Chained equations and more in multiple imputation in Stata 12

Yulia Marchenko

Associate Director, Biostatistics
StataCorp LP

2011 UK Stata Users Group Meeting

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, and 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)

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, mcerror` 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:

$$
\begin{array}{l}
X_1^{(t+1)} \text{ using } X_2^{(t)}, X_3^{(t)}, \ldots, X_q^{(t)} \\
X_2^{(t+1)} \text{ using } X_1^{(t+1)}, X_3^{(t)}, \ldots, X_q^{(t)} \\
\qquad \ldots \\
X_q^{(t+1)} \text{ using } X_1^{(t+1)}, X_2^{(t+1)}, \ldots, X_{q-1}^{(t+1)}
\end{array}
$$

- 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
  obs:           154                       Fictional heart attack data;
                                             bmi and age missing; arbitrary
                                             pattern
 vars:             6                       1 Sep 2011 10:11
 size:         1,848
--------------------------------------------------------------------------------
              storage  display     value
variable name   type   format      label      variable label
--------------------------------------------------------------------------------
attack          byte   %9.0g                  Outcome (heart attack)
smokes          byte   %9.0g                  Current smoker
age             float  %9.0g                  Age, in years
bmi             float  %9.0g                  Body Mass Index, kg/m^2
female          byte   %9.0g                  Gender
hsgrad          byte   %9.0g                  High school graduate
--------------------------------------------------------------------------------
Sorted by:
```

- Let's summarize missing values

```
. misstable summarize, generate(Mis_)
                                                          Obs<.
                                               +------------------------------
               |                               | Unique
      Variable |     Obs=.     Obs>.     Obs<. | values        Min         Max
  -------------+--------------------------------+------------------------------
           age |        12                 142 |    142   20.73613    83.78423
           bmi |        28                 126 |    126   17.22643    38.24214
  -----------------------------------------------------------------------------
```

- and explore missing-data patterns

```
. misstable patterns
   Missing-value patterns
     (1 means complete)

              |   Pattern
    Percent   |  1  2
  ------------+-------------
       77%    |  1  1
              |
       16     |  1  0
        5     |  0  1
        3     |  0  0
  ------------+-------------
      100%    |
  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

```
. 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 ...
Multivariate imputation                     Imputations =        5
Chained equations                                 added =        5
Imputed: m=1 through m=5                        updated =        0
Initialization: monotone                     Iterations =       50
                                                burn-in =       10
               age: linear regression
               bmi: linear regression
```

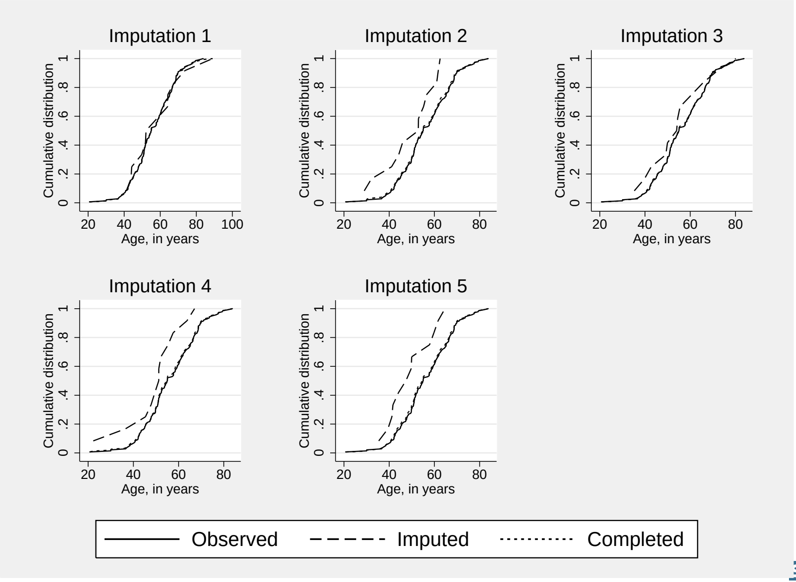
	Observations per m			
Variable	Complete	Incomplete	Imputed	Total
age	142	12	12	154
bmi	126	28	28	154

```
(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)
```

