

# Missing Data: An Introduction (with a focus on multiple imputation)

Workshop offered by the Mississippi Center for  
Supercomputing Research and the UM Office of  
Information Technology

John P. Bentley  
Department of Pharmacy Administration  
University of Mississippi

March 30, 2009

Acknowledgement: Parts of this presentation are adapted from presentations made by Dr. Hemant Tiwari of the UAB Department of Biostatistics and Dr. Gerald McGwin of the UAB Department of Epidemiology.

## Outline

- The problem of missing data (MD)
- Missing data mechanisms
- Methods for dealing with missing data
  - Traditional
  - Single imputation
  - Advanced methods
    - Maximum likelihood methods
    - **Multiple imputation – SAS example**

## **Disclaimer**

- Missing data are a pervasive problem in data analysis. Significant progress has been made with respect to developing advanced statistical methodologies for dealing with missing data. Numerous articles and books have been written on this topic (both review articles and original research). While I have a general appreciation and understanding of these methods, I am not an expert in this field.

## **The problem of missing data**

## The problem of missing data

- Most analytic methods assume a “full” matrix

ID	Sex	Race	Age	Income	Anxiety
1	1	2	35	1	0
2	2	1	62	3	5
3	1	1	17	4	2
4	1	1	48	2	4

- ...but often some observations are not made.

ID	Sex	Race	Age	Income	Anxiety
1	1	2	35		0
2	2	1	62	3	5
3	1	1	17		
4					

- Or complete cases are lost-to-follow-up

## The problem of missing data

Missing data prevalent in any field of research:

- Subjects move away
- Subjects refuse to answer sensitive questions
- Subjects become ill/terminate (mortality)
- The measurement instrument fails
- Data file becomes corrupt
- etc...

## Where is the missing data?

- Values that are missing are *unobserved*.
- There are actual underlying values that would have been observed if techniques to generate data or measure data had been better.
- Hence, the idea: the missing data are “out there” and can be ‘found’.

## The problem of missing data

What is the impact of missing data?

- Smaller effective  $N$  – less power
- Findings are non-representative...even of the sample
- Inferences are invalid
- Estimates are unstable, biased
- Errors are large – Type II errors
- Basically, you have a biased estimate and incorrect variance

## **Missing data mechanisms**

## **Missing data mechanisms**

- Missing Completely at Random (MCAR)
- Missing at Random (MAR)
- Not Missing at Random (NMAR), Non-ignorable (NI)

## Missing data mechanisms

- Missing Completely at Random (MCAR)
  - Suppose some data are missing on  $Y$ . These data are said to be MCAR if the probability that  $Y$  is missing is unrelated to  $Y$  or other variables  $X$  (where  $X$  is a vector of observed variables).

$$\Pr(Y \text{ is missing} | X, Y) = \Pr(Y \text{ is missing})$$

- MCAR is best situation to be in
- MCAR is a strong assumption
- If data are MCAR, complete data sample is a random sub sample of original target sample.

## MCAR

- Missingness does not depend on any values of any variables in the data set.
- Missingness depends on neither the values of the observed variables, nor on those of unobserved variables.
- Does not mean pattern is random, but that missingness does not depend on the data values.

## MCAR Examples

- Research assistant shuffles data sheets and arbitrarily discards some sheets.
- Suppose two variables to be measured, age and income. Everyone answers age, not everyone answers income:
  - If the probability that income is missing is the same for all individuals, regardless of age or income, then the data are MCAR.

## Missing data mechanisms

- Missing at Random (MAR)
  - Data on  $Y$  are missing at random if the probability that  $Y$  is missing does not depend on the value of  $Y$ , after controlling for other observed variables.

$$\Pr(Y \text{ is missing} | X, Y) = \Pr(Y \text{ is missing} | X)$$

- Weaker assumption than MCAR
- MAR is a moderate assumption
- Can test whether missingness on  $Y$  depends on  $X$
- In practice, cannot test whether missingness on  $Y$  depends on  $Y$

## MAR

- Missingness does not depend on the values of any of the missing, or unobserved variables, but might depend on values of the observed variables.
- That is to say, although the occurrence of the missing values themselves may be random, their missingness can be linked to the observed values of other variables in the data.
- Less restrictive than MCAR. More realistic assumption for data to meet.

