Multiple imputation using chained equations: Issues and guidance for practice (White, Royston, Wood, 2011)

Presentation by Adam Sadowski

University of Waterloo

December 11, 2017

Overview

1. Introduction
   - Missing Data
   - Multiple Imputation (MI)
   - Multiple Imputation by Chained Equations (MICE)
   - Example of Imputation for a Binary Variable
2. Imputation Model Specification
   - Variable Selection
   - Approaches to Non-Linear Terms
   - Model Building Issues
   - Example of Importance
3. Number of Imputations

Missing Data

We need to know how to handle missing data because it commonly occurs across all sorts of studies. Incorrect handling can bias our statistical inference.
Little and Rubin's framework [1]:

- Missing Completely at Random (MCAR): independent
- Missing at Random (MAR): independent of the unobserved conditional on the observed
- Missing Not at Random (MNAR): dependent on the unobserved regardless of the observed

Multiple Imputation

Imputation: replacing missing data with values
Multiple: m "complete" data sets imputed

Steps

1. Construct an imputation model: where among subjects with observed z, regress z on variables that have complete data
   - gives $\hat{\beta}$ and V
2. Repeate following m times
   - draw β^* from the posterior distribution (approx. by β^* as MVN($\hat{\beta}, V$))
   - imputations for z from its posterior predictive distribution and appropriate prob. distribution
   - proper imputation: incorporates estimation errors in model coefficients and prediction errors of imputed values
3. Now have m data sets ready for analysis: for parameters of interest, analyse each jth data set to get estimates $\hat{\theta}_j$ and its variance W_j

Multiple Imputation

Within-imputation variance is $W = \frac{1}{m} \sum_{j=1}^{m} W_j$

Between-imputation variance is $B = \frac{1}{m-1} \sum_{j=1}^{m} (\hat{\theta}_j - \hat{\theta})^2$

Final Estimate

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

Final Variance of Estimate

$$var(\hat{\theta}) = W + (1 + \frac{1}{m})B$$

correction factor accounts for simulation error in $\hat{\theta}$ [2]

Multiple Imputation by Chained Equations (MICE)

MICE is an approach to imputation for data sets with missing data among more than one variable.

Chained Equations: linked regression equations

Steps

1. Replace missing values by simple random sampling from respective variable's observed values
2. Construct imputation model: among subjects with observed x_1, regress x_1 on $x_2, ..., x_k$
   - draw x_1 from its posterior predictive distribution
3. Construct imputation model: among subjects with observed x_2, regress x_2 on $x_1, x_3, ..., x_k$
   - draw x_2 from its posterior predictive distribution
4. Repeat for all variables with missing data
5. Repeat steps 2-4 10 or 20 times
6. Repeat all the above m times to get m data sets

Example of Imputation for a Binary Variable

For a binary variable z with missing values, we construct a logistic regression imputation model, fit to subjects with the observed z.
We randomly draw β^* (β^* as MVN($\hat{\beta}, V$)). Then

$$\pi_i^* = [1 + exp(-\beta^* x_i)]^{-1}$$

for each missing observation z_i. We draw z_i as

$$z_i^* = \begin{cases} 1, & \text{if } u_i < \pi_i^*, \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

with u_i being a random draw from a uniform (0,1).
What if there is perfect prediction (where outcome is always 0 or 1)? This can result in infinite parameter estimates, but solvable by creating a few observations with very small weights ("augmentation").

Variable Selection

Say there is an outcome y that we have omitted as a predictor in the imputation model for z:

$$logit(Pr(z = 1 \mid x; \beta)) = \beta_0 + \beta_1 x$$

Then, subjects who have imputed z values and observed y will have uncorrelated z and y. But z and y may be correlated. This biases the coefficient of z in the analysis model toward 0.

Guiding Rule 1

Imputation models should include all variables used in all analysis models.

Variable Selection

- Recall MAR: conditional on observed data, prob. of data missing does not depend on unobserved data.
- MAR is more plausible if we include explanatory variables predictive of the missing variable (bias is then reduced).
- Generally having more variables in the imputation model does not cause bias and other resulting deficiencies are not important for practice [3].

Guiding Rule 2

Imputation models should include all predictors of z's value or missingness.

Approaches to Non-Linear Terms

Assume analysis involves interaction vx and outcome y. There are multiple approaches to imputing for vx.

- Separate: separate imputation of v within groups of categorical x
- Passive: say v is imputed from regression on z, y, x; x is imputed from regression on v, z, y; missing values of vx are then replaced by the $(v_{imputed} x_{imputed})$ values
- Improved Passive: v is imputed from regression on z, y, x, yx
- Just Another Variable: treat vx as just another variable
  - regress vx on v, z, y, x

Model Issues

Problem: We assumed analysis involves an interaction.

- In practice, when constructing imputation models we might not know which interaction terms will be needed in analysis.
- Including every possible interaction term in our imputation models can cause converge issues and "defeat the software".

Solution:

1. Start with a simple imputation model including predictors (Rule 2) and interaction terms of interest.
2. Use imputed data to check if the analysis model should have interaction terms.
3. Re-do imputations with these terms using one of the approaches just discussed (Rule 1).
4. Use the newly imputed data set for analysis or repeat Steps 2 and 3.

Model Issues

Problem: Resulting imputation models from the above solution's process can again be too complex for the software and model convergence.
Solution: Simplify the imputation model such that insubstantial bias is incurred.
Impasse: The authors still recommend following Rule 1: including all variables used in analysis. This contradicts their solution. They admit there is not yet a universal solution or rule of thumb for model form.

Example of Importance

- QRISK: Massive study with data on over a million people; evaluating risk factors for cardiovascular disease
- One predictor: cholesterol ratio of serum to HDL levels
- Serum and HDL levels were missing in over 60 percent of people.
- After MI, Hazard Ratio was 1 for 1 unit change in the ratio.
  - But cholesterol is known to predict cardiovascular disease.
- Authors omitted an outcome variable (event indicator) in the imputation model.
  - Event indicator was more important than log(survival time) since most data was censored.
  - Furthermore, the ratio was imputed passively (biasing its coefficient in the analysis model toward 0).
- After fixes: Hazard Ratio 1.17

Number of Imputations

- Authors recommend m imputed data sets in relation to the Monte Carlo (MC) error so that results are reproducible.
  - Definition: SD of results across repeated runs of MI

$$SD = \sqrt{B/m}$$

- Authors define adequate reproducibility as the following MC errors:
  1. of $\hat{\beta}$, 10 per cent of its SE
  2. of test-statistic, 0.1
  3. of P-value $= 0.05$, 0.01
- achieved using $m \geq 100 * FMI/m$, where FMI is the fraction of missing information

Result: Do as many imputations as the percentage of incomplete cases.

References

Paper

White IR, Royston P, Wood AM. Multiple imputation using chained equations: issues and guidance for practice. Statistics in medicine. 2011 Feb 20;30(4):377-99.

References:

1. Little RJA, Rubin DB. Statistical Analysis with Missing Data (2nd edn). Wiley: Hoboken, NJ, 2002.

2. Schafer JL, Olsen MK. Multiple imputation for multivariate missing-data problems: A data analyst's perspective. Multivariate behavioral research. 1998 Oct 1;33(4):545-71.

3. Schafer JL. Analysis of Incomplete Multivariate Data. Chapman and Hall: London, 1997.