A COMPARISON OF IMPUTATION TECHNIQUES IN THE THIRD NATIONAL HEALTH AND NUTRITION EXAMINATION SURVEY

Trena M. Ezzati-Rice[1], Meena Khare[1], Donald B. Rubin[2],
Roderick J. A. Little[3], Joseph L. Schafer[4]
[1]National Center for Health Statistics, [2]Harvard University,
[3]University of Michigan, [4]Pennsylvania State University
Trena M. Ezzati-Rice, 6525 Belcrest Road, Hyattsville, MD 20782

KEY WORDS: Missing data, item nonresponse

Introduction

The National Health and Nutrition Examination Survey (NHANES) is a periodic survey conducted by the National Center for Health Statistics (NCHS), Centers for Disease Control and Prevention. The NHANES is designed to provide national estimates of the health and nutritional status of the civilian noninstitutionalized population. Sociodemographic and medical history information are obtained through household interviews, while physical measurements, physiological tests, and biochemical measurements are collected through standardized physical examinations in mobile examination centers (MECs). The on-going Third NHANES or NHANES III is the seventh of an extensive series of periodic health and nutrition surveys that NCHS has conducted since 1960. The current NHANES III, with a sample of approximately 40,000 sample persons 2 months of age and older, has been divided into two 3-year national samples. Phase 1 was conducted from October 1988 to October 1991 while Phase 2 will continue until October 1994.

NHANES III is based on a complex, multistage area probability sample design and includes an oversample of children under 5 years of age, older Americans aged 60+ years, and both black and Mexican-American persons. Details of the sample design of NHANES III have been previously published (1).

NHANES III, like most sample surveys, experiences both total (unit) nonresponse and item nonresponse. The missing data problem for NHANES III is somewhat unique since sample persons can refuse to participate at three different stages of the data collection. Unit nonresponse rates for NHANES III-Phase 1 ranged from 0% for the screening interview (with about 7% of the screening data obtained from neighbors) to 14% for the household interview to 22% for the physical examination. It is common survey practice to compensate for unit nonresponse through weighting class adjustments (2-5). The adjustments to reduce potential nonresponse bias for NHANES III-Phase 1 have been previously described (6). In addition to unit nonresponse, various levels of item nonresponse occur in NHANES III. In Phase 1, item nonresponse of 1-5% occurred for the household interview questions. In addition, some components of the physical examination were not successfully completed for all sample persons. Furthermore, some examination components include a number of individual measurements (e.g., body measurements)--some of which may be missing. Item nonresponse rates for the individual components ranged from 5-8%. Generally, item nonresponse is handled by some type of imputation. Imputation methods fill in missing items with values from similar units in the dataset or with predicted values obtained from a model, thus making it possible to analyze the data as if it were complete. Some common methods of imputation used in surveys include deductive imputation, mean imputation, Hot Deck imputation, Cold Deck imputation, regression imputation, stochastic regression, multiple imputation, and composite imputation methods (7). Each of these imputation methods has relative advantages and disadvantages. The method of choice for a survey may depend upon particular circumstances including the type of survey data and availability of computer hardware and software. In addition to allowing complete data methods of analysis, multiple imputation allows one to assess the impact of missing data uncertainty on the variances and to revise estimates of variance to reflect the additional uncertainty (8). In previous NHANES surveys, imputation for item nonresponse was done on an *ad hoc* basis. The purpose of this paper is to describe research conducted to compare alternative missing data adjustment methods for selected survey components in NHANES III- Phase 1 based on single and multiple imputation methodology. The information contained in this paper, in part, is based on a special project carried out during 1992 and contained in a final report by Datametrics Research, Inc. (9).

Methods

For this investigation, two single imputation (SI) methods applying two closely related regression techniques were used (10). The first, "SI_1", involved predictive mean matching and the second, "SI_2", involved a Hot Deck regression procedure in which empirical residuals were added to the predicted values. The other method, multiple imputation or MI, was

based on a multivariate model for mixed normal and categorical data. Ten multiple imputations were generated by iterative simulations using the Gibbs sampler and an E-M type algorithm was used for parameter estimation (11,12). The dataset for this research was restricted to adults 17 years and older. The initial test dataset included selected demographic items and survey location information from the screener questionnaire as well as selected auxiliary variables from the household interview to impute for 6 target examination variables: height, weight, diastolic blood pressure, systolic blood pressure, HDL cholesterol, and total cholesterol.

