

# Recipes for multilevel imputation

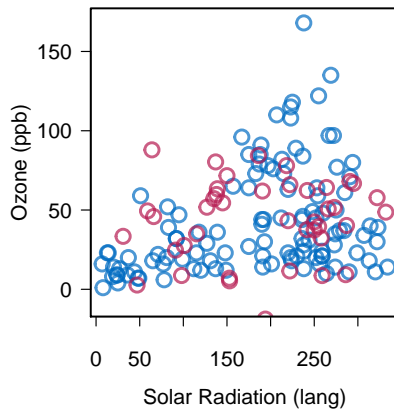
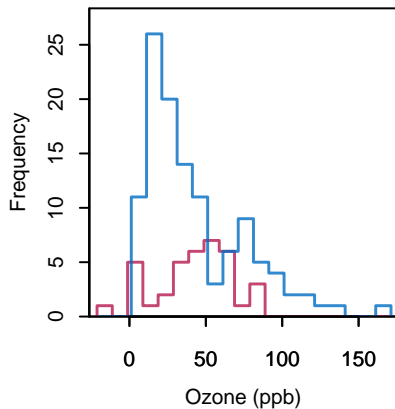
Stef van Buuren (Utrecht University)

April 9, 2019

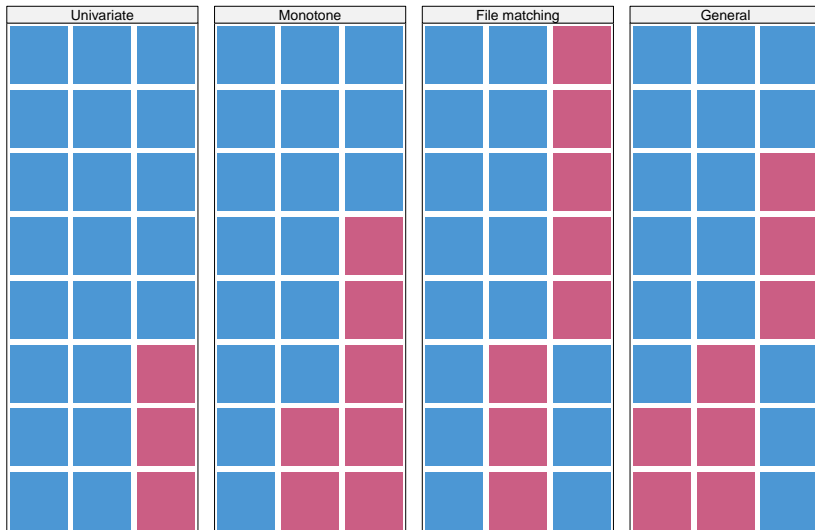
# Main question

*Can we use MICE-algorithm for multilevel data, and if so, how?*

# Stochastic regression imputation



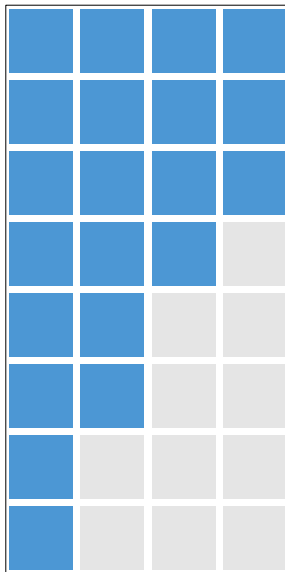
# Missing data patterns, multivariate



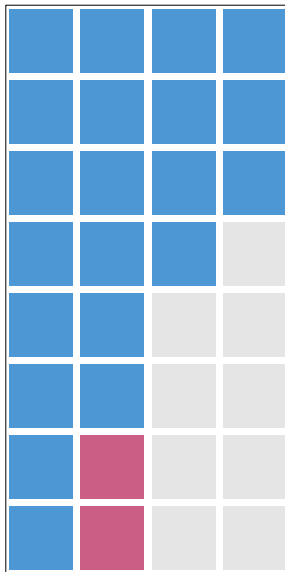
# Three general strategies

- ▶ Monotone data imputation
- ▶ Joint modeling
- ▶ Fully conditional specification (FCS)

