

An Introduction to  
Multiple Imputation (MI)  
Patrick M. Krueger

---

---

---

---

---

---

---

---

- Outline
- The missing data problem
  - Definitions
  - Conventional methods of dealing with missing data and their failings
  - The logic of Multiple Imputation (MI)
  - The imputation model
  - The estimation model and an example

---

---

---

---

---

---

---

---

- The Missing Data Problem
- Perfect world: "The best solution to the problem of missing data is to not have them"
  - In reality, missing data can result in:
    - Diminished sample size and inefficient estimates
    - Biased estimates of relationships
    - Wasted time and money, if we toss out observations

---

---

---

---

---

---

---

---

### Definitions: Types of Missing Data

- **Missing completely at random (MCAR):** The probability of having a missing value is unrelated to the respondent's other characteristics
- **Missing at random (MAR):** The probability of having a missing value is conditional on the respondent's other observed characteristics
- **Missing not at random (MNAR):** The probability of having a missing value depends on the respondent's other unobserved characteristics

---

---

---

---

---

---

---

---

### Conventional Methods and their Failings

- Listwise deletion
  - If data are MCAR, estimates will be unbiased but we lose efficiency
  - If data are not MCAR, estimates *may* be biased
- Dummy indicator for missing data
  - Simulation studies have demonstrated that this method routinely provides biased estimates, even when data are MCAR

---

---

---

---

---

---

---

---

### Conventional Methods and their Failings, cont'd

- Marginal mean imputation
  - Estimates will be biased even if data are MCAR
- Conditional mean imputation
  - If data are MCAR, estimates will be unbiased
  - Even if data are MCAR, standard errors are underestimated and test statistics are overestimated

---

---

---

---

---

---

---

---

### What is MI?

- Create multiple data sets, each with a different set of imputed values
- Each imputed value includes stochastic variation based on the variability in the posterior distribution of the estimate and the residuals
  - Thus, the imputed values vary across each of the MI data set
- MI relies on the MAR assumption; a weaker assumption than MCAR

---

---

---

---

---

---

---

---

### What is MI?

Observed	
y	x
1	4.2
3	8
2	5
6	.
7	.

---

---

---

---

---

---

---

---

### What is MI?

Observed	
y	x
1	4.2
3	8
2	5
6	.
7	.

m=1		m=2	
y	x	y	x
1	4.2	1	4.2
3	8	3	8
2	5	2	5
6	2	6	3.2
7	1.5	7	2

...

---

---

---

---

---

---

---

---

### Thinking about MI as Preserving Parameter Estimates

- MI is NOT about finding “the best” value for each missing value
- MI is about preserving unbiased estimates of population parameters:
  - means, variances, covariances, correlations, multivariate regression coefficients
  - We need to be right on average while accurately represent our uncertainty about missing values

---

---

---

---

---

---

---

---

### Thinking about Imputation as Replacement

- Statistically, estimates from a given study should not depend the particular sample drawn
- If re-sampling the population is not feasible, we can draw from the *estimated* distribution of values in the population
- Because we are no longer drawing from a population but an estimated distribution of the variables, we take multiple draws to represent our uncertainty

---

---

---

---

---

---

---

---

### The Imputation Model

- MI uses one model to create the multiple imputed data sets
  - MI models are iterative and stochastic models
- Estimation requires a second model
  - We estimate our substantive model using standard methods, in each data set
  - Combine our results from each data set using standard methods

---

---

---

---

---

---

---

---

### MI Models: Multivariate Normal (MVN) with Data Augmentation

- Strongest theoretical justification
- Implemented in SAS, Stata, and stand alone packages
- MVN assumes:
  - All variables have normal distributions
  - Each variable can be represented as a linear function of all other variables, together with a normal homoscedastic error term
  - Certain “tricks” exist for dealing with binary and categorical variables (see Allison 2002)

