Missing Data Imputation

Applying multiple imputation to survey data harmonized ex-post

Stef van Buuren, University of Utrecht & TNO

December 19, 2019; Building Multi-Source Databases for Comparative Analyses

Two problems in combining datasets

1. Different number of categories
2. Uncollected variables

1. Different number of categories

Trust in government

Control variables	Source trust in parliament scale length	C_TR_PARLI_SRC_SCALE_LENGTH	2 = 2-point scale 4 = 4-point scale 5 = 5-point scale 7 = 7-point scale 10 = 10-point scale 11 = 11-point scale -2 (.b) = not applicable
	Source trust in parliament scale direction	C_TR_PARLI_SRC_ASCEND	0 = descending 1 = ascending -2 (.b) = not applicable
	Source trust in parliament polarity	C_TR_PARLI_SRC_UNIPOLAR	0 = bipolar 1 = unipolar -2 (.b) = not applicable

Method 1: Stretch to finer scale

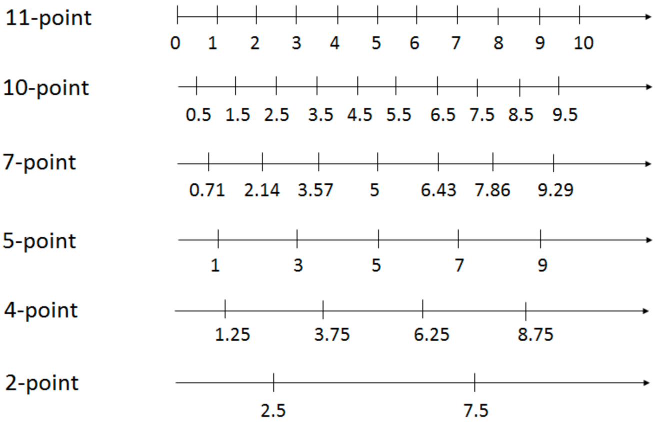


Figure 2. Transformation of source values into the target 0-10 scale

Method 2: Align ranges

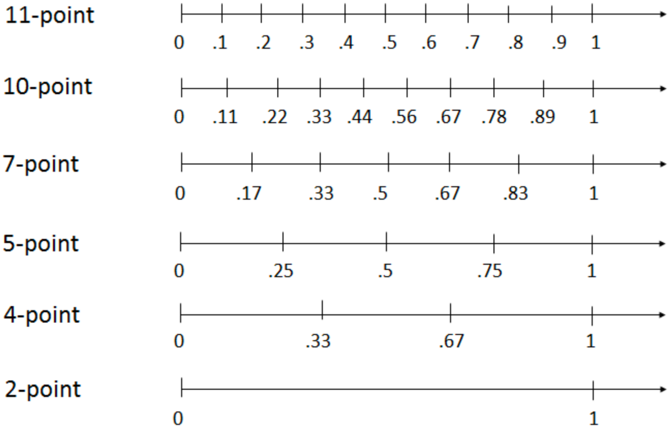


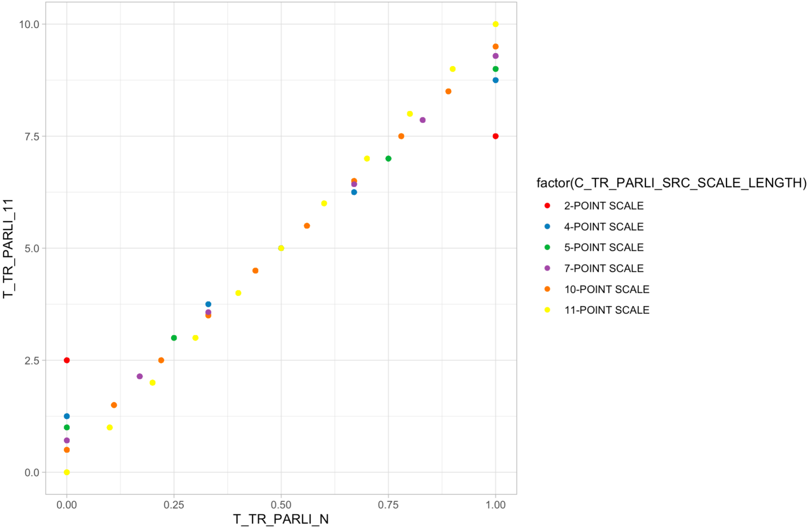
Figure 3. Transformation of source values into the target 0-1 scale

