A SIMPLE EVALUATION OF THE IMPUTATION PROCEDURES USED IN NSDUH

E.A. Grau[1], P. Frechtel[1], D.M. Odom[2], and D. Painter[3]
[1]Statistics Research Division, RTI International, RTP, NC 27709-2194
[2]Inveresk, Cary, NC 27513
[3]Substance Abuse and Mental Health Services Administration, Gaithersburg, MD 20877

Keywords: Predictive Mean Neighborhoods; Sequential Hot Deck; Predictive Mean Matching; Nearest Neighbor Imputation; Imputation Evaluation

Abstract

The National Survey on Drug Use and Health (NSDUH) is the primary source of information on substance use in the U.S. Since 1999, the Predictive Mean Neighborhoods (PMN) procedure has been used to impute missing values for many of the analytical variables. This method is a combination of two commonly used imputation methods: a nearest-neighbor hot deck and a modification of Rubin's predictive mean matching method. Although PMN has many practical advantages, it has not been formally evaluated. In this paper we discuss a simple simulation to evaluate PMN. Using only complete data cases, we induced random patterns of missingness in the data for selected outcome variables. Imputations were conducted using a simple version of PMN, a random allocation to use/nonuse based on the predictive mean from a model, and weighted and unweighted sequential hot deck. This process of inducing missingness and imputing missing values was repeated multiple times. The imputed values using PMN, model-only, and the hot deck methods were then compared with the true values that were found in the complete data, across the repeated iterations. In particular, we compared the number of matches between the two methods, as well as statistics derived from the data, such as drug prevalence estimates.

Introduction

In the National Survey on Drug Use and Health (NSDUH), imputation is used as a way of handling missing data in a number of key analytic variables. In survey years from 1999 onwards, a new method of imputation was introduced, called Predictive Mean Neighborhoods (PMN). This method, described in detail in Singh, Grau, and Folsom (2002), is a semiparametric approach to imputation, which combines the model-based attributes of Rubin's Predictive Mean Matching method (PMM) (Rubin, 1986) with the nonparametric nearest neighbor hot deck (NNHD). PMN enhances the PMM method in that it can be applied to both discrete and continuous variables either individually or jointly. PMN also enhances the NNHD method in that the distance function used to find neighbors is no longer ad hoc.

In the univariate version of PMN (denoted as UPMN), as in the case of predictive mean matching, the prediction model is fit to the data from complete respondents, but the predictive means for both recipients and donors are computed to obtain the distance between the two predictive means. In the multivariate version of PMN (denoted as MPMN), a set of predictive means can be obtained either from a single multivariate model, or from a series of univariate models. Fitting a multivariate model requires the specification of a covariance structure, which may not be straightforward, while fitting univariate models may preclude the ability to incorporate the correlations between the outcome variables in the models. A middle ground is to fit a sequence of univariate models, where models later in the sequence are conditioned on outcomes from models earlier in the sequence. (That strategy was pursued in the NSDUH, but is not pursued in this paper. Rather, a series of univariate models are fitted that do not incorporate the full covariance structure of the explanatory and response variables, except in the final assignment of imputed values.) Regardless of the method used to obtain the vector of predictive means, a neighborhood for picking a donor is defined using a vector delta or the Mahalanobis distance or both.

In some cases, covariates cannot be included in the models used to calculate the predictive means. This may be due to the fact that the covariates have missing values that cannot be easily imputed, or because the covariate is related to the response in such a way that convergence to a stable model is difficult. In some cases, these covariates may be strongly related to the response, so that excluding them from the model will result in a loss of potentially useful information. For both UPMN and MPMN, an additional feature is that these problematic covariates can still be included in the imputation process, by limiting donors and recipients in the hot deck step to have matching values for the covariate, if the covariate's values are nonmissing for both donor and recipient.

Although the advantages of PMN are apparent from the above description, it is necessary to evaluate whether the advantages provide a real "improvement" in the imputed values. "Improvement" can be measured in several ways. One way of evaluating PMN as a methodology would be to simulate data from a hypothetical population, and then compare estimates and their standard errors using a variety of imputation methodologies. Research is currently underway in this area, though this is not the method of evaluation used here. In the method described in this paper, no population is simulated from a set of modeled values. Rather, we are attempting to evaluate how well the imputed values match the sample, which in turn represents the actual finite civilian, noninstitutionalized population of the United States aged 12 years old or older. This will involve comparing imputed values with the actual responses of respondents.

