Recent Advances in Missing Data Methods:
Imputation and Weighting

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

MCHB/HRSA
September 27, 2012
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...
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
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
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
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
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
Missing data is a problem in almost every 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. Discussion will focus in particular on Multiple Imputation by Chained Equations, which is particularly useful for large datasets with complex data structures. Emphasis will be on providing practical tips and guidance for implementing multiple imputation and analyzing and interpreting multiply imputed data.
Missing data

- Missing data common, especially with administrative data (e.g., medical records) 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
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

- "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
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
- 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

- $P(R|Y, X)$ cannot be simplified
- e.g., probability of someone reporting their income depends on what their income is
- e.g., probability of reporting psychiatric treatment depends on whether or not they have received it
- 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 (implemented in Stata’s mi package, possibly others)
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
- Other references: Hedeker and Gibbons (1997), Resseguier et al. (2011; sensMICE package for R), Enders (2011),
Introduction and terminology
- Understanding types of missingness

Ways of handling missing data
- (Generally) improper ways of handling missing data...
- Better ways of dealing with missing data...

Implementing multiple imputation: MICE

Conclusions

Software

References
Inappropriate ways of handling missing data

- Ignoring it
- Complete case
- Single imputation
- Missing indicator approach
- Last observation carried forward
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!
  - 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)
- (Also called listwise deletion)
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
- Works well when values non-normally distributed or have bounds (ensures that imputed values in same range as observed values)
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
- Alternative sources of information
- 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
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 (and assumes MAR given those)

Graham (2008), Siddique et al. (2008)
Alternative sources of information

- In some cases, can utilize a secondary data source to get needed information.
- Or, for example, 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.
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)
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 \( \frac{100}{80} = \frac{1}{\frac{80}{100}} = 1.25 \)

Female respondents will get weight of \( \frac{100}{75} = \frac{1}{\frac{75}{100}} = 1.333 \)

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
• 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, 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
• Relatively simple (and widely accepted) way of handling attrition/unit non-response
Now on to item non-response . . .
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
- SPSS: Missing Values add-on module ("fully conditional specification" option)
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: PRAMS data

- Hypothetical analysis using NYC PRAMS data
- Interested in predicting postpartum depression ("depress") as a function of other predictors
- Predictors include demographics, information on the delivery, health measures, etc.
- About 40 possible predictors
- But lots of missingness
- Will go through some code and diagnostics here
- Also see Stuart et al. (2009), Azur et al. (2011) for another example, including code
Step 1: Rates of missingness

High rates of missingness for some variables

<table>
<thead>
<tr>
<th>Variable</th>
<th>% Missing</th>
</tr>
</thead>
<tbody>
<tr>
<td>BMI</td>
<td>0.2%</td>
</tr>
<tr>
<td>Maternal age</td>
<td>1.6%</td>
</tr>
<tr>
<td>Diabetes status</td>
<td>2%</td>
</tr>
<tr>
<td>Freq of drinking</td>
<td>3.8%</td>
</tr>
<tr>
<td>Sad</td>
<td>8.2%</td>
</tr>
<tr>
<td>Nhope</td>
<td>11.2%</td>
</tr>
<tr>
<td>Slow</td>
<td>11.1%</td>
</tr>
<tr>
<td>Income</td>
<td>21%</td>
</tr>
</tbody>
</table>
Missingness depends on observed characteristics

- Not MCAR: means of variables differ across those with missing vs. observed values on other items
- 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)
- May make sense to include some variables in imputation model, even if not going to be used in analyses ("auxiliary variables"; Collins et al. 2001)
  - 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
- With so many variables, can’t possibly do careful model selection for each one
- Some software packages allow stepwise selection to select imputation model for each variable (IVEWare, add-on packages for Stata’s ice: pred_eq, check_eq)
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)
- Can be fairly loose and include a lot, especially if using stepwise procedures
- In PRAMS, N \( \sim 1400 \), had about 40 variables in imputation procedure
- Note: Don’t need to specify which are dependent vs. independent
General guidance/issues

