




Causal inference and matching

Samy Alsharani, Alodie Boissonnet,
Charlotte, Gallezot, Gauthier Guinet,
Barnabé Mas, Chiara Régniez



Summary

1. Matching Motivation: Statistical Framework and Assumptions
2. Matching methods
3. Comparing Matching Methods
4. Application to Traumabase
5. Robustness to missing data

1. Matching Motivation

- Goal: Estimate the effect of a treatment
- 2 problems:
 - A person cannot be treated and not treated at the same time
 - There is an inherent bias linked with treatment assignment

1.1 Notation and Hypothesis

- Goal: Estimate the effect of a treatment

$$\mathbb{E}[Y^1 - Y^0]$$

- Average Treatment on the Treated

$$\mathbb{E}[Y^1 - Y^0 | A = 1]$$

- Hypotheses:

- Stable Unit Treatment Value Assumption (SUTVA) : observation on one unit should be unaffected by the particular assignment of treatments to the other units
- Ignorability : treatment randomly assigned among people with same characteristics
- Positivity

2. Matching methods

- Propensity score matching
- Coarsened exact matching
- Cardinality matching

2.1 Propensity Score Matching

- Matching method that groups the individual according to their propensity score
- The propensity score of an individual is its probability to be treated

PS estimation

- Logistic regression
- Step AIC
- Polynomial regression
- Machine learning algorithms

Matching on the PS

- kNN with various k
- Exact matching
- Caliper
- With replace

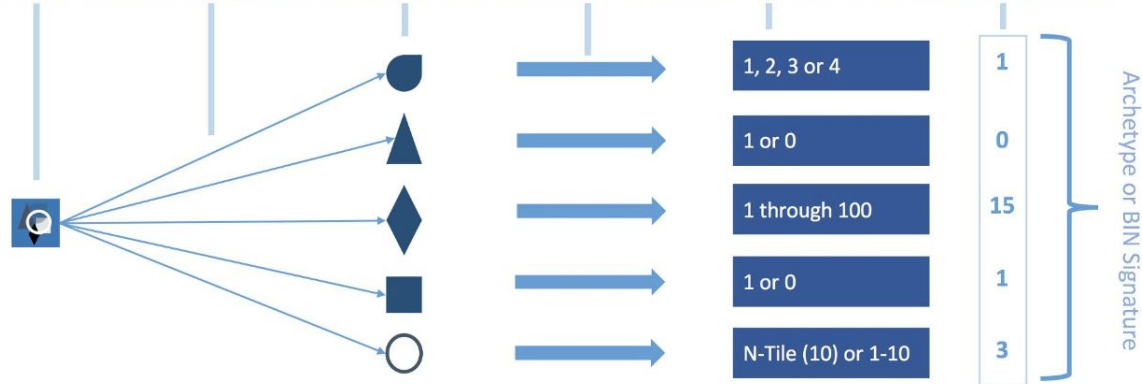
**ATT/ATE
estimation**

2.2 Coarsened Exact Matching

The CEM algorithm then involves three steps :

1. Temporarily coarsen each control variable in X (covariates) according to user-defined cutpoints, or CEM's automatic binning algorithm, for the purposes of matching.
2. Sort all units into strata, each of which has the same values of the coarsened X.
3. Prune from the data set the units in any stratum that do not include at least one treated

A member can be fairly represented by properties coarsened into values or BINS thus creating a BIN signature.



2.3 Cardinality Matching

- Matching method that maximizes the number of individuals matched under the covariates balance constraints
- The cardinality matching algorithm solves the integer programming problem:

$$\begin{array}{l} \text{maximize} \\ \mathbf{m} \end{array} \sum_{t \in \mathcal{T}} \sum_{c \in \mathcal{C}} m_{tc} \quad \left\{ \begin{array}{l} \sum_{c \in \mathcal{C}} m_{tc} \leq 1, \quad t \in \mathcal{T}, \\ \sum_{t \in \mathcal{T}} m_{tc} \leq 1, \quad c \in \mathcal{C}, \\ \left| \sum_{t \in \mathcal{T}} \sum_{c \in \mathcal{C}} m_{tc} (f_k(x_{tp}) - f_k(x_{cp})) \right| \leq \varepsilon_p \sum_{t \in \mathcal{T}} \sum_{c \in \mathcal{C}} m_{tc} \end{array} \right.$$

