**Multiple imputation using MI – A case study**

06.10.2022

What is multiple imputation?

* Multiple imputation (MI) is a simulation-based approach for analyzing incomplete data.
* MI replaces missing values with multiple sets of simulated values to complete the data, applies standard analyses to each completed dataset, and adjusts the obtained parameter estimates for missing-data uncertainty (Marchenko, 2009).

Assumptions of multiple imputation

* mi impute assumes that missing data are missing at random; that is, missing values do not carry any extra information about why they are missing than what is already available in the observed data.
* mi impute creates imputations by simulating from a (approximate) Bayesian posterior predictive distribution of the missing data, following Rubin’s recommendation (Marchenko, 2009).

Why use multiple imputation instead of complete case analyses?

* Complete case (CC) analysis is commonly used, but it reduces sample size and is perceived to lead to reduced statistical efficiency of estimates while increasing the potential for bias. As multiple imputation (MI) methods preserve sample size, they are generally viewed as the preferred analytical approach (Mukaka et al., 2016)

A case study – using the oasis data

Stata syntax are in yellow

Steps:

1. Check missing data

Local vars a b c d

Egen nmiss=rowmiss(`vars’)

Sum `vars’

1. Step up mi data

mi supports 4 styles (formats) for storing MI data:

* flongsep — full long and separate — imputed data are in separate files, one per imputation;
* flong — full long — original and imputed data are in one file, imputations are saved as extra observations;
* mlong — marginal long — original and imputed data are in one file, only observations containing imputed values are saved as extra observations. mlong is a memory-efficient version of flong;
* wide — wide — original and imputed data are in one file, imputations are saved as extra variables.

Some tasks are easier in one style than another. You can switch from one style to another during your mi session by using mi convert (Marchenko, 2009).

mi set mlong

1. register data

//register all missing variables that need imputation (those variables that have missing values)

mi register imputed race\_cat education medicaid year

//register regular variables (those variables that have complete cases)

mi register regular age female livearrange srh\_fp

1. impute data

//use linear regression for continuous variables

//use logistic regression for binary variables

//use multinomial logit model for categorical variables

// mi impute chained fills in missing values in multiple variables iteratively by using chained equations, a sequence of univariate imputation methods. It accommodates arbitrary missing-value patterns.

mi impute chained (regress) year (logit) education medicaid (mlogit) race\_cat= age female livearrange srh\_fp, add(10) rseed(1234)

1. mi estimate

mi estimate: logit srh\_fp age i.female i.livearrange i.race\_cat i.education i.medicaid year

References:

Ginkel et al. (2018). Rebutting Existing Misconceptions About Multiple Imputation as a Method for Handling Missing Data. Statistical Development and Applications. 297-308. <https://doi.org/10.1080/00223891.2018.1530680>

Marchenko, Y. (2009) <http://repec.org/usug2009/uk09_marchenko.pdf>

Mukaka et al. (2016) <https://trialsjournal.biomedcentral.com/articles/10.1186/s13063-016-1473-3#:~:text=Complete%20case%20(CC)%20analysis%20is,as%20the%20preferred%20analytical%20approach>.

Rubin. D. (1996). Multiple Imputation after 18+ Years. *Journal of the American Statistical Association, 91*, 434, 473-489. <http://www.yaroslavvb.com/papers/rubin-imputation.pdf>

Stata manuals. <https://www.stata.com/manuals/mi.pdf>

Stata manuals. <https://www.stata.com/manuals13/mimiimpute.pdf>