

The Problem of Missing Data:

Analysis of Incomplete  
Observations

CIRA M&B Seminar Series

Marilyn Stolar

20 Feb 2003

# QUIZ

Agree or Disagree:

1. Using complete cases only (ignoring cases with any missing values on any variables) is a bad approach to analyzing data with missing values.

2. If we don't know the values of the data that are missing or why they are missing, then no matter what statistical "trick" we use we cannot obtain valid results from an analysis of the data.

3. The values of a given variable that are missing must be a simple random sample from the complete (but unknown) set of values for that variable in order to obtain any valid results from an analysis of the data involving that variable.

4. There is no acceptable remedy for missing values on variables like “sex” or “race”, especially if they are variables of interest and not just covariates.

# Caveats

- All methods depend on problematic assumptions; can't know if they actually hold (untestable)
- Best solution—reduce missingness up front
- Preferred/principled methods require more “effort”, but are becoming more accessible with recent software enhancements

# Considerations

- Missing is better than false
- Room for error
- Ways to mitigate harm
- Assumptions → sensitivity analysis
  
- Goal is valid and efficient inference, not to estimate each missing value
- We want true coverage of Type I error, and to minimize Type II error
- Avoid non-productive “tricks”

Each method for the analysis of incomplete data has its respective strengths and weaknesses. These depend upon:

- the rate and pattern of missingness
- the type of model for the research question
- the number of variables in the analysis
- the types of variables where missingness occurs
- the underlying missingness mechanism

# Types of Non-response

- Unit non-response
  - All information is missing for sample member (person is missing)
- Item non-response
  - Information is missing on some variables for some persons

*only item non-response will be discussed today*

## Item Non-response

- Missing by design
- Doesn't apply
- Don't know
- Left blank by commission
- Left blank by omission
- Drop-out

# Item Attributes

- Repeatedly measured?
- Discrete or continuous?
- Outcome, explanatory or auxiliary variable?

# Some Statistical Approaches

Older:

- Listwise deletion
- Pairwise deletion
- Dummy variable adjustment
- Selection models
- (Single) Imputation\*

Newer:

- Maximum likelihood
- Multiple imputation
- Pattern-mixture models
- Weighting

## (Single) Imputation\*

- Unconditional mean imputation
- Unconditional distribution imputation (e.g. hot-decking)
- Conditional mean imputation
- Conditional distribution imputation
- Last value carried forward

# Patterns of Non-response

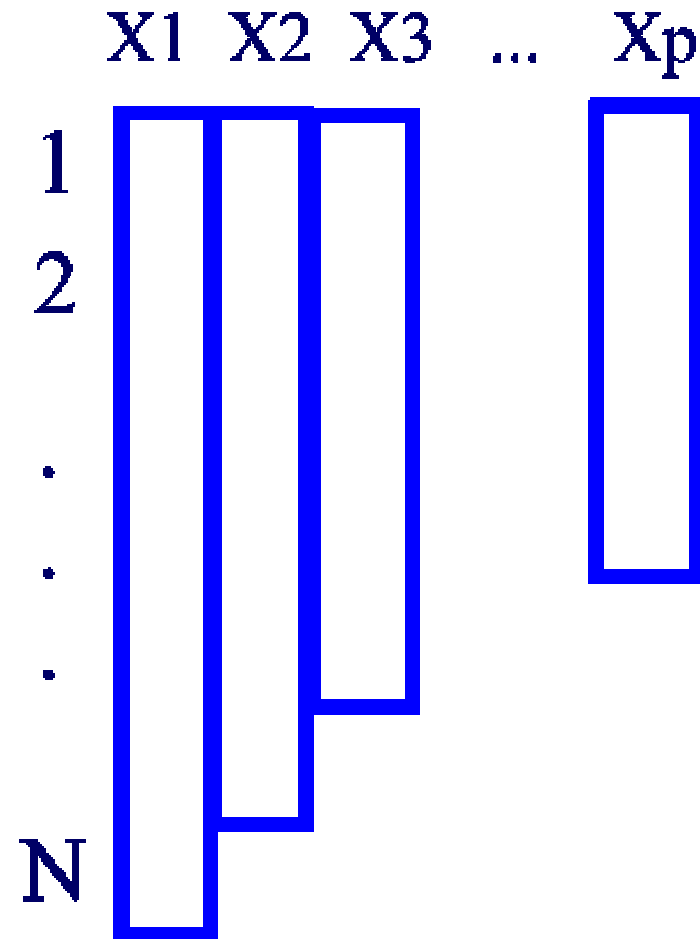
- Univariate Pattern
- Monotonic Pattern
- Arbitrary Pattern

# Univariate Pattern

$X_1$   $X_2$  ...  $X_p$   $Z_1$  ...  $Z_q$

1	observed	observed
2		
.		
.		
.		
N	??	

# Monotonic Pattern



# Arbitrary Pattern

$X_1$   $X_2$   $X_3$  ...  $X_p$

1	?			?
2		?		
.	?		?	
.		?		
N				?

# Missingness Mechanisms

- Missing Completely At Random (MCAR)
- Missing At Random (MAR)
- Missing Not At Random (MNAR)

# Missing Completely At Random (MCAR)

Suppose some data are missing on  $\mathbf{Y}$ .

