Missing Value Imputation

Problem - incomplete data

Two cases:

- Conscious missing (willingly)
- Unconscious missing (unwillingly)

Missing-data mechanisms

1. Missingness completely at random
2. Missingness at random
3. Missingness that depends on unobserved predictors
4. Missingness that depends on the missing value itself.

Example - case study

Study population earnings.

Input variables:

- sex
- race
- education
- age
- earnings
- police arrest

	sex	race	educ_r	r_age	earnings	police
[91,]	1	3	3	31	NA	0
[92,]	2	1	2	37	135.00	1
[93,]	2	3	2	40	NA	1
[94,]	1	1	3	42	3.00	1
[95,]	1	3	1	24	0.00	NA

Missingness completely at random

Probability of missingness is the same for all units

Example: if each survey respondent randomly decide whether of answer a variable or not (like rolling a dice)

Easiest case, data can we approximated with the mean of existing data.

Missingness at random

Probability of missingness depends on the other OBSERVED variables

Example: white people are more probable to not fill the “earnings” than black people

Missing data can be modeled according to existing data based on a model that account all the observed variables

Missingness that depends on unobserved predictors

Probability of missingness depends on the other UNOBSERVED variables

Example: suppose that people that have children are less likely to reveal their earnings, but having children is not observed.

Missing data is not at random, therefore it must be explicitly modeled, or the reliability of the data will be affected.

Missingness that depends on the missing value itself

Probability of missingness depends on the missing value itself, this is also called censoring.

Example: people with earning greater than 100.000 are less likely to reveal their earnings.

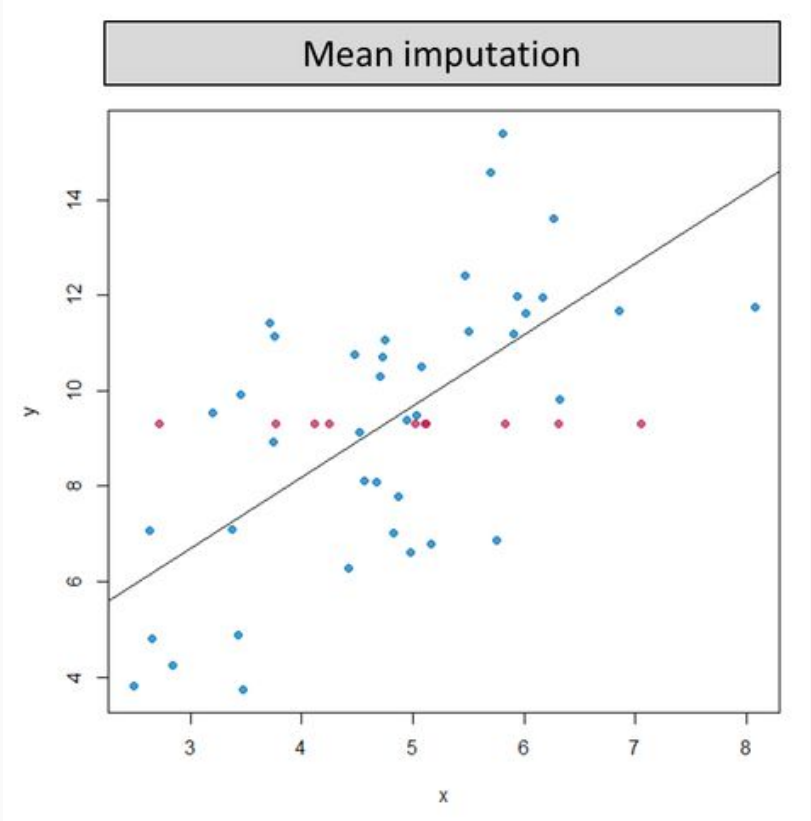
In order to reliably model the missing data, more predictors can be included so the case can fit in the missingness at random case.

