Handling Missing Data

Estie Hudes

Tor Neilands

UCSF Center for AIDS Prevention Studies

Part 2

February 27, 2015

Contents

1. Summary of Part 1
2. Multiple Imputation (MI) for normal data
3. Multiple Imputation (MI) for normal & non-normal data
4. Steps in multiple imputation
5. Software for performing multiple imputation
6. Example of basic multiple imputation-basic analysis
7. Extensions to more complex scenarios
8. Myths about MI with collaborator talking points
9. Conclusions regarding multiple imputation
10. Overall summary and conclusion of parts 1 and 2
11. Acknowledgements
12. References
13. Appendix: EM Algorithm and partial list of software

Summary of Part 1 (1)

- Missing data are a ubiquitous problem in applied research
- Incomplete data arise due to different mechanisms: MCAR, MAR, NMAR.
- Ad hoc methods for handling missing data assume MCAR and can result in parameter estimate bias and loss of power for hypothesis testing.

Summary of Part 1 (2)

- Methods that assume incomplete data arise from a MAR process are generally recommended over ad hoc methods that assume MCAR.
  - Inverse probability of censoring weights (IPCW)
  - Fully Bayesian estimation
  - Maximum likelihood (Part 1)
  - **Multiple Imputation (Part 2; this presentation)**

Multiple Imputation (1)

- Like the single imputation approaches discussed in Part 1 (e.g., mean substitution), in MI missing values are imputed and then used in standard statistical software routines.
- What is unique about MI: We impute **multiple** data sets to analyze, not a single data set as in single imputation approaches
  - Use the expectation-maximization (EM) algorithm to obtain starting values for MI (see Appendix for details on EM)
  - **The differences among the imputed data sets capture the uncertainty due to imputing values**
  - **The actual values in the imputed data sets are less important than analysis results combined across all data sets**

Multiple Imputation (2)

- Several MI advantages:
  - MI yields consistent, asymptotically efficient, and asymptotically normal estimators under MAR (same as direct ML)
  - MI-generated data sets may be used with any kind of software or model (but see White and Royston, *SIM*, 2011, v. 30, pp. 377-399, Table VIII for limitations on what kinds of statistics can be combined):
    - 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

Multiple Imputation (3)

- Imputation model vs. analysis model
  - Imputation model should include any variables and effects of interest plus *auxiliary variables* (i.e., variables that are strongly correlated with other variables that have incomplete data; variables that predict data missingness)
  - Analysis model should contain either only the variables and effects from the imputation model or a subset of the variables and effects from the imputation model.
- Texts that discuss MI in detail:
  - Rubin (1987, John Wiley and Sons): Non-response in surveys
  - J. L. Schafer (1997, Chapman & Hall): Modern and updated
  - Little & Rubin (2002, John Wiley and Sons): A classic, updated
  - P. Allison (2002, Sage Publications series # 136): A readable and practical overview of and introduction to MI and missing data handling approaches
  - Enders (2010): Readable and comprehensive

Multiple Imputation (4)

- Multivariate normal (MVN) imputation approach
  - MI approaches exist for multivariate normal data, categorical data, mixed categorical and normal variables, and longitudinal/clustered/panel data
  - The MVN imputation approach was popular in the 1990s and the first decade of the 2000s because it performs well in most applications, even with somewhat non-normal input variables (Schafer, 1997)
    - Variable transformations can further improve imputations
    - Some authors assert this method is as good as others for multivariate data with arbitrary patterns of missingness (Enders, 2010)

Multiple Imputation (5)

- Multivariate normal imputation approach
  - For each variable with missing data, estimate the linear regression of that variable on all other variables in the data set.
  - Using a Bayesian prior distribution for the parameters, typically noninformative, regression parameters are drawn from the posterior Bayesian distribution. Estimated regression equations are used to generate predicted values for missing data points.
  - Add to each predicted value a random draw from the residual normal distribution to reflect uncertainty due to incomplete data.

Multiple Imputation (6)

- Obtaining Bayesian posterior random draws is the most complex part of the procedure. Two approaches:
  - Data augmentation - implemented in the freeware program NORM, SAS PROC MI, and Stata's -mi impute mvn-. Uses a Markov-Chain Monte Carlo (MCMC) approach to generate the imputed values
  - Sampling Importance/Resampling (SIR) - implemented in Gary King's Amelia program; claimed to be faster than data augmentation-based approaches.
- "The relative superiority of these methods is far from settled" (Allison, 2002, p. 34)
- These methods work fairly well for non-MVN data as long as most variables are approximately normally distributed, there are no multi-category nominal variables, and the analysis models take into account potential assumption violations (e.g., non-normality).

