PhUSE US Connect 2018

Paper DH04

Missing Data Imputation and Its Effect on the Accuracy of Classification

Edwin Ponraj Thangarajan, PRA Health Sciences, Chennai, India
Giri Balasubramanian, PRA Health Sciences, Chennai, India

ABSTRACT

Multivariate data sets frequently have missing observations scattered throughout the data set. Many machine learning algorithms assume that there is no significance in the fact that an observation has an attribute value missing. A common approach in coping with these missing values is to replace the missing value using some plausible value, and the resulting completed data set is analyzed using standard methods. Programs evaluate the effect that some commonly used imputation methods have on the accuracy of classifiers in supervised leaning. The effect is assessed in simulations performed on several classical datasets where observations have been made missing at random in different proportions. This paper will describe the below,

- Missing Data Mechanisms
- Strategies of Handling Missing Data
- Method Used to Deal with Missing Values

INTRODUCTION

Missing data is a problem because nearly all standard statistical methods presume complete information for all the variables included in the analysis. Many of the multivariate data sets collected today would have unobserved or missing observations scattered throughout the data set. These missing values can have no particular pattern of occurrence. Despite the frequent occurrence of missing data, many machine learning algorithms assume that there is no particular significance in the fact that a particular observation has an attribute value missing: the value is simply unknown, and the missing value is handled in a simple way.

Most statistical procedures usually eliminate entire cases whenever they encounter missing data in any variable included in the analysis. For example, a regression analysis to predict home ownership based on age and educational background would ignore all cases where either of these variables had a missing response. So, although each individual variable may only have a small percent of missing data, when examined in combination, the total number of cases in the analysis is reduced drastically.

Case	Age	Gender	Home	Education	Occupation
1	.	Female	No	16	Non-professional
2	22	Male	No	.	Non-professional
3	39	Male	.	20	Professional
4	.	Female	Yes	.	Professional
5	40	.	Yes	16	Non-professional
6	22	Female	No	16	.
7	35	Male	Yes	18	Professional
8	39	Male	Yes	20	Professional

In this data set, data are missing for some respondents as shown by the ".". Any case with either missing data in age, home or education is ignored in the analysis. So, only respondents 5-8 would be included in your analysis which would drastically reduce the amount of data to be analyzed and increase the risk of a costly error.

Missing data can also lead to misleading results by introducing bias. Whenever segments of your target population do not respond, they become underrepresented in your data. In this situation, you end up not analyzing what you intended to measure. For example, suppose you surveyed a group of customers, but many people refused to answer the question about their age. If you calculate the average age based on the data you have, you would conclude that the average age of your customers is 39. However, some segments of customers may be under represented, so this conclusion could be incorrect. If every customer reported their age, you might get different results. For example, if

those who did not respond are younger, the actual average age of your customer base is 29. For this question, the "youth" segment of your customers is "under represented" and your conclusion would have been incorrect.

Case	Age	Gender
1	.	Female
2	.	Male
3	39	Male
4	.	Female
5	42	.
6	.	Female
7	37	Male
8	39	Male

Missing data can impact your results. Here, the average age is 39 when respondents with missing data are ignored.

Case	Age	Gender
1	21	Female
2	22	Male
3	39	Male
4	20	Female
5	42	Male
6	18	Female
7	37	Male
8	39	Male

With complete data, the mean age is 29 -a difference in generation. Missing data can seriously impact the conclusions you draw from your data.

Example of misrepresentation of statistical data:

- A statistical report: "The number of accidents taking place in the middle of the road is much less than the number of accidents taking place on its side. Hence it is safer to walk in the middle of the road." This conclusion is obviously wrong since we are not given the proportion of the number of accidents to the number of persons walking in the two cases.
- "The number of students taking up Mathematics Honors in a University has increased 5 times during the last 3 years. Thus, Mathematics is gaining popularity among the students of the university." Again, the conclusion is faulty since we are not given any such details about the other subjects and hence comparative study is not possible.
- "75% of the people who drink alcohol die before attaining the age of 80 years. Hence drinking is harmful for longevity of life." This statement, too, is incorrect since nothing is mentioned about number of persons who do not drink alcohol and die before attaining the age of 80 years.

Thus, statistical arguments based on incomplete data often lead to fallacious conclusions.

"INCOMPLETE DATA USUALLY LEADS US TO FALLACIOUS CONCLUSIONS."

With many classification algorithms, a common approach that is used is to replace the missing values in the data set with some plausible value and the resulting completed data set is analyzed using standard algorithms. The procedure that replaces the missing values using some value is known as imputation. For a layman, this idea conjures an image of a statistician making up the data. This is true only if one were to analyze the data as if the imputed data are real values.

The purpose of this paper is to clearly present the essential concepts and methods of imputation have on the accuracy of classification when classifying data to successfully deal with missing data using SAS.

MISSING DATA MECHANISMS

Varied solution and imputation methods were developed for the missing data problem. It is needed to identify the missing data mechanism before such solution and imputation methods. Because the decision which solution and imputation method would be applied to the data set depends upon the missing data mechanisms. Little and Rubin classified those mechanisms under three basic categories.

MCAR	Missing completely at random (MCAR). If missing observations is independent of both observed and unobserved values
MAR	Missing at random (MAR). If it is independent of unobserved values and dependent of observed values
NMAR	Not missing at random (NMAR). If it depends on both observed and unobserved values

Knowledge of the mechanism that led to the values being missing is important in choosing an appropriate analysis to use for the data. Hence it is important to consider how the classifier handles the missing data to avoid bias being introduced into the knowledge induced from that classifier.

STRATEGIES OF HANDLING MISSING DATA

There are several basic strategies that can be used to deal with missing data in classification studies. Some of these methods were developed in the context of sample surveys and can have some disadvantages in classification.

COMPLETE CASE ANALYSIS

Complete case analysis (also known as elimination) is an approach in which observations that have any missing attributes are deleted from the data set. This strategy may be satisfactory with lesser amounts of missing data. However, with enormous amounts of data, it is possible to lose considerable sample size. The critical concern with this strategy is that it can lead to biased estimates as it requires the assumption that the complete cases are a random subsample of the original observations. The completely recorded cases frequently differ from the original sample.

AVAILABLE CASE ANALYSIS

Available case analysis is another approach that can be used. As this procedure uses all observations that have values for a particular attribute, there is no loss of information as all cases are used. However, the sample base changes from attribute to attribute depending on the pattern of missing data, and hence any statistics calculated can be based on different numbers of observations. The main disadvantage to this approach is that the procedure can lead to covariance and correlation matrices that are not positive definite.

WEIGHTING PROCEDURES

Weighting Procedures are another approach to dealing with missing data. This approach is frequently used in the analysis of survey data. In survey data, the sampled units are weighted by their design weight which is inversely proportional to the probability of selection. Weighting procedures for non-response modify the weights in an attempt to adjust for non-response as if it were part of the sample design.

