Missing data and multiple imputation: A conceptual introduction

Steve Gregorich

CADC Scholars Meeting
Jan 13, 2009

About this talk

Introduction to concepts

Not an introduction to algorithms or computer programs

Intended to be non-technical, so I skirt many technical issues

This is a vast topic, so I'll skip some basic issues as well

Issues addressed

. Missing values

. Causes of missing values

. What to do about missing values?

. What is multiple imputation

. 3 steps in using multiple imputation
  1. Create multiple, plausible imputed data sets
  2. Fit substantive model to each imputed data set
  3. Combine results across substantive models

. Examples of multiple imputation

. Virtues of MI / Why use multiple imputation?

. What if assumptions are violated?

. Statistical alchemy?

. Reasonable Goals for analysis of incomplete data

Missing values

Data from surveys, experiments and observational studies
typically include missing values

To be a missing value...

. An <u>underlying</u> value must exist, and that value is truly unknown

The question/variable is applicable to the respondent

A legitimate value/response exists, but is unobserved

Missing values

General types of missing values

. Item-non response

. Unit non-response

Not observed at one or more waves in a longitudinal study, but may return

. Attrition—gone forever

Issues addressed

. Missing values

. Causes of missing values

. What to do about missing values?

. What is multiple imputation

. 3 steps in using multiple imputation

1. Create multiple, plausible imputed data sets
2. Fit substantive model to each imputed data set
3. Combine results across substantive models

. Examples of multiple imputation

. Virtues of MI/ Why use multiple imputation?

. What if assumptions are violated?

. Statistical alchemy?

. Reasonable Goals for analysis of incomplete data

Causes of missing values / Missing-data mechanisms

MCAR—Missing Completely at Random

This means that missing values are random events

Missingness is not related to anything

. Probability of missingness is unrelated to the unobserved values
. Whether or not X is observed does not depend on its true value

. Missingness is not related to values of any other variables

Causes of missing values / Missing-data mechanisms

MCAR—Missing Completely at Random

Examples

. Random events such as administrative error or computer crash

. Missingness by design

Points

. Strong assumption

. Often unrealistic unless MCAR is by design

. Assumption *can* be tested

Causes of missing values / Missing-data mechanisms

MAR—Missing at Random

Not to be confused with MCAR

. Missing values are random events, <u>conditional</u> on <u>observed</u> data

. Probability of missingness may depend upon observed data values,
but does not depend upon data values that are missing

Causes of missing values / Missing-data mechanisms

MAR—Missing at Random

<u>Made-up examples</u>

. The probability of missing income values depends on respondent sex,
but within each sex, the probability of missing income values
is unrelated to actual income

. In longitudinal study, participants drop out for reasons that depend upon
past recorded (and modeled) responses,
but not current or future responses

Causes of missing values / Missing-data mechanisms

MAR—Missing at Random

Worked, made-up example

100 men and 100 women are sampled and asked.

"How important is it to have an annual physical exam?"

Response options: Important / Not Important

Causes of missing values / Missing-data mechanisms

MAR—Missing at Random

Worked, made-up example: RESULTS:

. Women were more likely to report that annual exams are important

. Women were also more likely to have missing responses

. Within respondent sex, missingness is unrelated to agreement with the item

	100% observed data		incomplete data (MAR)	
	observed *N*	% 'important'	observed *N*	% 'important'
Men	100	50%	90	50%
Women	100	80%	60	80%
Total	200	65%	150	62%

What if missingness was determined by a continuous variable?

Causes of missing values / Missing-data mechanisms

MAR—Missing at Random

Strictly speaking, for the MAR assumption to hold, a variable representing the missingness mechanism must be completely observed *and* appropriately modeled

. To speak of a single missingness mechanism is often misleading—data values may be missing for a variety of reasons

. Example: Drop-out in a school-based sample may result from

(1) students moving out of the area (MCAR?),

(2) dropping out of school,

(3) substance use, etc.

Causes of missing values / Missing-data mechanisms

MAR—Missing at Random

Points

. MAR assumption is milder than MCAR

. MAR assumption cannot be tested

. MAR assumption may be met to varying degrees, not *all-or-none*

Causes of missing values / Missing-data mechanisms

NMAR—Not Missing at Random

The most difficult circumstance

. Probability of missingness depends upon quantities that are unobserved

Causes of missing values / Missing-data mechanisms

NMAR—Not Missing at Random

. Probability of missingness on *X* depends on the missing *X* values themselves, e.g.,

Income

In a study of substance use, substance users may more often skip measurement sessions because of their drug use

. Probability of missingness might depend on some other variable that is not observed or is not modeled

Causes of missing values / Missing-data mechanisms

NMAR—Not Missing at Random

Points

. Difficult to accommodate statistically

. NMAR assumption cannot be tested

Causes of missing values / Missing-data mechanisms

Initial Summary

	MCAR	MAR	NMAR
Missingness assumption	Random	Random, conditional on observed data	Systematically related to values that are missing
Assumption testable?	Yes	No	No
Requirements of assumption	Strong	Milder	Mildest
Implementation of modeling	Standard	More difficult	Most difficult *(too dicey?)*

Issues addressed

. Missing values

. Causes of missing values

. What to do about missing values?