Multiple Imputation (7)

- For non-normal variables or nominal variables with missing values, consider Multiple Imputation through Chained Equations (MICE), a variant of data augmentation
  - Uses a Gibbs sampler and switching regressions approach (Fully Conditional Specification - FCS) to generate the imputed values (van Buuren, 2007)
  - Treating the variable with the least amount of missing data as the first outcome, the approach uses a series of regression models to fill in missing values for that outcome. Then values for the next variable with the second least missing data are imputed using another regression equation, and so on for all variables with missing values.
  - The approach proceeds iteratively until a steady state is reached.

Multiple Imputation (8)

- In SAS, MICE is available via the FCS statement in PROC MI
- In Stata, MICE was originally implemented in the user-written Stata program -ice- and subsequently in Stata's official command -mi impute chained-
- There is less theoretical justification for the MICE approach relative to the MVN methods described previously. However, a key benefit of the MICE approach is that multivariate normality need not be assumed.
  - For example, in Stata supported distributions include linear regression (regular, truncated, interval), logistic (binary, ordinal, multinomial), Poisson, and negative binomial. (Vittinghoff et al. *Regression Methods in Biostatistics*, 2012, p. 448).
  - In SAS supported distributions include linear and logistic (binary, ordinal, multinomial) regression.

Multiple Imputation (9)

- Steps in using MI
  - Select variables for the imputation model. Use all variables in the analysis model, including any dependent variable(s), and any variables that are associated with variables that have missing data or the probability of those variables having missing data (auxiliary variables), in part or in whole. Be sure to include any interaction or polynomial terms in the imputation model.
  - Transform non-normal continuous variables to attain normality (e.g., skewed variables), especially if using the MVN imputation method. Consider the bootstrap option in Stata's -mi impute chained- command.
  - Select a random number seed to ensure replicable results.

Multiple Imputation (10)

- Steps in using MI
  1. Perform pre-imputation diagnostics.
  2. Choose the number of imputations to generate.
  3. Generate the imputations and evaluate whether the chosen number of imputations was sufficient; if not, generate more imputations.
     - Historically, the number of imputations was typically 5 to 10 because early literature showed > 90% coverage & efficiency in large sample scenarios with M = 5 to 10 imputations (Rubin, 1987; Schafer, 1999) and it was tedious to move data and results back and forth between dedicated imputation programs and statistical analysis programs.
     - But there the focus was on effectively estimating quantities such as means and proportions from national survey data. Newer recommendations focus on accurately estimating parameters and also on standard error efficiency and p-values for hypothesis testing in the context of regression analyses.

Multiple Imputation (11)

- A newer rule of thumb: Generate as many imputations as you have percentage of values missing. (e.g., for 50% missingness, generate at least 50 imputations). It is often as easy to generate and analyze 50 vs. 5 imputations with modern software and computing power. See http://www.statisticalhorizons.com/more-imputations for further discussion of this issue.
- Another approach is to use the FMI (Fraction of Missing Information: ”The ratio of information lost due to the missing data to the total information that would be present if there were no missing data”; Stata 13 mi.pdf, pp. 361 & 366) and set the number of imputations greater than or equal to 100 times the largest FMI value (Stata 13 mi.pdf, p. 48). Generate more imputations if the largest FMI across estimates indicates the originally chosen number was insufficient.

Multiple Imputation (12)

- Steps in using MI (continued):
  4. Produce the multiply imputed data sets
  5. Estimated parameters must be independent of initial values. Perform MI diagnostics to check the soundness of the imputations:
     - Pre-imputation: For MVN imputations, assess independence via autocorrelation and time series plots. For MICE/FCS approach, examine time series plots.
     - Post-imputation: In Stata, the add-in command -midiagplots- compares the distributions of the observed and observed+imputed values via plots for continuous variables and proportions for categorical variables. You can do this by hand in other software programs.

Multiple Imputation (13)

- Steps in using MI (continued):
  6. Back-transform any previously transformed variables
  7. Analyze each imputed data set using standard statistical approaches. If you generated *M* imputations (e.g., 50), you would perform *M* separate, but identical analyses (e.g., 50).
  8. Combine results from the *M* multiply imputed analyses (using SAS PROC MIANALYZE, Stata -*mi estimate*-, etc.) using Rubin's (1987) formulas to obtain a single set of parameter estimates and standard errors. Both *p*-values and confidence intervals may be generated.

