Multiple Imputations in Political Science

Sebastian Suarez

Abbreviations

MAR:
Missing at random

MCAR:
Missing completely at random

MCMC:
Markov chain Monte Carlo

MNAR:
Missing not at random

Reason For this Presentation

- Dataset From Intro to CP Paper
  - Merged two datasets
    - A lot of missing Values
  - Tried to Improve it using multiple imputations
    - Still insignificant results

- This is an introduction to Multiple Imputations

What are Multiple Imputations?

- Multiple imputation fills in missing values by generating plausible numbers derived from distributions of and relationships among observed variables in the data set (Li, et al., 2015).
- Multiple imputation provides accurate estimates of quantities or associations of interest, such as treatment effects in randomized trials, sample means of specific variables, correlations between 2 variables, as well as the related variances. In doing so, it reduces the chance of false-positive or false-negative conclusions (Li, et al., 2015).
- Multiple imputation is a simulation-based statistical. (Jakobsen, et al., 2017).

Multiple Imputation Steps (Jakobsen, et al., 2017)

- Imputation step.
  - An 'imputation' represents one set of plausible values for missing data
    - multiple imputation represents multiple sets of plausible values. When using multiple imputation, missing values are identified and are replaced by a random sample of plausible values imputations (completed datasets). Multiple completed datasets are generated via some chosen imputation model.

Multiple Imputation Steps (Jakobsen, et al., 2017)

- Completed-data analysis (estimation) step. The desired analysis is performed separately for each dataset that is generated during the imputation step. Hereby, for example, 50 analysis results are constructed.

Multiple Imputation Steps (Jakobsen, et al., 2017)

- Pooling step. The results obtained from each completed-data analyses are combined into a single multiple-imputation result. There is no need to conduct a weighted meta-analysis as all say 50 analysis results are considered to have the same statistical weight.
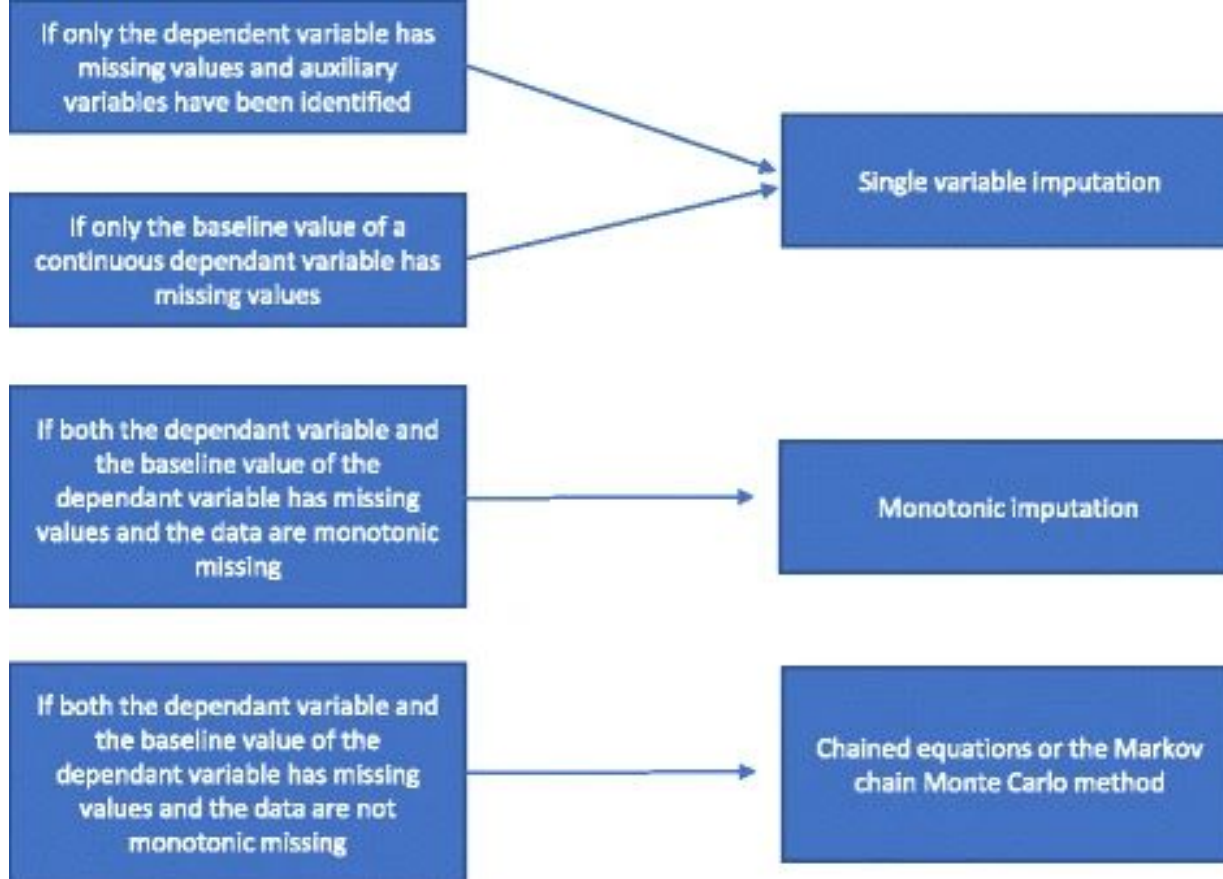
Why Use Multiple Imputations?

