Multiple Imputation in Stata — \texttt{mi} and \texttt{ice} commands

Henrik Støvring
(stovring@biostat.au.dk)

Department of Biostatistics

\textbf{SCHOOL OF PUBLIC HEALTH}
\textbf{FACULTY OF HEALTH SCIENCES}
\textbf{AARHUS UNIVERSITY}

November 19, 2010
Outline

Background and terminology

Generating imputed datasets

Brief list of introductory references
Background and setting

- Assume we “know” that
  - Data belongs to MAR category
  - More than one variable has missings
  - Missing data pattern is not monotone

- What is needed in order to proceed?
  - Prediction model for missing values
  - Tool to impute the missing values
  - Tool to combine estimates from analysis of each imputed dataset into an overall estimate
**Terminology**

**Dataset zero:** The original dataset with missing values

**Imputed dataset:** A copy of the original dataset with all missing values replaced with imputed values \((j = 1, \ldots, m)\)

**Imputed values:** (Randomly generated) values substituted for unobserved values

**Multiple imputation analysis:** The ordinary analysis on each imputed dataset AND combination of estimates into a single estimate

**Iteration:** Do a procedure (compute some numbers), update starting values with the result (the computed numbers), and repeat the procedure

**Passive variable:** A variable that depends on an imputed variable
Obtaining and installing

- Obtained in Stata with the commands
  
  . search ice
  . net sj 9-3 st0067_4

  and then click “(click here to install)"

- Background information in
  
  - help-file and references therein
  - Royston (2009)
Rationale

- **M** ultiple (Imputation)
  - Iterated: Repeat to achieve stability...
- **C** hained: In a specific order, one by one...
- **E** quations: Based on a set of regression equations

- Consists of two “steps”
  1. Estimate relationships between each variable to be imputed and predictive variables (covariates)
  2. Impute values from fitted model
Initial observations regarding \texttt{-ice-}

- Variables take turn in being predictor and predicted ("outcome"/ to be imputed)
- Variables to be imputed can be predicted from variables without missing values
- Allows regression types for categorical data
  - Logistic
  - Multinomial/Ordered logistic
- Allows imputation of interval censored data

\textbf{H Støvring} \quad \textbf{Stata, MI, and ICE}
Major options

- **m()**: Number of imputed datasets
- **eq()**: Equations used to predict from
- **cmd()**: Regression type used to model “dependent” variable
- **by()**: Impute separately for subsets of dataset
- **cycles()**: Number of cycles used to obtain estimated regression coefficients used in subsequent prediction/imputation
Useful options

- **dryrun** Checks syntax and equations for consistency, but doesn’t do any actual imputations
- **saving()** Save imputed dataset to named file
- **seed()** Fix random origin, i.e. make process reproducible
Other options

- Countless:
  - clear
  - match()
  - passive()
  - boot()

- **WARNING:** Package is often updated AND options sometimes changes definition!
Imputing values: running \texttt{–ice–}

- Two types of approaches (at least!):
  1. “Black box” — minimal specification, maximal complexity:
      \[
      \begin{align*}
      \text{. ice total	extunderscore noncompl health incmean edulvl sex, m(10) clear}
      \end{align*}
      \]
      where \texttt{sex} is a binary indicator for gender
  2. “Dedicated” — detailed specification, transparent modeling
      \[
      \begin{align*}
      \text{. ice total	extunderscore noncompl health logincmean edulvl sex, ///}
      \text{~ m(10) clear ///}
      \text{~ eq(logincmean: total	extunderscore noncomp edulvl sex, ///}
      \text{~ total	extunderscore nomcomp: incmean edulvl sex) ///}
      \text{~ passive(incmean:exp(logincmean))}
      \end{align*}
      \]
      [etc....]
Advice and cautions

- Always include outcome in predictions of covariates
- Always check the reported equations used by \texttt{ice-}
- Always check that imputed values are sensible
  (Barnard and Meng (1999), Meng (1994))
- When you estimate many parameters, $m$ should be large
- Build from simple to complex:
  Try it out in a simple setting which you understand!
What you get from `-ice-`

- Output report on how it all went
- A dataset consisting of
  - $m + 1$ sub-datasets: The original plus $m$ imputed datasets
  - Two new variables:
    - `_mi_`: Record identifier across all imputed datasets, $i = 1, \ldots, n$
    - `_mj_`: Identifier of imputed dataset, $j = 1, \ldots, m$
Importing data into \texttt{mi} command family

- Stata has its own suite of commands for multiple imputation analysis: \texttt{mi}
- Requires
  - Specific organization of datasets
  - Specific naming of variables \texttt{\_mi\_id} different from those of \texttt{\_ice\_id}
  - Registration of relation between variables
- Obtained with
  - \texttt{mi import ice, clear auto}
- Creates variables: \texttt{\_mi\_m, \_mi\_id, \_mi\_miss}
Formats for multiply imputed datasets

- Stata has four different types of formats:
  - **wide**: Adds new variables with imputed values
  - **mlong**: Adds a new record for each missing value per imputed dataset
  - **flong**: Adds a full dataset per imputed dataset
  - **flongsep**: Stores each of the imputed dataset in separate files

- Two first are efficient, computationally and storagewise
- Latter two are transparent
-mi estimate-

- General syntax is similar to `-bysort variable : -`:
  
  . mi estimate, post: regress outcome covar1 covar2

  . mi estimate, or post: logit binoutcome covar1 covar2

- Runs regression on each imputed dataset and combines results using Rubin’s rule

- Leaves behind results that allows use of `-mi test-

- This part is the easiest!
Interactions

- Consider ESS data
- Suppose we are interested in
  - **Outcome**: Health (binary)
  - **Covariates**: Age (3 categories), Income (4 categories) and their interaction
- Should interaction term be included in imputations?
- Yes!
Not including interactions in imputations?

- Consider all subjects with missing health
- If we impute health based on Age and Income alone:
  1. In imputed subjects, the interaction is non-existing in any subsequent analysis
  2. For the entire dataset, the interaction effect (if truly present) becomes diluted
- Same reasoning basically if for example Income category is missing:
  If interaction effect is truly present, then this should be allowed for in imputations of Income
- Remember to use `passive` option of `ice` command!

PMID: 16980149.

Can differences in medical drug compliance between European countries be explained by social factors: analyses based on data from the European Social Survey, round 2.

*BMC Public Health* 9, 145.

*Statistical Analysis with Missing Data.*
New York: Wiley.

Multiple-Imputation inferences with uncongenial sources of input.
Missing data: dial M for ???  
*J. Amer. Statist. Assoc.* 95(452), 1325–1330.

Can one assess whether missing data are missing at random in medical studies?  
PMID: 16768297.

Multiple imputation of missing values: Further update of ice, with an emphasis on categorical variables.  
*Stata Journal* 9(3), 466–477(12).
Inference and missing data. 
*Biometrika* 63, 581–92.

*Multiple imputation for nonresponse in surveys.* 
Wiley Series in Probability and Mathematical Statistics: 

Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. 
*BMJ* 338, b2393.
Thank you for your attention!

Slides prepared with \LaTeX and Beamer