

Missing Data? A Look at Two Imputation Methods

Anita Rocha, Center for Studies in Demography and Ecology
University of Washington, Seattle, WA

ABSTRACT

Statistical analyses can be greatly hampered by missing data. It represents a loss of information, but worse, it can introduce bias in the results of the analysis when the data is not missing at random. In the end can we make confident assertions? One way to help rectify the problem of missing data is to employ a sound method of imputation, a way to replace missing values with reasonable estimates. There are advantages and disadvantages to imputation. Here we look at two approaches. First, we consider the hot-deck method using SAS® macros developed by Stiller and Dalzell (1998). Next, we explore the multiple imputation method using SAS PROC MI and PROC MIANALYZE. The examples herein were run on SAS 9.1 for Windows requiring Base SAS and SAS/STAT.

INTRODUCTION

If a test is marginally powered and the statistical model of interest ignores observations from the input data set if any outcome or covariate value is missing, then having as “complete” a data set as possible may be key to performing analyses. Imputation is the replacement of missing values in data with reasonable estimates. The biggest obstacle to the effective application of imputation methods is bias, when an estimator’s long-run average (expectation) differs from the quantity being estimated. In the case of survey data, the “danger comes from the possible difference between responders and non-responders. When this difference is systematic, the results of analyses may be biased and false conclusions are easily drawn.” (Huisman 2000) Ideally, an imputation method will be plausible and consistent, reduce bias while preserving the relationship between items within the data, and can be evaluated for bias and precision (Sande 1982). By way of simple illustrations, we will look at two popular methods of imputation in this paper, the hot-deck method and the multiple imputation method.

BACKGROUND

With the public’s relatively strong trust in government in the 1940s and 1950s, the U.S. Census Bureau could garner high response rates, even percentages in the high 90s, for its Current Population Surveys (CPS). Even with this high response rate, methods were devised to estimate missing values. The most well known technique for replacing missing values during this era was the hot deck method, a simple approach given that computers were expensive and unreliable (Scheuren 2005). Hot deck involves sampling from a group of “like” records to fill-in missing values.

Early on, assessments of precision for the imputed hot deck estimates were not routinely performed although the variance properties had been developed (Hansen et al. 1953). As for bias, the implicit assumption was that the data were missing completely at random (MCAR) or conditionally at random (MAR), and for the latter case the conditional variables were used to create the hot deck bins for the random sampling. However, another form of bias can be introduced when values are missing not at random (MNAR). In the case of MNAR it is difficult if not impossible to accurately account for bias in estimates, even employing modeling augmentation. Let us review further these mechanisms of bias first defined by Rubin in 1977.

Definitions related to bias (Carpenter 2002)

Missing Completely at Random	MCAR	The probability of an observation being missing does not depend on observed or unobserved measurements.
Missing at Random (sometimes termed Missing Conditionally at Random)	MAR	Given the observed data, the missingness mechanism does not depend on the unobserved data. This is equivalent to saying that the behaviour of two units who share observed values have the same statistical behaviour on the other observations, whether observed or not.
Missing Not at Random (in the likelihood setting the missingness mechanism is termed Non-ignorable)	MNAR	When neither MCAR nor MAR hold, we say the data are Missing Not At Random. What this means is 1) even accounting for all the available observed information, the reason for observations being missing still depends on the unseen observations themselves, and 2) to obtain valid inference, a joint model of both the data and the missingness mechanism is required. Unfortunately, 1) we cannot tell from the data at hand whether the missing observations are MCAR, NMAR or MAR, and 2) in the MNAR setting it is very rare to know the appropriate model for the missingness mechanism.

Hence the central role of sensitivity analysis; we must explore how our inferences vary under assumptions of MAR, MNAR, and under various models. Unfortunately, this is often easier said than done, especially under the time and budgetary constraints of many applied projects.