Single imputation

1. Mean imputation
2. Last observation carried forward
3. Regression imputation
4. Matching methods

Mean Imputation

- It's one of the easiest way to input missing values
- Replace each missing value with the mean of the observed values for that variable



Mean Imputation

Downsides:

- May lead to distortion in the distribution of the valiable
- Underestimates the standard deviation

Advantages:

- Very easy to implement
- Low computational resources needed
- May give good results if data is (almost) missing completely at random

Last Observation Carried Forward (LOCF)

- Replaces the missing data with the last available data, which is carried forward

Example:

- Clinical drug trial

	Observation time						
Unit	1	2	3	4	5	6	...
1	3.8	3.1	2.0	? –> 2.0	? –> 2.0	? –> 2.0	
2	4.1	3.5	3.8	2.4	2.8	3.0	
3	2.7	2.4	2.9	3.5	? –> 3.5	? –> 3.5	

Last Observation Carried Forward (LOCF)

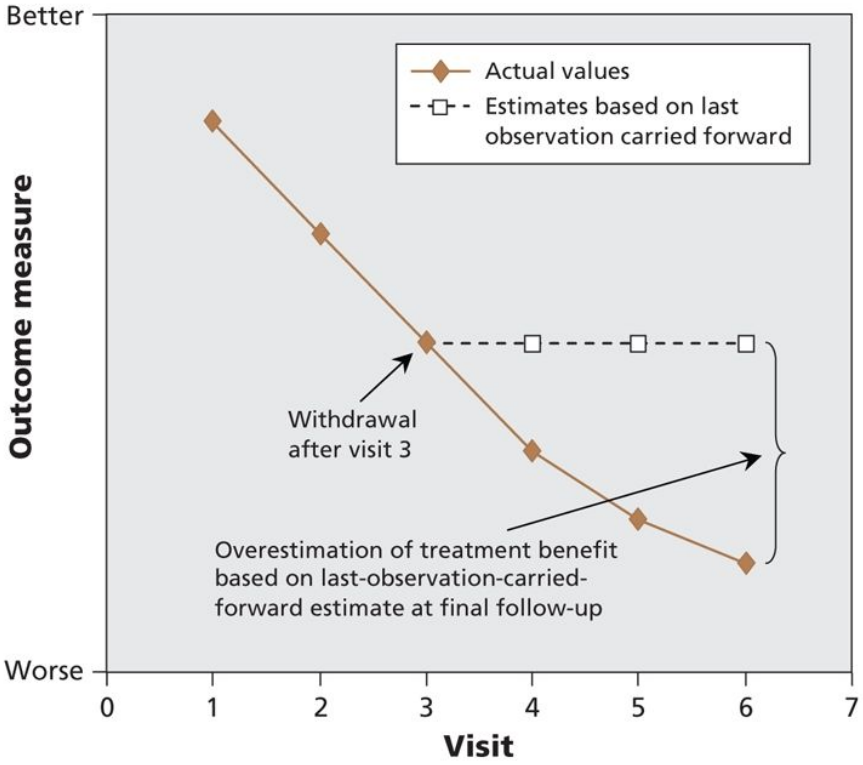


Figure 1: Potential impact of last-observation-carried-forward analysis in longitudinal randomized controlled trials in chronic progressive diseases.

Last Observation Carried Forward (LOCF)

Downsides:

- Means and covariance structure are seriously distorted
- It does not take into consideration the trend of previous versions

