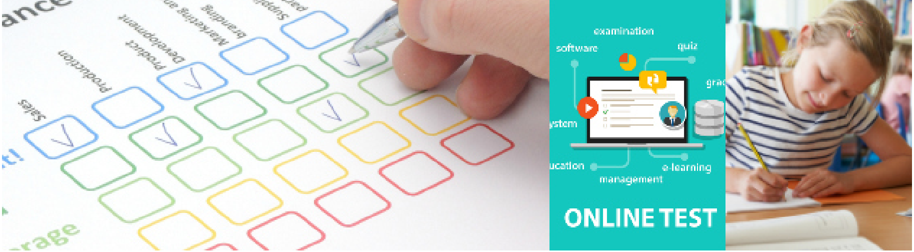


UiO : Centre for Educational Measurement
Det utdanningsvitenskapelige fakultet

Missing Data Treatment

A hand-on illustration using R package mice

Tony C. A. Tan

28 Feburary 2022

Structure

1 **Background**
- Summary
- Data Missing Mechanism

2 **The mice Package**
- mice Workflow
- mice Methods

3 **References**

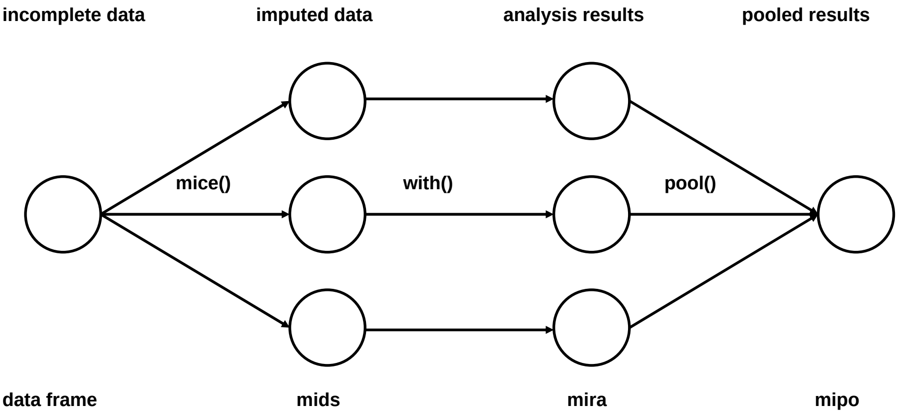
Summary

- Complete-case analyses:
  - ✗ Wasteful
  - ✗ Biased
- Two approaches:
  - ① Joint modelling (JM, Schafer, 1997)
  - ② Fully conditional specification (FCS)
  - ☞ FCS aka multivariate imputation by chained equations (MICE, van Buuren & Groothuis-Oudshoorn, 2011)
- Existing R packages:
  - ➤ Amelia, Hmisc, jomo, mi, mice, norm, norm2, pan
  - ✍ See Table 5.1, Kleinke et al. (2020) (p. 134) for popularity contest across various MI packages
  - ✍ See Table 6, Grund et al. (2018) (pp. 134–135) for missing data treatment for multilevel models

Data Missing Mechanism (Rubin, 1976)

- Missing completely at random (**MCAR**)
  - ✎ missingness of variables is independent of the variables considered in the study
  - ✓ no treatment required, complete-case analyses valid and unbiased
- Missing at random (**MAR**)
  - ✎ missingness depends exclusively on observable variables
  - ✓ the assumption behind most MI procedures, including mice
- Missing not at random (**MNAR**)
  - ✎ missingness depends on unobservable but important variables of interest in the study
  - ✓ exact treatment rather complicated (Rose, 2013)
  - ✓ in practice: introduce lots of covariates and hope MNAR $\approxeq$ MAR
- ● Ignorable = { MCAR, MAR }; Nonignorable = { MNAR }

mice Workflow (van Buuren & Groothuis-Oudshoorn, 2011)



incomplete data — imputed data — analysis results — pooled results

mice() — with() — pool()

data frame — mids — mira — mipo

mice Methods

Method	Description	Scale type	Default
`pmm`	Predictive mean matching	numeric	Y
`norm`	Bayesian linear regression	numeric	
`norm.nob`	Linear regression, non-Bayesian	numeric	
`mean`	Unconditional mean imputation	numeric	
`2L.norm`	Two-level linear model	numeric	
`logreg`	Logistic regression	factor, 2 levels	Y
`polyreg`	Multinomial logit model	factor, >2 levels	Y
`polr`	Ordered logit model	ordered, >2 levels	Y
`lda`	Linear discriminant analysis	factor	
`sample`	Random sample from the observed data	any	

References

van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, *45*(3), 1–67. https://doi.org/10.18637/jss.v045.i03

Grund, S., Lüdtke, O., & Robitzsch, A. (2018). Multiple imputation of missing data for multilevel models: Simulations and recommendations. *Organizational Research Methods*, *21*(1), 111–149. https://doi.org/10.1177/1094428117703686

Kleinke, K., Reinecke, J., Salfrán, D., & Spiess, M. (2020). *Applied multiple imputation: Advantages, pitfalls, new developments and applications in R*. Springer. https://doi.org/10.1007/978-3-030-38164-6

Rose, N. (2013). *Item nonresponses in educational and psychological measurement* [PhD Thesis, Friedrich-Schiller-Universität Jena]. Open Access Thesis and Dissertations. https://www.db-thueringen.de/servlets/MCRFileNodeServlet/dbt_derivate_00027809/Diss/NormanRose.pdf

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*(3), 581–592. https://doi.org/10.1093/biomet/63.3.581

Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. Chapman & Hall; CRC.

UiO : Centre for Educational Measurement
Det utdanningsvitenskapelige fakultet

Tony C. A. Tan

Missing Data Treatment

A hand-on illustration using R package mice

UNIVERSITAS OSLOENSIS · MDCCCXI ·