Missing data procedures for psychosocial research

Elizabeth Stuart

Mental Health Summer Institute 330.616
Johns Hopkins Bloomberg School of Public Health
Department of Mental Health
Department of Biostatistics
Department of Health Policy and Management
estuart@jhu.edu
www.biostat.jhsph.edu/∼estuart

June 15-16, 2015

Outline: Day 1

Outline: Day 2

- Will demonstrate some statistical software for creating and analyzing multiply imputed data
- Will at least briefly cover:
  - For creating imputations:
    - SAS: IVEWare, proc mi
    - Stata: mi suite of commands, ice
    - R: mice, mi
  - For analyzing multiply imputed data:
    - SAS: proc mianalyze
    - Stata: mi, mim, micombine
    - R: mitools, mi
    - Mplus
    - HLM
- Will have some time for individual work if you want to bring in your own dataset to try things out on

Outline

Course description

Nearly every study in mental health research suffers from at least some amount of missing data: either individuals who did not respond to an entire survey (termed "unit nonresponse" or "attrition") or individuals with partially observed data but some missing items (termed "item nonresponse"). Analyses that use just the individuals for whom data is observed can lead to bias and misleading results. This course will first discuss types of missing data, and implications of the misssingness on analyses. It will then cover solutions for dealing with both types of missing data. These solutions include weighting approaches for unit nonresponse and imputation approaches for item nonresponse. An emphasis wills be on practical implementation of the proposed strategies, including discussion of software to implement imputation approaches. This will focus on recently developed software to implement multiple imputation, such as IVEware for SAS and ICE for Stata. Examples will come from school-based prevention research as well as drug abuse and dependence. Course attendees are not expected to have extensive background in statistical methods; an emphasis will be on making the ideas accessible to a broad audience.

Missing data

- 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!

An example

- Hill et al. (2005): maternal employment and child development
- Abstract: "Our results demonstrate small but significant negative effects of maternal employment on children's cognitive outcomes for full-time employment in the 1st year postbirth as compared with employment postponed until after the 1st year. Multiple imputation yields noticeably different estimates as compared with a complete case approach for many measures."
- Missingness rate ranged from 0-60% across variables

Hill et al. (2005); Table 2

Table 2
Comparison of Treatment Effects

	Complete case regression[a]		MI regression[b]	
Analysis	TE	*SE*	TE	*SE*
PPVT–R, ages 3–4	1.50	1.44	1.58*	0.79
PIAT–M, ages 5–6	0.81	1.08	1.62*	0.57
PIAT–M, ages 7–8	0.78	1.01	0.98	0.54
PIAT–R, ages 5–6	−0.49	0.97	0.82	0.62
PIAT–R, ages 7–8	0.46	1.06	0.95	0.53
BPI				
Internalizing, age 5–6	0.10	0.18	−0.01	0.10
Internalizing, age 7–8	0.17	0.18	0.19	0.10
Externalizing, age 5–6	0.12	0.29	0.20	0.17
Externalizing, age 7–8	−0.04	0.29	0.08	0.19

Types of missing data

Will discuss two main types of missing data:

- "Unit nonresponse": when data for an entire "unit" (e.g., individual) is missing
  - e.g., did not respond at all to follow-up survey
  - Also called "attrition"
  - Usually handled using nonresponse weighting adjustments or maximum likelihood methods
- "Item nonresponse": when individual items are missing for an individual
  - e.g., someone answered most of the survey questions, but left a few blank
  - Usually handled using imputation approaches or maximum likelihood methods

A little notation

- X^{obs} denotes observed values
- X^{mis} denotes missing values
- Y denotes some observed outcome of interest
- R denotes missing data indicators
  - $R_{ij} = 1$ if person i has variable j missing, R_{ij} if that value observed

Four common methods for dealing with missingness

1. Complete-case analysis
   - Assumes data missing completely at random: can lead to very biased results
   - Often results in large reductions in sample size; reduced power
2. Simple (single) imputations
   - e.g., mean imputation, regression prediction imputation, hot-deck imputation
   - Doesn't account for uncertainty in imputations
3. Multiple imputation
   - Best imputation approach
   - Easy to use software now exists
4. Maximum likelihood methods
   - For some models (e.g., longitudinal models), maximum likelihood methods can take missing data into account
   - Use the observed data, standard errors accurately reflect the missing data

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
  - Analyses using only complete cases will not be biased, but may have low power
  - 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
  - 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

Of course those are assumptions...

- Never know which of them is correct
- Can do diagnostics/tests for whether missingness is MCAR vs. (MAR or NMAR) (Enders 2010, p. 18)
  - 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?
  - Little (1998; JASA): test for MCAR

- But never know for sure if missingness is MAR or NMAR...
  - Have to use substantive understanding of what might have led to missing values
  - e.g., Are those who had been arrested more likely to not respond to a question asking about previous arrests? (They may not want to lie, but also may not want to tell the truth...)
  - Helps to have a good understanding of the data collection process
  - If believe missingness is NMAR, have to posit some model for the missingness (e.g., that those with previous arrests are 10% more likely to not respond to that question)
  - Tailored for each research question
  - Siddique and Belin (2008): example of missing depression levels; simulations show value in using a variety of assumptions and models
  - More later . . .

Why is it hard to come up with guidelines regarding the % of missingness: The fraction of missing information

- Hard to say what % of missing data is too much (or too little to worry about)
  - Variable with 90% missing might be fine if really good predictors are observed
  - Variable with 15% missing might be very problematic if no good predictors are observed
- Instead: fraction of missing information
  - How much information is in the observed data regarding a particular parameter?
  - Can be estimated by examining variation across multiple imputations (more later...)

Outline

Inappropriate ways of handling missing data

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

Ignoring it...

- 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!
  - Makes assumption that missingness is MCAR
  - Often results in lots of cases dropped...decreased power and loss of representativeness (Little and Rubin, 200; 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...
- (Also called listwise deletion)

- Researchers will often compare characteristics of the people in the final sample with those in the original sample
  - This okay, but doesn't tell the whole story
  - Does give some evidence for generalizability of results, but what if the relationships differ?

Missing data indicator approach

- Sometimes people will create an indicator for the missingness and include that as an additional predictor in regression models
- Categorical variables: create an additional category
  - e.g., Gender: "male", "female", "missing"
- Continuous variables: create an additional variable, and impute the mean for the cases with missing values
  - In regression models, include both the variable itself (e.g., age) and the indicator for having age missing (e.g., mage)
- Doesn't work very well and can lead to bias (Vach and Blettner 1991, Donders et al. 2006, Greenland and Finkle 1995)
- (Note: This does actually work well within propensity score estimation context)

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
- Used often in FDA clinical trials
- But generally biased (Carpenter et al. 2004; Cook, Zeng, and Yi, 2004; Jansen et al. 2006)
- A simple form of single imputation (see next slides...)

Single imputation

Single imputation fills in ("imputes") each missing value with a "best-guess"

Ways of doing that prediction:

- 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 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 error term on (impute off the regression line)

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

- "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

Simple example: Observed data

Individual	Gender	Age	M_{IQ}	IQ Test Score
1	M	24	0	122
2	M	32	0	109
3	M	41	0	131
4	M	22	1	?
5	F	20	0	135
6	F	42	0	102
7	F	31	0	124
8	F	32	1	?

