Australian
National
University

Multiple imputation in three level models

Alice Richardson, NCEPH, ANU

Nidhi Menon, NCEPH, ANU

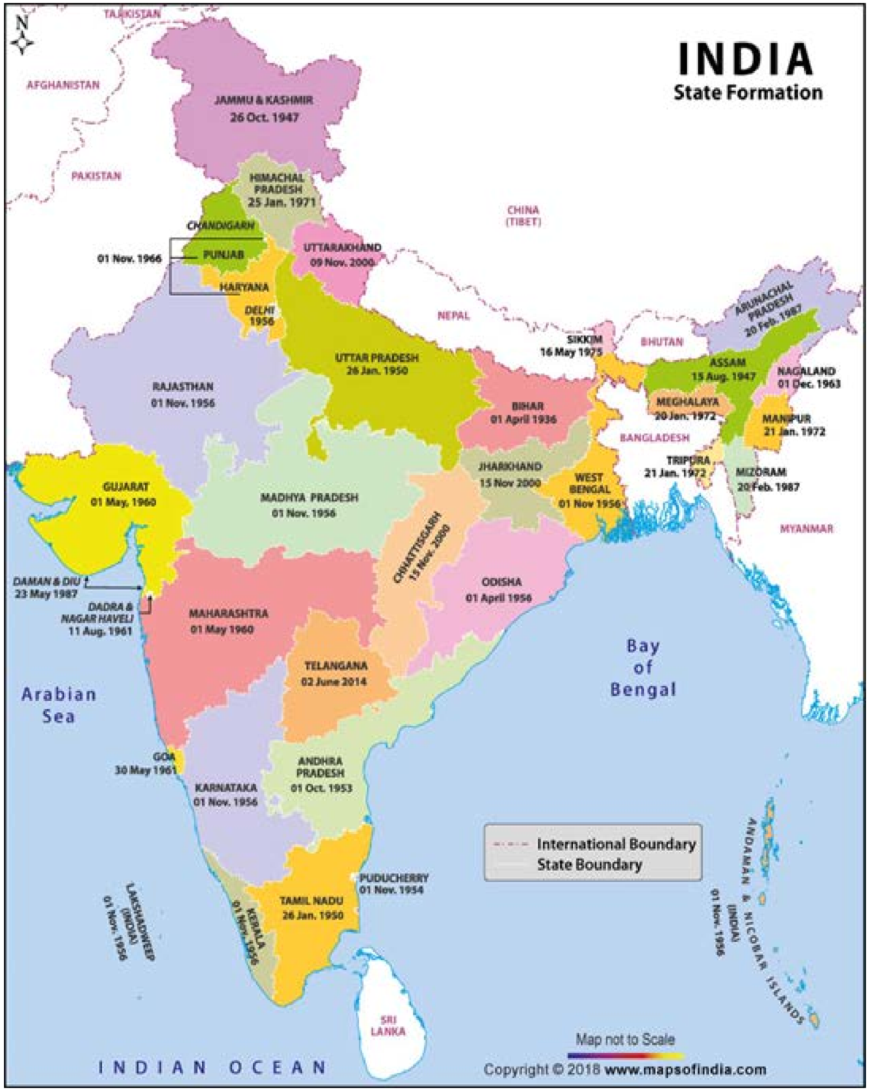
Introductions

Overview

- Data
- Models
- Simulation study
- Results
- Future work

Data

- National Family Health Survey 4, India 2015 – 2016
- Information on population, health and nutrition for each state and Union Territory
- Vital estimates of the prevalence of malnutrition, anaemia, hypertension, HIV, and high blood glucose levels through a series of biomarker tests and measurements



Structure of NFHS- 4 (2015-16)

- Individuals → Households → Districts

- A total of 3,604,509 occupied households were interviewed.

- Response rates for women and men are observed at 92%. This implies that unit nonresponse was approximately 8%. The same cannot be said for item non-response

- Measure the relative impact of individual and household risk factors for anaemia using variables such as