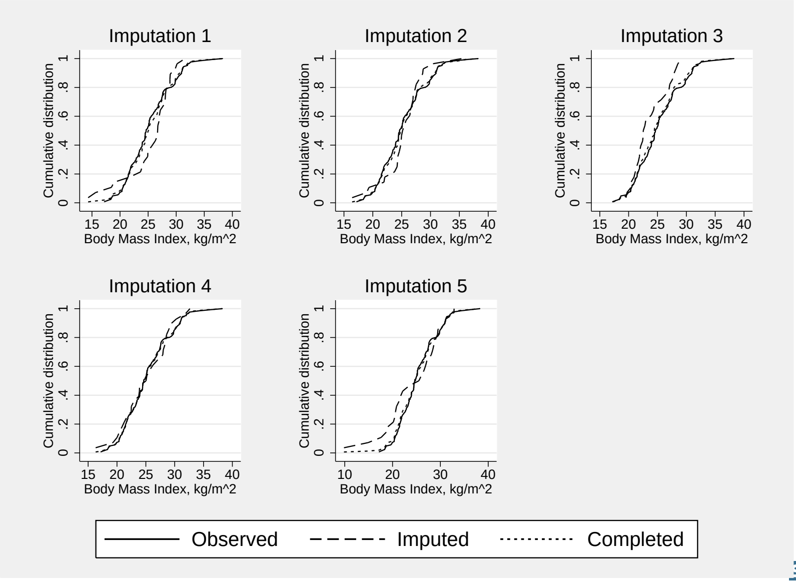
(*Continued on next page*)


Imputation 1
Imputation 2
Imputation 3
Imputation 4
Imputation 5
Cumulative distribution
0 .2 .4 .6 .8 1
Age, in years
20 40 60 80 100
Observed
Imputed
Completed

- 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*)


Imputation 1
Imputation 2
Imputation 3
Imputation 4
Imputation 5
Cumulative distribution
0 .2 .4 .6 .8 1
15 20 25 30 35 40
10 20 30 40
Body Mass Index, kg/m^2
Observed
Imputed
Completed

```
. mi estimate, mcerror cformat(%8.4f): logit attack smokes age bmi female hsgrad

Multiple-imputation estimates                   Imputations       =          5
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

------------------------------------------------------------------------------
      attack |     Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      smokes |    1.1326     0.3561     3.18   0.001       0.4347      1.8306
             |    0.0145     0.0009     0.04   0.000       0.0137      0.0155
             |
         age |    0.0372     0.0162     2.30   0.022       0.0054      0.0691
             |    0.0019     0.0003     0.12   0.007       0.0019      0.0021
             |
         bmi |    0.0935     0.0457     2.05   0.041       0.0039      0.1831
             |    0.0044     0.0011     0.11   0.011       0.0050      0.0048
             |
      female |   -0.1331     0.4171    -0.32   0.750      -0.9507      0.6844
             |    0.0195     0.0020     0.05   0.035       0.0209      0.0189
             |
      hsgrad |    0.1324     0.4019     0.33   0.742      -0.6553      0.9201
             |    0.0112     0.0007     0.03   0.021       0.0099      0.0126
             |
       _cons |   -5.2048     1.5652    -3.33   0.001      -8.2726     -2.1371
             |    0.0170     0.0163     0.03   0.000       0.0413      0.0304
------------------------------------------------------------------------------
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 ...
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
```

	Observations per m			
Variable	Complete	Incomplete	Imputed	Total
age	142	12	12	154
bmi	126	28	28	154

```
(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
```

Variable	Observations per m			
	Complete	Incomplete	Imputed	Total
age	142	12	12	154
bmi	126	28	28	154

```
(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
```

Variable	Observations per m Complete	Incomplete	Imputed	Total
age	142	12	12	154
bmi	126	28	28	154

```
(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 ...
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
```

Variable	Observations per m			
	Complete	Incomplete	Imputed	Total
age	142	12	12	154
bmi	126	28	28	154

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

               age: interval regression
               bmi: predictive mean matching
```

	Observations per m			
Variable	Complete	Incomplete	Imputed	Total
age	142	12	12	154
bmi	126	28	28	154

- 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
```

by() Variable	Observations per m Complete	Incomplete	Imputed	Total
female = 0				
age	106	10	10	116
bmi	95	21	21	116
female = 1				
age	36	2	2	38
bmi	31	7	7	38
Overall				
age	142	12	12	154
bmi	126	28	28	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)

              |   Pattern
    Percent   |  1  2  3  4
  ------------+-------------
       60%    |  1  1  1  1
              |
       14     |  1  1  1  0
       10     |  1  1  0  1
        7     |  0  0  1  1
        3     |  1  1  0  0
        2     |  1  0  1  1
        1     |  0  0  0  1
       <1     |  0  0  1  0
       <1     |  1  0  0  0
       <1     |  1  0  1  0
  ------------+-------------
      100%    |
  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 ...

