Chained equations and more in multiple imputation in Stata 12

Yulia Marchenko
Associate Director, Biostatistics
StataCorp LP

2011 Nordic and Baltic Stata Users Group Meeting
<table>
<thead>
<tr>
<th>Outline</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brief overview of MI</td>
</tr>
<tr>
<td>Brief history of MI in Stata</td>
</tr>
<tr>
<td>New official MI features in Stata 12</td>
</tr>
<tr>
<td>Multiple imputation using chained equations (MICE)</td>
</tr>
<tr>
<td>Overview</td>
</tr>
<tr>
<td>Advantages/Disadvantages</td>
</tr>
<tr>
<td>Incompatibility of conditionals</td>
</tr>
<tr>
<td>MICE versus MVN</td>
</tr>
<tr>
<td>Examples</td>
</tr>
<tr>
<td>Convergence</td>
</tr>
<tr>
<td>Concluding remarks</td>
</tr>
<tr>
<td>References</td>
</tr>
</tbody>
</table>

Chained equations and more in multiple imputation in Stata 12

Outline

Brief overview of MI
Brief history of MI in Stata
New official MI features in Stata 12
Multiple imputation using chained equations (MICE)
  Overview
  Advantages/Disadvantages
  Incompatibility of conditionals
  MICE versus MVN
  Examples
  Convergence
Concluding remarks
References
Multiple imputation (MI) is a principled, simulation-based approach for analyzing incomplete data.

MI procedure:
1) replaces missing values with multiple sets of simulated values to complete the data,
2) applies standard analyses to each completed dataset,
3) adjusts the obtained parameter estimates for missing-data uncertainty.

The objective of MI is not to predict missing values as close as possible to the true ones but to handle missing data in a way resulting in valid statistical inference (Rubin 1996).

MI is statistically valid if an imputation model is proper and the primary, completed-data analysis is statistically valid in the absence of missing data (Rubin 1987).
Chained equations and more in multiple imputation in Stata

Brief history of MI in Stata

User-written tools

Stata 7

- 2003 (Carlin et al. 2003): tools for analyzing multiply imputed data (mifit, miset, mido, mici, mitestparm, miappend, etc.)

Stata 8

- 2004 (Royston 2004): univariate imputation (uvis) and multivariate imputation using chained equations (mvis), analysis of multiply imputed data (micombine similar to Carlin’s mifit)
- 2005 (Royston 2005a, 2005b): ice replaces and extends mvis for imputation using chained equations
- 2007 (Royston 2007): updates for ice with an emphasis on interval censoring
- 2008: mira by Rodrigo Alfaro for analyzing MI data stored in separate files
Stata 9

- 2008 (Carlin et al. 2008): new framework for managing and analyzing MI data (the `mim:` prefix replaces `micombine`, `mifit`, and other earlier tools for analyzing and manipulating MI data)

- 2009 (Royston 2009, Royston et al. 2009): updates to `ice` and `mim`

`inorm` by John Galati and John Carlin for performing imputation using MVN
Stata 11

- 2009: an official suite of commands for creating (mi impute), manipulating (mi merge, mi reshape, etc.), and analyzing (mi estimate) MI data
  - mi provides 4 different styles of storing MI data, MI data verification, and extensive data-management support
  - mi impute provides a number of univariate imputation methods and multivariate imputation using MVN
  - the mi estimate: prefix, similar to mim:, analyzes MI data

Stata 12

- 2011: various additions to mi, including multivariate imputation using chained equations (mi impute chained)

See http://www.stata.com/support/faqs/stat/mi_ice.html for comparison of mi with user-written commands ice and mim
Multivariate imputation using chained equations (mi impute chained)

Four new univariate imputation methods of mi impute: truncreg, intreg, poisson, and nbreg

Conditional imputation within mi impute chained and mi impute monotone

Handling of perfect prediction via the new augment option during imputation of categorical data

Separate imputation for different groups of the data via the new by() option of mi impute
- **mi estimate, mceerror** estimates the amount of simulation error associated with MI results.
- New commands **mi predict** and **mi predictnl** to compute linear and nonlinear MI predictions.
- **misstable summarize, generate()** creates missing-value indicators for variables containing missing values.
MICE (van Buuren et al. 1999) is an iterative imputation method that imputes multiple variables by using chained equations, a sequence of univariate imputation methods with fully conditional specification (FCS) of prediction equations.