Method 3: Cohortwise transform to uniform

Table 22. Distribution-based transformation; example: TRUST IN PARLIAMENT, LITS/2/PL

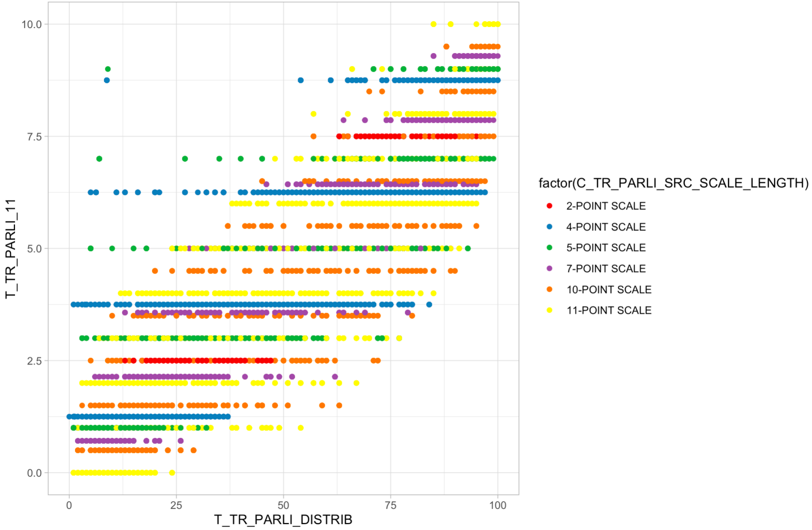
Source value k	**Distribution** X_k	**Cumulative distribution** $\sum_{i=1}^{k} X_i$	$\sum_{i=1}^{k-1} X_i$	$\sum_{i=1}^{k-1} X_i + \frac{X_k}{2}$	**Target value (rounded to integer)**
1	10.68	10.68	0	= 10.68/2 = 5.340	5
2	32.75	43.44	10.68	= 10.68 + 32.75/2 = 27.055	27
3	32.11	75.55	43.44	= 43.44 + 32.11/2 = 59.495	59
4	21.69	97.23	75.55	= 75.55 + 21.69/2 = 86.395	86
5	2.77	100	97.23	= 97.23 + 2.77/2 = 98.615	99

1 vs 2



T_TR_PARLI_11
10.0
7.5
5.0
2.5
0.0
0.00
0.25
0.50
0.75
1.00
T_TR_PARLI_N
factor(C_TR_PARLI_SRC_SCALE_LENGTH)
2-POINT SCALE
4-POINT SCALE
5-POINT SCALE
7-POINT SCALE
10-POINT SCALE
11-POINT SCALE

1 vs 3



10.0
7.5
5.0
2.5
0.0
T_TR_PARLI_11
0
25
50
75
100
T_TR_PARLI_DISTRIB
factor(C_TR_PARLI_SRC_SCALE_LENGTH)
2-POINT SCALE
4-POINT SCALE
5-POINT SCALE
7-POINT SCALE
10-POINT SCALE
11-POINT SCALE

Comments

- Untested assumptions
  - 1/2: equal distance between categories
  - 3: same percentile distribution over cohorts
- Arbitrary, not obvious which to choose
- Does not account for response behaviors
- Impact on conclusions

Levels of equivalence

1. **construct inequivalence**: no equivalent concepts across cohorts
2. **construct equivalence**: same concept is measured, but scales differ
3. **procedural equivalence**: common procedure to measure objects, but there is no underlying unit or ordering in the numbers
4. **unit equivalence**: same units but different anchors
5. **scalar equivalence**: same ratio scale across cohort