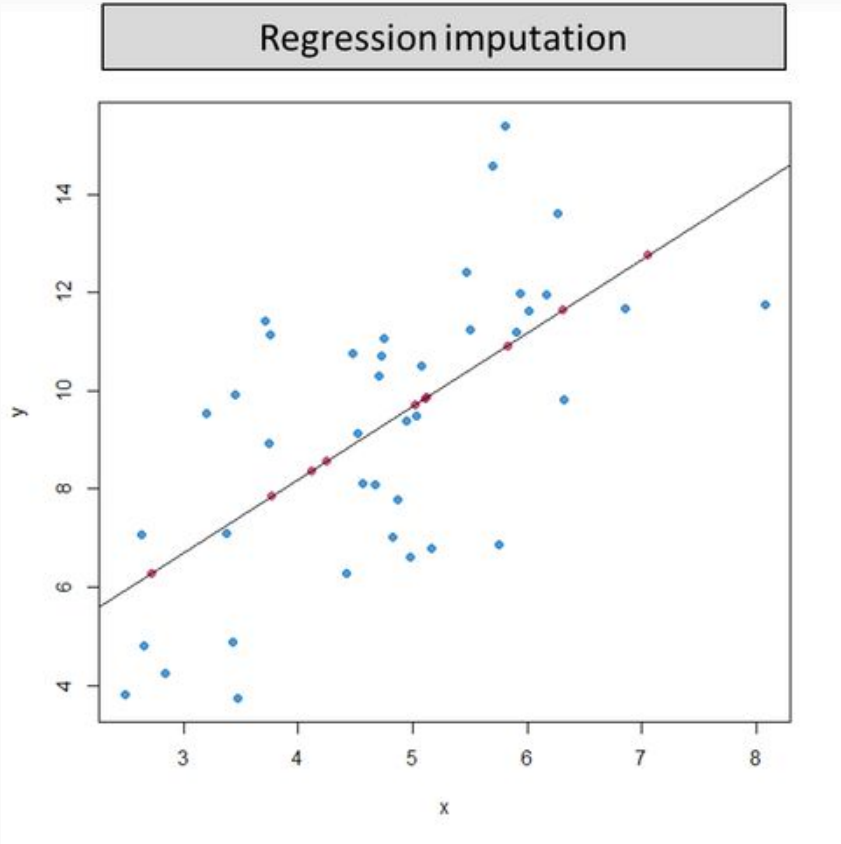
Advantages:

- No extra computational power needed
- Can perform good if data is missing completely at random

Regression imputation

- Complete observations is used to predict missing observations
- Use regression predictions to perform
  - Deterministic imputation
  - Random imputation
- Deterministic imputation (Single regression)
  - Assumes that imputed values fall on regression line with a non-zero slope
  - Implies a correlation of 1 between predictors and missing outcome variable
  - Will overestimate the correlations, the variances and covariances are underestimated

Deterministic imputation



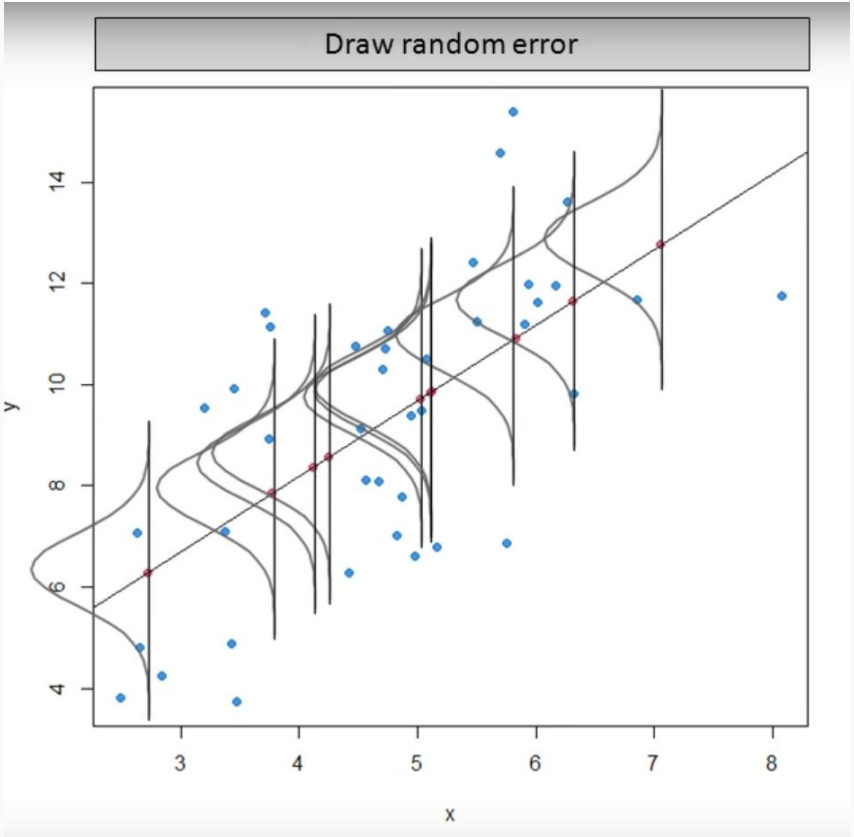
Deterministic imputation

- Advantages
  - It generates a complete data set
- Disadvantages
  - Inputs data with perfectly correlated scores
  - Overestimate correlation
  - Bias

