SERIAL IMPUTATION OF NHANES III WITH MIXED REGRESSION AND HOT-DECK TECHNIQUES

Trena M. Ezzati-Rice, Mansour Fahimi, David Judkins, and Meena Khare
Mansour Fahimi, Westat, Inc., 1650 Research Blvd., Rockville, MD 20850

Keywords: Empirical Residuals, Statistical Matching, Hot-Deck

1. Introduction

Missing data (item or unit) are among the most unavoidable problems in large-scale surveys such as the National Health and Nutrition Examination Survey (NHANES) III. Traditionally, unit nonresponse has been compensated for via weighting, i.e., some form of nonresponse adjustment procedure. For item nonresponse, however, there is a potpourri of available methods.

Despite the reality of missing values problem, however, a portion of proposed methods fall short with respect to practicality and implementation. This work reviews some regression-based techniques that were used to impute the missing values for six NHANES III examination (MEC) variables: measured height and weight (in log), systolic and diastolic blood pressures, and total serum and HDL cholesterol. For the purpose of this presentation, these variables will be referred to as LGHT, LGWT, D_SYS, D_DIAS, TCRESULT, and HDRESULT, respectively.

There are a number of stages in the NHANES III data collection that can result in nonresponse. The screening interview can be refused (although portions of the required information are usually then obtained from neighbors). The extended interview on self-perceived health and nutrition habits can be refused in total or in part. The examination by physician can be refused in whole or in part. Data can be lost for reasons other than refusal. In this paper, we focus on the problem of missing data from the physical exam where at least part of the exam has taken place.

The imputation for the above six variables was carried out in a sequential manner, using each of two techniques to fill in the missing data while attempting to preserve vital correlation and error structures. Both alternatives involved modeling the missing variables through various regression techniques. In one approach, empirical residuals were added to predicted values, while the other approach adopted a form of statistical matching where the item with the missing value was assigned the reported value from the item with the closest predicted value.

2. Data Background

Starting with 20,278 NHANES III records, those corresponding to nonadults (8,138 records) were deleted. Furthermore, it was decided that records for which all six MEC variables were missing (3,181) be deleted from the file as well. Consequently, the imputation procedure was carried out with a total of 8,959 adult records with an observed value for at least one of the target MEC variables.

3. Imputation Methods

Two separate imputation methods were investigated using independent imputation programs specifically developed for this purpose by Westat, Inc. For the first method, WESMATCH[1] was created, which is capable of performing various forms of Statistical Matching. This SAS macro has been used to perform a version of predictive mean matching imputation. For the second method, Hane-Deck[2], a customized imputation program was developed to perform a hybrid Hot-Deck imputation.

Common to both methods was the initial step in which a model was built by regressing the target variable on a carefully selected set of covariates. For this purpose, the missing values of all categorical covariates were presented as separate categories. With the regression coefficients estimated, a predicted value was computed for the target variable in question. For this step, all missing values of the continuous covariates were replaced by their averages within gender, age, and race categories. Hence, for all 8,959 records a predicted value was computed.

3.1 Method I

For this method a variation on the conventional regression method of imputation was used. Following

[1] WEStat's statistical MATCHing macro, Version 1.0, June 1992.

[2] Hane-Deck Imputation Macro, Version 1.1, June 1992, Westat, Inc.

the computation of predicted values for a target variable, in the second step, all missing values of that variable were imputed via WESMATCH. For this purpose, donors were selected based on the proximity of their predicted values. Specifically, all observations were first listed by ascending order of the computed predicted values. Then, when a record with a missing value was encountered, the observed value corresponding to the record with the closest predicted value with the least number of prior donations was selected. Each donor could donate three times before it was banned from further donation.

3.2 Method II

As mentioned earlier, the first step of the second method was identical to that of the first one. With a complete set of predicted values, the second step of this method consisted of imputation of all missing residuals via Hane-Deck. For this purpose, all records were sorted by gender, age, and race; and pools of donors and missing values were created within the resulting pools. Next, all missing residual values were imputed by selecting donors from the corresponding donor pools of residuals. The missing values of the target variable were then imputed by adding the imputed residuals to the corresponding predicted values.

4. Details of Imputation for Individual Variables

As stated earlier, each of the six target variables were imputed twice. These variables were imputed consecutively, incorporating the completed values of successive variables in subsequent imputations. For instance, completed LGHT was used to impute LGWT, completed LGHT and LGWT were used to impute D_SYS, etc. Of note, for consistency purposes, only one version of these imputed variables (method I) was used as predictor variables in subsequent imputations. In what follows, some details of the model-building process for each of these variables are described.