## MAR Example

- Returning to our previous example with age and income:
  - If the probability that income is missing varies, according to the age of the respondent, but does not vary according to the income of the respondent of a given age, then the data are MAR.

## Missing data mechanisms

- Not Missing at Random (NMAR)
  - Missing data mechanism is NMAR if MAR assumption cannot be met, the missing data mechanism must be modeled to get good parameter estimates
  - NMAR is a weakest assumption
  - NMAR is **nonignorable** (NI)

Effective estimation for NMAR requires very good prior knowledge about missing data mechanism

- Data contain no information about what model would be appropriate
- No way to test goodness of fit of missing data model
- Results often very sensitive to choice of model

## NMAR

- Missingness depends on the values of the missing, or unobserved variables.
- Pattern is non-random, non-ignorable, and arises due to the variable on which the data is missing.
- Not amenable to common handling techniques.

## NMAR Examples

- Returning to our previous example with age and income:
  - If the probability that income is recorded varies according to income within each age group (e.g., wealthy and poor subjects are less likely to answer the income question regardless of age), then the data are neither MAR nor MCAR, hence NMAR (or NI).
- If we are studying mental health and people have been diagnosed as depressed are less likely than others to report their mental status.

## Missing data mechanisms

- Ignorable
  - MCAR
  - In practice, “MAR” and “ignorable” are used interchangeably.
  - If missing data are ignorable, no need to model the missing data process.
  - Any general purpose method for handling missing data must assume that the missing data mechanism is ignorable.
  - The term “ignorable” does not mean that we do not have to worry about missing data, but that we do not have to model the missingness mechanism as part of the estimation process.

## Mechanisms summary

Mechanism		Can predict MD with
Missing Completely at Random	MCAR	- -
Missing at Random	MAR	$Y_{\text{obs}}$
Missing Not at Random	NMAR/NI	$Y_{\text{obs}}$ & $Y_{\text{miss}}$

## Methods for dealing with missing data

## **Solution to missing data**

- Prevent missing data in the first place; have strategy to obtain complete data.
- When missing data occur, consider the mechanism.
- Pragmatic solutions
  - Clear coding to distinguish missingness from inappropriate and other meaningful responses
  - Measure covariates

## **Goals of addressing MD**

- Represent not create
  - Not seeking optimal point prediction, but valid statistical inference (Rubin, 1996)
    - Properly reflect uncertainty
    - Preserve important aspects of data distributions
    - Preserve important relationships
- When implemented properly, some methods allow complete-data analytic techniques to be used, as if the data were never missing.

## Goals for handling MD

- Minimize bias
- Maximize use of available information
- Get good estimates of uncertainty; get appropriate standard error for hypothesis testing

## Methods

- Categories of methods
  - Traditional methods (available-case)
    - Listwise deletion
    - Pairwise deletion
  - Single imputation (deterministic imputation)
  - Advanced methods
    - Maximum likelihood methods
    - Multiple imputation
- A few other points
  - MD handling methods are not all equally good, and none are universally good
  - Most assume
    - Multivariate normal distribution
    - Without interactions or non-linear relationships
    - Independently drawn cases

## Traditional methods

### Listwise Deletion (LD)

- LD deletes the entire record which has missing data for any variable used in a particular analysis (thus, LD is also called complete-case analysis).
- This approach is implemented as the default in software packages such as SAS and SPSS.
- Strengths
  - Easy to implement
  - Works for any kind of statistical analysis
  - Does not introduce bias in parameter estimates if MCAR
  - Standard errors are appropriate; standard error may be large because loss of sample size, but are still appropriate
- Weakness
  - May introduce bias if MAR but not MCAR
  - Potential loss of power due to smaller sample size