Simple example: Mean imputation

Impute mean across all individuals (120.5)

Individual	Gender	Age	M_{IQ}	IQ Test Score
1	M	24	0	122
2	M	32	0	109
3	M	41	0	131
4	M	22	1	(120.5)
5	F	20	0	135
6	F	42	0	102
7	F	31	0	124
8	F	32	1	(120.5)

Simple example: Conditional mean imputation

Impute mean separately for males and females

Individual	Gender	Age	M_{IQ}	IQ Test Score
1	M	24	0	122
2	M	32	0	109
3	M	41	0	131
4	M	22	1	(120.7)
5	F	20	0	135
6	F	42	0	102
7	F	31	0	124
8	F	32	1	(120.3)

Simple example: Conditional mean imputation

Run regression model of IQ on gender and age, generate $\hat{IQ}$

$$IQ = 143.517 + .575 * \text{Male} - 0.725 * \text{Age}$$

Individual	Gender	Age	M_{IQ}	IQ Test Score
1	M	24	0	122
2	M	32	0	109
3	M	41	0	131
4	M	20	1	(129.6)
5	F	23	0	135
6	F	42	0	102
7	F	31	0	124
8	F	35	1	(118.2)

Simple example: Stochastic mean imputation

Run regression model of IQ on gender and age, generate $\hat{IQ}$

$$IQ = 143.517 + .575 * \text{Male} - 0.725 * \text{Age} + N(0, \sigma^2)$$

Individual	Gender	Age	M_{IQ}	IQ Test Score
1	M	24	0	122
2	M	32	0	109
3	M	41	0	131
4	M	20	1	(113.03)
5	F	23	0	135
6	F	42	0	102
7	F	31	0	124
8	F	35	1	(121.09)

Simple example: Hot-deck imputation

Based just on gender here...
For each person with a missing value, impute a value drawn randomly from the observed values of people with the same gender

Individual	Gender	Age	M_{IQ}	IQ Test Score
1	M	24	0	122
2	M	32	0	109
3	M	41	0	131
4	M	20	1	(109)
5	F	23	0	135
6	F	42	0	102
7	F	31	0	124
8	F	35	1	(124)

Predictive mean matching

- Another way of generating imputations
- Like a mix of MICE and hot-deck
- For each person with a missing value, generates a predicted value (using some model like in MICE) and then finds individuals with observed values but similar predictions, "takes" their observed value for the person with a missing value
- Works best for continuous variables and monotone missing data

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
- Getting information from another source
- Multiple imputation

Maximum likelihood approaches

- In some cases, maximum likelihood approaches exist
- Sometimes called "full information maximum likelihood"
- Directly maximize the likelihood function, $f(X, Y)$
  - Likelihood factors into two pieces: piece due to cases with fully observed data and piece due to cases with missing data
  - These two maximized together to get the maximum likelihood estimates (MLEs)
  - Often uses Expectation-Maximization (EM) algorithm (Dempster, Laird, and Rubin, 1977)
  - Can think of as iterating between generating imputations of missing values and estimating model of interest
  - Iterate until convergence

- 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
- But they don't always exist so not always a feasible option
- Mostly assume only missing outcomes, not missing predictors
- Another drawback is that you cannot use auxiliary information to improve the predictions; uses only the variables in the actual analysis
    - Assumes MAR given the variables in the model
- Exist mostly in structural equation modeling software (LISRESL, Mplus)
- Graham (2008), Siddique et al. (2008; a mixed-effects regression model), Enders (2010; Chapter 4)

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)
- Horton and Lipsitz (1999), Carpenter et al. (2006), Seaman and White (2013)

- A simple example...
- Imagine 100 males and 100 females in sample
- But only 80 males and 75 females respond
- Male respondents will get weight of $100/80 = (1/(80/100)) = 1.25$
- Female respondents will get weight of $100/75 = (1/(75/100)) = 1.333$
- So, e.g., a male respondent represents 1.25 males in the original sample
- These weights will make the 80 male and 75 female respondents represent the full sample of 200

- To implement weighting adjustments:
  - Fit model predicting response as a function of fully observed characteristics
  - Assign respondents a weight of 1/(p(response))
  - Use those weights in regression models and summary statistics
- Will weight the respondents to look like the full original sample
- Like survey sampling weights, except estimated instead of known
- Can be used for attrition as well as for original survey response

- Use model with many characteristics, generally measured at baseline (especially those predictive of response as well as the variables of primary interest)
- Treat the weights like you would survey sampling weights (e.g., using survey packages), run weighted models (e.g., pweight in Stata)
- Some concern about extreme weights
  - Check distribution of weights, respecify model if needed, trim outliers
  - Some do a "weighting class adjustment" where actually just form 5 subclasses based on the probabilities and everyone in each subclass gets the same weight
  - Extreme weights may also indicate extrapolation from complete to incomplete cases
- Relatively simple (and widely accepted) way of handling attrition/unit non-response

Proposal for creation of weights (Seaman and White, 2013, p. 287)

1. Identify a priori predictors of missingness. Exclude any that are likely not predictive of key variables of ultimate interest. Add any strongly predictive of primary outcome of interest.
2. Examine distribution of continuous predictors, transform as needed to avoid long tails.
3. Fit missingness model using full set of predictors. Consider lasso or other non-parametric models.
4. Check model fit using Hosmer-Lemeshow and/or Hinkley's method.
5. Check distribution of weights for complete and incomplete cases. If any zeros, simplify model. Check for extreme weights, modify model as needed.

Alternative sources of information

- In some cases, can utilize a secondary data source to get needed information
  - e.g., school records to get high school graduation
  - e.g., criminal records to get criminal activity
  - e.g., national death records to get death information
  - Can be resource intensive, especially if need to cover a lot of geographic areas/lots of schools
- In other cases, can calibrate numbers to known totals
  - e.g., issue of missing information about offender and incident in the Supplemental Homicide Reports (SHR)
  - Compare victim counts in SHR to similar data from NCHS, adjust as necessary (Fox and Zawitz 2004)
  - Wadsworth and Roberts (2008) evaluates four common techniques for dealing with this missingness that utilize supplemental info from police records

Multiple imputation

- Same idea as single imputation, but fills in each missing value multiple times
  - Like repeating the stochastic mean imputation multiple times
  - (Although could potentially use hot-deck or predictive mean matching as well; just do each multiple times and allow for randomness in which observed values selected each time)
- Three steps:
  1. Generate imputations: Create multiple (e.g., 10) "complete" data sets by filling in (imputing) the missing values
  2. Run analysis on each imputed data set (can be almost any analysis)
  3. Combine (pool) results from the imputed data sets using standard "combining rules" (Rubin 1987)
- In pooled result, 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

Simple example: Multiple imputation (1)

Run regression model of IQ on gender and age, generate $\hat{IQ}$

$$IQ = 143.517 + .575 * \text{Male} - 0.725 * \text{Age} + N(0, \sigma^2)$$

Individual	Gender	Age	M_{IQ}	IQ Test Score
1	M	24	0	122
2	M	32	0	109
3	M	41	0	131
4	M	20	1	(113.03)
5	F	23	0	135
6	F	42	0	102
7	F	31	0	124
8	F	35	1	(121.09)