IMPUTATION PROCEDURES

Imputation procedures in which the missing data values are replaced with some value is another commonly used strategy for dealing with missing value. These procedures result in a hypothetical 'complete' data set that will cause no problems with the analysis. Many machine learning algorithms are designed to use either complete case analysis or an imputation procedure.

Imputation methods often involve replacing the missing values with estimated values based on information that is in the data set. Many of the imputation methods are restricted to coping with one type of variable (i.e. either categorical or continuous) and make assumptions about the distribution of the data or subsets of variables. The performance of classifiers with imputed data in unreliable, and it is hard to distinguish situations in which the methods work from those in which they fail. When imputation is used, it is easy to forget that the data is incomplete. However, imputation methods are commonly used in classification algorithms. There are many options that are available for imputation.

Imputation using a model-based approach is another popular strategy for handling missing data. A predictive model is created to estimate the values to be imputed for the missing values. With regression imputation, the attribute with missing data is used as the response attribute, and the remaining attributes are used as input for the predictive model. Maximum likelihood estimation using the EM algorithm is one of the recommended missing data techniques in the methodological literature. This method assumes that the underlying model for the observed data is Gaussian.

Rather than imputing a single value for each missing data value, multiple imputation procedures are also commonly used. With this method, the missing values are imputed with values drawn randomly (with replacement) from a fitted distribution for that attribute. This is repeated a number, N, of times. The classifier is applied to each of the N "complete" data sets and the misclassification error is calculated. The misclassification error rates are averaged to provide a single misclassification error estimate and also estimate variances of the error rate. Iterative regression imputation is not restricted to data having a multivariate normal distribution and can cope with mixed data. For the estimation, regression methods are usually applied in an iterative manner where each iteration uses one variable as an outcome and the remaining variables as predictors. If the outcome has any missing values, the predicted values from the regression are imputed. Iterations end when all variables in the data frame have served as an outcome.

METHOD USED TO DEAL WITH MISSING VALUES

The imputations should condition on observed variables, be multivariate to preserve associations between missing variables and generally be draws rather than means.

EXAMINE PATTERNS OF MISSING DATA

Missing data can be informative. Sometimes missing values in one variable are related to missing values in another variable. Other times missing values in one variable are independent of missing values in other variables. As part of the exploratory phase of data analysis, the analyzer should investigate whether there are patterns in the missing data.

To find the number of missing values for numeric variables PROC MEANS procedure can be used in SAS. PROC MEANS creates a compact easy-to-read table that summarizes the number of missing values for each numerical

variable. The following statements use the N and NMISS options in the PROC MEANS statement to count the number of missing values in eight numerical variables in the SASHELP.HEART data set:

```
/* count missing values for numeric variables */
proc means data=Sashelp.Heart nolabels N NMISS;
       var AgeAtStart Diastolic Systolic Height Weight MRW Smoking Cholesterol;
run;
```

The MEANS Procedure

Variable	N	N Miss
AgeAtStart	5209	0
Diastolic	5209	0
Systolic	5209	0
Height	5203	6
Weight	5203	6
MRW	5203	6
Smoking	5173	36
Cholesterol	5057	152

The NMISS column in the table shows the number of missing values for each variable. There are 5209 observations in the data set. Three variables have zero missing values, and another three have six missing values. The Smoking variable has 36 missing values whereas the Cholesterol variable has 152 missing values.

These univariate counts are helpful, but they do not tell you whether missing values for different variables are related. For example, there are six missing values for the Height, Weight, and MRW variables. How many patients contributed to those six-missing value? Ten? Twelve? Perhaps there are only six patients, each with missing values for all three variables? If a patient has a missing Height, does that imply that Weight or MRW is missing also? To answer these questions, you must look at the pattern of missing values.

The MI procedure in SAS is used for multiple imputation of missing values. PROC MI has an option to produce a table that summarizes the patterns of missing values among the observations. The following call to PROC MI uses the NIMPUTE=0 option to create the "Missing Data Patterns" table for the specified variables.

```
ods select MissPattern;
proc mi data=Sashelp.Heart nimpute=0;
       var AgeAtStart Height Weight Diastolic Systolic MRW Smoking Cholesterol;
run;
```

The MI Procedure

Missing Data Patterns

											Group Means							
Group	AgeAtStart	Height	Weight	Diastolic	Systolic	MRW	Smoking	Cholesterol	Freq	Percent	AgeAtStart	Height	Weight	Diastolic	Systolic	MRW	Smoking	Cholesterol
1	X	X	X	X	X	X	X	X	5039	96.74	44.073030	64.826206	153.060131	85.411193	136.987498	119.864457	9.370907	227.410796
2	X	X	X	X	X	X	X	.	124	2.38	43.653226	64.137097	154.193548	83.459677	134.040323	124.387097	8.685484	.
3	X	X	X	X	X	X	.	X	8	0.15	42.875000	65.312500	158.125000	83.500000	131.625000	121.750000	.	210.125000
4	X	X	X	X	X	X	.	.	28	0.54	45.214286	65.482143	151.392857	87.714286	140.571429	115.928571	.	.
5	X	X	.	X	X	.	X	X	4	0.08	40.000000	63.687500	.	73.500000	126.500000	.	27.500000	256.500000
6	X	.	X	X	X	X	X	X	4	0.08	55.000000	.	154.000000	81.250000	129.000000	124.500000	10.250000	254.250000
7	X	.	.	X	X	.	X	X	2	0.04	34.000000	.	.	77.000000	125.000000	.	2.500000	201.500000

The table reports the number of observations that have a common pattern of missing values. In addition to counts, the table reports mean values for each group. Because the table is very wide, I have truncated some of the mean values. The first row counts the number of complete observation. There are 5039 observations that have no missing values. Subsequent rows report the number of missing values for variables, pairs of variables, triplets of variables, and so forth. The variables are analyzed from right to left. Thus, the second row shows that 124 observations have missing values for only the rightmost variable, which is Cholesterol. The third row shows that eight observations have missing values for only the Smoking variable. The fourth row shows that 28 observations have missing values for both Smoking and Cholesterol. The table continues by analyzing the remaining variables that have missing values, which are Height, Weight, and MRW. You can see that MRW is never missing by itself, but that it is always missing when Weight is missing. Height is missing by itself in four cases and is missing simultaneously with Weight in two cases. Notice that each row of the table represents a disjoint set of observations. Consequently, we can easily answer the previous questions about how many patients contribute to the missing values in Height and Weight. There are 10 patients: four are missing only the Height measurement, four are missing only the Weight measurement (which forces MRW to be missing), and two are missing both Height and Weight. This preliminary analysis has provided valuable information about the distribution of missing data in the SASHELP.HEART data set. Most patients (96.7%) have complete data. The most likely measurement to be missing is Cholesterol, followed by information about whether the patient smokes. You can see exactly how many patients are missing Height measurements, Weight measurements, or both. It is obvious that the MRW variable has a missing value if and only if the Weight variable is missing.

The "Missing Data Patterns" table from PROC MI provides a useful summary of missing values for each combination of variables. Examining patterns of missing values can lead to insight into the data collection process and is also the first step prior to modeling missing data by using multiple imputation.

