# MULTIPLE IMPUTATION

Nidhi Menon
Nidhi.menon@anu.edu.au

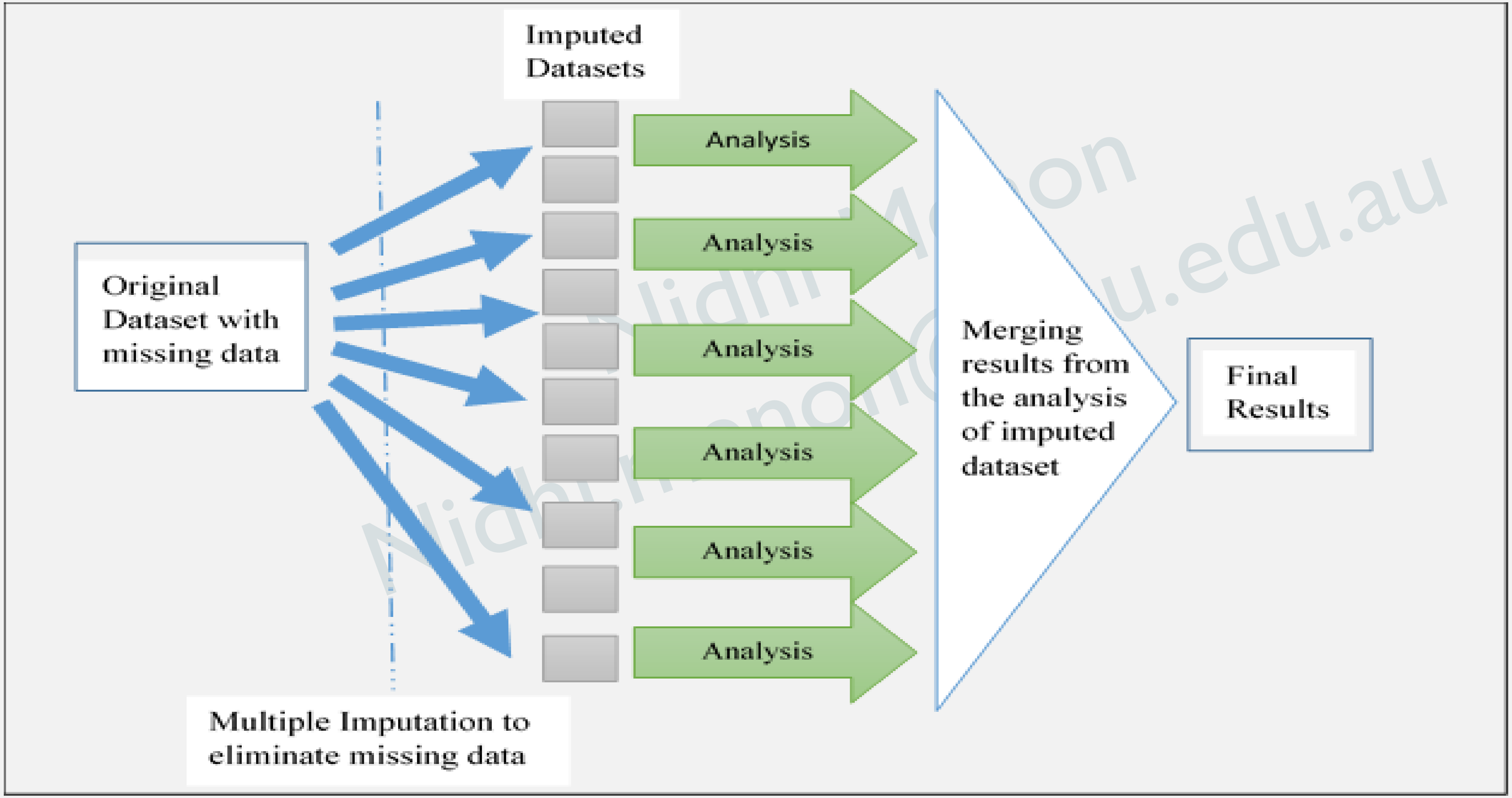
# STRATEGIES FOR HANDLING MISSING DATA

- Complete Case Analysis ( Available Case Analysis)
- Single Imputation
- Multiple Imputation