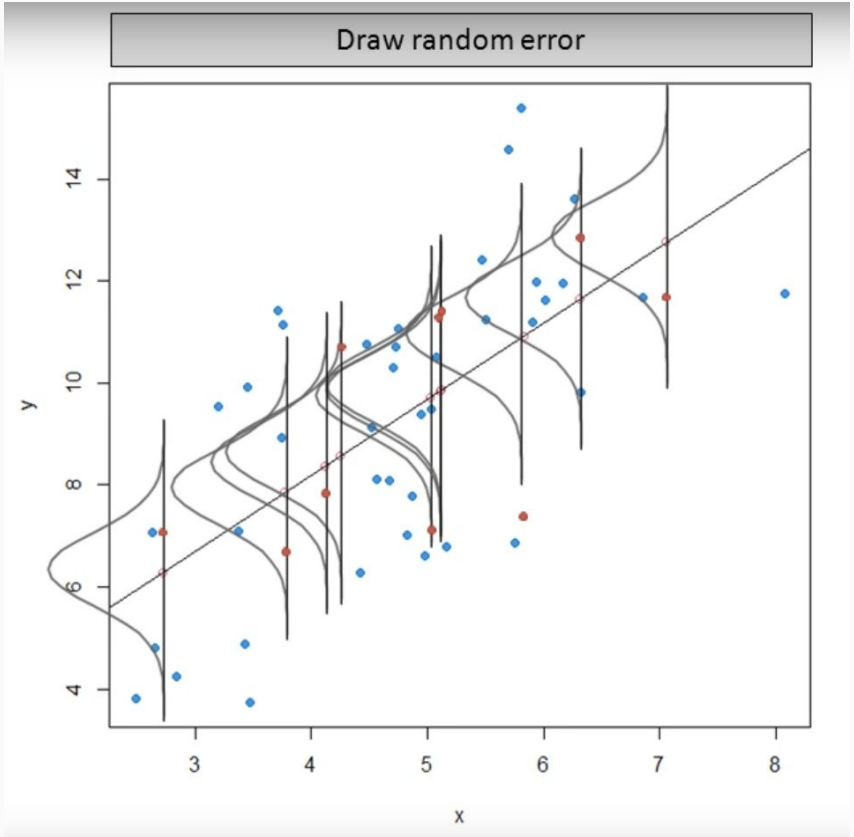
Regression imputation

- Random imputation (Stochastic regression)
  - Aims to reduce the bias by an extra step with a residual term
  - The residual term is normally distributed with a mean of zero
  - And with a variance equal to the residual variance from the regression of the predictor
  - The error of the normal distribution is added