Assumptions and patterns of missingness are used to determine which methods can be used to deal with missing data.

MEAN AND MEDIAN IMPUTATION

Imputation of the missing value by either the mean, median or mode for the attribute are commonly used imputations. These types of imputation ignore any relationships between the variables. For mean imputation, it is well known that this method of imputation will underestimate the variance covariance matrices for that data. Both mean and median imputation can only be used on continuous attributes. For categorical data, the mode is often imputed whilst using either mean or median imputation.

The easiest way to perform mean imputation in SAS is to use PROC STDIZE. The following statements are available in the STDIZE procedure.

PROC STDIZE	<options>; The PROC STDIZE statement invokes the STDIZE procedure.
BY	variables; You can specify a BY statement with PROC STDIZE to obtain separate analyses of observations in groups that are defined by the BY variables. When a BY statement appears, the procedure expects the input data set to be sorted in order of the BY variables. If you specify more than one BY statement, only the last one specified is used.
FREQ	variable; If one variable in the input data set represents the frequency of occurrence for other values in the observation, specify the variable name in a FREQ statement. PROC STDIZE treats the data set as if each observation appeared n times, where n is the value of the FREQ variable for the observation. Nonintegral values of the FREQ variable are truncated to the largest integer less than the FREQ value. If the FREQ variable has a value that is less than 1 or is missing, the observation is not used in the analysis.
LOCATION	variables; The LOCATION statement specifies a list of numeric variables that contain location measures in the input data set specified by the METHOD=IN option
SCALE	variables; The SCALE statement specifies the list of numeric variables that contain scale measures in the input data set specified by the METHOD=IN option.
VAR	variables; The VAR statement lists numeric variables to be standardized. If you omit the VAR statement, all numeric variables not listed in the BY, FREQ, and WEIGHT statements are used.
WEIGHT	variable; The WEIGHT statement specifies a numeric variable in the input data set with values that are used to weight each observation. Only one variable can be specified.

PROC STDIZE supports the REPONLY and the METHOD=MEAN options, which tells it to replace missing values with the mean for the variables on the VAR statement. To demonstrate mean imputation, the following statements randomly add missing values to the SASHELP.CLASS data set. The call to PROC STDIZE then replaces the missing values and creates a data set called IMPUTED that contains the results.

```
/* Create "original data" by randomly inserting missing values for some heights */
data Have;
       set sashelp.class;
       call streaminit(12345);
       Replaced = rand("Bernoulli", 0.4);/* indicator variable is 1 about 40% of time */
       if Replaced then Height = .;
run;

/* Mean imputation: Use PROC STDIZE to replace missing values with mean */
proc stdize data=Have out=Imputed
      oprefix=Orig_          /* prefix for original variables */
      reponly                /* only replace; do not standardize */
      method=MEAN;           /* or MEDIAN, MINIMUM, MIDRANGE, etc. */
   var Height;               /* you can list multiple variables to impute */
run;

proc print data=Imputed;
   format Orig_Height Height BESTD8.1;
   var Name Orig_Height Height Weight Replaced;
run;
```

Obs	Name	Orig_Height	Height	Weight	Replaced
1	Alfred	69	69	112.5	0
2	Alice	56.500	56.500	84.0	0
3	Barbara	65.300	65.300	98.0	0
4	Carol	62.800	62.800	102.5	0
5	Henry	63.500	63.500	102.5	0
6	James	.	61.500	83.0	1
7	Jane	59.800	59.800	84.5	0
8	Janet	.	61.500	112.5	1
9	Jeffrey	62.500	62.500	84.0	0
10	John	59	59	99.5	0
11	Joyce	51.300	51.300	50.5	0
12	Judy	64.300	64.300	90.0	0
13	Louise	.	61.500	77.0	1
14	Mary	66.500	66.500	112.0	0
15	Philip	.	61.500	150.0	1
16	Robert	.	61.500	128.0	1
17	Ronald	.	61.500	133.0	1
18	Thomas	57.500	57.500	85.0	0
19	William	.	61.500	112.0	1

The output shows that the missing data (such as observations 6 and 8) are replaced by 61.5, which is the mean value of the observed heights. For a subsequent visualization, a binary variable (Replaced) that indicates whether an observation was originally missing. The METHOD= option in PROC STDIZE supports several statistics. You can use METHOD=MEDIAN to replace missing values by the median, METHOD=MINIMUM to replace by the minimum value, and so forth.

Mean imputation, which is easy to implement, enables analysts to use every observation. However, mean imputation has three serious disadvantages that can lead to problems in your statistical analysis. Mean imputation is a univariate method that ignores the relationships between variables and makes no effort to represent the inherent variability in the data. In particular, when you replace missing data by a mean, you commit three statistical sins:

- Mean imputation reduces the variance of the imputed variables.
- Mean imputation shrinks standard errors, which invalidates most hypothesis tests and the calculation of confidence interval.
- Mean imputation does not preserve relationships between variables such as correlations.

HOT DECK IMPUTATION

Hot deck imputation is another imputation method that is commonly used, especially in survey samples, and it can cope with both continuous and categorical attributes. Hot deck imputation involves replacing the missing values using values from one or more similar instances that are in the same classification group. There are various forms of hot deck imputation commonly used. Random hot deck imputation involves replacing the missing value with a randomly selected value from the pool of potential donor values. Other methods known as deterministic hot deck imputation involve replacing the missing values with those from a single donor, often the nearest neighbor that is determined using some distance measure. Hot deck imputation has an advantage in that it does not rely on model fitting for the missing value that is to be imputed and thus is potentially less sensitive to model misspecification than an imputation method based on a parametric model.

The SURVEYIMPUTE SAS procedure provides several imputation methods for replacing missing values in an item by observed values from the same item. These imputation methods can be used to impute missing values for survey data.

PROC SURVEYIMPUTE	<options>; The PROC SURVEYIMPUTE statement invokes the SURVEYIMPUTE procedure. The DATA= option identifies the data set to be analyzed.
BY	variables; You can specify a BY statement with PROC SURVEYIMPUTE to obtain separate analyses of observations in groups that are defined by the BY variables. When a BY statement appears, the procedure expects the input data set to be sorted in order of the BY variables. If you specify more than one BY statement, only the last one specified is used.
CELLS	variables; The CELLS statement names the variables that identify the imputation cells. The imputation cells divide the data into groups of similar units. The combination of levels of CELLS variables defines the imputation cells. If you do not use this statement, then all observation units are assumed to be in one imputation cell.

