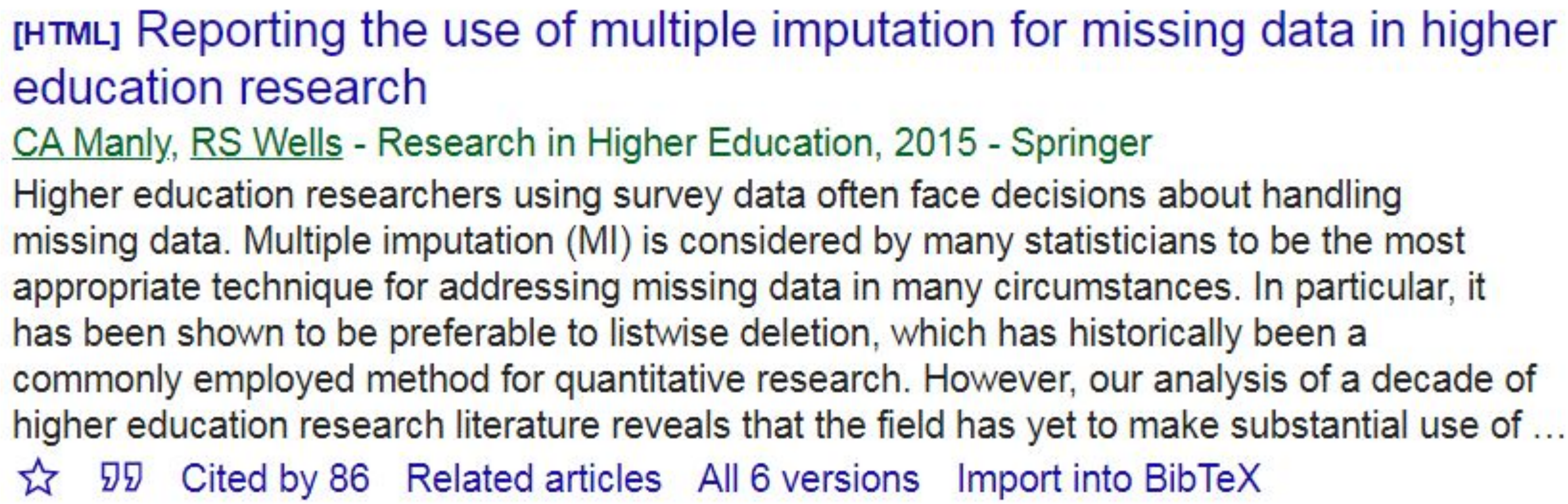
Missing Data:
Further DIscussion & Reporting guideline

ehsan.karim@ubc.ca

Oct 14, 2020

SPPH 504/007

Ref



[HTML] Reporting the use of multiple imputation for missing data in higher education research

CA Manly, RS Wells - Research in Higher Education, 2015 - Springer

Higher education researchers using survey data often face decisions about handling missing data. Multiple imputation (MI) is considered by many statisticians to be the most appropriate technique for addressing missing data in many circumstances. In particular, it has been shown to be preferable to listwise deletion, which has historically been a commonly employed method for quantitative research. However, our analysis of a decade of higher education research literature reveals that the field has yet to make substantial use of …

Content

1. Assumptions
2. Why MI is better?
3. Two MI procedures
4. How does MICE work?
5. How to estimate using MICE?
6. How to report?
7. What to do in complex scenarios

Assumptions

- MCAR
  - data are missing due to random reasons
  - Reasons are
    - neither due to observed nor
    - due to unobserved variables
- MAR
  - data are missing due to observed/measured/known reasons
  - can control for those observed variables to obtain unbiased estimates
- MNAR
  - Data is missing due to
    - either unobserved variables or
    - Missingness is due to the dependent variable itself

Multiple imputation (MI)

- Superior to ad hoc methods (complete case)
  - MI could be still reasonably used; even with <5 % data is missing
  - Why? Protecting against loss of power
- Superior to single imputation methods
  - a tool for representing missing-data uncertainty properly
  - Single imputation methods suffer from unrealistically lower SE
- Works under the MAR assumption.
- How it works?
  - Uses correlated nature of the data to improve imputation.
  - Focuses more on distributions of a variable rather than just the missing value.

MI types

1. Multiple imputation by chained equations (MICE)
    a. sequential regression imputation,
    b. fully conditional specification (FCS)
2. Multivariate normal imputation (MVN) for continuous variables only
    a. data augmentation (DA)

Assumption of MVN (joint multivariate normality) may not be realistic in some scenarios.

How does MICE works?

1. The initial iteration provides starting values for all missing data
2. Number of imputation (m) selected.
3. Imputation model selected, appropriate for the variable with missing.
4. Each variable is then imputed one by one sequentially from the distribution of the variables (not just one predicted value from a given model).
5. This induces some realistic randomness in the process.

How to estimate from MICE

1. Produce m number of imputed data
2. Apply procedure (say, logistic regression) on all imputed data and record m outputs (e.g., beta/SE)
3. Pool estimates to obtain a average estimate using Rubin's rule.
4. Due to induced randomness, SE represents some uncertainty in the imputation process.