Because of time constraints, the version of PMN evaluated here is a much simpler version of PMN than that used in the NSDUH. Results obtained here cannot be directly translated to the NSDUH imputation procedures. Because of these simplifications, the version of PMN evaluated in this paper did not account for one of the major advantages that PMN provides. In particular, PMN allows information from a large number of variables to be included in the imputation models. Some of these variables, in the version used in the NSDUH, could have had missing values that were replaced with provisionally imputed values. Variables within a multivariate set were modeled in sequence, so that models for variables that were farther along the sequence could potentially have a large number of covariates (corresponding to variables earlier in the sequence) with provisionally imputed values. Final imputed values for these variables were not obtained until the final multivariate imputation. This greatly increased the predictive power of the models, especially for variables that were later in the sequence. PMN had a clear advantage over the simple hot deck in this regard, since sparse donor classes prevented the use of more than a small number of covariates to define classing and sorting variables. For example, a major advantage of PMN as applied in the NSDUH was the ability to use drug use variables in the imputation of other drug use variables. For a drug such as cocaine, use of a large variety of other drug use information provided quite a bit of useful information in the imputation of missing values for cocaine. However, using provisionally imputed values gave rise to theoretical complications; we desired to avoid these complications in this study. Hence, all of the covariates used in both the PMN and hot deck methods were considered known (demographic variables, for the most part). Even with this limitation, it would have been possible to use drug use variables in the hot deck, or in the hot deck step of PMN. We did not do so, since the added advantage of PMN over the hot deck methods did not seem apparent with such a small group of drugs.

Methodology

Although PMN was applied across a wide variety of variables in the NSDUH, we focused our efforts on the imputation of lifetime usage of selected drugs. To be included in the study, respondents had to answer all three of the questions concerning lifetime usage of cigarettes, alcohol, and marijuana. Because of the strong relationship between age and drug use, imputations in the NSDUH are generally conducted separately within three age groups: 12 to 17 years old, 18 to 25 years old, and 26 years old or older. For this study, interest was focused on one of these age groups, 18 to 25 year olds. Missingness was induced in the data by randomly selecting 200 5% subsamples among respondents 18 to 25 years old. Seven possible missingness patterns were possible with these data:

1. Missing only cigarettes
2. Missing only alcohol
3. Missing only marijuana
4. Missing cigarettes and alcohol
5. Missing cigarettes and marijuana
6. Missing alcohol and marijuana
7. Missing cigarettes, alcohol, and marijuana

Due to computational restraints, this study focused only on patterns #3 (missing only marijuana), #6 (missing alcohol and marijuana), and #7 (missing all three drugs). Only the results from missingness pattern #7 are presented here; the conclusions obtained using the other missingness patterns did not differ substantially from those obtained using missingness pattern #7.

Two missingness mechanisms were considered in this study. In one the missing data were considered missing completely at random (MCAR), and in the other, the missingness mechanism depended upon marijuana use, and therefore would probably be considered not missing at random (NMAR). In the case where other information was not used to induce missingness, the missing data mechanism was ignorable, and could be considered MCAR.

Most imputation methods assume that the missing data are missing at random (MAR), which, in the words of Shafer (1997), allows the probability that a datum is missing to depend upon the datum itself, but only indirectly through quantities that are observed (e.g., the explanatory variables). The performance of the imputation methods was not evaluated with MAR data directly. However, a second missing data mechanism was also induced, that may have been MAR, but was more likely NMAR. In particular, 7% of marijuana users had their values for cigarettes, alcohol, and marijuana set to missing, and only 2½% of nonusers had the values for these variables set to missing, giving an overall missingness percentage of 5%. If covariates that were selected had been strongly related to the marijuana usage response, then the missing data mechanism could also be related to these auxiliary variables, making the missing data mechanism MAR. The more likely scenario, however, was that the selected covariates were not sufficiently related to the response to give us any information about the missing data mechanism, making it NMAR.

