Can Missing Patterns in Covariates Improve Imputation for Missing Data?

BigSurv18, Barcelona, Spain

Micha Fischer, Felicitas Mittereder
University of Michigan
Program in Survey Methodology

October 27, 2018

Background

Most (survey) data sets have missing data:

Treatment:

- Imputation of plausible values to receive a data set without missings (e.g., sequential imputation step)

Problems:

- Bias due to oversimplified models in the sequential imputation step
- Information that “respondent did not answer a question” is lost
- If missing data mechanism is Missing Not At Random (MNAR), the item missing pattern can be informative for the imputed values

Basic Idea

- Include "Missing" as own category in imputation model to improve imputation accuracy and therefore estimators from survey data
- Tree-based methods (e.g., random forest) can incorporate this additional information and account for complex interactions, (Doove, Van Buuren, and Dusseldorp 2014)
- Increasing efficiency by skipping sequential imputation steps

$\Rightarrow$ Easy to implement in current software

Previous Research

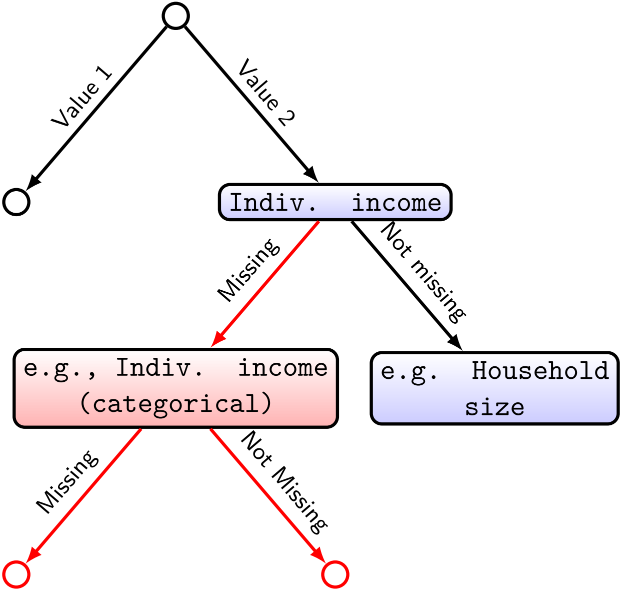
- Loh et al. (2018) use missing values in a regression and classification tree (GUIDE) to impute missing values
- Ding and Simonoff (2010) show that random forest can handle incomplete covariates by coding "missing" as its own category

Potentially New Approach

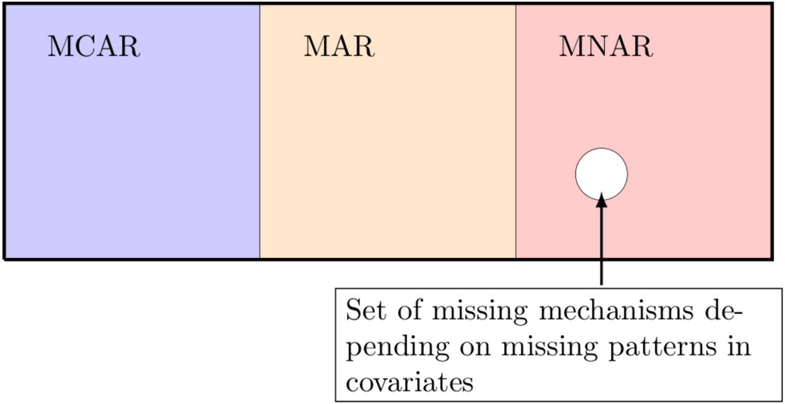
- Combine both approaches, "tree-based imputation" and "missing as its own category" with CART and random forest:
  - Categorical covariates: Treat "Missing" as its own category
  - Continuous covariates: Code "Missing" to arbitrary value "(far) away" from the actual data

⇒ One kind of pattern mixture model which partition data by patterns of missing values (Little 1993)

Example: Potential Part of Tree for Imputing Household Income



Assumption on Missing Data - Venn Diagram:



Theoretical Considerations

- Little (2019 - forthcomming):
  MAR (standard) vs. specific MNAR assumption in Loh et al (2018)
- Many covariates with missing values: potential advantage due to more available data
- Reasonable when missing due to "not applicable" cases

Simulation Study - Set Up (1)