. What is multiple imputation

. 3 steps in using multiple imputation
1. Create multiple, plausible imputed data sets
2. Fit substantive model to each imputed data set
3. Combine results across substantive models

. Examples of multiple imputation

. Virtues of MI/ Why use multiple imputation?

. What if assumptions are violated?

. Statistical alchemy?

. Reasonable Goals for analysis of incomplete data

What to do about missing values?

MCAR Modeling Methods—Complete Cases analysis

Same as Casewise (CW) / listwise (LW) deletion

Advantages

. Easy

. Maybe OK if < 5% of cases would be lost due to missing values

Disadvantages

. If data are not MCAR, bias may result

. Inefficient—discarded information

What to do about missing values?

Modeling methods that assume MCAR—Complete Cases analysis

Disadvantage #1: if data are not MCAR, bias may result

Worked, made-up example

N=1000 observations on two variables $X1$ and $X2$

. Mean = 0, Variance = 1, Correlation = .50

Assume

. X1 is completely observed

. . 50% of X2 values are MAR

. Higher values of $X1$ cause a higher probability of missingness on $X2$

	Observed Means	
	$X1$	$X2$
100% data (N=1000)	0	0
Complete cases (N=500)	-0.55	-0.27

What to do about missing values?

Complete Cases. *Disadvantage #2—inefficient*

0	$\boxed{?}$	1	0	1
0	1	0	1	0
1	1	1	0	$\boxed{?}$
0	1	1	0	1
1	0	1	0	1
0	0	0	1	0
0	1	$\boxed{?}$	1	1
0	1	0	0	1
0	1	0	0	1
1	0	1	$\boxed{?}$	1

- 10 cases
- 5 items
- 50 data points--"complete"
- 4 missing data points:
  $<$ 10% missing data points
- Complete cases n=6
  $\rightarrow$ 40% missing cases

What to do about missing values?

MCAR Modeling Methods—Pairwise (PW) deletion

<u>Advantages</u>

. Easy

<u>Disadvantages</u>

. If data are not MCAR, bias may result

. Correlation matrix may be non-positive definite

. There is no simple basis for estimating standard errors

What to do about missing values?

MCAR Modeling Methods—Reweighting

. More refined version of complete cases analysis—

Incomplete cases are removed

The remaining cases are weighted so that they resemble the full sample or the population of interest

<u>Advantages</u>

. Relatively easy

<u>Disadvantages</u>

. If data are not MCAR, bias may result

. Inefficient—discarded information

What to do about missing values?

MCAR Modeling Methods—Single imputation

. Unconditional mean imputation: impute the sample mean
This can lead to biased estimates, even when MCAR holds

There is no basis for estimating standard errors

. Conditional mean imputation: impute the respondent mean
Better, but still problematic

. Single regression imputation
Produces biased variance estimates, even when MCAR holds

There is no basis for estimating standard errors

. An overarching problem is that the imputed values are treated as *known*, so that standard errors and confidence intervals are too liberal (too small)

What to do about missing values?

MAR Modeling Methods—Likelihood methods
(e.g., mixed models, multilevel models, HLM, random coefficient models)

Advantages
Easy
Assume that missing outcomes are MAR

Disadvantages
. Can be inefficient
Cases with missing explanatory variables are dropped

Issues addressed

. Missing values

. Causes of missing values

. What to do about missing values?

. What is multiple imputation

. 3 steps in using multiple imputation
1. Create multiple, plausible imputed data sets
2. Fit substantive model to each imputed data set
3. Combine results across substantive models

. Examples of multiple imputation

. Virtues of MI/ Why use multiple imputation?

. What if assumptions are violated?

. Statistical alchemy?

. Reasonable Goals for analysis of incomplete data

What is multiple imputation?

