METHODOLOGY FOR MISSING DATA IN THE U.S. NATIONAL HEALTH AND NUTRITION EXAMINATION SURVEY

P.D. Williams, National Center for Health Statistics
and
H. Nisselson, Westat, Inc.

P.D. Williams, 3700 East-West Highway, Center Building, Hyattsville, MD 20782

KEY WORDS: Imputation, nonresponse, replication

1. Introduction

The U.S. National Health and Nutrition Examination Surveys (NHANES) are a series of surveys, sponsored by the National Center for Health Statistics (NCHS), that use both interview and physical examination procedures to collect a variety of medical and nutritional data, and related demographic, socio-economic, and morbidity information. This paper is based on the second survey, NHANES II, conducted during 1976-1980. A sample of 27,801 persons was selected in a multi-stage design with 64 primary sampling units (PSUs). In-person interviews, were completed for 91 percent of the sample. Interviewed persons were invited to participate in the examination phase of the survey. Approximately 80 percent of those interviewed were examined, representing 73 percent of the initial sample.

For some variables, especially those requiring blood and urine samples, data may be missing for a relatively large percent of examined persons. For example, among biochemical variables, missing data rates varied from 7 percent for hematocrit to 38 percent for serum vitamin A. A weighting adjustment was carried out for individuals not examined. However, typically no imputation was carried out for missing biochemical variables except hemoglobin.

2. Study objectives

This paper presents an evaluation of three methods for imputing missing data for each of eight blood and urine related variables. For two variables, an investigation is presented of the effect of imputation on the total variance of NHANES estimates, and methods for estimating the total variance.

3. Imputation methods

The three imputation methods used in the study are a "nearest neighbor" hot deck based on an ordered file, a stochastic regression method, and a random hot deck (Little and Rubin, 1987). The effectiveness of the imputation methods depends on the imputation classes used and, for the "nearest neighbor" and regression methods, the covariate available. Imputation was carried out within 18 classes defined by age, race and sex (ARS), for 10 pairs of variables and covariate (in parentheses) of choice for them; for example, hemoglobin (hematocrit) and transferrin saturation (serum iron). Two types of missing data cases may be encountered: the selected covariate was reported or was also missing. When the covariate of choice was missing, age was used as the covariate. Determining the imputation bias requires knowledge of the "true" values. Since these were not known for each variable-covariate pair, we created a data set of "synthetic nonrespondents" for whom values were reported in the examination survey to serve as the evaluation population, as follows. Within the ARS classes, for each individual with missing data a synthetic nonrespondent was selected who matched that individual's covariate value most closely. This procedure was pursued to simulate the unknown decision process which accounts for nonresponse. No individual was used as a synthetic nonrespondent more than once for any data set created unless the file in the imputation class was depleted. Values to impute for synthetic nonrespondents for a variable of interest were selected, according to the imputation methods, from the survey data file excluding both the individuals in the survey with missing values for the variable and their corresponding synthetic nonrespondents.

4. Total variances and components

Total variances of estimates of mean hemoglobin and transferrin saturation, including the between-imputation component, were computed by BRR methods with 32 replicates. An estimated sampling variance was computed from the 32 half-sample estimates with no imputation for missing data. A between-imputation component of variance was estimated from the variance of four estimates created by four replicate imputations (Herzog and Rubin, 1983). The total variance is the sum of these components (equation 1). A second estimate was done using independent imputations within each replication (equation 2). For the ordered hot deck, four "nearest neighbors" were selected without replacement, two forward in the data file and two backward, and randomized. For the random hot deck four donors were selected without replacement.