  - Systolic blood pressure
  - Number of household members
  - Toilet facilities

Simulation study setup

Variable	Description
Level 3	Number of higher level units e.g. districts(n_k = 14)
Level 2	Number of higher level units each nested within level 3 units. e.g. households (n_{jk} = 800)
Level 1	Number of lower level units each nested within level 2 units, further nested within level 3 units. e.g. household members(n_{ijk} = 4)
X	Predictor variable measured at level 1 corresponding to i^{th} individual in j^{th} household in k^{th} district
Z	Predictor variable measured at level 2 corresponding to j^{th} household in k^{th} district
W	Predictor variable measured at level 3 corresponding to k^{th} district
Y	Outcome variable measured at level 1 corresponding to i^{th} individual in j^{th} household in k^{th} district
$\bar{X}_{jk}$	Mean of X_{ijk} calculated for each level 2 unit
$\bar{Z}_{k}$	Mean of Z_{jk} calculated for each level 3 unit
$\bar{X}_{k}$	Mean of X_{ijk} calculated for each level 3 unit
$X_{ijk} - \bar{X}_{jk}$	Deviation of each level 1 observation (X_{ijk}) from the level 2 mean
$\bar{X}_{jk} - \bar{X}_{k}$	Deviation of each level 2 mean from the level 3 mean
$Z_{jk} - \bar{Z}_{k}$	Deviation of each level 2 observation (Z_{jk}) from the level 3 mean

Random Intercept Model

$$Y_{ijk} = \gamma_{000} + \gamma_{100}(X_{ijk} - \overline{X_{jk}}) + \gamma_{200}(\overline{X_{jk}} - \overline{X_k}) + \gamma_{010}(Z_{jk} - \overline{Z_k}) + \gamma_{001}(W_k) + u_k + r_{jk} + e_{ijk}$$

Assumed Values and Distributions for Data generation

Variables
$X \sim N(75.6, 17.14)$
$Z \sim Pois(\lambda = 3)$
$W \sim Pois(\lambda = 3)$

Error Terms
$u_k \sim N(0,1)$
$r_{jk} \sim N(0,1)$
$e_{ijk} \sim N(0,1)$

Constants
$\gamma_{000} = 2$
$\gamma_{100} = 2.5$
$\gamma_{200} = 2.0$
$\gamma_{010} = 2.5$
$\gamma_{001} = 2.5$
$i = 4$
$j = 800$
$k = 14$

MAR Mechanism in X, Z and W

We create 2 variables X_2 and Z_2 correlated to X and Z such that corr(X, X_2) = 0.5 and corr(Z, Z_2) = 0.5

- **Probability of MAR in X_{ijk}**
  - $p_i = \dfrac{e^{X2+\ \beta Ys}}{1+e^{X2+\ \beta Ys}}$; where $Y_s = \dfrac{(Y - E(Y))}{SD(Y)}$

- **Probability of MAR in Z_{jk}**
  - $p_i = \dfrac{e^{Z2+\beta Y's}}{1+e^{Z2+\beta Y's}}$; where $Y'_s = \dfrac{(\overline{Yjk} - E(\overline{Yjk}))}{SD(\overline{Yjk})}$

- **Probability of MAR in W_k**
  - $p_i = \dfrac{e^{2-0.85Y's}}{1+e^{2-0.85Y's}}$; where $Y'_s = \dfrac{(\overline{Yk} - E(\overline{Yk}))}{SD(\overline{Yk})}$

Multiple Imputation using Chained Equations (mice)

- To create multiple imputations y^* of y_{mis}
  1. Calculate the posterior distribution $P(\theta|y_{obs})$ of θ based on the observed data y_{obs};
  2. Draw a value θ^* from $P(\theta|y_{obs})$;
  3. Draw a value y^* from $P(y_{mis}|y_{obs}, \theta = \theta^*)$, the conditional posterior distribution of y_{mis} given $\theta = \theta^*$.
- Repeat 2 - 3 for all variables $\rightarrow$ first cycle
- Run cycles till convergence

Joint modelling (JoMo)

- Employs Bayesian estimation that views the missing values, residuals, and model parameters as random variables having a joint distribution. For iteration (t), the univariate draw steps are

$$y_{ij}^{(t)} \sim N\left(\beta_{0(y)}^{(t)} + \beta_{1(y)}^{(t)} x_{ij}^{(t-1)} + \beta_{2(y)}^{(t)} zj + u_{0j(y)}^{(t)}, \sigma^{2(t)}{}_{(y|xz)}\right)$$
$$x_{ij}^{(t)} \sim N\left(\beta_{0(x)}^{(t)} + \beta_{1(x)}^{(t)} y_{ij}^{(t)} + \beta_{2(x)}^{(t)} z_j + u_{0j(x)}^{(t)}, \sigma^{2(t)}{}_{(x|yz)}\right)$$

- One of the limiting factors of joint modelling is that it works best at the lowest level.

