



# Multiple Imputation: Methods and Applications

---

Jerry Reiter  
Department of Statistical Science  
Information Initiative at Duke  
Duke University  
[jreiter@duke.edu](mailto:jreiter@duke.edu)



# Plan today

---

- Overview of multiple imputation (MI)
  - Problems of missing data
  - Various solutions and their limitations
  - MI inferences
- Implementation of MI
- Example application of MI



# Types of missing data

---

- Unit nonresponse and item nonresponse
- Missing completely at random (MCAR)
- Missing at random (MAR)
- Not missing at random (NMAR)



# Mathematical formulation

---

- Let  $Y = (Y_{obs}, Y_{mis})$
- Let  $r_i = 1$  when data for unit  $i$  missing  
 $r_i = 0$  otherwise.
- Let  $R = (r_1, \dots, r_n)$
- Let  $\theta$  be the parameters associated with  $Y$
- Let  $\phi$  be the parameters associated with  $R$
- Assume  $\theta$  and  $\phi$  are distinct



# Mathematical formulation

---

- MCAR

$$f(R | Y, \theta, \phi) = f(R | \phi)$$

- MAR

$$f(R | Y, \theta, \phi) = f(R | Y_{obs}, \phi)$$

- NMAR

$$f(R | Y, \theta, \phi) = f(R | Y_{obs}, Y_{mis}, \phi)$$



# Implications for likelihood function for parameters

---

Likelihood function, no missing Y

$$L(\theta | Y_{obs}) \propto f(Y_{obs} | \theta)$$

Likelihood function, missing Y

$$\begin{aligned} L(\theta, \phi | Y_{obs}, R) &\propto f(Y_{obs}, R | \theta, \phi) \\ &= f(Y_{obs} | \theta) f(R | Y_{obs}, \phi) \end{aligned}$$



# Likelihood function: MCAR

---

$$f(Y_{obs}, R | \theta, \phi)$$

$$= \int f(Y_{obs}, Y_{mis} | \theta) f(R | Y_{obs}, Y_{mis}, \phi) dY_{mis}$$

■ MCAR:

$$= \int f(Y_{obs}, Y_{mis} | \theta) f(R | \phi) dY_{mis}$$

$$= f(Y_{obs} | \theta) f(R | \phi)$$



# Likelihood function: MAR

---

$$f(Y_{obs}, R | \theta, \phi)$$

$$= \int f(Y_{obs}, Y_{mis} | \theta) f(R | Y_{obs}, Y_{mis}, \phi) dY_{mis}$$

■ MAR:

$$= \int f(Y_{obs}, Y_{mis} | \theta) f(R | Y_{obs}, \phi) dY_{mis}$$

$$= f(Y_{obs} | \theta) f(R | Y_{obs}, \phi)$$





# Likelihood function: NMAR

---

$$f(Y_{obs}, R | \theta, \phi)$$

$$= \int f(Y_{obs}, Y_{mis} | \theta) f(R | Y_{obs}, Y_{mis}, \phi) dY_{mis}$$

- NMAR: Cannot simplify
- We cannot ignore the missing data when making inferences about  $\theta$



# How do you tell the typology?

---

In general, we don't know!!

- Rare that data are MCAR (unless planned)
- Possible that data are NMAR
- Compromise: assume data are MAR if we include enough variables in model for missing data indicators  $R$



# Likelihood function: MAR

---

Suppose have to include variables  $X$  to explain the reasons for missingness

$$f(Y_{obs}, R | X, \theta, \phi) = f(Y_{obs} | X, \theta) f(R | X, Y_{obs}, \phi)$$

*Practical implication for MAR:* include variables that explain missingness in model for  $Y$



# Strategies for handling item nonresponse

---

- Use complete/available cases analyses
  - Single imputation methods
  - Multiple imputation
  - Model-based methods
- 
- Weighting adjustments often used for unit nonresponse. Not readily used for item nonresponse.



# Complete/available cases analyses

---

- Consider income and education with item nonresponse for income.
- Estimate mean income and regression of income on education.
- What can happen when using available case analyses with different types of missing data?



# Summary of effects

---

Using complete/available cases when

- MCAR:  
unbiased when disregarding missing data;  
variance increase (losing partially complete data)
- MAR:  
bias when missing data mechanism not modeled;  
variance increase (losing partially complete data)
- NMAR:  
generally biased.



# Imputation methods

---

- Single imputation
  - (conditional) mean imputation
  - nearest neighbor imputation



# Mean imputation

---

Plug in the variable mean for missing values.

