Multiple Imputation for Missing Data

or, listwise deletion is evil*

Michele Claibourn, Research Data Services

November 3, 2015

Research Data Services in the Library

Research Data Services: `data.library.virginia.edu` – consulting, workshops, and more!

- StatLab, data analysis and statistical consulting
- Data discovery, Locating and acquring data
- GIS, using geospatial data and technologies
- Data management, documenting, sharing, archiving research data
- Research software, installing site-licensed research software

The Problem

If archaeologists threw away every piece of evidence, every tablet, every piece of pottery that was incomplete, we would have entire cultures that disappeared from the historical record. We would no longer have the Epic of Gilgamesh, or any of the writings of Sappho. It is a ridiculous proposition because we can take all the partial sources, all the information in each fragment, and build them together to reconstruct much of the complete picture without any invention (Honaker and King 2010).

Assumptions about Missingness

Given

$$D = \begin{array}{cccccc} \textit{ID} & \textit{Female} & \textit{Age} & \textit{Income} & \textit{DPid} & \textit{BondMarket} \\ 1 & 1 & 42 & 85000 & 7 & 5 \\ 2 & 0 & 32 & . & 5 & . \\ 3 & 0 & 27 & . & 2 & 1 \\ 4 & 1 & 72 & 25000 & 3 & 2 \\ 5 & 1 & 63 & 56000 & 4 & . \end{array}$$

and

$$M = \begin{array}{cccccc} \textit{ID} & \textit{Female} & \textit{Age} & \textit{Income} & \textit{DPid} & \textit{BondMarket} \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{array}$$

Assumptions about Missingness

Missingness Assumptions

Assumption	Acronym	M predicted by
Missing completely at random	MCAR	–
Missing at random	MAR	D_{obs}
Not missing at random/ aka Nonignorable	NMAR/ NI	D_{obs} and D_{mis}

Listwise Deletion = Not Good

Listwise deletion is

- Inefficient under all assumptions
- Biased under MAR/NMAR

Multiple imputation is

- More efficient under all assumptions
- Unbiased under MCAR or MAR
- Still biased under NMAR

Listwise Deletion = Not Good

Listwise deletion is

- Inefficient under all assumptions
- Biased under MAR/NMAR

Multiple imputation is

- More efficient under all assumptions
- Unbiased under MCAR or MAR
- Still biased under NMAR

MI is normally better than, and generally not worse than, complete case analysis. How do you know data are MAR and not NMAR? You don't.

Traditional Aproaches

- Listwise deletion, aka complete case analysis: exclude cases with missing data for variables used in analysis
- Mean imputation: substituting the variable's mean for missing observations
- Conditional mean imputation (regression imputation): subsitute a predicted value from a multivariate model for the observation (typically using the same variables used in the main analysis)
- Hot deck imputation: substitute the answer from a randomly selected similar unit for the missing value.
- Predictive mean matching: combination of regression and hot deck imputation
- Dummy indicator for missing data: add an extra dummy variable coded 1 for all missing values and 0 otherwise.
- Last value carried forward: longitudinal/repeated measures designs – substitute last observed value for missing value.