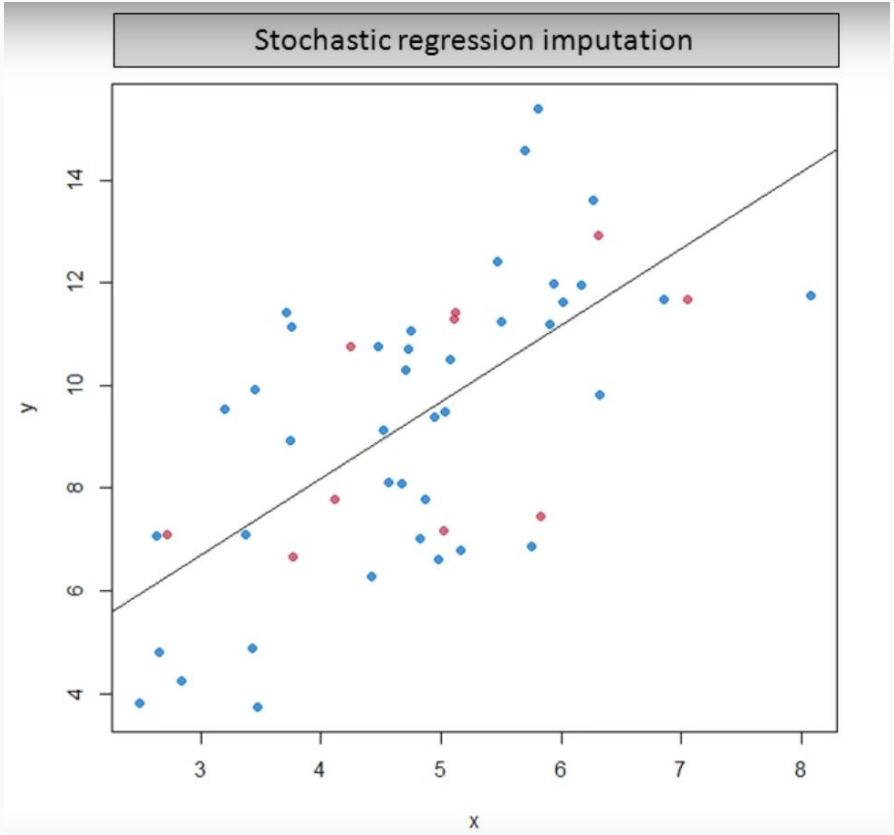
Random imputation



Random imputation



Random imputation



Random imputation

- Advantages
  - Most appropriate method
  - Input approximately equal results
  - It gives unbiased parameter under an MAR data mechanism
- Disadvantages
  - Under estimate standard error

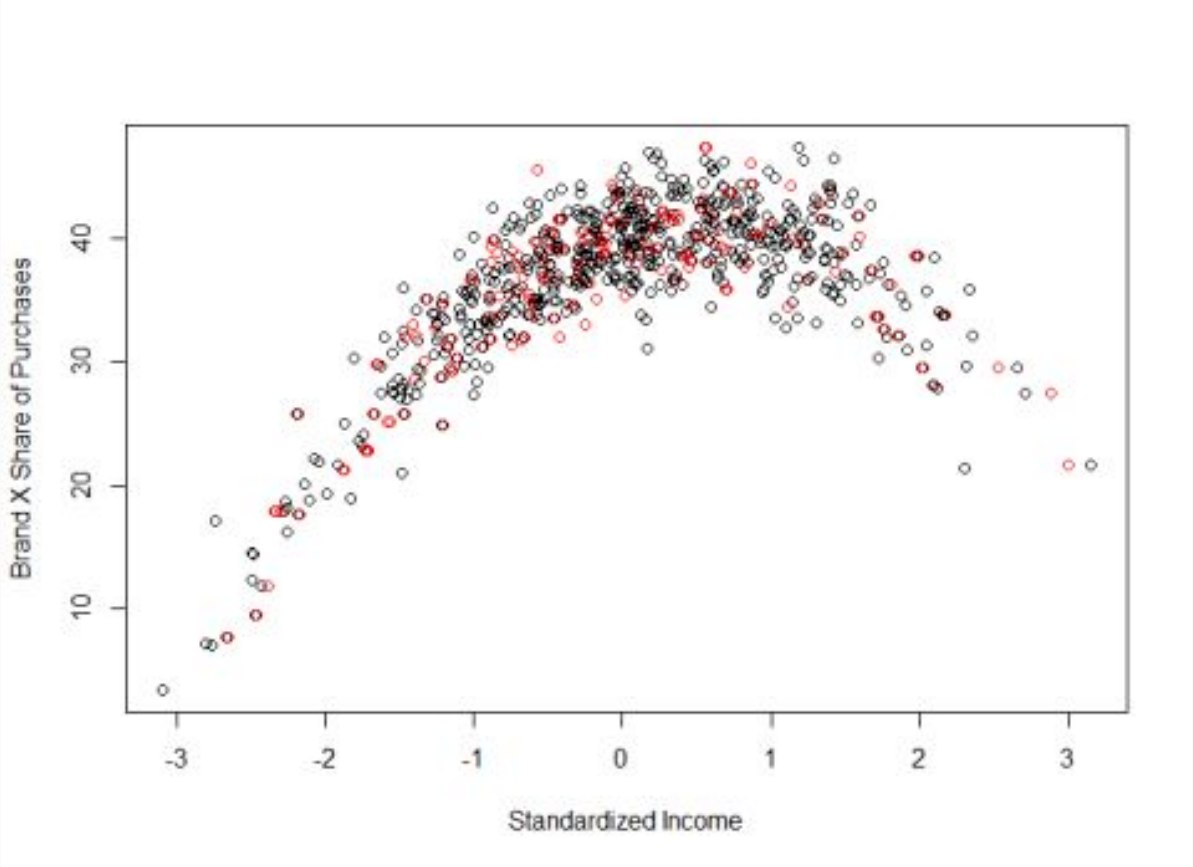
Matching methods (Hot-deck imputation)