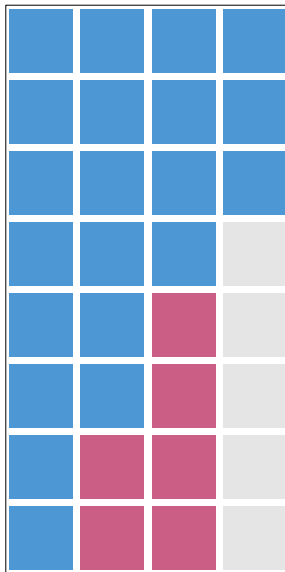
# Imputation of monotone pattern



# Imputation of monotone pattern

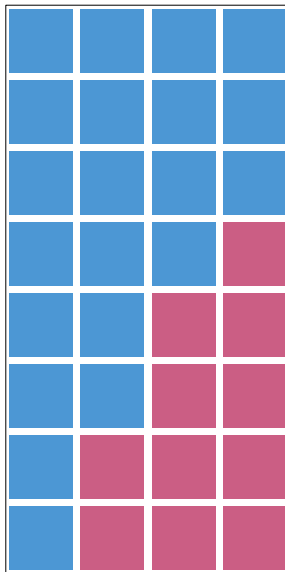


# Imputation of monotone pattern

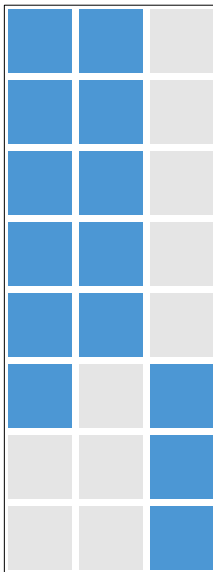




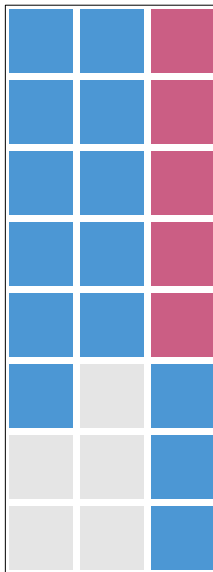
# Imputation of monotone pattern



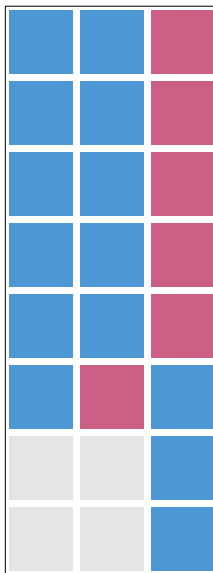
# Imputation by joint modelling



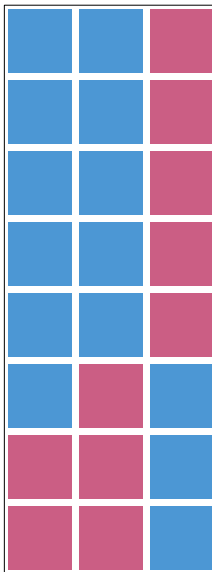
# Imputation by joint modelling



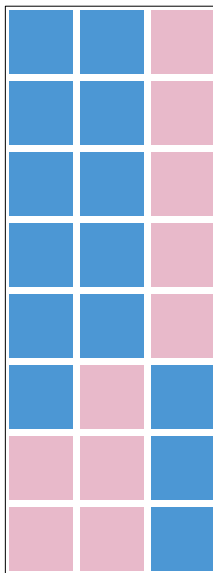
# Imputation by joint modelling



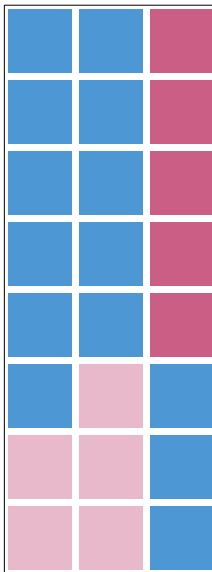
# Imputation by joint modelling



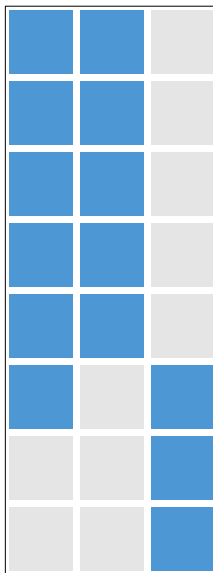
# Imputation by joint modelling - next iteration