## Traditional methods

### Pairwise Deletion (PD)

- Calculates each step of the analysis separately using the cases that have data available for that step. Therefore, a case with data missing on one variable will be used only in steps that do not involve that variable.
- For example, you can create a missing data correlation (or covariance) matrix – the  $r$ 's between all possible pairs of variables are calculated using complete cases for each pair of variables. Then use this matrix for subsequent analyses.
- Strengths
  - Approximately unbiased if MCAR
  - Appears to use all available information
- Weakness
  - Standard errors incorrect
  - May break down (i.e., covariance matrix not positive definite)
  - May be less efficient than LD in some cases

## Single imputation methods

- Any method that substitutes an estimated value for missing value. For example:
  - Replace missing values with means on that variable.
  - Regression imputation (i.e., replace with conditional means - the predicted value obtained by regressing the missing variable on other variables).
  - Stochastic regression: The predicted value from a regression plus a random residual value.
  - Hot deck: Divide sample into homogeneous strata on observed variables. Within each stratum pick “donor” units with observed values to fill in missing values for other units.
  - Interpolation and extrapolation: An estimated value from other observations from the same individual.
- Single imputation is popular because it is conceptually simple and because the resulting sample has the same number of observations as the full data set.
- It can be very tempting when complete-case analysis eliminates a large proportion of the data set.

## Historical Note: Hot deck imputation

- Hot deck imputation goes back over 50 years and was used by the Census Bureau and others.
- Example: Suppose that in the 1950 census a young black male resident of block X was not available or refused to participate. The census bureau would simply take a stack of Hollerith cards (known as “IBM” cards) that came from young black males in census block X, reach in the pile, and pull one out. That card was substituted for missing card in the analysis.

## Single imputation methods

- Problems
  - Often leads to biased parameter estimates (unless the data are MCAR) – The bias is often worse than with complete-case analysis, especially for mean imputation
  - Usually leads to standard errors that are biased downward (underestimate standard errors).
    - Treats imputed values as the true values, ignores variability in imputation. Since the imputed observations are themselves estimates, their values have corresponding random error.
    - The extra source of error is ignored, resulting in too-small standard errors and too-small  $p$ -values.
- Although single imputation is conceptually simple, it is usually difficult to do well in practice. Therefore, these imputation methods are not satisfactory in most circumstances.

## Advanced methods

- Maximum likelihood estimation
  - Produces point estimates
- Multiple random imputation
  - Produces point estimates and measure of uncertainty (i.e., confidence limits)

## ML with ignorable missing data

- **Maximum likelihood estimation 101**
- Let  $p(y|\theta)$  be the probability of observing  $y$ , given  $\theta$ . For a sample of  $n$  independent observations, the likelihood function is

$$L(\theta|\mathbf{y}) = \prod_{i=1}^n p(y_i|\theta)$$

- We then seek an estimate of the parameter of interest, in this case  $\theta$ , that maximizes this function.
- Properties of MLEs:
  - Consistent (implies approximately unbiased in large samples)
  - Asymptotically efficient
  - Asymptotically normal

## ML with ignorable missing data

- Now suppose we have two variables  $X$  and  $Y$ , and there is ignorable missing data on  $X$ . Let  $p(x,y|\theta)$  be the joint probability function. The likelihood for the entire sample with  $m$  complete cases is

$$L(\theta|\mathbf{y},\mathbf{x}) = \prod_{i=1}^m p(x_i, y_i|\theta) \prod_{i=m+1}^n g(y_i|\theta)$$

- This likelihood may be maximized like any other.

## ML with ignorable missing data

- Example:

	Scaring		Data missing for 10 Drug and 15 Placebo
	Yes (1)	No (2)	
Drug (1)	38	55	
Placebo (2)	32	62	

The parameters are  $p_{11}, p_{12}, p_{21}, p_{22}$ . If we exclude cases with missing values, the likelihood is

$$(p_{11})^{38}(p_{12})^{55}(p_{21})^{32}(p_{22})^{62}$$

If we allow for missing data, the likelihood is

$$\left[ (p_{11})^{38}(p_{12})^{55}(p_{21})^{32}(p_{22})^{62} \right] \left[ (p_{11} + p_{12})^{10}(p_{21} + p_{22})^{15} \right]$$