Multiple Imputation (14)

- The MI point estimate is the mean:

$$\overline{Q} = \frac{1}{M}\sum_{i=1}^{M} Q_i$$

- The MI variance estimate is the sum of Within and Between imputation variation:

$$V = \overline{W} + (1 + \frac{1}{M})B$$

where

$$\overline{W} = \frac{1}{M}\sum_{i=1}^{M} W_i$$

$$B = (1 - \frac{1}{M})\sum_{i=1}^{M} (Q_i - \overline{Q})^2$$

- (Q_i and W_i are the parameter estimate and its variance in the ith imputed dataset)

Multiple Imputation (15)

- Steps in using MI (continued)
  - Rules for combining parameter estimates and standard errors
    - A parameter estimate is the mean of the parameter estimates from the multiple analyses you performed.
    - The standard error is computed as follows:
      - Square the standard errors from the individual analyses.
      - Calculate the variance from the squared *SE*s across the *M* imputations.
      - Add the results of the previous two steps together, applying a small correction factor to the variance in the second step, and take the square root. (see previous slide)

    ✓-mi estimate- in Stata and PROC MIANALYZE in SAS do this combining automatically.

    ✓There is a separate *F*-statistic available for multiparameter inference (i.e., multi-DF tests of several parameters at once). -mi test- in Stata is available as a post-estimation command for this purpose.

Multiple Imputation (16)

- How should dependent variables be handled?
  - Given that the goal of MI is to reproduce all the relationships in the data as closely as possible, this can only be accomplished if all the dependent variable(s) are included in the imputation process. Not including a DV is akin to assuming that its relationships with other variables are zero.
  - Exceptions (Acock, 2012):
    - When all Xs are complete and the only missing values are on Y, imputing the missing Y values adds no information to the estimate of Y from X.

Multiple Imputation (17)

- Multiple Imputation-then-Delete (MID) Strategy:
- Perform MI as usual, including all variables. Then drop observations which had an imputed outcome values (Von Hippel, Regression with missing Y's: An improved strategy for analyzing multiple imputed data. *Sociological Methodology, 37, 83–117,* 2007).
  - The quality of the standard errors improves with deletion of imputed Y-values.
  - The advantage of MID is slight for small to moderate amounts of missingness, but if, for instance, attrition is high in a longitudinal study, use of MID can confer a substantial advantage.
  - However, if auxiliary variables related to X and Y are included, additional information may be gained by imputing Y values. If auxiliary variables are used in the imputation phase, retain the imputed Y-values.

Multiple Imputation (18)

- SAS: PROC MI and PROC MIANALYZE
  - PROC MI produces imputations: For arbitrary missingness patterns, MCMC assuming MVN and FCS via chained equations methods are available
  - PROC MIANALYZE combines results from analyses of imputed data into a single results set
- MI Software from Stata:
  - -mi impute mvn- generates imputations under the assumption of multivariate normality
  - -mi impute chained- generates imputations via the MICE approach
  - -mi estimate- is used to perform analyses and combine results
    - Post-estimation commands are available following -mi estimate- (e.g., -mi test-).
  - Two very useful entries in the Stata MI manual can be viewed by issuing *findit mi workflow* and *findit mi glossary* and then go to the appropriate place in the MI manual (e.g, for Stata 13: p. 343 & 353, respectively)

Multiple Imputation (19)

- SAS and Stata also offer predictive mean matching (PMM) for continuous variables. The Stata documentation describes PMM as an approach that "combines the standard linear regression and the nearest-neighbor imputation approaches."
  - PMM guarantees imputed values are within the range of observed values and may be useful when continuous variables are non-normal.
  - PMM is less useful when imputed values may lie outside the range of observed values.
- Many other programs offer multiple imputation options. A partial list of programs is shown in the appendix.

Stata MI Example (1)

Steps in using Stata to perform MI via ICE/FCS (the steps and logic are similar for other programs):

- Describe patterns of missing data
- Let Stata know which variables are to be imputed: *register* analysis variables as *imputed*, *regular,* or *passive*
- Do a *dry run* to make sure prediction equations are as desired
- Generate trace plots to evaluate the adequacy of number of the number of burn-in iterations
- Generate the multiple imputation data sets.

Stata MI Example (2)

Steps to perform MI via ICE/FCS (continued):

- After imputed values have been generated, use

  -midiagplots- to compare the observed+imputed values' distributions with those of the original observed values. They should be similar.
- Perform desired inferential analyses (typically some sort of regression model or models, though not always), on the imputed data using -mi estimate-.
- Perform any desired post-estimation commands using Stata's -mi- post-estimation features, e.g., -mi test-.

MI Example 1: Tobacco and Bars Data

[Same as ML Example 1 from Part 1]

- Dr. Pam Ling and her research group at the UCSF Center for Tobacco Control Research and Education (CTCRE) administered a brief survey to 1,217 young adult bar patrons in San Francisco. The design features clustered data from participants gathered within bars using a 3-form survey design with planned missingness. Variables available for analysis include:
  - Number of days in past 30 the participant (PPT) smoked (the outcome; *n* = 1145)
  - Age in years from 18-26 (*n* = 1217)
  - Race (White, Latino, Black, A/PI, Other; *n* = 1207)
  - Male gender dummy variable (*n* = 1217)
  - Sexual Orientation (Gay, Straight, Bi, Other; *n* = 1212)
  - Participant considers self a smoker (0 = no; 1 = yes; *n* = 858)
  - Social network smoking: Sum of ordinal items asking how many friends, partying companions, and coworkers smoke (*n* = 616)
  - Extraversion index: Sum of ordinal outgoingness items (*n* = 801)
- This example demonstrates a linear regression analysis of data from a planned missingness design using multiply-imputed data sets in Stata.
  - For illustrative purposes, venue clustering is not accounted for in this example.