The distinction among these mechanisms, argues Donald Rubin, might be particularly important to the analysis of survey data because it is “rare that the missing values occur completely at random, as they might in some experimental contexts; in surveys, it is often reasonable to suspect that nonrespondents systematically differ from respondents” (Rubin 1987).

Although understanding MCAR, MAR and MNAR is crucial, it is beyond the scope of this paper to explore in depth the mechanisms of missing values. None of the methods presented here alone will address certain forms of bias in a sample. Rather, the early solution to reduce bias was to minimize the number of missing values in the first place by double sampling (Scheuren 2005). Why worry about possible bias introduced by missing value patterns when there are few missing values to begin with? In double sampling the respondents that didn’t complete a survey the first time were asked a second time for a response, perhaps with a different follow-up method than the original survey administration. For example, the first round of surveys could be mailed while the second round targeting non-responders could be administered by a telephone call. Double sampling and even greater measures to minimize missing data are still used today.

Despite these efforts, missingness in survey data remains a huge issue. With technological advancements in survey dissemination, today’s wary public is less likely to respond, partially or at all, after being inundated with surveys along with other competing demands on their attention.

In the 1970s, another popular method took root called multiple imputation (Rubin 1977). Multiple imputation represents a broad range of methodologies, all having in common the characteristic that the imputation method is performed multiple times. With greater computing capabilities, these more intensive methods became feasible and affordable. Both hot-deck and multiple imputation methods remain in use today so we’ll explore each in turn.

HOT-DECK

The hot-deck method begins by designating several characteristics to help choose a plausible estimate to replace a missing value. For example, in the portion of a sorted data set shown below, where each observation represents an individual, there is one individual, ID=089, with a missing value for INCOME. There are two other individuals that precisely match the individual with ID=089 on other characteristics such as SEX, AGE, MARITAL (marital status), and CHILDREN (number of children). Those are individuals with IDs=088 and 090.

ID	SEX	AGE	MARITAL	CHILDREN	INCOME		INCOME	FINCOME
...								
087	F	37	M	3	29.4		29.4	0
088	F	38	M	3	28.5		28.5	0
089	F	38	M	3	.	>	28.5	1
090	F	38	M	3	30.6		30.6	0
091	F	38	S	3	25.7		25.7	0
...								

The hot-deck algorithm makes a reasonable “guess” and will populate the missing INCOME value using one of the other two matching candidate’s values, either 28.5 or 30.6, by way of a random draw. In the example shown, the value for INCOME from ID=088 is randomly drawn replacing the missing INCOME value. An accompanying indicator FINCOME signifies the value 28.5 for ID=089 has been imputed. Subsequent analyses can now employ the imputed, more “complete” version of INCOME variable.

Hot-deck is easy to use and conceptually straightforward. However, one can imagine with this simple example why the hot-deck method tends to underestimate the true variance of parameters, at least in its simplest application. Note that the resultant data set now has, by design, two identical values for imputed INCOME. The additional duplicate INCOME value “pulls” the estimate of the sample mean of imputed INCOME towards 28.5, thus reducing the estimated variance contribution from this observation. In other words, with hot-deck you get more of the same in your resultant data set and so you get less estimated variability. It’s fair to assume that this may not produce the standard errors we’d expect had we begun with the complete, non-missing data.

HOT-DECK EXAMPLE USING SAS MACROS

John Stiller and Donald R. Dalzell of the US Census Bureau published a paper in the 1998 *Proceedings SUGI 23* that outlined a hot-deck implementation strategy, including efficient SAS macros for implementation. These macros, once understood, are elegant, easy to use and fast. I refer the interested reader to the Stiller and Dalzell paper itself, but an introduction to the method might jump-start comprehension for the uninitiated.

Here's a brief example of how to use the Stiller and Dalzell SAS macros to implement the hot-deck method for imputation. A researcher recently asked us to use the hot-deck method to impute missing household income values for mothers in the NLSY data (the National Longitudinal Surveys of Youth from the US Bureau of Labor Statistics). The following mother's characteristics (also called "input variables" by Stiller and Dalzell) were used to find individual matches: marital status, race, highest completed grade in school, and age.