CLASS	variable < (options) > . . . < variable < (options) > > < / options >; The CLASS statement names the classification variables for the analysis.
CLUSTER	variables; The CLUSTER statement names variables that identify the first-stage clusters in a clustered sample design. First-stage clusters are also known as primary sampling units (PSUs). The combinations of categories of CLUSTER variables define the clusters in the sample. If you also use the STRATA statement, clusters are nested within strata.
ID	variable; The ID statement names a variable in the DATA= input data set to identify observation units.
IMPJOINT	<variables>; The IMPJOINT statement specifies the names of variables that are to be imputed jointly for the fully efficient fractional imputation (FEFI) method. If you do not use the IMPJOINT statement, then all the variables that you specify in the VAR statement are imputed jointly. You can use multiple IMPJOINT statements. The levels of the variables in the IMPJOINT statement describe a nonparametric imputation model for the expectation maximization (EM) step for the fractional imputation.
OUTPUT	<OUT=SAS-data-set ><OUTJKCOEFS=SAS-data-set >< keyword=name...keyword=name >; The OUTPUT statement creates a SAS data set that contains the imputed data. You must use the OUTPUT statement to store the imputed data in a SAS data set.
REPWEIGHTS	variables; The REPWEIGHTS statement names variables that provide replicate weights.
STRATA	variables; The STRATA statement names one or more variables that form the strata in a stratified sample design. The combinations of levels of STRATA variables define the strata in the sample, where strata are nonoverlapping subgroups that were sampled independently.
VAR	variables; The VAR statement names the analysis variables to be imputed. The analysis variables can be either character or numeric. The categorical variables in the VAR statement, which can be either character or numeric, must also be specified in the CLASS statement. Only variables that you specify in the VAR statement will be imputed.
WEIGHT	variable; The WEIGHT statement names the variable that contains the sampling weights. This variable must be numeric, and the sampling weights must be positive numbers. If an observation has a weight that is nonpositive or missing, then PROC SURVEYIMPUTE omits that observation from the analysis

The below example illustrates the approximate Bayesian bootstrap hot-deck imputation method by using a simulated data set from a fictitious survey of drug abusers. A stratified clustered sample of drug abuse treatment centers is taken from a list of available treatment centers. The list is first stratified based on geographic locations. From each strata, two or three treatment centers are sampled as the primary sampling units (PSU). Data are collected from individual patients within the selected treatment centers. The survey collects information about the substances that the patients used (such as drugs, alcohol, and marijuana) along with insurance information and treatment information.

The data set contains 736 observation units in 35 PSUs and 10 strata. The sum of the weights is 19,600. Therefore, the survey data represent a population of 19,600 patients from the study area. Some participants did not respond to all questions. The data set contains missing values in many variables.

To impute the missing items, you first need to decide whether to impute within imputation cells. Imputation cells divide the data into groups of similar units such that the recipient units share similar characteristics with the donor units in the same group. For example, it is reasonable to believe that different age groups, races, and income categories might have different responses to the drug abuse survey. You can use these characteristics to create imputation cells. Characteristics for imputation cells might come from the same survey or might come from other sources such as census data or previous surveys. In this example, assume that the imputation cells are available as a variable called ImputationCell in the data set.

The data set DrugAbuse contains the following items:

DRUGABUSE	
STRATA	stratum identification
PSU	PSU identification (treatment centers)
OBSWEIGHT	observation weight for patients
IMPUTATIONCELL	imputation cell identification
AGE	age, in years
SEX	1 for female and 2 for male
RACE	1 for white, 2 for black, and 3 for others
INSURANCE	1 if the patient has any insurance, and 2 otherwise

DRUG	1 if the patient used any drugs in the past three months, and 2 otherwise
ALCOHOL	1 if the patient consumed any alcohol in the past month, and 2 otherwise
TREATMENT	1 if the patient is being treated for the first time, and 2 otherwise

```
data DrugAbuse;
   input Strata PSU ObsWeight ImputationCell Age Sex Race Insurance
         Drug Alcohol Treatment;
   datalines;
 1    1     5        1      74  1    1      3       2      2        1
 1    1     5        1      20  0    3      1       2      2        1
 1    1     5        3      42  1    2      1       1      2        1
 1    1     5        3      65  1    3      2       1      2        1
 1    1     5        2      53  1    1      1       1      1        1
 1    1     5        3      49  1    1      1       1      2        1
 1    1     5        3      51  0    2      1       2      1        1
 1    1     5        2      77  0    3      1       1      1        1
 1    1     5        2      26  1    1      1       1      2        1
 1    1     5        3      28  0    3      1       1      1        1
 1    1     5        1      71  1    1      1       2      2        1
 1    1     5        2      72  1    1      3       2      2        1
 1    1     5        3      24  1    1      1       1      1        1
 1    1     5        2      65  1    1      2       1      1        2
 1    1     5        3      47  1    1      1       1      1        1
 1    1     5        2      37  1    1      2       1      2        1
 1    1     5        2      46  1    1      3       1      1        1
 1    1     5        2      52  1    1      1       1      2        2
 1    1     5        3      60  0    3      1       1      2        1
 1    1     5        1      31  0    1      1       1      2        1
 1    2     5        1      23  0    3      3       1      1        1
 1    2     5        2      78  1    1      .       .      .        .
 1    2     5        2      29  1    1      1       1      1        1
 1    2     5        2      21  .    .      .       .      .        .

   ... more lines ...

10    4  55.5556      1      40  0    3      2       1      1        2
10    4  55.5556      1      32  1    3      1       2      2        1
10    4  55.5556      3      68  0    1      2       2      1        2
10    4  55.5556      3      35  1    1      2       1      2        2
;
run;
```

The following statements request that the missing items be imputed by using the approximate Bayesian bootstrap hot-deck imputation method:

```
proc surveyimpute data=DrugAbuse method=hotdeck(selection=abb)
                  ndonors=5 seed=773269;
   var Sex Race Insurance Drug Alcohol Treatment;
   cells ImputationCell;
   output out=DrugAbuseABB;
run;
```

The PROC SURVEYIMPUTE statement invokes the procedure, the DATA= option specifies the input data set DrugAbuse, the METHOD= option requests the hot-deck imputation method, the METHOD=HOTDECK (SELECTION=ABB) option requests the approximate Bayesian bootstrap method, the NDONORS= option requests five donor units for every recipient unit, and the SEED= option specifies the random number generator seed. The VAR statement specifies the variables that are to be imputed, the CELLS statement identifies the imputation cell variable ImputationCell, and the OUT= option in the OUTPUT statement names the output data set DrugAbuseABB.

The SURVEYIMPUTE Procedure

Imputation Information	
Data Set	WORK.DRUGABUSE
Imputation Method	HOTDECK
Selection Method	ABB
Random Number Seed	773269

Number of Observations Read	736
Number of Observations Used	736

Missing Data Patterns

Group	Sex	Race	Insurance	Drug	Alcohol	Treatment	Freq	Sum of Weights	Unweighted Percent	Weighted Percent	Group Means					
											Sex	Race	Insurance	Drug	Alcohol	Treatment
1	X	X	X	X	X	X	681	681	92.53	92.53	0.566814	1.491924	1.718062	1.292217	1.716593	1.201175
2	X	X	X	X	X	.	34	34	4.62	4.62	0.500000	1.441176	1.588235	1.294118	1.588235	.
3	X	X	.	.	.	.	13	13	1.77	1.77	0.692308	1.230769	.	.	.	.
4	.	.	.	.	.	.	8	8	1.09	1.09	.	.	.	.	.	.