- 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)
  - White et al. (2011) recommend $m = 100 \times FMI$ (FMI = fraction of missing information)
    - Since FMI hard to estimate, but Bodner’s approximation says $FMI < \%$ missing cases, approximate $m = 100 \times (\%$ missing cases)
    - e.g., 20% missing cases would imply $m = 20$
  - White et al. (2011) also argue that for reproducibility may need $m > 100$
  - Stata command to check if you’ve done enough imputations: mim mcerror
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; see Enders (2010, Chp. 5))
Variables that are functions of others

- In this case, dependent variable (depress) is a function of 3 other variables
- Best to impute the 3 variables separately, then re-create the depress summary variable after the imputations are created
  - Keeps the imputation process more general (e.g., in case you want to use the individual variables for something)
  - Retains observed values on the 3 individual variables (e.g., someone who is missing one would be missing for depress but we don’t want to throw away their 2 observed values on the other two variables)
- Once imputations can created, can manipulate those datasets basically in the same ways you normally would
- (A bit more on this later)
Implementing MICE

- In R using the mi package: `imp <- mice(data, pred=pred, maxit=10, m=10, seed=92385)`

- In Stata using ice:
  ```
  ice modeprt2 batch2 wtanl i.stratumc mmhbp vagdel m.race2 bornus plural2 married2 o.matdeg o.pnc /// lga2 sga o.kotelchuck2 o.prelb o.gestwk bmi wrknow o.income2 medicaid2 wicpreg /// o.pp_drink m.diabetes2 preexer infert_tx m.feelpg2 o.smoke o.strs inficu o.los2 o.bfeed2 back_sleep2 cosleep2 /// ppvchk i.inf_age2 o.sad o.nhope o.slow matage totcnt, saving(impute, replace) m(10) boot(mmhbp vagdel race2 /// bornus matdeg pnc lga2 sga kotelchuck2 prelb gestwk wrknow income2 medicaid2 wicpreg pp_drink diabetes2 /// preexer infert_tx feelpg2 smoke strs inficu los2 bfeed2 back_sleep2 cosleep2 ppvchk sad nhope slow matage) /// seed(1285964)
  ```
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)
- Can also do diagnostics that assess the imputation models themselves (Su et al., 2009)
- 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
Graphical summaries

- Bivariate scatterplots of observed and imputed values
- Residual plots, for observed and imputed values
- Density plots of observed and imputed values
- Next slide: Example from R mi package
Numerical diagnostics

- Some packages automatically print out some diagnostics
- Model fit information to assess imputation models
- Comparisons of original and imputed values (Abayomi et al., 2008)
  - 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
At a basic level . . .

- Can compare means of variables pre and post imputation, make sure reasonable
- Note that changes are not necessarily bad! (May in fact be fixing the problem!)

<table>
<thead>
<tr>
<th>Variable</th>
<th>Mean before</th>
<th>Mean after</th>
</tr>
</thead>
<tbody>
<tr>
<td>BMI</td>
<td>24.97</td>
<td>24.97</td>
</tr>
<tr>
<td>Freq of drinking</td>
<td>0.66</td>
<td>0.66</td>
</tr>
<tr>
<td>Sad</td>
<td>1.18</td>
<td>1.17</td>
</tr>
<tr>
<td>Nhope</td>
<td>0.56</td>
<td>0.57</td>
</tr>
<tr>
<td>Slow</td>
<td>1.23</td>
<td>1.21</td>
</tr>
<tr>
<td>Income</td>
<td>2.99</td>
<td>2.75</td>
</tr>
</tbody>
</table>
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, mi package
  - 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 ($B$)

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

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

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

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

Degrees of freedom for $t$ distribution can also be calculated (Enders, 2010, p. 231+)

See Schafer (1997) or Little and Rubin (2002) for details
Ran models predicting small for gestational age (sga; binary variable) as a function of other predictors

Note: Analyses not meant to have substantive meaning or to present actual findings on these associations

Purely illustrative examples to demonstrate the methods
PRAMS data analysis: Complete case (sga)

```
.svy: logit sga strs i.feelpg2 vagdel i.race2 income2 medicaid2 wicpreg matdeg bmi
bmi sq smoke2 matage matagesq infert_tx (running logit on estimation sample)
```