- Point estimates of means OK under MCAR.
- Variances and covariances underestimated.
- Distributional characteristics altered.
- Regression coefficients inaccurate.

Similar problems for plug-in conditional means.





# Nearest neighbor imputation

---

Plug in donors' observed values.

- For each nonrespondent, find a respondent who “looks like” the nonrespondent.
- Common metrics: Statistical distance, adjustment cells, propensity scores.



# Nearest neighbor imputation

---

Plug in donors' observed values.

- Point estimates of means OK under MAR.
- Variances and covariances underestimated.
- Distributional characteristics OK.
- Regression coefficients OK under MAR.



# Multiple imputation

---

- Fill in data sets several times with imputations. Analyze repeated data sets.
- Imputations drawn from probability models for missing data.

# *Observed Data*

<i>x</i>	<i>y</i>
✓	?
?	✓
✓	✓
✓	✓
✓	✓
✓	✓
?	✓

# *MI Datasets (1,...,m)*

<i>x</i>	<i>y</i>	<i>x</i>	<i>y</i>	<i>x</i>	<i>y</i>
✓	✓	✓	✓	✓	✓
✓	✓	✓	✓	✓	✓
✓	✓	✓	✓	✓	✓
✓	✓	✓	✓	✓	✓
✓	✓	✓	✓	✓	✓
✓	✓	✓	✓	✓	✓
✓	✓	✓	✓	✓	✓



# Inferences from multiply-imputed datasets

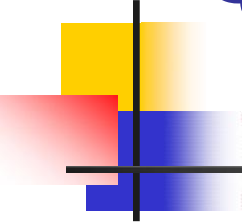
---

Rubin (1987)

- Estimand:  $Q = Q(X, Y)$
- In each imputed dataset  $d_i$

$$q_i = Q(d_i) \quad u_i = U(d_i)$$

# Quantities needed for inferences



---

$$\bar{q}_m = \sum_{i=1}^m q_i / m$$

$$b_m = \sum (q_i - \bar{q}_m)^2 / (m - 1)$$

$$\bar{u}_m = \sum_{i=1}^m u_i / m$$



# Inferences with multiply-imputed data

---

- Estimate of  $Q$ :  $\bar{q}_m$

- Estimate of variance is

$$T_m = (1 + 1/m)b_m + \bar{u}_m$$

- Use t-distribution inference for  $Q$ :

$$\bar{q}_m \pm t_{1-\alpha/2} \sqrt{T_m}$$



# Explanation of MI variance

---

$$T_m = (1 + 1/m)b_m + \bar{u}_m$$

- Consider  $m = \infty$ . Rubin (1987) shows

$$\begin{aligned} \text{Var}(Q | Y_{obs}) &= \text{Var}(E(Q | Y_{obs}, Y_{mis})) \\ &\quad + E(\text{Var}(Q | Y_{obs}, Y_{mis})) \\ &= b_\infty + \bar{u}_\infty \end{aligned}$$





# Methods for Wald tests or likelihood ratio tests

---

- Better degrees of freedom  
(Barnard and Rubin, 1999 *Biometrika*)
- Multi-component significance tests

Li et al. (1991, *JASA*)

Meng and Rubin (1992, *Biometrika*)

Reiter (2007, *Biometrika*)



# Pros and Cons

---

- Advantages

- Straightforward estimation of uncertainty
- Flexible modeling of missing data

- Disadvantages (?)

- Extra data sets to manage
- Explicitly model-based



# Concluding remarks

---

- Ignoring missing data is risky.
- Single imputation procedures at best underestimate uncertainty and at worst fail to capture multivariate relationships.
- Multiple imputation recommended (implementation in next part of course).



# Resources for learning more

---

- Little and Rubin (2002), *Statistical Analysis with Missing Data*, Wiley.
- Schafer (1997), *Analysis of Incomplete Multivariate Data*, CRC Press.
- Reiter and Raghunathan (2007), “The multiple adaptations of multiple imputation,” *J. Amer. Statist. Assoc.*



# Plan for today

---

- Overview of multiple imputation (MI)
- Implementation of MI
  - Approaches and models
  - Checking adequacy of imputations
- Example application of MI

# Observed Data

$x$	$y$
✓	?
?	✓
✓	✓
✓	✓
✓	✓
✓	✓
?	✓

# MI Datasets (1,...,m)

$x$	$y$	$x$	$y$	$x$	$y$
✓	✓	✓	✓	✓	✓
✓	✓	✓	✓	✓	✓
✓	✓	✓	✓	✓	✓
✓	✓	✓	✓	✓	✓
✓	✓	✓	✓	✓	✓
✓	✓	✓	✓	✓	✓
✓	✓	✓	✓	✓	✓