Imputation Summary		
Observation Status	Number of Observations	Sum of Weights
Nonmissing	681	681
Missing	55	55
Missing, Imputed	55	55
Missing, Not Imputed	0	0

UnitID	Recipient	Strata	PSU	ObsWeight	ImputationCell	Age	Sex	Race	Insurance	Drug	Alcohol	Treatment
20	0	1	1	5	1	31	0	1	1	1	2	1
21	0	1	2	5	1	23	0	3	3	1	1	1
22	1	1	2	5	2	78	1	1	1	1	1	1
22	2	1	2	5	2	78	1	1	1	1	1	1
22	3	1	2	5	2	78	1	1	2	1	2	1
22	4	1	2	5	2	78	1	1	3	1	2	1
22	5	1	2	5	2	78	1	1	1	1	2	2
23	0	1	2	5	2	29	1	1	1	1	1	1
24	1	1	2	5	2	21	1	2	2	1	2	2
24	2	1	2	5	2	21	1	3	1	1	1	1
24	3	1	2	5	2	21	1	2	2	1	2	1
24	4	1	2	5	2	21	1	1	2	1	2	1
24	5	1	2	5	2	21	1	2	2	2	1	1
25	0	1	2	5	3	85	0	1	3	1	2	1

Some selected observations from the output data set are displayed. The output data set DrugAbuseABB contains the unit identification, the recipient index, and all the variables from the input data set DrugAbuse. Units that are complete respondents have one row, but units that are incomplete respondents have five rows in the output data set. For example, unit 21 is a complete respondent, so it has only one row in the output data set and its Recipient value is 0. Unit 22 is an incomplete respondent, so it has five rows in the output data set and its Recipient values range from 1 to 5.

MULTIPLE IMPUTATION

Multiple Imputation (MI) is not simply a technique for imputing missing data. It is also a method for obtaining estimates and correct inferences for statistics ranging from simple descriptive statistics to the parameters of complex multivariate models. The imputed values are draws from a distribution, so they inherently contain some variation. Thus, multiple imputation (MI) solves the limitations of single imputation by introducing an additional form of error based on variation in the parameter estimates across the imputation, which is called "between imputation error". It replaces each missing item with two or more acceptable values, representing a distribution of possibilities.

PROC MI in SAS is a simulation-based procedure. Its purpose is not to re-create the individual missing values as close as possible to the true ones, but to handle missing data to achieve valid statistical inference. There are several imputation methods in PROC MI. The method of choice depends on the pattern of missingness in the data and the type of the imputed variable, as the table below summarizes:

MISSING DATA PATTERN	TYPE OF IMPUTED VARIABLE	AVAILABLE METHODS
MONOTONE	Continuous	Parametric method that assumes multivariate normality • Monotone regression • Monotone predicted mean matching
		Nonparametric method that assumes propensity scores • Monotone propensity score
	Ordinal	Monotone logistic regression
	Nominal	Monotone discriminant function
ARBITRARY	Continuous	Markov Chain Monte Carlo (MCMC) full-data imputation
		MCMC monotone-data imputation
		Fully conditional specification (FCS), which assumes the existence of a joint distribution for all variables. FCS regression
		FCS predicted mean matching
	Ordinal	FCS logistic regression
	Nominal	FCS discriminant function

PROC MI has the below major features,

- Multivariate normal for continuous variable
- Flexible continuous specification (FCS) or Sequential regression
    - Linear regression or predictive mean matching for continuous variables
    - Logistic for binary and ordinal logistic for polytomous; discriminant analysis for classification variables
- Several nice options and print out missing data pattern and other descriptive information
- Stores all the completed data sets in a single data set (use the system variable _IMPUTE_ to select one imputed data set)

The following statements are available in the MI procedure.

PROC MI	<options>; The PROC MI statement invokes the MI procedure
BY	variables; You can specify a BY statement with PROC MI to obtain separate analyses of observations in groups that are defined by the BY variables. When a BY statement appears, the procedure expects the input data set to be sorted in order of the BY variables. If you specify more than one BY statement, only the last one specified is used.
CLASS	variables; The CLASS statement specifies the classification variables in the VAR statement. Classification variables can be either character or numeric. The CLASS statement must be used in conjunction with either an FCS or MONOTONE statement.
EM	<options>; The expectation-maximization (EM) algorithm is a technique for maximum likelihood estimation in parametric models for incomplete data
FCS	<options>; The FCS statement specifies a multivariate imputation by fully conditional specification methods. If you specify an FCS statement, you must also specify a VAR statement.
FREQ	variable; To run a procedure on an input data set that contains observations that occur multiple times, you can use a variable in the data set to represent how frequently

	observations occur and specify a FREQ statement with the name of that variable as its argument (variable) when you run the procedure.
MCMC	<options>; The MCMC statement specifies the details of the MCMC method for imputation.
MNAR	options; The MNAR statement imputes missing values by using the pattern-mixture model approach, assuming the missing data are missing not at random (MNAR).
MONOTONE	<method <(<imputed <= effects>> </ options>)>> <…method <(<imputed <= effects>> </ options>)>> ; The MONOTONE statement specifies imputation methods for data sets with monotone missingness. You must also specify a VAR statement, and the data set must have a monotone missing pattern with variables ordered in the VAR list.
TRANSFORM	transform (variables</ options>)<…transform (variables</ options>)> ; The TRANSFORM statement lists the transformations and their associated variables to be transformed. The options are transformation options that provide additional information for the transformation.
VAR	variables; The VAR statement lists the variables to be analyzed. The variables can be either character or numeric. If you omit the VAR statement, all continuous variables not mentioned in other statements are used. The VAR statement is required if you specify either an FCS statement, a MONOTONE statement, an IMPUTE=MONOTONE option in the MCMC statement, or more than one number in the MU0=, MAXIMUM=, MINIMUM=, or ROUND= option.

The below example, the pattern of missingness is arbitrary and all variables with missing values are continuous. We choose MCMC full-data imputation, which uses a single chain to create 5 imputations. The posterior mode, the highest observed-data posterior density, with a noninformative prior, is computed from the expectation-maximization (EM) algorithm and is used as the starting value for the chain.

```
proc mi data=Sashelp.Heart nimpute=5 out=ImputedHeart seed=55555;
        mcmc;
        var AgeAtStart Height Weight Diastolic Systolic MRW Smoking Cholesterol;
run;
```

Variance Information

Variable	Variance			DF	Relative Increase in Variance	Fraction Missing Information	Relative Efficiency
	Between	Within	Total				
Height	0.000000292	0.002464	0.002464	5205.1	0.000142	0.000142	0.999972
Weight	0.000324	0.160593	0.160982	5154.3	0.002425	0.002422	0.999516
MRW	0.000206	0.076675	0.076922	5120.9	0.003220	0.003215	0.999357
Smoking	0.000345	0.027823	0.028236	4023.5	0.014860	0.014748	0.997059
Cholesterol	0.017539	0.387542	0.408588	1154.9	0.054308	0.052766	0.989557