For each missing data mechanism, missing values were induced, and then imputed 200 times, using four frequentist approaches to imputation: weighted sequential hot deck (WSHD), unweighted sequential hot deck (USHD), model-based random allocation, and PMN. The hot deck procedures are described in detail in the appendix to this manuscript. For both of the hot deck methods, the data sets were sorted using a serpentine sort. After all the imputations had been completed, the following tests were performed:

a) Compare the number of cases where the imputed value is same as the original response for the three methods. Conduct statistical tests to compare the number of matches between the methods.
b) Compare the mean proportion of lifetime users of the imputed values to the mean proportion of lifetime users for the actual responses among the 5% subsamples, across the four methods.

Specifics on Imputation Methods

The PMN, model-based random allocation, and WSHD methods are design-based methods, so that the original survey sampling weights needed to be used. Since this study was limited to respondents who answered each of the three lifetime usage questions (cigarettes, alcohol, and marijuana), the initial set of sampling weights were adjusted to reallocate the weights from the "real" item nonrespondents to other respondents. The weights were calibrated using an item response propensity model, which measures the probability of responding to these three questions. The variables in the response propensity model were age, gender, race, marital status (married or not married), employment status (employed full time or not employed full time), census region, and population category of the segment in which the respondent lived.[1]

Once the initial sampling weights were adjusted to account for the "real" item nonresponse, missingness was induced for each of the 5% subsamples, 200 times. Each of the imputation methods was then applied to each of the missingness patterns. Details for each method are given below, describing a process that was implemented 400 times (2 missing data mechanisms, 200 iterations):

PMN and Model-based Random Allocation

The weights for the artificially induced nonrespondents were reallocated to the 95% of respondents that were not part of the subsample. Weights were calibrated using an item response propensity model with the same set of covariates as given in the item response propensity model previously described. Predicted probabilities were obtained using logistic regression models, with the weights appropriately adjusted. Covariates for these models included age and age squared (both centered to avoid multicollinearity problems), gender, marital status[2], employment status[3], education level[4], state rank[5], and census region. For the model-based random allocation, a

[1] A segment is a group of census blocks that are the primary sampling unit within the multistage cluster sample used in the NSDUH. The variable discussed here categorizes segments into 3 levels: segment in MSA of over 1 million persons, segment in MSA of less than 1 million persons, or segment not in MSA

[2] Married, widowed, divorced/separated; never married

[3] Employed full-time, employed part-time, unemployed, other

[4] Less than high school, high school graduate, some college, college graduate

[5] The states were ranked by unweighted proportion of lifetime users in the sample. The top 17 states were assigned rank 1, the next 17 rank 2, and the last 17 rank 3.

uniform(0,1) random variable was generated for each respondent, and the imputed value was assigned by comparing the random value to the predicted probability of the lifetime usage variables that were missing.

The hot-deck step of PMN was a simple random imputation based on the same predicted probability that was used in the model-based random allocation. If only one lifetime usage indicator was missing, the donor was selected from a neighborhood comprised of the item respondents with the 30 closest predictive means to the recipient's predictive mean. In cases where more than one lifetime usage indicator was missing, the donor in the hot deck was selected from a neighborhood comprised of the item respondents with the 30 closest Mahalanobis distances.

Weighted and Unweighted Sequential Hot Deck

In the hot deck methods used for missingness pattern #7, education level and employment status (defined as in the predictive mean model) were used to define imputation classes, and race and age were sorting variables. The imputation classes were defined by choosing the variables in the predictive mean models of PMN with the highest level of significance. In the sequential hot deck, the order of the sorting variables was crucial, with variables lower down in the sorting order having minimal impact in cases where the number of respondents with such attributes is sparse within the given imputation class. The order of the sorting variables was determined by looking at the levels of significance of the variables in the predictive mean models of PMN.

Specifics on Evaluation Criteria

The methods were compared on two levels. On one level, we compared the individual imputed values with the values that were actually observed in the original sample. On another level, we compared the estimates of prevalence of drug use among respondents in the subsample.

Comparison of individual imputed values (the "match test")

For each iteration, counts were obtained of the number of exact matches for each drug among imputed cases for all four methods. The numbers of matches for each method were analyzed as a multivariate analysis of variance, with the drugs defining the multiple variables and the methods acting as "treatments." In those instances where significant differences existed, the methods were evaluated using the Tukey multiple comparison procedure. The measurement variable used was the proportion of matches between the imputed value and the value that had been set to missing. Although these proportions were not normally distributed, the deviation from normality was not considered serious enough to affect the conclusions obtained by using the normality assumption. In part, this was due to the fact that the mean proportion of matches, especially in the case of cigarettes and marijuana, was not close to 0 or 1.