- To overcome this limitation, JoMo uses separate Gibbs samplers – one for each level with missingness.

Passive imputation – Impute then Transform approach

Derived Variables $\bar{X}_{jk}$, $\bar{Z}_k$ and $\bar{X}_k$ were recalculated using the imputed values of X and Z.

d_st_c =($\boldsymbol{X_{ijk}}$- $\bar{X}_{jk}$), d_c_sc =($\bar{X}_{jk}$- $\bar{X}_k$)and d_z =($\boldsymbol{Z_{jk}}$ -$\bar{Z}_k$)

```
impList<- miceadds::mids2datlist(imp)
within(impList,{
      Xbar_jk<-clusterMeans(Xijk_20,level2)
      d_st_c20 <- (Xijk_20 - Xbar_jk)
      Xbar_k<-clusterMeans(Xijk_20, level3)
      d_c_sc <- (Xbar_jk - Xbar_k)
            })
```

Overview of methodology used for imputation

- Gelman and Hill approach
- Create two different datasets for individual and group level data
- Group level dataset includes aggregate forms of individual level measurements when imputing for missing values in this level.

- 20% and 50% MAR introduced in level-1 and level-2 covariates separately and combined.

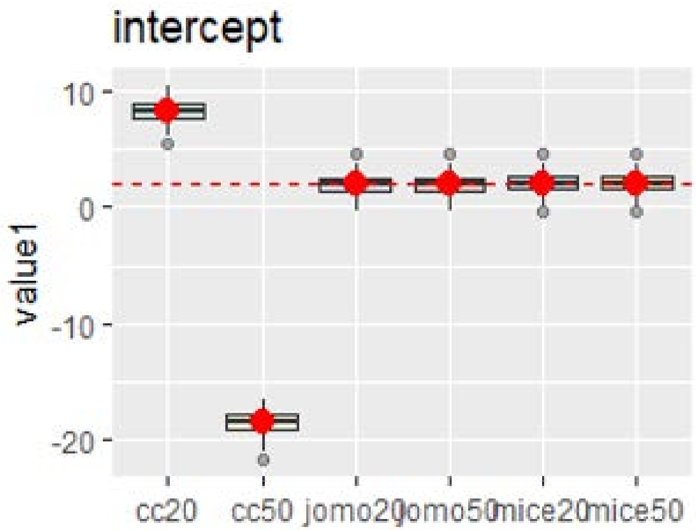
Overview of methodology used for imputation (cont.)

- Levels 2 and 3 were combined to identify a unique clustering variable to identify each observation in the dataset. This was done in the imputation model to overcome the software limitation of defining only one clustering variable.

- Performance of MICE and JoMo were compared with complete case analysis.

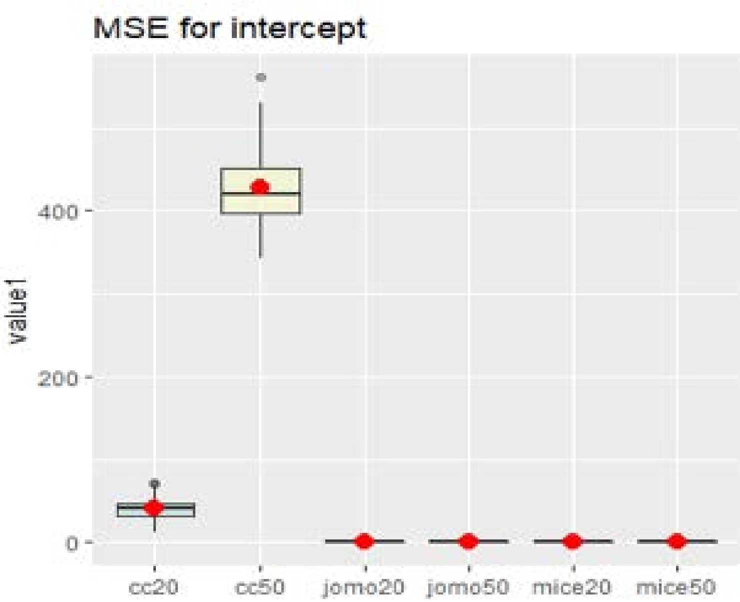
- Measures to assess performance
  - Comparison of distribution of imputed v/s observed data
  - Mean Squared Errors
  - Relative Bias