# Two general approaches

---

- Joint modeling
  - Posit a multivariate model for all the data
  - Estimate the model, usually with Bayesian methods
  - Impute from the joint model
- Sequential modeling
  - Estimate a sequence of conditional models
  - Impute from each model



# Desiderata

---

- Incorporate all sources of uncertainty in imputations, including uncertainty in parameter estimates.
- Want models that accurately describe the distribution of missing values.
- Important to keep in mind that imputation model used only for cases with missing data.
  - 30% missing values.
  - Model that is “80% good” (“20% bad”)
  - Completed data are only “6% bad”





# Some joint modeling techniques: Continuous data

---

- Multivariate normal data:

R: NORM, Amelia II

SAS: proc MI

Stata: MI command

- Mixtures of multivariate normal distributions:

R: EditImpCont (also does editing)



# Some joint modeling techniques: Categorical Data

---

- Multinomial data:
  - R: CAT -- log-linear model
  - NPBayesImpute – latent class model
- Mixed data:
  - R: MIX -- general location model
- Many other joint models, but often without open source software



# Sequential regression models

---

- Suppose data include  $Y_1, Y_2, Y_3, \text{etc.}$
- Fill in plausible starting values, e.g., simulate from regressions based on complete cases.
- Regress  $Y_1 | Y_2, Y_3, \text{etc.}$  using completed data. Impute new values of  $Y_1$  from this model.
- Repeat for  $Y_2 | Y_1, Y_3, \text{etc.}$
- Repeat for  $Y_3 | Y_1, Y_2, \text{etc.}$



# Sequential regression models

---

- Repeat for all variables with missing data.
- Cycle through steps many times.
  - Usually 5 times is a default but there is not theory underpinning this default
- Final dataset is one multiple imputation
- Repeat entire process  $m$  times



# Existing software for sequential regression approach

---

Free downloads of

- MICE for Stata, R (also, MI for R).
- IVEWARE for SAS.
  
- Can specify many types of conditional models and include constraints on values.



# Comparison to joint modeling

---

- Advantages

- Often easier to specify reasonable conditionals than a joint model.
- Complex MCMC not needed.

- Disadvantages

- Labor intensive to specify models.
- Incoherent conditionals can cause odd behaviors (e.g., order matters).
- Theoretical properties difficult to assess.



# Scenario in which current imputation approaches struggle

---

- Thousands of units, dozens of variables
- Numerical and categorical data
- Skewed or multi-modal distributions
- Complicated relationships
- Many public uses
- Lots of missing data

Aside: not required for MI to be useful



# A pie-in-the-sky vision for imputation generators

---

- An ideal imputation generator would
  - preserve as many relationships as possible
  - handle diverse data types
  - be computationally feasible for large data
  - be easy to implement with little tuning by the agency
- Existing methods don't always meet these desiderata





# Possible solutions

---

- Convert Bayesian mixture models into joint imputation engines.
  - Si and Reiter (2013, *JEBS*)
  - Manrique-Vallier and Reiter (2014, *JCGS*; 2018, *JASA*)
  - Kim et al. (2015, *JASA*)
  - Murray and Reiter (2016, *JASA*)
- Convert machine learning methods into sequential imputation engines.
  - CART (Burgette and Reiter, 2010)
  - Random forests (Caiola and Reiter, 2010)



# Bayesian mixture for MI

---

Consider high dimensional categorical data:

- Log-linear models
  - Difficult to specify and fit in data with high dimensions and complex dependencies (high order interactions).
  - Random zeros and separation present problems.
- Chained equations (MICE, IVEWARE)
  - Logistic and multinomial regressions suffer from similar problems as log-linear models.
  - Not derived from formal joint models and so can exhibit incoherent behavior.



# Notation for categorical data

---

- Data:  $p$  variables on  $n$  individuals.
- For  $j = 1, \dots, p$ , variable  $j$  takes on values in  $(1, \dots, d_j)$ .
- Let  $X_{ij}$  be value of categorical variable  $j$  for person  $i$ .