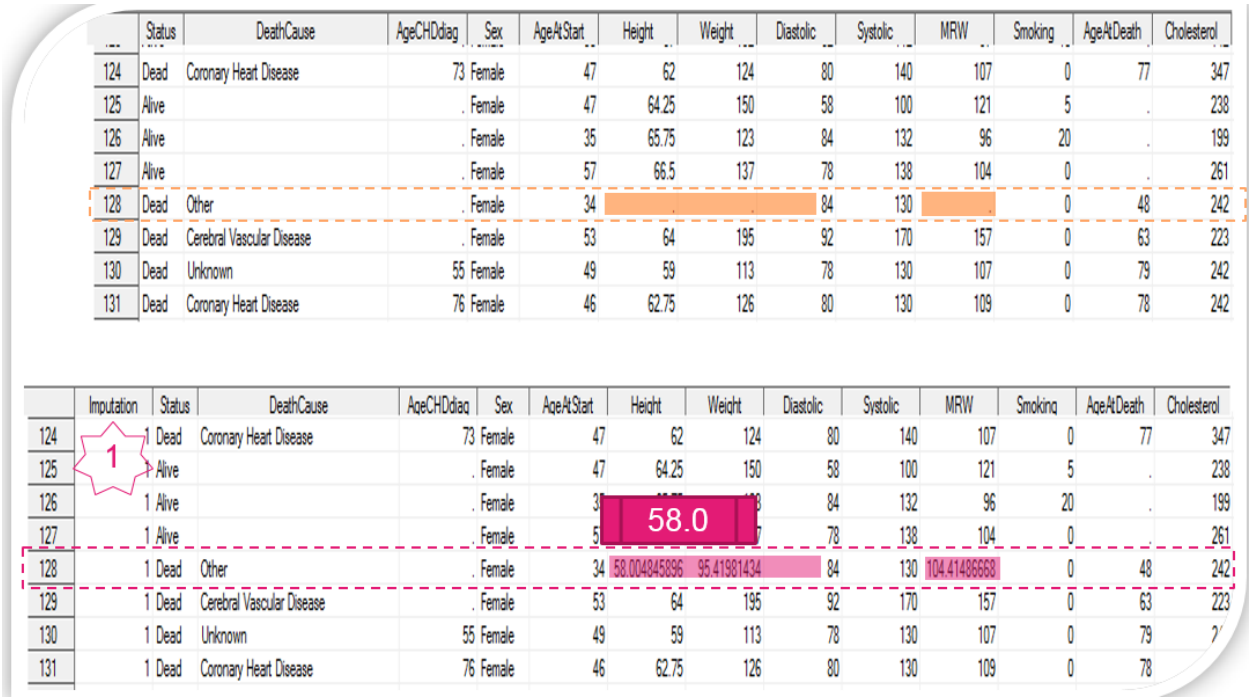
Parameter Estimates

Variable	Mean	Std Error	95% Confidence Limits		DF	Minimum	Maximum	Mu0	t for H0: Mean=Mu0	Pr > |t|
Height	64.811979	0.049637	64.7147	64.9093	5205.1	64.811077	64.812496	0	1305.71	<.0001
Weight	153.062451	0.401226	152.2759	153.8490	5154.3	153.044720	153.083799	0	381.49	<.0001
MRW	119.945077	0.277348	119.4014	120.4888	5120.9	119.929548	119.961558	0	432.47	<.0001
Smoking	9.369757	0.168036	9.0403	9.6992	4023.5	9.340612	9.387299	0	55.76	<.0001
Cholesterol	227.483481	0.639209	226.2293	228.7376	1154.9	227.294263	227.645529	0	355.88	<.0001

The output variance information presents the between-imputation variance, within-imputation variance, and total variance for combining complete-data inferences for each variable, along with the degrees of freedom for the total variance, relative increase in variance, fraction missing information and relative efficiency. The parameter estimates output summarizes basic descriptive statistics for the imputed values by variable.

The imputed data sets are stored in the ImputedHeart data set, with the index variable _Imputation indicating the imputation numbers. The imputed values show in the output dataset for height variable as 58.0 (obs 128), 64.5 (obs 5337), 60.9 (obs 10546), 59.7 (obs 15755) and 61.5 (obs 20964).

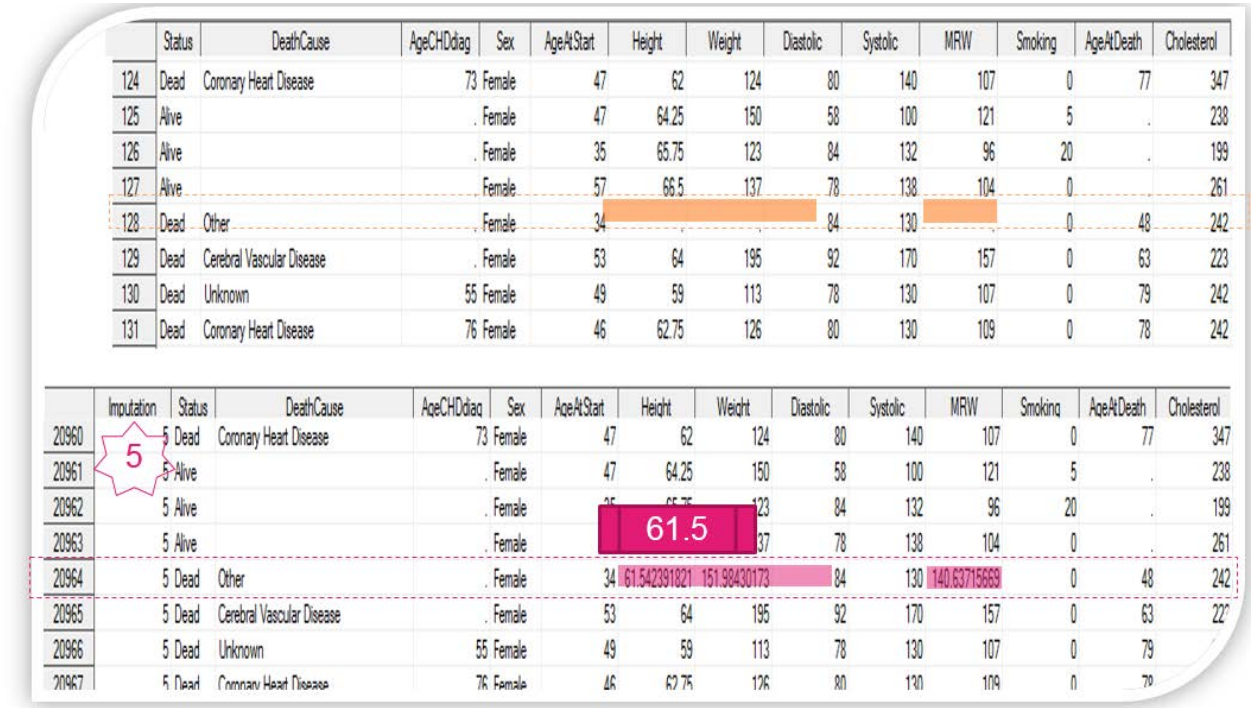
The below figure represents the imputed value for height variable by an imputation number 1.