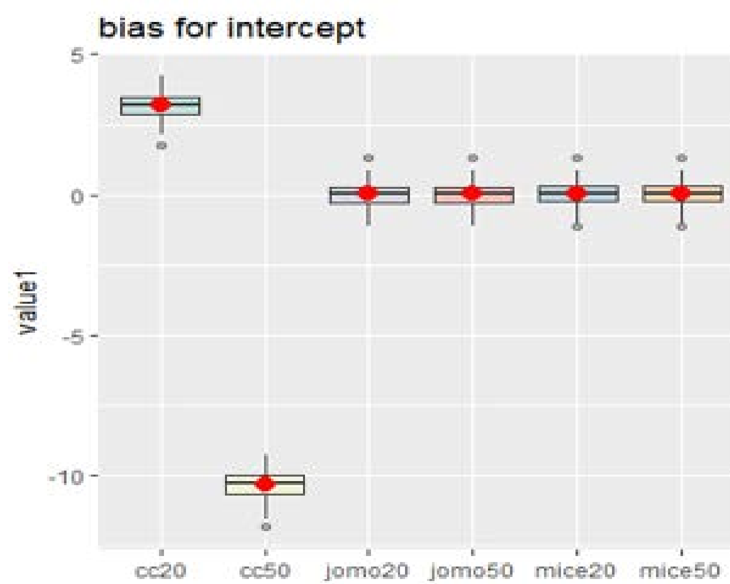
MAR in Xijk & MAR in Zjk



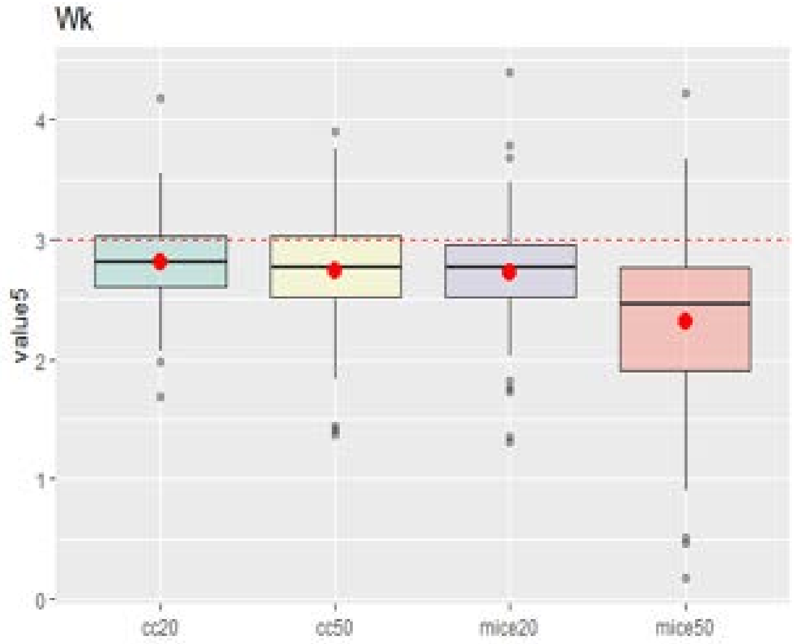
Large variation in the parameter estimated of the intercept, large MSE and bias for intercept was observed when CC was used to analyse for MAR in level 1 variables (X_{ijk}) only.

In analysis for MAR in level 2 variables (Z_{jk}), we observed that all 3 methods performed well across different scenarios (results not shown).