Simple example: Multiple imputation (2)

Run regression model of IQ on gender and age, generate $\hat{IQ}$

$$IQ = 143.517 + .575 * \text{Male} - 0.725 * \text{Age} + N(0, \sigma^2)$$

Individual	Gender	Age	M_{IQ}	IQ Test Score
1	M	24	0	122
2	M	32	0	109
3	M	41	0	131
4	M	20	1	(131.93)
5	F	23	0	135
6	F	42	0	102
7	F	31	0	124
8	F	35	1	(120.16)

Simple example: Multiple imputation (3)

Run regression model of IQ on gender and age, generate $\hat{IQ}$

$$IQ = 143.517 + .575 * \text{Male} - 0.725 * \text{Age} + N(0, \sigma^2)$$

Individual	Gender	Age	M_{IQ}	IQ Test Score
1	M	24	0	122
2	M	32	0	109
3	M	41	0	131
4	M	20	1	(111.35)
5	F	23	0	135
6	F	42	0	102
7	F	31	0	124
8	F	35	1	(126.93)

The goal of MI, or actually any procedure to deal with missing data

The goal is not to predict the missing values or get the missing values close to the true values . . . goal is to obtain valid statistical inferences accounting for the missing data

Outline

How to create multiple imputations

Two main approaches:

- Joint model of all variables
  - e.g., assume multivariate normal distribution of all of the variables
  - Fit using the observed cases
  - Use to predict (multiple times) the missing values
  - Sometimes multivariate normal model used even with categorical variables (imputations are then rounded back to categories), but this can be severely biased (Horton, Lipsitz, and Parzen, 2003; Allison 2005)
  - Software: Norm, mix, SAS proc mi
- Multiple imputation by chained equations (MICE)
  - Model each variable one at a time as a function of the other variables
  - Allows for much more flexible models for each variable (e.g., counts/binary/continuous)
  - Doesn't necessarily imply a proper joint distribution, but doesn't seem to be a big problem in practice
  - Will discuss and illustrate software tomorrow

An aside . . . monotone missing data

- Monotone missing data has a structure where the variables can be sorted such that:
  - If variable X is missing, then the rest of the variables are missing
- e.g., longitudinal data with drop out (and once people drop out they don't come back)
- Data missingness pattern looks like a staircase
- When have a monotone structure, MI easy to implement–can just impute in order from least missing to most missing
- Doesn't require iteration, which we'll see MICE does require

Multiple Imputation by Chained Equations (MICE)

- MICE procedure allows modeling each variable one at a time
  - Fit model of each variable, conditional on all others
  - Iterate fitting model and imputing each variable
  - Allows bounds (e.g., age started smoking)
  - Incorporates restriction to subpopulations (e.g., age started smoking)
- Also called "fully conditional specification" (FCS)
- Raghunathan et al. (2001)

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

Software to implement MICE procedure

- SAS and stand-alone: IVEWare
- Stata: mi suite of commands (mi impute chained), ice
- R/Splus: 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

See Azur et al. (2011) for a tutorial; also some great websites available (links given later)

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
- 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; see White et al., 2011 for examples)
  - If non-normal, can also use predictive mean matching (White et al., 2011)
- With so many variables, can't possibly do careful model selection for each one; some packages will do stepwise selection

What variables should be included?

- Any variables that will be used in subsequent analyses
  - Otherwise its associations with other variables will be attenuated in analyses
- Any higher-order effects that are of interest in the analysis phase
- Any other special features of the data (e.g., survey weights)
- If have a categorical variable (like race) should keep it as categorical and impute that way, rather than breaking it into individual dummy variables before the imputation
- That said, can't make the models too big or may run into convergence problems
- Note: Don't need to specify which are dependent vs. independent

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; see Enders (2010; Chapter 5))

Interactions and non-linear terms

- Any interaction or non-linear terms (e.g., X^2) that will be in the analysis need to also be included in the imputation model
- Treat them as "just another variable" (White, Royston, and Wood): create a variable that is the interaction term or X^2 to use in the imputation process
- For interactions with binary or group variables, could also impute each group separately (e.g., male/female or race groups); this allows all possible interactions
  - In Stata, can use "by()" option

Imputation and analysis compatibility

- Imputation model should be more general ("bigger") than analysis model that will be used: otherwise risk finding null effects simply because data imputed assuming no relationship between variables
- Imputation model may need to explicitly set some relationships to 0
- Basically, include all variables and associations of interest in the analyses
- Difficult to include too many interactions in models; limits the analyses that can be done
- Also termed "congeniality"

Model specification

- IVEWare and ice in Stata allow the use of stepwise selection to select the imputation model for each variable
  - mi_ice gives a wrapper to use ice within the mi suite of commands (http://www.stata.com/support/faqs/statistics/mi-versus-ice-and-mim/)
- Uses some criteria (e.g., # of predictors, minimum marginal R^2)
  - Smaller minimum marginal R^2 will lead to more variables being included
- CMHI: Used minimum additional $R^2 = 0.01$ (also did sensitivity analysis trying 0.005)
- 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)
  - May have convergence problems if try to run with all predictors included; stepwise often more feasible computationally

Bounds

- e.g., scales with minimum and maximum values
- IVEWare can easily handle bounds on variables
  - Specify in IVEWare code using BOUNDS statement
- Other packages don't have easy ways of doing this: often handled using post-hoc rounding or predictive mean matching
  - Should check how often imputations are outside the correct range
  - In Stata, "truncreg" can be used, but often has convergence problems; predictive mean matching may be a better choice

Restrictions

- IVEWare can also handle variables that are only defined for a subset of the sample
- e.g., instruments only given to children above a certain age
- e.g., skip patterns, where only children who endorse a particular question are asked follow-up questions
  - e.g., "How many days did you drink alcohol in past 30 days?" only asked of those who said they had drunk any alcohol in the past 30 days
- Specify in IVEWare using RESTRICT command
  - Will give a fake value for individuals for whom the value is meaningless
  - Important to recognize this in imputed datasets: set values back to missing, to avoid confusion

Questions to address in CMHI example

- How big/inclusive to make the models?
  - CMHI: Tried a few minimum R^2 values to assess sensitivity
  - Couldn't include too many predictors
  - Did include interactions of primary interest (e.g., race*gender)
- Force some variables into the models?
  - e.g., site
  - CMHI: didn't force any; let data decide which site indicators to include
- CMHI: In end, about 6 predictors included in each model

How many imputations to generate?

- Conventional advice has been 5-10, but more (e.g., 40) may be better in terms of power (Graham, Olchowski, & Gilreath (2007))
- White et al. (2011) recommend $m = 100 * FMI$ (FMI=fraction of missing information)
  - Since FMI hard to estimate, but Bodner's approximation says FMI $<$ % missing cases, approximate $m = 100*$(% missing cases)
  - e.g., 20% missing cases would imply $m = 20$
- White et al. (2011) also argue that for reproducibility may need $m > 100$
- Need to balance that with computational issues
- In CMHI, did 10
- Note: SAS seems better able to handle large datasets and large numbers of imputations than Stata

Stata's mcerr option

From https://www.ssc.wisc.edu/sscc/pubs/stata_mi_estimate.htm

- "mcerr" option in "mi estimate" command will give an estimate of the Monte Carlo error in estimation results
  - Leaves out one imputation at a time
- White, Royston, and Wood guidelines:
  1. The Monte Carlo error of a coefficient should be less than or equal to 10% of its standard error
  2. The Monte Carlo error of a coefficient's T-statistic should be less than or equal to 0.1
  3. The Monte Carlo error of a coefficient's P-value should be less than or equal to 0.01 if the true P-value is 0.05, or 0.02 if the true P-value is 0.1
- If those conditions are not met, you should increase the number of imputations.