Patterns of Missingness

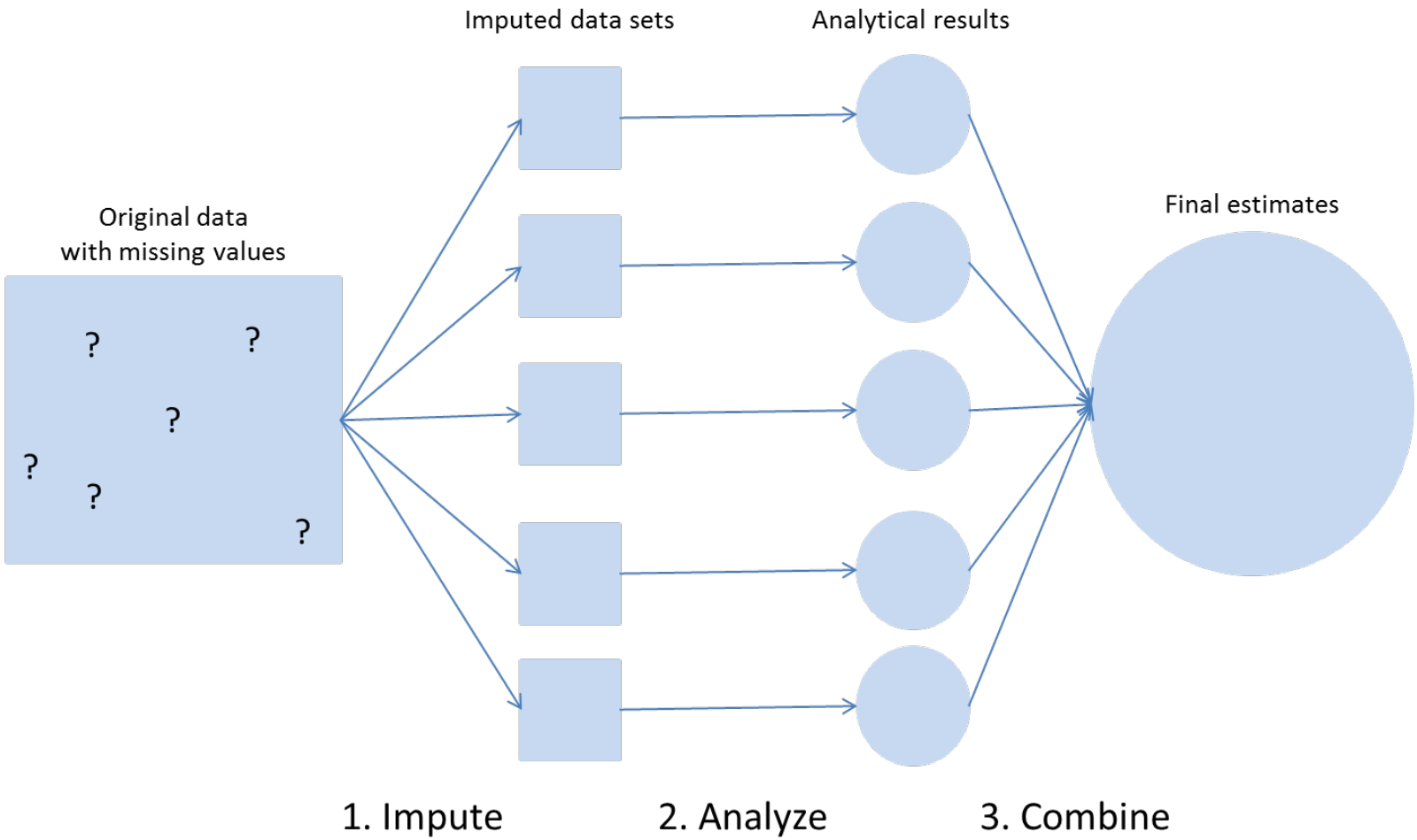
- Number of missing – which variables have a lot of missing values?
- Fraction of incomplete cases
- Patterns of missingness – are there groups of subjects with little information available?
- Is Missingness MCAR? Model missingness as function of other variables

Patterns of Missingness

- Number of missing – which variables have a lot of missing values?
- Fraction of incomplete cases
- Patterns of missingness – are there groups of subjects with little information available?
- Is Missingness MCAR? Model missingness as function of other variables

Can demonstrate data are not MCAR, but it's generally impossible to distinguish between MAR and NMAR. MI methods assume MAR. Happily, MI is often unbaised with NMAR data (Schafer & Graham 2002).

MI Conceptually



MI Models

MI models rely on iterative methods. Two main approaches:

1. Multivariate Normal (MVN)/Joint Modeling:
   - Stronger theoretical justification, procedures have known statistical properties
   - Assumes all variables have multivariate normal distribution
   - Lacks flexibility
   - Implemented in Stata (`mi impute/mi estimate`), R (`Amelia`), SAS (`proc mi/proc mianalyze`)
2. Imputation by Chained Equations (MICE)/Fully Conditional Specification:
   - Semi-parametric, sepcifies multivariate model by a series of conditional models
   - Flexible, easily applied to a variety of data types
   - Statistical properties are difficult to establish
   - Implemented in Stata (`mi impute chained/mi estimate`), R (`mi, mice`), SAS (`IVEWare`), SPSS (`impute method=fcs`)