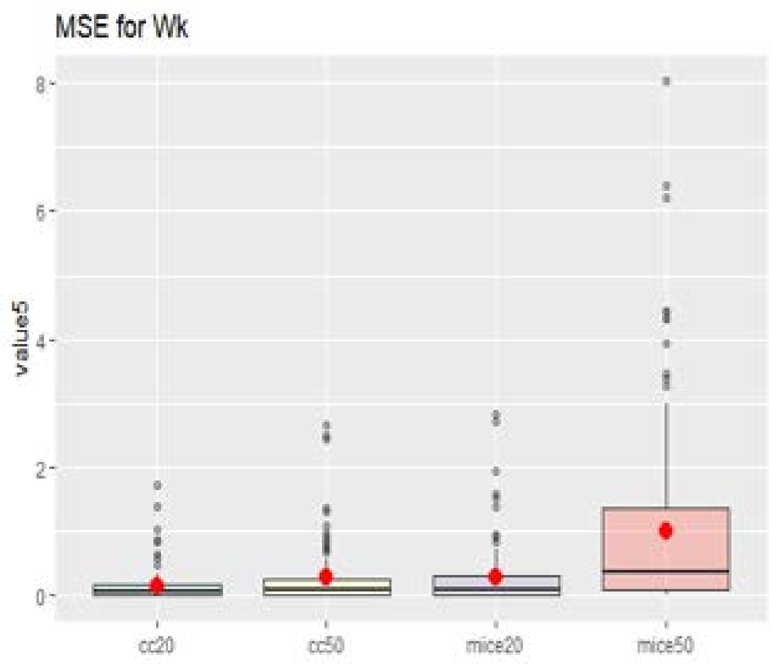




20% and 50% MAR in Wk

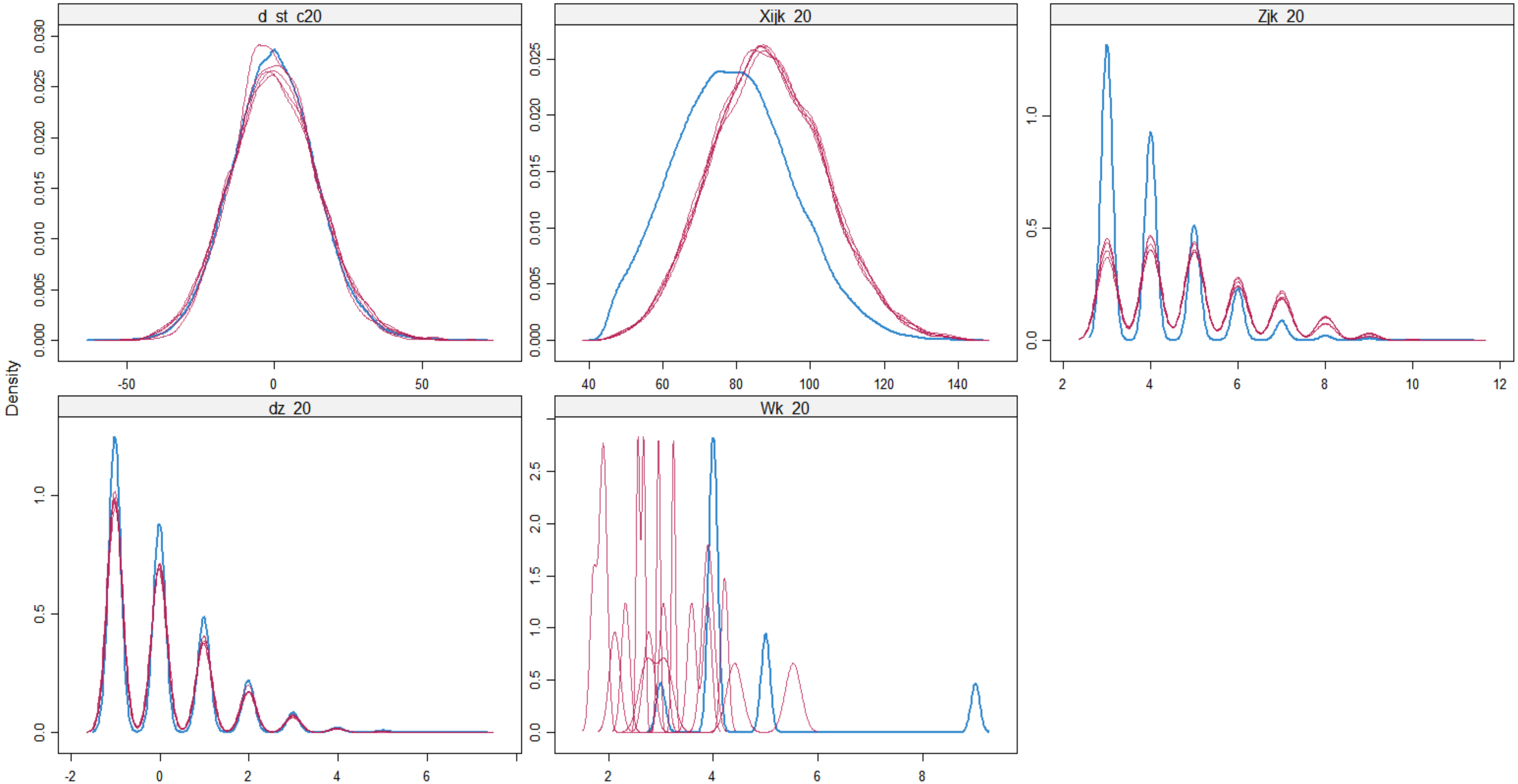


- Two methods were compared- CC and MICE with 20% and 50% MAR in Wk.
- A large variability in MSE was observed for 50% MAR in Wk using MICE