- Non-respondents are matched to resembling respondents
- Two hot-deck approaches are
  - The distance function approach (Nearest neighbor approach)
  - The matching pattern method
- Hot-deck imputation is especially common in survey research

Matching methods (Hot-deck imputation)

- The distance function approach
  - Imputes the missing value with the case with the smallest squared distance
- The matching pattern method
  - The sample is stratified in separate homogenous groups
  - Missing values are drawn from cases in the same group

Hot-deck imputation



Matching methods (Hot-deck imputation)

- Advantages
  - It generates a complete data set
- Disadvantages
  - Not well suited for estimating measures of association
  - Produce substantially biased estimates of correlation and regression coefficients

Multiple imputation

What is multiple imputation ?

A statistical technique for analyzing incomplete data sets for which some entries are missing. Each missing value is replaced by a set of possible values.

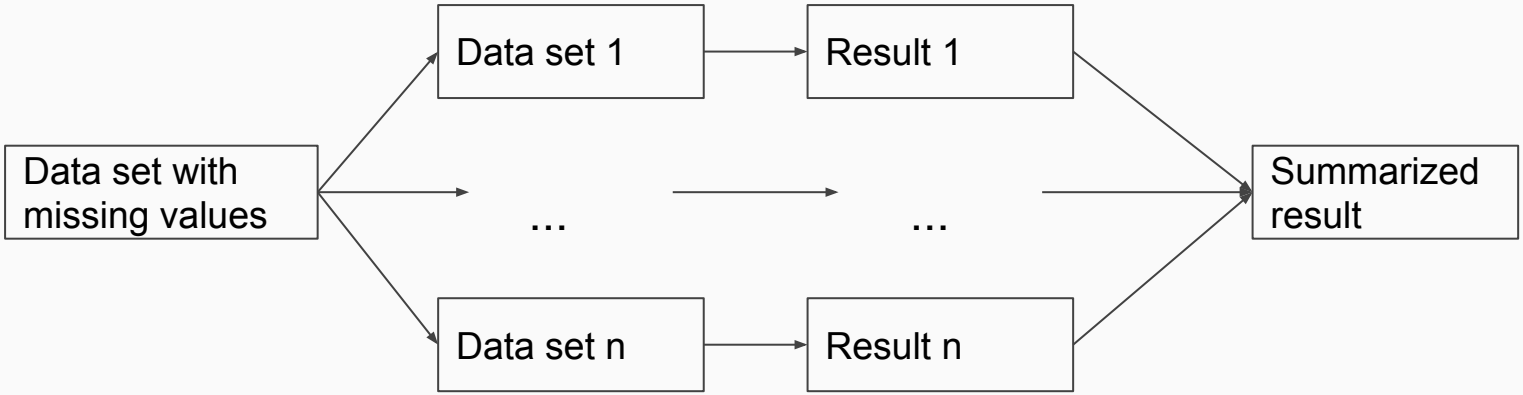
Multiple imputation

What is multiple imputation ?

When do we use it?