---

---

---

---

---

---

---

---

### MVN with DA: How it works

1. Choose starting values for means & covariance matrix
2. Use means & covariances to obtain regression estimates; each variable with missing data is regressed on all other variables
3. Regressions provide expected values for missing data. Predicted values add a draw from the residual normal distribution to impute missing data (this is the DA step)
4. Use “complete” data set with observed and imputed values to recalculate means and covariances
5. Use random draw from the posterior distribution of means & covariances to start again at step 2. Cycle through routine until model converges. Imputations from final iteration are used to create the complete data set

---

---

---

---

---

---

---

---

### MI Models: Imputation with Chained Equations (ICE)

- Less theoretical justification, but simulation studies suggest it works well
- Implemented in .ado file in Stata, and in stand alone packages
- Accommodates non-normal variables better than MVN
  - No need for “special tricks” required by MVN

---

---

---

---

---

---

---

---

### ICE: How it Works

1. Randomly order variables with missing data and replicate observed values across the missing cases
2. Regress each variable with missing data on all other variables with appropriate regression model
3. Use draws from posterior distributions of residuals and expected values to predict the missing values
4. Update the missing values with the predicted values
5. Repeat steps 2-4 with every other variable with missing data
6. Cycle through steps 2-5 until the model converges

---

---

---

---

---

---

---

---

### Rules of Thumb

- Use at least 5 data sets. Use more to:
  - Better preserve statistical power
  - Achieve more stable estimates when missing rates are high
- Imputation model should be at least as comprehensive as the estimation model
  - The imputation model must include all of the variables & interactions you want to use in the estimation model
  - Include any variables associated with the mechanism of missingness or with the variables to be imputed
- DO include the dependent variable in the imputation model
- Do NOT round imputed values (say, when using mvn model to predict a dummy variable)

---

---

---

---

---

---

---

---

### Benefits of MI?

- Preserves sample size for more efficient estimates even under MCAR
- Preserves representativeness of sample & reduces bias
  - MI relies on MAR assumption—this is most plausible if you have a strong imputation model

---

---

---

---

---

---

---

---

### Drawbacks of MI

- It can be cumbersome to create multiply imputed data sets for sophisticated models
- MI alone cannot deal with MNAR
  - Best defense is including as many variables as possible in the imputation model
- Due to the stochastic element, different analysts could use the same data and find different results

---

---

---

---

---

---

---

---

### Estimation Model: Example

- Estimate models in each data set like you would in any other data
- Use “Rubin’s rules” to combine estimates across data sets
  - Take the mean of the coefficients and “average” the standard errors
  - Most software automates this process

---

---

---

---

---

---

---

---

### References

Introductory Readings

Paul D. Allison. 2002. *Missing Data*. Thousand Oaks, CA: Sage

A. Rogier T. Donders, Geert J.M.G. van der Hijden, Theo Stijnen, Karel G.M. Moons. 2006. “Review: A gentle introduction to imputation of missing values.” *Journal of Clinical Epidemiology* 59: 1087-1091.

John W. Graham. 2009. “Missing data analysis: Making it work in the real world.” *Annual Review of Psychology* 60:549-576

---

---

---

---

---

---

---

---

### References

ICE in Stata

findit ice

Patrick Royston. 2009. "Multiple imputation of missing values: further update of ice, with an emphasis on categorical values." *The Stata Journal* 9(3):466-477

Patrick Royston. 2007. "Multiple imputation of missing values: further update of ice, with an emphasis on interval censoring." *The Stata Journal* 7(8):445-464

Patrick Royston. 2005. "Multiple imputation of missing values: update." *The Stata Journal* 5(2):188-201

Patrick Royston. 2004. "Multiple imputation of missing values." *The Stata Journal* 4(3):227-241

---

---

---

---

---

---

---

---

### References

Technical

Donald B. Rubin. 1987. *Multiple imputation for nonresponse in surveys*. New York: Wiley

Joe L. Schafer. 1997. *Analysis of incomplete multivariate data*. London: Chapman and Hall.

---

---

---

---

---

---

---

---