Three Form Design (N=1217)

	Venue ID	Days smoked in past 30 (continuous)	Age in years (continuous)	Race (categorical)	Male gender (binary)	Sexual Orientation (categorical)
Form X	Yes	Yes	Yes	Yes	Yes	Yes
Form Y	Yes	Yes	Yes	Yes	Yes	Yes
Form Z	Yes	Yes	Yes	Yes	Yes	Yes

	Do you consider yourself to be a smoker? (binary n=858)	How many people in your social network smoke? (continuous n=616)	Extraversion Index (continuous n=801)	Smokes within 30 min of waking (Binary Yes/No; can be used as an auxiliary variable)
Form X	Yes	no	Yes	Yes
Form Y	no	Yes	Yes	Yes
Form Z	Yes	Yes	no	Yes

MI Stata Example 1: Results

- *M* = 75 imputations were used based on the FMI output from Stata's -mi estimate- command.
- There are no significant effects for age, race, male gender, and extraversion.
- There is a significant overall difference for LGBT status, with gay participants having a lower mean number of smoking days and bisexual and other ppts having a higher mean number of smoking days relative to the straight participant reference group ($F[3, 892] = 6.52$, $p = .0002$).
  - The finding for gay participants is contrary to the literature showing that LGBT groups have higher rates of smoking
  - Additional analyses suggest a suppression effect such that R^2 increases when the gay sexual orientation variable is omitted from the model.
- Self-identification as a smoker is positively associated with the number of days smoked ($B = 14.42$, $p < .001$).
- More smoking occurring in one's social network is associated with more days smoked ($B = .40$, $p = .01$).
- Results are substantively the same as what was obtained with FIML using Stata -sem- command.
- With MI, however, fitting many other types of models (e.g., GEEs) is possible.

MI Stata Example 1: Comparison with FIML

Effect	FIML	MI
Age	.105 (.159), p=.508	.109 (.159), p=.489
Male	.194 (.523), p=.711	.201 (.543), p=.711
Self-Reported Smoker	14.43 (.671), p<.001	14.42 (.767), p<.001
Network Smoking	.414 (.150), p=.006	.402 (.155), p=.010
Extraversion	-.037 (.211),p=.859	-.031 (.212),p=.885
Race	$\chi^2(4)=5.74$, p=.219	F(4, 1010)=1.41, p=.229
African American	-.943 (.632), p=.136	-.969 (.647), p=.135
Latino	.815 (1.17), p=.487	.693 (1.20), p=.562
Asian/Pacific Islander	-.597 (.818), p=.465	-.579 (.823), p=.482
Other	1.10 (.921), p=.233	1.14 (.918), p=.215
LGBT	$\chi^2(3)=21.68$, p=.0001	F(3, 892)=6.52, p=.0002
Gay	-2.75 (.990), p=.005	-2.67 (.978), p=.006
Bisexual	3.05 (1.15), p=.008	3.20 (1.19), p=.007
Other	3.00 (1.25), p=.016	2.78 (1.35), p=.040

MI Stata Example 1: Non-Convergences Remedies

- Convergence can be an issue, especially for MICE methods.
- You can diagnose the source model for a non-convergent imputation running the imputation models one by one. In Stata you can use the command -set more off- and then specifying the -noisily- option in -mi impute chained-.
- Some possible remedies (all were demonstrated in this example):
  - Use the -augment- option to address zero cell problems for logistic models involving binary, ordinal, or nominal outcomes
  - Use PMM rather than ordinal logistic for ordinal variables with many levels
  - Set the maximum number of iterations for iterative estimation methods to some reasonable number (e.g., 50)
  - Treat ordinal variables as continuous when predicting other variables (as long as you are comfortable assuming a constant decreasing or increasing linear association between the ordinal predictor and the various outcomes)
  - Shift the order in which the chained equations are run, placing the most difficult equations (typically multinomial models) last

MI Extensions (1)

- MI works well in relatively basic scenarios such as the linear models examples just presented. What about more complex analyses? Let's consider a few common situations.
- Non-linearity and interaction
  - If there is one focal grouping variable with a few levels (e.g., male vs. female; intervention vs. control) and there are sufficient data within each group, consider generating separate imputations by group and then combining the imputed datasets for analyses. This approach allows for different means, variances, and covariances by group.
  - For all other situations, include each product and polynomial term in the imputation model as "just another variable" (JAV). (see White et al Statistics in Medicine article mentioned on the next slide)
  - Stata has an -mi passive- command that will ensure consistency of derived variables across imputed data sets. For instance, a variable *ab* defined as the product of variables *a* and *b* will be equal to *a***b* in all imputed data sets.