That is, to get one set of imputed values, iterate over $t = 0, 1, \ldots, T$ and impute:

- $X_1^{(t+1)}$ using $X_2^{(t)}, X_3^{(t)}, \ldots, X_q^{(t)}$
- $X_2^{(t+1)}$ using $X_1^{(t+1)}, X_3^{(t)}, \ldots, X_q^{(t)}$
- \[ \ldots \]
- $X_q^{(t+1)}$ using $X_1^{(t+1)}, X_2^{(t+1)}, \ldots, X_{q-1}^{(t+1)}$
MICE is also known as FCS and SRMI, sequential regression multivariate imputation (Raghunathan et al. 2001)

MICE can handle variables of different types

MICE can handle arbitrary missing-data patterns

MICE can accommodate certain important characteristics (data ranges, restrictions within a subset) of the observational data

Being an iterative method, MICE requires checking of convergence

MICE requires careful modeling of conditional specifications

See White et al. (2011) for practical guidelines about using MICE
The variable-by-variable specification of MICE makes it easy to build complicated imputation models for multiple variables.

Unlike sequential monotone imputation, MICE does not require monotone missing-data patterns.

MICE accommodates variables of different types by using an imputation method appropriate for each variable.

MICE allows different sets of predictors when imputing different variables.

MICE allows to impute missing values within the observed (or pre-specified) ranges of the data.

MICE can handle imputation of variables defined only on a subset of the data—conditional imputation.

MICE can incorporate functional relationships among variables.
- MICE lacks formal theoretical justification

- In particular, its theoretical weakness is possible incompatibility of fully conditional specifications for which no proper joint multivariate distribution exists

- The variable-by-variable specification of MICE also makes it easy to build models with incompatible conditionals
MICE is similar in spirit to a Gibbs sampler but is not a true Gibbs sampler except in rare cases.

A set of fully conditional specifications may be incompatible, that is, it may not correspond to any proper joint multivariate distribution (e.g., Arnold et al. 2001).

For example, $X_1|X_2 \sim N(\alpha_1 + \beta_1 X_2, \sigma_1^2)$ and $X_2|X_1 \sim N(\alpha_2 + \beta_2 \ln X_1, \sigma_2^2)$ are incompatible.

See, for example, van Buuren (2006, 2007) for the impact of incompatible conditionals on final MI results—only minor impact was found in the examples considered.
MICE uses a sequential (variable-by-variable) approach for imputation; MVN (Schafer 1997) uses a joint modeling approach based on a multivariate normal distribution.

MICE has no theoretical justification (except in some particular cases); MVN does.

MICE can handle variables of different types; MVN is intended for continuous variables and requires normality (Schafer [1997] and Allison [2001] note that MVN can be robust to departures from normality and can sometimes be used to model binary and ordinal variables).

MICE can incorporate important data characteristics such as ranges and restrictions within a subset of the data; in general, MVN cannot.

In practice, the quality of imputations from either of the methods should be examined.

See, for example, Lee and Carlin (2010) for a recent comparison of MVN and MICE.
Consider fictional data recording heart attacks

. use mheart8
(Fictional heart attack data; bmi and age missing; arbitrary pattern)
. describe

Contains data from mheart8.dta

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>obs:</td>
<td>154</td>
<td>Fictional heart attack data; bmi and age missing; arbitrary pattern</td>
</tr>
<tr>
<td>vars:</td>
<td>6</td>
<td>1 Sep 2011 10:11</td>
</tr>
<tr>
<td>size:</td>
<td>1,848</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>variable name</th>
<th>storage type</th>
<th>display format</th>
<th>value label</th>
</tr>
</thead>
<tbody>
<tr>
<td>attack</td>
<td>byte</td>
<td>%9.0g</td>
<td>Outcome (heart attack)</td>
</tr>
<tr>
<td>smokes</td>
<td>byte</td>
<td>%9.0g</td>
<td>Current smoker</td>
</tr>
<tr>
<td>age</td>
<td>float</td>
<td>%9.0g</td>
<td>Age, in years</td>
</tr>
<tr>
<td>bmi</td>
<td>float</td>
<td>%9.0g</td>
<td>Body Mass Index, kg/m^2</td>
</tr>
<tr>
<td>female</td>
<td>byte</td>
<td>%9.0g</td>
<td>Gender</td>
</tr>
<tr>
<td>hsgrad</td>
<td>byte</td>
<td>%9.0g</td>
<td>High school graduate</td>
</tr>
</tbody>
</table>