2.3 Cardinality Matching

- 3 steps :
 - Choose the covariate balance rules
 - Calculation of the largest group of individuals respecting balance constraints
 - New pairing among the matched group
- Variants : Matching One to many / Many to many
- Parameters: Balance covariates, choice of solver

3. Comparing Matching Methods

Methodology inspired by Resa, María de los Angeles et Zubizarreta, José R. Evaluation of subset matching methods and forms of covariate balance, *Statistics in Medicine* 2016 :

- Needed as no theoretical results to compare matching
- Challenge of “empirical experiments” $Y = f(\mathbf{X}) + Z + \varepsilon$ where $\varepsilon \sim \mathcal{N}(0, 4)$

Results comparison using:

- New covariate distribution (in mean, ratio of Variance divergence KS)
- Know ATT

3. Comparing Matching Methods

Methodology inspired by Resa, María de los Angeles et Zubizarreta, José R. Evaluation of subset matching methods and forms of covariate balance, *Statistics in Medicine* 2016 :

Covariate effect complexity



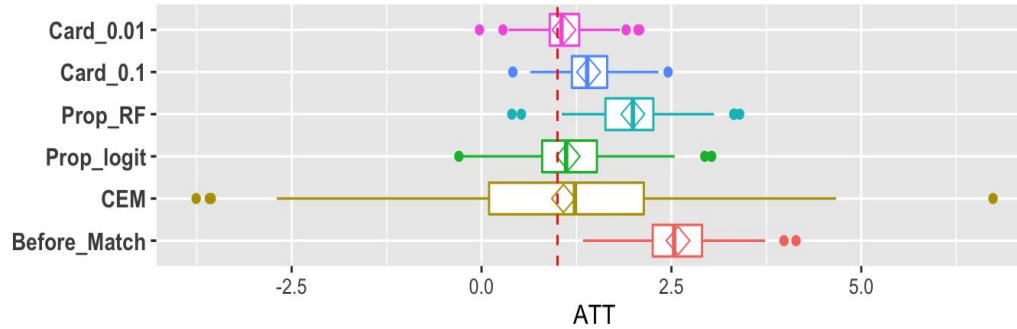
Linear	Additive	Additive with interactions	Diff. in mean
			Diff. in mean, Variance
			Diff. in mean, variance, Correlation