Survey: Logistic regression

<table>
<thead>
<tr>
<th>Linearized</th>
</tr>
</thead>
<tbody>
<tr>
<td>sga</td>
</tr>
<tr>
<td>-------------</td>
</tr>
<tr>
<td>strs</td>
</tr>
<tr>
<td>feelpg2</td>
</tr>
<tr>
<td>3</td>
</tr>
<tr>
<td>4</td>
</tr>
<tr>
<td>5</td>
</tr>
<tr>
<td>vagdel</td>
</tr>
<tr>
<td>race2</td>
</tr>
<tr>
<td>3</td>
</tr>
<tr>
<td>4</td>
</tr>
</tbody>
</table>

Number of strata = 2
Number of obs = 897
Population size = 76003.841
Design df = 895
F(19, 877) = 1.98
Prob > F = 0.0077
PRAMS data analysis: Single imputation (sga)

```
.svy: logit sga strs i.feelpg3 vagdel i.race3 income2 medicaid2 wicpreg matdeg bmi bmisq smoke2 matage matagesq infert_tx (running logit on estimation sample)
```

Survey: Logistic regression

Number of strata = 2  Number of obs = 1436
Number of PSUs = 1436  Population size = 115091.01
Design df = 1434  F( 19, 1416) = 2.29
Prob > F = 0.0013

|             | Linearized Coef. | Std. Err. | t    | P>|t|  | [95% Conf. Interval] |
|-------------|------------------|-----------|------|------|----------------------|
| sga         |                  |           |      |      |                      |
| strs        | 0.0906505        | 0.1258239 | 0.72 | 0.471| -0.1561681 - 0.337469 |
| feelpg3     |                  |           |      |      |                      |
| 2           | 0.1569055        | 0.3307433 | 0.47 | 0.635| -0.4918871 - 0.865981 |
| 3           | 0.322266         | 0.3297164 | 0.98 | 0.329| -0.3245122 - 0.969442 |
| 4           | 0.2078305        | 0.296976  | 0.70 | 0.484| -0.3747235 - 0.793844 |
| vagdel      | -0.2975287       | 0.1860587 | -1.60| 0.110| -0.6625051 - 0.0674478 |
| race3       |                  |           |      |      |                      |
| 2           | -0.1242725       | 0.2946427 | -0.42| 0.673| -0.7022494 - 0.4537043 |
| 3           | -0.5491899       | 0.2595075 | -2.12| 0.034| -1.058245 - 0.0401349 |
PRAMS data analysis: Multiple imputation (sga)

```
.mi estimate: svy: logit sga strs i.feelpg2 vagdel i.race2 income2 medicaid2 wicpreg matdeg bmi bmisq smoke2 matage matage sq infert_tx

Multiple-imputation estimates Imputations = 10
Survey: Logistic regression Number of obs = 1436
Number of strata = 2 Population size = 115091
Number of PSUs = 1436
Average RVI = 0.0714
Complete DF = 1434
DF adjustment: Small sample DF: min = 100.19
avg = 971.24
max = 1391.55
Model F test: Equal FMI F( 19, 1374.4) = 2.18
Within VCE type: Linearized Prob > F = 0.0023

sga | Coef. Std. Err. t P>|t| [95% Conf. Interval]
-------------+----------------------------------------------------------------
strs | .114504 .135696 0.84 0.399 -.1521104 .3811184
feelpg2 |  3 | .1470457 .3411933 0.43 0.667 -.5223776 .816469
      |  4 | .3355484 .337919 0.99 0.321 -.3274445 .9985413
      |  5 | .2565546 .3094421 0.83 0.407 -.35068 .8637892

E. Stuart (JHSPH) MCHB September 27, 2012 61 / 103
```
Complete case analysis drops a lot of subjects (almost 30%)

Output from MI models looks about the same; easy to read it just as you would standard regression output!