Comparison of prevalence estimates (the "mean test")

The prevalence for each substance (i.e., proportion of substance users) was determined for each iteration. We evaluated the imputation methods by comparing the prevalence estimates obtained by looking at imputed cases with the prevalence estimates from the actual responses within the subsample. The prevalence estimates from the actual responses were considered the "control," and estimates obtained from the imputed cases were compared to the control using Dunnett's test. A significant result indicated that the estimate obtained by the method in question deviated significantly from the estimate that would have been obtained from the actual responses.

Results

This evaluation showed that using PMN provided a modest advantage in the match test, and, in the MCAR case, using WSHD provided a modest advantage in the mean test. In each of these cases, however, the degree of difference between the methods, though significant, was not substantial. Not surprisingly, all methods did badly in the mean test when the data were NMAR.

We had expected that the model-based methods, especially PMN, would provide significantly better imputations; reasons for this lack of a strong indicator of differences between the methods are provided in the Discussion section. The strength of the WSHD method in the mean test was particularly puzzling. In general, however, it is clear that this study was not broad enough to capture the advantage that would be provided by the model-based methods.

Tables 1 and 2 show the modest advantage provided by PMN over the other methods with the match test. The table presents the proportion of imputed values that matched the original reported

value. In all cases, significant differences between the methods were found using a multivariate analysis of variance; individual differences between the methods were measured using Tukey's multiple comparison test. Methods that were not significantly different from each other share the same letter in the "Tukey's test" column. It is clear that, in order to reproduce the actual responses that would have occurred, some hot deck method (including PMN) is necessary.

Table 1. Results of Comparison of Individual Imputed Values (the "match test"), MCAR.

Substance	Method	Mean proportion of matches	Tukey's test
Cigarettes	PMN	0.624	A
	USHD	0.623	A
	WSHD	0.614	B
	Model[6]	0.551	C
Alcohol	PMN	0.801	A
	USHD	0.797	B
	WSHD	0.789	C
	Model	0.684	D
Marijuana	PMN	0.532	A
	WSHD	0.520	B
	USHD	0.520	B
	Model	0.491	C

Table 2. Results of Comparison of Individual Imputed Values (the "match test"), NMAR.

Substance	Method	Mean proportion of matches	Tukey's test
Cigarettes	USHD	0.667	A
	PMN	0.667	A
	Model	0.665	A
	WSHD	0.653	B
Alcohol	PMN	0.840	A
	USHD	0.836	B
	Model	0.834	B
	WSHD	0.824	C
Marijuana	PMN	0.549	A
	Model	0.543	B
	USHD	0.539	C
	WSHD	0.531	D

The mean test does not attempt to reproduce individual values; rather, the overall estimate is what must be reproduced. Table 3 shows a slight advantage of WSHD over the other methods when using the mean test in the MCAR case, and Table 4 shows that all methods did not compare well with the control mean in the NMAR case. The fact that the subsample contained a disproportionate number of marijuana users meant that the remaining cases were relatively bereft of marijuana users. Since users of marijuana were, by and large, also users of alcohol and cigarettes, the remaining cases were also, to a lesser degree, relatively bereft of alcohol and cigarette users. Clearly, the hot deck methods could not account for this. The models were also not sufficiently sensitive to detect the fact that the missingness was related to marijuana use. Methods that provided estimates that deviated significantly from those of the actual responses are starred in the "Dunnett's test" column.

Table 3. Results of Comparison of Prevalence Estimates (the "mean test"), MCAR.

Substance	Method	Subsample Prevalence Estimate	Dunnett's Test
Cigarettes	Control	0.711	
	PMN	0.717	**
	Model	0.721	**
	USHD	0.723	**
	WSHD	0.710	
Alcohol	Control	0.867	
	PMN	0.873	**
	Model	0.875	**
	USHD	0.878	**
	WSHD	0.866	
Marijuana	Control	0.537	
	PMN	0.547	**
	Model	0.553	**
	USHD	0.549	**
	WSHD	0.535	

[6] The entry "Model" refers to Model-based random allocation

Table 4. Results of Comparison of Prevalence Estimates (the "mean test"), NMAR.