Data complexity

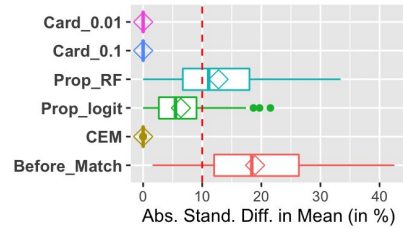


Scenario 1

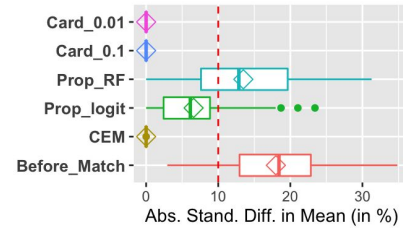
ATT estimation, scenario 1



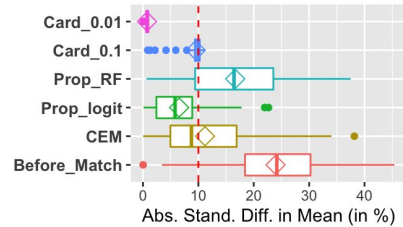
X1 balance



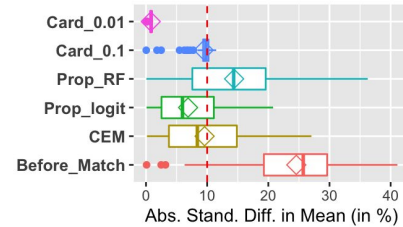
X3 balance



X5 balance



X7 balance

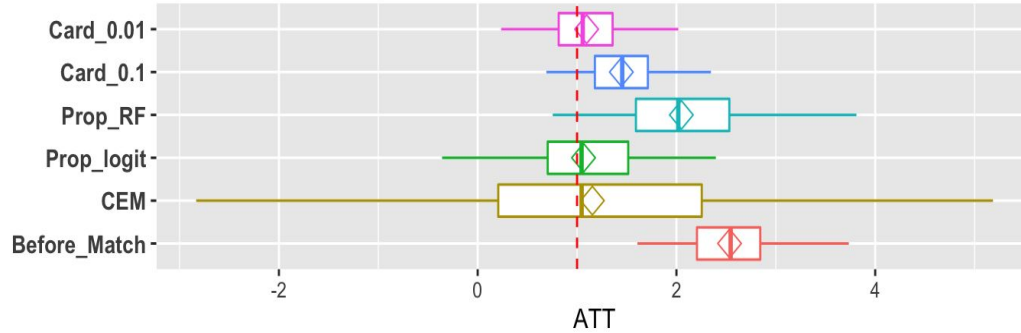


Scenario 1

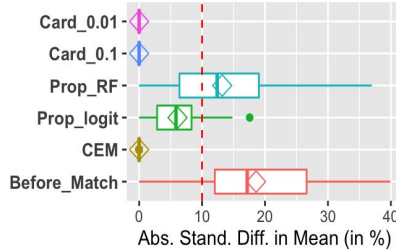
		Sdiff (in %)	Ratio Variance	K-S
X1	Cardinality 0.01	0.000000	1.000000	1.00000
	Cardinality 0.1	0.000000	1.000000	1.00000
	Propensity RF	14.57333	1.029241	0.22362
	Propensity Logit	6.124590	1.004150	0.55663
	CEM	0.000000	1.000000	0.09282
	Before Matching	20.73910	1.045730	0.09287
X3	Cardinality 0.01	0.000000	1.000000	1.00000
	Cardinality 0.1	0.000000	1.000000	1.00000
	Propensity RF	13.712944	1.872816	0.20781
	Propensity Logit	6.103638	1.063425	0.57964
	CEM	0.000000	1.000000	0.10780
	Before Matching	17.481523	2.132238	0.10813
X5	Cardinality 0.01	0.7863993	1.0099669	0.74640
	Cardinality 0.1	9.4089805	1.0074040	0.44845
	Propensity RF	14.1941137	0.9965678	0.16823
	Propensity Logit	6.2045843	1.0051326	0.24743
	CEM	9.3606365	1.0145340	0.11892
	Before Matching	23.3615836	0.9945550	0.11893
X7	Cardinality 0.01	0.7635358	1.0300577	0.73253
	Cardinality 0.1	9.6307802	1.0258699	0.46059
	Propensity RF	15.3117998	1.0146033	0.16968
	Propensity Logit	5.7594552	1.0298047	0.26992
	CEM	8.6104224	0.9986833	0.11090
	Before Matching	23.0940371	0.0116181	0.10934