Two of these characteristics were categorical variables, marital status and race, and two were considered continuous variables, highest completed grade and age. One way to differentiate between categorical and continuous variables when applying this hot-deck method is to think of the categorical variables as helping to distribute observations into bins. In this case, marital status has three categories (1=Never married, 2=Married spouse present, 3=Other) and race has three categories (1=Hispanic, 2=Black, 3=Non-Hispanic, Non-Black). So there are $3 \times 3 = 9$ bins that represent unique combinations of the marital status and race categories. Each observation, assuming there are no missing characteristic values, can be distributed into its corresponding bin.

Categorical input variables designate "bins" (part of the hot-deck matrix) Data is sorted within bins by continuous variables

1=Never Married 1=Hispanic	1=Never Married 2=Black	1=Never Married 3=Non-Hispanic, Non-Black
2=Married spouse present 1=Hispanic	2=Married spouse present 2=Black	2=Married spouse present 3=Non-Hispanic, Non-Black
3=Other 1=Hispanic	3=Other 2=Black	3=Other 3=Non-Hispanic, Non-Black

Next, within each bin, the observations are sorted by the remaining continuous variables, highest grade completed and age. The hot-deck method is then applied to each bin independently, replacing a missing INCOME value by randomly choosing an INCOME value that comes from observations that match on the continuous variables. However, there are unavoidable complications. If within a bin no precise characteristic match is found between an observation that has a missing household INCOME value and another candidate observation (donor), then the closest match possible is found, a "nearest neighbor" approach. To add to the complexity, no candidate INCOME value can be used as a substitute for missing INCOME more than once. So for each bin, up to four candidate values are held in what the authors' term a hot-deck matrix until each, in turn, is used to replace a missing value or discarded, as appropriate.

At this point, I leave it to the reader to review and implement the Stiller and Dalzell macros. Besides being useful, they're a nice primer to using macros and arrays as matrices. Later, we'll revisit the results for this example when we compare them to our next method, multiple imputation.

MULTIPLE IMPUTATION

Another way to substitute plausible data for missing values is a method called multiple imputation. Multiple imputation, as its name implies, repeats the imputation over and over again, usually at least three times. The multiple imputation method, unlike the hot-deck, tends not to underestimate variance because it replaces missing values by drawing randomly from its sampling distribution, likely better quantifying the uncertainty.

From the SAS OnLineDoc 9.1.2, there are three distinct steps to performing a multiple imputation.

1. The missing data are filled in m times to generate m complete data sets.
2. The m complete data sets are analyzed using standard statistical analyses.
3. The results from the m complete [analyses] are combined to produce inferential results.

For step 1, any reasonable imputation method could be used as long as there is a random component to the algorithm, thus producing different versions of the imputed data sets. Note that hot-deck imputation would qualify and could be the method employed in Step 1 to create multiple "complete" data sets. Moreover, the process for

combining results, Step 3, is essentially the same no matter what imputation method or statistical analyses are performed for Steps 1 and 2. So multiple imputation actually refers to a broad range of methods having the common attributes that there are multiple versions of the imputed data sets, each data set is analyzed the same way, and results are combined into a single version of estimates. Now we look at how SAS in particular implements multiple imputation.

STEP 1: MISSING DATA ARE IMPUTED TO GENERATE MULTIPLE COMPLETE DATA SETS

There are numerous imputation methods available in SAS. From the SAS OnLineDoc 9.1.2, the following table outlines options for PROC MI.