Idea

- Generalize transformation one –> many
- Learn relations from the data

Crisp coding

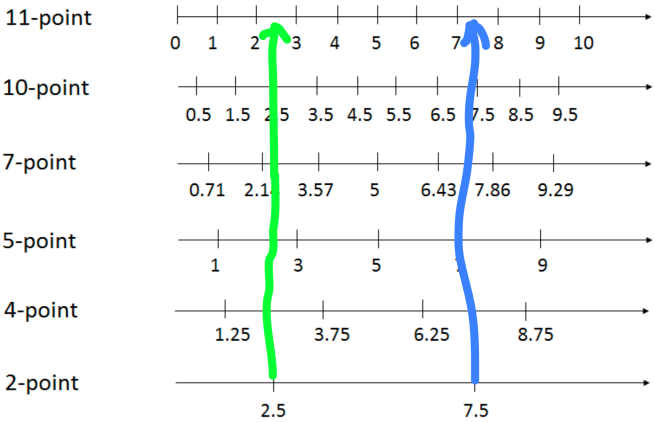


Figure 2. Transformation of source values into the target 0-10 scale

Fuzzy coding

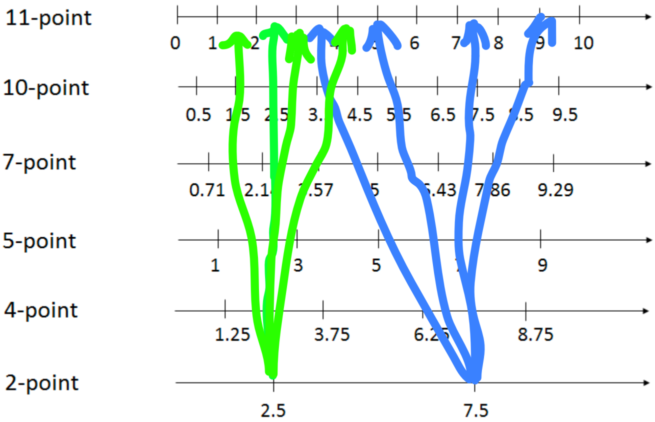


Figure 2. Transformation of source values into the target 0-10 scale

Trust parlement: 0-10 scale, 0-1 scale

0-10	no	yes	no	yes
0	3	0	.02	.00
1	0	0	.00	.00
2	6	1	.05	.01
3	10	0	.08	.00
4	12	1	.09	.01
5	30	8	.24	.05
6	32	12	.25	.07
7	25	54	.20	.33
8	8	44	.06	.27
9	0	33	.00	.20
10	1	8	.01	.05
	127	161	1.00	1.00

Example: Walking disability in two countries

- Uses the `walking` data in `mice`

Item HAQ8 measured in Antonia

Are you able to walk outdoors on flat ground?

Cat	Label	Count
0	Without any difficulty	242
1	With some difficulty	43
2	With much difficulty	15
3	Unable to do	0
NA	Missing	6
	Total	306

Antonia statistic (Mean disability):
$(242 \times 0 + 43 \times 1 + 15 \times 2)/300 = 0.243$

Item GARS9 measured in Belmark

Can you, fully independently, walk outdoors (if necessary with a cane)?

Cat	Label	Count
0	Yes, no difficulty	145
1	Yes, with some difficulty	110
2	Yes, with much difficulty	29
3	No, only with help from others	8
NA	Missing	0
	Total	292

Belmark statistic: proportion no difficulty (PND): $145/292 = 0.50$

Problem

- We want to compare walking problems between Antonia and Belmark
- What to do?

The easy way: Equate all categories

Country	*Mean*	*PND*
Antonia	.24	.80
Belmark	.66	.50
Difference	-.42	.30

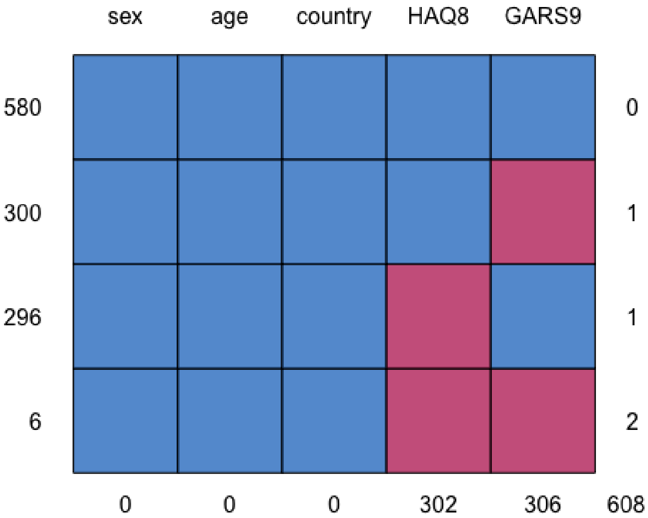
- Both *Mean* and *PND* indicate more walking problems in Belmark
- Differences are large
- Assumes that we can perfectly map *HAQ8* into *GARS9*, and vice versa.
- That is, the correlation is 1.0: Is that realistic?