# Mixture (latent class) model

---

- Assume each person belongs to one of  $K$  latent classes. Let  $z_i \in (1, \dots, K)$  indicate the class membership.
- Mixture model for multiple imputation (Vermunt et al. 2008, *Sociol. Method.*)

$$X_{ij} \mid z_i, \pi, \omega \sim \text{Discrete}(\omega_{z_i 1}, \dots, \omega_{z_i d_j})$$

$$z_i \mid \pi \sim \text{Discrete}(\pi_1, \dots, \pi_K)$$



# Generating imputations: A Gibbs sampler

---

- Given completed data, sample parameters from common distributions (Dirichlet, categorical).
- Given parameter draws, straightforward to create completed datasets:
  - Draw latent class indicator for each individual.
  - Given latent class indicator, draw each  $X_{ij}$  from independent discrete distributions.
- Computationally efficient since using independent multinomial draws.



# Software implementations

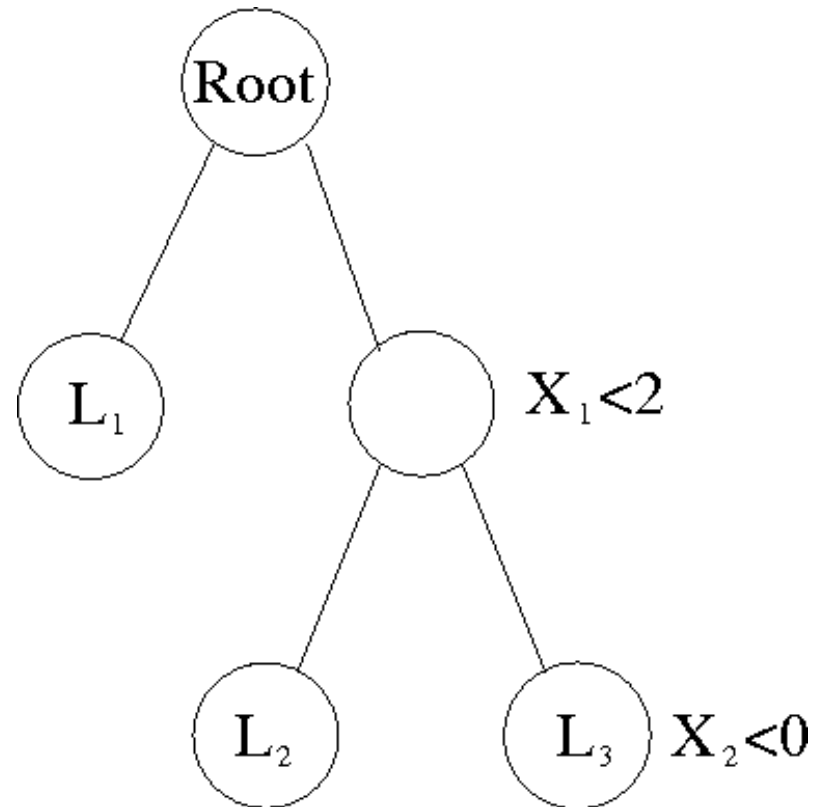
---

- R code implementing the categorical data imputations with or without structural zeros on CRAN as “NPBayesImpute.”
- R code implementing mixtures of multivariate normals with constraints on variables on CRAN as “EditImputeCont.”

# Sequential regression via CART: Overview of CART

Goal: Describe  $f(Y | X)$

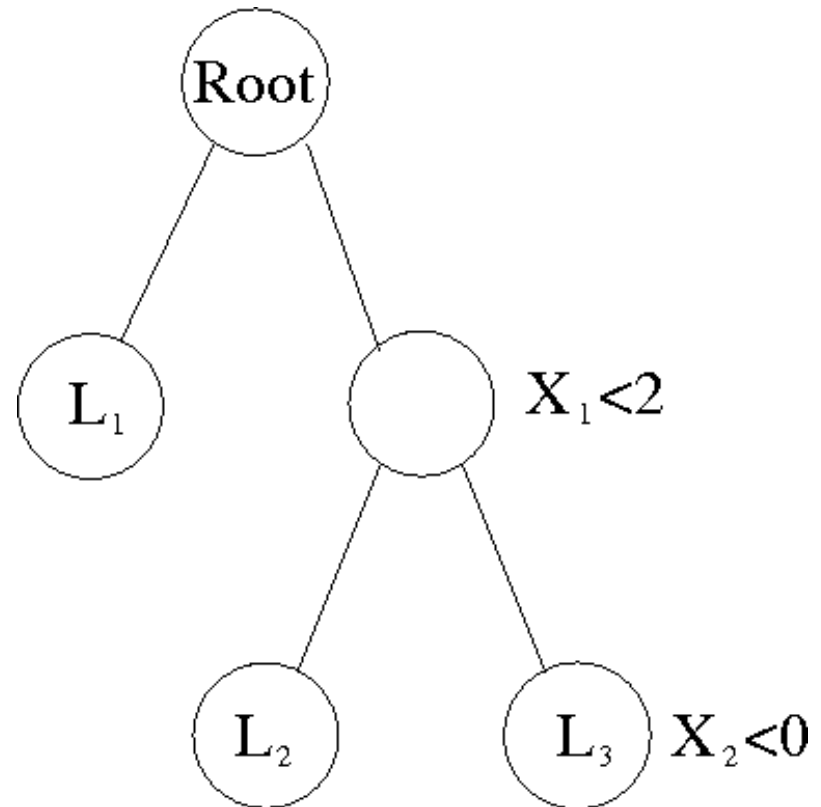
- Partition  $X$  space so that subsets of units formed by partitions have relatively homogenous  $Y$ .
- Partitions from recursive binary splits of  $X$ .
- Free routines in R.