Pattern of Missingness	Type of Imputed Variable	Recommended Methods	Notes
Monotone	Continuous	Regression	Parametric: assumes multivariate normality
		Predicted Mean Matching	Parametric: assumes multivariate normality
		Propensity Score	
Monotone	Classification (Ordinal)	Logistic Regression	Binary or ordinal response
Monotone	Classification (Nominal)	Discriminant Function Method	Binary or nominal response
Arbitrary	Continuous	MCMC Full-Data Imputation	Impute all missing values
		MCMC Monotone-Data Imputation	Impute enough missing values to achieve monotone missing patterns – then use second method

Monotone missingness, a criteria that if met provides greater flexibility in selecting an imputation method, requires that the first variable is at least as observed as the second, which is at least as observed as the third, and so on. The following example from Rubin (1987 p171) is helpful.

Monotone pattern of missingness (1 = observed, 0 = missing)

	Variables						
	Y1	Y2	Y3	Y4	Y5	Y6	Y7
Observations	1	1	1	1	1	1	1
	1	1	1	1	1	1	0
	1	1	1	1	1	0	0
	1	1	1	0	0	0	0
	1	1	0	0	0	0	0

Here a '1' means that the value for this variable has been observed and recorded, while a '0' means that the value is missing. Notice the "nesting" quality of the data. Suppose the values in the table above represent responses to survey questions. Then every respondent answered the questions corresponding to Y1 and Y2, while a subset of those answered Y3, a subset of those answered Y4 and Y5, a subset of those answered Y6, and finally a subset of those answered Y7. Hence the data is said to have monotone missingness.

MULTIPLE IMPUTATION EXAMPLE USING PROC MI

In the sample data from the NLSY as in the previous example, we can safely assume monotone missingness since for simplicity it has been pre-selected to have missing values only in the INCOME variable. But this does not preclude the use of PROC MI when there is more than one variable with missing values. INCOME is a continuous outcome, so consider the following use of PROC MI with variables MARC (marital status), RACEM (race), AGEM (age), HGCR (last grade completed), and HHINC (household income).

```
proc mi data=total out=totali nimpute=3 simple;
  var marc racem agem hgcr hhinc;
run;
```

The option NIMPUTE specifies the number of imputed "complete" data sets requested. The more missing values that are to be imputed, the larger number of complete data sets should be requested. Typically 3 is the minimum number and often 5 is a common selection. Also, the option SIMPLE instructs SAS to display univariate and correlation statistics. Here, since no method was specified, the default method is an EM algorithm to find maximum

likelihood estimates for a multivariate normal distribution. Random draws are taken from this distribution to populate the missing values. Recall that in our case only HHINC has missing values as the values for other variables are complete.

PROC MI generates an output data set that includes a variable called `_IMPUTATION_` that distinguishes the imputation data sets from each other. So if the original data set contained 100 observations and `NIMPUTE=3`, the output data set from PROC MI will contain 300 observations with `_IMPUTATION_` equal to 1, 2, or 3, respectively, for each data set.

Next, we look at a slightly more complicated implementation of this procedure.

It is clear from descriptive statistics and a histogram that the income variable HHINC is not normally distributed but is skewed with a sample median that is clearly less than the sample mean. The other variables violate this assumption to a far lesser degree. To mitigate the effects on estimation from the non-normality of HHINC, a log transformation can be used. The TRANSFORM statement specifies the variables to be transformed, in this case HHINC, before the imputation process and then back-transformed to create imputed values. The option `C=1` below is chosen as an offset so that for incomes equal to zero (yes we have some zero incomes!) the logarithm can be computed as $\log(\text{HHINC} + C)$.

One issue with the above implementation of PROC MI is that the categorical variables MARC and RACEM are treated as continuous. In SAS 9.1, the use of a CLASS statement is experimental and accounts for categorical variables.