## ML for multivariate normal data

- Multivariate normality implies
  - All variables are normally distributed
  - All conditional expectation functions are linear
  - All conditional variance functions are homogenous
- Several ways to get ML estimates with missing data
  - Factoring the likelihood for monotone missing data
  - EM algorithm
  - Direct maximization of the likelihood

## ML for multivariate normal data

- EM algorithm (two-step procedure)
  1. Expectation (E): Find the expected value of the log-likelihood for the observed data, based on current parameter values; missing values are replaced with the conditional expectation of the missing data given the observed data and initial estimate of the covariance matrix.
  2. Maximize (M): Maximize the expected likelihood to get new parameter estimates; complete-data ML estimation problem.
  3. Repeat 1 and 2 until convergence. Log-likelihood must increase at each step.
- SPSS implements the EM algorithm in its Missing Value Analysis add-on module; SAS can also do EM in its PROC MI procedure.

## ML for multivariate normal data

- Direct ML (i.e., full information ML)
  - Directly maximize the likelihood for the specified model
  - With no missing values, the likelihood for multinormal data is
$$L(\mu, \Sigma) = \prod_i f(y_i | \mu, \Sigma)$$
  - With missing values, the likelihood becomes
$$L(\mu, \Sigma) = \prod_i f(y_i | \mu_i, \Sigma_i)$$
  - If data are missing for individual  $i$ , then  $y_i$  deletes the missing values,  $\mu_i$  deletes the corresponding means, and  $\Sigma_i$  deletes the corresponding rows and columns. The likelihood can be maximized by conventional methods.
- Analysis of the full, incomplete data set using maximum likelihood estimation is available in AMOS. AMOS is a structural equation modeling package, but it can run multiple linear regression models.

## ML for multivariate normal data

- Limitations of ML with Missing Data
  - Requires estimation of a model for the joint distribution of all the variables
    - Often only interested in conditional distributions
    - May not be robust
  - Models and software may not be readily available
    - Good software and/or models for linear and log-linear models
    - But very limited for Cox regression, Poisson regression or logistic regression (with continuous predictors)

## Multiple imputation

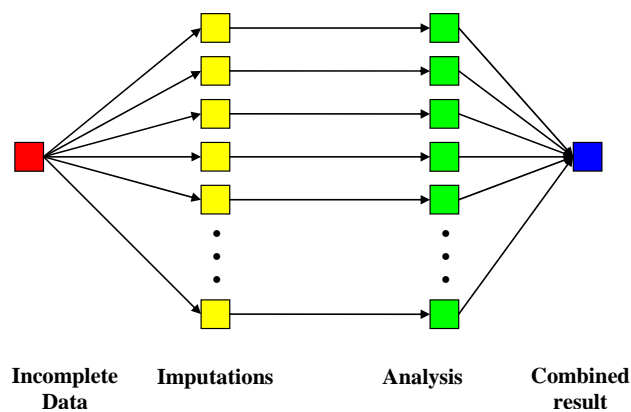
- Multiple imputation (MI) fills in estimates for the missing data.
- However, to capture the uncertainty in those estimates, MI imputes the values multiple times. Because it uses an imputation method with error built in, the multiple estimates should be similar, but not identical.
- The result is multiple data sets with identical values for all of the non-missing values and slightly different values for the imputed values in each data set.
- The statistical analysis of interest, such as ANOVA or logistic regression, is performed separately on each data set, and the results are then combined.
- Because of the variation in the imputed values, there should also be variation in the parameter estimates, leading to appropriate estimates of standard errors and appropriate  $p$ -values.

## Multiple imputation

- Step 1. Impute missing values using an appropriate model that incorporates random variation. Do  $m$ -times to create  $m$  complete data sets. Each data set will have slightly different values for the imputed data because of the random component.
- Step 2. Analyze the each  $m$  completed data sets using standard complete-data methods. Each set of parameter estimates will differ slightly because the data differs slightly.
- Step 3. The last step is to integrate or combine  $m$  analysis to get single result. This involves averaging the values of the parameter estimates across the  $m$ -samples to produce a single point estimate and variance.