For the regression method the principal analysis was based on the simple linear model. The imputed value is the predicted value plus a random residual. For generating replicate sets of imputed values, it was assumed that the

$$V_{T1}^2 = \mathrm{Var}(\theta_r) = \mathrm{Var}(\tilde{\theta}) + V_B^2(\hat{\theta}) \qquad (1)$$

where

$$V_S^2 = \mathrm{Var}(\tilde{\theta}) = \sum_{\alpha=1}^{32} \frac{(\tilde{\theta}_\alpha - \tilde{\theta})^2}{31}$$

$$V_B^2(\hat{\theta}) = \sum_{i=1}^{4} \frac{(\hat{\theta}_i - \hat{\theta})^2}{3}$$

$\tilde{\theta}_\alpha$ = estimate of parameter (such as mean cholesterol) within the α-th half sample, with-imputation;

$\bar{\tilde{\theta}}$ = mean of $\tilde{\theta}_\alpha$'s;

$\hat{\theta}_i$ = estimate of the parameter for i-th imputation

$\hat{\theta}$ = mean of $\hat{\theta}_i$'s.

$$V_{T2}^2 = \mathrm{Var}(\hat{\hat{\theta}}_r) = \sum_{\alpha=1}^{32} \frac{(\hat{\hat{\theta}}_\alpha - \hat{\hat{\theta}})^2}{31} \qquad (2)$$

where

$\hat{\hat{\theta}}_\alpha$ = Estimate of parameter within the α-th half sample, with imputation

$\hat{\hat{\theta}}$ = mean of the $\hat{\hat{\theta}}_\alpha$'s.

regression coefficients are normally distributed with mean and variance equal to those computed from the survey data file. Three additional regression coefficients were sampled from this distribution to create the replicate imputations. A total variance was then computed as the sum of the sampling variance and the between-imputation component.

5. Discussion

An indication of the size of the nonresponse problem for the variables in this study may be obtained from Table 1. The greatest impact appears to be among blacks in most age groups. This creates particular problems as even with over-sampling of blacks the numbers are often small reducing the size of the donor groups for imputation.

Table 2 gives the percent bias for each medical variable without regard to the age-race-sex categories. Tables 3 and 4 average the effect across the ARS categories and across all medical variables respectively. The ordered hot deck imputation procedure seems to consistently out perform the other procedures for each of the three cross sections presented (i.e., total/medical variable, average for medical variable/ARS, and average for medical variable/ARS). Within each procedure the addition of information on highly correlated covariate reduces the bias but findings are obscured somewhat by small numbers as it is most common for both correlates to be missing when one is not present. The poor performance of the regression procedure is likely to be the result of a poor choice for the model. Much improvement may be realized through just the addition of an intercept.

Table 5 and 6 give indications of design effects for the different procedures and between variance estimation procedure (T1 vs T2). Procedure T2 yields smaller estimates of total variance as one might expect. The between imputation component of variance was small for each procedure with the largest effect being in classes where the frequencies are small and choice of donor very limited.

6. Summary of findings

Despite incomplete analysis, imputing for missing data for the eight variables studied is generally more satisfying than not imputing. Consider the bias of an unadjusted linear estimate across imputation classes, such as an unadjusted mean, factored into two components. Imputation: (1) achieves a completed data set with the relative weighting between-imputation classes adjusted for differences in missing data rates by class; and (2) leads to within-imputation class means generally closer to the means of the synthetic nonrespondents than are the ARS class means. Where the latter is not the case, the differences generally are not large. It must be recognized that the imputation bias for an individual's ARS class may be relatively large. However, such instances in the study were associated with small classes and high missing data rates. The ordered "nearest neighbor" hot deck was generally the most satisfactory among the imputation methods studied, regardless of the size of the covariate correlation. We also believe that it is the simplest of the methods to implement with a number of variables being of interest. Regression models that did not include an intercept performed worst of all. As expected, the bias when the covariate of choice was reported is less than when it was also missing. However, if the variable of interest was missing it was likely that the selected covariate was also missing. Since the interview and medical history provide an opportunity to ask questions that might help improve imputation, we investigated the possible use of medical history responses to define imputation

classes. We found that response rates and the mean value of variable of interest varied according to whether the person ever had or believes that they had a diagnosis for a related condition. For most of the ARS imputation classes and variables, the between-imputation variance was small. However, in a few instances the estimated total sampling error including that component was 10 to 20 percent higher than the pure sampling error.