Sorted by:
Let's summarize missing values

. misstable summarize, generate(Mis_)

<table>
<thead>
<tr>
<th>Variable</th>
<th>Obs=.</th>
<th>Obs&gt;</th>
<th>Obs&lt;</th>
<th>Unique values</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>age</td>
<td>12</td>
<td>142</td>
<td>142</td>
<td>142</td>
<td>20.73613</td>
<td>83.78423</td>
</tr>
<tr>
<td>bmi</td>
<td>28</td>
<td>126</td>
<td>126</td>
<td>126</td>
<td>17.22643</td>
<td>38.24214</td>
</tr>
</tbody>
</table>

and explore missing-data patterns

. misstable patterns

Missing-value patterns
(1 means complete)

<table>
<thead>
<tr>
<th>Percent</th>
<th>Pattern</th>
<th>1</th>
<th>2</th>
</tr>
</thead>
<tbody>
<tr>
<td>77%</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>1</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>100%</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Variables are (1) age (2) bmi
- Declare the storage style

  . mi set wide

- Register variables

  . mi register imputed age bmi
  . mi register regular attack smokes female hsgrad
**Impute age and bmi using regression imputation**

```plaintext
.mi impute chained (regress) age bmi = attack smokes female hsgrad, add(5) rseed(27654)
```

Conditional models:
- age: `regress age bmi attack smokes female hsgrad`
- bmi: `regress bmi age attack smokes female hsgrad`

Performing chained iterations ...

<table>
<thead>
<tr>
<th>Multivariate imputation</th>
<th>Imputations = 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chained equations</td>
<td>added = 5</td>
</tr>
<tr>
<td>Imputed: m=1 through m=5</td>
<td>updated = 0</td>
</tr>
<tr>
<td>Initialization: monotone</td>
<td>Iterations = 50</td>
</tr>
<tr>
<td></td>
<td>burn-in = 10</td>
</tr>
</tbody>
</table>

- age: linear regression
- bmi: linear regression

<table>
<thead>
<tr>
<th>Variable</th>
<th>Complete</th>
<th>Incomplete</th>
<th>Imputed</th>
</tr>
</thead>
<tbody>
<tr>
<td>age</td>
<td>142</td>
<td>12</td>
<td>12</td>
</tr>
<tr>
<td>bmi</td>
<td>126</td>
<td>28</td>
<td>28</td>
</tr>
</tbody>
</table>

(complete + incomplete = total; imputed is the minimum across m of the number of filled-in observations.)
Compare distributions of the imputed, completed, and observed data for age. (*midiagplots* is a forthcoming user-written command; see Marchenko and Eddings (2011) for how to create MI diagnostic plots manually)

```
. midiagplots age, m(1/5) combine
(M = 5 imputations)
(imputed: age bmi)
```

*(Continued on next page)*
Chained equations and more in multiple imputation in Stata

- Multiple imputation using chained equations
- Example 1: MI diagnostics

Observed Imputed Completed

Yulia Marchenko (StataCorp) November 17, 2011 20 / 45
Compare distributions of the imputed, completed, and observed data for bmi

. midiagplots bmi, m(1/5) combine
(M = 5 imputations)
(imputed: age bmi)

(Continued on next page)
Chained equations and more in multiple imputation in Stata 12

Multiple imputation using chained equations

Example 1: MI diagnostics

Imputation 1

Cumulative distribution

0 .2 .4 .6 .8 1

Body Mass Index, kg/m^2

15 20 25 30 35 40

Imputation 2

Cumulative distribution

0 .2 .4 .6 .8 1

Body Mass Index, kg/m^2

15 20 25 30 35 40

Imputation 3

Cumulative distribution

0 .2 .4 .6 .8 1

Body Mass Index, kg/m^2

15 20 25 30 35 40

Imputation 4

Cumulative distribution

0 .2 .4 .6 .8 1

Body Mass Index, kg/m^2

15 20 25 30 35 40

Imputation 5

Cumulative distribution

0 .2 .4 .6 .8 1

Body Mass Index, kg/m^2

10 20 30 40

---  Observed  ---  Imputed  ---  Completed
. mi estimate, merror cformat(\%8.4f): logit attack smokes age bmi female hsgrad

