with missing values for multivariate imputation.

In the application of the PMN to the 1999 NHSDA, the multivariate error distribution could not be estimated since a multivariate model was not fit, and therefore a residual could not be randomly selected to add to the predictive mean. This process can be mimicked, however. The response variables are arranged in a hierarchy, and the univariate conditional means given the variables earlier on in the hiearchy are estimated using parametric models. The conditional residual distributions are estimated nonparametrically based on the neighborhood of donors with approximately equal conditional means. Drawing a record at random gives rise to (approximately) the conditional mean plus random residual. Note that in this process, the conditional residual distribution is specific to the incomplete record. Moreover, the parametric model for the predictive mean is used to assist in finding the neighborhood as well as in finding the variance estimate adjusted for imputation.

If it turns out that the donor set for MPMN is sparse, then the UPMN can be used as an alternative. Assuming that the donor set (i.e., the set of complete records which are in a small neighborhood of the

recipient with respect to all the elements of the predictive mean) is not sparse, then having a single record to fill in all the missing values in an incomplete record is desirable as it preserves the relationships among the variables of interest. Moreover, if the predictive mean vector includes both missing and nonmissing variables (this could easily happen when models are fitted in a univariate manner under a hierarchy), then one can also ensure that the predictive mean vector for the donor record is not only close to recipient with respect to missing variables, but also with respect to the nonmissing ones. Although the nonmissing values would not be replaced by the corresponding values of the donor, some degree of correlation between missing and nonmissing variables is expected to be preserved because of the closeness between the donor and the recipient. The reason for this is that the predictive mean vector consists of conditional means (the drug use covariates in the conditioning set appear earlier on in the hierarchy), and therefore being close to the conditional means should help in preserving the correlation between outcome variables of the recipient record.

Application of PMN to Person-Pair Data

The procedure for implementing UPMN and MPMN can be summarized with the following six steps. Steps 2 through 5, and sometimes 6, are cycled through each of the variables in the order determined by Step 1. Steps 4-5 (Steps 4-6 when applicable) could be considered a variant of a random nearest neighbor Hot Deck.

Step 1: Hierarchy definition. The first step is to determine the order in which variables are modeled, so that variables early in the hierarchy may be used for modeling the conditional predictive mean, i.e., they have the potential to be part of the set of covariates for variables later in the hierarchy. Note that not all variables in the hierarchy may be missing for a particular incomplete record. Nevertheless, models are developed for all the variables in a univariate fashion for reasons mentioned earlier. For the person-pair data in the NHSDA, the pair relationships, multiplicity counts, and household-level pair domain counts are all closely related, and could be imputed in one multivariate set when a pair was selected with the pair relationships imputed first, followed by the multiplicity counts, and the household counts. For households where pairs were not selected, a univariate imputation of household-level pair domains could be implemented. Using the sequence of variables determined by this step, cycle through steps 2 through 5, and sometimes 6.

For each variable:

Step 2: Setup for model building and hot deck assignment. For each model that is fitted, two groups must be created, complete and incomplete data respondents (item respondents and item non-respondents). Complete data respondents have complete data across the variables of interest, and incomplete data respondents encompass the remainder of respondents. If the final assignment will be multivariate, then complete data respondents must have complete data across all the variables in the multivariate response vector. Models will be constructed using complete data respondents only.

Step 3: Sequential hierarchical modeling. Build the model using the complete data respondents only, with weights adjusted for item non-response. The sequence depends upon the hierarchy determined in Step 1.

Step 4: Computation of predictive means and delta neighborhoods. Once the model has been fitted, the predictive means for item respondents and item nonrespondents are calculated using the model coefficients. For models with a multivariate predictive mean vector (such as with a polytomous logit model), a single element out of that vector must be chosen, so that each respondent has exactly one predictive mean value.[1] This predictive mean is the matching variable in a random nearest neighbor hot deck. It can either come directly from the model, it can be adjusted to account for the conditioning on the time period, or if it is the predicted value based on a model with a transformed response variable, it can be back-transformed to the original units.

For each item non-respondent, a distance is calculated between the predictive mean of the item non-respondent and the predictive means of every item respondent. Those item respondents whose predictive means are "close" (within a predetermined value delta) of the item non-respondent are considered part of the "delta neighborhood" for the item non-respondent, and are potential donors. If the number of item respondents who qualify as donors is greater than some number, say *k*, then only those item respondents with the smallest *k* distances are eligible to be donors.

The pool of donors is further restricted to

[1]Alternatively, one could perform a provisional MPMN just using the predicted probabilities from the polytomous model. The final MPMN would be built based on probabilities from the polytomous model, as well as predictive means for the other variables in the multivariate set. See Step 6 for a description of the MPMN.

satisfy constraints to make imputed values consistent with the pre-existing non-missing values of the item non-respondent. It is not possible, for example, for a donor household to have a larger multiplicity count than a recipient's household size. Other constraints, called likeness constraints, are placed upon the pool of donors to make the attributes of the neighborhood as close to that of the recipient as possible. For example, when imputing pair relationships or pair counts, it would be advantageous for donor pairs and recipient pairs to have ages as similar as possible. A small value of delta could also be thought of as a likeness constraint. Whenever insufficient donors are available to meet the likeness constraints, including the preset small value of delta, the constraints are loosened in priority order according to their perceived importance. As a last resort, if an insufficient number of donors are available to meet the logical constraints given the loosest set of likeness constraints allowable, then a donor is found using a sequential hot deck, where matching is done on the predictive mean. (Even though weights would not be used to determine the donor in the sequential hot deck, "unweighted" is not an accurate characterization of the imputation process, since weighting has already been incorporated in the calculation of the predicted mean.)