# DEVELOPMENT OF MULTIPLE IMPUTATION

- **1987:** Inception Donald. B. Rubin
- **1987:** 1st edition of Statistical Analysis with Missing Data by Little and Rubin
- **1997:** NORM, Schafer
- **1999:** MICE- concept, Van Buuren
- **2002:** SAS implemented the mi routine
- **2008:** mice in R, Van Buuren
- **2011:** Amelia was released in R
- **2012:** MI in Stata
- **2016:**mice in multilevel data, Ian White
- **2017:** jomo in R, Carpenter & Quartagno
- **2018:** micemd, Vinvent Audigier

# MULTIPLE IMPUTATION



# CHOICES TO MAKE BEFORE YOU IMPUTE

## Mechanism of Missingness

# CHOICES TO MAKE BEFORE YOU IMPUTE

Mechanism

Structure of Imputation Model

# CHOICES TO MAKE BEFORE YOU IMPUTE

Mechanism

Structure

Selecting Predictors for the Imputation Model

# CHOICES TO MAKE BEFORE YOU IMPUTE

Structure

Mechanism

Predictors

Imputing Derived Variables

# CHOICES TO MAKE BEFORE YOU IMPUTE

- Mechanism
- Structure
- Derived Variables
- Predictors
- Number of Imputations

# CHOICES TO MAKE BEFORE YOU IMPUTE

Mechanism

Structure

Predictors

Derived Variables

No. of imputations

Order of Imputations

# RECIPE FOR IMPUTATION

1. Define the most general analytic model to be applied to imputed data
2. The target variable is the variable with the missing values
3. Select a method that imputes close to the data
4. Include all level-1 variables and their cluster means
5. Include all level-2 predictors
6. Include any interactions implied by the model
7. Exclude any terms involving the target variable

# METHODS OF IMPUTATION – FCS/ MICE

- Can be used for datasets containing both continuous and categorical data.
- Defines an imputation model on a variable by variable basis –> great for datasets with complex structures
- The method also allows the researcher to account for the complexities observed in the data, in the imputation model.
- Consider a scenario with 3 partially missing covariates namely $X_1$,$X_2$ and $X_3$ and outcome variable Y is complete. Here, $X_1 = [X_1^{mis} ; X_1^{obs}]$ ; $X_2 = [X_2^{mis} ; X_2^{obs}]$ & $X_3 = [X_3^{mis} ; X_3^{obs}]$

*Iteration (1):*

$$\theta_1^{(1)} \sim f(\theta_1).f(x_1^{obs} | x_2^{obs}, x_3^{obs}, y, \theta_1)$$

$$x_1^{mis(1)} \sim f(x_1^{mis} | x_2^{obs}, x_3^{obs}, y, \theta_1^{(1)})$$

$$\theta_2^{(1)} \sim f(\theta_2).f(x_2^{obs} | x_1^{obs(1)}, x_3^{obs}, y, \theta_2)$$

$$x_2^{mis(1)} \sim f(x_2^{mis} | x_1^{obs(1)}, x_3^{obs}, y, \theta_2^{(1)})$$

$$\theta_3^{(1)} \sim f(\theta_3).f(x_3^{obs} | x_1^{obs(1)}, x_2^{obs(1)}, y, \theta_3)$$

$$x_3^{mis(1)} \sim f(x_3^{mis} | x_1^{obs(1)}, x_2^{obs(1)}, y, \theta_3^{(1)})$$

# METHODS OF IMPUTATION – JOMO

- Defines a multivariate joint model for all variables in the dataset for imputation of missing values in the outcome.
- Outcome here refers to the variables in the model with missing values and not the outcome of the analysis model.
- Suppose we have variables $Y_1$ and $Y_2$ are partially observed and $Y_3$ and $Y_4$ are variables with no missing data, then the simplest joint model is the multivariate normal model given by:

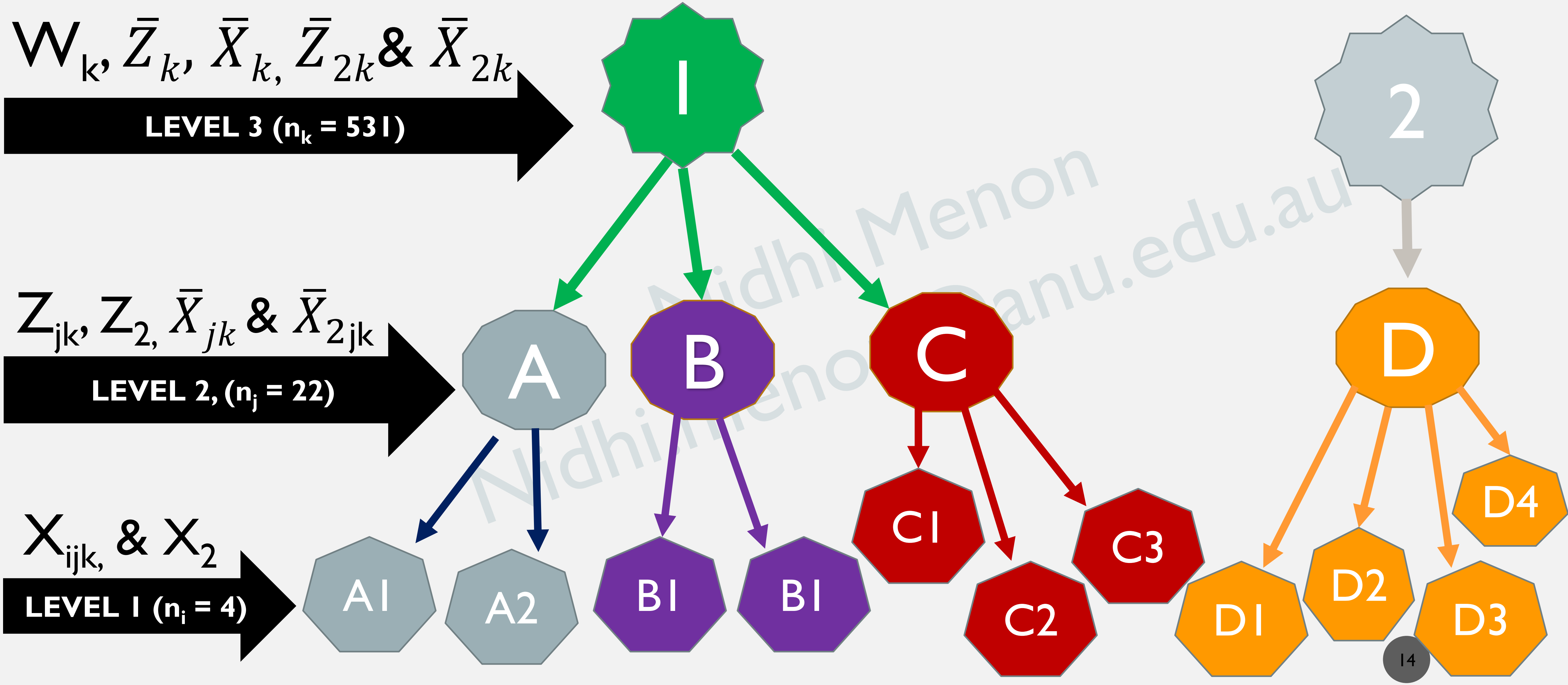
$$Y_{1,i} = \beta_{01} + \beta_{11} Y_{3,i} + \beta_{21} Y_{4,i} + e_{1,i}$$
$$Y_{2,i} = \beta_{01} + \beta_{11} Y_{3,i} + \beta_{21} Y_{4,i} + e_{2,i}$$
$$\begin{pmatrix} e_{1,i} \\ e_{2,i} \end{pmatrix} \sim N(0, \Omega)$$

- Jomo used the Gibbs Sampling approach by consistently drawing new values for all parameters i.e. the fixed

effects (β), the covariance matrix and the missing data.