Some key characteristics of the two imputation methods as applied to the NHANES III-Phase 1 test dataset are shown in Table 1. Under the MI model, data were imputed for both unit and item nonresponse, while the two SI methods were used to impute for only item nonresponse for height, weight, systolic and diastolic blood pressure, and total and HDL cholesterol. Using exploratory graphical analyses, some selected outliers were excluded for MI, while for SI the only exclusions were persons missing all six examination variables and survey location. Consequently, MI was used to multiply impute over 4000 values among 27 variables (11,12), while the two SI methods were used to impute only 690 values for the six examination variables. Under the SI model, the missing values were filled in sequentially which is appropriate for data such as from NHANES III with a monotone missing data pattern. The MI algorithm which can handle any missing data pattern was used to impute all missing values simultaneously. Both methods essentially used linear regression models to predict missing values. MI then added normal deviates, while the SI methods added noise in the form of residuals from matched cases.

Results

The results presented focus on single and multiple imputations for item nonresponse among examined persons only. Table 2 shows relative differences by race/ethnicity between the SI_1 imputed and observed estimates and between the SI_1 and SI_2 values, where SI_1 was the statistical mean matching method and SI_2 was the Hot Deck residual regression method. The last column of the table shows that the amount of missing data imputed ranged from 4 to 14%. The two singly imputed values resulted in extremely small differences even when imputing for 14% missing data, so from a methodological standpoint it does not appear that one method is superior to the other. However, Fahimi has suggested that there are some computational advantages of the empirical residual method (10). The imputed values were very similar to the observed values with the exception of mean weight for Mexican-Americans where the imputed values were slightly higher than the observed values.

Table 3 shows the relative difference in the observed and the imputed values from 10 imputations for the same six examination variables by race/ethnicity. Again, all the differences were less than 1% except for mean weight among Mexican-Americans.

For comparison of estimates from the SI and MI models, the MLE estimates (MI_0) from MI were compared to the predictive mean matching (SI_1) estimates (Table 4). For each variable, the relative differences were all less than 1% when datasets with imputations generated by both single and multiple imputation methods were compared. Some graphical analyses were also conducted to compare the marginal distributions of the imputed and observed data. Boxplots of observed and imputed data from SI and MI for each of the six variables showed that all three imputation methods (the two SI methods and MI) preserved the median and quartile distributions. Some significant outliers for the SI methods were the result of the use of preliminary data which had not been completely edited. For MI, a number of extreme values were excluded through exploratory graphical analyses, thus MI resulted in slightly more compact distributions. We also examined the data to see how well the imputed datasets maintained the relationship between variables. Bivariate scatterplots of height vs. weight, diastolic blood pressure vs. systolic blood pressure, and HDL cholesterol vs. total cholesterol showed that the marginal distributions were very similar for all three imputation methods.

Variance estimates from the imputed datasets were also examined. Variances were computed using the Taylor linearization method in SUDAAN (13). The design effects (DEFFs) varied across the multiply-imputed datasets reflecting the instability of the SUDAAN standard errors. The DEFFs also varied by race/ethnicity due to the oversampling of the two minority populations in NHANES III. Nevertheless, the DEFFs were remarkably similar for the single and multiply imputed datasets (Table 5).

To examine the effect of the number of imputations on the variance estimates, we also looked at the efficiency of MI as described by Rubin (8). Rubin points out that imputing multiple draws leads to some loss of efficiency relative to asymptotically efficient procedures such as maximum likelihood. However, the loss of efficiency tends to zero as m (the number of imputations) tends to infinity. Assuming proper imputation from a correct model, the variance of estimates from MI increases asymptotically by the factor, $(1 + \frac{\gamma}{m})$, where γ is the fraction of missing information. Most of the NHANES III missing data is

less than 15%, although higher levels of missing data can be found in some survey components or among some subgroups. Thus, if the amount of missing data is 15%, that is γ equals 0.15, the increase in variance would be 15% with single imputation and 3% with five multiple imputations. With 10 and 15 multiple imputations, the relative decrease in the increase in variance is not as great as for the five multiple imputations when compared to single imputation. Therefore, for most of the NHANES III-Phase 1 data, five multiple imputations would likely be sufficient as recommended by Rubin and Little (9).

