

Introduction to Multiple Imputation

Hsueh-Sheng Wu
CFDR Workshop Series
October 17, 2016

BGSU



Center for
**Family and
Demographic** Research

Outline

- Importance of Analyzing Missing data
- Three Mechanisms Underlying Missing data
- Strategies of Handling Missing Data
- Obtain Estimates from Imputed Data
- What is multiple imputation?
- Decisions on multiple imputation:
 - Do I really need to do multiple imputation (MI)?
 - What format of data do I want to work with?
 - What variables should be used in multiple imputation?
 - How many data sets need to be created?
 - Should I use MVN or MICE?
 - How to specify a MICE imputation model?
 - Do I impute data accurately?
 - How do I analyze imputed data?
- Possible imputation problems
- Conclusions
- Additional information

Importance of Analyzing Missing Data

- Missing data post several dilemma in analyzing data:
 - If you use the original data and exclude people with missing data from the analysis, you do not use all the information contained in the data.
 - If you replace missing values with other values and run the analysis, you can use all the information in the data set.
 - When the results differ for these two data sets, which result should you trust?
- Failure to adequately analyze missing data results in:
 - insufficient statistical power
 - upward or downward biases in parameter estimates
 - under- or over-estimated standard errors of the parameters
 - inaccurate findings

Three Mechanisms Underlying Missing Data

Assuming that we have a data set that contains one Y variable and many X variables:

- Missing completely at random (MCAR): No X variables in the data sets can predict whether the values in a variable (e.g., Y) will be missing. Also, the variable, Y, has missing value not because of the unobserved value of Y itself.
- Missing at random (MAR): X variables in the data sets can predict whether the values in Y will be missing.
- Missing not at random (MNAR): If the value of the variable, Y, or variables other than Xs determines whether the value of Y will be missing

Strategies of Handling Missing Data

- Delete cases
 - Pairwise deletion
 - Listwise deletion
- Substitution and imputation
 - Mean substitution
 - Hot deck imputation
 - Cold deck imputation
 - Regression
 - Multiple imputation

Obtain Estimates from Imputed Data

- Mean of the estimate obtained from m imputed data sets

$$\bar{Q} = \frac{1}{m} \sum_{j=1}^m \hat{Q}_j$$

- Standard error of the estimate obtained from m imputed data sets
 - within-imputation variance

$$\bar{U} = \frac{1}{m} \sum_{j=1}^m U_j$$

Obtain Estimates from Imputed Data (Con.)

- Between-imputation variance

$$B = \frac{1}{m-1} \sum_{j=1}^m (\hat{Q}_j - \bar{Q})^2$$

- Total variance

$$T = \bar{U} + \left(1 + \frac{1}{m}\right) B$$

- Standard error of the estimate

$$\sqrt{T}$$

What Is Multiple Imputation (MI)?

- The goal of MI is to obtain the accurate parameter estimates for relations of interest.
- The missing data are imputed m times to create m multiple data files.
- Analysis is conducted on each of m imputed data sets.
- The analysis results from each imputed data set are combined to obtain the final result.

Do I Really Need to Do Multiple Imputation?

You will need to do multiple imputation if many respondents will be excluded from the analytic sample due to their missing values and if the missing values of one variable can be predicted by other variables in the data file (i.e., the missing on random (MAR) assumption)

–Look at the patterns of missingness:
misstable sum, all

–Decide if multiple imputation can help you:
misstable sum, gen(miss_)

logit miss_urban female race edu exp wage
ttest female, by(miss_urban)

What Format of Data Do I Want to Work With?

Four types of data format: flong, flongsep, mlong, and wide

Original data						
	a	b	c			
1	1	2	3			
2	4	.	.			
flong						
	a	b	c	_mi_miss	_mi_m	_mi_id
1	1	2	3	0	0	1
2	4	.	.	1	0	2
3	1	2	3	.	1	1
4	4	4.5	8.5	.	1	2
5	1	2	3	.	2	1
6	4	5.5	9.5	.	2	2

What Format of Data Do I Want to Work With? (Cont.)

flongsep						
Original data						
	a	b	c	_mi_miss		_mi_id
1	1	2	3	0		1
2	4	.	.	1		2
First imputed data						
	a	b	c			_mi_id
1	1	2	3			1
2	4	4.5	8.5			2
Second imputed data						
	a	b	c			_mi_id
1	1	2	3			1
2	4	5.5	9.5			2

What Format of Data Do I Want to Work With? (Cont.)

Use “mi convert” to change between formats within Stata:

```
use style_flong.dta, clear
```

```
mi convert flongsep example, clear  
list
```

```
use _1_example, clear  
list
```

```
use _2_example, clear  
list
```

```
mi convert mlong, clear
```

```
mi convert wide, clear
```

What Variables Should Be Used in Multiple Imputation?