Step 3: Diagnosing and assessing imputations

- With so many variables, it is hard to carefully check each model to determine that it is reasonable
- 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)
- Standard packages have very limited diagnostics
- Note: Differences don't mean something is wrong! Could be because of differences in the types of people with observed vs. missing data

Graphical summaries

- 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

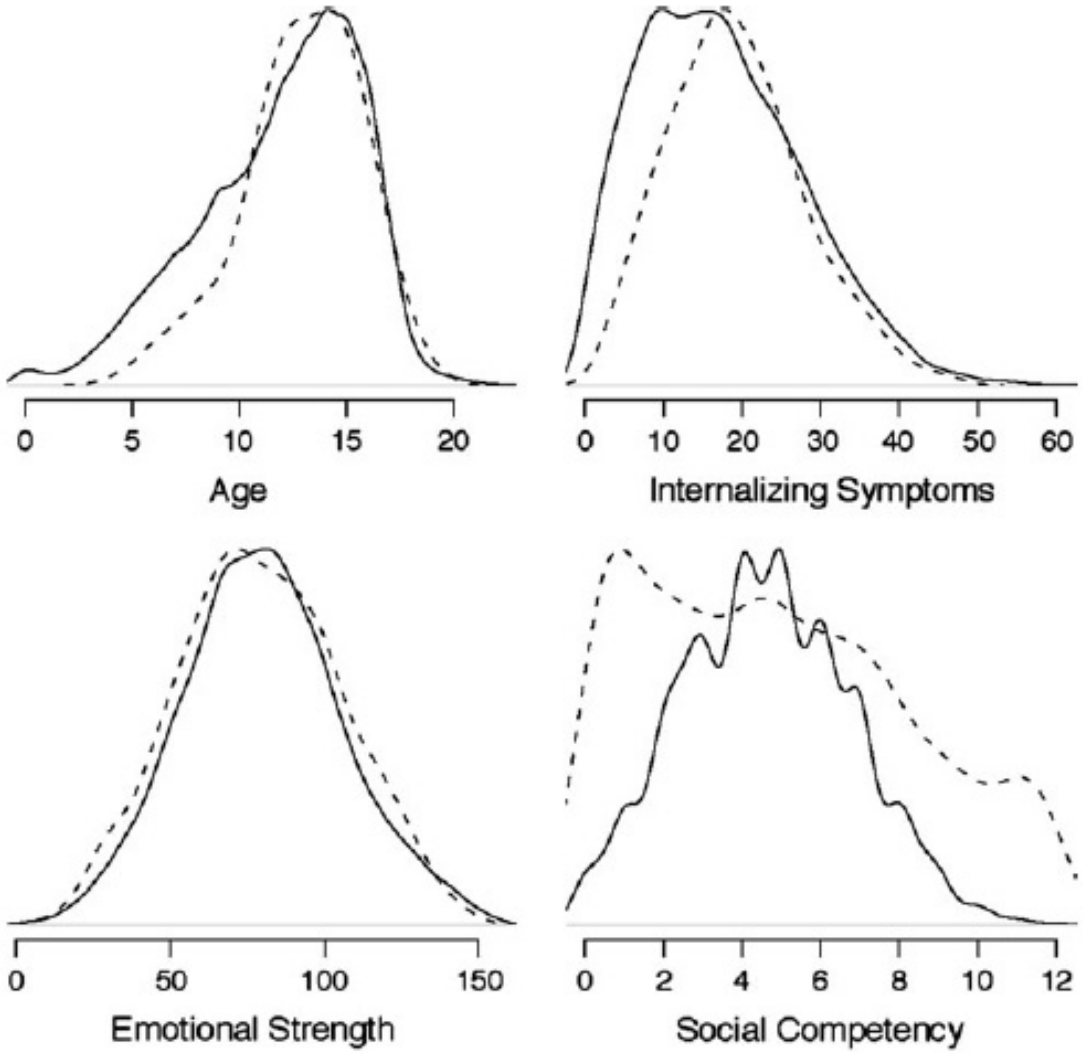

Figure 1. Comparison of observed and imputed values for 4 representative variables. For each variable, the solid line shows the density plot of observed values and the dashed line the density plot of imputed values. Age is expressed in years. The other measures are the Child Behavior Checklist (CBCL)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

Numerical diagnostics

- IVEWare automatically prints out some diagnostics, including of convergence
- Other packages show similar diagnostics; will go through those tomorrow
- Shows coefficients from each regression model
  - Unperturbed (the estimates themselves) as well as perturbed (estimates plus random error)
  - Big difference between unperturbed and perturbed may indicate a lot of uncertainty in the prediction models

	Unperturbed Coefficient	Perturbed Coefficient
Intercept	0.186	0.183
Famabu	-0.590	-0.584
Parntab	-0.840	-0.815

- IVEware also outputs summary statistics on original and imputed values
  - Should check that imputations look reasonable
  - Make sure values being imputed are in the correct ranges
- Example: functional impairment scale (0-3)
- The "4" is a placeholder for missing by design

	Observed		Imputed		Combined	
Code	n	%	n	%	n	%
0	762	10.4	126	6.9	888	9.7
1	1258	17.1	147	8.0	1405	15.3
2	1259	17.1	475	25.9	1734	18.9
3	4073	55.4	799	43.6	4872	53.0
4	0	0.0	286	15.6	286	3.1
Total	7352	100.0	1833	100.0	9185	100.0

An example of a problem...number of times binge drank in past 30 days

	Observed	Imputed	Combined
Number	710	8475	9185
Minimum	0	0	0
Maximum	30	4.50e+015	4.50e+011
Mean	1.96	5.1e+011	4.90e+011
Std Dev	4.25	4.89e+013	4.70e+013

- Diagnosis: Not enough information in data; decided not to impute
- Could have also tried simplifying imputation model (e.g., restrict # of predictors for this variable)
- Alternatively could specify bounds on variable; if so, check how often predictions at the bounds

In the CMHI data...

