PharmaSUG China 2019 - Paper SP-003

The Diagnosis and Handling of Missing Data – MCMC in Multiple Imputation

Ying Yao, Boehringer-Ingelheim;

ABSTRACT

Even in a well-designed and controlled study, missing data occurs in almost all research. Missing data can reduce the statistical power of a study and can produce biased estimates, leading to invalid conclusions. This manuscript reviews the problems and types of missing data, along with the techniques for handling missing data. The mechanisms by which missing data occur are illustrated, and the methods for handling the missing data are went through as well. Among them MCMC method in MI – multiple imputation will be introduced with details and SAS examples.

INTRODUCTION

Missing data (or missing values) is defined as the data value that is not stored for a variable in the observation of interest. The problem of missing data is relatively common in almost all research and can have a significant effect on the conclusions that can be drawn from the data. Accordingly, some studies have focused on handling the missing data, problems caused by missing data, and the methods to avoid or minimize such in medical research. However, most researchers have drawn conclusions based on the assumption of a complete data set in the past decades. The general topic of missing data has attracted more attention in the field of clinical trials recently, and the mechanism of how to deal with missing data become a growing appeal to lots of researchers.

In this paper, we would like to display the pattern of missing data that frequently happen in clinical trials and clarify the reasons why they happen, and discuss several ways of prevention for the missing data cases. Then we will review the theory of missing data and methodology regarding how to identify and then how to handle the missing data with examples in SAS. In the last section, we will go through the technics and the pre-requisites for applying and modeling to make sure the methods are used properly.

Multiple imputation (MI) is an approach for handling missing values in a dataset that allows researchers to use the entirety of the observed data. Although MI has become more prevalent in political science, its use still lags far behind complete case analysis – also known as list-wise deletion – which remains the default treatment for missing data in Stata, R, SAS, and SPSS.

BACKGROUND

Missing data present various problems. First, the absence of data reduces statistical power, which refers to the probability that the test will reject the null hypothesis when it is false. Second, the lost data can cause bias in the estimation of parameters. Third, it can reduce the representativeness of the samples. Fourth, it may complicate the analysis of the study. Each of these distortions may threaten the validity of the trials and can lead to invalid conclusions.

Common reasons for missing data include survey structure that deliberately results in missing data (questions asked only of women), refusal to answer (sensitive questions), insufficient knowledge (month of first words spoken), and attrition due to death or loss of contact with respondents in longitudinal surveys. Missing data can be categorized as unit non-response (entire survey is missing) or item non-response (some questions are missing within a survey).

TYPES OF MISSING DATA

Rubin first described and divided the types of missing data according to the assumptions based on the reasons for the missing data. In general, there are three types of missing data according to the mechanisms of missingness.

MCAR: Missing completely at random (MCAR) is defined as when the probability that the data are missing is not related to either the specific value which is supposed to be obtained or the set of observed responses. MCAR is an ideal but unreasonable assumption for many studies performed in the field of clinical trials. However, if data are missing by design, because of an equipment failure or because the samples are lost in transit or technically unsatisfactory, such data are regarded as being MCAR. The statistical advantage of data that are MCAR is that the analysis remains unbiased. Power may be lost in the design, but the estimated parameters are not biased by the absence of the data.

MAR: Missing at random (MAR) is a more realistic assumption for the studies performed in the anesthetic field. Data are regarded to be MAR when the probability that the responses are missing depends on the set of observed responses, but is not related to the specific missing values which is expected to be obtained.As we tend to consider randomness as not producing bias, we may think that MAR does not present a problem. However, MAR does not mean that the missing data can be ignored. If a dropout variable is MAR, we may expect that the probability of a dropout of the variable in each case is conditionally independent of the variable, which is obtained currently and expected to be obtained in the future, given the history of the obtained variable prior to that case.

MNAR: If the characters of the data do not meet those of MCAR or MAR, then they fall into the category of missing not at random (MNAR). The cases of MNAR data are problematic. The only way to obtain an unbiased estimate of the parameters in such a case is to model the missing data. The model may be incorporated into a more complex one for estimating the missing values.

Analysts often make assumptions about the nature of missing data including categorizations. PROC MI and PROC MIANALYZE both use the MAR assumption for all analyses.