	Status	DeathCause	AgeCHDdiag	Sex	AgeAtStart	Height	Weight	Diastolic	Systolic	MRW	Smoking	AgeAtDeath	Cholesterol
124	Dead	Coronary Heart Disease	73	Female	47	62	124	80	140	107	0	77	347
125	Alive		.	Female	47	64.25	150	58	100	121	5	.	238
126	Alive		.	Female	35	65.75	123	84	132	96	20	.	199
127	Alive		.	Female	57	66.5	137	78	138	104	0	.	261
128	Dead	Other	.	Female	34			84	130		0	48	242
129	Dead	Cerebral Vascular Disease	.	Female	53	64	195	92	170	157	0	63	223
130	Dead	Unknown	55	Female	49	59	113	78	130	107	0	79	242
131	Dead	Coronary Heart Disease	76	Female	46	62.75	126	80	130	109	0	78	242

	Imputation	Status	DeathCause	AgeCHDdiag	Sex	AgeAtStart	Height	Weight	Diastolic	Systolic	MRW	Smoking	AgeAtDeath	Cholesterol
124	1	Dead	Coronary Heart Disease	73	Female	47	62	124	80	140	107	0	77	347
125	1	Alive		.	Female	47	64.25	150	58	100	121	5	.	238
126	1	Alive		.	Female	3			84	132	96	20	.	199
127	1	Alive		.	Female	5	58.0		78	138	104	0	.	261
128	1	Dead	Other	.	Female	34	58.004845896	95.41981434	84	130	104.41486668	0	48	242
129	1	Dead	Cerebral Vascular Disease	.	Female	53	64	195	92	170	157	0	63	223
130	1	Dead	Unknown	55	Female	49	59	113	78	130	107	0	79	2
131	1	Dead	Coronary Heart Disease	76	Female	46	62.75	126	80	130	109	0	78	

The below figure represents the imputed value for height variable by an imputation number 5.



	Status	DeathCause	AgeCHDdiag	Sex	AgeAtStart	Height	Weight	Diastolic	Systolic	MRW	Smoking	AgeAtDeath	Cholesterol
124	Dead	Coronary Heart Disease	73	Female	47	62	124	80	140	107	0	77	347
125	Alive		.	Female	47	64.25	150	58	100	121	5	.	238
126	Alive		.	Female	35	65.75	123	84	132	96	20	.	199
127	Alive		.	Female	57	66.5	137	78	138	104	0	.	261
128	Dead	Other	.	Female	34			84	130		0	48	242
129	Dead	Cerebral Vascular Disease	.	Female	53	64	195	92	170	157	0	63	223
130	Dead	Unknown	55	Female	49	59	113	78	130	107	0	79	242
131	Dead	Coronary Heart Disease	76	Female	46	62.75	126	80	130	109	0	78	242

	Imputation	Status	DeathCause	AgeCHDdiag	Sex	AgeAtStart	Height	Weight	Diastolic	Systolic	MRW	Smoking	AgeAtDeath	Cholesterol
20960	5	Dead	Coronary Heart Disease	73	Female	47	62	124	80	140	107	0	77	347
20961	5	Alive		.	Female	47	64.25	150	58	100	121	5	.	238
20962	5	Alive		.	Female	3		123	84	132	96	20	.	199
20963	5	Alive		.	Female		61.5	37	78	138	104	0	.	261
20964	5	Dead	Other	.	Female	34	61.542391821	151.98430173	84	130	140.63715669	0	48	242
20965	5	Dead	Cerebral Vascular Disease	.	Female	53	64	195	92	170	157	0	63	22
20966	5	Dead	Unknown	55	Female	49	59	113	78	130	107	0	79	
20967	5	Dead	Coronary Heart Disease	76	Female	46	62.75	126	80	130	109	0	78	

The data set can now be analyzed using standard statistical procedures with _Imputation_ as a BY variable.

Multiple imputation often assumes that missing values are missing at random (MAR), and the following example use the MI procedure to impute missing values under this assumption.

Sample Trial Data (First 10 Observations)

Obs	Trt	y0	y1
1	0	11.4826	.
2	0	10.009	10.8667
3	0	11.3643	10.666
4	0	11.3098	10.8297
5	0	11.3094	.
6	1	10.3815	10.5587
7	1	11.2001	13.7616
8	1	9.7002	10.346
9	1	10.0801	.
10	1	11.2667	11.0634

Suppose that a pharmaceutical company is conducting a clinical trial to test the efficacy of a new drug. The trial is conducted with two groups that have an equal number of patients: a treatment group that receives the new drug and a placebo control group. The variable TRT is an indicator variable, with a value of 1 for patients in the treatment group and a value of 0 for patients in the control group. The variable Y0 is the baseline efficacy score, and the variable Y1 is the efficacy score at a follow-up visit. Suppose in a trial that the variables TRT and Y0 are fully observed and the variable Y1 contains missing values in both the treatment and control groups.

```
proc mi data=Mono2 seed=14823 out=outmi;
   class Trt;
   monotone reg;
   var Trt y0 y1;
run;
```

The MI Procedure

Model Information	
Data Set	WORK.MONO2
Method	Monotone
Number of Imputations	5
Seed for random number generator	14823

Monotone Model Specification	
Method	Imputed Variables
Regression	y0 y1

Missing Data Patterns							
						Group Means	
Group	Trt	y0	y1	Freq	Percent	y0	y1
1	X	X	X	7	70.00	10.747371	11.156014
2	X	X	.	3	30.00	10.957367	.

Variance Information							
	Variance				Relative	Fraction	
Variable	Between	Within	Total	DF	Increase in Variance	Missing Information	Relative Efficiency
y1	0.016696	0.156903	0.176938	6.5119	0.127691	0.118863	0.976779

Parameter Estimates									
Variable	Mean	Std Error	95% Confidence Limits		DF	Minimum	Maximum	Mu0	t for H0: Mean=Mu0	Pr > |t|
y1	11.117555	0.420640	10.10758	12.12753	6.5119	10.932723	11.275095	0	26.43	<.0001

The MAR assumption should be examined after the analysis of the imputed dataset.

The following example generate imputed dataset for a specified sequence of shift parameters, which adjust the imputed values for observations in the treatment group (TRT=1)

```
proc mi data=Mono2 seed=14823
out=outmi;
  class Trt;
  monotone reg;
  mnar adjust( y1/shift=-1.53
       adjustobs=(Trt='1'));
  var Trt y0 y1;

run;
```

Model Information

Data Set	WORK.MONO2
Method	Monotone
Number of Imputations	5
Seed for random number generator	14823

Monotone Model Specification

Method	Imputed Variables
Regression	y0 y1

Missing Data Patterns

Group	Trt	y0	y1	Freq	Percent	Group Means y0	Group Means y1
1	X	X	X	7	70.00	10.747371	11.156014
2	X	X	.	3	30.00	10.957367	.