A third country: Citrus

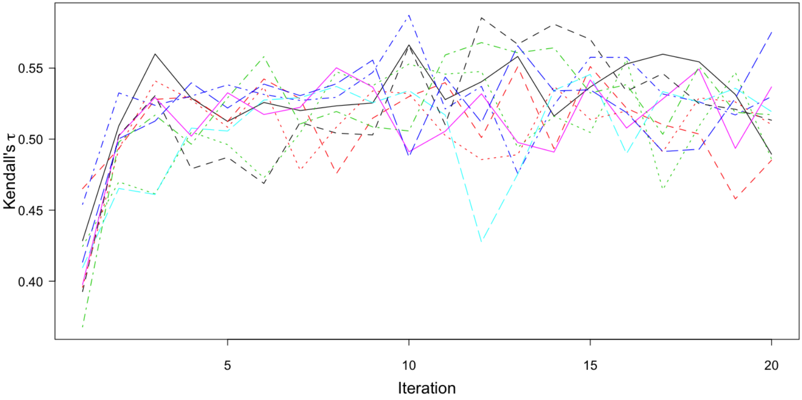
	GARS9				
HAQ8	0	1	2	3	Total
0	256	90	6	4	356
1	26	90	20	0	136
2	6	40	28	10	84
3	0	0	2	2	4
NA	2	0	2	0	4
Total	290	220	58	16	584

- Not symmetric: *HAQ8* appears more difficult than *GARS9*
- Kendall's $\tau = 0.57$, not 1.00
- Are there consequences for the comparison?

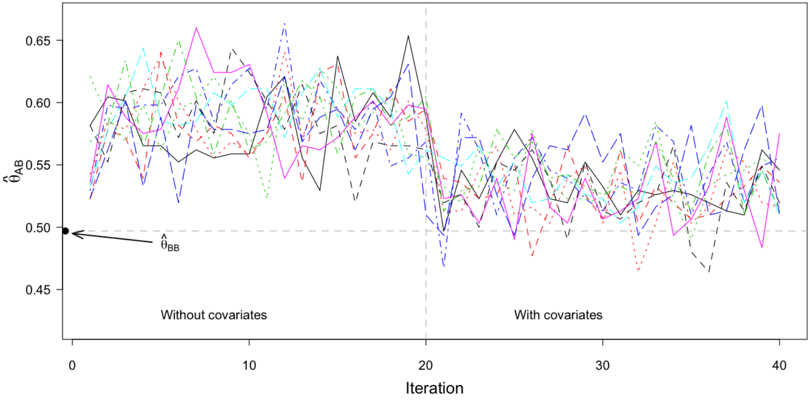
How to impute: data structure



How to impute: Kendall's τ for imputed HAQ8 and GARS9



Result: Proportion No difficulty (*PND*) in Antonia



Results: equating <–> MI <–> MI + age + sex

Country	*Mean*	*PND*	*Mean*	*PND*	*Mean*	*PND*
Antonia	.24	.80	.24	.59	.24	.53
Belmark	.66	.50	.45	.50	.45	.50
Difference	-.42	.30	-.22	.09	-.22	.03

- Antonia is still doing better, but the effects are much smaller
- Correction has more effect on "proportion of no difficulty" (*PND*) (10 times smaller!)

Conclusions

- **Simple equating exaggerates differences between countries**, e.g., $+30$ percent points instead of $+3$ percent points
- Overstated differences may spur inappropriate interventions, sometimes with substantial financial consequences
- Worse problems exist for the different number of categories, popular *crisp recoding*
- Advised remedy: multiple imputation, including covariates
- More detail: https://stefvanbuuren.name/fimd/sec-codingsystems.html
- Code: https://github.com/stefvanbuuren/fimdbook/blob/master/R/fimd.R

2. Uncollected variables

Analysis of individual patient data (IPD) is very popular. It has many advantages over meta-analysis of aggregate data, e.g.,

- Consistent inclusion/exclusion criteria
- Missing data can be treated at the patient level
- Verifies original analysis
- Removal of duplicate subjects
- Consistent correction for confounders

Problem: Studies collect different variables

Missing data in IPD

- **Systematically missing**: Not collected, missing for all in study
- **Sporadically missing**: Collected, but missing for some in study
- Can be at level-1 or level-2 of the analysis