Multiple-imputation estimates
Logistic regression
Number of obs = 154
Average RVI = 0.0338
Largest FMI = 0.0866

DF adjustment: Large sample
DF: min = 574.54
avg = 1370395.93
max = 7973220.18

Model F test: Equal FMI
F( 5, 9595.8) = 3.53

Within VCE type: OIM
Prob > F = 0.0035

| Coefficient | Std. Err. | t     | P>|t| | [95% Conf. Interval] |
|-------------|-----------|-------|------|----------------------|
| attack      |           |       |      |                      |
| smokes      | 1.1326    | 0.3561| 3.18 | 0.001                |
|             | 0.0145    | 0.0009| 0.04 | 0.000                |
| age         | 0.0372    | 0.0162| 2.30 | 0.022                |
|             | 0.0019    | 0.0003| 0.12 | 0.007                |
| bmi         | 0.0935    | 0.0457| 2.05 | 0.041                |
|             | 0.0044    | 0.0011| 0.11 | 0.011                |
| female      | -0.1331   | 0.4171| -0.32| 0.750                |
|             | 0.0195    | 0.0020| 0.05 | 0.035                |
| hsgrad      | 0.1324    | 0.4019| 0.33 | 0.742                |
|             | 0.0112    | 0.0007| 0.03 | 0.021                |
| _cons       | -5.2048   | 1.5652| -3.33| 0.001                |
|             | 0.0170    | 0.0163| 0.03 | 0.000                |

Note: values displayed beneath estimates are Monte Carlo error estimates.
Impute bmi using predictive mean matching instead

```
.mi impute chained (regress) age (pmm) bmi = attack smokes female hsgrad, replace
```

Conditional models:
```
age: regress age bmi attack smokes female hsgrad
bmi: pmm bmi age attack smokes female hsgrad
```

Performing chained iterations ...

<table>
<thead>
<tr>
<th>Multivariate imputation</th>
<th>Imputations = 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chained equations</td>
<td>added = 0</td>
</tr>
<tr>
<td>Imputed: m=1 through m=5</td>
<td>updated = 5</td>
</tr>
<tr>
<td>Initialization: monotone</td>
<td>Iterations = 50</td>
</tr>
<tr>
<td></td>
<td>burn-in = 10</td>
</tr>
</tbody>
</table>

age: linear regression
bmi: predictive mean matching

<table>
<thead>
<tr>
<th>Variable</th>
<th>Complete</th>
<th>Incomplete</th>
<th>Imputed</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>age</td>
<td>142</td>
<td>12</td>
<td>12</td>
<td>154</td>
</tr>
<tr>
<td>bmi</td>
<td>126</td>
<td>28</td>
<td>28</td>
<td>154</td>
</tr>
</tbody>
</table>

(complete + incomplete = total; imputed is the minimum across m of the number of filled-in observations.)
Omit hsgrad from the prediction equation for bmi

```
. mi impute chained (regress) age ///
> (pmm, omit(hsgrad)) bmi ///
> = attack smokes female hsgrad, replace
```

Conditional models:
```
age: regress age bmi attack smokes female hsgrad
bmi: pmm bmi age attack smokes female
```

Performing chained iterations ...

Multivariate imputation Imputations = 5
Chained equations added = 0
Imputed: m=1 through m=5 updated = 5
Initialization: monotone Iterations = 50
burn-in = 10

age: linear regression
bmi: predictive mean matching

<table>
<thead>
<tr>
<th>Variable</th>
<th>Complete</th>
<th>Incomplete</th>
<th>Imputed</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>age</td>
<td>142</td>
<td>12</td>
<td>12</td>
<td>154</td>
</tr>
<tr>
<td>bmi</td>
<td>126</td>
<td>28</td>
<td>28</td>
<td>154</td>
</tr>
</tbody>
</table>

(complete + incomplete = total; imputed is the minimum across m of the number of filled-in observations.)
Or, include `hsgrad` in the prediction equation for `age`

```
    . mi impute chained (regress, include(hsgrad)) age ///
        (pmm) bmi ///
        = attack smokes female, replace
```

Conditional models:

```
    age: regress age bmi hsgrad attack smokes female
    bmi: pmm bmi age attack smokes female
```

Performing chained iterations ...