MNAR Adjustments to Imputed Values

Imputed Variable	Observations	Shift
y1	Trt = 1	-1.5300

Variance Information

Variable	Variance Between	Variance Within	Variance Total	DF	Relative Increase in Variance	Fraction Missing Information	Relative Efficiency
y1	0.016696	0.197003	0.217038	6.7103	0.101699	0.096154	0.981132

Parameter Estimates

Variable	Mean	Std Error	95% Confidence Limits		DF	Minimum	Maximum	Mu0	t for H0: Mean-Mu0	Pr > |t|
y1	10.964555	0.465873	9.853226	12.07588	6.7103	10.779723	11.122095	0	23.54	<.0001

In the MI procedure the MNAR statement used to impute missing values under various MNAR scenarios and examine the results. If the results under MNAR differ from the results under MAR, then you should question the conclusion under the MAR assumption.

MAXIMUM LIKELIHOOD IMPUTATION

This method to get the variance-covariance matrix for the variables in the model based on all the available data points, and then use the obtained variance- covariance matrix to estimate the regression model.

PROC MI uses the default algorithm (EM) to do maximum likelihood of the means and the covariance matrix and, then, it considers these estimates as starting values for multiple imputation algorithms. The NIMPUTE=0 suppresses the multiple imputation and EM statement estimates and writes the means and covariance matrix into SAS data set EMImputedHeart

```
proc mi data=Sashelp.Heart nimpute=0;
       em out=EMImputedHeart;
       var AgeAtStart Height Weight Diastolic Systolic MRW Smoking Cholesterol;
run;
```

The MI Procedure

Model Information

Data Set	SASHELP.HEART
Method	MCMC
Multiple Imputation Chain	Single Chain
Initial Estimates for MCMC	EM Posterior Mode
Start	Starting Value
Prior	Jeffreys
Number of Imputations	0
Number of Burn-in Iterations	200
Number of Iterations	100
Seed for random number generator	290896000

Missing Data Patterns

Group	AgeAtStart	Height	Weight	Diastolic	Systolic	MRW	Smoking	Cholesterol	Freq	Percent	Group Means AgeAtStart	Height	Weight	Diastolic	Systolic	MRW	Smoking	Cholesterol
1	X	X	X	X	X	X	X	X	5039	96.74	44.073030	64.826206	153.060131	85.411193	136.987498	119.864457	9.370907	227.410796
2	X	X	X	X	X	X	X	.	124	2.38	43.653226	64.137097	154.193548	83.459677	134.040323	124.387097	8.685484	.
3	X	X	X	X	X	X	.	X	8	0.15	42.875000	65.312500	158.125000	83.500000	131.625000	121.750000	.	210.125000
4	X	X	X	X	X	X	.	.	28	0.54	45.214286	65.482143	151.392857	87.714286	140.571429	115.928571	.	.
5	X	X	.	X	X	.	X	X	4	0.08	40.000000	63.687500	.	73.500000	126.500000	.	27.500000	256.500000
6	X	.	X	X	X	X	X	X	4	0.08	55.000000	.	154.000000	81.250000	129.000000	124.500000	10.250000	254.250000
7	X	.	.	X	X	.	X	X	2	0.04	34.000000	.	.	77.000000	125.000000	.	2.500000	201.500000

Initial Parameter Estimates for EM

TYPE	_NAME_	AgeAtStart	Height	Weight	Diastolic	Systolic	MRW	Smoking	Cholesterol
MEAN		44.068727	64.813185	153.086681	85.358610	136.909580	119.957525	9.366518	227.417441
COV	AgeAtStart	73.529838	0	0	0	0	0	0	0
COV	Height	0	12.835792	0	0	0	0	0	0
COV	Weight	0	0	836.101866	0	0	0	0	0
COV	Diastolic	0	0	0	168.301098	0	0	0	0
COV	Systolic	0	0	0	0	563.568435	0	0	0
COV	MRW	0	0	0	0	0	399.336335	0	0
COV	Smoking	0	0	0	0	0	0	144.755816	0
COV	Cholesterol	0	0	0	0	0	0	0	2019.201302

EM (MLE) Parameter Estimates

TYPE	_NAME_	AgeAtStart	Height	Weight	Diastolic	Systolic	MRW	Smoking	Cholesterol
MEAN		44.068727	64.812354	153.072405	85.358610	136.909580	119.949917	9.369819	227.405950
COV	AgeAtStart	73.515722	-4.047092	23.354557	30.630758	77.213740	35.060972	-17.295472	105.390745
COV	Height	-4.047092	12.831202	53.597287	-0.662474	-6.070537	-9.750197	12.393431	-12.836556
COV	Weight	23.354557	53.597287	836.128050	123.013989	181.028016	443.321192	30.799915	95.621641
COV	Diastolic	30.630758	-0.662474	123.013989	168.268788	245.120542	99.821910	-10.148263	106.846696
COV	Systolic	77.213740	-6.070537	181.028016	245.120542	563.460244	171.972983	-26.582956	212.219624
COV	MRW	35.060972	-9.750197	443.321192	99.821910	171.972983	399.260989	-30.154493	123.324234
COV	Smoking	-17.295472	12.393431	30.799915	-10.148263	-26.582956	-30.154493	144.746562	-6.171769
COV	Cholesterol	105.390745	-12.836556	95.621641	106.846696	212.219624	123.324234	-6.171769	2018.939258

CONCLUSION

Missing data introduces complexity in statistical inference about unknown parameter or the population quantity. The complete data model assumption and missing data mechanism are needed to construct inference from incomplete data. The mean and median imputation have a similar mean percentage of observations correctly classified. The hot deck imputation has the highest mean percentage of observations correctly classified. Multiple imputation approach serves as the most general-purpose method, but not the best absolute methodology. From a practical point of view, a carefully chosen imputation process and a decently efficient statistical method for analysis of complete data will not have many problems.

REFERENCES

1. Data Science - Innovative Developments in Data Analysis and Clustering by Francesco Palumbo, Angela Montanari, Maurizio Vichi and the topic from Lynette A. Hunt
2. Fundamentals of Mathematical Statistics by S.C. Gupta and V.K Kapoor
3. SAS Blogs by Rick Wicklin
4. Missing Data Analysis in Practice, Trivellore Raghunathan
5. Multiple Imputation of Missing Data Using SAS by Patricia Berglund and Steven Heeringa
6. SAS/STAT 14.1 Online Documentation
7. Missing data: the hidden problem, SPSS White Paper
8. Dealing with missing data: Key assumptions and methods for applied analysis, Marina Soley-Bori

ACKNOWLEDGMENTS

I would like to express my special thanks of all authors in the references section.

CONTACT INFORMATION

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

Edwin Ponraj Thangarajan, Giri Balasubramanian
PRA Health Sciences
40, II Main Road, R.A. Puram
Chennai - 600 028, Tamilnadu, India
Email: ThangarajanEdwin@prahs.com, BalasubramanianGiri@prahs.com
Web: www.prahs.com