# Imputation by joint modelling - next iteration

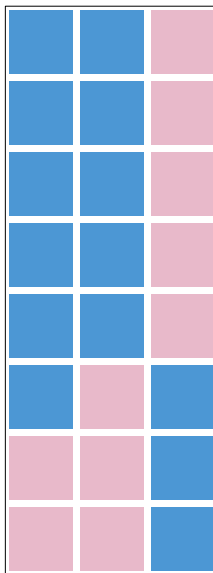


# Imputation by fully conditional specification

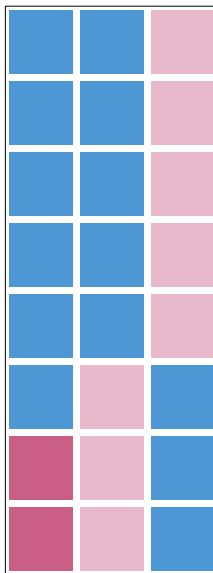




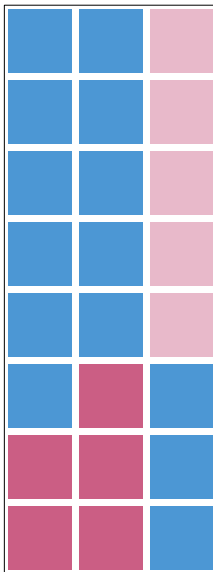
# Imputation by fully conditional specification



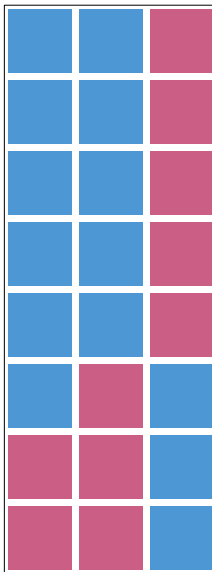
# Imputation by fully conditional specification



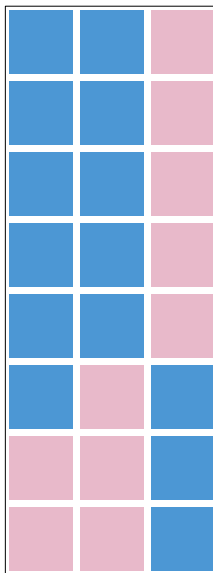
# Imputation by fully conditional specification



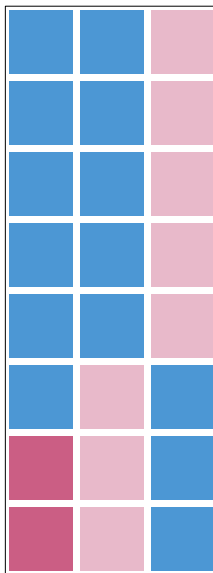
# Imputation by fully conditional specification



# Imputation by fully conditional specification - next iteration



# Imputation by fully conditional specification - next iteration



# Which predictors?

1. Include all variables that appear in the complete-data model
2. Include variables related to the nonresponse
3. Include variables that explain a considerable amount of variance
4. Remove from variables selected in steps 2 and 3 those variables that have too many missing values within the subgroup of incomplete cases

*Does this recipe also apply to multilevel data?*

- ▶ Brandsma and Knuver, Int J Ed Res, 1989.
- ▶ Extensively discussed in Snijders and Bosker (2012), 2nd ed.
- ▶ 4106 pupils, 216 schools, about 4% missing values



## brandsma data subset

```
library(mice)
d <- brandsma[, c("sch", "lpo", "sex", "den")]
head(d, 2)
```

```
##    sch lpo sex den
## 1    1  NA  1   1
## 2    1  50  1   1
```

- ▶ sch: School number, cluster variable,  $C = 216$ ;
- ▶ lpo: Language test post, outcome at pupil level;
- ▶ sex: Sex of pupil, predictor at pupil level (0-1);
- ▶ den: School denomination, predictor at school level (1-4).

