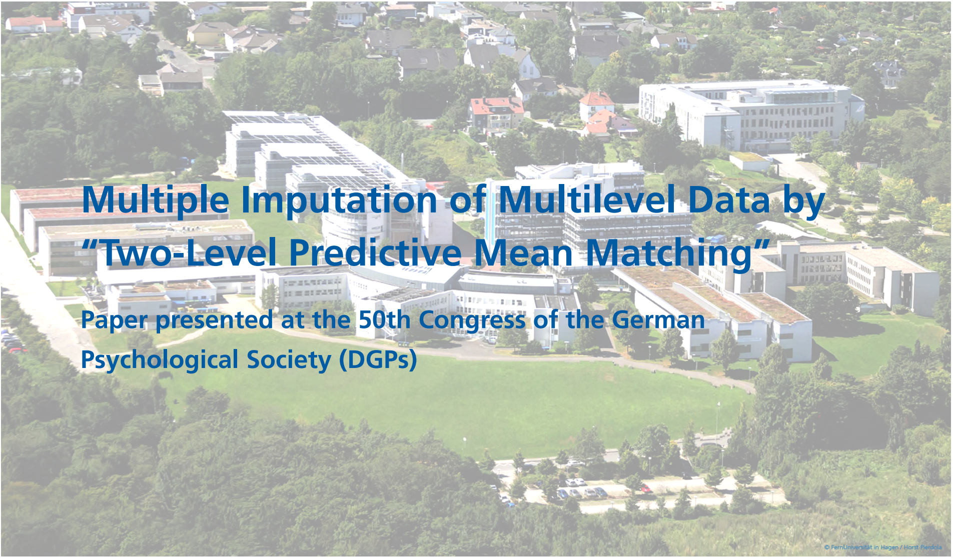
FernUniversität in Hagen

Institute of Psychology



© FernUniversität in Hagen / Horst Pierdola

Multiple Imputation of Multilevel Data by "Two-Level Predictive Mean Matching"

Paper presented at the 50th Congress of the German Psychological Society (DGPs)

Psychological Methods and Evaluation
Kristian Kleinke

Overview / Motivation

- Predictive mean matching (PMM) is a robust state-of-the-art hot deck imputation procedure and the default setting in many MI software packages.
- Currently existing two-level MI procedures are not very robust against violations of model assumptions (such as normality / homoscedasticity)
- For the time being, van Buuren (2013) recommends to use flat-file PMM even for the imputation of clustered or panel data. He also suggests to generalize the PMM idea to multilevel data.
- I propose a two-level predictive mean matching procedure to impute incomplete clustered data.

Multiple imputation – the general idea

3 steps:

1. impute
2. analyse
3. combine

Predictive mean matching

- PMM is one of the standard techniques to create the multiple imputations
- PMM is usually applied within a chained equations MI setting (e.g. van Buuren & Groothuis-Oudshoorn, 2011)
- PMM imputes an **observed value**, whose value predicted by a linear regression model is among a set of *k* values (the so-called donor pool) closest to the value predicted for the missing one

Predictive Mean Matching – advantages and pitfalls

van Buuren (2012, Chap. 3): “The method works best with large samples, and provides imputations that possess many characteristics of the complete data. Predictive mean matching cannot be used to extrapolate beyond the range of the data, or to interpolate within the range of the data if the data at the interior are sparse. Also, it may not perform well with small datasets. Bearing these points in mind, predictive mean matching is a great all-around method with exceptional properties” (p.74)

Two-level PMM – The algorithm

Let y be an incomplete dependent variable in a two-level regression model.

1. Fit two-level linear mixed effects model (using the available data) and estimate model parameters θ
2. **Bayesian regression** (cf Rubin, 1987, p. 169): Draw new parameters θ^* from the proper posterior. **Alternative**: use **approximate Bayesian bootstrap** and compute new parameters θ^*
3. Compute predicted values for the observed part $\hat{y}_{obs}$ using parameters θ and for the missing part $\hat{y}_{mis}$ using parameters θ^* [1]
4. compute all distances: $\delta_{ij} = |\hat{y}_i^{obs} - \hat{y}_j^{mis}|$ for each incomplete case in y
5. For each missing value in y, find k observations[2] with closest predicted values, randomly sample one of these, and take its observed value in y as the imputation

[1] with $\hat{y}_{obs}$ being the predicted value of y based on the subset of predictors X_{obs} from the imputation model that has complete y values, and $\hat{y}_{mis}$ being the predicted value of y, based on the subset of predictors X_{mis} from the imputation model that has missing y values.

