

Recent Advances in Missing Data Methods: Multiple Imputation by Chained Equations

Elizabeth A. Stuart

Johns Hopkins Bloomberg School of Public Health
Department of Mental Health
Department of Biostatistics
estuart@jhsph.edu
www.biostat.jhsph.edu/~estuart

AcademyHealth Annual Research Meeting
June 27, 2010

- 1 Introduction and terminology
 - Understanding types of missingness

- 1 Introduction and terminology
 - Understanding types of missingness
- 2 Ways of handling missing data
 - (Generally) improper ways of handling missing data...
 - Better ways of dealing with missing data...

- 1 Introduction and terminology
 - Understanding types of missingness
- 2 Ways of handling missing data
 - (Generally) improper ways of handling missing data...
 - Better ways of dealing with missing data...
- 3 Implementing multiple imputation: MICE

- 1 Introduction and terminology
 - Understanding types of missingness
- 2 Ways of handling missing data
 - (Generally) improper ways of handling missing data...
 - Better ways of dealing with missing data...
- 3 Implementing multiple imputation: MICE
- 4 Conclusions

- 1 Introduction and terminology
 - Understanding types of missingness
- 2 Ways of handling missing data
 - (Generally) improper ways of handling missing data...
 - Better ways of dealing with missing data...
- 3 Implementing multiple imputation: MICE
- 4 Conclusions
- 5 Software

- 1 Introduction and terminology
 - Understanding types of missingness
- 2 Ways of handling missing data
 - (Generally) improper ways of handling missing data...
 - Better ways of dealing with missing data...
- 3 Implementing multiple imputation: MICE
- 4 Conclusions
- 5 Software
- 6 References

- 1 Introduction and terminology
 - Understanding types of missingness
- 2 Ways of handling missing data
 - (Generally) improper ways of handling missing data...
 - Better ways of dealing with missing data...
- 3 Implementing multiple imputation: MICE
- 4 Conclusions
- 5 Software
- 6 References

Course description

Missing data is a problem in almost every health services research study, and standard ways of dealing with missing values, such as complete case analysis, are generally inappropriate. This session will discuss the drawbacks of traditional methods for dealing with missing data and describe why newer methods, such as multiple imputation, are preferable. Panelists will focus in particular on Multiple Imputation by Chained Equations, which is particularly useful for large datasets with complex data structures, as is often encountered in health services research. Emphasis will be on providing practical tips and guidance for implementing multiple imputation and analyzing and interpreting multiply imputed data, including sample R, SAS, and Stata code. Level of Difficulty: Intermediate. Participants should have an understanding of regression analysis and the principles of statistical inference.

- Missing data common, especially with administrative data or sensitive surveys
- Advanced methods have been developed to handle missing data
- But how do we actually implement those methods?
- What are the implications for analyses?

Why should you pay attention?

Ignoring or inappropriately handling missing data may lead to...

- Biased estimates
- Incorrect standard errors
- Incorrect inferences/results

Lots of reasons for missingness...

- Non-response/attrition
- Data entry errors
- Administrative data with missing values
- Lost survey forms
- Individuals not wanting to disclose (or not knowing) particular information
- Note: sometimes entire variables are missing in that they are “latent”; we will generally not be talking about those types of variables

More formally... “Missing data mechanisms”

Need to understand what led to missing values

- **Missing Completely at Random (MCAR):** Missingness is totally random; does not depend on anything
 - $P(R|Y, X) = P(R|Y, X^{obs}, X^{mis}) = P(R|\psi)$
 - Cases with missing values a random sample of the original sample
 - No systematic differences between those with missing and observed values
 - Generally unrealistic, although may be reasonable for things like data entry errors

- **Missing At Random (MAR):** Missingness depends on observed data

- $P(R|Y, X) = P(R|Y, X^{obs}, \psi)$
- e.g., women more likely to respond than men
- So there are differences between those with observed and missing values, but we observe the ways in which they differ
- Can use weighting or imputation approaches to deal with the missingness
- This is probably the assumption made most frequently
- Satisfied for data missing by design
- Including a lot of predictors in the imputation model can make this more plausible

- **Not Missing At Random (NMAR):** Missingness depends on unobserved values
 - $(R|Y, X)$ cannot be simplified
 - e.g., probability of someone reporting their income depends on what their income is
 - e.g., probability of reporting prior arrests depends on whether or not they had previously been arrested
 - e.g., probability of reporting prior arrests depends on whether or not they are left-handed, and we do not observe left-handedness for anyone
 - i.e., even among people with the same values of the observed covariates, those with missing values on Y have a different distribution of Y than do those with observed Y
 - So we can't just use the observed cases to help impute the missing cases
 - Unfortunately no easy ways of dealing with this...have to posit some model of the missing data process
 - Siddique and Belin (2008), Hedeker and Gibbons (1997)

Of course those are assumptions...