MICE

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

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 with expected value plus draw from the posterior distribution fo the residuals
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 with expected value plus draw from the posterior distribution fo the residuals
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 with expected value plus draw from the posterior distribution of the residuals
5. Iterate Steps 2-4, cycling through the routine many times so that model converges (hopefully)
6. Repeat to get multiple imputations

Available Imputation Models

Imputation models in the `mice` package:

Method	Description	Scale type	Default
pmm	Predictive mean matching	numeric	Y
norm	Bayesian linear regression	numeric	
norm.nob	Linear regression	numeric	
mean	Mean imputation	numeric	
2l.norm	Two-level linear model	numeric	
logreg	Logistic regression	factor, 2	Y
polyreg	Polytomous regression	factor, > 2	Y
lda	Linear discriminant analysis	factor	
sample	Random draw from data	any	

Implementing MICE

1. Examine missingness in data
2. Impute data
   - Which imputation model?
   - What goes in the imputation model?
   - How many imputations?
3. Check the imputed values
   - Examine distributions
   - Check for convergence
4. Analyze multiply imputed datasets
5. Pool results

Checking Imputations

Imputation diagnostics are a matter of active research, but at a minimum, you should

1. Check the imputed values:
   - Does the distribution of imputed values seem reasonable given the distribution of observed values?
   - Frequency tables for binary, categorical, ordinal variables
   - Means, standard deviations, shape (histogram, density plot) for continuous variables
   - Check imputations separately, as well

Checking Imputations

Imputation diagnostics are a matter of active research, but at a minimum, you should

1. Check the imputed values:
   - Does the distribution of imputed values seem reasonable given the distribution of observed values?
   - Frequency tables for binary, categorical, ordinal variables
   - Means, standard deviations, shape (histogram, density plot) for continuous variables
   - Check imputations separately, as well
2. Check for convergence
   - The first iteration is often atypical, and because iterations are correlated, subsequent iterations may be atypical
   - Examine the trace line plot (value of estimate at each iteration)
   - On convergence, the lines should intermingle

Analyzing MI data sets, Pooling estimates

After impuation, peform the analyses you would have performed on the original data on each of the imputed data sets. Then combine the results.[1]

Call the estimated quantity of interest q, in each data set j. To combine estimates across m data sets (Rubin 1987):

- Overall point estimate:

$$\bar{q} = \frac{(1)}{m} \sum_{j=1}^{m} q_j$$

- Overall standard error: If SE_{q_j} is the estimated standard error of the quantity, q_j from data set j, and S_q^2 is the sample varaince across the m point estimates (e.g., $S_q^2 = \sum_{j=1}^{m}(q_j - \bar{q}^2/(m-1))$, then the variance of the MI point estimate is

$$(SE_q)^2 = \frac{1}{m} \sum_{j=1}^{m} (SE_{q_j})^2 + S_q^2(1 + 1/m)$$

References

Cited works

- Honaker, J. and G. King. 2010. "What to Do about Missing Values in Time-Series Cross-Section Data."
- Rubin, D. B. 1987. *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley.
- Schafer, J. L., J. W. Graham. 2002. "Missing Data: Our View of the State of the Art." *Psychological Methods* 7: 141-177.

Resources

- Azur, M. J., E. A. Stuart, C. Frangakis, P. J. Leaf. 2011. "Multiple Imputation by Chained Equations: What Is It and How Does It Work?" *International Journal of Methods in Psychiatric Research* 20: 40-49.
- Lee, K. J., and J. B. Carlin. 2010. "Multiple Imputation for Missing Data: Fully Conditional Specification Versus Multivariate Normal Imputation." *American Journal of Epidemiology* 171: 624-632.
- Little, R. J., D. Rubin. 2002. *Statistical Analysis with Missing Data.* New York: Wiley.
- Rubin, D. B. 1996. "Multiple Imputation after 18+ Years." *Journal of the American Statistical Association* 91: 473-489.
- White, I. R., P. Royston, and A. M. Wood. 2011. "Multiple Imputation using Chained Equations: Issues and Guidance for Practice." *Statistics in Medicine* 30: 377-399.