- **Better Alternative to listwise deletion**
  - **It ignores the valuable information carried by partially observed units, and it can introduce bias in regression coefficient estimates (Arel-Bundock, et al., 2018).**

Types of Multiple Imputations

- single value regression analysis
- 2) monotonic imputation
- 3) chained equations or the Markov chain Monte Carlo (MCMC) method

UNIVERSITY OF MARYLAND   FEARLESSLY FORWARD

Jakobsen, J.C., Gluud, C., Wetterslev, J. *et al.* When and how should multiple imputation be used for handling missing data in randomised clinical trials – a practical guide with flowcharts. *BMC Med Res Methodol* 17, 162 (2017). https://doi.org/10.1186/s12874-017-0442-1

UNIVERSITY OF MARYLAND

FEARLESSLY FORWARD

10

Single Variable Imputation

A single variable regression analysis includes a dependent variable and the stratification variables used in the randomisation.

- The stratification variables often that which are correlated with the outcome. When using a continuous dependent variable, a baseline value of the dependent variable may also be included.
- if only the dependent variable has missing values and auxiliary variables are *not* identified, a complete case analysis should be performed and no specific methods ought to be used to handle the missing data
- If auxiliary variables have been identified, a single variable imputation may be performed. If there are significant missingness on the baseline variable of a continuous variable, a complete case analysis may provide biased results

Multiple Imputations are Not a Panacea

- **The cases that can use it are limited**
  - There cannot be violations to the multiple imputations assumptions
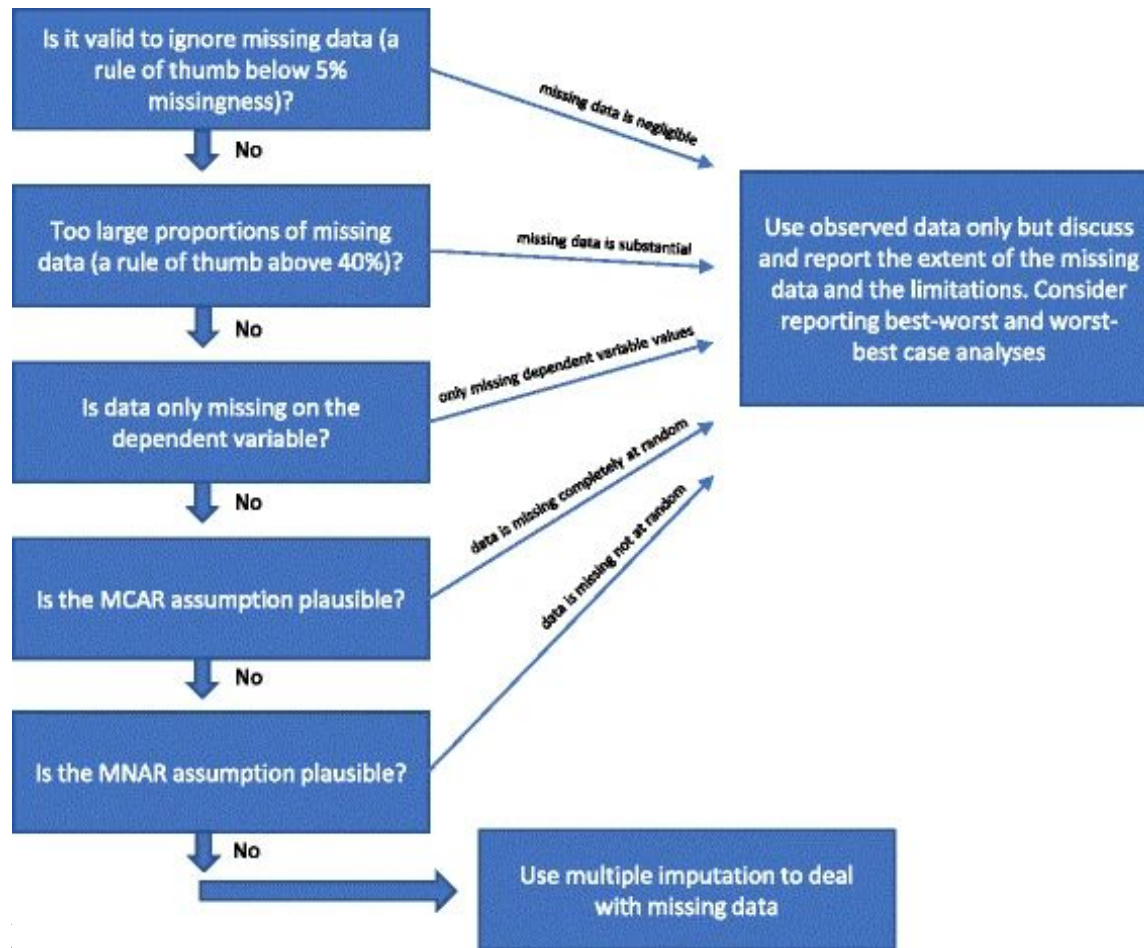
Jakobsen, J.C., Gluud, C., Wetterslev, J. *et al.* When and how should multiple imputation be used for handling missing data in randomised clinical trials – a practical guide with flowcharts. *BMC Med Res Methodol* 17, 162 (2017). https://doi.org/10.1186/s12874-017-0442-1