- Never know which of them is correct
- Can do diagnostics for whether missingness is MCAR vs. (MAR or NMAR)
 - Does the probability of missingness depend on other variables?
 - e.g., are the mean ages of people with missing and non-missing values of drug use behavior different?
 - e.g., in a logistic regression predicting missingness on some variable, are there other variables that are significant predictors?
- But never know for sure if missingness is MAR or NMAR...
 - Have to use substantive understanding of what might have led to missing values
 - Helps to have a good understanding of the data collection process

- 1 Introduction and terminology
 - Understanding types of missingness
- 2 Ways of handling missing data
 - (Generally) improper ways of handling missing data...
 - Better ways of dealing with missing data...
- 3 Implementing multiple imputation: MICE
- 4 Conclusions
- 5 Software
- 6 References

Inappropriate ways of handling missing data

- Ignoring it
- Complete case
- Single imputation
- Missing indicator approach
- Last observation carried forward

- Common approach is to “ignore” it; just run models without doing anything about missingness
- Then what is done will depend on the defaults of the software
- Usually will be the same as complete-case analyses, discussed next

Complete case analysis

- Restrict analyses to individuals with observed data
- Generally bad!
 - Assumes missingness is MCAR
 - Often results in lots of cases dropped...decreased power and loss of representativeness (Little and Rubin, 2002; page 42)
 - Generally leads to biased results
- Is also model-dependent...will mean that different analyses may use different subsets of the data (unless do big restriction at the beginning)
- Very common...

Single imputation: Fill in (“impute”) each missing value

Ways of doing that imputation:

- Mean
- Regression prediction (“conditional mean imputation”)
 - e.g., impute mean within categories of observed covariates (gender, race, etc.)
 - e.g., fit regression model among observed cases, use to predict response for individuals with missing values

$$\hat{Y}_i = \hat{\alpha} + \hat{\beta}X_i$$

- Regression prediction plus error (“stochastic regression imputation”)
 - Like regression prediction, but also add random error

$$\hat{Y}_i = \hat{\alpha} + \hat{\beta}X_i + e_i, e_i \sim N(0, \hat{\sigma}^2)$$

- “Hot-deck”
 - For an individual with missing data, find individuals with the same observed values on other variables, randomly pick one of their values as the one to use for imputation
- Predictive mean matching
 - Like a combination of regression prediction and hot-deck
 - Take observed value from someone with similar predicted value

Other (inappropriate) strategies

- Missing data indicator
 - Do simple imputation and include indicator of missingness as an additional predictor in regression models
 - Doesn't work very well and can lead to bias (Vach and Blettner 1991, Donders et al. 2006, Greenland and Finkle 1995)
- Last observation carried forward
 - For longitudinal studies
 - If someone drops out of study, the last value observed for them is "carried forward" (copied) to later time points
 - But generally biased (Carpenter et al. 2004; Cook, Zeng, and Yi, 2004; Jansen et al. 2006)

Summary of single imputation approaches

- Best are regression prediction plus error or hot-deck (based on categorical versions of all of the variables observed)
- Can be reasonable, especially if not a lot of missing data, e.g., $< 5\%$ (Graham 2008)
- BUT...results in overly precise estimates
 - Analyses following single imputation do not know that some of the values have been imputed
 - Simply treats all of the values as observed values
 - So does not take into account the uncertainty in the imputations
- Anti-conservative...results will have more significance, narrower confidence intervals, than they should (Donders et al. 2006)
 - Higher Type I error rates
- So what to do instead?

Appropriate ways of handling missingness

- Maximum likelihood
 - Weighting
 - Multiple imputation
-
- Remember: Goal is not to get correct predictions of missing values; goal is to obtain accurate parameter estimates for relationships of interest

Maximum likelihood

- In some cases, maximum likelihood approaches exist
- Directly maximize the likelihood function, $f(X, Y)$
- Use observed values, take missingness into account
- e.g., longitudinal analyses that use the observations available for each person and correctly account for the missing observations
- When ML methods exist, can work very well (e.g., Mplus, LISREL)
- But they don't always exist so not always a feasible option
- Another drawback is that you cannot use auxiliary information to improve the predictions; uses only the variables in the actual analysis
- Graham (2008), Siddique et al. (2008)

Nonresponse weighting

- Often used to deal with attrition
- Generate model predicting non-response given observed covariates
- Weight respondents by their inverse probability of response
 - Weights the respondents up to represent the full sample
 - Same idea as survey sampling weights
- Use analysis methods that allow for weights (e.g., survey packages)
- Works well for simple missing data patterns (e.g., attrition)
- Oh and Scheuren (1983), Kalton and Flores-Cervantes (2003), Horton and Lipsitz (1999), Carpenter et al. (2006)

Multiple imputation

- Same idea as single imputation, but fills in each missing value multiple times
 - Like repeating the stochastic mean imputation multiple times
- Creates multiple (e.g., 10) “complete” data sets
- Analyses then run separately on each dataset and results combined across datasets
 - Standard “combining rules” (Rubin 1987)
 - (Software will do this for you)
- Total variance a function of within-imputation variance and between-imputation variance
 - Takes into account the uncertainty in the imputations
- Also nice because very general: same set of imputations can be used for many analyses
 - “Imputer” may be different from “analyst”