LGHT: Starting with a set of potential covariates deemed correlated with LGHT, various models were examined. The selected model included the AGE, SEX, RACE, LGWT, and self-reported log values of height and weight as predictors. The emerging model displayed an R-squared of 0.632.

LGWT: Using the completed LGHT, the selected model for this case contained AGE, SEX, RACE, LGHT, and self-reported log values of height and weight with a resulting R-squared of 0.922.

D_SYS: Using the completed LGHT, the selected model for this case contained the following set of covariates: AGE, SEX, RACE, A_B1, A_D1, A_R3, HBP4, LGHT, D_DIAS, and self-reported systolic and diastolic blood pressures (Section 7 provides a description of these variables). The emerging model displayed an R-squared of 0.676. Note that for this case LGWT did not appear significant.

D_DIAS: Using the completed values of LGHT and D_SYS, the selected model for this case contained the following set of covariates: AGE, SEX, RACE, A_B1, A_D1, A_R3, HBP4, LGHT, D_SYS, and self-reported systolic and diastolic blood pressures with a resulting R-squared of 0.503.

TCRESULT: The task of selecting an appropriate model for this variable was the most time consuming of all. After unsatisfactory results with numerous models, it was decided to take a closer look at the distribution of this variable. Consequently, it was decided to eliminate a total of 8 severe outliers (those with a value larger than 400). Furthermore, due to skewness, a log transformation was applied to both cholesterol variables prior to model-building. Finally, using the completed LGHT, D_SYS, and D_DIAS, the selected model for this case contained the following set of covariates: AGE, MARRIED, A_D1, A_E7, LGHT, D_SYS, D_DIAS, and LGHDL, where the latter represented the natural log of HDRESULT. The resulting R-squared was 0.249.

HDRESULT: For this case too, the above outlier records were excluded and the log transformation was applied. Using the completed LGHT, D_SYS, D_DIAS, and TCRESULT, the selected model contained the following set of covariates: SEX, RACE, MARRIED, A_F10, A_R3, ALCOHOL, LGHT, LGWT, D_SYS, and LGTCR where the latter represented the natural log of TCRESULT. The resulting R-squared was 0.212.

5. Results

The following tables summarize some descriptive statistics for the six target MEC variables. Table 1 contains simple statistics, while the second table shows the correlation matrix for the original set, Method I imputed, and method II imputed variables, respectively.

Table 1. Descriptive statistics for the six MEC variables before and after imputations

MEC	*Original*		*Method I*		*Method II*	
Variable	*Mean*	*Std. Dev.*	*Mean*	*Std. Dev.*	*Mean*	*Std. Dev.*
LGHT	5.11	0.06	5.11	0.06	5.11	0.06
LGWT	4.28	0.22	4.27	0.23	4.28	0.23
D_SYS	123.95	20.81	124.21	20.79	124.18	20.85
D_DIAS	72.56	11.73	72.56	11.70	72.57	11.74
TCRESULT	205.57	45.41	205.32	45.29	205.23	45.31
HDRESULT	51.62	15.42	51.73	15.39	51.74	15.50

Table 2. Correlation among the six MEC variables before and after imputations

Original *Method I* *Method II*	*LGHT*	*LGWT*	*D_SYS*	*D_DIAS*	*TCRESULT*	*HDRESULT*
LGHT	1.00	0.45	-0.02	0.18	-0.10	-0.16
	1.00	0.44	-0.01	0.18	-0.10	-0.15
	1.00	0.44	-0.01	0.18	-0.10	-0.16
LGWT		1.00	0.16	0.32	0.10	-0.32
		1.00	0.14	0.31	0.09	-0.31
		1.00	0.14	0.31	0.09	-0.31
D_SYS			1.00	0.48	0.27	-0.02
			1.00	0.47	0.27	-0.02
			1.00	0.45	0.27	-0.02
D_DIAS				1.00	0.18	-0.07
				1.00	0.18	-0.06
				1.00	0.18	-0.07
TCRESULT					1.00	0.10
					1.00	0.09
					1.00	0.09
HDRESULT						1.00
						1.00
						1.00

5.1 Alternative Imputation of Cholesterol Measurements

Following the completion of the imputation of all six MEC variables, for methodological curiosity, it was decided to impute the missing values of the two cholesterol variables, TCRESULT and HDRESULT, using different regression models for different subsets of the data.

Of the 692 records with a missing value for the TCRESULT, 691 had a missing value for the HDRESULT as well. Previously, a model was developed by regressing the TCRESULT on a set of covariates including the HDRESULT. With regression coefficients estimated, predicted values were computed for all records. For this purpose, whenever the HDRESULT was missing, its average within gender, race, and age categories was used.