Some point estimates vary across the 3 approaches, some also vary in significance

It is the case that sometimes results won’t change after doing MI

But sometimes they will, and you don’t know if they will or won’t until you do it (and should trust the MI results more)

May think single imputation results “best” because they might match our expectations about sga and stresses during pregnancy (strs), but be careful of that interpretation

Single imputation approach has false precision; standard errors smaller than they should be
Example: mim command in Stata (other data)

mim: logit dsmmood sex age

Multiple-imputation estimates (logit)  Imputations = 10
Logistic regression  Minimum obs = 9185
  Minimum dof = 4.8

|          | Coef. | Std. Err. | t    | P>|t|  | [95% Conf. Int.] | MI.df |
|----------|-------|-----------|------|------|-----------------|-------|
| dsmmood  |       |           |      |      |                 |       |
| 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   |
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:
  \[ \gamma = \frac{(r + 2)/(df + 3)}{r + 1} \]
  \[ r = \frac{(1 + 1/m) \times B}{U} \]

- \( r \) is the relative increase in variance due to the nonresponse/missing data
- Alternatively (Enders, 2010, p. 225): \( FMI = \frac{B + B/m + 2/\nu + 3}{T} \)
  - \( \nu \) 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.
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)
- 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
### Table VIII. Common statistics that can and cannot be combined using Rubin’s rules (equations (1) and (2)).

<table>
<thead>
<tr>
<th>Statistics that can be combined without any transformation</th>
<th>Mean, proportion, regression coefficient, linear predictor, C-index, area under the ROC curve</th>
</tr>
</thead>
<tbody>
<tr>
<td>Statistics that may require sensible transformation before combination</td>
<td>Odds ratio, hazard ratio, baseline hazard, survival probability, standard deviation, correlation, proportion of variance explained, skewness, kurtosis</td>
</tr>
<tr>
<td>Statistics that cannot be combined</td>
<td>$P$-value, likelihood ratio test statistic, model chi-squared statistic, goodness-of-fit test statistic</td>
</tr>
</tbody>
</table>
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
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 if you are really worried about NMAR?  
  ● Sensitivity analyses (e.g., 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 $\alpha$'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)
- MLWwiN macro for imputing multi-level data: www.missingdata.org.uk
- Yucel (2008), Reiter et al. (2006)
What if imputing data within a study 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)
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 weighting

- Fairly easy to implement
- Allows inferences to be made for original full sample
- An established way of handling attrition/non-response
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
If just have a simple attrition problem, use non-response weights. But if things more complex . . .

If rates of missingness low (e.g., 1-2%), consider doing single imputation (e.g., regression prediction with noise)

Otherwise use multiple imputation
  Very general, flexible

With either approach, it may (or may not) change the results, but won’t know until you do it, and using multiply imputed data should be more accurate and correct

Either approach also allows you to make statements about the whole sample, not just about the respondents/those with complete data (and uses that same sample for all analyses; sample not dependent on the variables you happen to use in a given analysis)
MICE approaches can be useful, especially for large datasets with lots of categorical variables

- Make imputation models very general: lots of terms and interactions (little cost to including lots of potential predictors)
- 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
General resources

- Code for many packages in Horton and Kleinman (2007)
Software for MICE: R

- mi package (Su et al., 2011)
  - 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/clumped 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/hmtl/mice.htm](http://web.inter.nl.net/users/S.van.Buuren/mi/hmtl/mice.htm)
  - For creating and analyzing multiply imputed data
  - Multiple imputation by chained equations
  - Some good diagnostics, added functionality recently
  - Can incorporate bounds, restrictions
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
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 mianalyze
  - 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
  - Note: New in Stata 12, the ”mi” package can do the ice procedure
  - For pre-Stata 12 users: can easily go between ice and mi procedures using “mi import ice” and “mi export ice” commands so can use mi’s procedures, like for analyzing MI data
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
- http://www.ats.ucla.edu/stat/Stata/library/ice.htm
- http://www.ats.ucla.edu/stat/stata/seminars/missing_data/mi_in_stata_
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)
- 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)
Software for SPSS

- Missing values add-on package: Creates imputations and analyzes MI data
- Monotone or MICE approaches
- Option to include all two-way interactions in prediction models

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
References: General references on missing data and MICE

- http://missingdata.org.uk/
- http://www.stat.psu.edu/~jls/mifaq.html
References: Books


References: Guidance for dealing with missing data


References: Tutorials and software for implementing MI

- www.multiple-imputation.com
- MI FAQ’s: http://www.stat.psu.edu/~jls/mifaq.html


References: Survey weighting for nonresponse

References: Not missing at random models