# Model of scientific interest

Predict  $lpo$  from the

- ▶ level-1 predictor  $sex$
- ▶ level-2 predictor  $den$

## Level notation - Bryk and Raudenbush (1992)

$$\text{lpo}_{ic} = \beta_{0c} + \beta_{1c}\text{sex}_{ic} + \epsilon_{ic} \quad (1)$$

$$\beta_{0c} = \gamma_{00} + \gamma_{01}\text{den}_c + u_{0c} \quad (2)$$

$$\beta_{1c} = \gamma_{10} \quad (3)$$

- ▶  $\text{lpo}_{ic}$  is the test score of pupil  $i$  in school  $c$
- ▶  $\text{sex}_{ic}$  is the sex of pupil  $i$  in school  $c$
- ▶  $\text{den}_c$  is the religious denomination of school  $c$
- ▶  $\beta_{0c}$  is a random intercept that varies by cluster
- ▶  $\beta_{1c}$  is a sex effect, assumed to be the same across schools.
- ▶  $\epsilon_{ic} \sim N(0, \sigma_\epsilon^2)$  is the within-cluster random residual at the pupil level

# Where are the missings?

In single level data, missingness may be in the outcome and/or in the predictors

With multilevel data, missingness may be in:

1. the outcome variable;
2. the level-1 predictors;
3. the level-2 predictors;
4. the class variable.

## Univariate missing, level-1 outcome

	lpo	sex	den
1			
1			
1			
2			
2			
3			
3			
3			

# Univariate missing, level-1 predictor, sporadically missing

	lpo	sex	den
1			
1			
1			
2			
2			
3			
3			
3			

# Univariate missing, level-1 predictor, systematically missing

	lpo	sex	den
1			
1			
1			
2			
2			
3			
3			
3			

## Univariate missing, level-2 predictor

	lpo	sex	den
1			
1			
1			
2			
2			
3			
3			
3			



# Multivariate missing

	lpo	sex	den
1	Observed	Observed	Missing
1	Missing	Missing	Missing
1	Missing	Observed	Missing
2	Observed	Observed	Missing
2	Observed	Observed	Missing
3	Observed	Missing	Observed
3	Observed	Observed	Observed
3	Missing	Observed	Observed

# Nine challenges in multilevel imputation (1 of 3)

1. For small clusters the within-cluster mean and variance are unreliable estimates, so the choice of the prior distribution becomes critical.
2. For a small number of clusters, it is difficult to estimate the between-cluster variance of the random effects.
3. In applications with systematically missing data, there are no observed values in the cluster, so the cluster location cannot be estimated.

## Nine challenges in multilevel imputation (2 of 3)

4. The variation of the random slopes can be large, and some methods have difficulty handling this.
5. The error variance  $\sigma_{\epsilon}^2$  may differ across clusters (heteroscedasticity), whereas the standard model assumes equal error variances.
6. The residual error distributions can be far from normal, e.g., for categorical data.

## Nine challenges in multilevel imputation (3 of 3)

7. The model may contain aggregates of the level-1 variables, such as cluster means, which need to be taken in account during imputation.
8. The model may contain interactions, or other nonlinear terms.
9. It may not be possible to fit the multilevel model, or there are convergence problems.