Multivariate imputation                     Imputations =          5
Chained equations                                 added =          5
Imputed: m=1 through m=5                        updated =          0

Initialization: monotone                     Iterations =         50
                                                burn-in =         10

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

------------------------------------------------------------------
                   |               Observations per m             
                   |----------------------------------------------
          Variable |   Complete   Incomplete   Imputed |     Total
-------------------+-----------------------------------+----------
               age |        124           30        30 |       154
               bmi |        130           24        24 |       154
            smokes |        140           14        14 |       154
           hightar |        135           19        19 |       154
------------------------------------------------------------------

(complete + incomplete = total; imputed is the minimum across m
 of the number of filled-in observations.)
```

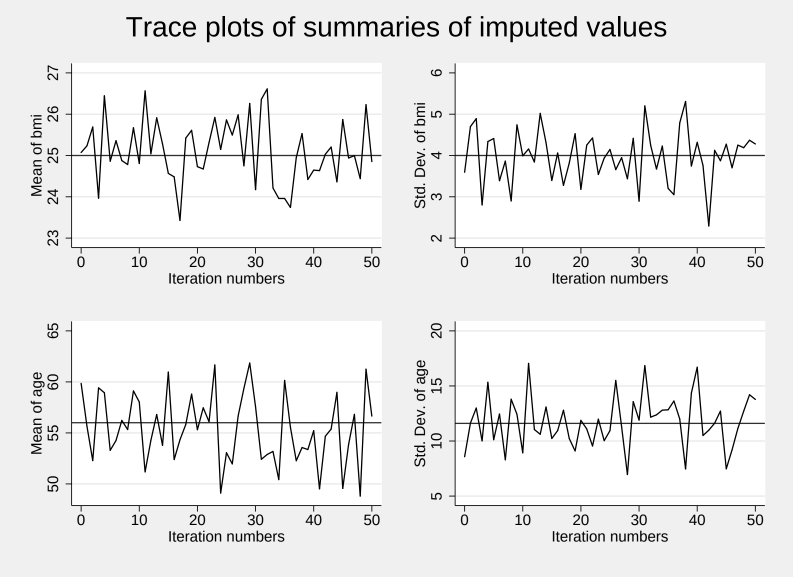
- MICE is an iterative method—its convergence needs to be evaluated
- Recall imputation model for age and bmi from example 2 (here we use 3 nearest neighbors with PMM)
- Let's explore the convergence of MICE

```
. 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
Mean of bmi
23
24
25
26
27
0
10
20
30
40
50
Iteration numbers
Std. Dev. of bmi
2
3
4
5
6
0
10
20
30
40
50
Iteration numbers
Mean of age
50
55
60
65
0
10
20
30
40
50
Iteration numbers
Std. Dev. of age
5
10
15
20
0
10
20
30
40
50
Iteration numbers

- 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

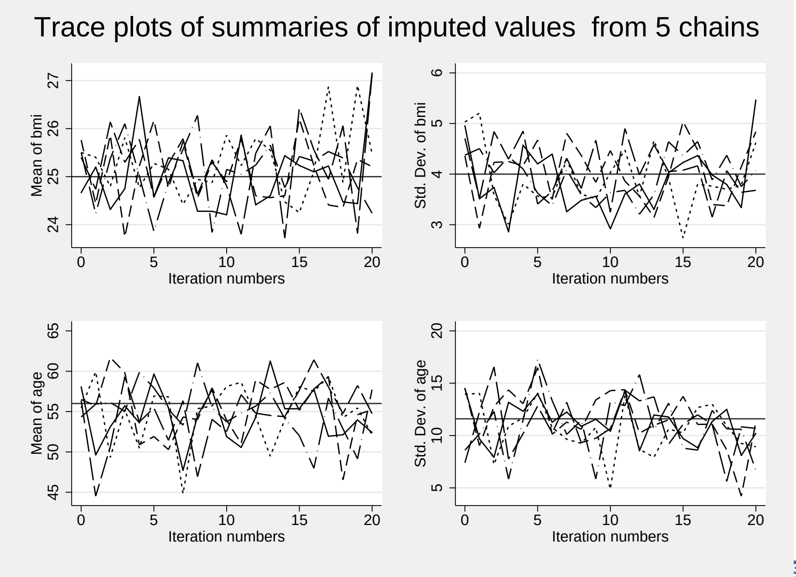
```
. webuse mheart8s0, clear
(Fictional heart attack data; bmi and age missing; arbitrary pattern)
. qui mi impute chain (regress) age (pmm, knn(3)) bmi = attack smokes female hsgrad,
>  add(5) burnin(20) savetrace(impstats, replace)
```