Imputation of IPD data

- In general, we need multilevel imputation models
- Historically, most techniques were suited only for sporadically missing
- Wish to preserve between-study heterogeneity in errors: `mice::mice.impute.2l.norm()`
- More recently, two types of models, level-1:
  - generalization to systematically missing: `mice::mice.impute.2l.lmer()` (Jolani, 2016)
  - 2-stage models: `micemd::mice.impute.2l.2stage.norm()` (Resche-Rigon 2016)

brandsma 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

```r
library(mice)
head(brandsma[, c(1:6, 9:10, 13)], 3)
```

```
##   sch pup   iqv   iqp sex    ses lpr lpo den
## 1   1   1 -1.35 -3.72   1 -17.67  33  NA   1
## 2   1   2  2.15  3.28   1     NA  44  50   1
## 3   1   3  3.15  1.27   0  -4.67  36  46   1
```

brandsma data subset

```r
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)

$$\texttt{lpo}_{ic} = \beta_{0c} + \beta_{1c}\texttt{sex}_{ic} + \epsilon_{ic} \quad (1)$$
$$\beta_{0c} = \gamma_{00} + \gamma_{01}\texttt{den}_c + u_{0c} \quad (2)$$
$$\beta_{1c} = \gamma_{10} \quad (3)$$

- lpo_{ic} is the test score of pupil i in school c
- sex_{ic} is the sex of pupil i in school c
- den_c is the religious denomination of school c
- β_{0c} is a random intercept that varies by cluster
- β_{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

Level 2 equations: interpretation

The first level-2 model

$$\beta_{0c} = \gamma_{00} + \gamma_{01}\mathsf{den}_c + u_{0c},$$

describes the variation in the mean test score between schools as a function of

- the grand mean γ_{00},
- a school-level effect γ_{01} of denomination, and a
- school-level random residual $u_{0c} \sim N(0, \sigma^2_{u_0})$

The second level 2 model

$$\beta_{1c} = \gamma_{10},$$

specifies β_{1c} as a fixed effect equal in value to γ_{10}

Unknown parameters

$$\texttt{lpo}_{ic} = \beta_{0c} + \beta_{1c}\texttt{sex}_{ic} + \epsilon_{ic} \tag{4}$$

$$\beta_{0c} = \gamma_{00} + \gamma_{01}\texttt{den}_c + u_{0c} \tag{5}$$

$$\beta_{1c} = \gamma_{10} \tag{6}$$

The unknowns to be estimated are the fixed parameters:

- γ_{00},
- γ_{01}, and
- γ_{10},

and the variance components:

- σ_ϵ^2 and
- $\sigma_{u_0}^2$.

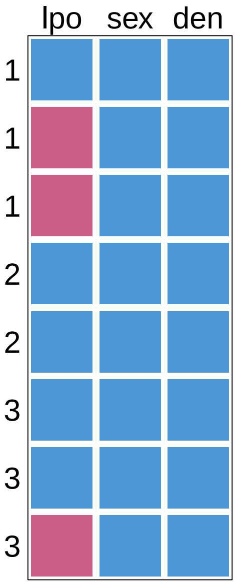
Where are the missings?

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

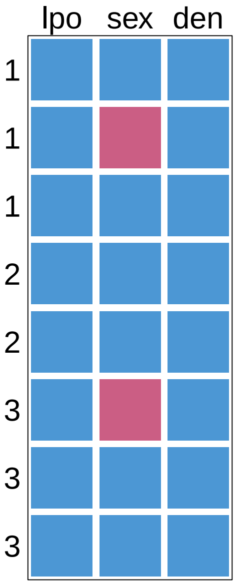
With multilevel data, missingness may be in:

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

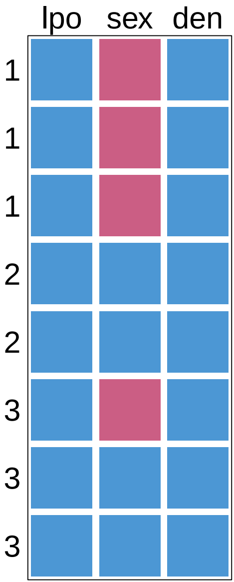
Univariate missing, level-1 outcome



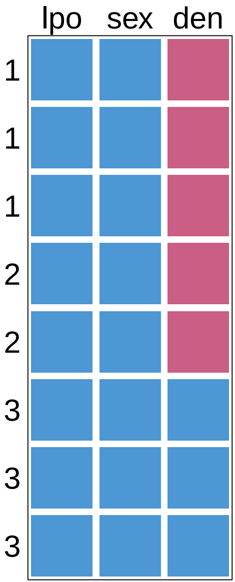
Univariate missing, level-1 predictor, sporadically missing



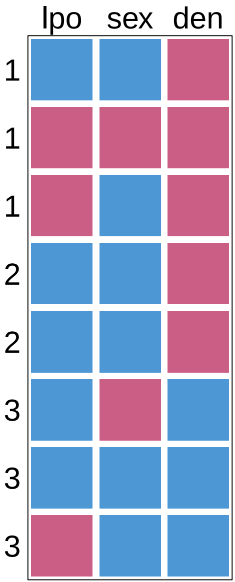
Univariate missing, level-1 predictor, systematically missing



Univariate missing, level-2 predictor



Multivariate missing



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 σ_ϵ^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)

Ad hoc solutions

1. Listwise deletion: Generally not recommended
2. Single-level imputation: Biases ICC downwards.

- Conducting multiple imputation with the wrong model (e.g., single-level methods) can be more hazardous than listwise deletion.

