# Imputation Strategy for a Health and Nutrition Survey

Pippa Simpson[1], Jeffrey Gossett[1], ChanHee Jo[1], Margaret Bogle[2]

UAMS[1], USDA[2]

**Abstract**

The FOODS 2000 is a cross-sectional telephone survey that was conducted in the spring of 2000 in the Lower Mississippi Delta regions of Arkansas, Louisiana, and Mississippi to assess health and nutrition. Missing data is a common problem in survey data. The purpose of the survey was to investigate the relationships between health and nutrition factors adjusting for control variables such as income, race, weight, and nutritional assistance program participation. It is particularly troublesome to have missing control variables. If we use all available cases, the sample size varies considerably. We investigate single and multiple imputation (MI) strategies. The FOODS 2000 is a complex weighted sample. We discuss strategies for incorporating predetermined weights in the imputation. This work was funded under the Lower Mississippi Delta Nutrition Intervention Research Initiative, USDA grant # 6251-53000-003-00D.

**Introduction**

Missing data is an unpleasant but commonly occurring problem in many scientific investigations, and is ubiquitous in survey data. Talks and papers about missing data have escalated over the last few years. But missing in the main has been the discussion of how to deal with missing data when a weighted analysis is required. We investigate imputation in the context of a weighted nutrition survey. There is the situation where a variable or variables have some missing values. For example, income, race or body weight may be missing for various reasons. This is called item non-response. The extreme of item no response is unit non-response, where no variables have values recorded for a selected respondee. Often different strategies are used for these two forms of non-response. For unit non-response, the most common methods are to use donors, based on a hot deck approach or to adjust weights. For item non response, a hot/ cold deck approach, nearest neighbor or modeling approach can be taken. Much progress has been made in the development of statistical methods to handle missing data appropriately (Rubin 1976; Little and Rubin 1987; Laird 1988; Little 1992; Schafer 1997). Still we are long way from being able to handle the practical situations correctly and efficiently, particularly when we have a weighted sample.

The aim of this paper is to show that despite limitations in imputation methodology for surveys, analyses with imputed data results, yield more comprehensible than when missing data is ignored

**Background**

To impute or not to impute is the question. Although people feel uncomfortable with making overt assumptions that may not be totally or even partially verifiable, they often opt, without realizing, to make much more restrictive and worse assumptions. One must carefully balance throwing out data or generating replacement values. In many situations, researchers without realizing are throwing away data and, in doing so are obtaining biased results. The default of many programs is to only use responders for which there are no missing values. This is a complete or available case analysis. An annoying factor of this type of analysis is that unless all the cases with any missing values are deleted before any analyses we may have different numbers of subjects for each sub analyses. This is confusing and can give misleading results unless the pattern really is not related to any other variable. When we ignore all cases for which there may be missing values, we are assuming that the missing pattern happens completely at random and does not depend on any other values; that is the data are missing completely at random (MCAR). This is a very sweeping assumption since often the missing data pattern is dependent on other variables such as age. Thus, if the respondents respond differentially to questions, depending on values of some other variables, our estimates may be biased with a complete or available case analysis.

If the missing pattern is related to other variables, then we say that the data are missing at random (MAR). An example would be where income is likely to be missing for people with high education. When we impute we can take this into account and make a far less sweeping assumption. If we use single or MI to fill in the missing data, our estimates are biased if the model is incorrect (or inappropriate) or if the missing data mechanism is neither missing at random or missing completely at random. Unfortunately, we really have no way of testing the missing at random assumption required for imputation.

When we impute we must be careful to hypothesize a reasonable explanation for the missing pattern. We can do so by investigating relationships of the missing pattern via modeling. For example with the missing status of a variable as an outcome we can use logistic regression modeling to look at the variables that might be related to the pattern. We can use *a priori* knowledge to select reasonable values to fill in the "holes" in our dataset. We can check after imputation that our values give us margins that fit known criteria (Barnard and Meng (1999)). We can also make alternative reasonable assumptions and see how the results are affected, thereby doing a sensitivity analysis.