- 1 Introduction and terminology
 - Understanding types of missingness
- 2 Ways of handling missing data
 - (Generally) improper ways of handling missing data...
 - Better ways of dealing with missing data...
- 3 Implementing multiple imputation: MICE**
- 4 Conclusions
- 5 Software
- 6 References

Traditional approach to creating multiple imputations

Joint model of all variables

- e.g., multivariate normal distribution
- Fit using the observed cases
- Used to predict (multiple times) the missing values
- Sometimes multivariate normal model 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
- Software: norm, mix, SAS proc mi, Stata mi

Newer approach: Multiple imputation by chained equations (MICE)

- Fit model of each variable, conditional on all others
- Iterate fitting model and imputing each variable
- Models used depend on type of variable (categorical/continuous/binary)
- Raghunathan et al. (2001), van Buuren et al. (2006)
- Also called “fully conditional specification” or “sequential regression multiple imputation”

Example of MICE

3 variables: X_1 (binary), X_2 (continuous), X_3 (ordinal)

Steps in MICE:

- 1 Do simple imputations to fill in missing values for X_1 , X_2 , X_3
- 2 Using cases with observed X_1 , fit logistic regression model of $X_1 \sim X_2 + X_3$; predict missing values of X_1
- 3 Using cases with observed X_2 , fit normal regression model of $X_2 \sim X_1 + X_3$; predict missing values of X_2
- 4 Using cases with observed X_3 , fit proportional odds regression model of $X_3 \sim X_1 + X_2$; predict missing values of X_3
- 5 Iterate Steps 2-4
- 6 Repeat Step 5 to get multiple imputations

Pros and cons of MICE

- Benefits
 - Can more easily work in large datasets
 - Models can more accurately reflect distribution of each variable
 - Allows bounds (e.g., age started smoking)
 - Incorporates restriction to subpopulations (e.g., age started smoking)
- Drawbacks
 - Potentially less principled than joint mode
 - Doesn't necessarily imply a proper joint distribution
 - (Although this doesn't seem to be a big problem in practice)

Software to implement MICE

- SAS and stand-alone: IVEWare
 - Stata: ice
 - R: mice, mi
-
- IVEWare used to create multiple imputations of National Health Interview Survey (NHIS) for public-use

Steps to implementing MI methods

- 1 Examine rates and patterns of missingness, and any predictors of missingness
- 2 Generate imputations
- 3 Diagnose and assess imputations
- 4 Analysis

Motivating example: The CMHI Evaluation

- Goal: Develop service systems to provide comprehensive mental health services to children and their families
- Since 1993, the Center for Mental Health Initiatives (CMHI) has funded 126 grantees and served over 83,000 children
- Monitoring data available
 - 9,186 youth
 - In 45 sites
 - 396 variables to be imputed (demographics, behavior, substance use, delinquency, etc.)
- But lots of missingness
- Data will be imputed and then publicly released, for potentially broad (and diverse) use
- Work done under NIMH grant to Johns Hopkins University and Macro International (NIMH 1R01MH075828-01A1)
- Stuart et al. (2009)

Step 1: Rates of missingness in CMHI data

High rates of missingness for some variables

Variable	% Missing
Date of birth	1.7
Sex	1.7
Race	10.8
Family income	11.9
DSM-IV diagnoses	23.8
% of day in special ed	40.0

Also varies across sites

Missingness depends on observed characteristics

Table 1. Comparison of Characteristics of Children With Observed and Missing Values on the Internalizing Symptoms Scale

Characteristic	Children With Missing Internalizing Scale Values, %	Children With Observed Internalizing Scale Values, %	P Value (2-Sided)
American Indian	18.3	6.3	0.00
Caucasian	49.5	61.0	0.00
Hispanic	10.9	13.0	0.01
Conduct disorder	15.8	8.9	0.00
Eligible for Medicaid	74.5	69.2	0.00
ADHD	36.3	42.0	0.00
Currently receiving services	59.6	65.8	0.00

- Not MCAR
- Also varies a lot across sites
- Don't have reason to think missingness is NMAR so comfortable with MAR

Step 2: Generate imputations

- Need to specify model for each variable, conditional on all other variables
- Check to see if transformations make sense (e.g., to look more normally distributed)
- May make sense to include some variables in imputation model, even if not going to be used in analyses (“auxiliary variables”)
 - i.e., include more rather than fewer variables in imputation procedure
 - There may be some that are very predictive of missing values, even if they aren't of primary interest in analyses
- Often difficult to do careful model selection for each variable

Model specification

- IVEWare allows the use of stepwise selection to select the imputation model for each variable
- Uses some criteria (e.g., # of predictors, minimum marginal R^2)
 - Smaller minimum marginal R^2 will lead to more variables being included
 - Worth trying a couple of different values (He et al., 2009)
- CMHI: Used minimum additional $R^2 = 0.01$
 - Also did sensitivity analysis with 0.005
- Note: Not all packages have this feature
 - By default, mice (R) and ice (Stata) use all variables as predictors for all other variables
 - Can also specify particular models (see documentation)
 - For ice (Stata), additional add-on function does allow stepwise selection: `pred_eq`, `check_eq`

