

# **Treatment of Missing Data in Social Work Research: Methods for Multiple Imputation**

**Roderick A. Rose, M.S.  
Mark W. Fraser, Ph.D.  
School of Social Work  
University of North Carolina at Chapel Hill**

**January 14, 2006  
Society for Social Work and Research Conference  
San Antonio, Tx**

# Introduction

- Our purposes in this presentation are to
  - Review the missing data mechanism
  - Describe a strategy for determining the best method for handling missing data
  - Describe a promising way to develop a model for multiple imputation if that is the chosen strategy.

# Summary of conclusions and recommendations

- Whether the data are missing at random or not is not a relevant concern (because it is not observable)
- Establish an analytical model before developing a strategy
- Determine:
  - Whether imputation is necessary
  - Whether the data will support an imputation
- Watch out for high rates of missingness among the data points

# **Causes, Consequences, and Treatments of Missing Data**

## The Missing Data Mechanism: What Are the Causes for Missing Information?

- The “reason” the data are missing.
- MCAR, MAR, or NMAR (non-ignorable).
- Generally, estimator bias is the consequence.
- However, the consequences vary
- Hence, the treatments differ based on the missing mechanism.
- A single variable can be characterized by any or all of the mechanisms.

# MCAR

- Missing data points are drawn completely at random from unconditional probability distributions.
- No patterns in the data can describe the missingness.
- Consequence: complete randomness = no bias.
- Solution: Use Listwise deletion of cases and assume parameter estimates are unbiased
- Problems: MCAR is unlikely in most scenarios; an untenable assumption<sup>[1]</sup>.

# MAR

- Missing data points are drawn conditionally on observed variables.
- Patterns in the data can describe missingness, BUT only among observed variables.
- E.g., assume the higher the education level, the less likely a respondent was to report income. If education level was observed, then missing information on income may be MAR.

# MAR (ctd.)

- Consequence: depends on strategy; bias is a potential threat that can be eliminated or reduced using the correct strategy.
- Solutions:
  - When only the dependent variable is missing then conventional ML estimates will be unbiased.
  - In nearly all other cases, use multiple imputation: conduct multiple random draws from conditional probability distributions[2].
- *Problem: Cannot be demonstrated from the data-it must be proven that (unobserved) patterns in the missing data **are not associated with missingness.***

# Non-ignorable/NMAR

- Patterns in the unobserved variables can determine their own missingness.
- Missing data points are drawn conditionally on unobserved variables.
- Consequences: bias, bias, bias.
- Solution:
  - Conduct the joint probability modeling of both data and the missingness mechanism.
  - Estimate both analytical and missingness parameters
- Problem:
  - This strategy is probably out of reach for most researchers.

# Here's the good news!!

- Multiple imputation can be used under certain non-ignorable conditions. Yay!!
- Even though MI will not completely eliminate bias it may be able to reduce it.
  - E.g., if income is missing for many respondents, and higher income respondents were less likely to respond, then the missing data are non-ignorable.
  - MI may still produce less-biased estimates if education is used to impute income, under the established finding that education and income are highly correlated.

# Key Citation

- “The crucial assumption made by ignorable methods is not that the propensity to respond is completely unrelated to the missing data, but that this relationship can be explained by data that are observed.”
- Schafer, 1997, p. 27.
- Interpretation:
  - Given that MAR cannot be observed or be assumed to perfectly characterize any missing data, forgo the question of “mechanism”.
  - Instead, focus on reducing bias in potentially non-ignorable situations.

# **How to Handle Missing Data**

# Analytical Model: X and Y

- The missing data mechanism *effectively* depends on characteristics of the model and the data.
  - All analytical variables must be included in an imputation model
  - If a variable associated with missingness is available, the data may be MAR; if not available, the data are non-ignorable.
- The analytical model (model for hypothesis testing) must therefore be established before deciding on a strategy.
- The analytical model contains X and Y variables:
  - X: set of fully-observed variables; no missing data points. *This includes the treatment condition variable.*
  - Y: contain both observed (YOBS) and missing (YMIS) data points.

# Establish a Strategy

- Determine how to handle the missing data in the analytical model--delete, or use MI to impute[3].
- The first requires MCAR.
  - A likelihood ratio test for MCAR exists (Little, 1988).
  - It is extremely unlikely that your data will pass this test.
- We will focus on developing an imputation model in order to reduce bias in a non-ignorable situation.

# Imputation Model

- MI uses information in observed data to develop conditional probability distributions.
  - Multiple random draws from these distributions are made.
  - More information = less bias.
- The data set will likely contain variables that will not be used in the analysis.
- MI literature recommends using *all variables* in imputation model.
  - This may be unrealistic.
  - Selection of appropriate imputation variables may be necessary.

# Define Z

- Z: non-analytical variables that may contain both observed (ZOBS) and missing (ZMIS) data points.
- Z variables are not necessarily required, but using them may reduce bias, i.e. they explain missingness patterns and missing values.
- Preference should be given to Z variables with:
  - Low rates of missingness.
  - High association with missingness of YMIS and YMIS itself.

## Z: Preferred low rates of missingness

- Higher rates of missingness may cause imputation to fail
- Problems of high percentages will be similar to problems of low sample size in analyses.
- Less information is available for the imputation.

# Z: High association with missingness of YMIS

- Observable relationships between Z variables and the missingness patterns in YMIS can help establish an imputation model if MI is chosen[4].
- Use bivariate and/or multivariate exploratory methods.
- Multivariate methods useful in limited situations where there are Z variables with low missingness and only a few YMIS variables.
- Use bivariate instead (pairwise deletion).