Weighting which is obtained from raking estimates to reflect population totals and account for total-non response is imputation in itself. The assumption behind

generation of the weights is a MAR pattern. If the generation of the weights does not take into account the missing pattern appropriately then using weights, may not give us a representative sample. Moreover, weights will not deal with missing items. Hence there will still be the need to have a strategy to deal with missing items. The model on which imputation is based should take account of design variables, variables related to the variable to be imputed and variables that affect the missing pattern. However, complete information about the design may not be available due to disclosure issues and/or limitations on the size of the dataset. Weights are effectively design variables which are often adjusted for non response. It could be argued that the model on which this adjustment is made should be consistent with any imputation. This may only, at best, be approximately possible, due to limited information. Nor may it be the best strategy given that often unit non response adjustment does not and/or cannot use all available information.

Single Imputation methods have the benefit of generating one complete data set. One could substitute a value for each missing value using means, medians, conditional means, or possibly a hot-deck approach with a random draw replacement. When done well the estimates are unbiased. Unfortunately, single imputation has drawbacks. It treats missing values as if known in the complete data analysis. This does not reflect uncertainty about predictions of the unknown missing values, and it will bias estimated variances of parameter estimates toward zero.

The MI Strategy (Rubin 1987) replaces each missing value with a set of plausible values that represent the uncertainty about the right value. It has the disadvantage of requiring multiple copies of the data set, which must be independently analyzed. It requires a model for the missing data mechanism. It is difficult to test whether an imputation model is correct. Selecting imputation model covariates is difficult. To do a proper imputation model, we should include as many variables (combinations of variables) as possible. According to (Van Buren, Boshuizen and Knook 1999), we should include: variables in our complete data model, variables associated with missingness of imputed variables, and variables correlated to imputed values. This could possibly result in hundreds of variables needed in a survey, not including variable interactions. Furthermore, it isn't clear whether variables should be transformed to another scale. Should we include external data sources such as median estimates of income in some geographical area?

There are arguments for and against MI. It is a more complex procedure and there will be no unique final dataset. The increase in variance of estimates due to the multiple draws can be seen as a negative if it does not reflect true variation. The pros are that it does reduce bias for estimating distributions, and the standard deviations are less biased. Each imputed data set is analyzed separately. The procedures for combining estimates are fairly straight forward.

When we impute we will usually want to impute all missing items and if the data is non monotone, or has no obvious pattern, some iterations will be necessary. Specifying the best model for imputation of all items may not be feasible.

Software for imputation for weighted survey data is limited. SAS has recently provided increased capabilities for dealing with missing data with its MI and MIANALYZE procedures. SOLAS, SPlus, SPSS and Stata have limited capabilities for imputation as well. In fact freeware, at present, deals better with weighted surveys. Schafer offers a model based MI approach (on which SAS has based their work). MICE and IVEWARE regression based procedures. IVEWARE is based on SAS, which is a plus for us. In addition IVEWARE allows for weights and specification of some but not all complex sample design features. IVEWARE was used for the MI of Family Income and Personal Earnings in the National health Interview Survey. Hence IVEWARE was used for imputation of the income in FOODS 2000. In the future, we plan to look at a hot-deck method incorporating weights. In the generation of imputed datasets complex design is still not properly taken into account.

## Methods

The survey - FOODS 2000. FOODS 2000, a cross-sectional telephone survey using list-assisted random-digit dialing, included respondents three years of age and older in 36 Delta counties. The lower Mississippi Delta region has historically high rates of poverty, poor education, limited access to health care, and high chronic disease burden. As is the case with most surveys, there is reluctance for some respondents to share sensitive personal information such as race, income, whether any member of the household is participating in nutritional assistance programs and so forth

Several Instruments were used in 3-phase interview. Weighting was used to handle unit and partial non-response. Person level sample weights range from 39 to 1876. For the dietary intake and Trailer questionnaires, the data were weighted to represent the population. Similarly, the 1662 adults who completed the FAOS questionnaire were weighted to represent the sample

population of adults. Although weighting was helpful for non response we still needed to deal with item non-response. Weights for each item would not be feasible. ,
Our item non response in no case exceeded 20% and we modeled the missing data simultaneously producing 5 imputed data sets. Decision about models for imputation is similar to modeling in general. It is important to decide which variables should be included in what form. We decided that we would not transform any before imputation, despite the fact that several nutrition variables and income tend to be skewed. For this paper we only included main effects in the model. Clearly, consideration of possible interactions is important. Calculated variables requiring special decisions were:

- Body mass index (BMI). This is calculated from self-reported heights and weights. We chose to impute the missing heights and weights. However, we could have imputed the BMI since this was ultimately the variable of interest. In fact BMI is used to categorize people as normal, overweight and obese and if that was only of interest, then it might have been better to impute categories.
- SF-12. The SF-12 is an instrument with 12 questions that purports to measure quality of life via the physical, mental, and total health scores. The scores are derived as linear combinations of recoded versions of the 12 questions. And the total is only of interest. We used a two step strategy that when questions were missing we imputed these and obtained a score, then for those who were missing all items we imputed a score. .

