


Missing Data and Multiple Imputations

Yossi Levy

16.12.2020

Missing Completely At Random


- There is no relationship between missingness and either observed or unobserved data.
- Examples:
  - The patient decided to move to Hawaii
  - "Missing by design", e.g. rotating panel study
  - Study is terminated at a common scheduled date before all subjects have complete follow-up.

MCAR essentials


- The observed data can be thought of as a random sample of the complete data
- In particular, complete cases can be regarded as a random sample from the target population.
- All methods for analysis that yield valid inferences in the absence of missing data will also yield valid inferences when the analysis is based on all available data
- Therefore, “Complete cases” analysis is valid, yet inefficient
- It may be possible to check the validity MCAR under certain assumptions

Missing Not At Random


- The missingness depends on both observed and unobserved data

- Examples:
  - Individuals who are heavier are less likely to report their weight
  - Device sensitivity: if it can measure only values that are above S, anything below that is missing
  - In RCT, subjects from the control group are more likely to withdraw from study

MNAR essentials

- The missing data mechanism cannot be ignored when the goal is to make inferences about the distribution of the complete data
- Any valid inferential method under MNAR requires specification of a model for the missing data mechanism

Missing At Random

- The missingness depends on observed data but not on unobserved data

- Examples
  - Those from a higher socioeconomic status may be less willing to provide salary information (but we know their SES status)
  - A study protocol requires that a subject be removed from the study as soon as the value of an outcome variable falls outside of a certain range of values

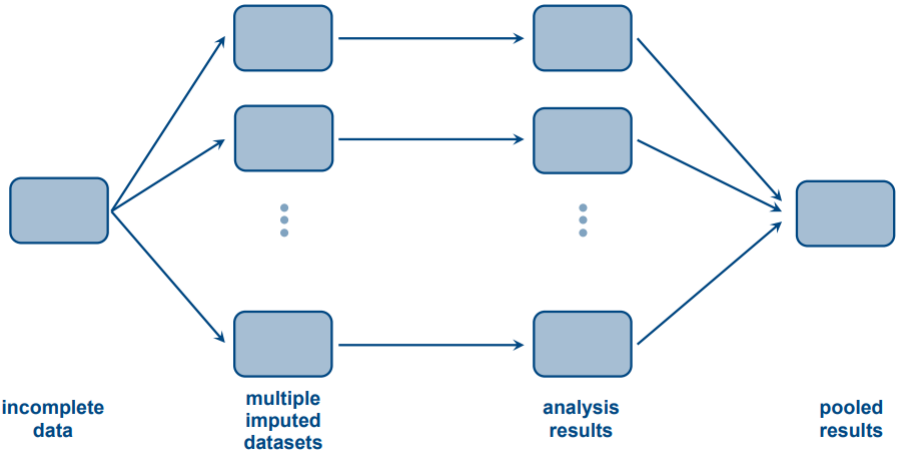
MAR essentials

- Complete cases are a biased sample from the target population
- Consequently, an analysis restricted to the “completers” is not valid
- However, the conditional distribution of the missing values is the same as the distribution of the completers data and the population data
- Therefore, the missing values can be validly “predicted” or “extrapolated” using the observed data

Checking the missingness mechanism

- The assumption of MAR can be tested against the alternative hypothesis of MCAR, under the assumptions that the data is not MNAR
- The assumption of MAR can be tested against the alternative hypothesis of MNAR only when a specific MNAR model is assumed

The Multiple Imputations principle



Combining MI analyses

$$\hat{\beta}^{MI} = \frac{1}{m}\sum_{j=1}^{m}\hat{\beta}^{(j)} \qquad V^{MI} = \bar{V} + (1+\frac{1}{m})\cdot B$$

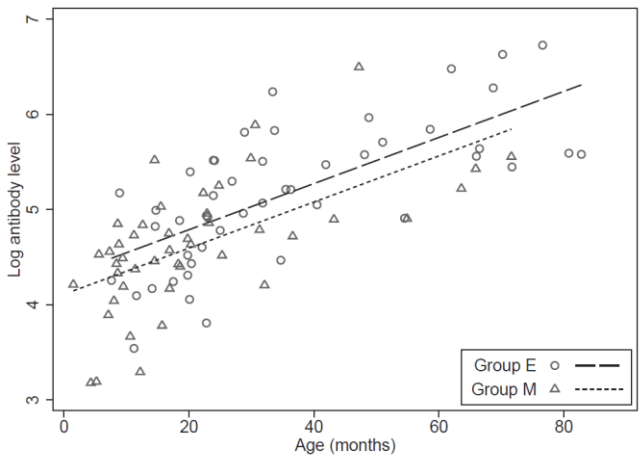
where $\bar{V} = \frac{1}{m}\sum_{j=1}^{m} Var(\hat{\beta}^{(j)})$ $\qquad B = \frac{1}{m-1}\sum_{j=1}^{m}(\hat{\beta}^{(j)} - \hat{\beta}^{MI})^2$