## Multiple imputation

General Idea:



## Multiple imputation

- Let  $m$  = the number of data sets imputed and analyzed,

$\hat{Q}_i$  = estimate of the parameter from the  $i^{\text{th}}$  set

$\hat{u}_i$  = variance estimate of  $Q$  from the  $i^{\text{th}}$  set

- The point estimate from the multiple imputations is the average of the estimates from  $m$ -analysis and is given by:

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

## Multiple imputation

The total variance estimate of the point estimate is the sum of within imputation variance

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

and between imputation variance

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

Total Variance:  $T = \frac{1}{m} \sum_{i=1}^m \hat{u}_i + \frac{m+1}{m} \left( \frac{1}{m-1} \sum_{i=1}^m (Q_i - \bar{Q})^2 \right)$

## Combining inferences from MI

MI Inferences are based on the approximation

$$T^{-1/2} (Q - \bar{Q}) \sim t_{df}$$

where

$$df = (m-1) \left[ 1 + \frac{\bar{U}}{(1+m^{-1})B} \right]^2$$

and

$$\bar{Q} \pm t_{df, 1-\alpha/2} \sqrt{T}$$

is the 100(1- $\alpha$ )% confidence interval about Q.

## MI efficiency

Rubin (1987) showed that the efficiency of an estimate based on  $m$  imputations is approximately equal to

$$(1 + \gamma m^{-1})^{-1}$$

where

$$\gamma = \frac{r + 2/(df + 3)}{r + 1}$$

$$r = \frac{T - \bar{U}}{\bar{U}} = \frac{(1 + m^{-1})B}{\bar{U}}$$

$\gamma$  = estimate of the fraction of missing information about Q

$r$  = relative increase in variance due to nonresponse

## MI efficiency

Number of imputations needed: Usually, only 3-10 imputations may be needed.

	$\gamma$				
<b><i>m</i></b>	<b>0.1</b>	<b>0.3</b>	<b>0.5</b>	<b>0.7</b>	<b>0.9</b>
<b>3</b>	<b>97</b>	<b>91</b>	<b>86</b>	<b>81</b>	<b>77</b>
<b>5</b>	<b>98</b>	<b>94</b>	<b>91</b>	<b>88</b>	<b>85</b>
<b>10</b>	<b>99</b>	<b>97</b>	<b>95</b>	<b>93</b>	<b>92</b>
<b>20</b>	<b>100</b>	<b>99</b>	<b>98</b>	<b>97</b>	<b>96</b>

From Schafer's Webpage

## MI efficiency

### The Bottom Line

- Many are surprised by the claim that only 3-10 imputations may be needed.
- Unless the rate of missing information is very high, in most situations there is simply little advantage to producing and analyzing more than a few imputed datasets.

## **Advantages of MI**

- Shafer (1997) has shown through simulation study that MI is generally robust to departures from normality and to model misspecification when the amounts of missing data are not large.
- The performance of multiple imputation in a variety of missing data situations has been well studied and it has been shown to perform favorably (Graham et al., 1997; Graham & Schafer, 1999; Schafer & Graham, 2002).

## **Advantages of MI From McCleary, 2002**

- Introduces appropriate random error
  - makes it possible to get approximately unbiased estimates of all parameters in question
  - more valid than ad hoc approaches to missing data
- Repeated imputation allows for reliable estimates of standard errors.
- Uses all available data
  - preserves sample size
  - maintains statistical power

## **Advantages of MI From McCleary, 2002**

- It can be used with any kind of data and a variety of analyses.
- Hence, multiple imputation is an attractive choice as a solution to missing data problems since it represents a good balance between quality of results and ease of use.

## **Complications in MI**

- Interactions and nonlinearities in MI
  - MI is very good in estimating the main effects of the variables with missing data.
  - May not be so good for estimating interaction effects.