MAR Modeling Methods—Multiple imputation

3 Steps in using multiple imputation

1. Create multiple, plausible imputed data sets
   The imputation model

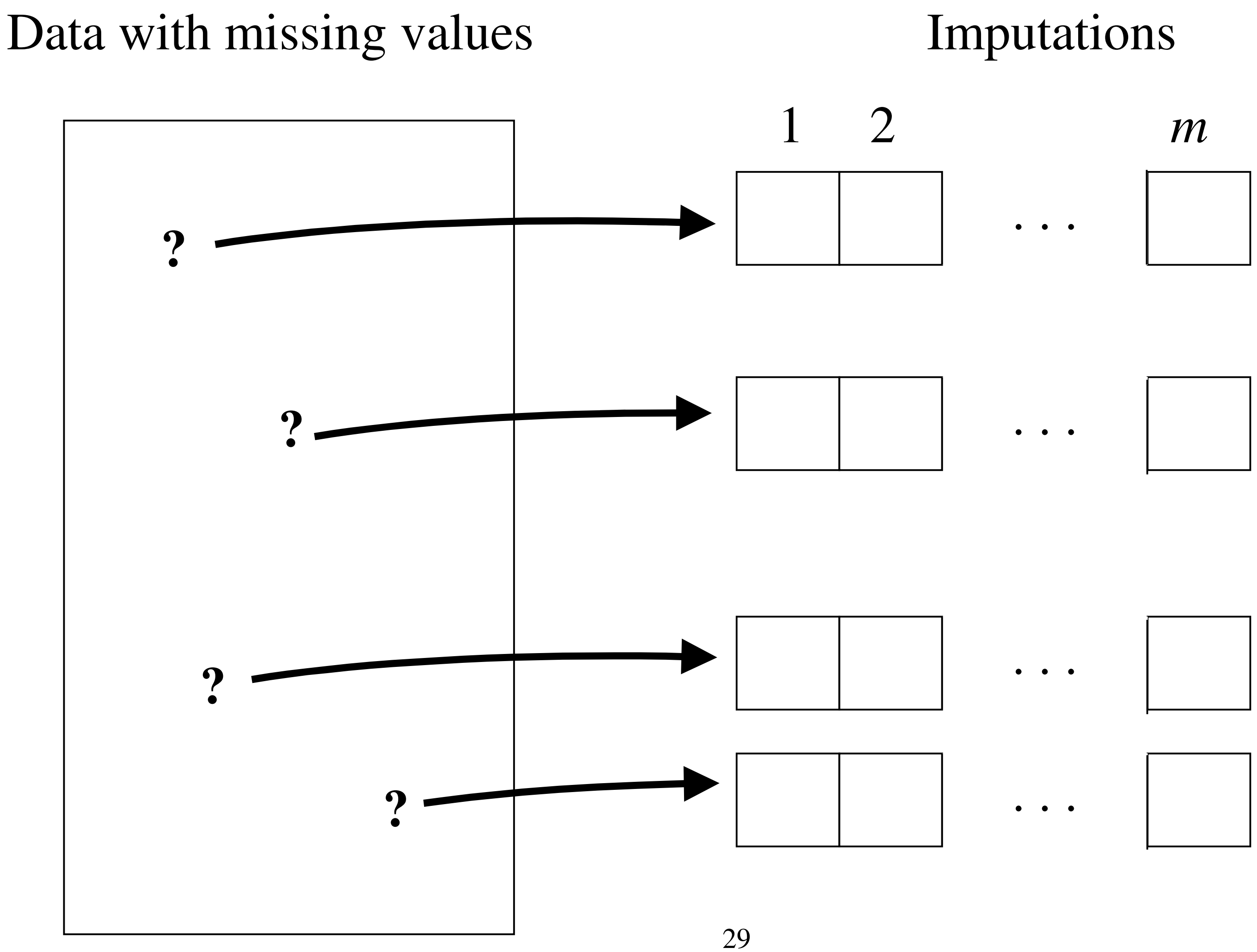
2. Fit the substantive model to each imputed data set

3. Combine results across substantive models

What is multiple imputation?

MAR Modeling Methods—Multiple imputation

Step 1. Create multiple, plausible imputed data sets



What is multiple imputation?

MAR Modeling Methods—Multiple imputation

Step 1. Create multiple, plausible imputed data sets by fitting a 'multiple imputation model'

There are several more general algorithms, e.g.,

. Markov Chain Monte Carlo (MCMC) or Data Augmentation (DA)

. Hot Deck

. Sampling importance/resampling (SIR)

Most general algorithms

. Assume the data are MAR and

. Allow for complex patterns of missing data

What is multiple imputation?

Step 2. Fit 'substantive model' to each imputed data set

Save parameter and standard error estimates

What is multiple imputation?

MAR Modeling Methods—Multiple imputation

Step 3. Combine results across imputed data sets

Average corresponding parameter estimates across imputed data sets

These are the parameter estimates from multiple imputation

Compute the parameter standard errors.

Formal summarization of the parameter and standard errors estimated from the separate imputed data sets

What is multiple imputation?