Summary

Three imputation methods were developed and compared to assess potential imputation strategies for NHANES III-Phase 1 data. For the subset of data evaluated, values generated from two separate single imputation methods exhibited nearly identical distributions. In addition, the single and multiple imputation methods exhibited similar point estimates. Also, both methods preserved the marginal distribution of the variables and the relationship between them.

Before specific recommendations on imputation strategies for NHANES III can be made, a number of important issues must be addressed and additional research undertaken. Although software to generate the single imputations for this project was developed in SAS and is fairly easy to implement, methods like Hot-Deck imputation are complex and difficult to implement in multivariate datasets. On the other hand, the model-based multiple imputation technique works well in the multivariate survey setting, but it requires specialized computing. Therefore, the development and implementation of appropriate MI software is critical. Further research is needed to enhance the MI model used in this investigation so as to include more variables and to extend the model to additional NHANES III survey components. Other important areas of future research include a simulation study to assess the validity of the methods and the development of better methods of variance estimation such as model-based variances. Finally, as briefly discussed by Johnson *et al.*(14), the issue of how to disseminate multiply imputed datasets on public-use files must be addressed.

References

1. Ezzati, T., Massey, J., Waksberg, J., Chu, A., and Maurer, K. Sample Design: Third National Health and Nutrition Examination Survey. National Center for Health Statistics. Vital Health Statistics. Series 2, No. 113, 1992.
2. Madow, WG., Olkin, I., and Rubin, DB, eds. Incomplete Data in Sample Surveys, Volume 2, New York: Academic Press, 1983.
3. Cox, B. Weighting Survey Data for Analysis. Presentation for the ASA Continuing Education Program. 1991 Joint Statistical Meetings, August, 1991.
4. Kalton, G. and Kaspryzk, D. The Treatment of Missing Survey Data. Survey Methodology, Vol. 12, No. 1, 1-17, Statistics Canada, 1986.
5. Little, RJA. Survey Nonresponse Adjustments for Estimates of Means. International Statistical Review 1986: 54:139-157.
6. Ezzati, T. and Khare, M. Nonresponse Adjustments in a National Health Survey. 1992 Proceedings of the Survey Research Methods Section of the American Statistical Association, 339-344.
7. Little, RJA and Rubin DB. "Statistical Analysis with Missing Data." John Wiley & Sons, New York, 1987.
8. Rubin, DB. "Multiple Imputation for Nonresponse in Surveys." John Wiley & Sons, New York, 1987.
9. Little, RJA and Rubin, DB. Final Report on NHANES Imputation Project. Datametrics Research, Inc., Waban, MA, December, 1992.
10. Ezzati-Rice, TM; Fahimi, M; Judkins, D; Khare, M. Serial Imputation of NHANES III with Mixed Regression and Hot-Deck Techniques. 1993 Proceedings of the Survey Research Methods Section of the American Statistical Association. In press.
11. Schafer, JL; Khare, M; Ezzati-Rice, T. Multiple Imputation of Missing Data in NHANES III. Proceedings of the 1993 Annual Research Conference, U.S. Bureau of the Census, Washington, D.C. In press.
12. Khare, M; Little, RJA; Rubin, DB; Schafer, JL. Multiple Imputation of NHANES III. 1993 Proceedings of the Survey Research Methods Section of the American Statistical Association. In press.
13. SUDAAN: Software for Survey Data Analysis. Research Triangle Institute, 1992.
14. Johnson, CL; Curtin, LR; Ezzati-Rice, TM;Khare, M; Murphy, RS. Single Versus Multiple Imputation: The NCHS Perspective. 1993 Proceedings of the Survey Research Methods Section of the American Statistical Association. In press.

Table 1. Comparison of Single and Multiple Imputation Methods for NHANES III-Phase 1

Characteristic	Single Imputation	Multiple Imputation
Data Imputed	Examination item nonresponse only	All unit and item nonresponse
How Imputed	Sequentially	All simultaneously
Model	Linear regression	Multivariate normal
Number of Variables	6	27
Number of Imputed Values	690	>4000
Exclusions	Persons missing all 6 variables and PSU location	Selected outliers

Table 2. Relative Difference in Estimates between Two Single Imputation Methods* (Examined persons only)