- Imputation and analysis compatibility
 - Imputation model should be more general than analysis model that will be used: otherwise risk finding null effects simply because data imputed assuming no relationship between variables
 - May want to force some variables into the models even if do stepwise
- How many imputations to generate?
 - Conventional advice has been 5-10, but more (e.g., 40) may yield increased power (Graham, Olchowski, & Gilreath, 2007)
 - Note: SAS seems better able to handle large datasets and large numbers of imputations than Stata

- Importance of auxiliary variables
 - Can be very beneficial to include “auxiliary variables:” not of interest in the analysis in and of themselves, but might help with the imputations
 - Collins et al. (2003) show that not much cost to including these extra variables and they can help a lot
 - Including a lot of variables can also make MAR assumption more reasonable
 - (No easy way to incorporate this extra information in maximum likelihood approaches)

Step 3: Diagnosing and assessing imputations

- Try to identify potentially problematic variables
- Two types of comparisons:
 - Before and after imputation
 - Across two imputation sets with slightly different settings (e.g., different criteria in the stepwise model)
- Most packages have very limited diagnostics
 - R's mi and mice packages have the most
- Note: Differences don't mean something is wrong! Could be because of differences in the types of people with observed vs. missing data

- Bivariate scatterplots of observed and imputed values
- Residual plots, for observed and imputed values
- Density plots of observed and imputed values
 - Example from Stuart et al. (2009); Figure 1

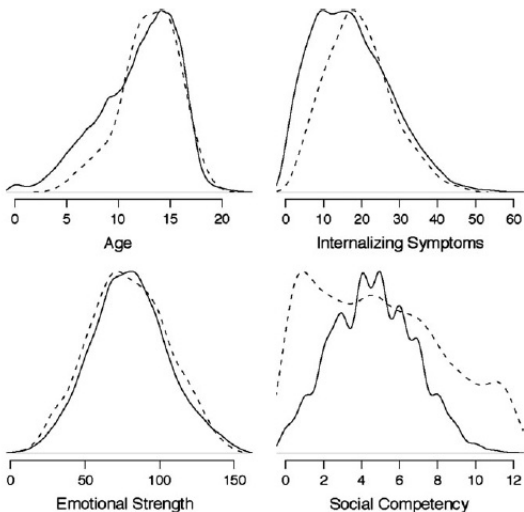


Figure 1. Comparison of observed and imputed values for 4 representative variables. For each variable, the solid line shows the density plot of observed values and the dashed line the density plot of imputed values. Age is expressed in years. The other measures are the Child Behavior Checklist (CBCL) internalizing syndrome score (33), the Behavioral and Emotional Rating Scale (34), and the CBCL total social

- Some packages automatically print out some diagnostics
- Model fit information to assess imputation models
- Comparisons of original and imputed values
 - Should check that imputations look reasonable (e.g., compare means, correlation coefficients)
 - Make sure values being imputed are in the correct ranges
- Potentially flag problematic variables to warn future data analysts

- Cross-validation approach of imposing random missingness after imputation; impute again, see how well it recovers values (Gelman, King, and Liu, 1998)
- Posterior predictive checks (He et al., 2009)
 - Compare estimates from the complete data (observed plus imputed) to estimates from simulated data generated solely from the models
 - May help identify parameters for which the imputation was not appropriate

Step 4: Analyses

- Combining rules allow the combination of results across the multiply imputed data sets (Rubin 1987)
 - Account for both within- and between-imputation variance
- Run analysis separately within each “complete” dataset, then combine across datasets
- Software packages have automated version of this for many models
 - Stata: `mim`, `mifit`
 - SAS: `proc mianalyze`
 - HLM: multiple imputation options
 - Mplus: multiple imputation command
- For other models, may need to do it “by hand”
- Final results just need to report the combined estimates; interpret as you would a standard regression

The math behind the combining

- \hat{Q}_j = estimate of scalar quantity of interest (e.g., regression coefficient) from complete dataset j
- U_j = standard error of \hat{Q}_j
- Overall estimate just the average of the estimates from each complete dataset

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

- For the overall variance, first calculate the average within-imputation variance (U) and the between-imputation variance (B)

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

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

- The total variance of \bar{Q} is then

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

- Degrees of freedom for t distribution can also be calculated
- See Schafer (1997) or Little and Rubin (2002) for details

Example: mim command in Stata

```
mim: logit dsmmood sex age
```

```
Multiple-imputation estimates (logit)  
Logistic regression
```

```
Imputations =      10  
Minimum obs  =     9185  
Minimum dof  =       4.8
```

```
-----  
dsmmood|  Coef.      Std. Err.      t    P>|t|    [95% Conf. Int.]    MI.df  
-----+-----  
sex|    .470223   .060136     7.82   0.000   0.346444   0.594003    25.3  
age|    .099599   .019677     5.06   0.004   0.04845   0.150748     4.8  
cons| -2.05873    .257295    -8.00   0.001  -2.72791  -1.38954     4.8  
-----
```