- Results looked better than expected
- Not very many problematic variables
- Each site generally had a problem with $< 5\%$ of the variables
- Each variable generally had a problem with $< 2\%$ of the sites
- Some variables and sites more problematic
  - e.g., Race imputations for Vermont site
- Most difficult variables: rare outcomes and those that are conditional on others

Some differences between imputations and observed values

Variable	Rate among imputations	Rate among observed values	Significance (p-value)
American Indian	18.3	6.3	*** (.000)
Caucasian	49.5	61.0	*** (.000)
Conduct disorder	15.8	8.9	*** (.000)
Eligible for Medicaid	74.5	69.2	*** (.000)
Has ADHD	36.3	42.0	*** (.000)
Parental history of psych hosp	38.9	42.3	** (.050)
Convicted of a crime	38.5	33.0	** (.046)
% of day in special ed	35.3	36.4	(.761)

Note: * Sig at 10% level, ** Sig at 5% level, *** Sig at 1% level

Sensitivity analysis

- Also helps to do sensitivity analyses
- Change imputation settings slightly, see how different the imputations/final models are
- e.g., for CMHI, changed the stepwise selection criteria to include more variables
- Nice if imputations not very sensitive to small changes like that

More advanced diagnostics

- 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, micombine
  - SAS: proc mianalyze
  - HLM: multiple imputation options
  - Mplus: multiple imputation command
- For other models, may need to do it "by hand"

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

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

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

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

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

- The total variance of $\overline{Q}$ is then

$$T = \overline{U} + (1 + \frac{1}{m})B$$

- Degrees of freedom for t distribution can also be calculated (Enders, 2010, p. 231+)
- See Schafer (1997) or Little and Rubin (2002) for details

Calculating the fraction (%) of missing information

- Captures how much information there is in the data about a particular parameter
- Compares the within-imputation and between-imputation variance
- To calculate:

$$\gamma = \frac{(r+2)/(df+3)}{r+1}$$

$$r = \frac{(1+1/m) * B}{\overline{U}}$$

- r is the relative increase in variance due to the nonresponse/missing data
- Alternatively (Enders, 2010, p. 225): $FMI = \frac{B+B/m+2/(\nu+3)}{T}$
  - ν is a degrees of freedom value; goes to infinity as m goes to infinity

- FMI will typically be lower than % of missing values because of correlations in the data
- Quantifies influence of missing data on the standard errors
- Also can be used as diagnostic: should pay more attention to variables that have high FMI when doing imputation diagnostics

Example: mim command in Stata

```
mim: logit dsmmood sex age

Multiple-imputation estimates (logit)                 Imputations =       10
Logistic regression                                   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
------------------------------------------------------------------------------
```

Post-imputation results make more sense (Stuart et al. 2009; Table 2)

Table 2. Bivariate Associations Between Predictors and Outcomes: Comparison of Complete-Case and Imputation Results

Outcome/ Independent Variable	Predictor/ Independent Variable	Complete-Case Coefficient (*P* Value)	Imputation Coefficient (*P* Value)	% Missing on Independent Variable
High emotional strength	Age at first smoking	−0.02 (0.00)	0.02 (0.19)	18
Functional impairment	Age at first smoking	0.03 (0.00)	−0.06 (0.04)	18
Clinical internalizing problems	Age at first smoking	−0.01 (0.08)	−0.07 (0.00)	18
Clinical internalizing problems	Ever smoked	−0.04 (0.46)	0.25 (0.00)	18
Clinical internalizing problems	Ever drank alcohol	−0.18 (0.00)	0.09 (0.13)	18

Differences pre- and post-imputation

- Results sometimes will, sometimes won't differ if compare complete-case analyses with analyses after MI
  - Won't know until you try whether or not it will matter!
  - Of course will depend on % of missingness, as well as on the missing data mechanisms
- MI results should be preferred

Some quantities cannot be easily combined using Rubin's rules

- Likelihood ratio tests hard to combine; better to use Wald tests with multiply imputed data (White et al., 2011)
  - Stata's mi package can combine subsets of coefficients or linear or nonlinear hypotheses (mi test, mi testtransform, mim: testparm)
  - See code here for combining likelihood ratio tests: http://www.stefvanbuuren.nl/publications/MICE%20V1.0%20Manual%20TNO00038%202000.pdf
- If model you are running not part of standard combining software, can just send point estimates and variances to a few functions
  - e.g., R: mitools, Stata: mi estimate (option cmdok)
- Combining R^2 values: http://www.ats.ucla.edu/stat/stata/faq/mi_r_squared.htm
- Combining rules assume normality so some parameters work better when transformed (Enders, 2010, p. 220+); e.g., correlation coefficient

White et al. (2011), Table VIII

Table VIII. Common statistics that can and cannot be combined using Rubin's rules (equations (1) and (2)).

Statistics that can be combined without any transformation	Mean, proportion, regression coefficient, linear predictor, C-index, area under the ROC curve
Statistics that may require sensible transformation before combination	Odds ratio, hazard ratio, baseline hazard, survival probability, standard deviation, correlation, proportion of variance explained, skewness, kurtosis
Statistics that cannot be combined	*P*-value, likelihood ratio test statistic, model chi-squared statistic, goodness-of-fit test statistic

Model fitting strategies with multiply imputed data

- Variable selection models such as stepwise selection not straightforward for outcome models
- One possibility: Stack data and run stepwise models on all n*m observations (White et al., 2011)
- If not sure which interactions will be included in analysis model, one strategy outlined by White et al. (2011, p. 381):
  1. "Produce a provisional and relatively simple imputation model, including non-linear terms of key scientific interest, but omitting all other non-linear terms."
  2. "Use the imputed data to build and check an analysis model, including investigating the need for non-linear terms. Note that these model checks are conservative when relevant non-linear terms were omitted from the imputation model."
  3. "If any convincing non-linear terms are found, then recreate the imputations including the non-linear terms . . ."
  4. "Use the revised imputed data set to estimate the parameters of the final analysis model."

Outline

Sometimes MAR is not a reasonable assumption

- If missingness related to unobserved variables, data is really NMAR
- i.e., people who are the same on observed characteristics but some are missing and some are not missing may have different values of the missing values
- e.g., substance abuse treatment studies, where those with most use also most likely to drop out of study
- e.g., depression treatment studies, where those who we can't find at follow-up may be more (or less) depressed than those who don't

Turning NMAR into MAR . . .

- Sometimes can make NMAR data more like MAR by collecting data on the potential causes of missingness
- Schafer and Graham (2002) recommend including survey question that asks respondents to report their likelihood of dropping out of the study before the next measurement occasion
- Or sometimes can supplement original data with other data sources
  - New data collection, at least on a subset of the study sample
  - Existing data from other sources
- e.g., Jackson et al. (2010) use proxy data to assess validity of MAR assumption in study of therapy for patients with schizophrenia
  - Use additional data on caregiver reported outcome and the number of contact attempts

Strategies for dealing with NMAR data

- Include beliefs about NMAR structure in imputations
  - Impute under MAR, then add a constant to the imputed values to acknowledge fact that true values may be higher (or lower) than what is predicted under MAR; repeat this for various sizes of the constant and assess sensitivity of results (Enders 2010, Section 10.2)
  - Non-ignorable approximate Bayesian bootstrap (see below . . .)
- Model the missingness: Allow for relationship between missing values and the missingness indicator (e.g., a joint model of missingness and the data: p(Y,R); Enders (2011), Enders (2010, Chapter 10))
  - Selection models: Combine substantive analysis with model predicting response probabilities
  - Pattern mixture models: Estimates substantive analyses separately within groups defined by missing data patterns
- Note: None of these solutions really solve the problem in the sense that they all rely on models and untestable assumptions. Best strategy might be to try a few.