- The current draw of missing values is combined with the observed data to make the first imputed dataset

# DEMONSTRATION



# ANALYSIS MODEL

$$Y_{ijk} = \beta_{0jk} + \beta_{1jk}(X_{ijk} - \bar{X}_{jk}) + e_{ijk};\ e_{ijk} \sim N(0, \sigma^2)$$

$$\beta_{0jk} = \gamma_{00k} + \gamma_{01}(Z_{jk} - \bar{Z}_k) + \gamma_{02}(\bar{X}_{jk} - \bar{X}_k) + r_{0jk};\ r_{0jk} \sim N(0, \tau_r)$$

$$\beta_{1j} = \gamma_{10}$$

$$\gamma_{00k} = \delta_{000} + \delta_{001}(W_k) + u_{00k};\ u_{00k} \sim N(0, \tau_u)$$

Substituting, we get

$$Y_{ijk} = \delta_{000} + \gamma_{10}(X_{ijk} - \bar{X}_{jk}) + \gamma_{02}(\bar{X}_{jk} - \bar{X}_k) + \gamma_{01}(Z_{jk} - \bar{Z}_k) + \delta_{001} W_k + u_{00k} + r_{0jk} + e_{ijk}$$

Here, $i = 1, 2, ...n_{jk}, j = 1, 2, ...n_k$ & $k = 1, 2, ...K$.

# DEMONSTRATION

| | |
|---|---|
| $e_{ijk}$ | $rnorm(n = i * j * k, mean = 0, sd = 1)$ |
| $r_{0jk}$ | $rep(rnorm(n = k * j, mean = 0, sd = 1), each = i)$ |
| $u_{00k}$ | $rep(rnorm(n = k, mean = 0, sd = 1), each = i * j)$ |
| *Coefficients* | $g_{000} = 2,\ g_{100} = 2.5,\ g_{200} = 2.5,\ g_{010} = 2, g_{001} = 3$ |
| $Y_{ijk}$ | $\delta_{000} + \gamma_{10}(X_{ijk} - \bar{X}_{jk}) + \gamma_{02}(\bar{X}_{jk} - \bar{X}_k) + \gamma_{01}(Z_{jk} - \bar{Z}_k) + \delta_{001}W_k + u_{00k} + r_{0jk} + e_{ijk}$ |
| $X_{ijk}$ | $rsn(n = i * j * k, xi = 70, omega = 20, alpha = 10)$ |
| $Z_{jk}$ | $rep(rtpois(j * k, lambda = 3, a = 2, b = 25), each = i)$ |
| $W_k$ | $rep(rtpois(k, lambda = 3, a = 2, b = 10), each = i * j)$ |
| *Correlated Variables:* | |
| Z | $rtruncnorm(n = i * j * k, a = 59, b = 396, mean = 114.8, sd = 14)$ |
| $X_2$ | $(0.6) * X_ijk + sqrt(1 - 0.6) * Z; correlation = 0.63$ |
| $Z_1$ | $rep(rtpois(j * k, lambda = 5.8, a = 1, b = 41), each = i$ |
| $Z_2$ | $(0.87) * Zjk + sqrt(1 - 0.87) * Z1; correlation = 0.7$ |

# ANALYSIS ON ORIGINAL DATASET

Table 3: Analysis of the complete (simulated) dataset

| *Fixed Effect* | *Coefficient* | *se* | *p-value* |
|---|---|---|---|
| Intercept | 1.851 | 0.618 | $< 0.001$ |
| $X_{ijk} - \bar{X}_{jk}$ | 2.5 | $4.385e^{-04}$ | $< 0.001$ |
| $\bar{X}_{jk} - \bar{X}_k$ | 2.498 | $1.767e^{-03}$ | $< 0.001$ |
| $Z_{jk} - \bar{Z}_k$ | 2.001 | $8.341e^{-02}$ | $< 0.001$ |
| $W_k$ | 3.070 | $3.712e^{-02}$ | $< 0.001$ |
| *Random Effects* | *Variance* | *Std. Dev* | |
| Level 3 effect ($u_{00k}$) | 0.9969 | 0.9984 | |
| Level 2 effect ($r_{0jk}$) | 1.0354 | 1.0176 | |
| Level 1 effects ($e_{ijk}$) | 0.9996 | 0.9998 | |