Multivariate imputation Imputations = 5
Chained equations added = 0
Imputed: m=1 through m=5 updated = 5
Initialization: monotone Iterations = 50
    burn-in = 10

age: linear regression
bmi: predictive mean matching

<table>
<thead>
<tr>
<th>Variable</th>
<th>Complete</th>
<th>Incomplete</th>
<th>Imputed</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>age</td>
<td>142</td>
<td>12</td>
<td>12</td>
<td>154</td>
</tr>
<tr>
<td>bmi</td>
<td>126</td>
<td>28</td>
<td>28</td>
<td>154</td>
</tr>
</tbody>
</table>

(complete + incomplete = total; imputed is the minimum across m of the number of filled-in observations.)
What if relationship between age and bmi is curvilinear?

```
. mi impute chained (regress, include(hsgrad (bmi^2))) age ///
>   (pmm) bmi ///
>   = attack smokes female, replace
```

Conditional models:
- age: `regress age bmi hsgrad (bmi^2) attack smokes female`
- bmi: `pmm bmi age attack smokes female`

Performing chained iterations ...

<table>
<thead>
<tr>
<th>Multivariate imputation</th>
<th>Imputations =</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chained equations</td>
<td>added =</td>
<td>0</td>
</tr>
<tr>
<td>Imputed: m=1 through m=5</td>
<td>updated =</td>
<td>5</td>
</tr>
<tr>
<td>Initialization: monotone</td>
<td>Iterations =</td>
<td>50</td>
</tr>
<tr>
<td></td>
<td>burn-in =</td>
<td>10</td>
</tr>
</tbody>
</table>

age: linear regression
bmi: predictive mean matching

<table>
<thead>
<tr>
<th>Variable</th>
<th>Complete</th>
<th>Incomplete</th>
<th>Imputed</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>age</td>
<td>142</td>
<td>12</td>
<td>12</td>
<td>154</td>
</tr>
<tr>
<td>bmi</td>
<td>126</td>
<td>28</td>
<td>28</td>
<td>154</td>
</tr>
</tbody>
</table>

(complete + incomplete = total; imputed is the minimum across m of the number of filled-in observations.)
What if unobserved values of age are known to lie in $[20, 84]$?

```
. generate age_l = cond(age==., 20, age)
. generate age_u = cond(age==., 84, age)
. mi impute chained (intreg, ll(age_l) ul(age_u) include(hsgrad)) age ///
> (pmm) bmi ///
> = attack smokes female, replace
```

Conditional models:
- age: intreg age bmi hsgrad attack smokes female , ll(age_l) ul(age_u)
- bmi: pmm bmi age attack smokes female

Performing chained iterations ...

<table>
<thead>
<tr>
<th>Variable</th>
<th>Complete</th>
<th>Incomplete</th>
<th>Imputed</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>age</td>
<td>142</td>
<td>12</td>
<td>12</td>
<td>154</td>
</tr>
<tr>
<td>bmi</td>
<td>126</td>
<td>28</td>
<td>28</td>
<td>154</td>
</tr>
</tbody>
</table>
Impute age and bmi separately for males and females

```
.mi impute chained (regress) age (pmm) bmi = attack smokes hsgrad,
> > replace by(female, noreport)
```

Multivariate imputation Imputations = 5
Chained equations added = 0
Imputed: m=1 through m=5 updated = 5
Initialization: monotone Iterations = 50
burn-in = 10

age: linear regression
bmi: predictive mean matching

<table>
<thead>
<tr>
<th>by()</th>
<th>Variable</th>
<th>Complete</th>
<th>Incomplete</th>
<th>Imputed</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>female = 0</td>
<td>age</td>
<td>106</td>
<td>10</td>
<td>10</td>
<td>116</td>
</tr>
<tr>
<td></td>
<td>bmi</td>
<td>95</td>
<td>21</td>
<td>21</td>
<td>116</td>
</tr>
<tr>
<td>female = 1</td>
<td>age</td>
<td>36</td>
<td>2</td>
<td>2</td>
<td>38</td>
</tr>
<tr>
<td></td>
<td>bmi</td>
<td>31</td>
<td>7</td>
<td>7</td>
<td>38</td>
</tr>
<tr>
<td>Overall</td>
<td>age</td>
<td>142</td>
<td>12</td>
<td>12</td>
<td>154</td>
</tr>
<tr>
<td></td>
<td>bmi</td>
<td>126</td>
<td>28</td>
<td>28</td>
<td>154</td>
</tr>
</tbody>
</table>