Can also set up as sensitivity analysis

- Impute assuming MAR, and then specify sensitivity parameters to see how much they would change results
- Sensitivity parameter characterizes difference in variable between those with missing and non-missing values
- Obtain new adjusted results
- sensMICE package works with mice for R: http://lertim.fr/Members/rgiorgi/DossierPublic/fonctions-r-s/

Resseguier et al. (2011); Table

TABLE. Example of Sensitivity Analysis on the Odds Ratio Estimating the Association Between Viral Load and Poor Mental Health Using the R Data Set "CHAIN"

	No.	% of Viral Load[a]	aOR (95% CI)[b]
	Viral load as a binary variable		
Complete cases	353		
<400 c/mL	188		1.00
≥400 c/mL	165	46.7	1.66 (0.94 – 2.93)
Multiple imputation			
MAR	508		
≥400 c/mL		50.0	2.01 (1.21 – 3.35)
MNAR ($\theta = 1.2$)	508		
≥400 c/mL		50.8	1.73 (1.04 – 2.85)
MNAR ($\theta = 1.5$)	508		
≥400 c/mL		52.1	1.73 (1.05 – 2.83)
MNAR ($\theta = 2.0$)	508		
≥400 c/mL		53.9	1.75 (1.03 – 2.97)

SAMON software

- www.missingdatamatters.org [Dan Scharfstein and others; account required to download software]
- Global sensitivity analysis to determining at what point results from an RCT would change given NMAR missingness on outcomes
- For monotone missingness, where people drop out of the study over time
- http://www.biostat.jhsph.edu/~dscharf/missingdatamatters/samon_1.0_userDoc.pdf

Creating imputations under a NMAR assumption

- Siddique and Belin (2008): "non-ignorable approximate Bayesian bootstrap" (ABB)
- Main idea: Use hot deck/predictive mean matching procedure, but vary probabilities of selection of each subject based on their outcome
  - e.g., draw high values with higher probability (see next slide)
- Recommend using a different ABB for each imputed data set; results will then average over possible missing data mechanisms (acknowledges the uncertainty in the missing data mechanism)
- Motivating example: depression treatment intervention study

Siddique and Belin (2008), Figure 1

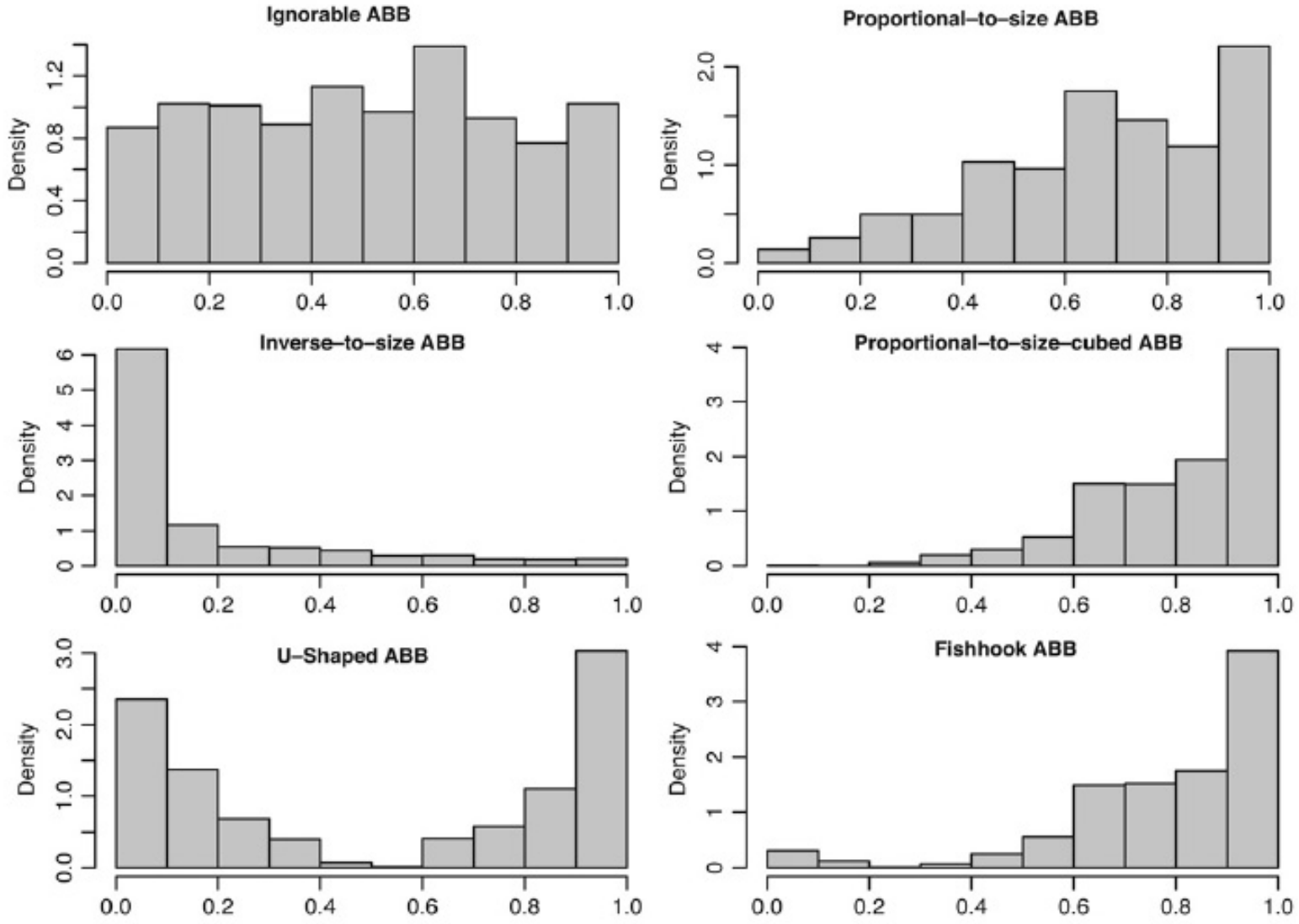

Fig. 1 Histograms of $n = 1000$ observations simulated from a uniform distribution that have been weighted by six different ARB types that place weigh

Selection models

- $p(Y, R) = p(R|Y)p(Y)$
- Two part model
  1. substantive regression model
  2. response model
- Allows for correlation in the errors of the two models: this is what relaxes MAR
- Relies on multivariate normality; can be sensitive to model form, variables included, etc.
- Heckman (1976, 1979)

Pattern mixture models

- $p(Y, R) = p(Y|R)p(R)$
- Estimates model separately within groups defined by missing data patterns
- Often requires smoothing across those groups: not enough data to really estimate separate models in each group
- A simplified description is that model includes parameters for missingness (e.g., missing data indicators)
- Relies on unestimable parameters
- Hedeker and Gibbons (1997)

Outline

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
- e.g., in CMHI data the CBCL scale has 113 items....we imputed the overall scales
- e.g., in CMHI data the delinquency scale has 25 items and each of interest to researchers...we imputed the individual items
- For more options and information, see Enders (2010, Section 9.6)

What about raw or standardized scores?