# INTRODUCING MISSING DATA

- Probability of MAR in $X_{ijk}$ was determined by model

$$\frac{e^{X_{2s}+\beta Y_s}}{1+e^{X_{2s}+\beta Y_s}};\ where\ Y_s = \frac{(Y - E(Y))}{SD(Y)} \& \ X_{2s} = \frac{X_2 - E(X_2)}{SD(X_2)}$$

- Probability of MAR in $Z_{jk}$ was determined by model

$$\frac{e^{Z_{2s}+\beta Y_s'}}{1+e^{Z_{2s}+\beta Y_s'}};\ where\ Y_s' = \frac{(\bar{Y}_{jk} - E(\bar{Y}_{jk}))}{SD(\bar{Y}_{jk})} \& \ Z_{2s} = \frac{Z_2 - E(Z_2)}{SD(Z_2)}$$

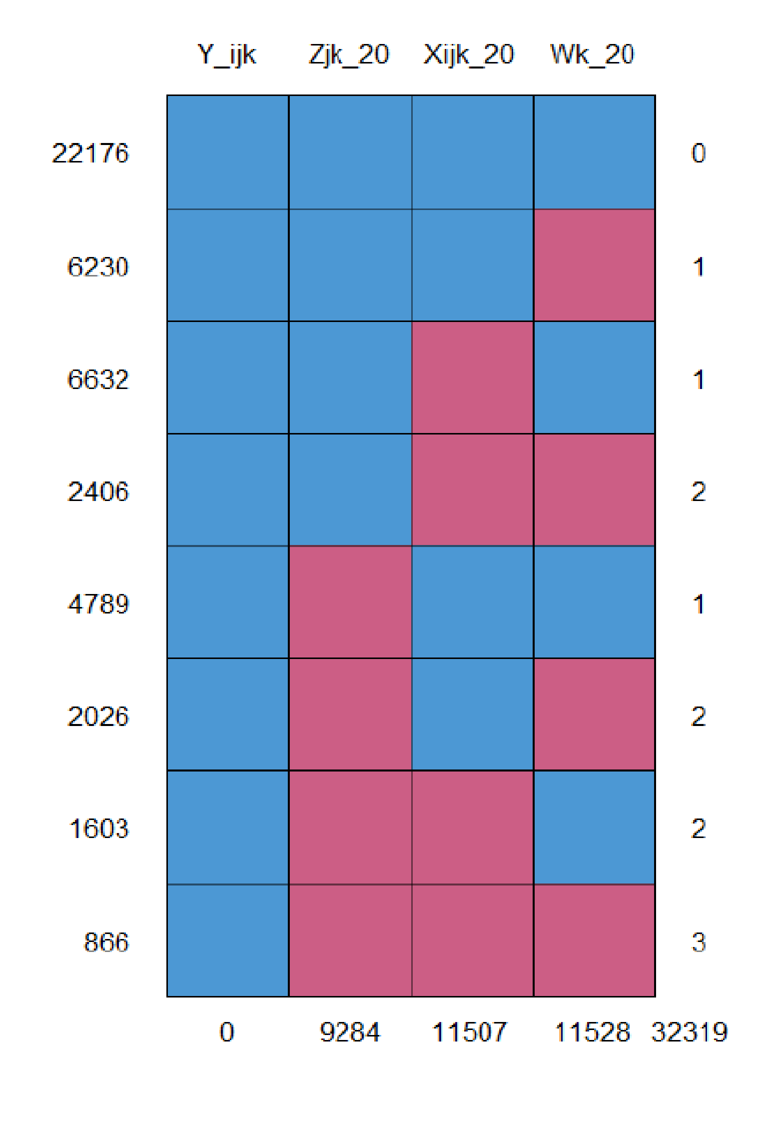
- Probability of MAR in $W_k$ was determined by model

$$\frac{e^{2+\beta Y_s'}}{1+e^{2+\beta Y_s'}};\ where\ Y_s' = \frac{(\bar{Y}_k - E(\bar{Y}_k))}{SD(\bar{Y}_k)}$$