We can, for example, specify an imputation method. HHINC is continuous and the missing value pattern satisfies the monotone assumption, so an ordinary least squares regression model is specified in the MONOTONE statement with HHINC as the dependent variable and the other characteristics as independent variables. Consider the following modifications.

```
proc mi data=total out=totali nimpute=3 simple;
  class marc racem;
  var marc racem agem hgcr hhinc;
  monotone reg(hhinc=marc racem agem hgcr);
  transform log(hhinc / c=1);
run;
```

Now that we are satisfied with this imputation method, let's begin the analysis phase.

STEP 2: USE STANDARD STATISTICAL ANALYSES WITH IMPUTED DATA SETS

First we sort by `_IMPUTATION_` since it will be used in a BY statement in our modeling procedure. Then we fit the model. In this example, we used a PROC MIXED, but again, nearly any reasonable modeling procedure can be accommodated by multiple imputation.

```
proc sort data=totali;
  by _imputation_;
run;

proc mixed data=totali;
  class marc racem;
  model hhinc = agem hgcr marc racem / intercept solution;
  by _imputation_;
  ods output SolutionF=mxparms;
run;
```

Note that we request a data set MXPARMs in the ODS output statement containing the parameter estimates and their standard errors. These data set will be used in phase three to combine estimates. Note that different analysis procedures will require different instructions to obtain the output data sets that will be useful in the Step 3. See the SAS OnLine Doc for a really nice selection of representative examples.

STEP 3: COMBINE RESULTS

The final step is to combine the results from all the imputed data sets, yielding valid statistical inferences that account for the uncertainty of missing values. The ideas behind combining estimates are simple. From the SAS OnLine Doc 9.1, we see that the combined point estimate is just an average across the estimates resulting from the analysis from step 2. Suppose we have computed m different sets of the point and variance estimates for a parameter Q . Suppose that $\hat{Q}_i$ and $\hat{U}_i$ are the point and variance estimates from the i^{th} imputed data set, $i=1, 2, \dots, m$. Then $\bar{Q}$, the point estimate from multiple imputation, is as follows.

$$\bar{Q} = \frac{1}{m} \sum_{i=1}^m \hat{Q}_i$$

Now suppose that $\bar{W}$ is the within-imputation variance and that B is the between-imputation variance. These can be estimated as follows.

$$\bar{W} = \frac{1}{m} \sum_{i=1}^m \hat{W}_i$$

$$B = \frac{1}{m-1} \sum_{i=1}^m (\hat{Q}_i - \bar{Q})^2$$

Finally, the variance estimate associated with Q is the total variance.

$$T = \bar{W} + \left(1 + \frac{1}{m}\right) B$$

Note that the between-imputation variance adds to the total variance, thus provided an accounting for the variance involved in estimating missing values. This alone improves upon the hot-deck method for imputation.

The procedure MIANALYZE will compute these estimates easily.

```
proc mianalyze parms=mxparms;
  class marc racem;
  modeleffects Intercept agem hgcr marc racem;
run;
```

The PARMS= option identifies the input data set containing the estimated model parameters. In this case, MXPARMs is specified as the data set name, just as it was created in step 2. The MODELEFFECTS statement names the model parameters as shown in the MXPARMs data set. Finally, the CLASS statement identifies our categorical variables as marriage (MARC) and race (RACEM).

COMPARING RESULTS FOR THE NLSY EXAMPLE

Here are the results for the NLSY data analyses, comparing three different methods: 1) the original data without imputation leaving the missing values "as is" (ORIG), 2) hot-deck method of imputation (HD), and 3) multiple imputation method employing regression (MI). In an ordinary least squares regression, the dependent variable is INCOME and the covariates are listed below.

Parameter	Original Est	Orig StdErr	HD Est	HD StdErr	MI Est	MI StdErr
N	3842		4873		4873	
Intercept	21.2	2.4	20.3	2.1	28.9	4.2
AGEM	-1.2	0.1	-1.0	0.1	-1.5	0.2
HGCR	1.4	0.1	1.2	0.1	1.4	0.2
MARC=1	4.3	1.1	4.7	0.9	3.5	1.4
MARC=2	2.6	1.1	3.0	0.9	3.6	1.8

RACE=1	-3.2	0.6	-3.4	0.5	-3.0	1.0
RACE=2	-6.7	0.5	-6.8	0.4	-6.6	0.7

(Some might argue this analysis isn't plausible considering that the same covariates listed were those used to categorize and otherwise model the imputed income values, and yet this type of approach has been done by numerous researchers. I think this is a non-trivial objection and will consider it further.)