Scenario 6

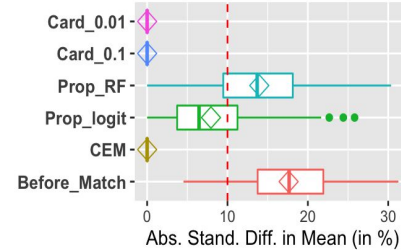
ATT estimation, scenario 6



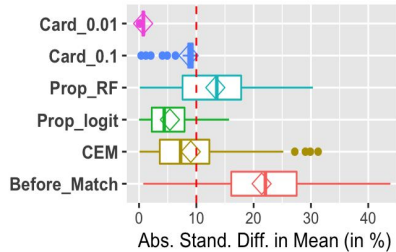
X1 balance



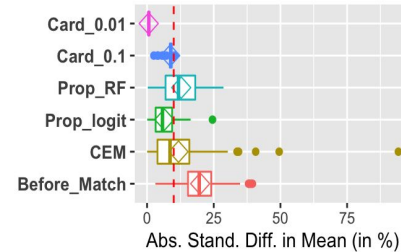
X3 balance



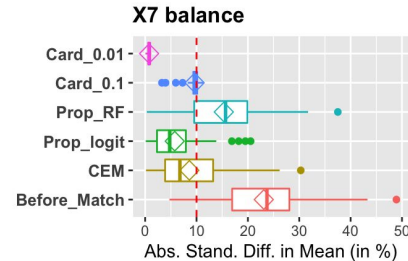
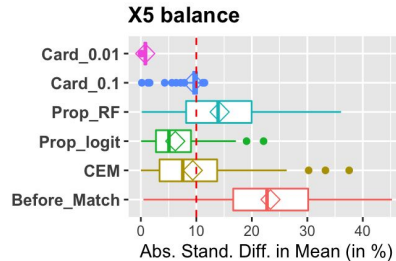
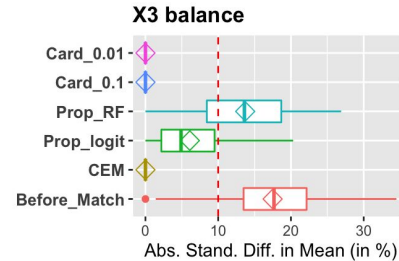
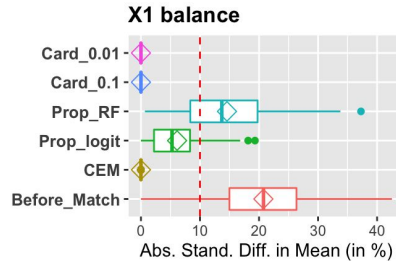
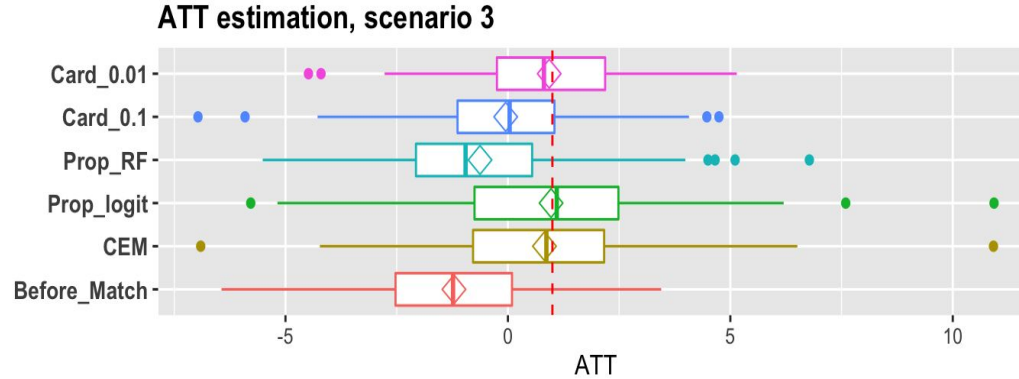
X5 balance