[2] k can be set by the user; default is $k = 5$

Monte Carlo Simulations – Overview

3 Simulations

- **Simulation 1 & 2: large sample scenarios**
  50 groups with group sizes $n_g > 100$
- **Simulation 3: "smaller" sample scenario**
  50 groups with $n_g = 30$; $N = 1500$

(Parametric) assumptions of the multilevel regression models were met. Missing data were introduced in the following way:

$$
\begin{aligned}
p_{ij} &= \text{invlogit}(I + CM) \\
v_{ij} &\sim \mathcal{U}(0, 1) \\
y_{ij} &= \text{NA, if } v_{ij} < p_{ij}.
\end{aligned}
$$

Monte Carlo Simulations

Simulation 1:

- 1 individual level predictor, 1 group level predictor
- 50 groups with sample sizes between 100 and 400; $N = 13350$
- MAR missingness in y depending on **group level** predictor
- 30.39% missing data in y
- average ICC: .23

Simulation 2:

- 1 individual level predictor, 1 group level predictor
- 50 groups with sample sizes between 100 and 400; $N = 12917$
- MAR missingness in y depending on **individual level** predictor
- 39.79% missing data in y
- average ICC: .23

Imputation setup: $m = 5$; $k = 5$

Table 1: Simulation 1 – Results

	β_0	β_1	β_2	τ_{00}	τ_{11}	τ_{01}
Q	1.000	0.750	0.500	0.500	0.300	0
	Complete Data					
$\widehat{Q}$	1.000	0.751	0.504	0.487	0.295	-0.001
	2l.pmm					
$\widehat{Q}$	1.001	0.753	0.506	0.487	0.294	0.002
BIAS	-0.001	-0.003	-0.006	0.013	0.006	-0.002
$\mathrm{SD}_{\widehat{Q}}$	0.071	0.045	0.077	0.050	0.032	0.159
CR	95.900	95.100	94.900	NA	NA	NA
WID	0.298	0.179	0.306	NA	NA	NA
	2l.norm					
$\widehat{Q}$	1.001	0.751	0.504	0.490	0.296	-0.001
BIAS	-0.001	-0.001	-0.004	0.010	0.004	0.001
$\mathrm{SD}_{\widehat{Q}}$	0.071	0.044	0.076	0.049	0.032	0.155
CR	94.300	93.800	93.400	NA	NA	NA
WID	0.277	0.171	0.279	NA	NA	NA
	2l.pan					
$\widehat{Q}$	1.001	0.751	0.504	0.487	0.296	0.000
BIAS	-0.001	-0.001	-0.004	0.013	0.004	0.000
$\mathrm{SD}_{\widehat{Q}}$	0.071	0.044	0.076	0.049	0.031	0.157
CR	94.300	93.800	93.400	NA	NA	NA
WID	0.277	0.171	0.279	NA	NA	NA

Table 2: Simulation 2 – Results

	β_0	β_1	β_2	τ_{00}	τ_{11}	τ_{01}
Q	1.000	0.750	0.500	0.500	0.300	-0.200
			Complete Data			
$\widehat{Q}$	1.001	0.751	0.501	0.486	0.297	-0.197
			2l.pmm			
$\widehat{Q}$	1.004	0.754	0.504	0.486	0.292	-0.212
BIAS	-0.004	-0.004	-0.004	0.014	0.008	0.012
$SD_{\widehat{Q}}$	0.073	0.045	0.076	0.052	0.034	0.159
CR	96.000	96.700	95.100	NA	NA	NA
WID	0.305	0.196	0.308	NA	NA	NA
			2l.norm			
$\widehat{Q}$	1.002	0.751	0.495	0.491	0.296	-0.181
BIAS	-0.002	-0.001	0.005	0.009	0.004	-0.019
$SD_{\widehat{Q}}$	0.072	0.044	0.075	0.052	0.034	0.150
CR	94.000	95.500	93.300	NA	NA	NA
WID	0.279	0.173	0.277	NA	NA	NA
			2l.pan			
$\widehat{Q}$	1.002	0.751	0.501	0.487	0.298	-0.191
BIAS	-0.002	-0.001	-0.001	0.013	0.002	-0.009
$SD_{\widehat{Q}}$	0.072	0.044	0.074	0.052	0.033	0.153
CR	93.800	95.200	93.900	NA	NA	NA
WID	0.277	0.174	0.274	NA	NA	NA