# CART for MI of Y for complete X

Goal: Impute  $Y \mid X$ .

- Grow large tree.
- For any  $X$ , trace down tree until reach appropriate leaf.
- Draw  $Y$  from leaf using Bayesian bootstrap.







# Experience with CART as MI engine

---

- Use CART as conditional models in sequential imputation routines
- Models interactions and complex distributions automatically.
- Can outperform MICE with GLMs for data with complex structure.
- Can struggle with heavy tailed distributions.



# Software implementations

---

- R code implementing CART sequential imputation available from supplemental material of Burgette and Reiter (2010), although not being maintained.
- Now an option for CART imputation in MICE package in R.



# What if imputation and analysis model do not match?

---

- Imputation model more general than analysis model: inferences conservative.
- Imputation model less general than analysis model: inferences invalid.



# General advice on specifying imputation models

---

- For sequential modeling, include all variables related to outcome and missing data (Schafer, 1997).
- Include design information in models (Reiter *et al.* 2006, *Surv. Methodol.*).



# Evaluating the fit of imputation models

---

- Graphics of imputed and observed values (Abayomi *et al*, 2008, *JRSS-C*)
  - Imputed values don't look like observed values: \*maybe\* poor imputation models
  - Useful as a sensibility check
- Model-specific diagnostics (Gelman *et al*. 2005, *Biometrics*)
  - Residual plots with marked observed and imputed values



# Evaluating the fit of imputation models

---

- Posterior predictive checks (He *et al.* 2010, *Stat. Meth. Med. Res.*)
  - Fill in the missing data to create  $d_i$
  - Use same imputation model to generate new values of entire data set, including observed and missing data. Call this  $d_{i,rep}$
  - Compute statistic of interest, say  $S$ , using both  $d_i$  and  $d_{i,rep}$ .
  - Compare  $S$  from  $d_i$  and  $d_{i,rep}$ .



# Evaluating the fit of imputation

---

- **Very different values of  $S$ :**

Imputation model generates data that do not look like the completed data (with respect to  $S$ ). May want to improve imputation model.

- **Similar values of  $S$ :**

Imputation model generates data that look like the completed data (with respect to  $S$ ). Imputations reasonable.



# Interpreting posterior checks

---

- Practical issues when interpreting posterior predictive checks
  - Variables with very high rates of missing data can have small differences, since completed data and replicates both use model heavily. It's hard to get much useful out of the checks in this case.
  - Don't worry about variables with few missing values.
  - Consider size of deviations in statistics – meaningless deviations may not matter.





# Concluding remarks

---

- Sequential modeling strategies offer flexible imputations.
- Newer imputation methods being developed
  - Mixture models.
  - Machine learning
- Imputation model diagnostics still challenging. Posterior predictive checks can offer useful information.



# Plan for today

---

- Overview of multiple imputation (MI)
- Implementation of MI
- Examples of applications of MI
  - Simple example with NHANES data
  - R script available online



# Simple illustration

---

- Simple example using data that come with the MICE software package
- Dataset from NHANES includes 25 cases measured on 4 variables
- Only 13 cases with complete data
- We will use multiple imputation to make completed datasets and do analyses



# Concluding remarks

---

- Multiple imputation (MI) is a flexible method for handling missing data
- Sequential regression imputation techniques are useful as MI engines
- We discussed MI for MAR data. When data are NMAR life much harder – get experts in missing data on your team.



## For more...

---

- Triangle Census Research Network: one of eight nodes in the NSF-Census NCRN network
- Developing methodology for confidential data dissemination, handling missing values, and combining information
- Information, papers, and software from our group posted at

[www.sites.duke.edu/tcrn/](http://www.sites.duke.edu/tcrn/)