- 1 Introduction and terminology
 - Understanding types of missingness
- 2 Ways of handling missing data
 - (Generally) improper ways of handling missing data...
 - Better ways of dealing with missing data...
- 3 Implementing multiple imputation: MICE
- 4 Conclusions**
- 5 Software
- 6 References

- Isn't imputation “making up” data?
 - No! It is creating our best guesses at the missing values
 - In fact non-imputation methods (e.g., complete case analysis) generally rely on much stronger assumptions
 - Also important to note that we aren't assuming that we are imputing the correct values...generating the imputations only as an intermediate step to estimating the model parameters of real interest
- What if the imputation model is wrong?
 - Usually it's fine; most results indicate that MI still works well even if the imputation models are not correct (Schafer 1997)
 - Can help the situation by, for example, taking logs to make data more normally distributed when using linear regression

- Are there guidelines for how much missingness is “too much”?
 - Unfortunately, no
 - And remember that if there is not good information in the data to do imputations (i.e., not much that is predictive of the missing values), MI will take that into account by making the imputations very variable
 - Good results have been found with over 40% missingness
 - Key quantity is the fraction of missing information (Schafer 1997), which combines the % missing with how correlated the missing variable is with observed values
- What if you are really worried about NMAR?
 - Sensitivity analyses (e.g., Brame and Paternoster, 2003; Siddique and Belin 2008)
 - Pattern mixture models (fitting model separately for each missing data pattern; e.g., Hedeker and Gibbons, 1997)

- Should I impute a scale or the individual items?
 - Impute the scale if: (1) over half of the individual items observed if any are observed, (2) items have high α 's, and (3) the item-total correlations are similar across items (Graham, 2008)
 - Otherwise (and if have the code to recreate the scales), impute the items
- Should I impute raw or standardized scores?
 - Assuming you have the ability to recreate the standardized scores...
 - Impute whichever one looks more normally distributed

- What about data with a multilevel structure?
 - Not a lot of guidance on this
 - If analysis will have only random intercepts, can include cluster indicators as possible predictors (this done in CMHI data)
 - If analysis will have random intercepts and slopes (i.e., if going to look at relationships between variables separately for different clusters), impute separately within each cluster or include cluster*variable interactions in imputation model (Graham, 2008)
 - Yucel (2008), Reiter et al. (2006)
- What if imputing data within a study estimating causal effects?
 - Very hard to impute “treatment” indicator...maybe just drop people missing treatment status
 - Important to include the outcome in the imputation model, and in fact best to include lots of interactions between treatment status, covariates, and outcomes in imputation model (Moons et al., 2006)
 - (Want to make sure not to impose a treatment effect on the imputations)

- Should I include variables that are predictive of the missingness or predictive of the missing values?
 - Ideally would be inclusive and include any variables that may be related to the missingness AND/OR the values themselves
 - If can't do that (e.g., small samples), better to include variables predictive of the missing values
- What should I do if some analysis I want to do isn't covered by any of the existing packages that analyze multiply imputed data?
 - If just exploratory (e.g., regression diagnostics, graphics), run it on 2-3 of the imputed datasets separately and see how consistent the results are. If results consistent, just go with them. If not consistent, rethink imputations: why are they so variable?
 - If want to actually estimate models, will need to write code to do the combining across datasets yourself
 - The `mitools()` package for R gives some examples of this, makes it easy if you can send it coefficient estimates and their associated variances

Overall lessons

- Missing data can have serious implications for analyses
- Requires making assumptions about the missingness and missing values
- Best approach: Minimize the amount of missing data up front
 - Invest substantial resources in following up individuals
 - Design surveys to encourage full response
 - Explore alternative data sources (e.g., administrative records) as necessary
- Important to have a good understanding of the missing data process
 - Why were some cases missing?
 - How plausible is MAR? Are we worried about NMAR?
 - Can we collect additional data that will inform about the missingness?
 - e.g., for attrition, can ask in earlier waves about individual's likelihood of answering subsequent surveys
 - Is it possible to follow-up a subsample of those who initially did not respond?