- Generating covariables: $X_i \sim N(0,1)$ and $Z_i \sim N(0,1)$
- Generating response indicators for X and Z:

$$R_{Z,i} \sim Ber(p_Z) \text{ and } R_{Z,i} = \begin{cases} 1 \text{ for } p_Z \geq u_{Z,i}, \\ 0 \text{ for } p_Z < u_{Z,i} \end{cases}$$

where $u_{Z,i} \sim Unif(0,1)$

Simulation Study - Set Up (2)

- Generating outcome variable Y:
  $Y_i = \beta_0 + \beta_1 X_i + \beta_2 Z_i + \beta_3 R_{Z,i} + \beta_4 X_i R_{Z,i} + \epsilon_i$
  where $\epsilon_i \sim N(0,1)$
- Generating response indicator for Y:
  $P(R_{Y,i} = 1) = logit^{-1}(p_Y + \delta_1 X_i + \delta_2 Z_i + \delta_3 R_{Z,i} + \delta_4 Y_i)$

with p_Y as the baseline response rate,

$$\text{and } R_{Y,i} = \begin{cases} 1 \text{ for } P(R_{Y,i} = 1) \geq u_{Y,i}, \\ 0 \text{ for } P(R_{Y,i} = 1) < u_{Y,i} \end{cases}$$

where $u_{Y,i} \sim Unif(0,1)$

$$\Rightarrow Y_{\text{obs},i} = \begin{cases} Y_i \text{ if } R_{Y,i} = 1, \\ missing \text{ if } R_{Y,i} = 0 \end{cases} \qquad Z_{\text{obs},i} = \begin{cases} Z_i \text{ if } R_{Z,i} = 1, \\ missing \text{ if } R_{Z,i} = 0 \end{cases}$$

Simulation Study - Data Structure

Table 1: Resulting data structure

Y_{obs}	X_{obs}	Z_{obs}
Y	X	Z
Y	X	miss
miss	X	Z
miss	X	miss

Simulation Study - Procedure

- Single imputation with linear models, CART and random forest using $X_{obs,i}, Y_{obs,i}, Z_{obs,i}$
- Additionally, CART and random forest using "Missing" information as own category
- Assessment on RMSE of regression coefficiences after imputation in a substantive model - Here:

$$Y_{imp,i} \sim X_{obs,i} + Z_{imp,i}$$

Simulation - Parameters

Table 2: Implemented parameter values

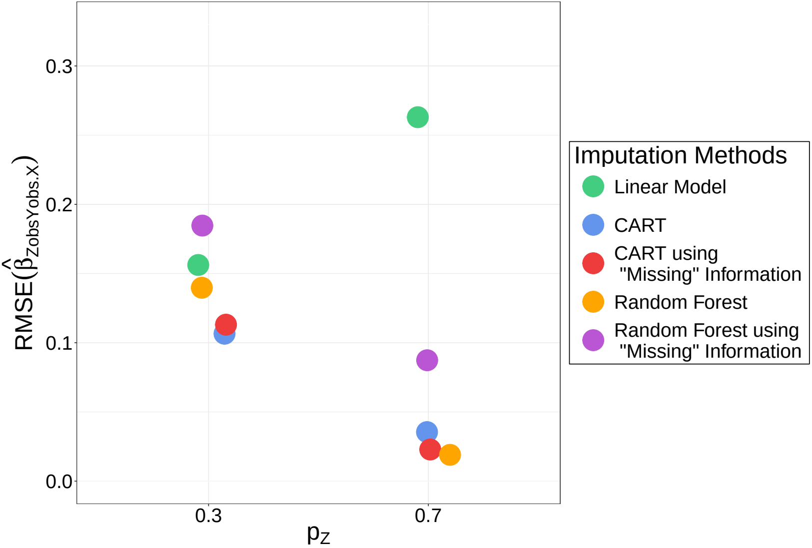
Y	Parameter	β_1	β_2	β_3	β_4
	Values	1	1	$\{0; 2\}$	$\{0; 2\}$
Response propenstiy	Parameter	δ_1	δ_2	δ_3	δ_4
	Values	1	$\{0; 1\}$	$\{0; 1\}$	$\{0; 1\}$
Baseline response	Parameter	p_Y	p_Z		
	Values	0.5	$\{0.3; 0.7\}$		