See Van Buuren (2018)

# Fully conditional specification

$$\text{lpo}_{ic} \sim N(\beta_0 + \beta_1 \text{den}_c + \beta_2 \text{sex}_{ic} + u_{0c}, \sigma_\epsilon^2) \quad (4)$$

$$\text{sex}_{ic} \sim N(\beta_0 + \beta_1 \text{den}_c + \beta_2 \text{lpo}_{ic} + u_{0c}, \sigma_\epsilon^2) \quad (5)$$

# Theoretical problem with FCS

Conditional expectation of  $\text{sex}_{ic}$  in a random effects model depends on

- ▶  $\text{lpo}_{ic}$ ,
- ▶  $\overline{\text{lpo}}_i$ , the mean of cluster  $i$ , and
- ▶  $n_i$ , the size of cluster  $i$ .

Resche-Rigon & White (2018) suggest the imputation model

- ▶ should incorporate the cluster means of level-1 predictors
- ▶ be heteroscedastic if cluster sizes vary

# Univariate continuous multilevel imputation in mice

Table 7.2: Overview of methods to perform univariate multilevel imputation of continuous data. Each of the methods is available as a function called `mice.impute.[method]` in the specified R package.

Package	Method	Description
<i>Continuous</i>		
mice	2l.lmer	normal, lmer
mice	2l.pan	normal, pan
miceadds	2l.continuous	normal, lmer , blme
micemd	2l.jomo	normal, jomo
micemd	2l.glm.norm	normal, lmer
mice	2l.norm	normal, heteroscedastic
micemd	2l.2stage.norm	normal, heteroscedastic
<i>Generic</i>		
miceadds	2l.pmm	pmm, homoscedastic, lmer
micemd	2l.2stage.pmm	pmm, heteroscedastic, mvmeta

Figure 1

# Univariate binary and count multilevel imputation in `mice`

Table 7.3: Methods to perform univariate multilevel imputation of missing discrete outcomes. Each of the methods is available as a function called `mice.impute.[method]` in the specified R package.

Package	Method	Description
<i>Binary</i>		
<code>mice</code>	<code>2l.bin</code>	logistic, <code>glmer</code>
<code>miceadds</code>	<code>2l.binary</code>	logistic, <code>glmer</code>
<code>micemd</code>	<code>2l.2stage.bin</code>	logistic, <code>mvmeta</code>
<code>micemd</code>	<code>2l.glm.bin</code>	logistic, <code>glmer</code>
<i>Count</i>		
<code>micemd</code>	<code>2l.2stage.pois</code>	Poisson, <code>mvmeta</code>
<code>micemd</code>	<code>2l.glm.pois</code>	Poisson, <code>glmer</code>
<code>countimp</code>	<code>2l.poisson</code>	Poisson, <code>glmmPQL</code>
<code>countimp</code>	<code>2l.nb2</code>	negative binomial, <code>glmmadmb</code>
<code>countimp</code>	<code>2l.zihnb</code>	zero-infl neg bin, <code>glmmadmb</code>

Figure 2



# Univariate level-2 imputation in mice

Table 7.4: Overview of `mice.impute.[method]` functions to perform univariate multilevel imputation.

Package	Method	Description
<i>Level-2</i>		
mice	2lonly.mean	level-2 manifest class mean
miceadds	2l.groupmean	level-2 manifest class mean
miceadds	2l.latentgroupmean	level-2 latent class mean
mice	2lonly.norm	level-2 class normal
mice	2lonly.pmm	level-2 class pmm
miceadds	2lonly.function	level-2 class, generic
miceadds	ml.lmer	$\geq 2$ levels, generic

Figure 3

# General imputation/modeling sequence - START SIMPLE

1. Pick a simple complete-data model
2. Create imputations using an imputation template
3. Check the imputes (convergence/plausibility)
4. Estimate parameters
5. Make complete-data model more realistic, go to 1.