Benefits of multiple imputation

- Yields accurate standard errors and inferences
- Allows the use of auxiliary variables to improve imputations
- Can be used for very general settings
 - Can impute a dataset and then use for lots of different analyses (Stuart et al. 2009)
 - Imputer and analyst can be different people
 - Analyses run on “complete” data sets and so any type of analysis can be run

Lessons for doing imputation

- If rates of missingness low (e.g., 1-2%), consider doing single imputation (e.g., regression prediction with noise)
- Make imputation models very general: lots of terms and interactions (little cost to including lots of potential predictors)
- MICE can be a very useful method for dealing with missing data
- Compare distributions of data pre- and post-imputation
 - Determine ways to summarize the results across variables
- If others will be using the imputed data, make clear documentation
 - Specify models used, interactions included
 - Highlight potentially problematic variables

Outline

- 1 Introduction and terminology
 - Understanding types of missingness
- 2 Ways of handling missing data
 - (Generally) improper ways of handling missing data...
 - Better ways of dealing with missing data...
- 3 Implementing multiple imputation: MICE
- 4 Conclusions
- 5 Software**
- 6 References

- <http://web.inter.nl.net/users/S.van.Buuren/mi/html/software.htm>
- Code for many packages in Horton and Kleinman (2007)
 - <http://www.math.smith.edu/muchado-appendix.pdf>

- mi package
 - <http://cran.r-project.org/web/packages/mi/index.html>
 - For creating and analyzing multiply imputed data
 - Multiple imputation by chained equations incorporated with predictive mean matching
 - Lots of good diagnostics
 - Automatically determines the correct model (e.g., linear vs. multinomial logit)
 - Goal is for the software to handle many complexities automatically (like collinearity, perfect prediction)
 - Eventually want to include procedures to explicitly handle time-series data and multilevel/clustered data
 - Not a lot of functionality currently, but check back soon...should be good!

- mice package

- <http://web.inter.nl.net/users/S.van.Buuren/mi/html/mice.htm>
- For creating and analyzing multiply imputed data
- Multiple imputation by chained equations
- Some good diagnostics, added functionality recently
- Can incorporate bounds, restrictions

Analyzing MI data in R

- mice and mi packages have built in commands
- mitools package
 - Run `imputationList()` command to combine the imputed datasets (could be from mice or from another package)
 - Use the `with()` command to run analysis on each complete dataset in the `imputationList` object
 - Use `micombine()` command on the results from the `with()` command to get results pooled across the complete datasets
 - Very general: Can run as long as you have the estimates and variances from each complete dataset
 - <http://cran.r-project.org/web/packages/mitools/mitools.pdf>
- Zelig package
 - Can run almost any model
 - Just say `data=mi(dataset1, dataset2, ...)` in the command

- IVEWare (stand-alone as well)
 - <http://www.isr.umich.edu/src/smp/ive/>
 - For creating multiple imputations
 - Multiple imputation by chained equations
 - More details below
- proc mianalyze
 - <http://www.sas.com/rnd/app/papers/mianalyzev802.pdf>
 - For analyzing multiply imputed data
 - Can be run on data imputed using proc mi or imputed using another package
 - Horton and Kleinman (2007) appendix shows code for reading multiply imputed data into SAS and running mianalyze

Code for IVEWare

```
proc printto print='C:\myoutdir\summerinst.lst' NEW;
run;

LIBNAME MYLIB1 'C:\myindir';
LIBNAME MYLIB2 'C:\myoutdir';
options set = SRCLIB "C:\Program Files\SAS\srclib"
sasautos=('!SRCLIB' sasautos) maautosource;
run;
%IMPUTE (NAME=MYSETUP,DIR=C:\myoutdir,SETUP=NEW);
DATAIN MYLIB1.sinst;
DATAOUT MYLIB2.impsinst1;

DEFAULT categorical;

COUNT totchild totadu susa5 ;

CONTINUOUS age bersraw ctotcomr ctotraw cintraw cextraw
ytotraw yintraw yextraw ars a1_rs ar1_s a1_r1_s;
```

```
DROP cohort totrole;

TRANSFER childid axis_1a ;

RESTRICT susa5(susa1=1, atleast11=1) susb13a(atleast11=1)
susb13b(susb13a=1,atleast11=1) ds1(atleast11=1)
ds2(atleast11=1) ds3(atleast11=1) ;

BOUNDS susa5(>=0,<=40) ctotraw(>=0) cintraw(>=0) cextraw(>=0),
ytotraw(>=0) yintraw(>=0) yextraw(>=0) ctotcomr(>=0)
bersraw(>=0) ;

MINRSQD .01;
ITERATIONS 10;
SEED 24578
MULTIPLES 1;

print coef;
run;

proc printto;
run;
```

- ice
 - <http://ideas.repec.org/c/boc/bocode/s446602.html>
 - <http://www.ats.ucla.edu/stat/Stata/library/ice.htm>
 - For creating multiple imputations
 - Multiple imputation by chained equations
 - More details below
- For analyzing mi data: micombine, mim, mi estimate
- Note: Stata's new multiple imputation procedure "mi" does not do MICE (Stata 11)
 - But can easily go between ice and mi procedures using "mi import ice" and "mi export ice" commands

Using ice in Stata

- To install ice: `ssc install ice, replace`
- To install mim: `ssc install mim, replace`
- Main command:


```
ice cohort sex age income totchild totadu nrace3 nrace5 nrace7  
totrole bersraw ctotcomr ctotraw cintraw cextraw ytotraw yintraw  
yextraw i.siteid, clear;
```
- Default is to let each variable be regressed on all other variables
 - Often run into convergence/collinearity issues
 - Can also specify particular regression models for each variable
 - Not as feasible as IVEware for large datasets
- <http://www.ats.ucla.edu/stat/Stata/library/ice.htm>