- 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)

Data                                long   ->   wide
-----------------------------------------------------------------------------

Number of obs.                       105   ->      21
Number of variables                    6   ->      21
j variable (5 values)                  m   ->   (dropped)
xij variables:
                                age_mean   ->   age_mean1 age_mean2 ... age_mean5
                                bmi_mean   ->   bmi_mean1 bmi_mean2 ... bmi_mean5
                                  age_sd   ->   age_sd1 age_sd2 ... age_sd5
                                  bmi_sd   ->   bmi_sd1 bmi_sd2 ... bmi_sd5
-----------------------------------------------------------------------------
--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
24
25
26
27
Std. Dev. of bmi
3
4
5
6
Mean of age
45
50
55
60
65
Std. Dev. of age
5
10
15
20
0
5
10
15
20
Iteration numbers

- Stata 12's `mi` provides multivariate imputation using chained equations, `mi impute chained`, among other new features
- MICE is a very powerful and flexible imputation tool. Its flexibility, however, must be used with caution.
- MICE has no formal theoretical justification but provides ways of capturing important data characteristics
- MICE is an iterative imputation method so its convergence needs to be evaluated
- As with any imputation method, the quality of imputations needs to be evaluated after MICE
- 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

Allison, P. D. 2001. *Missing Data*. Thousand Oaks, CA: Sage.

Arnold, B. C., E. Castillo, and J. M. Sarabia. 2001. Conditionally specified distributions: An introduction. *Statistical Science* 16: 249—274.

Carlin, J. B., J. C. Galati, and P. Royston. 2008. A new framework for managing and analyzing multiply imputed data in Stata. *Stata Journal* 8: 49—67.

Carlin, J. B., N. Li, P. Greenwood, and C. Coffey. 2003. Tools for analyzing multiple imputed datasets. *Stata Journal* 3: 226—244.

Lee, K. J., and J. B. Carlin. 2010. Multiple imputation for missing data: Fully conditional specification versus multivariate normal imputation. *American Journal of Epidemiology* 171: 624—632.

Marchenko, Y. V., and W. D. Eddings. 2011. A note on how to perform multiple-imputation diagnostics in Stata. http://www.stata.com/users/ymarchenko/midiagnote.pdf.

Raghunathan, T. E., J. M. Lepkowski, J. Van Hoewyk, and P. Solenberger. 2001. A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey Methodology* 27: 85—95.

Royston, P. 2004. Multiple imputation of missing values. *Stata Journal 4*: 227—241.

Royston, P. 2005a. Multiple imputation of missing values: Update. *Stata Journal* 5: 188—201.

Royston, P. 2005b. Multiple imputation of missing values: Update of ice. *Stata Journal* 5: 527—536.

Royston, P. 2007. Multiple imputation of missing values: Further update of ice, with an emphasis on interval censoring. *Stata Journal* 7: 445—464.

Royston, P. 2009. Multiple imputation of missing values: Further update of ice, with an emphasis on categorical variables. *Stata Journal* 9: 466—477.

Royston, P., J. B. Carlin, and I. R. White. 2009. Multiple imputation of missing values: New features for mim. *Stata Journal* 9: 252—264.

Rubin, D. B. 1987. *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley.

Rubin, D. B. 1996. Multiple imputation after 18+ years. *Journal of the American Statistical Association* 91: 473—489.

Schafer, J. L. 1997. *Analysis of Incomplete Multivariate Data*. Boca Raton, FL: Chapman & Hall/CRC.

van Buuren, S. 2007. Multiple imputation of discrete and continuous data by fully conditional specification. *Statistical Methods in Medical Research* 16: 219—242.

van Buuren, S., H. C. Boshuizen, and D. L. Knook. 1999. Multiple imputation of missing blood pressure covariates in survival analysis. *Statistics in Medicine* 18: 681—694.

van Buuren, S., J. P. L. Brand, C. G. M. Groothuis-Oudshoorn, and D. B. Rubin. 2006. Fully conditional specification in multivariate imputation. *Journal of Statistical Computation and Simulation* 76: 1049—1064.

White, I. R., P. Royston, and A. M. Wood. 2011. Multiple imputation using chained equations: Issues and guidance for practice. *Statistics in Medicine* 30: 377—399.