- Imputation model should definitely include dependent variables, independent variables, and some auxiliary variables (i.e., interaction terms or squared terms of independent variables, and weight variables), and maybe some other auxiliary variables (i.e., variables not in your analytic models).
- These auxiliary variables might help with the imputations as they make MAR assumption more reasonable (Collins et al., 2003). Using auxiliary variables is easy when MICE, but not MVN, is used.
- If you analyze a scale score, you should impute scale items and then generate the scale score unless (1) over half of the individual scale items are observed, (2) items have high value of internal consistency, and (3) the item-total correlations are similar across items (Graham, 2008).
- Stata can impute data and take into account the weighting issues at the same time.

How Many Data Sets Need to Be Created?

There is not consensus on this question.

- Stata suggests 20 imputed data sets.
- Schaffer suggested 5 imputed data sets
- Conventional advice has been 5-10, but more (e.g., 40) may yield increased power (Graham, Olchowski, & Gilreath, 2007).

Should I Use MVN or MICE?

Stata provides two approaches for imputing missing data:

- (1) Multiple imputation with the multivariate normal model (MVN)
- (2) Multiple Imputation by Chained Equations (MICE)

MVN:

- Assume a joint multivariate normal distribution of all variables.
- Directly maximize the parameter estimate using the observed cases and maximum likelihood method.
- Sometimes multivariate normal model is used even with categorical variables, but this can be severely biased (Horton, Lipsitz, and Parzen, 2003; Allison 2005).
- Can't easily handle complexities such as skip patterns, bounds restrictions, complex designs

MICE

- Fit model of each variable, conditional on all others.
- Models used depend on types of variables (categorical/continuous/binary). Researchers have more flexibility in specifying imputation models for different variables and for different subpopulations.
- Doesn't necessarily imply a proper joint distribution like MVN does, but this doesn't seem to be a big problem in practice.

How Do I Specify a MICE Imputation Model?

A MICE model will need at least the following information:

- The attribute of variables: regular, imputed, and passive.
- Variables that will be used to generate imputed values on variables
- Regression models that link variables together, including linear regress (regress), predictive mean matching (pmm), truncated regression (truncreg), interval regression (intreg), logistic regression (logit) ordered logistic regression (ologit), multinomial logistic regression(mlogit), poisson regression (poisson), and negative binomial regression (nbreg)
- Variables that modify the relations among variables, such as group indicators or weights
- The random seed number
- The number of imputed data sets

How Do I Specify a MICE Imputation Model?

Example 1. A imputation model

```
use http://www.stata-press.com/data/r13/mheart8s0, clear  
mi describe  
misstable pattern
```

```
mi impute chained //  
(pmm, knn(3)) bmi //  
(reg) age //  
= attack smokes hsgrad female, add(5) rseed(123)
```

Example 2. using weights in the imputation model

```
use http://www.stata-press.com/data/r13/mheart8s0, clear  
gen wgt =1  
gen pwgt = 1/154
```

```
mi impute chained //  
(pmm, knn(3)) bmi //  
(reg) age //  
= attack smokes hsgrad female [pweight=pwgt], add(5) rseed(123)
```

Example 2. using weight different imputation model (Continued)

```
use http://www.stata-press.com/data/r13/mheart8s0, clear
gen wgt =1
gen pwgt = 1/154

mi svyset [pw=pwgt]

mi impute chained                               ///
(pmm, knn(3)) bmi                               ///
(reg) age                                       ///
= attack smokes hsgrad female [pweight=pwgt], add(5) rseed(123)
```

How Do I Specify a MICE Imputation Model?

Example 3. Custom prediction equations

```
use http://www.stata-press.com/data/r13/mheart8s0, clear
mi impute chained //
(pmm, knn(3) omit(hsgrad)) bmi //
(regress) age //
= attack smokes hsgrad female, add(5) rseed(123)
```

Example 4. Imputing variables of different types

```
use http://www.stata-press.com/data/r13/mheart9s0, clear
mi describe
mi misstable nested

mi impute chained //
(pmm, knn(3) omit(hsgrad)) bmi //
(regress) age //
(logit if smokes==1, omit(smokes)) hightar //
= attack smokes hsgrad female, add(5) rseed(123)
```

How Do I Specify a MICE Imputation Model?

Example 5. Conditional imputation

use <http://www.stata-press.com/data/r13/mheart10s0>, clear

mi describe

mi misstable nested

```
mi impute chained (pmm, knn(3) omit(hsgrad)) bmi           ///
(regress) age                                             ///
(logit, cond(if smokes==1) omit(i.smokes)) hightar       ///
(logit) smokes                                           ///
= attack hsgrad female, add(5) rseed(123)
```

Example 6. Including expressions of imputation variables

```
mi impute chained                                       ///
(pmm, knn(3) omit(hsgrad) incl((age^2))) bmi           ///
(pmm) age                                               ///
(logit, cond(if smokes==1) omit(i.smokes)) hightar     ///
(logit) smokes                                         ///
= attack hsgrad female, replace
```

How Do I Specify a MICE Imputation Model?