How to report a MICE analysis? _ I

- Rates of missingness in the variables
    - % of missing data in the complete case
        - "90% observations ... in our cohort had no missing values, for any of the variables included in these models"
    - %s of missing in all variables
        - "ranged from nearly 0 for some demographic variables to as high as 12 %"
        - "Missing data for the primary outcome, primary exposure, and all covariates were <1%, 16%, and <10%, respectively"
- Describe why (the reasons) the data are missing
    - Non-response? Stigma?
    - Randomly refused to answer, missed filling out by mistake?
    - Fear of government getting information about immigration / income?

How to report a MICE analysis? _ II

- Mention under what assumption you are operating
    - Is MAR a reasonable/plausible assumption in your data?
        - "... under the assumption that missing values are MAR"
- Report m
    - "To deal with missing data, we performed 20 imputations of the dataset ..., without individuals who died, using the 'multiple imputations with chained equations' (MICE) method"
    - "This procedure was repeated 20 times, and final results for parameters and standard errors accounted for the variation of the estimates across the 20 replications"
- Report software/version
    - "All imputations and analyses were conducted in R version 3.2.3"
    - "Stata 13's 'mi impute chained' command"

How to report a MICE analysis? _ III

- Imputation model
    - Report covariates in the imputation model
        - "The missing values of variables are imputed based on the regression model that uses observed values of all the variables"
    - Adding outcome is encouraged in general
    - General advice is to add covariates that are highly correlated with the variable that has missing (so that MAR assumption is plausible)
    - If auxiliary variables are added, report them
- Imputation model-specifications
    - Interaction, polynomial between continuous variables
    - Binary / categorical variables may be imputed separately before imputation continuous variables

How to report a MICE analysis? _ IV

- Report pooling method
  - Rubin's rule is the most popular one. There are others as well. Be specific which one you used.
    - "Analyses run on each dataset were pooled according to Rubin's (1987) rules"
- Report if you see any discrepancies in the following
  - Compare observed and imputed Values (histogram/tabulating)
    - "Imputed values compare reasonably to observed values."
  - Check whether different imputations are producing very different results or not
- Is the reporting enough for reproducing?
  - Note: RStudio is not a statistical package; it is an IDE

What to do? - 1: Outcome is missing

- Multiple Imputation, then Deletion (MID) is the most commonly cited approach
  a. Use the imputed outcome to impute other variables
  b. When imputation complete, delete the imputed outcomes
  c. MID operates under the assumption that imputed outcomes have nothing to add in the regression; but just adds noise to the analysis
     - ■ To keep those imputed outcomes, one may need to justify whether they really add anything to the analysis
       - Same is true for imputed exposure variable. Many suggests that both may just add random error/noise.
     - ■ Subject to some controversy in recent years
     - ■ Other suggest imputing outcome as usual (not imputing assumes no tx effect).

What to do? - II: Imputation model not perfect

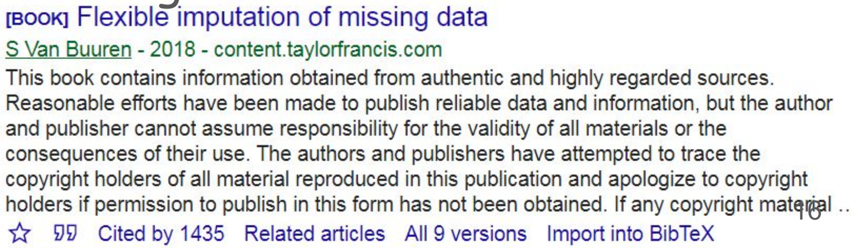
- In observational setting it is hard to come-up with an imputation model that is completely “correct”.
- Even in that scenario, it is suggested to do multiple imputation. If the data has some correlated covariates (usually we have that), under MAR, it will be better than doing complete case analysis.
- But have to be honest and realistic about the confidence in the results.

What to do? - III: How much missing is too much?

- No absolute answer available.
- It depends on how many useful covariates you have, and how likely is MAR.
- Even with more than 40% missing, some studies produces reasonable results.

What to do? - IV: MI for Longitudinal data

- A topic of recent interest
- Search for “Multiple Imputation for Multilevel Data”
- Some R packages available
  - mitml
  - Within mice look for methods starting with 2l.*
    - 2l.norm
    - 2l.pan
  - Within miceadds look for methods starting with 2l.*
    - 2l.pmm
  - CALIBERrfimpute



[BOOK] Flexible imputation of missing data

S Van Buuren - 2018 - content.taylorfrancis.com

This book contains information obtained from authentic and highly regarded sources. Reasonable efforts have been made to publish reliable data and information, but the author and publisher cannot assume responsibility for the validity of all materials or the consequences of their use. The authors and publishers have attempted to trace the copyright holders of all material reproduced in this publication and apologize to copyright holders if permission to publish in this form has not been obtained. If any copyright material ...

☆ 99 Cited by 1435 Related articles All 9 versions Import into BibTeX

Thanks!

ehsan.karim@ubc.ca

www.ehsankarim.com