See <https://stefvanbuuren.name/fimd/sec-mlguidelines.html>

# Seven imputation templates, increasing complexity

1. Intercept-only model, missing outcomes
2. Random intercepts, missing level-1 predictor
3. Random intercepts, contextual model
4. Random intercepts, missing level-2 predictor
5. Random intercepts, interactions
6. Random slopes, missing outcomes and predictors
7. Random slopes, interactions

# 1 Intercept-only model, missing outcomes (model)

$$\text{lpo}_{ic} = \beta_{0c} + \epsilon_{ic} \quad (6)$$

$$\beta_{0c} = \gamma_{00} + u_{0c} \quad (7)$$

# 1 Intercept-only model, missing outcomes (imputation)

```
d <- brandsma[, c("sch", "lpo")]
pred <- make.predictorMatrix(d)
pred["lpo", "sch"] <- -2
imp <- mice(d, pred = pred, meth = "2l.pmm", m = 10, maxit
           print = FALSE, seed = 152)
```

# 1 Intercept-only model, missing outcomes (analysis)

```
library(lme4)
```

```
## Loading required package: Matrix
```

```
fit <- with(imp, lmer(lpo ~ (1 | sch), REML = FALSE))  
summary(pool(fit))
```

```
##               estimate std.error statistic    df p.value  
## (Intercept)      40.9      0.322      127 3368      0
```

# 1 Intercept-only model, missing outcomes (variances)

```
library(mitml)
testEstimates(as.mitml.result(fit), var.comp = TRUE)$var.co
```

##	Estimate
## Intercept~~Intercept sch	18.021
## Residual~~Residual	63.306
## ICC sch	0.222

## 2 Random intercepts, missing level-1 (model)

$$\text{lpo}_{ic} = \beta_{0c} + \beta_{1c}\text{iqv}_{ic} + \epsilon_{ic} \quad (8)$$

$$\beta_{0c} = \gamma_{00} + u_{0c} \quad (9)$$

$$\beta_{1c} = \gamma_{10} \quad (10)$$

- ▶ Missing values in both `lpo` and `iqv`



## 2 Random intercepts, missing level-1 (imputation)

```
d <- brandsma[, c("sch", "lpo", "iqv")]
pred <- make.predictorMatrix(d)
pred["lpo", ] <- c(-2, 0, 3)
pred["iqv", ] <- c(-2, 3, 0)
imp <- mice(d, pred = pred, meth = "2l.pmm", seed = 919,
            m = 10, print = FALSE)
```

- ▶ Impute lpo from iqv *and* the cluster means of iqv
- ▶ Impute iqv from lpo *and* the cluster means of lpo
- ▶ Alternative: Use `mitml::panImpute()` or `mitml::jomoImpute()`

## 2 Random intercepts, missing level-1 (predictorMatrix)

```
pred
```

```
##      sch lpo iqv  
## sch    0   1   1  
## lpo   -2   0   3  
## iqv   -2   3   0
```

## 2 Random intercepts, missing level-1 (analysis)

```
fit <- with(imp, lmer(lpo ~ iqv + (1 | sch), REML = FALSE))
summary(pool(fit))
```

	estimate	std.error	statistic	df	p.value
## (Intercept)	40.96	0.2378	172	3337	0
## iqv	2.52	0.0525	48	2127	0

```
testEstimates(as.mitml.result(fit), var.comp = TRUE)$var.co
```

	Estimate
## Intercept~~Intercept sch	9.479
## Residual~~Residual	40.862
## ICC sch	0.188

## 4 Random intercepts, missing level-2 predictor (model)

$$\text{lpo}_{ic} = \beta_{0c} + \beta_{1c}\text{iqv}_{ic} + \epsilon_{ic} \quad (11)$$

$$\beta_{0c} = \gamma_{00} + \gamma_{01}\text{den}_c + u_{0c} \quad (12)$$

$$\beta_{1c} = \gamma_{10} \quad (13)$$

- ▶ Missing values in lpo, iqv and den
- ▶ For den the imputation model uses school level aggregates