Overall: 154
Consider heart attack data containing `hightar`, an indicator for smoking high-tar cigarettes.

```
. webuse mheart10s0
(Fict. heart attack data; bmi, age, hightar, & smokes missing; arbitrary pattern)
. mi describe
```

- Style: mlong
- last mi update 25mar2011 11:00:38, 66 days ago

- Obs.: complete 92
- incomplete 62 (M = 0 imputations)
- total 154

- Vars.: imputed: 4; bmi(24) age(30) hightar(19) smokes(14)
- passive: 0
- regular: 3; attack female hsgrad
- system: 3; _mi_m _mi_id _mi_miss

(there are no unregistered variables)
Explore missing-data patterns

```
. mi misstable patterns

Missing-value patterns
(1 means complete)

<table>
<thead>
<tr>
<th>Percent</th>
<th>Pattern</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>60%</td>
<td></td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>14</td>
<td></td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>10</td>
<td></td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>7</td>
<td></td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td></td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>&lt;1</td>
<td></td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>&lt;1</td>
<td></td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>&lt;1</td>
<td></td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>100%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Variables are (1) smokes (2) hightar (3) bmi (4) age
```

```
. mi misstable nested

1. smokes(14) -> hightar(19)
2. bmi(24)
3. age(30)
```
Impute `hightar` conditionally on `smokes`; check prediction equations prior to imputation (option `dryrun`)

```
. mi impute chained ///
  > (regress) age ///
  > (pmm) bmi ///
  > (logit) smokes ///
  > (logit, conditional(if smokes==1) omit(i.smokes)) hightar ///
  > = attack hsgrad female, dryrun
```

Conditional models:

- `smokes`: logit `smokes` `bmi` `age` `attack` `hsgrad` `female`
- `hightar`: logit `hightar` `bmi` `age` `attack` `hsgrad` `female`, conditional(if `smokes==1`)
- `bmi`: pmm `bmi` i.smokes i.hightar `age` `attack` `hsgrad` `female`
- `age`: regress `age` i.smokes i.hightar `bmi` `attack` `hsgrad` `female`
Prediction equations are as intended; proceed to imputation

```
. mi impute chained ///
> (regress) age ///
> (pmm) bmi ///
> (logit) smokes ///
> (logit, conditional(if smokes==1) omit(i.smokes)) hightar ///
> = attack hsgrad female, add(5)
```

Performing chained iterations ...

<table>
<thead>
<tr>
<th>Multivariate imputation</th>
<th>Imputations = 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chained equations</td>
<td>added = 5</td>
</tr>
<tr>
<td>Imputed: m=1 through m=5</td>
<td>updated = 0</td>
</tr>
<tr>
<td>Initialization: monotone</td>
<td>Iterations = 50</td>
</tr>
<tr>
<td></td>
<td>burn-in = 10</td>
</tr>
</tbody>
</table>

Conditional imputation:
- hightar: incomplete out-of-sample obs. replaced with value 0
  - age: linear regression
  - bmi: predictive mean matching
  - smokes: logistic regression
  - hightar: logistic regression

<table>
<thead>
<tr>
<th>Variable</th>
<th>Complete</th>
<th>Incomplete</th>
<th>Imputed</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>age</td>
<td>124</td>
<td>30</td>
<td>30</td>
<td>154</td>
</tr>
<tr>
<td>bmi</td>
<td>130</td>
<td>24</td>
<td>24</td>
<td>154</td>
</tr>
<tr>
<td>smokes</td>
<td>140</td>
<td>14</td>
<td>14</td>
<td>154</td>
</tr>
<tr>
<td>hightar</td>
<td>135</td>
<td>19</td>
<td>19</td>
<td>154</td>
</tr>
</tbody>
</table>

(complete + incomplete = total; imputed is the minimum across m of the number of filled-in observations.)
MICE is an iterative method—it's convergence needs to be evaluated

Recall imputation model for age and bmi from example 2 (here we use 3 nearest neighbors with PMM)