INTRODUCTION OF MULTIPLE IMPUTATION

Imputation methods can be defined as simple or multiple. Though simple imputation is attractive and often used to impute missing data, the focus of this paper is use of multiple imputation methods in SAS. This is due to the ability of the multiple imputation process to incorporate statistically sophisticated techniques and draw from distributions of “plausible” values while accounting for the variability introduced by the process of selecting a value for the missing data point (Rubin, 1987). Simple imputation methods such as inserting a mean value or a value selected from a similar type of respondent are attractive due to ease of concept and implementation but do not account for the variability introduced by the imputation process. They also tend to distort the variable distribution once imputation is complete. Given these limitations, multiple imputation is generally considered a preferred method for dealing with missing data.

Every approach to MI follows the same two steps: (1) replace the missing values in the data with values that preserve the relationships expressed by the observed data; and (2) use independently drawn imputed values to create several datasets, and use the variation across these datasets to inflate model standard errors so that they reflect our uncertainty about the parametric imputation model.

In SAS processing, the robust, flexible option in many practical problems and the major focus of this book is to address missing values within the MI framework for estimation and inference. This approach consists of a three-step process: 1) formulation of the imputation model and imputation of missing data using PROC MI, 2) analysis of complete data sets using standard SAS procedures (that assume the data are identically and independently distributed or from a simple random sample) or SURVEY procedures for analysis of data from a complex sample design, and 3) analysis of the output from the two previous steps using PROC MIANALYZE. Many types of missing data patterns and analytic models can be handled within this framework, making it, in our opinion, the preferred option for dealing with most missing data problems.

In practice, however, there are many ways to implement MI, and these approaches differ greatly in the assumptions they make about the structure and distribution of the data. Now let’s take a look on the various methods in multiple imputation.

METHODS IN MULTIPLE IMPUTATION

Depending on the pattern of missing data and variable types, PROC MI provides three primary classes of methods for generating the multiple imputations. If the pattern of missing data is univariate or monotonic, the monotone option is the method of choice. For an arbitrary multivariate pattern of missing data, the choice is between the MCMC or the FCS methods. Table 2.1 summarizes methods available in SAS (v9.4) according to the pattern of missing data and the type of variable being imputed.

Missing Data Pattern	Variable Type	Method
Monotone	Continuous Binary/Ordinal Nominal	Liner regression, predictive mean matching, propensity score Logistic regression Discriminant function
Arbitrary	Continuous	With continuous covariates: MCMC monotone methods MCMC full data imputation
Arbitrary	Continuous	With mixed covariates: FCS regression FCS predictive mean
Arbitrary	Binary/Ordinal Nominal	FCS logistic regression FCS discriminant function

Table 1. SAS PROC MI Imputation Methods

MCMC - MARKOV CHAIN MONTE CARLO

In the common situation where the missing data problem is multivariate, has an arbitrary pattern of missing values, and may include variables of differing type (continuous, nominal, binary, ordinal), it is analytically difficult or impossible to evaluate the true expression for the joint posterior distribution, p(θ |Yobs). In such cases, statisticians have devised iterative simulation techniques that permit us to approximate draws from the analytically intractable complex joint posterior.

The MCMC method is such an algorithm to simulate the joint posterior, p(θ |Yobs), for arbitrary data patterns in which the underlying complete data are assumed to follow a multivariate normal distribution.

Assuming a multivariate normal distribution for all variables, f(y| θ)=MVN(μ , Σ) and a noninformative or Jeffries prior distribution for the parameters μ and Σ. In thecase of complete data, the posterior distribution p(μ , Σ |Y) can be derived under Bayes' Rule.

However, for an arbitrary pattern of missing data where individual cases are missing different combinations of the variables in Y={Y1,⋯,Yp}, the same posterior—now conditional only on the observed data—is difficult or impossible to derive in a closed form. The PROC MI MCMC full data imputation method uses an iterative Markov chain Monte Carlo method to simulate draws from the posterior, p(μ , Σ |Yobs).

2 STEPS IN MCMC

Here provides a detailed description of the MCMC algorithm. We will describe the algorithm in general terms. The MCMC algorithm involves an iterative sequence of paired I-steps and P-steps.

I-Step

At each iteration of the simulation (t=1,⋯,T), the MCMC algorithm draws imputations from the current iteration's predictive distribution, f(Y(t+1) mis| Yobs, μ (t), Σ (t)). The imputation proceeds case by case, taking into account the pattern of missing variables for the case. For example, the predictive posterior for a case with the observed/missing pattern of Yi=(Y1, . ,Y3, .,Y5) is different from that for Yi=(Y1, . , . ,Y4,Y5). For efficiency, MCMC uses the SWEEP operator (Goodnight 1979)—a computationally convenient way to estimate linear regression parameters from Σ (t)—to derive the

conditional distributions needed to simulate the predictive posterior for each possible pattern of missing data.