## 4 Random intercepts, missing level-2 (imputation)

```
d <- brandsma[, c("sch", "lpo", "iqv", "den")]
meth <- make.method(d)
meth[c("lpo", "iqv", "den")] <- c("2l.pmm", "2l.pmm",
                                  "2lonly.pmm")

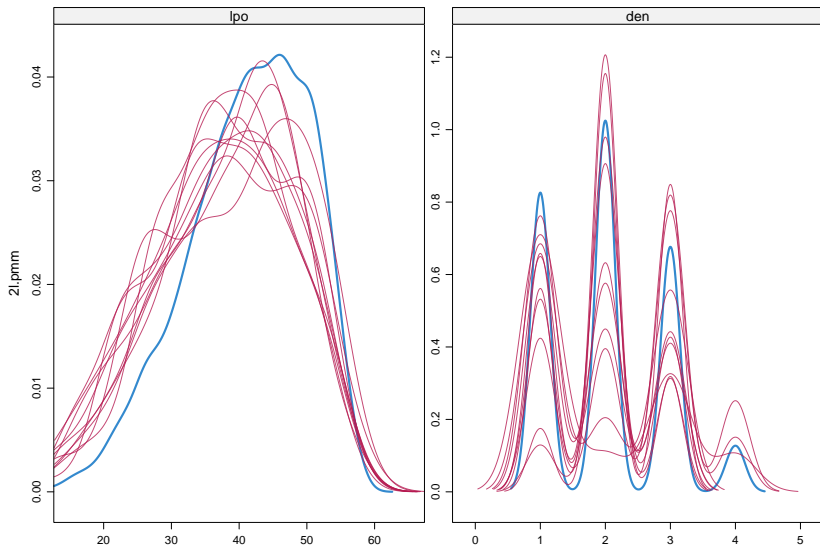
pred <- make.predictorMatrix(d)
pred["lpo", ] <- c(-2, 0, 3, 1)
pred["iqv", ] <- c(-2, 3, 0, 1)
pred["den", ] <- c(-2, 1, 1, 0)
imp <- mice(d, pred = pred, meth = meth, seed = 418,
            m = 10, print = FALSE)
```

## 4 Random intercepts, missing level-2 (predictorMatrix)

```
pred
```

```
##      sch lpo iqv den
## sch    0   1   1   1
## lpo   -2   0   3   1
## iqv   -2   3   0   1
## den   -2   1   1   0
```

## 4 Random intercepts, missing level-2 (density)



## 4 Random intercepts, missing level-2 (analysis)

##	estimate	std.error	statistic	df	p.val
## (Intercept)	40.071	0.4549	88.09	187	0.0000
## iqv	2.516	0.0532	47.34	1242	0.0000
## as.factor(den)2	2.041	0.5925	3.45	430	0.0006
## as.factor(den)3	0.234	0.6519	0.36	285	0.7194
## as.factor(den)4	1.843	1.1642	1.58	1041	0.1137

##	Estimate
## Intercept~~Intercept sch	8.621
## Residual~~Residual	40.761
## ICC sch	0.175



## Recipe: Missing level-1

---

### Recipe for a level-1 target

---

1. Define the most general analytic model
  2. Select a 21 method that imputes close to the data
  3. Include all level-1 variables
  4. Include the disaggregated cluster means of level-1 variables
  5. Include all level-1 interactions implied by analytic model
  6. Include all level-2 predictors
  7. Include all level-2 interactions implied by analytic model
  8. Include all cross-level interactions implied by analytic model
  9. Include predictors related to the missingness and the target
  10. Exclude any terms involving the target
-

## Recipe: Missing level-2

---

### Recipe for a level-2 target

---

1. Define the most general analytic model
  2. Select a 2lonly method that imputes close to the data
  3. Include the cluster means of all level-1 variables
  4. Include the cluster means of all level-1 interactions
  5. Include all level-2 predictors
  6. Include all interactions of level-2 variables
  7. Include predictors related to the missingness and target
  8. Exclude any terms involving the target
-

*Can we use MICE-algorithm for multilevel data, and if so, how?*

- ▶ Hot spot of current research
- ▶ Multilevel imputation: more complex, but doable
- ▶ Start simple, take small steps
- ▶ Build upon templates and modeling recipes
- ▶ Study  
<https://stefvanbuuren.name/fimd/sec-mlguidelines.html>
- ▶ Gain confidence at each step
- ▶ Start playing around. . .