$\Rightarrow$ MAR situation if $\beta_4 = \delta_4 = 0$

$\Rightarrow$ MNAR situation if $\beta_4 \neq 0 \vee \delta_4 \neq 0$

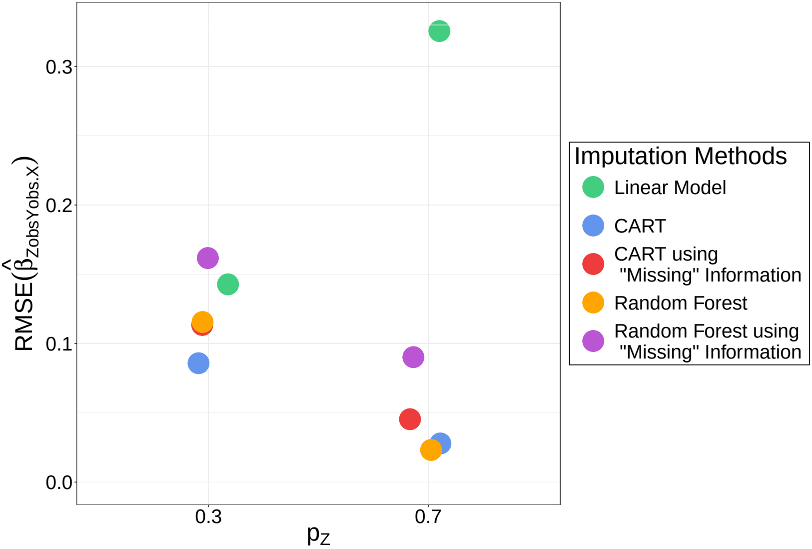
Simulation Results - MAR

Parameter values: $\beta_3 = 2$, $\beta_4 = 0$ and $\delta_2 = \delta_4 = 0$, $\delta_3 = 1$



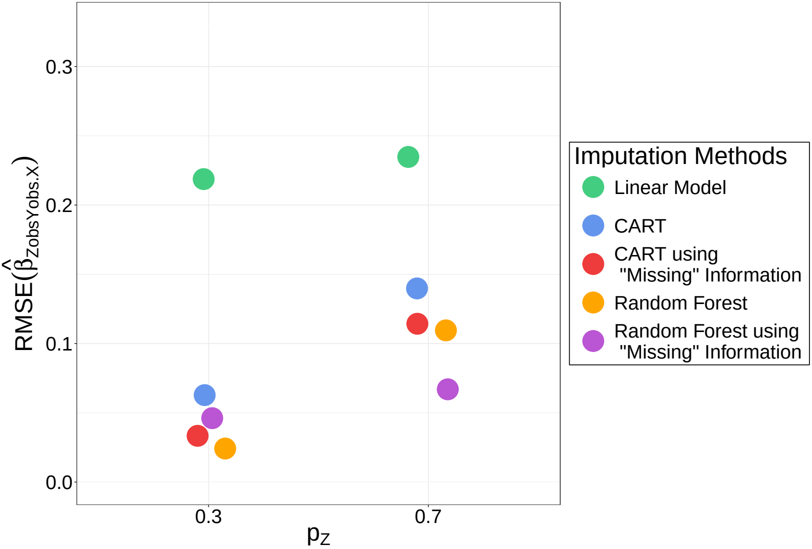
Simulation Results - MNAR

Parameter values: $\beta_3 = \beta_4 = 2$ and $\delta_2 = \delta_4 = 0$, $\delta_3 = 1$



Simulation Results - MNAR

Parameter values: $\beta_3 = \beta_4 = 2$ and $\delta_2 = \delta_3 = \delta_4 = 1$



Future Research

1. Simulation extension for non-normal distributed variables (e.g., binary variables)
2. Evaluation on survey data linked to administrative records
3. Accounting for imputation uncertainty in variance estimation

Thank you for your attention!

Any questions?

fmitter@umich.edu
michaf@umich.edu

References

Ding, Yufeng, and Jeffrey S Simonoff. 2010. "An Investigation of Missing Data Methods for Classification Trees Applied to Binary Response Data." *Journal of Machine Learning Research* 11 (Jan): 131–70.

Doove, LL, Stef Van Buuren, and Elise Dusseldorp. 2014. "Recursive Partitioning for Missing Data Imputation in the Presence of Interaction Effects." *Computational Statistics & Data Analysis* 72. Elsevier: 92–104.

Little, Roderick. 1993. "Pattern-Mixture Models for Multivariate Incomplete Data." *Journal of the American Statistical Association* 88 (421). Taylor & Francis Group: 125–34.

———. 2019. "On Algorithmic and Modeling Approaches to Imputation in Large Data Sets." *Statistica Sinica*.

Loh, Wei-Yin, John Eltinge, Moon Jung Cho, and Yuanzhi Li. 2018. "CLASSIFICATION and Regression Trees and Forests for Incomplete Data from Sample Surveys." *Statistica Sinica*.