Substance	Method	Subsample Prevalence Estimate	Dunnett's Test
Cigarettes	Control	0.816	
	PMN	0.721	**
	Model	0.717	**
	USHD	0.726	**
	WSHD	0.707	**
Alcohol	Control	0.926	
	PMN	0.875	**
	Model	0.869	**
	USHD	0.880	**
	WSHD	0.865	**
Marijuana	Control	0.764	
	PMN	0.546	**
	Model	0.538	**
	USHD	0.550	**
	WSHD	0.526	**

Discussion and Further Research

As stated in the introduction, the PMN imputation method evaluated here did not correspond directly to the imputation method used in the NSDUH. We had expected PMN to perform better than the other methods; however, it is clear that this evaluation did not sufficiently highlight the advantages that PMN provides. In summary, there are two major advantages to PMN that were not incorporated here:

1. PMN allows the use of models, which provide a way of including information from a relatively large set of variables, appropriately weighted according to their relative importance. Because provisionally imputed values were not allowed in the determination of predictive means in this study, other variables in the multivariate set were not included as covariates in the models. In addition, only three drugs were considered, which precluded the possibility of including a large number of variables within the same multivariate set.

2. In the hot deck step of PMN, covariates that could not be included in the models could still be used in the determination of the imputed value, by constraining donors according to the values of these covariates. This step was also not taken advantage of in this study.

In general, the model-based methods have an advantage over the hot-deck methods, in that the hot-deck methods cannot incorporate a large number of variables without having sparse donor cells. This study did not pick that up. Clearly, further study is needed. In order to fully establish whether PMN is superior or not to these other methods, a study must be set up that allows for the incorporation of these extra advantages.

References

Chromy, J. R. (1979). Sequential sample selection methods. In *Proceedings of the American Statistical Association, Survey Research Methods Section* (pp. 401-406). Alexandria, VA: American Statistical Association.

Cox, B. G. (1980). The weighted sequential hot deck imputation procedure. In *Proceedings of the American Statistical Association, Survey Research Methods Section* (pp. 721-726). Washington, DC: American Statistical Association.

Iannacchione, V. (1982). Weighted sequential hot deck imputation macros. In *Proceedings of the Seventh Annual SAS Users Group International Conference*. Cary, NC: SAS Corporation.

Little, R. J. A., & Rubin, D. B. (1987). *Statistical analysis with missing data*. New York: John Wiley & Sons.

Rubin, D. B. (1986). Statistical matching using file concatenation with adjusted weights and multiple imputations. *Journal of Business and Economic Statistics, 4* (1), 87-94.

Schafer, J. L. (1997). *Analysis of incomplete multivariate data.* Boca Raton: Chapman and Hall.

Singh, A.C., Grau, E.A., and Folsom, R.E., Jr. (2002) Predictive Mean Neighborhood Imputation with application to the person-pair data of the National Household Survey on Drug Abuse. *Proceedings of the American Statistical Association, Survey Research Methods Section for 2001*.

Williams, R. L., & Chromy, J. R. (1980). SAS sample selection MACROS. In *Proceedings of the Fifth International SAS Users Group International Conference* (pp. 382-396). Cary, NC: SAS Corporation.

Appendix: The Hot Deck Method of Imputation

A.1 Introduction

Typically, with the hot-deck method of imputation, missing responses for a particular variable (called the "base variable" in this appendix) are replaced by values from similar respondents with respect to a number of covariates (called "auxiliary variables" in this appendix). If "similarity" is defined in terms of a single predicted value from a model, these covariates can be represented by that value. The respondent with the missing value for the base variable is called the "recipient," and the respondent from whom values are borrowed to replace the missing value is called the "donor."

A step that is common to all hot-deck methods is the formation of imputation classes, which is discussed in Section A.2. This is followed by a discussion of the sorting methods used for sequential hot decks in Section A.3. Specific details of the unweighted and weighted methods are given in Sections A.4 and A.5 respectively. With each type of hot-deck imputation, the identities of the donors are generally tracked. For more information on the general hot-deck method of item imputation, see Little and Rubin (1987, pp. 62-67).