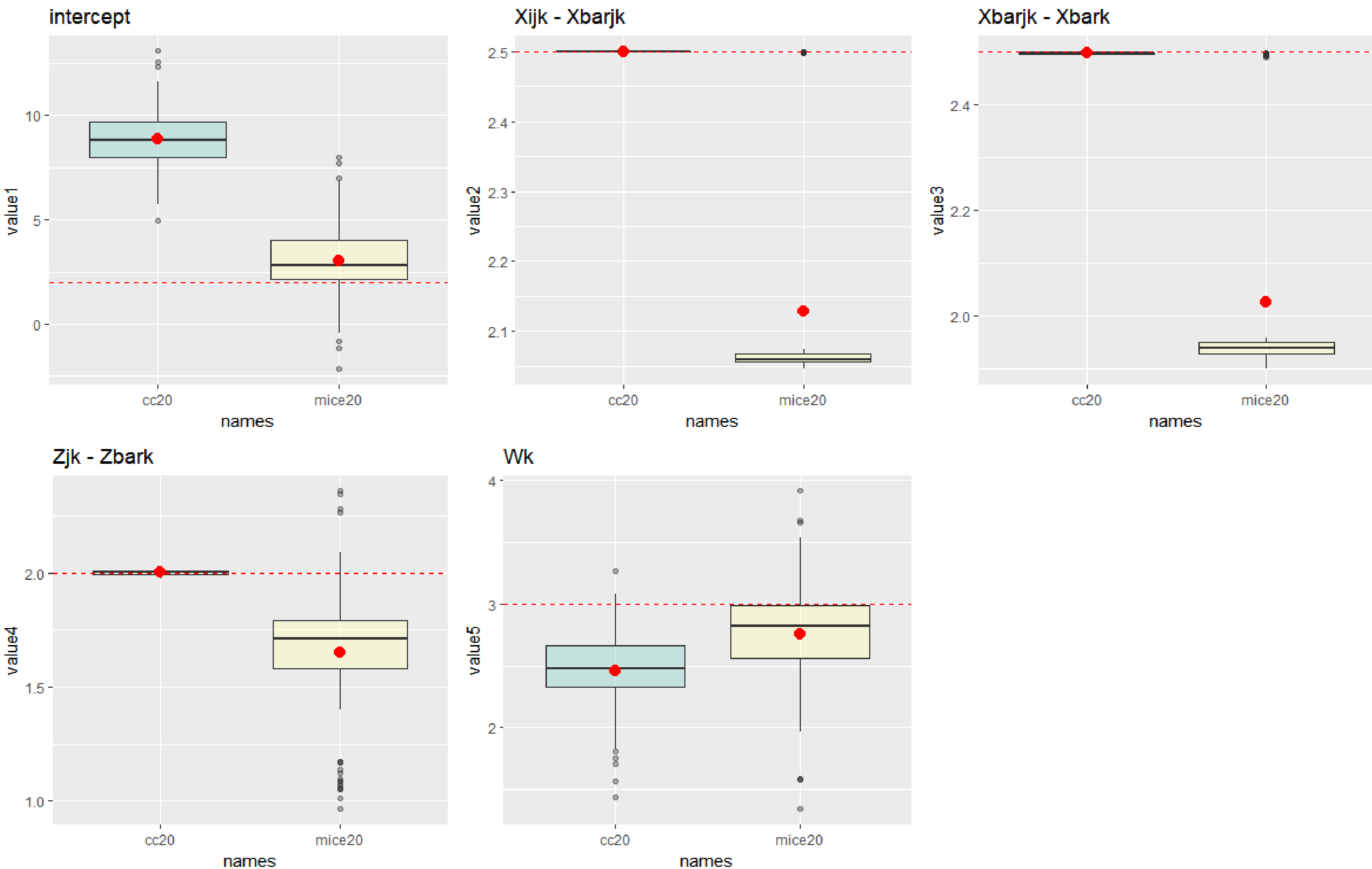


- MI using JoMo for MAR in level 3 variables is currently being investigated

Introducing 20% MAR in X, Z and W



MAR in X, Z & W – Parameter Estimates



Future work

- Replication of MAR in W using JoMo
- Performance of MI procedures with varying number of households/districts and household/district sizes
- Performance of MI with varying ICC values

- Contact us!
- Alice.Richardson@anu.edu.au
Nidhi.Menon@anu.edu.au