- *MI can be used when the data is missing completely at random, missing at random or missing not at random.*
- *When we want more precise results*

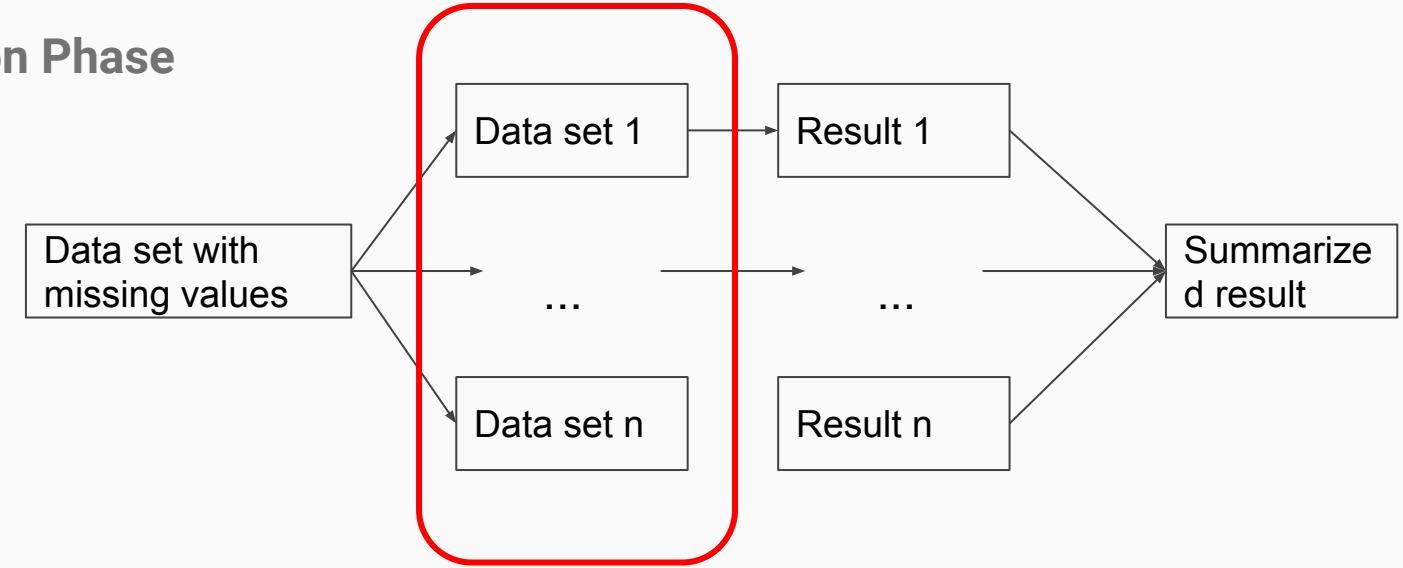
Multiple imputation



Multiple imputation

Steps:

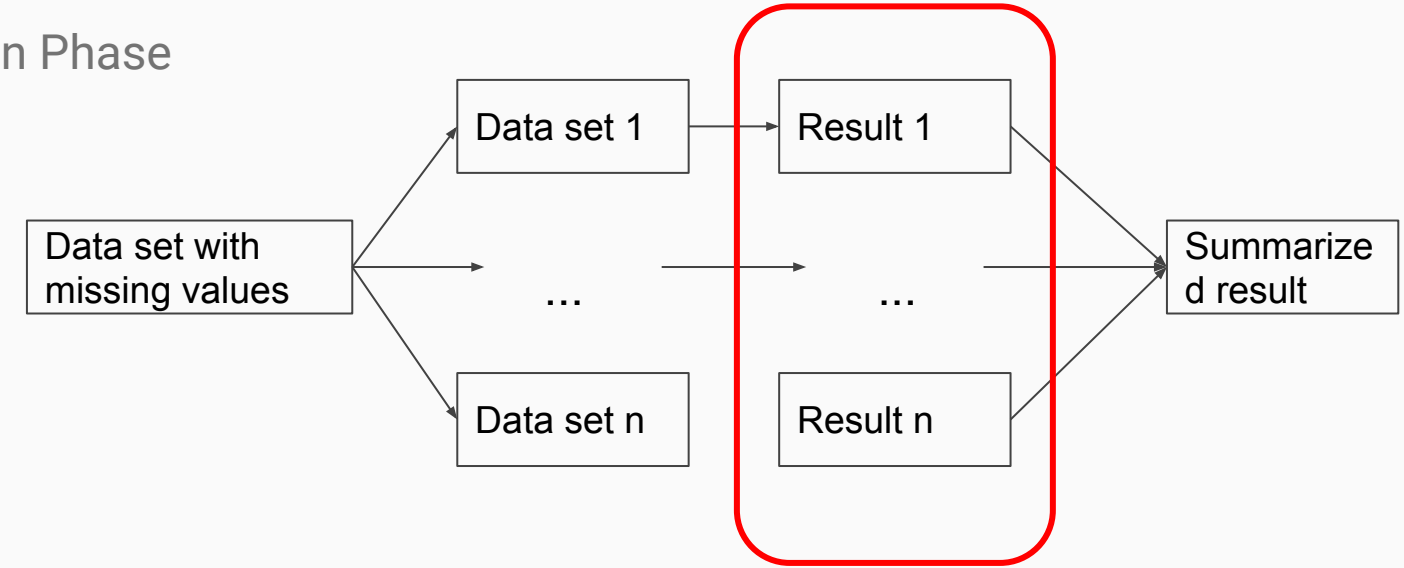
1. **Imputation Phase**



Multiple imputation

Steps:

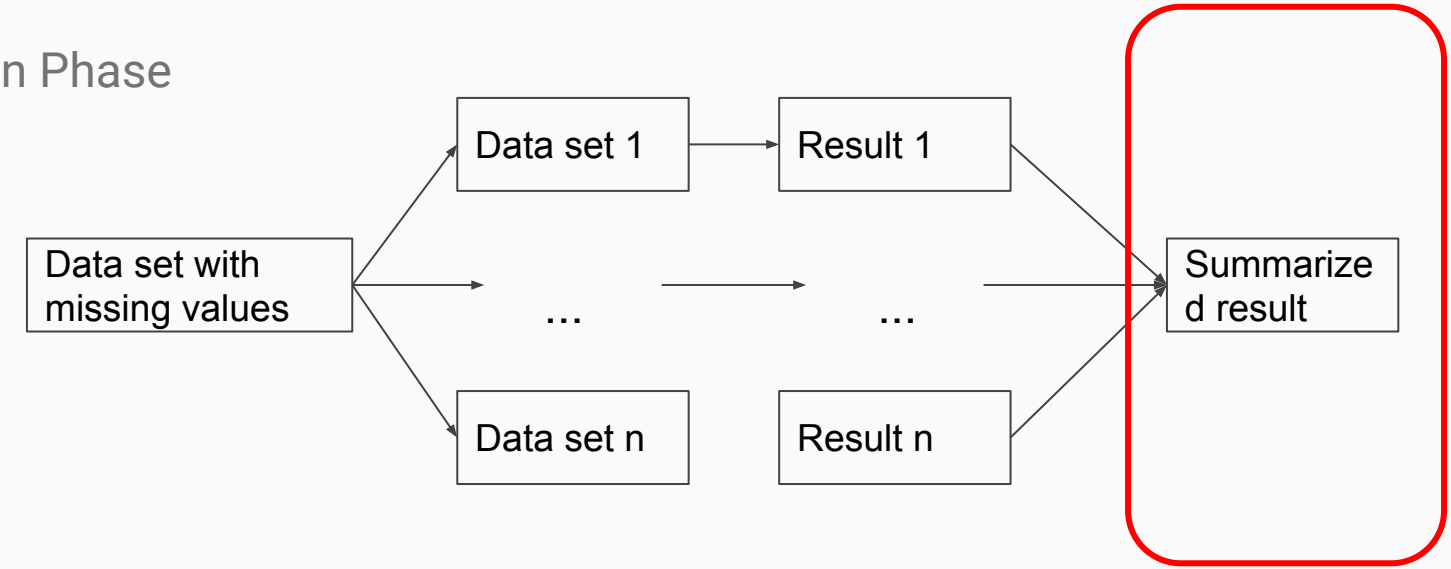
1. Imputation Phase
2. **Analysis**



Multiple imputation

Steps:

1. Imputation Phase
2. Analysis
3. **Pooling**



Multiple imputation (MICE method)

MICE Method steps :

1. Perform simple imputation
2. Set back to missing the imputed values
3. Perform regression for the missing values
4. Set the predicted values
5. Repeat 2-4 for another variable
6. Repeat 2-4 until convergence