Examination Variable	Relative Difference (%)		Percent Imputed
	$(SI_1 - SI_2)/SI_1$	$(Obs - SI_1)/Obs$	
Average Systolic Blood Pressure			
White/Other	0.00	-0.08	3.8
Black	-0.08	0.08	5.5
Mexican-American	0.17	-0.17	8.8
Average Diastolic Blood Pressure			
White/Other	0.00	0.00	3.9
Black	0.13	-0.13	5.7
Mexican-American	0.00	-0.14	8.6
Height			
White/Other	-0.06	0.06	2.5
Black	0.06	-0.06	4.0
Mexican-American	0.00	0.06	4.2
Weight			
White/Other	0.00	-0.13	4.8
Black	-0.13	-0.39	5.8
Mexican-American	0.00	-0.98	7.7
Total Cholesterol			
White/Other	0.15	0.24	5.8
Black	-0.20	0.00	14.0
Mexican American	-0.15	0.20	5.1
HDL Cholesterol			
White/Other	0.39	-0.20	7.0
Black	0.00	0.00	14.3
Mexican-American	-0.20	0.20	6.0

*SI_1 = predictive mean matching; SI_2 = Hot Deck with empirical residuals.

Table 3. Relative Difference in Estimates due to Multiple Imputation (m = 10) (Examined persons only)

Examination Variable	Relative Difference (%)*	Percent Imputed
Average Systolic Blood Pressure		
White/Other	-0.08	3.9
Black	0.04	5.9
Mexican-American	-0.05	8.9
Average Diastolic Blood Pressure		
White/Other	0.06	4.0
Black	-0.05	6.1
Mexican-American	0.07	8.8
Height		
White/Other	-0.05	2.7
Black	-0.14	4.5
Mexican-American	-0.08	4.4
Weight		
White/Other	-0.13	5.0
Black	-0.62	6.3
Mexican-American	-1.00	8.0
Total Cholesterol		
White/Other	0.21	5.9
Black	-0.13	14.4
Mexican-American	0.05	5.3
HDL Cholesterol		
White/Other	-0.05	7.2
Black	0.04	14.7
Mexican-American	-0.13	6.2

*Relative difference = (Obs - MI10)/Obs

Table 4. Relative Difference in Estimates from Single and Multiple Imputation Models* (Examined persons only)

Examination Variable	Relative Difference (%)		Percent Imputed
	$(MI_0 - SI_1)/MI_0$	$(Obs - MI_0)/Obs$	
Average Systolic Blood Pressure			
White/Other	0.07	-0.15	3.7
Black	0.07	0.01	5.5
Mexican-American	0.03	-0.20	8.7
Average Diastolic Blood Pressure			
White/Other	0.03	-0.03	3.9
Black	-0.08	-0.05	5.7
Mexican-American	-0.22	0.08	8.6
Height			
White/Other	0.12	-0.07	2.5
Black	0.10	-0.15	4.0
Mexican-American	0.14	-0.08	4.2
Weight			
White/Other	-0.03	-0.11	4.8
Black	0.22	-0.61	5.8
Mexican-American	0.03	-1.01	7.7
Total Cholesterol			
White/Other	-0.02	0.27	5.8
Black	0.00	0.00	14.0
Mexican American	0.08	0.11	5.1
HDL Cholesterol			
White/Other	-0.02	-0.18	7.0
Black	-0.07	0.07	14.3
Mexican-American	0.24	-1.04	6.0

*SI_1 = predictive mean matching; MI_0 = MLE estimate

Table 5. Comparison of Design Effects from Single and Multiple Imputation Methods (Examined persons only)

Examination Variables	Observed	Single (SI_1)	Multiple (MI_0)
Average Systolic Blood Pressure			
White/Other	1.5	1.6	1.6
Black	1.0	1.1	1.1
Mexican-American	1.0	1.1	1.1
Average Diastolic Blood Pressure			
White/Other	1.9	1.9	1.9
Black	1.3	1.3	1.3
Mexican-American	1.7	1.9	2.2
Height			
White/Other	1.3	1.3	1.3
Black	1.0	1.0	1.0
Mexican-American	1.1	1.2	1.1
Weight			
White/Other	1.1	1.1	1.1
Black	1.5	1.5	1.5
Mexican-American	1.2	1.2	1.3
Total Cholesterol			
White/Other	1.1	1.1	1.2
Black	1.5	1.5	1.5
Mexican-American	1.6	1.6	1.7
HDL Cholesterol			
White/Other	1.3	1.3	1.3
Black	1.5	1.4	1.6
Mexican-American	1.8	1.8	1.8

*SI_1 = predictive mean matching; MI_0 = MLE estimate