## Software

- Several software packages are available for performing multiple imputation. Below are just a few summaries:
  - Schafer's free, stand-alone windows program **NORM**. **NORM** is a Windows program for multiple imputation of incomplete multivariate data and is an excellent companion to his book.
    - <http://www.stat.psu.edu/~jls/>
  - As of version 17, **SPSS** offers multiple imputation of missing values for both categorical and continuous variables. This feature is part of the **Missing Values Analysis** add-on module.
  - **SAS** has two procedures, **PROC MI** and **PROC MIANALYZE**. It provides three methods for creating the imputed data sets: the regression, the propensity score, and Markov Chain Monte Carlo (MCMC) methods. These imputed data sets can be analyzed with appropriate standard procedures, and then MIANALYZE procedure can be used to combine the results.

## One additional note

- As with ML estimation, multiple imputation requires that the missing data mechanism is ignorable.
- If the mechanism is ignorable, resulting estimates (i.e., regression parameters and standard errors) will be unbiased with no loss of power.
- But what if the cause of missingness is not MAR? Should these methods be used when MAR assumptions are not met?
  - **YES! These Methods Work!**
- Multiple causes of missingness
  - Only small part of missingness may be NMAR.
- MAR & NMAR are widely misunderstood concepts
- John Graham argues
  - that **the** cause of missingness is **never** purely MNAR
  - and that **the** cause of missingness is virtually never purely MAR either

## One additional note

- MAR and MNAR form a continuum
- Pure MAR and pure MNAR are just theoretical concepts
  - Neither occurs in the real world
- MAR vs. MNAR ***NOT*** dimension of interest
- All missing data situations are partly MAR and partly MNAR
- Sometimes it matters ...
  - bias affects statistical conclusions
- Often it does not matter
  - bias has minimal effects on statistical conclusions
- Methods designed to handle NMAR missingness are **NOT** always better than MAR methods.

## PROC MI and PROC MIANALYZE

## Overview

- **Multiple imputation** is a strategy for dealing with data sets with **missing values**.
- You replace each missing value with a **set of plausible values** that represent the **uncertainty** about the right value to impute.
- You create **multiple imputed data sets**, analyze them with **standard analyses**, and then **combine the results**. You produce valid statistical inferences that properly reflect the uncertainty due to the missing values.

## Overview

- **PROC MI** creates multiply imputed data sets for incomplete  $p$ -dimensional multivariate data.
- It offers **three methods** for creating the imputed data sets:
  - the **regression method**
  - the **propensity score method**
  - the **Markov Chain Monte Carlo (MCMC)** method  
(**Note:** We could spend another couple of seminars on these methods.)
- The procedure creates an **output data set** containing  $m$  imputed versions of the original data. In each version, the missing values are replaced with imputed values.

## Overview

- After analyzing your imputed data with standard procedures, you use **PROC MIANALYZE** to combine the results.
- The **MI** procedure was introduced in Release 8.1. Among others, a new **TRANSFORM** statement enables you to transform variables before imputation and back-transform these variables before combining inferences and creating output data sets (it is production in Release 9.0).
- As mentioned earlier, the basic assumption is missing at random (MAR).

## Overview

- The most generally applicable imputation method available in PROC MI is the MCMC algorithm which is based on the multivariate normal model.
- We will use this procedure in our example (linear regression model with 2 predictors) – all predictors are continuous.
- For categorical predictors, see “Imputation of Categorical Variables with PROC MI” by Allison.  
<http://www2.sas.com/proceedings/sugi30/113-30.pdf>

## Getting Started with PROC MI

The **Fitness** data set has been altered to contain an arbitrary pattern of missingness.

```
*----- 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.  |
*-----*
```

Assume the data are **multivariate normally distributed** and the missing data are missing at random (**MAR**).

## Getting Started with PROC MI

```
data FitMiss;
  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
;
```

## Getting Started with PROC MI

```
proc mi data=FitMiss seed=501213 mu0=50 10 180 out=outmi;  
  var Oxygen RunTime RunPulse;  
run;  
  
proc print data=outmi (obs=10);  
  title 'First 10 Observations of the Imputed Data Set';  
run;
```

- By default, the procedure uses the Markov Chain Monte Carlo (**MCMC**) method with a single chain to create 5 imputations and specifies initial mean and covariance estimates calculated by EM algorithm.
- In a Markov chain, the information in the current iteration influences the state of the next iteration. The **MI** procedure takes **200 burn-in iterations** before the first imputation (to eliminate the series of dependence on the starting value of the chain and to achieve the stationary distribution) and **100 iterations between imputations** (to eliminate the series of dependence between the two imputations).
- This reflected in the **Model Information** table of the output.