- Passive imputation: for variables that are a direct function of others (e.g., interactions)
 - Need to make sure the imputations are consistent with each other
 - “passive” option
 - `passive(sexxrace1: sex*nrace1 \ sexxrace3: sex*nrace3)`
- Specify regression model to be used
 - e.g., default for categorical is multinomial logit (unordered), but what if want to use ordered logit?
 - “cmd” option
 - `cmd(income:ologit)`

- Specify predictors in particular regression models
 - ice doesn't do stepwise, so what if want to use simpler model (not include all variables as predictors)?
 - "eq" option
 - `eq(income: sex cintraw, cextraw: nrace1 nrace2)`
 - (Note: of course the models in previous line make no sense; no reason to do that, but this could be useful to, e.g., exclude certain predictors from particular models)
 - Can also use user-written `pred_eq` and `check_eq` functions to facilitate stepwise models within context of ice
- Impute categorical variables as categories, but when predictors use series of dummy variables
 - "sub" option
 - `passive(\ inc1:(income==1) \ inc2:(income==2)) sub(income: inc1 inc2)`
 - (Assuming just 2 levels of income variable)

- 1 Introduction and terminology
 - Understanding types of missingness
- 2 Ways of handling missing data
 - (Generally) improper ways of handling missing data...
 - Better ways of dealing with missing data...
- 3 Implementing multiple imputation: MICE
- 4 Conclusions
- 5 Software
- 6 References

References: General references on missing data and MICE

- <http://missingdata.org.uk/>
- <http://www.stat.psu.edu/~jls/mifaq.html>
- Allison, P.D. (2002) *Missing Data in Quantitative Applications in the Social Sciences*. Thousand Oaks, CA. Sage.
- Carpenter, J. (2006). Missing Data Example Analysis. Available at <http://www.lshtm.ac.uk/msu/missingdata/example.html>
- Schafer, J.L. (1997) *Analysis of Incomplete Multivariate Data*. Chapman & Hall, London.
- Little, R.J.A. and Rubin, D.B. (2002) *Statistical Analysis with Missing Data, 2nd Edition*. J. Wiley & Sons, New York.
- van Buuren S, Brand JPL, Groothuis-Oudshoorn K, Rubin DB (2006). Fully conditional specification in multivariate imputation. *Journal of Statistical Computation and Simulation*, 76(12), 1049-1064
- van der Heidjen, G.J.M.G., Donders, A.R.T., Stijnen, T., and Moons, K.G.M. (2006). Imputation of missing values is superior to complete case analysis and the missing-indicator method in multivariable diagnostic research: A clinical example. *Journal of Clinical Epidemiology* 59: 1102-1109.
- Wadsworth, T., and Roberts, J.M. (2008). When missing data are not missing: A new approach to evaluating supplemental homicide report imputation strategies. 

References: Guidance for dealing with missing data

- Carpenter, J. (2006). Annotated Bibliography on Missing Data [accessed July 30, 2006]. Available online at <http://www.lshtm.ac.uk/msu/missingdata/biblio.html>
- Carpenter, J.R. and Kenward, M.G. (2007). Missing data in randomised controlled trials: A practical guide. Final report available at http://www.pcpoh.bham.ac.uk/publichealth/methodology/docs/invitations/Final_Report_RM04_JH17_mk.pdf.
- Graham, J.W. (2008). Missing data analysis: Making it work in the real world. *Annual Review of Psychology* 60(6): 1-28.
- He, Y., Zaslavsky, A.M., Landrum, M.B., Harrington, D.P., and Catalano, P. (2009). Multiple imputation in a large-scale complex survey: A practical guide. *Statistical Methods in Medical Research* 1-18.
- Lavori, P. et al. (2008). Missing data in longitudinal clinical trials Part A: Design and Conceptual Issues. *Psychiatric Annals* 38(12): 784-792.
- Schafer, J.L., & Graham, J.W. (2002). Missing data: Our view of the state of the art. *Psychological Methods* 7(2): 147-177.
- Siddique J. et al. (2008). Missing data in longitudinal trials Part B Analytic Issues. *Psychiatric Annals* 38(12): 793-801.

- Kenward, M.G. and Carpenter, J. (2007). Multiple imputation: Current perspectives. *Statistical Methods in Medical Research* 16: 199-218.
- Raghunathan, T.E., Lepkowski, J.M., Van Hoewyk, J., & Solenberger, P. (2001). A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey Methodology* 27: 85-95.
- Rubin, D.B. (1976) Inference and missing data. *Biometrika*, 63, 581-592.
- Rubin, D.B. (1987) *Multiple Imputation for Nonresponse in Surveys*. J. Wiley & Sons, New York.
- Rubin, D.B. (1996) Multiple imputation after 18+ years (with discussion). *Journal of the American Statistical Association*, 91, 473-489.