References

1. Herzog, T.N. and Rubin, D.B. (1983). "Using Multiple Imputation to Handle Nonresponse in Sample Surveys" in Incomplete Data in Sample Surveys, Vol. 2, New York, Academic Press.

2. Little, Roderick J.A., and Rubin, D.B. (1987). Statistical Analysis with Missing Data, New York, Wiley & Sons.

3. Williams, P.D., Fullwood, R., and Johnson, C. "Response Problems and Imputation Processes for the National Health and Nutrition Examination Survey," Proceedings of the Social Statistics Section, American Statistical Association, 1983.

Table 1. Average missing data rate for five variables, based on persons examined; number of persons examined, total rate and rate with covariate reported or covariate also missing, by age, race, and sex* (hemoglobin, transferrin saturation, serum zinc, serum copper, and serum lead)

Race, sex, and age	Number of persons examined (unweighted counts)	Missing data rate (percent)		
		Total	Covariate not missing	Covariate missing
Total sample	19,868	13.5	6.6	6.9
Under 4 years				
White	2,023	20.5	3.8	16.7
Black	443	18.6	4.3	14.3
4-7 years				
White males	871	23.6	6.5	17.2
White females	821	26.6	6.4	20.3
Black males	174	27.1	8.4	18.7
Black females	210	25.0	6.2	18.9
8-14 years				
White males	903	15.2	7.0	8.2
White females	842	15.8	6.5	9.3
Black males	175	22.4	8.2	14.2
Black females	178	18.7	9.1	9.6
15-44 years				
White males	2,843	8.9	6.5	2.4
White females	2,986	9.6	6.7	2.8
Black males	408	9.4	3.9	13.2
Black females	469	14.6	11.0	3.5
45+ years				
White males	2,748	9.0	6.9	2.1
White females	3,068	9.0	6.1	2.8
Black males	319	13.2	9.7	3.6
Black females	387	15.0	10.7	4.2

*Whites and blacks. Excludes 454 other persons.

Table 2. Percent bias, based on three imputation procedures: Ordered hot deck, regression, and random hot deck; frequency of one variable and two variables missing, by variable imputed

Variable imputed (covariate in parenthesis)	Frequency		Imputation procedure					
	one variable missing	two variables missing	Ordered hot deck		Regression	Random hot deck		Correlation coefficient
			1 var. mis.	2 var. mis.	1 var. mis.	1 var. mis.	2 var. mis.	
			Percent bias					
Hemoglobin (hematocrit)	522	1288	-0.4	0.1	-0.2	-1.0	0.2	0.93
Transferrin saturation (iron)	1867	1597	-0.0	-11.1	0.4	-1.1	-0.2	0.91
Serum zinc (albumin)	1775	1322	0.1	0.2	0.1	-0.8	-0.3	0.25
Serum copper (albumin)	1897	1319	0.3	1.7	-2.4	-0.1	-2.0	-0.31
Serum lead (eryth. protoporphy.)	503	1289	1.8	-3.3	-20.0	-7.5	-1.3	0.16
Carboxyhemoglobin* (cigarette smoking)	403	37	5.6	-6.2	-51.8	6.8	32.3	0.71
Glucose tolerance (skinfold)*	1746	22	0.7	-4.6	-24.4	0.6	5.6	0.08
Glucose tolerance (wt/ht^2)*	1759	10	0.8	3.4	-3.0	0.8	4.8	0.26
Serum cholesterol (wt/ht^2)*	264	4	-0.9	-13.2	0.4	0.5	-6.4	0.16
Serum cholesterol (skinfold)*	259	9	1.8	3.6	-20.4	6.3	-13.7	0.16

*Limited to persons aged 15 and over

Table 3. Average percent bias for age-race-sex groups, based on three imputation procedures: Ordered hot deck, regression, and random hot deck; frequency of one variable and two variables missing, by variable imputed