## Getting Started with PROC MI

### Model Information

Data Set	WORK.FITMISS
Method	MCMC
Multiple Imputation Chain	Single Chain
Initial Estimates for MCMC	EM Posterior Mode
Start	Starting Value
Prior	Jeffreys
Number of Imputations	5
Number of Burn-in Iterations	200
Number of Iterations	100
Seed for random number generator	501213

A summary of the missing data patterns is then given in the **Missing Data Patterns** table (next page).

# Getting Started with PROC MI

## Missing Data Patterns

Group	Oxygen	Time	Pulse	Freq	Percent
1	X	X	X	21	67.74
2	X	X	.	4	12.90
3	X	.	.	3	9.68
4	.	X	X	1	3.23
5	.	X	.	2	6.45

## Missing Data Patterns

-----Group Means-----

Group	Oxygen	RunTime	RunPulse
1	46.353810	10.809524	171.666667
2	47.109500	10.137500	.
3	52.461667	.	.
4	.	11.950000	176.000000
5	.	9.885000	.

# Getting Started with PROC MI

The **Parameter Estimates** table summarizes the descriptive statistics for the imputed data sets.

## Parameter Estimates

Variable	Mean	Std Error	95% Confidence Limits		DF
Oxygen	47.094040	1.011116	45.0139	49.1742	25.549
RunTime	10.572073	0.255870	10.0477	11.0964	27.721
RunPulse	171.787793	2.091776	167.3478	176.2278	15.753

## Parameter Estimates

Variable	Minimum	Maximum	Mu0	t for H0: Mean=Mu0	Pr >  t
Oxygen	46.783898	47.395550	50.000000	-2.87	0.0081
RunTime	10.526392	10.599616	10.000000	2.24	0.0336
RunPulse	170.774818	173.122002	180.000000	-3.93	0.0012

## Getting Started with PROC MI

A listing of the first 10 observations of the imputed values show they have a different precision than the original data. This can be corrected by using the **ROUND=** option, e.g., **ROUND=0.001 0.01 0.1**; would be appropriate here.

Obs	_Imputation_	Oxygen	RunTime	Run Pulse
1	1	44.6090	11.3700	178.000
2	1	45.3130	10.0700	185.000
3	1	54.2970	8.6500	156.000
4	1	59.5710	8.0747	155.925
5	1	49.8740	9.2200	176.837
6	1	44.8110	11.6300	176.000
7	1	42.8857	11.9500	176.000
8	1	46.9992	10.8500	173.099
9	1	39.4420	13.0800	174.000
10	1	60.0550	8.6300	170.000

## PROC MIANALYZE: Overview

- **PROC MIANALYZE** combines the results of the analyses of imputations and generates valid statistical inferences.
- Typical sequence of steps:
  1. Run **PROC MI** to generate a data set with *m* sets of imputed values.
  2. Run a statistical procedure, e.g., **PROC REG**, using **BY \_IMPUTATION\_**; and create an output data set with the combined results, e.g., *m* sets of regression estimates.
  3. Run **PROC MIANALYZE** using the output data set from Step 2 as input.