P-Step

After each I-step, the parameter values for the predictive distribution are updated by draws from the completed data posterior, p(μ , Σ |Yobs,Y(t+1) mis).

In theory, if the chain of MCMC I-step/P-step pairs is allowed to continue for many iterations, the algorithm will converge so that the imputation draws for the missing values will simulate draws from the true joint posterior, p(Ymis| μ , Σ , Yobs). Once a sufficient burn in period of iterations has passed, the m=1,⋯,M repetitions can be taken as a successive series of systematic draws in the single imputation chain. Another option is to use m=1,⋯,M MCMC runs in parallel chains and obtain each repetition of the multiple imputation as single draws from each of the independent MCMC chains. The single chain option is the default in SAS, but users may optionally request a multiple chain approach in applications of the MCMC method.

SAS EXAMPLES

The *Fitness1* data set is constructed from the *Fitness* data set and contains three variables: *Oxygen*, *RunTime*, and *RunPulse*. Some values have been set to missing, and the resulting data set has an arbitrary pattern of missingness in these three variables.

```
*----------------- Data on Physical Fitness -----------------*
| These measurements were made on men involved in a physical |
| fitness course at N.C. State University.                   |
| Only selected variables of                                 |
| Oxygen (oxygen intake, ml per kg body weight per minute),  |
| Runtime (time to run 1.5 miles in minutes), and            |
| RunPulse (heart rate while running) are used.              |
| Certain values were changed to missing for the analysis.   |
*------------------------------------------------------------*;
data Fitness1;
   input Oxygen RunTime RunPulse @@;
   datalines;
44.609  11.37  178     45.313  10.07  185
54.297   8.65  156     59.571    .      .
49.874   9.22    .     44.811  11.63  176
  .     11.95  176          .  10.85    .
39.442  13.08  174     60.055   8.63  170
50.541    .      .     37.388  14.03  186
44.754  11.12  176     47.273    .      .
51.855  10.33  166     49.156   8.95  180
40.836  10.95  168     46.672  10.00    .
46.774  10.25    .     50.388  10.08  168
39.407  12.63  174     46.080  11.17  156
45.441   9.63  164       .      8.92    .
45.118  11.08    .     39.203  12.88  168
45.790  10.47  186     50.545   9.93  148
48.673   9.40  186     47.920  11.50  170
47.467  10.50  170
;
```

Suppose that the data are multivariate normally distributed and that the missing data are missing at random. This example uses the MCMC method to impute missing values for a data set with an arbitrary

missing pattern. The following statements invoke the MI procedure and specify the MCMC method with six imputations:

```
proc mi data=Fitness1 seed=21355417 nimpute=6 mu0=50 10 180 ;
   mcmc chain=multiple displayinit initial=em(itprint);
   var Oxygen RunTime RunPulse;
run;
```

The "Model Information" table in Output 1 describes the method used in the multiple imputation process. When you use the CHAIN=MULTIPLE option, the procedure uses multiple chains and completes the default 200 burn-in iterations before each imputation. The 200 burn-in iterations are used to make the iterations converge to the stationary distribution before the imputation.

The MI Procedure

Model Information	
Data Set	WORK.FITNESS1
Method	MCMC
Multiple Imputation Chain	Multiple Chains
Initial Estimates for MCMC	EM Posterior Mode
Start	Starting Value
Prior	Jeffreys
Number of Imputations	6
Number of Burn-in Iterations	200
Seed for random number generator	21355417

Output 1. Model Information

By default, the procedure uses a noninformative Jeffreys prior to derive the posterior mode from the EM algorithm as the starting values for the MCMC method.
The "Missing Data Patterns" table in Output 2 lists distinct missing data patterns with corresponding statistics.

Missing Data Patterns

Group	Oxygen	RunTime	RunPulse	Freq	Percent	Group Means		
						Oxygen	RunTime	RunPulse
1	X	X	X	21	67.74	46.353810	10.809524	171.666667
2	X	X	.	4	12.90	47.109500	10.137500	.
3	X	.	.	3	9.68	52.461667	.	.
4	.	X	X	1	3.23	.	11.950000	176.000000
5	.	X	.	2	6.45	.	9.885000	.

Output 2. Missing Data Patterns

When you use the ITPRINT option within the INITIAL=EM option, the procedure displays the "EM (Posterior Mode) Iteration History" table in Output 3.