Problems with Multiple Imputations

- The main way MI can go wrong is if the imputations are poorly constructed.
  - A very useful technique is to compare the distribution of the imputed values with the distribution of the observed values.
  - If we spot a implausible values in the imputation, something is wrong
  - Sometimes the imputed data leans more towards a value because of the data used for the imputations
    - We need to explain why
    - Different assessment for categorical variables

Problems with Multiple Imputations

- Data is not randomly missing
  - It is difficult to use an appropriate technique
  - Our best suggestion is to consider doing a sensitivity analysis in which imputed values are varied away from the values imputed under missing at random. Examples of how to do this are given in White et al (White, et al., 2022)
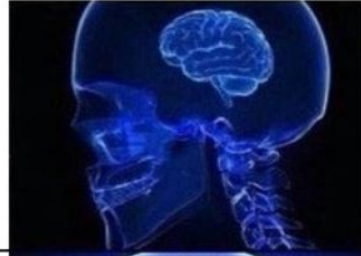
How to Avoid Problems

- Compare the Imputation data to complete cases
  - Not expected to be very different
    - If that happens, we need to explain
- Five imputed datasets have traditionally been suggested to be sufficient on theoretical grounds, but 50 datasets (or more) seem preferable to reduce sampling variability from the imputation process

Take Away

Bibliography and Resources

Li P, Stuart EA, Allison DB. Multiple Imputation: A Flexible Tool for Handling Missing Data. JAMA. 2015 Nov 10;314(18):1966-7. doi: 10.1001/jama.2015.15281. PMID: 26547468; PMCID: PMC4638176.

Jakobsen, J.C., Gluud, C., Wetterslev, J. *et al*. When and how should multiple imputation be used for handling missing data in randomised clinical trials – a practical guide with flowcharts. *BMC Med Res Methodol* 17, 162 (2017). https://doi.org/10.1186/s12874-017-0442-1

Missing Data Part II: Multiple Imputation & Maximum Likelihood Richard Williams, University of Notre Dame, https://www3.nd.edu/~rwilliam/

Lall, R. (2016). How multiple imputation makes a difference. *Political Analysis*, *24*(4), 414-433.

Arel-Bundock, V., & Pelc, K. J. (2018). When can multiple imputation improve regression estimates?. *Political Analysis*, *26*(2), 240-245.

https://stats.oarc.ucla.edu/r/faq/how-do-i-perform-multiple-imputation-using-predictive-mean-matching-in-r/

UNIVERSITY OF MARYLAND   FEARLESSLY FORWARD

Presentation Title