MI Extensions (2)

- However, approaches that "fix" the imputed values to make them consistent within imputed data sets (including Stata's passive method) can lead to biased regression coefficients. (see http://www.utexas.edu/lbj/sites/default/files/file/news/Transform%20then%20impute.pdf).
- Some passive imputation approaches are less biased than others; see White, Royston, and Wood, 2011, "Multiple imputation using chained equations: Issues and guidance for practice," 2010, *Statistics in Medicine*, 30, 4, 377-399, (http://onlinelibrary.wiley.com/doi/10.1002/sim.4067/pdf), for an excellent treatment of these and other issues involved in using multiple imputation via chained equations.
- See http://www.ssc.wisc.edu/sscc/pubs/stata_mi_intro.htm for an excellent "how to" introduction to doing MI in Stata that covers some of these same issues.

- Unsupported estimation commands in Stata: First try -cmdok- option of -mi estimate-; if that doesn't work, try the older user-written imputation combining commands (e.g., -micombine-)

MI Extensions (3)

- Variable and model selection: Applying standard variable and model selection methods to multiple imputed data sets may result in different models and variables being chosen across different imputed data sets.
  - For variable selection, it is better to use the MI data and inference based on Rubin's rules for combining imputed data sets rather than using the listwise data set. See: Wood AM, White IR, Royston P. "How should variable selection be performed with multiply imputed data?" *Statistics in Medicine*, 2008, 27, 3227-46.
  - There is also a MI-LASSO technique available via a SAS macro program that treats the multiple regression coefficients for a given variable across the imputed data sets as a single group to yield consistent variable selection across MI data sets. See: Chen Q, Wang S. "Variable selection for multiply-imputed data with application to dioxin exposure study," *Statistics in Medicine,* 2013;32:3646-59.
  - For model selection in the context of GEE, the mean QIC statistic across multiply imputed data sets works well. Shen C-W, Chen Y-H. See: "Model selection of generalized estimating equations with multiply imputed longitudinal data," *Biometrics* 2013; 55(6), 899-911.

MI Extensions (4)

- Survival analysis with missing covariate data: A common practice is to include the event indicator and the log of the time-to-event in the imputation phase. A less biased approach uses the Nelson-Aalen estimator of the baseline hazard function *H(T)* and the event indicator in the imputation phase. See: White IR, Royston P. "Imputing missing covariate values for the Cox model," *Statistics in Medicine*, 2009;28:1982-98.
- Item-level missingness: Is it better to impute at the item-level or the scale-level for survey scales with multiple items? Analyses based on scale-level imputations yield unbiased estimates, but so do item-level imputations and item-level imputations may have more efficiency and thus more power. Item-level imputations are recommended whenever feasible, but sometimes item-level imputations are infeasible when hundreds of items across multiple scales are present. Gottscall AC, West SG, Enders CK. "A Comparison of Item-Level and Scale-Level Multiple Imputation for Questionnaire Batteries," *Multivariate Behavioral Research*, 2012;47:1-25.

MI Extensions (5)

- Clustered Data Structures
  - If you have a limited number of fixed time points in a longitudinal design transform the "long" clustered data structure to a "wide" format in which multiple time points are cast as multiple variables, perform MI, and retransform the imputed data sets into the "long" form for analysis.
  - Limited number of clusters (e.g., < 30) in hierarchically structured data sets: Include dummy variables for *K*-1 clusters, where *K* is the number of clusters (Graham et al, 2009, *Annual Review of Psychology*).
  - Large numbers of clusters: Consider M*plus*, which can impute ordinal and continuous variables under two-level and three-level multilevel data structures.
- NMAR situations - rely on a priori knowledge of missingness mechanism
  - Pattern-mixture models
  - Selection models (e.g., Heckman's model)
  - MI-based sensitivity analyses available in SAS PROC MI (Vittinghoff et al., *Regression Methods for Biostatistics*, 2012, p. 463)

MI: Myths and Colleague Talking Points (1)

- *MI is making up data.*

  Like direct ML, MI is preserving and utilizing the available information to obtain the best point estimates, standard errors, and *p*-values. It is making best use of all of the data that investigators worked so hard to get in the first place. Also, we don't focus on the individual imputed data sets singly; they are just a means to the end of getting optimal regression estimates and standard errors – the fluctuation of imputed values across the multiple data sets quantifies the inherent uncertainty in imputing missing values.
- *There are too much missing data to use MI*.