MAR Modeling Methods—Multiple imputation

A bit of strategy

The imputation model versus the substantive model

. Variables in the imputation model should be a superset of the substantive model, including any interaction terms and the outcome

. Use a rich imputation model with auxiliary variables.
Good to have more variables in the imputation model than the substantive model

. Good to include variables that are related to missingness
This will help to make the MAR assumption more plausible

What is multiple imputation?

MAR Modeling Methods—Multiple imputation

<u>Advantages relative to methods that assume MCAR</u>
. MAR assumption
. efficiency
. reduced bias

<u>Disadvantages relative to methods that assume MCAR</u>
. more difficult to implement

Issues addressed

. Missing values

. Causes of missing values

. What to do about missing values?

. What is multiple imputation

. 3 steps in using multiple imputation
  1. Create multiple, plausible imputed data sets
  2. Fit substantive model to each imputed data set
  3. Combine results across substantive models

. Examples of multiple imputation

. Virtues of MI / Why use multiple imputation?

. What if assumptions are violated?

. Statistical alchemy?

. Reasonable Goals for analysis of incomplete data

Examples of multiple imputation

Multiple imputation when data are MCAR

<u>Example</u>

N=1000 observations on two variables *X*1 and *X*2
. Mean = 0, Variance = 1, Correlation = .50

Assume
. *X*1 is completely observed
. 50% of *X*2 values are MCAR

	Observed Means	
	X1	X2
100% data (N=1000)	0 (.032)	0 (.032)
Complete cases (N=500)	-0.01 (.046)	-0.01 (.045)
Multiple Imputation	0 (.032)	0 (.043)

Examples of multiple imputation

Multiple imputation when data are MAR

<u>Example</u>

N=1000 observations on two variables X1 and X2

. Mean = 0, Variance = 1, Correlation = .50

Assume

. X1 is completely observed

. 50% of X2 values are MAR

. Higher values of X1 cause a higher probability of missingness on X2

	Observed Means	
	X1	X2
100% data (N=1000)	0 (.032)	0 (.032)
Complete cases (N=500)	-0.55 (.036)	-0.27 (.041)
Multiple Imputation	0 (.032)	0.01 (.047)

Examples of multiple imputation

Multiple imputation when data are NMAR

<u>Example</u>

N=1000 observations on two variables X1 and X2
. Mean = 0, Variance = 1, Correlation = .50

Assume
. X1 is completely observed
. 50% of X2 values are NMAR
. Higher values of X2 cause a higher probability of missingness on X2

	Observed Means	
	X1	X2
100% data (N=1000)	0 (.032)	0 (.032)
Complete cases (N=500)	-0.28 (.041)	-0.56 (.036)
Multiple Imputation	0 (.032)	-0.44 (.033)

Issues addressed

. Missing values

. Causes of missing values

. What to do about missing values?

. What is multiple imputation

. 3 steps in using multiple imputation

1. Create multiple, plausible imputed data sets
2. Fit substantive model to each imputed data set
3. Combine results across substantive models

. Examples of multiple imputation

. Virtues of MI / Why use multiple imputation?

. What if assumptions are violated?

. Statistical alchemy?

. Reasonable Goals for analysis of incomplete data

Virtues of MI/ Why use multiple imputation?

Milder assumption about missingness mechanism than ad hoc methods

Separation of imputation and substantive models
. Large imputation model can make MAR assumption more reasonable

More efficient than ad hoc methods, such as complete cases

Principled basis for estimating standard errors

Can use any analysis technique that is appropriate for complete data

One set of imputed data sets may be used for different substantive models

Can be highly efficient with small numbers of imputed data sets
. I use 20 imputed data sets. In practice that is always sufficient

What if the assumptions of the MI model are violated?

Distributional assumptions

Assumptions about the missingness mechanisms

. What if missing data are not MAR?

. Unless data are MCAR (testable) you'll never really know

. The MAR assumption may not seem plausible in many applications

E.g., drop-out in a longitudinal study may be related to current data values

What to do...

. Avoid unplanned missing data

. Rich imputation model—try to inform about missingness mechanism

. Partially observed mechanisms are helpful

Statistical alchemy?

Isn't multiple imputation just making up the data?

Multiple imputation is nothing other than a way to representing missing data uncertainty

Multiple imputation replaces missing values with plausible values, then averages across that uncertainty

Note that Complete Cases analysis assumes <u>no</u> uncertainty about missing values

Reasonable goals for analysis of incomplete data

The only really good solution to the missing data problem is not to have missing data. (P. Allison)

Make the best inferences using all of the observed data
. not to predict or recover the missing data,
. not to obtain the same results as you would have with complete data

Observed data provide indirect evidence about likely values of unobserved data

Missing values are a source of variability *to be averaged over* (Schafer)