The model used was an explicit regression model, where we included variables expected to be correlated to either the missing value and/or missing data mechanism and design variables.

Imputation Algorithm:

1. Logical imputation of variables. Calculate number supported by income based on household enumeration.
2. Impute missing SF-12 questions based on results to other questions.
3. Main Imputation.
4. Repeat Steps 2 and 3 five times to obtain 5 replicates.
5. Calculate derived Variables.

After creating the 5 imputation data sets, we fit 6 models to each data set. The first two had obesity and hypertension (present or not) and were logistic models. The other four were SF-12 Mental and physical Health Scores, Protein and healthy eating index and were linear regression models. All models were fit using SUDAAN.

**Results**

For obesity, the complete case and MI datasets yielded similar results with race playing a more important role with the MI sets. This is understandable since it is known that there are a higher prevalence of obesity in African Americans, there were no significant differences for the hypertension model and for SF-12 Mental and Physical Health Score. For Protein consumption age and race made a significant difference in the MI analysis. In table 1, partial results are presented. It can be seen that consistent with expected results, Caucasians eat more protein than African Americans and there is an increase in protein consumption up to about 45 and then a decline. Finally for the HEI, with the MI analysis, it was found that males do not eat as well as females. In all models there were more significant results and these results were justifiable.

Table 1: Regression Model Parameter Estimates for Protein Consumption

| Model Parameter | MI Beta(SE) | Complete Case Beta (SE) | MI P-value | Complete Case P-value |
|---|---|---|---|---|
| Intercept | 50.078(7.193) | 63.011(8.267) | 0 | 2.17E-10 |
| **Race:White** | **4.335(2.107)** | **4.509(2.261)** | **0.007441** | **0.050695** |
| Sex:Male | 4.046(2.013) | 24.755(2.133) | 0 | 5.65E-17 |
| **Age:18to34** | **17.035(4.153)** | **8.789(4.771)** | **0.001883** | **0.070395** |
| **Age:35to44** | **19.267(4.417)** | **7.87(5.029)** | **0.005378** | **0.122849** |
| **Age:45to64** | **19.617(4.459)** | **5.852(5.008)** | **0.028357** | **0.247238** |
| Age:65to74 | 15.662(3.977) | 2.902(5.173) | 0.278986 | 0.576946 |

Conclusion

We found that the results were similar for Complete case (CC) and MI sets. Where there were differences the imputed values yielded more comprehensible results. We used a less than perfect approach for modeling the missing data. We could not model totally consistent with the weights. We used only main effects. Yet we obtained better results.

## References


Barnard J, Meng XL. Applications of multiple imputation in medical studies: from AIDS to NHANES. *Statistical Methods in Medical Research* 1999;8:17-36.

Laird NM. Missing data in longitudinal studies. *Statistics in Medicine* 1988;7:305-315

Little RJA. Regression with missing X's: a review. *Journal of the American Statistical Association* 1992;87(420):1227-1237.

Little RJA, Rubin DB. Statistical analysis with missing data. New York: Wiley, 1987.

Raghunathan TE, Solenberger PW, Van Hoewyk J. IVEware: imputation and variance estimation software. MI: Institute for Social Research, University of Michigan, 2002.

Rubin DB. Inference and missing data, *Biometrika* 1976;63(3):581-592.

Rubin DB. Multiple imputation for nonresponse in surveys. New York: Wiley, 1987.

Schafer JL. Analysis of incomplete multivariate data. London: Chapman and Hall, 1997.

Van Buuren S, Boshuizen HC, Knook DL. Multiple imputation of missing blood pressure covariates in survival analysis. *Statistics in Medicine* 1999; 18:681-694.