  ML and MI are needed most when the sample size is small. For instance, simulation studies have shown MI outperforms complete-case analysis at *N*s as low as 50 with 50% of the data missing. (Graham & Schafer, 1999, in R. Hoyle (ed) *Statistical Strategies for Small Sample Research*, pp. 1-29)

MI: Myths and Colleague Talking Points (2)

- *Reviewers will never accept a paper or grant proposal using MI.*

  In the late 1990s and early 2000s when MI was relatively new, involved explanations and justifications were sometimes needed to convince skeptical reviewers. Now the techniques have been around for decades, have entered the mainstream, and practical articles and textbooks have been written (see References section). With its entry into supported software routines of major software companies like SPSS, SAS and Stata, MI is now part of the normal analysis landscape.
- *MI is too complicated and time-consuming to be worthwhile*.

  MI (and ML) gives optimal answers in terms of best point estimates, standard errors and *p*-values. Plus it maximizes our chances, in a legitimate way, to find interesting and significant results. While there is still more work involved in using MI relative to listwise deletion, modern software routines and computing power and make it ever faster and more convenient to use.

Multiple Imputation Summary

- MI is flexible: imputed datasets can be analyzed using many parametric and non-parametric techniques. Imputed datasets are also available for performing informal diagnostics and data explorations.
- MI is available in SAS, Stata, SPSS, and many other stand-alone (e.g., M*plus*) and integrated software programs (e.g., R)
- Multiple imputation is non-deterministic: you get a different result each time you generate imputed data sets (unless the same random number seed is used each time)
- It is easy to include auxiliary variables in the imputation model to improve the quality of imputations
- Compared with ML, large numbers of variables may be handled more easily
- MI may be used in sensitivity analyses to evaluate NMAR missingness (Vittinghoff et al., *Regression Methods for Biostatistics*, 2012, p. 463)

Overall Conclusions (Parts 1 and 2)

- Planning ahead can minimize missing cross-sectional responses and longitudinal loss to follow-up
- Use of ad hoc methods is not harmful for small amounts of missing data (e.g., < 5%; see Roth, 1994), but otherwise can lead to biased results and loss of power for hypothesis testing
- Modern methods are readily available for MAR data
  - Direct ML: most convenient for models that are supported by available software and when parametric assumptions are met
  - Multiple Imputation: Available and effective for most remaining situations
- Imputation strategies for clustered data and non-linear analyses are available, but are more complicated to implement
- Models for NMAR data are available, but are still more complicated and rest on tenuous assumptions regarding how the data came to be missing. Sensitivity analyses may be useful to gauge the tenability of the MAR assumption.

Acknowledgements

- NIMH P30 MH062246 (Lightfoot, PI) Methods Core supported time to prepare examples and slides.
- Pamela Ling, MD: Tobacco and bars data; three-form design table
- Mallory Johnson, PhD: MBSR data set
- Steve Gregorich, PhD: Overall mentoring on missing data; review of previous slides
- Melissa Krone and Jesse Canchola: Contributions to early versions of this presentation in the early 2000s.

References (1)

- Acock, A. C. (2012). What to do about missing values. APA handbook of research methods in psychology, Vol 3: Data analysis and research publication. H. Cooper, P. M. Camic, D. L. Long et al., American Psychological Association. **3:** 27-50.
- Allison, P. (2002). Missing Data. Thousand Oaks, CA, Sage Publications.
- Collins, L. M., et al. (2001). "A comparison of inclusive and restrictive strategies in modern missing data procedures." Psychological Methods, **6**(4): 330-351.
- Chen, Q. and S. Wang (2013). "Variable selection for multiply-imputed data with application to dioxin exposure study." Statistics in Medicine **32**(21): 3646-3659.
- Davey, A. and Savla, J. (2010). Statistical Power Analysis with Missing Data: A Structural Equation Modeling Approach. New York: Routledge Academic.
- Dempster, A. P., et al. (1977). "Maximum Likelihood from Incomplete Data via the EM Algorithm." Journal of the Royal Statistical Society Series B - Statistical Methodology, **39**(1): 1-38.
- Enders, C.K. (2010). Applied Missing Data Analysis. New York, NY: Guilford.
- Gottscall, A. C., et al. (2012). "A Comparison of Item-Level and Scale-Level Multiple Imputation for Questionnaire Batteries " Multivariate Behavioral Research, **47**(1): 1-25.
- Graham, J.W. (2012). Missing Data: Analysis and Design. New York: Springer.
- Graham, J. W. (2009). "Missing Data Analysis: Making It Work in the Real World." Annual Review of Psychology, **60**: 549-576.
- Graham, J. W., et al. (1996). "Maximizing the usefulness of data obtained with planned missing value patterns: An application of maximum likelihood procedures." Multivariate Behavioral Research, **31**(2): 197-218.
- Graham, J. W. and J. L. Schafer (1999). On the Performance of Multiple Imputation for Multivariate Data with Small Sample Size. Statistical Strategies for Small Sample Research. R. H. Hoyle. Thousand Oaks, CA, Sage Publications**:** 1-32.
- Hartley, H. O. (1958). "Maximum likelihood estimation from incomplete data." Biometrics, **14**(174-194).