Let’s explore the convergence of MICE

. webuse mheart8s0
   (Fictional heart attack data; bmi and age missing; arbitrary pattern)
. set seed 38762
. mi impute chained (regress) age (pmm, knn(3)) bmi = attack smokes female hsgrad, 
   > chainonly burnin(50) savetrace(impstats)
Conditional models:
   age: regress age bmi attack smokes female hsgrad
   bmi: pmm bmi age attack smokes female hsgrad, knn(3)
Performing chained iterations ...
Note: no imputation performed.
Trace plots of means and standard deviations of imputed values

. use impstats
(Summaries of imputed values from -mi impute chained-)
. tsset iter
    time variable: iter, 0 to 50
    delta: 1 unit
. tsline bmi_mean, name(gr1) nodraw yline(25)
. tsline bmi_sd, name(gr2) nodraw yline(4)
. tsline age_mean, name(gr3) nodraw yline(56)
. tsline age_sd, name(gr4) nodraw yline(11.6)
. graph combine gr1 gr2 gr3 gr4, title(Trace plots of summaries of imputed values) > rows(2)

(Continued on next page)
Trace plots of summaries of imputed values
MICE uses separate independent chains to obtain imputations

Use `add()` instead of `chainonly` in combination with `savetrace()` to save summaries of imputed values from multiple chains

```
. webuse mheart8s0, clear
(Fictional heart attack data; bmi and age missing; arbitrary pattern)
. qui mi impute chain (regress) age (pmm, knn(3)) bmi = attack smokes female hsgrad,
   >  add(5) burnin(20) savetrace(impstats, replace)
```
Trace plots of means and standard deviations of imputed values from multiple chains

. use impstats, clear
(Summaries of imputed values from -mi impute chained-)
. reshape wide *mean *sd, i(iter) j(m)
(note: j = 1 2 3 4 5)

<table>
<thead>
<tr>
<th>Data</th>
<th>long</th>
<th>-&gt;</th>
<th>wide</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of obs.</td>
<td>105</td>
<td>-&gt;</td>
<td>21</td>
</tr>
<tr>
<td>Number of variables</td>
<td>6</td>
<td>-&gt;</td>
<td>21</td>
</tr>
<tr>
<td>j variable (5 values)</td>
<td>m</td>
<td>-&gt;</td>
<td>(dropped)</td>
</tr>
<tr>
<td>xij variables:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>age_mean</td>
<td>-&gt;</td>
<td>age_mean1 age_mean2 ... age_mean5</td>
<td></td>
</tr>
<tr>
<td>bmi_mean</td>
<td>-&gt;</td>
<td>bmi_mean1 bmi_mean2 ... bmi_mean5</td>
<td></td>
</tr>
<tr>
<td>age_sd</td>
<td>-&gt;</td>
<td>age_sd1 age_sd2 ... age_sd5</td>
<td></td>
</tr>
<tr>
<td>bmi_sd</td>
<td>-&gt;</td>
<td>bmi_sd1 bmi_sd2 ... bmi_sd5</td>
<td></td>
</tr>
</tbody>
</table>

--more--
. tsset iter
time variable:  iter, 0 to 20
delta:  1 unit
. tsline bmi_mean*, name(gr1) nodraw legend(off) ytitle(Mean of bmi) yline(25)
. tsline bmi_sd*, name(gr2) nodraw legend(off) ytitle(Std. Dev. of bmi) yline(4)
. tsline age_mean*, name(gr3) nodraw legend(off) ytitle(Mean of age) yline(56)
. tsline age_sd*, name(gr4) nodraw legend(off) ytitle(Std. Dev. of age) yline(11.6)
. graph combine gr1 gr2 gr3 gr4, title(Trace plots of summaries of imputed values > from 5 chains) rows(2)

(Continued on next page)
Trace plots of summaries of imputed values from 5 chains

Mean of bmi

Std. Dev. of bmi

Mean of age

Std. Dev. of age

Iteration numbers
1. Stata 12’s `mi` provides multivariate imputation using chained equations, `mi impute chained`, among other new features.
2. MICE is a very powerful and flexible imputation tool. Its flexibility, however, must be used with caution.
3. MICE has no formal theoretical justification but provides ways of capturing important data characteristics.
4. MICE is an iterative imputation method so its convergence needs to be evaluated.
5. As with any imputation method, the quality of imputations needs to be evaluated after MICE.
6. Careful modeling is required with MICE to avoid incompatible conditionals, although a few simulation studies suggest the impact of incompatible conditionals on final MI inference is minor.