**Scenario 1:** 20% Missing in both $X_{ijk}$ and $Z_{jk}$

**Scenario 2:** 20% Missing in both $X_{ijk}$, $Z_{jk}$ and $W_k$

**Scenario 3:** 50% Missing in both $X_{ijk}$ & $Z_{jk}$, and 20% Missing in $W_k$



## Available Case Analysis

| Scenario - 3 | (n = 11,083) | | |
|---|---|---|---|
| *Fixed Effect* | *Coefficient* | *se* | *p-value* |
| Intercept | -218 | 0.703 | $< 0.001$ |
| $X_{ijk} - \bar{X}_{jk}$ | 2.495 | $1.801e^{-03}$ | $< 0.001$ |
| $\bar{X}_{jk} - \bar{X}_k$ | 2.597 | $8.866e^{-02}$ | $< 0.001$ |
| $Z_{jk} - \bar{Z}_k$ | 2.024 | $2.159e^{-02}$ | $< 0.001$ |
| $W_k$ | 2.784 | 0.143 | $< 0.001$ |
| *Random Effects* | *Variance* | *Std. Dev* | |
| Level 3 effect ($u_{00k}$) | 1.012 | 1.006 | |
| Level 2 effect ($r_{0jk}$) | 11.701 | 3.421 | |
| Level 1 effects ($e_{ijk}$) | 1.010 | 1.005 | |