While inclusion of the HDRESULT variable in the regression model was well justified from a model-building point of view, questions were raised regarding the effect of using the average value of this variable for creation of a predicted value for TCRESULT with a missing HDRESULT. Particularly, it was prognosticated that a number of the resulting empirical residuals might be of improper magnitude.

5.2 TCRESULT

In order to remedy the above potential problem, a new regression model was developed that did not include HDRESULT as a predictor, eliminating the need for use of its crudely imputed value (cell averages) for construction of predicted values. This model included AGE, MARRIED, A_D1, A_E7, LGWST, and the Method I imputed values of LGHT, D_SYS, and D_DIAS with a resulting R-square of 0.231. Of note, aside from elimination of HDRESULT, this model-building process and the subsequent steps were identical to those used during the previous imputations of TCRESULT.

With a predicted value computed for all 8,959 records, the two alternative imputed values were created using methods I and II, respectively. For reference purposes, these imputed values will be labeled as the result of methods III and IV. Table 3 summarizes some simple statistics, while Table 4 presents the correlation matrix for the original TCRESULT and its four imputed values.

5.3 HDRESULT

Using the completed values of TCRESULT (WESTC3) the process of imputing HDRESULT was repeated once again. The two resulting imputed variables (WESHD3 and WESHD4) were created via methods I and II. Analogously, Table 5 summarizes some simple statistics, while Table 6 presents the correlation matrix for the original HDRESULT and its four imputed values.

Table 3. Simple Statistics for the Original and Imputed values of TCRESULT

Variable	*Sample*	*Mean*	*SD*	*Min*	*Max*
TCRESULT	8,267	205.570	45.397	77	702
Method I	8,959	205.323	45.297	77	702
Method II	8,959	205.226	45.314	77	702
Method III	8,959	205.317	45.293	77	702
Method IV	8,959	205.148	45.340	77	702

Table 4. Correlation for the Original and Imputed values of TCRESULT

	TCRES	*Mtd.*	*Mtd.*	*Mtd.*	*Mtd.*
Variable	*ULT*	*I*	*II*	*III*	*IV*
TCRESULT	1.000	1.000	1.000	1.000	1.000
Method I	1.000	1.000	0.942	0.942	0.942
Method II	1.000	0.942	1.000	0.947	0.943
Method III	1.000	0.942	0.948	1.000	0.943
Method IV	1.000	0.942	0.943	0.943	1.000

Table 5. Simple Statistics for the Original and Imputed values of HDRESULT

Variable	*Sample*	*Mean*	*SD*	*Min*	*Max*
HDRESULT	8,186	51.622	15.423	2.000	191.00
Method I	8,959	51.735	15.391	2.000	191.00
Method II	8,959	51.738	15.496	2.000	191.00
Method III	8,959	51.726	15.448	2.000	191.00
Method IV	8,959	51.698	15.381	2.000	191.00

Table 6. Correlation for the Original and Imputed values of HDRESULT

	HDRES	*Mtd.*	*Mtd.*	*Mtd.*	*Mtd.*
Variable	*ULT*	*I*	*II*	*III*	*IV*
HDRESULT	1.000	1.000	1.000	1.000	1.000
Method I	1.000	1.000	0.931	0.930	0.931
Method II	1.000	0.931	1.000	0.927	0.926
Method III	1.000	0.930	0.927	1.000	0.928
Method IV	1.000	0.931	0.927	0.928	1.000

6. Conclusions

Upon comparing the simple statistics as summarized in Table 1, it is evident that the two imputed values exhibit nearly identical distributions. Moreover, Tables 3 and 5 seem to indicate the same conclusion for the alternative methods of imputing the cholesterol measurements. Further support for these conclusions can be drawn from the corresponding correlation matrices. Consequently, there does not seem to be any methodological grounds for discriminating among these two (and the alternative) imputation methods. Computationally, however, method II (empirical residuals using the entire data set) does possess a few appealing features.

In addition to computational ease, Method II does benefit from utilization of a well-tested imputation macro, Hane-Deck. This customized macro has a number of fine tuning options that enable the user to impose various types of control on the imputation process (e.g., formation of donor pools, definition of eligible donors, maximum number of donation for each donor). Moreover, Hane-Deck can be used for imputation of categorical variables.

On the negative side, with Method II (as with any other prediction) there does exist the possibility of assigning impossible values to a missing item. In fact, when imputing LGHT and LGWT, a handful of individuals did receive imputed values that were beyond reasonable limits. There are, however, a number of trivial remedies for this problem.

7. Description of Variables

A_B1: Self-perceived health status with 5 categories.

A_D1: Indicator for ever being diagnosed with diabetes.

A_E7: Indicator for ever being diagnosed with high cholesterol.

A_F10: Indicator for ever having a heart attack.

A_R3: Current smoking indicator.

HBP4: Indicator for high blood pressure.