A.2 Formation of Imputation Classes

When there is a strong logical association between the base variable and certain auxiliary variables, the dataset is partitioned by the auxiliary variables and imputation procedures are implemented independently within classes defined by the cross of the auxiliary variables.

A.3 Sorting the File

Within each imputation class, the file is sorted by auxiliary variables relevant to the item being imputed. The sort order of the auxiliary variables is chosen to reflect the degree of importance of the auxiliary variables in their relation to the base variable being imputed (i.e., those auxiliary variables that are better predictors for the item being imputed are used as the first sorting variables). In general, two types of sorting procedures have been used in NSDUH imputation procedures:

- **Straight Sort**. A set of variables is sorted in ascending order by the first variable specified; then within each level of the first variable, the file is sorted in ascending order by the second variable specified; and so forth.

- **Serpentine Sort**. A set of variables is sorted so that the direction of the sort (ascending or descending) changed each time the value of a variable changed.

The serpentine sort has the advantage of minimizing the change in the entire set of auxiliary variables every time any one of the variables changes its value.

A.4 Unweighted Sequential Hot Deck

The unweighted sequential hot deck method involves three basic steps. After the imputation classes are formed, the file is appropriately sorted and read sequentially. Each time an item respondent is encountered (i.e., the base variable is nonmissing), the base variable response is stored, updating the donor response. Any subsequent nonrespondent that is encountered receives the stored donor response, creating the statistically imputed response. A starting value is needed if an item nonrespondent is the first record in a sorted file. Typically, the response from the first respondent on the sorted file is used as the starting value. Due to the fact that the file is sorted by relevant auxiliary variables, the preceding item respondent (donor) closely matches the neighboring item nonrespondent (recipient) with respect to the auxiliary variables.

A.4 Weighted Sequential Hot Deck

The steps taken to impute missing values in the weighted sequential hot deck are equivalent to those of the unweighted sequential hot deck. The details on the final imputation, however, differ with the incorporation of sampling weights. The first step, as always, is the formation of imputation classes, followed by the appropriate sorting of the variables. The assignment of imputed values is necessarily more complex than in the unweighted case.

The procedure described below follows directly from Cox (1980). Specifically, once the imputation classes are formed, the data is divided into two data sets: one for respondents and one for nonrespondents. Scaled weights $v(j)$ are then

derived for all nonrespondents using the following formula:

$$v(j) = w(j)s(+)/w(+);\ j = 1, 2, \ldots n,$$

where n is the number of nonrespondents, $w(j)$ is the sample weight for the j^{th} nonrespondent, $w(+)$ is the sum of the sample weights for all the nonrespondents, and $s(+)$ is the sum of the sample weights for all the respondents. The respondent data file is partitioned into zones of width $v(j)$, where the imputed value for the j^{th} nonrespondent is selected from a respondent in the corresponding zone of the respondent data file.

This selection algorithm is an adaptation of Chromy's (1979) sequential sample selection method, which could be implemented using the Chromy-Williams sample selection software (Williams & Chromy, 1980). Furthermore, Iannacchione (1982) revised the Chromy-Williams sample selection software, so that each step of the weighted sequential hot deck is executed in one macro run.

Benefits of Weighted Sequential Hot-Deck

With the unweighted sequential hot-deck imputation procedure, for any particular item being imputed, there is the risk of several nonrespondents appearing next to one another in the sorted file. An imputed value could still be found for those cases, since the algorithm would select the previous respondent in the file; however, some modifications are required in the sorting procedure to prevent a single respondent from being the donor for several nonrespondents. With the weighted sequential hot-deck method, on the other hand, this problem does not occur because the weighted hot deck controls the number of times a donor can be selected. In addition, the weighted hot deck allows each respondent the chance to be a donor since a respondent is randomly selected within each zone of width $v(j)$.

The most important benefit of the weighted sequential hot-deck method, however, is the elimination of bias in the estimates of means and totals. This type of bias is particularly present when either the response rate is low or the covariates explain only a small amount of variation in the specified variable, or both. In addition, many surveys sample subpopulations at different rates, and using the sample weights allows, in expectation, the imputed data for the nonrespondents to have the same mean (for the specified variables) as the respondents. In other words, the weighted hot deck preserves the respondents' weighted distribution in the imputed data (Cox, 1980).