References (2)

- Jamshidian, M. and R. I. Jennrich (2000). "Standard errors for EM estimation." Journal of the Royal Statistical Society Series B - Statistical Methodology, **62**(2): 257-270.
- Little, R. J. A. and D. B. Rubin (2002). Statistical Analysis with Missing Data. New York, John Wiley and Sons.
- Roth, P. L. (1994). "Missing data: A conceptual review for applied psychologists." Personnel Psychology, **47**(3): 537-560.
- Rubin, D. B. (1976). "Inference and missing data (with discussion)." Biometrika, **63**: 581-592.
- Rubin, D. B. (1987). Multiple Imputation for Nonresponse in Surveys. New York, John Wiley and Sons.
- Schafer, J. L. (1997). Analysis of Incomplete Multivariate Data. New York, Chapman and Hall.
- Schafer, J. L. and J. W. Graham (2002). "Missing data: Our view of the state of the art." Psychological Methods, **7**(2): 147-177.
- Shen, C.-W. and Y.-H. Chen (2013). "Model selection of generalized estimating equations with multiply imputed longitudinal data." Biometrics, **55**(6): 899-911.
- Van Buuren, S. (2007). "Multiple imputation of discrete and continuous data by fully conditional specification." Statistical Methods in Medical Research, **16**(3): 219-242.
- Von Hippel, P. T. (2007). "Regression with missing Y's: An improved strategy for analyzing multiple imputed data." Sociological Methodology, **37**: 83-117.
- Vittinghoff, E., et al. (2012). Regression Methods in Biostatistics. New York, NY, Springer
- Wood, A. M., et al. (2008). "How should variable selection be performed with multiply imputed data?" Stat Med, **27**(17): 3227-3246.
- White, I. R. and P. Royston (2009). "Imputing missing covariate values for the Cox model." Statistics in Medicine **28**(15): 1982-1998.
- White, I. R., et al. (2011). "Multiple imputation using chained equations: Issues and guidance for practice." Stat Med, **30**(4): 377-399.
- Zhao Y, Long Q. Multiple imputation in the presence of high-dimensional data *Statistical Methods in Medical Research,* 2013. http://smm.sagepub.com/content/early/2013/11/25/0962280213511027.full

Appendix: EM Algorithm (1)

- EM algorithm proceeds in two steps to generate ML estimates for incomplete data: *Expectation* and *Maximization*. The steps alternate iteratively until convergence is attained.
- Seminal article by Dempster, Laird, & Rubin (1977), *Journal of the Royal Statistical Society, Series B,* 39, 1-38. Early treatment by H.O. Hartley (1958), *Biometrics*, 14(2), 174-194.
- Goal is to estimate *sufficient statistics* that can then be used for substantive analyses. In normal theory applications these would be the mean vector and variance-covariance matrix of the variables.

Appendix: EM Algorithm (2)

- Example from Allison, 2002, pp. 19-20: For a normal theory regression scenario, consider four variables X1 through X4 that have some missing data on X3 and X4.
- Starting Step (0):
  - Generate starting values for the means and covariance matrix. One can use the usual formulas with listwise or pairwise deletion.
  - Use these values to calculate the linear regression of X3 on X1 and X2. Similarly for X4.
- Expectation Step (1):
  - Use the linear regression coefficients and the observed data for X1 and X2 to generate imputed values of X3 and X4.

Appendix: EM Algorithm (3)

- Maximization Step (2):
  - Use the newly imputed data along with the original data to compute new estimates of the sufficient statistics (e.g., means, variances, and covariances)
    - Use the usual formula to compute the mean
    - Use modified formulas to compute variances and covariances that correct for the usual underestimation of variances that occurs in single imputation approaches.
- Cycle through the expectation and maximization steps until convergence is attained (sufficient statistic values change slightly from one iteration to the next).