Variable imputed (covariate in parenthesis)	Average frequency per group		Imputation procedure					
	one variable missing	two variables missing	Ordered hot deck		Regression	Random hot deck		Correlation coefficient
			1 var. mis.	2 var. mis.	1 var. mis.	1 var. mis.	2 var. mis.	
			Average percent bias					
Hemoglobin (hematocrit)	29	72	1.5	1.5	1.6	1.9	2.8	0.93
Transferrin saturation (iron)	104	89	3.2	12.2	4.7	9.3	14.8	0.91
Serum zinc (albumin)	99	73	2.3	2.4	4.2	2.9	2.7	0.25
Serum copper (albumin)	105	73	3.7	3.3	11.8	3.6	5.9	-0.31
Serum lead (eryth. protoporphy.)	28	72	12.9	8.6	32.7	11.5	6.3	0.16
Carboxyhemoglobin* (cigarette smoking)	51	7	9.3	--	63.5	45.4	--	0.71
Glucose tolerance (skinfold)*	218	3	3.8	--	30.3	1.4	--	0.08
Glucose tolerance (wt/ht^2)*	220	2	4.9	--	4.0	3.2	--	0.26
Serum cholesterol (wt/ht^2)*	33	1	9.5	--	10.3	7.8	--	0.16
Serum cholesterol (skinfold)*	32	1	5.3	--	24.2	8.6	--	0.16

*Limited to persons aged 15 and over

Table 4. Average percent bias for five variables based on three imputation procedures: Ordered hot deck, regression, and random hot deck, frequency of one variable and two variables missing, by age, race, and sex*

Race, sex, and age	Frequency		Imputation procedure				
			Ordered hot deck		Regression	Random hot deck	
	1 variable missing	2 variables missing	1 variable missing	2 variables missing	1 variable missing	1 variable missing	2 variables missing
			Average percent bias				
Total sample	1,312	1,363	.5	3.3	4.6	2.1	.8
Under 4 years							
White	77	338	4.4	3.3	9.2	5.3	4.0
Black	19	63	6.0	5.5	9.9	11.1	10.6
4-7 years							
White males	56	149	4.3	3.9	6.9	2.9	1.9
White females	52	166	3.2	4.5	8.1	5.3	2.4
Black males	15	33	2.9	6.1	14.8	13.7	5.8
Black females	13	40	3.1	2.3	10.6	6.4	4.9
8-14 years							
White males	64	74	4.1	5.5	5.2	5.5	2.9
White females	43	79	1.8	4.3	10.0	2.4	3.6
Black males	14	25	2.8	5.3	15.7	4.5	6.0
Black females	16	17	14.4	12.1	14.6	9.6	8.9
15-44 years							
White males	185	69	.8	4.2	10.7	5.4	3.5
White females	201	85	2.4	4.1	9.6	4.1	3.3
Black males	38	15	5.0	4.8	14.0	2.3	4.8
Black females	52	17	5.2	10.2	22.5	9.0	12.5
45+ years							
White males	189	59	1.3	3.8	10.2	3.8	3.2
White females	189	87	6.6	4.2	6.3	2.3	1.9
Black males	31	11	9.7	8.1	12.4	10.5	4.9
Black females	42	16	7.3	7.4	7.1	3.8	8.4

*Hemoglobin, transferrin saturation, serum zinc, serum copper, and serum lead.

Table 5. "Design effects" for estimates of variance of mean transferrin saturation after imputation with covariate, based on three imputation procedures: Ordered hot deck, regression, random hot deck by age, race, and sex (covariate is iron, r = .91)