V^{MI} components: within imputations and between imputations

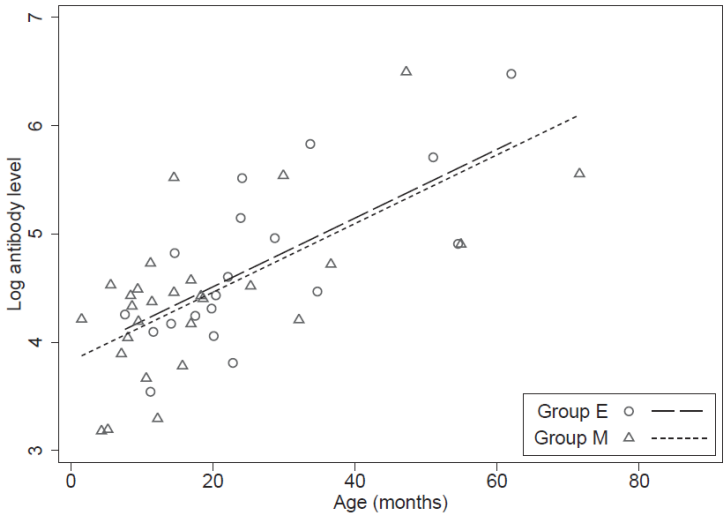
Inference is based on the assumption that $\frac{\hat{\beta}^{MI} - \beta}{\sqrt{V^{MI}}}$ follows either a standard normal or a t distribution

Illustrative example

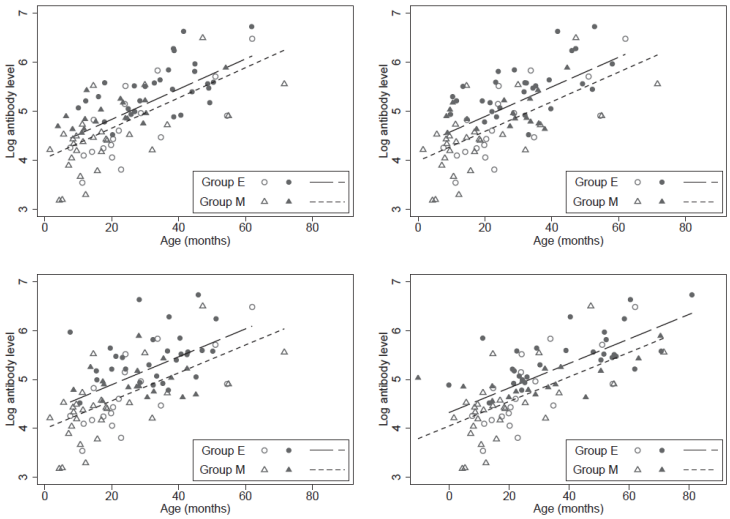
	Unadjusted	ANCOVA
Group diff. (s.e.)	-0.542 (0.145)	-0.194 (0.117)
Age effect (s.e)	- -	0.0242 (0.0029)



What happens when 50% of age values are missing?



Multiple imputations: m=4



MI results

	Unadjusted	ANCOVA
Group diff.	-0.542	-0.194
(s.e.)	(0.145)	(0.117)
Age effect	-	0.0242
(s.e)	-	(0.0029)

	Complete cases	Imp. 1	Imp. 2	Imp. 3	Imp. 43	MI (m=100)
Group diff.	-0.051	-0.179	-0.303	-0.328	-0.275	-0.211
(s.e.)	(0.167)	(0.121)	(0.117)	(0.123)	(0.113)	(0.147)
Age effect	0.0318	0.0308	0.0303	0.0285	0.0250	0.0293
(s.e.)	(0.0051)	(0.0039)	(0.0039)	(0.0042)	(0.0029)	(0.0045)

MI basic principles


- The MI paradigm is Bayesian in its nature
- Observed data depend on parameter(s) β
- Distribution of β is estimated from the observed data
- Missing data is sampled/simulated by using the β estimate

Limitations/considerations


- Bayesian analysis is predicated on the assumption that the proposed models are correct.
- Therefore, model checking is an essential feature of sound statistical analysis
- Even Bayesians should avoid using statistical methods that can be expected to perform poorly when considered within a framework of repeated sampling

MI methods

- Regression-based imputation – useful when only one variable needs imputation
- Imputation under a joint model for the observed data and the missing data – various Monte Carlo methods
- Imputation using fully conditional specification - MICE
  - Intuition: apply the univariate regression approach to each of the variables that has missing values
  - Application: a series of regression models are run whereby each variable with missing data is modeled conditional upon the other variables in the data

MICE process

- Impute all missing values using simple imputations
- Cycle across variables with missing values
  - Select a variable to be imputed and reset the imputed values to missing
  - Use an appropriate regression model to impute the missing value based on all other data
- Re-iterate until convergence

Example

Variables with missing data: age, sex, number of lesions at baseline (NoL)

1. Single impute all missing data
   2. Reset imputed age values to missing and re-impute using ***linear regression*** model and imputed values of sex and NoL
   3. Reset imputed sex values to missing and re-impute using ***logistic regression*** and imputed values of age and NoL
   4. Reset imputed NoL values to missing and re-impute using ***Poisson regression*** and imputed values of age and sex
5. Reiterate steps 2-4 until convergence, to get an imputed data set

Repeat the whole process to get more imputed data sets