3. Include dummy per cluster: Fixed effects generally unbiased, but the random effects are not. Biases ICC upwards.

Three general strategies

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

Fully conditional specification

$$\dot{\texttt{lpo}}_{ic} \sim N(\beta_0 + \beta_1 \texttt{den}_c + \beta_2 \texttt{sex}_{ic} + u_{0c}, \sigma_\epsilon^2) \tag{7}$$

$$\dot{\texttt{sex}}_{ic} \sim N(\beta_0 + \beta_1 \texttt{den}_c + \beta_2 \texttt{lpo}_{ic} + u_{0c}, \sigma_\epsilon^2) \tag{8}$$

Theoretical problem with FCS

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

- $\texttt{lpo}_{ic}$,
- $\overline{\texttt{lpo}}_i$, 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

Methods for 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
Continuous		
`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
Generic		
`miceadds`	`2l.pmm`	pmm, homoscedastic, `lmer`
`micemd`	`2l.2stage.pmm`	pmm, heteroscedastic, `mvmeta`

Methods for 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
Binary		
`mice`	`2l.bin`	logistic, `glmer`
`miceadds`	`2l.binary`	logistic, `glmer`
`micemd`	`2l.2stage.bin`	logistic, `mvmeta`
`micemd`	`2l.glm.bin`	logistic, `glmer`
Count		
`micemd`	`2l.2stage.pois`	Poisson, `mvmeta`
`micemd`	`2l.glm.pois`	Poisson, `glmer`
`countimp`	`2l.poisson`	Poisson, `glmmPQL`
`countimp`	`2l.nb2`	negative binomial, `glmmadmb`
`countimp`	`2l.zihnb`	zero-infl neg bin, `glmmadmb`

Methods for multilevel imputation in `mice`

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

Package	Method	Description
Level-2		
`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`	≥ 2 levels, generic

In practice: start simple, empty model

```r
d <- brandsma[, c("sch", "lpo")]
pred <- make.predictorMatrix(d)
pred["lpo", "sch"] <- -2
imp <- mice(d, pred = pred, meth = "2l.pmm", m = 10,
            maxit = 1, print = FALSE, seed = 152)
```

Analysis

```r
library(lme4)
```

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

```r
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
```

Variance components

```r
library(mitml)
```

```
## *** This is beta software. Please report any bugs!
## *** See the NEWS file for recent changes.
```

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

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

Now start adding model terms

https://stefvanbuuren.name/fimd/sec-mlguidelines.html

Recipe: Missing level-1

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

Uncollected variables: conclusion

- No need restrict analysis to least common denominator
- Impute systematically missing data with multilevel imputation model
  - Either 2-stage or generalized multilevel
- Technically still challenging, but doable (in `Stata` or `R`)
- More detail: https://stefvanbuuren.name/fimd/ch-multilevel.html

Wrap up

1. Different number of categories
2. Uncollected variables

- I believe multiple imputation provides substantial progress over
  - ad-hoc recoding strategies
  - restriction to observed data
- Of course, we always need MAR assumptions, but these are often natural for combined data
- Still experimental, more experience is needed
- Long-term vision: data-combination as a **information translation service** for distributed data