Race, sex, and age	Number of persons examined (unweighted counts)*	Mean transferrin saturation (after imputation)**	Imputation procedure					
			Ordered hot deck		Regression		Random hot deck	
			V^2_{T2} / V^2_S	V^2_{T2} / V^2_{T1}	V^2_{T2} / V^2_S	V^2_{T2} / V^2_{T1}	V^2_{T2} / V^2_S	V^2_{T2} / V^2_{T1}
Under 4 years								
White	2,023	21.7	1.26	.75	1.01	.88	1.01	.72
Black	443	19.1	1.02	1.16	1.01	.56	1.01	1.25
4-7 years								
White males	871	23.1	1.00	.96	1.03	.70	1.21	.83
White females	821	23.8	1.10	.97	1.06	.88	1.04	1.35
Black males	174	21.4	1.60	.60	1.06	1.33	1.22	.74
Black females	210	21.9	1.04	1.04	1.03	1.39	1.02	.69
8-14 years								
White males	903	24.2	1.02	1.05	1.01	.93	1.05	1.41
White females	842	25.4	1.04	1.22	1.02	1.05	1.20	.97
Black males	175	22.6	1.09	1.23	1.04	1.02	1.13	1.22
Black females	178	23.9	1.11	1.12	1.02	.74	1.13	.94
15-44 years								
White males	2,843	29.8	1.06	1.11	1.01	.85	1.05	1.04
White females	2,986	26.9	1.02	.91	1.00	.86	1.05	.95
Black males	408	29.6	1.09	.89	1.02	.91	1.31	1.35
Black females	469	24.1	1.01	1.16	1.13	.92	1.08	.83
45+ years								
White males	2,748	28.6	1.05	1.30	1.00	1.01	1.13	1.02
White females	3,068	26.1	1.04	1.01	1.02	.98	1.03	1.22
Black males	319	25.0	1.09	.83	1.01	1.01	1.20	1.14
Black females	387	23.4	1.37	.60	1.02	.77	1.05	.76

*Excludes 454 other persons. V^2_{T2} is estimated total variance based on independent imputation by half sample.
**Average of 4 iterations.

Table 6."Design effects" for estimates of variance of mean hemoglobin after imputation with covariate, based on three imputation procedures: Ordered hot deck, regression, random hot deck by age, race, and sex (covariate is hematocrit, r = .93)

Race, sex, and age	Number of persons examined (unweighted counts)*	Mean hemoglobin (after imputation)**	Imputation procedure					
			Ordered hot deck		Regression		Random hot deck	
			V_{T2}^2 / V_S^2	V_{T2}^2 / V_{T1}^2	V_{T2}^2 / V_S^2	V_{T2}^2 / V_{T1}^2	V_{T2}^2 / V_S^2	V_{T2}^2 / V_{T1}^2
Under 4 years								
White	2,023	12.1	1.30	.79	1.01	.85	1.01	1.52
Black	443	11.6	1.10	.95	1.01	.93	1.05	.99
4-7 years								
White males	871	12.6	1.22	.97	1.00	.97	1.04	.71
White females	821	12.6	1.07	.83	1.00	.93	1.16	.90
Black males	174	12.0	1.48	.94	1.00	.94	1.08	.85
Black females	210	12.1	1.01	1.08	1.00	.93	1.08	
8-14 years								
White males	903	13.5	1.00	.88	1.00	.89	1.04	1.08
White females	842	13.2	1.02	1.06	1.01	.96	1.04	.99
Black males	175	12.8	1.07	.95	1.00	.93	1.01	.98
Black females	178	12.4	1.05	1.00	1.00	1.00	1.04	1.11
15-44 years								
White males	2,843	15.3	1.00	.97	1.01	.93	1.01	1.15
White females	2,986	13.4	1.00	.95	1.00	.88	1.00	.68
Black males	408	14.4	1.01	.90	1.00	.98	1.01	1.28
Black females	469	12.6	1.02	1.14	1.00	1.08	1.03	.88
45+ years								
White males	2,748	15.1	1.00	.91	1.01	.91	1.01	.95
White females	3,068	13.7	1.01	.99	1.00	1.02	1.04	1.01
Black males	319	14.1	1.02	1.04	1.00	1.04	1.06	.85
Black females	387	12.9	1.04	.95	1.00	.93	1.01	1.10

*Excludes 454 other persons. V_{T2}^2 is estimated total variance based on independent imputation by half sample.
**Average of 4 iterations.