Step 5: Assignment of imputed values using a univariate predictive mean (UPMN). Using a simple random draw from the neighborhood developed in Step 4, a donor is chosen for each item non-respondent. If only one response variable is to be imputed, then the assignment step is just a simple replacement of a missing value by the value of the donor.

The assignment step is multivariate if several response variables are associated with a single predictive mean, provided more than one of those response variables is missing. In that case, all of the missing values will be imputed using the same donor. If there is more than one response variable associated with a single predictive mean, but not all of them are missing, only the missing values are replaced by those of the donor. The resulting imputed values are provisional if a multivariate neighborhood (MPMN) step is called for; otherwise, these values are final.[2]

Go to step 6 if the variables for which steps 2-5 have been completed are part of a complete multivariate set, for which MPMN is to be applied. Go to step 2 if the variables for which steps 2-5 are completed are not part of a complete multivariate set, and other variables are still to be imputed. Otherwise, the process is finished.

Step 6: Determination of multivariate predictive mean neighborhood and assignment of imputed values (MPMN). With the multivariate predictive mean neighborhood, the neighborhood is defined based on a vector of predictive means, rather than from a single predictive mean as in the univariate case. This vector may encompass a sub-vector of predictive means from a single categorical model (as with a polytomous logit model), in addition to scalar predictive means from any number of models with continuous response variables. For each item non-respondent, a distance is calculated between the elements of this vector of predictive means where the observed values are missing, and the corresponding elements of the vector for every item respondent.

The subset of elements that are used to determine a neighborhood for a particular item non-respondent depends on the missingness pattern of that item non-respondent.[3]

A neighborhood that results from this vector of distances can be constrained by a multivariate pre-set delta, where the distances associated with each element of the predictive mean vector must each be less than pre-set delta associated with that element. From the donors that remain, a single neighborhood can be created out of a vector of differences by converting that vector to a scalar, called the Mahalanobis distance, which is given by

$$(\mu_R - \mu_{NR})^T \Sigma^{-1} (\mu_R - \mu_{NR})$$

where μ_R refers to the predictive mean (sub-)vector for a given item respondent, and μ_{NR} is the predictive mean (sub-)vector for a given item non-respondent. The matrix Σ is the variance-covariance matrix of the predictive means, calculated using the sub-vector of predictive means associated with each missingness pattern, using complete data respondents within each imputation class. The Mahalanobis distance is only calculated for those respondents meet the delta constraint. The neighborhood is determined by selecting the *k* smallest Mahalanobis distances within this subset of item respondents for a given item non-respondent.

[2] If the variable is part of a multivariate set upon which MPMN is to be applied, and provisional values are not needed for subsequent models, Steps 4b (creation of delta neighborhood) and 5 could be skipped.

[3] Alternatively, one could use the entire predictive mean vector to determine the neighborhood, regardless of the missingness pattern.

If some of the variables in the response vector are not missing, then only those that are missing are replaced. However, logical constraints must be placed on the multivariate neighborhood, so that imputed values are consistent with pre-existing non-missing values. In addition to the multivariate delta, likeness constraints are used to make the donors in the neighborhood as much like the recipient as possible. These can be loosened if insufficient donors are available. As with the univariate assignments, a donor is randomly drawn from the neighborhood for each item non-respondent.

Comparison of PMN with Other Available Methods

The PMN method has some similarity with the predictive mean matching method of Rubin (1986) except that for the donor records the observed variable value and not the predictive mean is used for computing the distance function. Thus Rubin's method is not entirely suitable for discrete variables because the predictive mean is on a continuous scale. However, with predictive mean vector defined for both discrete and continuous variables, the PMN method is not restricted to continuous variables.

The well known method of nearest neighbor imputation is also similar to PMN, except that the distance function is in terms of the original predictor variables and would often require scaling for discrete variables. Also for this method it is generally hard to decide about the relative weights for different predictor variables.

In comparison to other model-based methods, discrete and continuous variables could be handled jointly and relatively easily in MPMN by using the idea of univariate (conditional) modeling in a hierarchical manner. In MPMN, one could objectively assign differential weights to different elements of the predictive mean vector depending on the variability of predictive means in the data set via the Mahalanobis distance. For a given predictive mean, differential weights were given to different predictor variables according to their association with the variable targeted for imputation (the response variable). Thus the response variable was also used (indirectly) in defining the distance function and the neighborhood. This feature was useful in finding similar records for the neighborhood.

Concluding Remarks and Further Work

The PMN methodology has been widely used for the imputation of a variety of variables in the NHSDA, including both continuous and categorical variables with one or more levels. The models were fit using standard modeling procedures in SAS and SUDAAN, while SAS macros were used to implement the hot deck step, including the restrictions on the neighborhoods. Although creating a different neighborhood for each item non-respondent was computationally intensive, the method was implemented successfully.

The imputations team at RTI is currently implementing a series of simulations to evaluate the new method, comparing it against the unweighted sequential hot deck used earlier and a simpler model-based method.

References

Chromy, JR and Singh, AC (2001) *ASA Proc. Surv. Res. Meth. Sec.*

Little RJA and Rubin DB (1987), *Statistical Analysis with Missing Data*, John Wiley and Sons, New York: 1987

Rubin DB (1986), "Statistical Matching Using File Concatenation with adjusted Weights and Multiple Imputations," *J. Business Econ. Statist* **4**, 87-94

Schafer, J. (1997). Analysis of incomplete multivariate data, Chapman and Hall, London.