EM (Posterior Mode) Iteration History

Iteration	-2 Log L	-2 Log Posterior	Oxygen	RunTime	RunPulse
0	254.482800	282.909549	47.104077	10.554858	171.381669
1	255.081168	282.051584	47.104077	10.554857	171.381652
2	255.271408	282.017488	47.104077	10.554857	171.381644
3	255.318622	282.015372	47.104002	10.554523	171.381842
4	255.330259	282.015232	47.103861	10.554388	171.382053
5	255.333161	282.015222	47.103797	10.554341	171.382150
6	255.333896	282.015222	47.103774	10.554325	171.382185
7	255.334085	282.015222	47.103766	10.554320	171.382196

Output 3. EM(Posterior Mode) Iteration History

When you use the DISPLAYINIT option in the MCMC statement, the "Initial Parameter Estimates for MCMC" table in Output 4 displays the starting mean and covariance estimates used in the MCMC method. The same starting estimates are used in the MCMC method for multiple chains because the EM algorithm is applied to the same data set in each chain. You can explicitly specify different initial estimates for different imputations, or you can use the bootstrap method to generate different parameter estimates from the EM algorithm for the MCMC method.

Initial Parameter Estimates for MCMC

TYPE	_NAME_	Oxygen	RunTime	RunPulse
MEAN		47.103766	10.554320	171.382196
COV	Oxygen	24.549967	-5.726112	-15.926036
COV	RunTime	-5.726112	1.781407	3.124798
COV	RunPulse	-15.926036	3.124798	83.164045

Output 4. Initial Parameter Estimates for MCMC

Output 5 and Output 6 display variance information and parameter estimates, respectively, from the multiple imputation.

Variance Information

Variable	Variance			DF	Relative Increase in Variance	Fraction Missing Information	Relative Efficiency
	Between	Within	Total				
Oxygen	0.051560	0.928170	0.988323	25.958	0.064809	0.062253	0.989731
RunTime	0.003979	0.070057	0.074699	25.902	0.066262	0.063589	0.989513
RunPulse	4.118578	4.260631	9.065638	7.5938	1.127769	0.575218	0.912517

Output 5. Variance Information

Parameter Estimates										
Variable	Mean	Std Error	95% Confidence Limits		DF	Minimum	Maximum	Mu0	t for H0: Mean=Mu0	Pr > |t|
Oxygen	47.164819	0.994145	45.1212	49.2085	25.958	46.858020	47.363540	50.000000	-2.85	0.0084
RunTime	10.549936	0.273312	9.9880	11.1118	25.902	10.476886	10.659412	10.000000	2.01	0.0547
RunPulse	170.969836	3.010920	163.9615	177.9782	7.5938	168.252615	172.894991	180.000000	-3.00	0.0182

Output 6. Parameter Estimates

CONCLUSION

The Markov chain Monte Carlo (MCMC) method originated in physics as a tool for exploring equilibrium distributions of interacting molecules. In statistical applications, it is used to generate pseudorandom draws from multidimensional and otherwise intractable probability distributions via Markov chains. A Markov chain is a sequence of random variables in which the distribution of each element depends only on the value of the previous element.

In MCMC simulation, you constructs a Markov chain long enough for the distribution of the elements to stabilize to a stationary distribution, which is the distribution of interest. Repeatedly simulating steps of the chain simulates draws from the distribution of interest. See Schafer (1997) for a detailed discussion of this method.

REFERENCES

Book Patricia Berglund, Steven Heeringa. *Multiple Imputation of Missing Data Using SAS*. SAS Institute Inc., Cary, NC, USA.

Article in conference proceedings Yang, Yuan. *Multiple Imputation Using SAS Software*, SAS Institute Inc.

Website "Why MCMC works." https://freethoughtblogs.com/reprobate/2018/02/17/why-mcmc-works/.

Website "The prevention and handling of the missing data." https://www.researchgate.net/publication/237061322_The_prevention_and_handling_of_the_missing_data

Website "A zero-math introduction to Markov Chain Monte Carlo method."https://towardsdatascience.com/a-zero-math-introduction-to-markov-chain-monte-carlo-methods-dcba889e0c50

Website "THE MI PROCEDURE."

https://support.sas.com/documentation/cdl/en/statug/63962/HTML/default/viewer.htm#statug_mi_sect050.htm

RECOMMENDED READING

- *Base SAS® Procedures Guide*
- *Multiple Imputation of Missing Data Using SAS®*