Example 7. Imputing on subsamples

use <http://www.stata-press.com/data/r13/mheart8s0>, clear

```
mi impute chained //  
(pmm, knn(3) omit(hsgrad)) bmi //  
(regress) age //  
= attack smokes hsgrad, by(female) dryrun
```

```
mi impute chained //  
(pmm, knn(3) omit(hsgrad)) bmi //  
(regress) age //  
= attack smokes hsgrad, add(5) rseed(123) by(female, noreport) dots
```

Do I Impute Data Accurately?

This area is under-developed. However, after imputing data, you can look at the values of the variables to identify two possible problems.

(1) The value of variables in the data set do not vary the way you had anticipated

mi vary

(2) The imputed value of a variable exceed the range of observed values of the variable

mi xeq 0: misstable sum, all

mi xeq 0: tab age

mi xeq 1/5: tab age if miss_age

How Do I Analyze Imputed Data?

After imputing the data, you can use “mi estimate” to analyze the data and do five tests:

(1) Test what variables can predict the outcome variable
mi estimate: logistic attach smokes age bmi hsgrad
mi estimate: svy:logistic attach smokes age bmi hsgrad

(2) Test what variables are simultaneously significant predictors of the outcome variable

mi estimate: regress y x1 x2 x3 x4
mi test x2 x3 x4

(3) Test the equality between two regression coefficients

mi estimate (diff: b[x1]- b[x2]): regress y x1 x2 x3 x4
mi testtransform diff

mi estimate, saving(myresults): regress y x1 x2 x3 x4
mi estimate (diff: b[x1]- b[x2]) using myresults
mi testtransform diff

How Do I Analyze Imputed Data? (Continued)

(4) Test the equality between multiple pairs of coefficients
mi estimate (diff1: $b[x1] - b[x2]$) (diff2: $b[x1] = b[x3]$) using myresults
mi testtransform diff1 diff2

(5) Test nonlinear hypotheses
mi estimate (diff: $b[x1]/b[x2] - b[x3]/b[x4]$) using myresults
mi testtransform diff

Possible Imputation Problems

- Syntax errors
 - Use the dryrun option to check the accuracy of the syntax
- Converge problems
 - Mis-specified regression model can lead to convergence problem. You can run imputation model for each variable and identify the accurate regression model.
 - When you try to impute a data file with many categorical variables and small number of observations, the imputation model may not converge. You can consider collapsing categories.
- Perfect prediction
 - When the imputed variable is a categorical variable, Stata may find that some variables can perfectly predict the imputed variable. You can use the “augment” option to solve this problem.

Conclusions

- Multiple Imputation helps keep as many observations as possible in the analytic analysis.
- Theoretically, if multiple imputation models are specified correctly, researchers should be able to get unbiased parameter estimates when analyzing imputed data.
- When checking the imputation model, please check for the accuracy of your codes, the functional form between the predictors and the imputed variable, and if such functional form should differ for different sub-populations.
- Doing multiple imputation can be time-consuming if you have big data files, many variables, lots of categorical variables, and complex functional forms between variables. Thus, you should start small by doing multiple imputation for a small number of variables and then expand.
- If you have problems doing multiple imputation, please send an email to me (wuh@bgsu.edu) or drop by my office. I am glad to help.

Additional Information

- Azur, M; Stuart, E.; Frangakis, C.; Leaf, P (2011) Multiple Imputation by Chained Equation: What is it and how does it work? *International Journal of Methods in Psychiatric Research*, 20,40-49.
- Royston, P & White, I. (2011) Multiple Imputation by Chained Equations (MICE): Implementation in Stata. *Journal of Statistical Software*, 45,1-20.
- The Manual of Stata Multiple Imputation
- Youtube Videos:
 - Recent Advances in missing Data Methods: Imputation and Weighting - Elizabeth Stuart (<https://www.youtube.com/watch?v=xnQ17bbSeEk>)
 - Multiple imputation in Stata®: Setup, imputation, estimation--regression imputation (<https://www.youtube.com/watch?v=i6SOIq0mjuc&index=1&list=PLN5lSkQdgXWmhjxC5eopeRJwpI9G7Kp5w>)
 - Multiple imputation in Stata®: Setup, imputation, estimation--predictive mean matching (<https://www.youtube.com/watch?v=c75E2LBGoBQ&index=2&list=PLN5lSkQdgXWmhjxC5eopeRJwpI9G7Kp5w>)
 - Multiple imputation in Stata®: Setup, imputation, estimation--logistic regression (<https://www.youtube.com/watch?v=QVvTpPx2LyU&index=3&list=PLN5lSkQdgXWmhjxC5eopeRJwpI9G7Kp5w>)
 - Multiple Imputation in Stata: (http://www.ssc.wisc.edu/sscc/pubs/stata_mi_intro.htm)
 - Recent Advances in Missing Data Methods: Multiple Imputation by Chained Equations (<http://www.academyhealth.org/files/2010/sunday/StuartE.pdf>)