Appendix: EM Algorithm (4)

- EM Advantages:
  - One only needs to assume incomplete data arise from MAR process, not MCAR
  - Fast (relative to MCMC-based multiple imputation approaches)
  - Applicable to a wide range of data analysis scenarios
  - Uses all available data to estimate sufficient statistics
  - Fairly robust to non-MVN data
  - Provides a single, deterministic set of results
  - May be all that is needed for non-inferential analyses (e.g., Cronbach's alpha or exploratory factor analysis)
  - Lots of software (commercial and freeware)

Appendix: EM Algorithm (5)

- EM Disadvantages:
  - Produces correct parameter estimates, but standard errors for inferential analyses will be biased downward because analyses of EM-generated data assume all data arise from a complete data set without missing information. The analyses of the EM-based data do not properly account for the uncertainty inherent in imputing missing data.
    - There are various numerical methods by which appropriate standard errors may be generated for EM-based parameter estimates (Jamshidian & Jennrich, 2000)
    - Bootstrapping may also be used to overcome this limitation
- As with ML, EM algorithms must be available in software or programmed. An alternative, Multiple Imputation (MI), covers situations where ML and EM are neither available nor practical.

Appendix: MI Software (6)

- Available imputation software for multiple imputation (partial list):
  - NORM - for MV normal data (J. L. Schafer)
    - Windows freeware; S-Plus MISSING library; R (add-in file)

    http://sites.stat.psu.edu/~jls/misoftwa.html
  - CAT, MIX, and PAN - for categorical data, mixed categorical/normal data, and longitudinal or clustered panel data respectively (J. L. Schafer)
    - S-Plus MISSING library; R (add-in file)
  - LISREL - http://www.ssicentral.com
  - M*plus*: Version 7 supports continuous normal, binary, and ordinal variables via several different methods of imputation. A unique feature of M*plus* is its ability to generate imputations for hierarchically nested or clustered (i.e., multilevel) data sets.
  - SPSS: AMOS will perform multiple imputation for continuous normal, binary, ordered categorical, and censored variables. Available as part of the UCSF library's desktop SPSS license.
    - MI for ordered categorical variables creates probit-normal scores for both the observed and imputed values.
  - SPSS: Missing Values Analysis (MVA) optional module. Available in the UCSF library's desktop SPSS license.
- Other SAS:
  - IVEWare: http://www.isr.umich.edu/src/smp/ive/ (Stand-alone version also available)

Appendix: MI Example 2: Longitudinal Analysis (7)
[Same as ML Example 3 from Part 1]

- Duncan, Moskowitz, Neilands, Dilworth, Hecht & Johnson developed a Mindfulness-Based Stress Reduction (MBSR) intervention to reduce symptoms experienced and bother/distress from taking antiretroviral HIV medications. [*J. Pain Symptom Mange.* 43(2), 161-171, 2012.]
- *N*=76 people living with HIV who were actively taking ART and reported distress from ART-related side effects were randomly assigned to an MBSR program or a wait-list control (WLC) standard care condition.
- Study retention was adequate, with 86% (*N*=65) completing three-month follow-up assessments and 93% (*N*=71) completing assessments at the six-month follow-up.
- Primary Independent Variable: Intervention group (0 = control; 1 = MBSR)
- Dependent Variables: Sum of the number of symptoms reported (log-transformed) and average bother attributed to symptoms
- Analysis: Use SAS PROC MI to generate multiple imputations using the fully-conditional specification (FCS) method. Impute data sets separately by randomization group. Then fit doubly-multivariate repeated measures model using PROC MIXED to each imputed data set. Summarize the key results from those analyses.

Appendix: MI Example 2: Results (8)

- The MBSR group*time interaction is statistically significant (F[2,5633.3]=3.01, p=.0493).
- Tests of group differences for number of symptoms are significant at 3 months (difference=.35, p = .0312) and 6 months (difference=.39, p=.0178).
- Tests of group differences for symptom bother are significant at 3 months (difference=.47, p = .0455) but not at 6 months (difference=.45, p=.0613).
- We reach the same conclusion using MI that we did with the ML analysis presented in Part 1.
  - PMM MI less vulnerable to distribution assumption violations.
  - Missing independent variables can be addressed.
  - More complicated to implement.

Appendix: MI Example 2 Results Comparison (9)

Effect	Listwise	FIML	MI
Randstat*Time	F(2, 120)=2.26, p=.11	F(2, 132)=3.87, p=.02	F(3, 5633)=3.02, p=.049
Sx difference (SE)			
Baseline	.08 (.11), p=.48	.03 (.10), p=.76	.03 (.10), p=?*
3 Months	.31 (.17), p=.07	.33 (.16), p=.04	.35 (.16), p=.03
6 Months	.30 (.18), p=.10	.38 (.17), p=.02	.39 (.16), p=.02
Bother difference (SE)			
Baseline	-.05 (.16), p=.76	-.03 (.14), p=.85	-.03 (.14), p=?*
3 Months	.43 (.26), p=.10	.47 (.24), p=.048	.47 (.24), p=.046
6 Months	.35 (.27), p=.20	.43 (.24), p=.08	.45 (.24), p=.06
			* Not produced by SAS when there are no missing data