Although there are shifts in the point estimates in the table above, most telling are the trends in the standard errors. Note that the standard errors are consistently lower for the original estimates compared to the estimates from the multiple imputation method. This reflects, at least in part, the additional estimated between-imputation variance that was added to the total variance. Also note the frequent lower standard error estimates for the hot-deck method compared with the original, again suggesting that adding "more of the same" may be contributing to the dampening of the hot-deck variance estimates.

CONCLUSION

By no means have we exhausted the topic of imputation in this paper. There are many articles, for example, devoted to various randomization methods used to impute missing data due to loss to follow-up in the longitudinal data collection setting. The references below may offer resources to those that want a broad introduction.

Although the mechanisms of bias were introduced, there was little said about how to think specifically about bias when applying an imputation method. Yet thoughtful consideration of bias is pivotal to implementing a reasonable method of imputation. It may be decided, after careful deliberation, that imputation is inappropriate for a particular analysis.

Another important step not discussed here is the conduct of sensitivity analyses. Typically, this requires that the analyst look critically at her assumptions. It may involve varying the imputation method by developing a range of possible choices for implementation. Then the analyst assesses how robust her estimates are against the range of methods employed. This takes time, but adds confidence, and is often cited as a crucial component to successful analysis in the presence of missing values.

REFERENCES

- Carpenter, J, and Kenward, M (2002) website: Missing Data <http://www.lshtm.ac.uk/msu/missingdata/start.html>, sponsored by the Research Development Initiative (RDI) and the Economic and Social Research Council (ESRC).
- Hansen, M.H., Hurwitz, W.N., and Madow, W. (1953) *Sample Survey Methods and Theory*, New York: Wiley.
- Huisman, Mark (2000) Imputation of Missing Item Responses: Some Simple Techniques, *Quality & Quantity* 34: 331-351.
- Rubin, D.B. (1977) Inference and Missing Data, *Biometrika* 63: 581-592.
- Rubin, D.B. (1987) *Multiple Imputation for Nonresponse in Surveys*, New York: Wiley.
- Sande, I.G. (1982) Imputation in surveys: Coping with reality. *The American Statistician* 36: 145-152.
- SAS Institute, Inc. website: OnLineDoc 9.1.2 <http://support.sas.com/onlinedoc/912/docMainpage.jsp>, Copyright 2002-2004, Cary, NC.
- Scheuren, Fritz (2005) Multiple Imputation: How It Began and Continues. *The American Statistician* 59: 315-319.
- Stiller, John and Dalzell, Donald R. (1998) Hot-deck Imputation with SAS® Arrays and Macros for Large Surveys, *Proceedings SUGI 23*, SAS Institute, Inc (<http://www2.sas.com/proceedings/sugi23/Stats/p246.pdf>).

ACKNOWLEDGMENTS

The author would like to acknowledge support from the Center for Studies in Demography and Ecology, University of Washington, Box 353412, Seattle, WA 98195-3412 (<http://csde.washington.edu>).

A good portion the author's work with the hot-deck method was sponsored by the grant "Give and take: Child agency in resource allocation", grant number NICHD R01HD045635-01, and principal investigator Jennifer Romich.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Anita Rocha
Center for Studies in Demography and Ecology
University of Washington
Box 353412
Seattle, WA 98195-3412
Phone: 206.616.6687
E-mail: alrocha@u.washington.edu
Web: <http://csde.washington.edu/>

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Other brand and product names are trademarks of their respective companies.