References: Tutorials and software for implementing MI

- www.multiple-imputation.com
- MI FAQ's: <http://www.stat.psu.edu/~jls/mifaq.html>
- http://www.ats.ucla.edu/stat/stata/seminars/missing_data/mi_in_stata_pt1.htm
- Azur, M., Frangakis, C., and Stuart, E.A. (2008). Disparities Among Children Served by the CMHS Childrens Services Program: Overview of Multiple Imputation and Using Multiply Imputed Data. General documentation on creating and using multiply imputed data. Available at <http://www.biostat.jhsph.edu/~estuart/MIDocumentationFinal.pdf>
- Donders ART, van der Heijden GJMG, Stijnen T, Moons KGM.(2006). Review: A gentle introduction to imputation of missing values. *Journal of Clinical Epidemiology* 1087-1091.
- Horton, N. & Kleinman, K.P. (2007). Much ado about nothing: A comparison of missing data methods and software to fit incomplete data regression models. *The American Statistician* 61(1): 79-90. Software appendix: <http://www.math.smith.edu/muchado-appendix.pdf>

- Lunt, M. (2008). A guide to imputing missing data with Stata. <http://personalpages.manchester.ac.uk/staff/mark.lunt/mi.html>
- Raghunathan, T.E., Solenberger, P.W., & Van Hoewyk, J.V. (2002). IVEWare: Imputation and Variance Estimation Software User's Guide. Ann Arbor, MI: Institute for Social Research, University of Michigan. www.isr.umich.edu/c/smp/ive/
- Schafer, J.L. (1999). Multiple imputation: A primer. *Statistical methods in medical research* 8(1): 3-15.
- Stuart, E.A., Azur, M., Frangakis, C.E., and Leaf, P. (2009). Multiple imputation with large datasets: A case study of the Children's Mental Health Initiative. *American Journal of Epidemiology* 169(9): 1133-1139.

References: Survey weighting for nonresponse

- Kalton, G., and Flores-Cervantes, I. (2003). Weighting Methods. *Journal of Official Statistics*, 19, pp. 81-97.
- Kalton, G. and Kasprzyk, D. (1986). The Treatment of Missing Survey Data. *Survey Methodology*, Vol. 12, No. 1: 1-16.
- Oh, H. and Scheuren, F. (1983). Weighting Adjustment for Unit Nonresponse. Chap. 13 in vol. 2, part 4 of *Incomplete Data in Sample Surveys*. New York: Academic Press

References: Diagnostics

- Abayomi, K., Gelman, A., & Levy, M. (2008). Diagnostics for multivariate imputations. *Applied Statistics* 57(3), 273-291.
- Gelman A, King G, & Liu C (1998). Not Asked and Not Answered: Multiple Imputation for Multiple Surveys. *Journal of the American Statistical Association* 93(443), 846:857.

- Allison, P.D. (2005). Imputation of Categorical Variables with PROC MI [accessed July 30, 2006]. Available online at <http://www2.sas.com/proceedings/sugi30/113-30.pdf>
- Carpenter, J., Kenward, M., Evans, S., and White, I. (2004). Last Observation Carry-Forward and Last Observation Analysis. *Statistics in Medicine* 23: 32413244.
- Collins LM, Schafer JL, Kam CK. (2001). A comparison of inclusive and restrictive strategies in modern missing-data procedures. *Psychological Methods* 6(4):330351.
- Cook, R. J., Zeng, L., and Yi, G. Y. (2004). Marginal Analysis of Incomplete Longitudinal Binary Data: A Cautionary Note on LOCF Imputation. *Biometrics* 60: 820828.
- Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum Likelihood From Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society, Series B* 39: 122.

- Fox, J. A. and M. W. Zawitz (2004). Homicide Trends in the United States: Weighting and Imputation Procedures for the 1976-2002 Cumulative Data File. Washington, DC. Bureau of Justice Statistics.
- Greenland S and Finkle WD. (1995). A Critical Look at Methods for Handling Missing Covariates in Epidemiologic Regression Analyses. *American Journal of Epidemiology* 142:1255-64.
- Moons, K.G.M., Donders, R.A.R.T., Stijnen, T., and Harrell, F.E., Jr. (2006). Using the outcome for imputation of missing predictor values was preferred. *Journal of Clinical Epidemiology* 59: 1092-1101.
- Reiter, J. P., Raghunathan, T. E., and Kinney, S. (2006), The importance of the sampling design in multiple imputation for missing data, *Survey Methodology*, 32.2, 143 - 150.
- Siddique, J. and Belin, T.R. (2008). Using an approximate Bayesian bootstrap to multiply impute nonignorable missing data. *Computational Statistics and Data Analysis* 53: 405-415.
- Vach W and Blettner M. (1991). Biased Estimation of the Odds Ratio in Case-Control Studies due to the Use of Ad Hoc Methods of Correcting for Missing Values for Confounding Variables. *American Journal of Epidemiology* 134:895-907.