- Assuming you have the ability to recreate the standardized scores...
- Impute whichever one looks more normally distributed
- e.g., in CMHI data internalizing raw scores looked more normally distributed than standardized scores so imputed the raw scores

Multilevel or clustered data

- Not a lot of guidance on this
- If analysis will have only random intercepts, can just 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)
- MLWwiN and REALCOM impute macros for imputing multi-level data: www.missingdata.org.uk (inc. sample code)
- Yucel (2008)

In studies estimating causal effects

- Impute covariates and outcomes together, include lots of interactions between treatment status, covariates, and outcomes in imputation model (Want to make sure not to impose a treatment effect on the imputations)
- Although some people may balk at including outcome in imputation process, better to impute them than to leave it out, which would assume no treatment effect (Moons et al., 2006; Sterne et al., 2009)
- That said . . . some recommendation to include only those with observed treatment and outcome in the outcome analyses (but still use the outcome and treatment when creating the imputations)
    - Using imputed treatment status and imputed outcomes in analyses may just add noise/random error (White et al., 2011)

Longitudinal data

- Makes sense to convert data into "wide" format so observed time points can be used to help impute missing time points
- In Stata: "mi reshape" can be used to help convert the imputed data back into long format

Survey data

- In general, want to incorporate information about the survey design (e.g., strata, PSU) in the imputation if possible
  - May capture relevant information about individuals
- Can sometimes run weighted imputation models
- But at a minimum can also include the weight and other survey design variables in the imputation model (as predictors)
- In Stata, can run "mi estimate: svy: COMMAND" to run models on survey data that has been imputed
- For creating imputations, "svy" cannot be used with "mi impute chained" (can use weights by specifying [pweight=weight])
  - So run weighted models, and include strata or PSU variables as predictors in the imputation model

FAQ's

- 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 is "planned missingness"? (Enders 2010, p. 21+)
  - Study design where purposefully only select data from some individuals
  - Simple example: Instead of asking everyone all 100 questions, ask everyone 50 questions and randomly select half to get the other 20 and half to get the other 30
  - Known to be missing at random (by design)
  - Can help save resources
  - Useful for long surveys and longitudinal designs (follow up a subset at each time point)

- 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

Outline

National Research Council recommendations

- Report on how to handle missing data in clinical trials (Little et al., 2012).
- Recommendations:
  - Continue collecting data on everyone, even those who discontinue treatment
  - Limit missing data (see next slide)
  - Do not use complete case analysis or single imputation methods (such as last observation carried forward)
  - Use approaches such as multiple imputation or maximum likelihood
  - Do sensitivity analysis to assess robustness to underlying assumptions

Little et al. (2012), Table 1

Table 1. Eight Ideas for Limiting Missing Data in the Design of Clinical Trials.

Target a population that is not adequately served by current treatments and hence has an incentive to remain in the study.
Include a run-in period in which all patients are assigned to the active treatment, after which only those who tolerated and adhered to the therapy undergo randomization.
Allow a flexible treatment regimen that accommodates individual differences in efficacy and side effects in order to reduce the dropout rate because of a lack of efficacy or tolerability.
Consider add-on designs, in which a study treatment is added to an existing treatment, typically with a different mechanism of action known to be effective in previous studies.
Shorten the follow-up period for the primary outcome.
Allow the use of rescue medications that are designated as components of a treatment regimen in the study protocol.
For assessment of long-term efficacy (which is associated with an increased dropout rate), consider a randomized withdrawal design, in which only participants who have already received a study treatment without dropping out undergo randomization to continue to receive the treatment or switch to placebo.
Avoid outcome measures that are likely to lead to substantial missing data. In some cases, it may be appropriate to consider the time until the use of a rescue treatment as an outcome measure or the discontinuation of a study treatment as a form of treatment failure.

Little et al. (2012), Table 2

Table 2. Eight Ideas for Limiting Missing Data in the Conduct of Clinical Trials.

Select investigators who have a good track record with respect to enrolling and following participants and collecting complete data in previous trials.
Set acceptable target rates for missing data and monitor the progress of the trial with respect to these targets.
Provide monetary and nonmonetary incentives to investigators and participants for completeness of data collection, as long as they meet rigorous ethical requirements.[15,16]
Limit the burden and inconvenience of data collection on the participants, and make the study experience as positive as possible.
Provide continued access to effective treatments after the trial, before treatment approval.
Train investigators and study staff that keeping participants in the trial until the end is important, regardless of whether they continue to receive the assigned treatment. Convey this information to study participants.
Collect information from participants regarding the likelihood that they will drop out, and use this information to attempt to reduce the incidence of dropout.
Keep contact information for participants up to date.

Selecting MI vs. ML (Enders 2010, Section 11.4)

- If same variables used in both, both methods should give similar results and both can yield accurate standard errors and inferences. In fact, asymptotically they are the same (Seaman and White, 2013). But each has advantages ...
- Advantages of MI:
  - Easy use of auxiliary variables
  - Easier handling of incomplete predictor variables (MI doesn't care if a variable is predictor or outcome); some ML methods will still drop cases with missing covariates
  - Better for handling missingness on individual items within a scale
  - More flexible; can be used for almost any analysis
  - Can also be used in context where imputer and analyst are different (e.g., imputer may have access to more data)

- Advantages of ML:
  - Easier for estimating moderating effects
  - SEM models: often handle missingness automatically. In contrast, pooling SEM fit indices after MI not straightforward
  - Fewer "procedural ambiguities" and open questions of implementation
  - Often easier to implement than MI

Selecting MI vs. weighting

- Advantages of MI:
  - More flexible, can handle any missing data pattern
  - Can use variables in imputation model that are not fully observed (weighting requires predictors of response be fully observed)
  - Generally more efficient (Seaman and White, 2013), because uses more information
- Advantages of weighting:
  - Computationally simple for unit nonresponse
  - Potentially easier to specify the nonresponse model than the imputation model(s)
  - May be easier to see when extrapolating from complete to incomplete cases (Seaman and White, 2013); will see large weights
- Can also use them together!
  http://missingdata.lshtm.ac.uk/talks/RSS_2012_04_18_seaman.pdf

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

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 (e.g., Fumagalli et al., 2013)
  - 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?

Outline

General resources

- http://www.stefvanbuuren.nl/mi/Software.html
- Code for many packages in Horton and Kleinman (2007)
  - http://www.math.smith.edu/muchado-appendix.pdf

Software for MICE: R

- mi package (Gelman et al., 2015)
  - http://cran.r-project.org/web/packages/mi/index.html
  - http://cran.r-project.org/web/packages/mi/vignettes/mi_vignette.pdf
  - 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)

- mice package (van Buuren, 2015)
  - http://cran.r-project.org/web/packages/mice/index.html
  - For creating and analyzing multiply imputed data
  - Multiple imputation by chained equations
  - Some good diagnostics, added functionality recently
  - Can incorporate bounds, restrictions, passive imputation (variables that are functions of other variables)
  - See program mice-R.R, output (mice-R.out)

Analyzing MI data in R

- mice and mi packages have built in commands
  - See sample code in R-mi.R and R-mice.R
- mitools package (sample code in R-mice.R)
  - 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

Software for MICE: SAS

- 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 mi
  - Uses multivariate normal
  - As of Version V9.3, can also do MICE
- 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