Monte Carlo Simulations

Simulation 3:

- “smaller” sample
- 1 individual level predictor, 1 group level predictor
- 50 groups, constant group size $n_g = 30$
- $N = 1500$
- average ICC: .23
- 30.33% missing data in y, depending on individual level predictor
- $m = 5$; $k = 5$

Table 3: Simulation 3 – Results

	β_0	β_1	β_2	τ_{00}	τ_{11}	τ_{01}
Q	1.000	0.750	0.500	0.500	0.300	0
			Complete Data			
$\widehat{Q}$	1.001	0.750	0.499	0.487	0.294	-0.003
			2l.pmm			
$\widehat{Q}$	1.010	0.766	0.508	0.487	0.280	-0.050
BIAS	-0.010	-0.016	-0.008	0.013	0.020	0.050
$\text{SD}_{\widehat{Q}}$	0.080	0.058	0.080	0.062	0.053	0.228
CR	94.400	93.700	94.400	NA	NA	NA
WID	0.321	0.232	0.320	NA	NA	NA
			2l.norm			
$\widehat{Q}$	1.000	0.750	0.500	0.504	0.297	0.037
BIAS	0.000	0.000	0.000	-0.004	0.003	-0.037
$\text{SD}_{\widehat{Q}}$	0.079	0.057	0.079	0.057	0.052	0.199
CR	94.000	95.000	95.500	NA	NA	NA
WID	0.307	0.229	0.307	NA	NA	NA
			2l.pan			
$\widehat{Q}$	1.000	0.750	0.500	0.490	0.308	0.017
BIAS	0.000	0.000	0.000	0.010	-0.008	-0.017
$\text{SD}_{\widehat{Q}}$	0.079	0.057	0.079	0.060	0.045	0.196
CR	94.000	95.000	95.500	NA	NA	NA
WID	0.307	0.229	0.307	NA	NA	NA

Summary and future research. . .

- **two-level predictive mean matching works as well as currently available two-level procedures, when model assumptions are fully met**
- I assume that two-level PMM will be more robust against violations of (distributional) assumptions in comparison to currently existing imputation procedures – this however remains to be tested.
- Implement more flexible and adaptive donor selection strategies like the ones proposed by Schenker & Taylor (1996) or Siddique & Belin (2008)
- For very large data sets, the algorithm is quite slow (!). Test some of the ideas (fast / partitioned PMM) proposed by Vink, Lazendic, & van Buuren (2015) for large data sets.

Contact

Dr. Kristian Kleinke

University of Hagen
Institute of Psychology
Universitätsstr. 33
D–58097 Hagen

http://e.feu.de/kleinke
kristian.kleinke@feu.de

References

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York: Wiley.

Schenker, N., & Taylor, J. M. (1996). Partially parametric techniques for multiple imputation. *Computational statistics & data analysis*, *22*(4), 425–446.

Siddique, J., & Belin, T. R. (2008). Multiple imputation using an iterative hot-deck with distance-based donor selection. *Statistics in medicine*, *27*(1), 83–102. doi: 10.1002/sim.3001

van Buuren, S. (2012). *Flexible imputation of missing data*. Boca Raton: Chapman & Hall / CRC.

van Buuren, S. (2013, May). *Multiple imputation of multilevel data.* Paper presented at the conference on Recent advances in multiple imputation, with emphasis on dealing with deviations from MAR or exchangeability, Utrecht, NL.

van Buuren, S., & Groothuis-Oudshoorn, K. (2011). MICE: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, *45*(3), 1–67.

Vink, G., Lazendic, G., & van Buuren, S. (2015). Partitioned predictive mean matching as a multilevel imputation technique. *Psychological Test and Assessment Modeling*, *57*(4), 577–594.