These data are said to be **MCAR** if the probability that  $\mathbf{Y}$  is missing is unrelated to the missing values of  $\mathbf{Y}$  or other variables  $\mathbf{X}$ .

e.g. simple random sample of  $\mathbf{Y}$ s are missing

- $\Pr(Y \text{ is missing} \mid X, Y) = \Pr(Y \text{ is missing})$
- MCAR is best situation to be in
- Complete data sample is a SRS of full sample (if no missings)  $\rightarrow$  no bias
- Sets of variables may always be missing together completely at random
- The probability that Y is missing may be related to irrelevant Z
- *How do we know?*

# Missing At Random (MAR)

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

e.g. “MCAR within strata”

- $\Pr(Y \text{ is missing} \mid X, Y) = \Pr(Y \text{ is missing} \mid X)$
- Much weaker assumption than MCAR
- Can test whether missingness on Y depends on X
- Cannot test whether missingness on Y depends on value of Y (*we rely on an assumption we cannot test—is it plausible?*)
- Practically the same as “ignorable missingness”, i.e. no need to model the missing data mechanism in the analysis of the data

# Not Missing At Random (NMAR)

- If the MAR assumption is violated, **the missing data mechanism must be modeled** to get good parameter estimates (e.g. Heckman's selection model)
- Requires good prior knowledge about causes of missingness
- Data contain no info on what would be appropriate
- No way to test goodness of fit of missing data model
- Results may be very sensitive to choice of model → sensitivity analysis is critical to data analysis process

# Preliminary Data Analysis

- Missing data pattern
- % missing on each variable
- % of missings per subject
- Missings together
- Correlations between variables
- Other?

# Listwise Deletion

(aka Complete Case Analysis)

- Analyze only those subjects with observed responses to all measures.
- (+) Easy to implement, BUT: Be careful when fitting nested models—when you discard a variable you may gain subjects—different subjects=models no longer nested
- (+) Any model can be fit using standard software
- (+) If MCAR holds, no bias is introduced into parameter estimates; standard errors of estimates are appropriate, but increased

- If doing any kind of regression, missing only on predictors will be robust to MNAR when

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

example: Y=num kids X=income

special case: logistic regression (w/o interactions)

- (-) may lose power, increase Type II error

“Rule of Thumb”: If proportion of subjects with incomplete cases is less than .05, then listwise deletion will be relatively robust

but: you may lose a substantial subgroup of interest

# Pairwise Deletion

## aka Available Case Analysis

- For linear models, means and covariance matrix (moments) determine estimates.

Estimate each moment with all available non-missing cases.

Substitute moment estimates into equations for parameter estimates.

- (+) Approx unbiased if MCAR
- (+) Appears to use all information in data
- (-) Standard errors incorrect

# Pairwise Deletion

- May break down
- May be less efficient than listwise in some cases (High correlations)

# Dummy Variable Adjustment

- Replace all missings on a given predictor variable **X** with a constant (0 or observed mean is good)
- Create a dummy variable **D** for missing vs observed
- Include both **X** and **D** among predictors
- Add a category to a discrete variable to indicate missing
- Produces biased coefficient estimates/redefines the parameters

## (Single) Imputation

- Replace each missing value with a single value
- May yield biased parameter estimates
- Standard error estimates be biased (too small)—ignores uncertainty in replacement values

# Maximum Likelihood

- Choose **parameter** estimates which, if true, would maximize the probability of observing what has, in fact, been observed
- (+) ML estimates have nice statistical properties (consistent, asymptotically efficient and normal)
- (-) Model joint distribution of all variables—data should reasonably adhere to joint distribution that is chosen
- EM vs FIML
- (+/-) Software
- (+/-) Auxiliary information

# Multiple Imputation

- Same properties as ML
- Works with any kind of data or model
- Software
- Captures uncertainty in imputed values in estimates of standard errors
- Get a different result every time you use it
- Complex—many approaches, many steps, many decisions
- Auxiliary variables—*“when in doubt, don’t leave it out”*

# ML versus MI

- In either approach, it is possible to add auxiliary variables solely for the purpose of improving the missing data procedure.
- Simulation studies show that inclusive strategy is greatly preferred—increased efficiency and reduced bias for almost no cost
- ML encourages restrictive strategy, whereas MI makes inclusive strategy cheap and easy

# Software

- Various (Single) Imputation methods:  
SOLAS  
S-plus library Hmisc (Frank Harrell), SPSS
- Maximum Likelihood  
EM Algorithm:  
LEM, LOGLIN for discrete variables  
SAS MI, LISREL, EQS, SPSS  
FIML:  
AMOS, Mplus, Mx

- Multiple Imputation

SOLAS

SAS MI / MIANALYZE

S-plus libraries NORM, CAT, PAN, MIX—Joseph  
Schafer

MICE

# Useful Links:

Joe Schafer's website—freeware S-plus programs  
for performing MI

[www.stat.psu.edu/~jls/](http://www.stat.psu.edu/~jls/)

[www.stat.psu.edu/~jls/mifaq.html](http://www.stat.psu.edu/~jls/mifaq.html)

[www.stat.psu.edu/~jls/misoftwa.html](http://www.stat.psu.edu/~jls/misoftwa.html)

# Useful Links:

Paul Allison's homepage—SAS macros to accompany the MI and MIANALYZE procedures  
[www.ssc.upenn.edu/~allison/](http://www.ssc.upenn.edu/~allison/)

# Useful Links:

MICE Freeware, good list of references

[www.multiple-imputation.com](http://www.multiple-imputation.com)