Software for MICE: Stata

- 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 enditemize
  - For analyzing mi data: micombine, mim, mi estimate
  - Stata 11: mi suite of commands
  - Stata 12: mi suite now integrated with ice using "mi impute chained" and "mi ice" commands
  - (Pre-Stata 12, can easily go between ice and mi functions using "mi import ice" and "mi export ice" commands so can use mi's procedures, like for analyzing MI data)

Using mi commands in Stata

- https://www.ssc.wisc.edu/sscc/pubs/stata_mi_intro.htm
- Will walk through sample code (Stata-mi.do)

Key steps in using Stata's mi suite

1. Set data as mi
2. Register variables (imputed, regular, passive)
3. Check the imputation models (can use the "dryrun" option to easily see the models that will be used)
   - May need to simplify models until you get them to run
4. Impute!
5. Check convergence
6. Check the imputed values
7. Manage the mi data
8. Run outcome models ("mi estimate")

Using ice in Stata

- To install ice: ssc install ice, replace
- Can run ice through the mi suite in two ways:
  - mi_ice wrapper, which runs ice: in Stata, type net from http://www.homepages.ucl.ac.uk/~ucakjpr/stata
  - mi impute chained
- mi impute chained supports factor variables
- mi_ice supports stepwise model selection
- Sample code at the end of the Stata-mi.do code

Details

[mi] 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/seminars/missing_data/mi_in_stata_pt2.htm
- http://www.ats.ucla.edu/stat/Stata/library/ice.htm
- http://www.stata.com/support/faqs/statistics/mi-versus-ice-and-mim/

Options in ice

- 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)
  - "stepwise" option will run stepwise selection (and can specify particular types of stepwise models)
- 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)

Software for SPSS

- ftp://public.dhe.ibm.com/software/analytics/spss/documentation/statistics/20.0/en/client/Manuals/IBM_SPSS_Statistics_Algorithms.pdf
- http://support.spss.com/ProductsExt/SPSS/ESD/17/Download/User%20Manuals/English/SPSS%20Missing%20Values%2017.0.pdf
- Missing values add-on package: Creates imputations and analyzes MI data
- Monotone or MICE approaches

Software for MPlus

From Stef van Buuren's page (http://www.stefvanbuuren.nl/mi/Software.html):

> *Mplus Version 6 implements routines to generate, analyze and pool multiply imputed data. Multivariate imputations can created under a joint model based on the variance-covariance matrix (default) or by a form of conditional specification. Mplus embeds multiple imputation using an unrestricted imputation model that is specified behind the scenes (called H1 imputation). It is possible to specify a custom imputation model in conjunction with the Bayesian estimator (called H0 imputation).*

Analyzing MI data in Mplus

- Create text file with column list of file names with the multiply imputed datasets
- Reference that file in the "DATA: FILE IS" statement
- Specify "IMPUTATION" in the "TYPE" statement

Subset of data from CMHI

- Subset of kids from longitudinal follow-up sample (N=9,551)
- From 45 sites
- Baseline data only
- For illustrative use only! A subset of the data to help models run quickly.
- Data: sinst-subset.dta, sinst-subset.csv

Variables

- Demographics: age,sex, race (nrace3), family income (income), number of kids in family (totchild), number of adults in family (totadult)
- Role scales: total roles (school, home, community) (totrole)
- Child Behavior Checklist (CBCL) scales: Total competency raw (ctotcomr), total problem (ctotraw), internalizing (cintraw), externalizing (cextraw)
- Youth Self Report (YSR) scale: Total problem score (ytotraw)

Sample programs

- Stata: Stata-mi.do, Stata-mi.log
- R: R-mice.R, R-mice.out, R-mi.R, R-mi.out
- IVEWare: iveware-sas.sas

A note on the concordance between packages

- The 3 packages to create imputations (IVEware, mice, ice) yield somewhat different imputations
  - Slightly different procedures
  - In our example, different variables used because of computing limitations in R and Stata
  - IVEware allowed the use of the largest set of variables
- The packages to analyze multiply imputed data (zelig, glm.mids, MIcombine, micombine, mim) generate the same results when given the same imputed datasets
  - This reassuring!
  - (They are all using the same combining rules)

Outline

References: General references on missing data

- http://missingdata.org.uk/
- http://www.stat.psu.edu/∼jls/mifaq.html
- http://missingdata.lshtm.ac.uk (esp. "Bibliography" link)
- https://www.ssc.wisc.edu/sscc/pubs/stata_mi_ex.htm
- Google group: https://groups.google.com/forum/#!forum/missing-data

- Allison, PD. (2002) *Missing Data in Quantitative Applications in the Social Sciences*. Thousand Oaks, CA. Sage.
- Carpenter, J. (2006). Missing Data Example Analysis [accessed December 19, 2006]. Available online at http://www.lshtm.ac.uk/msu/missingdata/example.html
- Little, R.J. et al. (2012). The prevention and treatment of missing data in clinical trials. *New England Journal of Medicine* 367(14): 1355-1360.
- 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. *Criminology* 46(4): 841-870.

References: Books

- Carpenter, J.R., and Kenward, M.G. (2013). *Multiple imputation and its application*. Wiley.
- Enders, C.K. (2010). *Applied missing data analysis*. Guilford Press.
- 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. (2012). *Flexible Imputation of Missing Data.* Chapman & Hall/CRC, Boca Raton, FL.

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.
- 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.

References: Statistical basis for MI

- 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
- Azur, M., Stuart, E.A., Frangakis, C.M., and Leaf, P.J. (in press). Multiple imputation by chained equations: What is it and how does it work? Forthcoming in *International Journal of Methods in Psychiatric Research.*
- 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.edurs/c/smp/ive/
- Schafer, J.L. (1999). Multiple imputation: A primer. *Statistical methods in medical research* 8(1): 3-15.
- Sterne, J.A.C., et al. (2009). Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *British Medical Journal* 338:b2393.
- 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.
- White, I.R., Royston, P., and Wood, A.M. (2011). Multiple imputation using chained equations: Issues and guidance for practice. *Statistics in Medicine* 30: 377-399. (Includes Stata code snippets).

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
- Seaman, S.R., and White, I.R. (2013). Review of inverse probability weighting for dealing with missing data. *Statistical Methods in Medical Research* 22: 278-295.

References: Not missing at random models

- Enders, C.K. (2011). Missing not at random models for latent growth curve analysis. *Psychological Methods* 16(1): 1-16.
- Jackson, D., White, I.R., and Leese, M. (2010). How much can we learn about missing data? An exploration of a clinical trial in psychiatry. *Journal of the Royal Statistical Society Series A* 173(3): 593-612.
- Resseguier, N., Giorgi, R. & Paoletti, X. (2011). Sensitivity Analysis When Data Are Missing Not-at-random. *Epidemiology* 22(2): 282.
- 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.

References: Other

- Abayomi, K., Gelman, A., & Levy, M. (2008). Diagnostics for multivariate imputations. *Applied Statistics* 57(3), 273-291.
- 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 theEMAlgorithm. *Journal of the Royal Statistical Society, Series B* 39: 122.

- 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.
- Hill, J.L., Waldfogel, J., Brooks-Gunn, J., and Han, W-J. (2005). Maternal employment and child development: A fresh look using newer methods. *Developmental Psychology* 41(6): 833-850.
- 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.