# Example: Logistic regression

- Objective: to predict YMIS using a logistic model—model a variable representing YMIS on Z.
- Generate a single variable or a set of variables indicating where Y are missing[5].
- Logistic model will delete any observations with variables in ZMIS—potentially biasing the results of an exploratory analysis.
- High missingness among Z variables or a large number of variables in YMIS can hamper efforts to use multivariate methods.

# Z: High association with YMIS

- Non-sample information about relationships between observed variables and unobserved characteristics may help determine missingness of YMIS and YMIS.
- Results of previous studies demonstrating relationships between variables, e.g., association of education with income in previous example.
- Qualitative data that may explain why respondents left whole or partial surveys blank.

# **Selection of Appropriate Imputation Variables**

# Using all Z Variables

- MI literature recommends using all Z variables, but this may be a problem.
  - Multicollinearity among X, Y and Z considered to be a problem for the ML algorithms used in imputation[6].
- In our experience, too many variables, relative to number of non-missing observations, cause imputation to fail.

# Tests on Simulated Data

- Test 1: Tested for convergence with collinear data in SAS PROC MI:
  - Simulated high and perfectly collinear data but with relatively few collinear relationships.
  - Contrasted models with high/low item count (no. of variables used) and high/low missingness.
  - Contrasted with imputation convergence from a data set with very high collinearity, high item count and a high degree of missingness.
- Test 2: Compared *estimates* in high and perfect collinear simulations with estimates from non-collinear data, and estimates from highly collinear with high item count.

# Result: Multicollinearity may be misunderstood

- Collinearity itself, with sufficient degrees of freedom, results in convergence.
  - Estimates were nearly the same in the non-collinear, highly-collinear and perfectly-collinear models with high and low missingness and low item count.
  - Most simulation imputations converged
  - Except for high item count, high missingness; extra items contributed very high multicollinearity (VIF)
- Collinear data produce equivalent estimates (test 2) except with high item count and consequently extremely high collinearity
  - Tested on low missingness (high missingness did not converge)

# Conclusions: Steps to Imputation

1. Do not assume you should do MI
  - MI reduces bias in MAR-NMAR situations
  - Listwise deletion is better when data are MCAR
  - Note that ML may work under some circumstances with MAR data; for example, if only the dependent variable is in YMIS
2. Establish and document the X, Y and Z variables in your imputation model
  - Use all X variables that are available (no missing)  
Include all Y variables in the imputation model  
(missing data points on variables in analytical model)
  - Include Z variables in the model with relatively low rates of missingness and that are high informative of YMIS or the missingness of YMIS

# Conclusions: Steps to Imputation

3. Multivariate methods of exploratory research for establishing a good imputation model may be impugned by the missing data itself.
  - Be wary of large models (too many variables, too many missing data points) and models with a very small degrees of freedom.
  - Alternatively, use methods that pairwise delete only (usually, bivariate methods)

# Endnotes

- [1] Even if your data fail to demonstrate patterns that would explain the missingness, there may be a systematic explanation that is not measured.
- [2] Under *certain circumstances*, ML models will produce unbiased estimates when data are MAR. The criteria are beyond the extent of this presentation, but are discussed extensively in Allison, 2001 and also briefly in Raudenbush, 2001.
- [3] All other options—mean substitution, for example—are inferior and may actually increase bias.
- [4] “Prediction of missingness” insufficient for imputation (relationships among variables). Imputation also uses relationships among variables.
- [5] One indicator variable set to 1 on any observation on which any variable is in YMIS (zero otherwise); or, one indicator variable for each in YMIS set to 1 on any observation missing only that variable (zero otherwise).
- [6] As discussed on the MI listserv, the EM and MCMC algorithms may both experience trouble in the presence of multicollinearity.

# Appendix

- Do not use mean substitution because it reduces SEs, increasing the chances of observing a significant effect when one does not exist (Type I error). See Allison, 2001.
- Do not use SPSS MVA, because it relies on four faulty methods that may increase bias (von Hippel, 2004).
- Do not use SOLAS, because it relies on propensity scores, which provide insufficient information for imputation (relationship with missingness only, not with the missing values) (Allison, 2000).
- Do not use SAS version 8.2 experimental release of PROC MI, because it has been observed to impute the values (i.e., **replace**) *of observed variables*.

# References

- Allison, P. D. (2001). Missing Data. (Sage University Papers Series on Quantitative Applications in the Social Sciences, 07-136). Thousand Oaks, CA: Sage.
- Allison, P. D. (2000). Multiple imputation for missing data: a cautionary tale. Sociological Methods and Research 28: 301-309.
- Little, R. J. A., and Rubin, D. B. (2002). Statistical Analysis with Missing Data (2nd Edition). Hoboken, NJ: Wiley-Interscience.
- Raudenbush, S. W. (2001). Comparing personal trajectories and drawing causal inferences from longitudinal data. Annual Review of Psychology 52: 501-525.
- Rubin, D. B. (1996). Multiple imputation after 18+ years. Journal of the American Statistical Association 91(434): 473-489.
- \*\*Rubin, D. B. (1976). Inference and missing data. Biometrika 63(3): 581-592.
- Rubin, D. B. (1987). Multiple Imputation for Nonresponse in Surveys. New York, New York: Wiley.
- Schafer, J. L. (1997). Analysis of Incomplete Multivariate Data. Boca Raton, FL: Chapman Hall/CRC.
- Von Hippel, P. T. (2004). Biases in SPSS Missing Values Analysis. The American Statistician 58: 160-164.