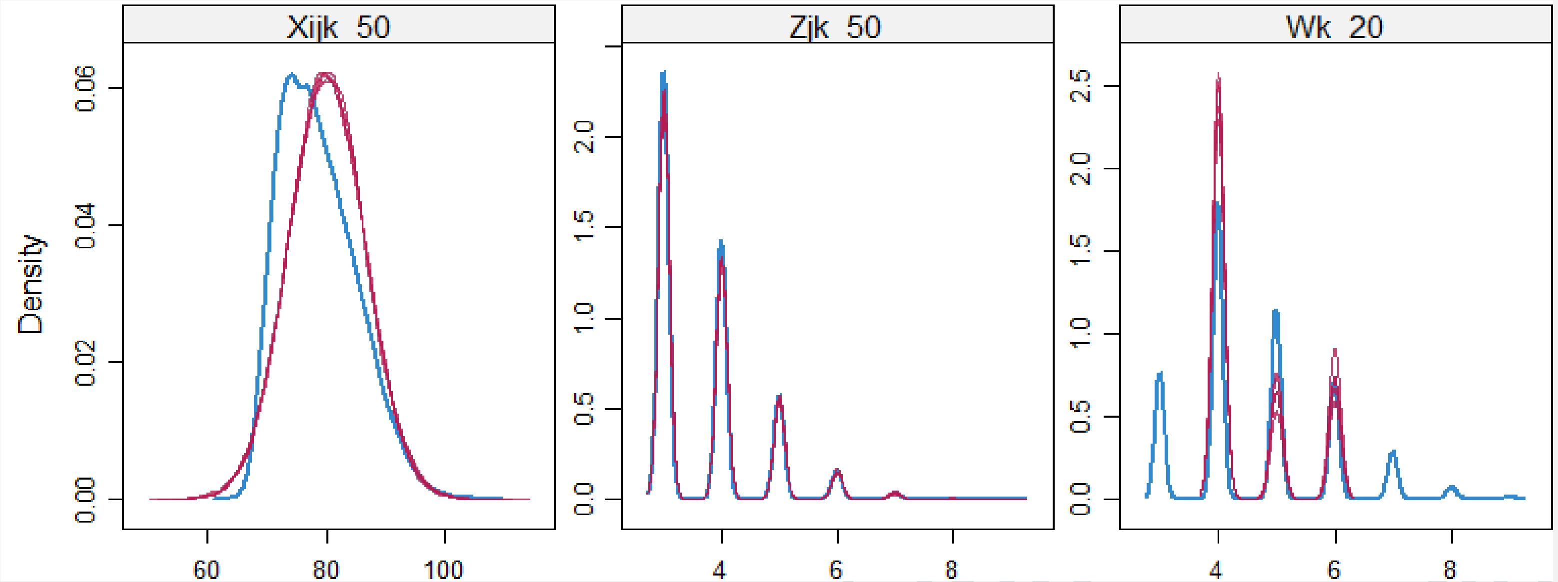


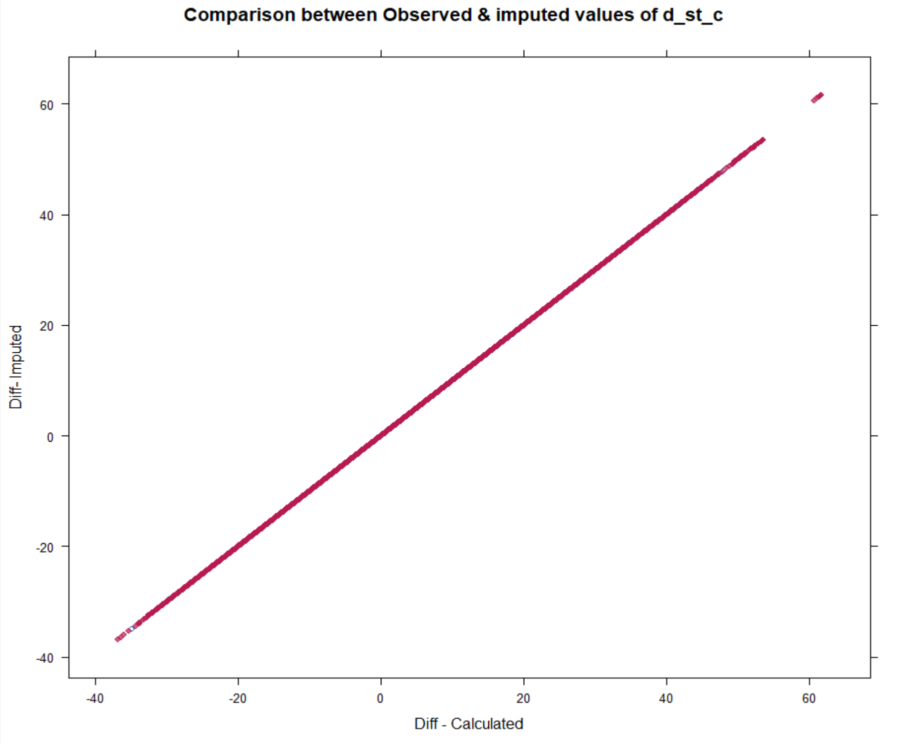
## JOMO

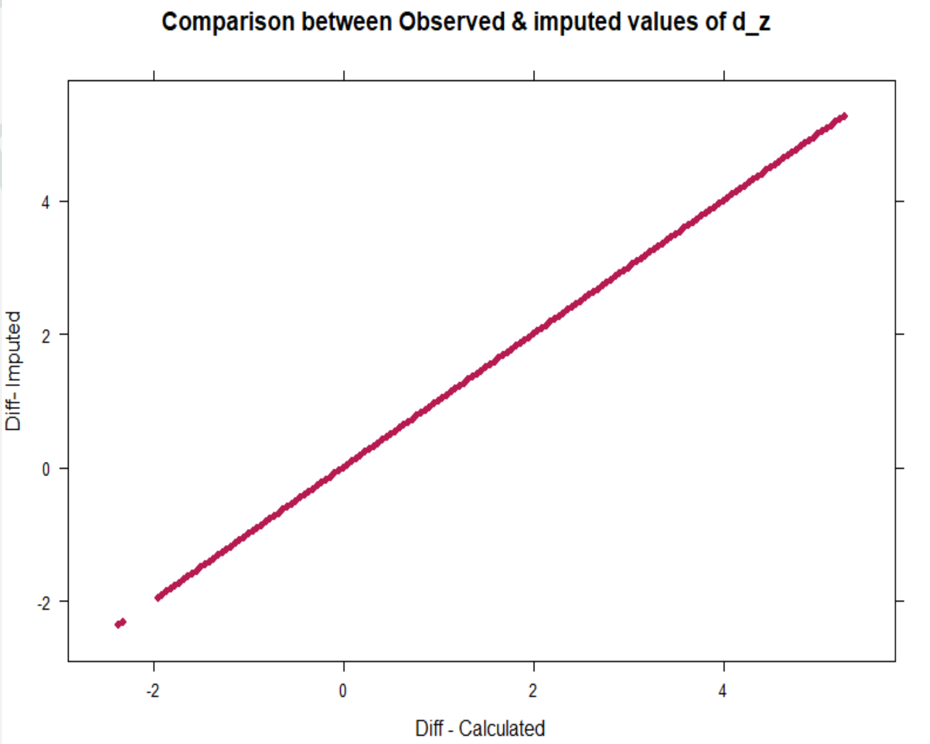
| Scenario - 2 | | | |
|---|---|---|---|
| *Fixed Effect* | *Coefficient* | *se* | *p-value* |
| Intercept | 1.655 | 0.178 | $< 0.001$ |
| $X_{ijk} - \bar{X}_{jk}$ | 1.775 | 0.008 | $< 0.001$ |
| $\bar{X}_{jk} - \bar{X}_k$ | 1.7637 | 0.0109 | $< 0.001$ |
| $Z_{jk} - \bar{Z}_k$ | 1.763 | 0.0781 | $< 0.001$ |
| $W_k$ | 3.0701 | 0.0404 | $< 0.001$ |
| *Random Effects* | *Variance* | | |
| Level 3 effect ($u_{00k}$) | 2.5263 | | |
| Level 2 effect ($r_{0jk}$) | $1.3098e - 13$ | | |
| Level 1 effects ($e_{ijk}$) | 1.0943 | | |

## MICE

| Scenario - 3 | | | |
|---|---|---|---|
| *Fixed Effect* | *Coefficient* | *se* | *p-value* |
| Intercept | 2.877 | 0.7640 | $< 0.001$ |
| $X_{ijk} - \bar{X}_{jk}$ | 1.876 | 0.0396 | $< 0.001$ |
| $\bar{X}_{jk} - \bar{X}_k$ | 2.407 | 0.0432 | $< 0.001$ |
| $Z_{jk} - \bar{Z}_k$ | 2.9681 | 0.2130 | $< 0.001$ |
| $W_k$ | 2.5492 | 0.1730 | $< 0.001$ |
| *Random Effects* | *Variance* | *Std. Dev* | |
| Level 3 effect ($u_{00k}$) | $2.0894e^{-14}$ | | |
| Level 2 effect ($r_{0jk}$) | 0.7294 | | |
| Level 1 effects ($e_{ijk}$) | 7.622 | | |


Xijk 50
Zjk 50
Wk 20
Density
0.00
0.02
0.04
0.06
60
80
100
0.0
0.5
1.0
1.5
2.0
4
6
8
0.0
0.5
1.0
1.5
2.0
2.5
4
6
8


Comparison between Observed & imputed values of d_st_c
Diff- Imputed
Diff - Calculated
-40
-20
0
20
40
60
-40
-20
0
20
40
60


Comparison between Observed & imputed values of d_z
Diff- Imputed
Diff - Calculated
-2
0
2
4
-2
0
2
4

# THANK YOU!

# DISADVANTAGES

*Disadvantages of JoMo*:- Considering a joint model on variables subject to missingness may not always be feasible or even realistic. For example, consider a survey with items targeted at different sub-populations; e.g. item asking respondents when was their last pap smear or item asking respondents the number of cigarettes smoked in the last 24 hours. This could apply to even questionnaires with a skip pattern. Imposing a joint distribution when a joint distribution may not even exist is not practicable. There are several cases when a joint modelling strategy may not work such as when variables have nominal, count or semi-continuous variables (Yucel, 2008). Thus researchers must remain cautious when choosing the right method of imputation bearing these factors in mind.

*Disadvantages of FCS*:- Although an appealing method of imputation, FCS is not without its limitations. The method is based on the assumption that the data is missing at random (MAR). Secondly, each conditional distribution needs to be specified separately. This would result in substantial modelling especially for datasets with many variables. The technique is more computationally challenging compared to joint modelling.