## PROC MI & MIANALYZE: Code

```
proc mi data=FitMiss noprint out=outmi seed=3237851;
  var Oxygen RunTime RunPulse;
run;

proc reg data=outmi outest=outreg covout noprint;
  model Oxygen = RunTime RunPulse;
  by _Imputation_;
run;

proc print data=outreg(obs=8);
  var _Imputation_ _Type_ _Name_
      Intercept RunTime RunPulse;
  title 'Parameter Estimates from Imputed Data Sets';
run;

proc mianalyze data=outreg;
  modeleffects Intercept RunTime RunPulse;
run;
```

## PROC MIANALYZE: Results

Parameter Estimates from Imputed Data Sets

Obs	_Imputation_	_TYPE_	_NAME_	Intercept	RunTime	RunPulse
1	1	PARMS		86.544	-2.82231	-0.05873
2	1	COV	Intercept	100.145	-0.53519	-0.55077
3	1	COV	RunTime	-0.535	0.10774	-0.00345
4	1	COV	RunPulse	-0.551	-0.00345	0.00343
5	2	PARMS		83.021	-3.00023	-0.02491
6	2	COV	Intercept	79.032	-0.66765	-0.41918
7	2	COV	RunTime	-0.668	0.11456	-0.00313
8	2	COV	RunPulse	-0.419	-0.00313	0.00264

# PROC MIANALYZE: Results

The MIANALYZE Procedure

## Model Information

Data Set WORK.OUTREG  
Number of Imputations 5

## Variance Information

Parameter	-----Variance-----			DF
	Between	Within	Total	
Intercept	45.529229	76.543614	131.178689	23.059
RunTime	0.019390	0.106220	0.129487	123.88
RunPulse	0.001007	0.002537	0.003746	38.419

## Variance Information

Parameter	Relative Increase in Variance	Fraction Missing Information	Relative Efficiency
Intercept	0.713777	0.461277	0.915537
RunTime	0.219051	0.192620	0.962905
RunPulse	0.476384	0.355376	0.9336410

# PROC MIANALYZE: Results

## Parameter Estimates

Parameter	Estimate	Std Error	95% Confidence Limits		DF
Intercept	90.837440	11.453327	67.14779	114.5271	23.059
RunTime	-3.032870	0.359844	-3.74511	-2.3206	123.88
RunPulse	-0.068578	0.061204	-0.19243	0.0553	38.419

## PROC MIANALYZE: Results

Parameter Estimates					
Parameter	Minimum	Maximum	Theta0	t for H0: Parameter=Theta0	Pr >  t
Intercept	83.020730	100.839807	0	7.93	<.0001
RunTime	-3.204426	-2.822311	0	-8.43	<.0001
RunPulse	-0.112840	-0.024910	0	-1.12	0.2695

## For more information

### Visit the SAS website:

General information on Multiple Imputation is SAS:

<http://support.sas.com/rnd/app/da/new/dami.html>

SUGI paper by SAS developer:

<http://support.sas.com/rnd/app/papers/multipleimputation.pdf>

SAS Documentation on MI & MIANALYZE

<http://support.sas.com/rnd/app/papers/miv802.pdf>

<http://support.sas.com/rnd/app/papers/mianalyze.pdf>

## Recommendations

- Proportion of missing obs < 5%
  - Any imputing method probably OK
- Proportion of missing obs 5-10%
  - Substitute with constant (e.g., mean) if low correlation to other variables
  - Single imputation OK
  - Multiple imputation better

## Recommendations

- Proportion of missing data > 15%
  - Use multiple imputation
  - Caution that imputation might be ineffective
- Reason for missingness is more important than number of missing values.

## Resources (Literature)

- Allison (2002). *Missing Data*. Sage Publication.
- Horton and Lipsitz. (2001). Multiple imputation in practice: Comparison of software packages for regression models with missing variables. *American Statistician*, 55, 244-254.
- Little & Rubin (2002). *Statistical analysis with missing data*, 2<sup>nd</sup> edition. Wiley-Interscience.
- McKnight et al. (2007). *Missing Data: A Gentle Introduction*. The Guilford Press.
- Rubin (1996). Multiple imputation after 18+ years. *Journal of American Statistical Association* 91: 473-489.
- Schafer and Graham. (2002) Missing data: Our view of the state of the art. *Psychological Methods*, 7, 147-177.

## Resources (Websites)

- **Joe Shafer**  
<http://www.stat.psu.edu/~jls/index.html>
- **Paul Allison**  
<http://www.ssc.upenn.edu/~allison/>
- **General Info on Multiple Imputation**  
<http://www.multiple-imputation.com>