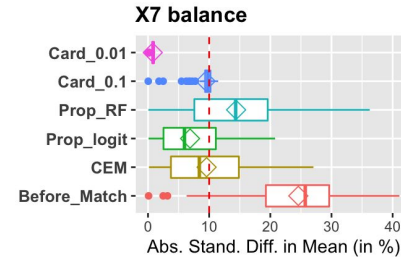
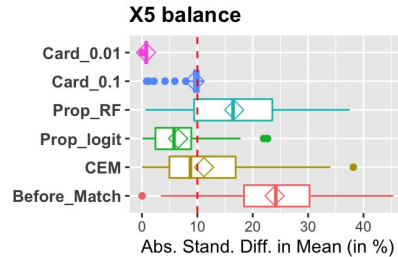
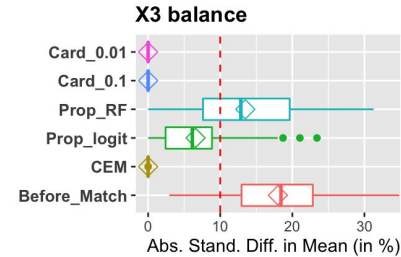
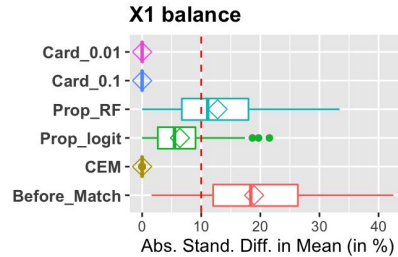
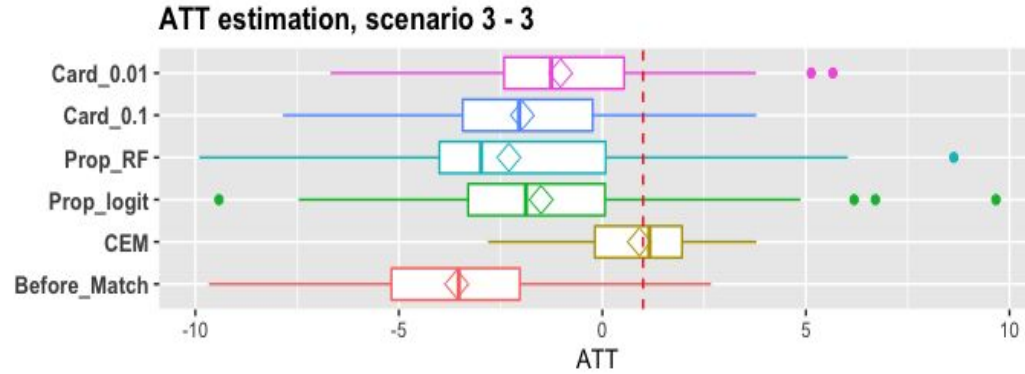
X7 balance



Scenario 3



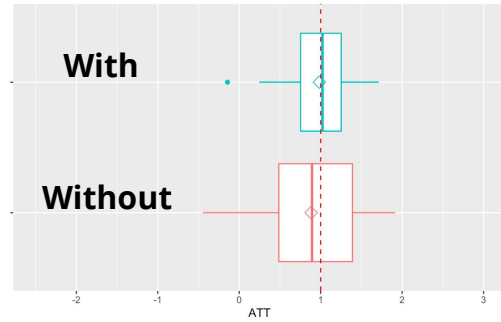
Scenario 9



Propensity results

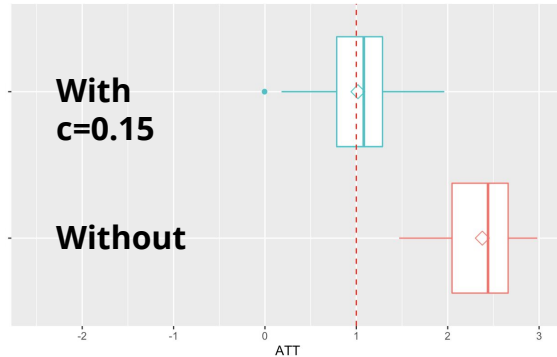
Replacement

ATT estimation, scenario X - Y



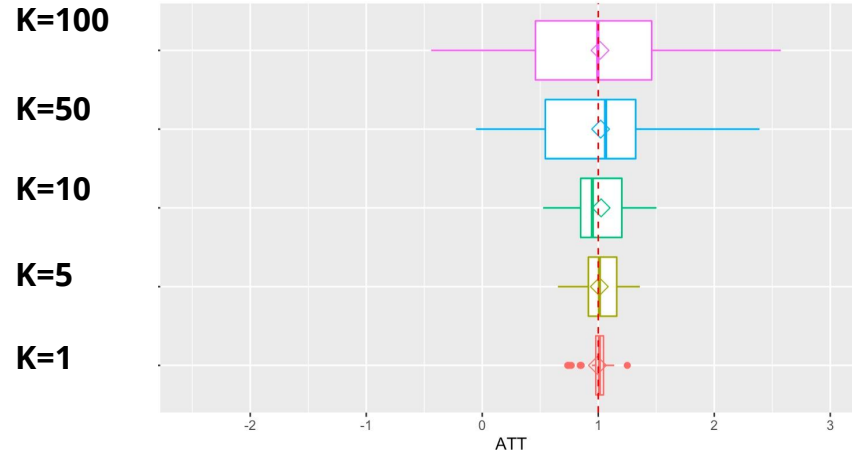
Caliper

ATT estimation, scenario X - Y



KNN with various K

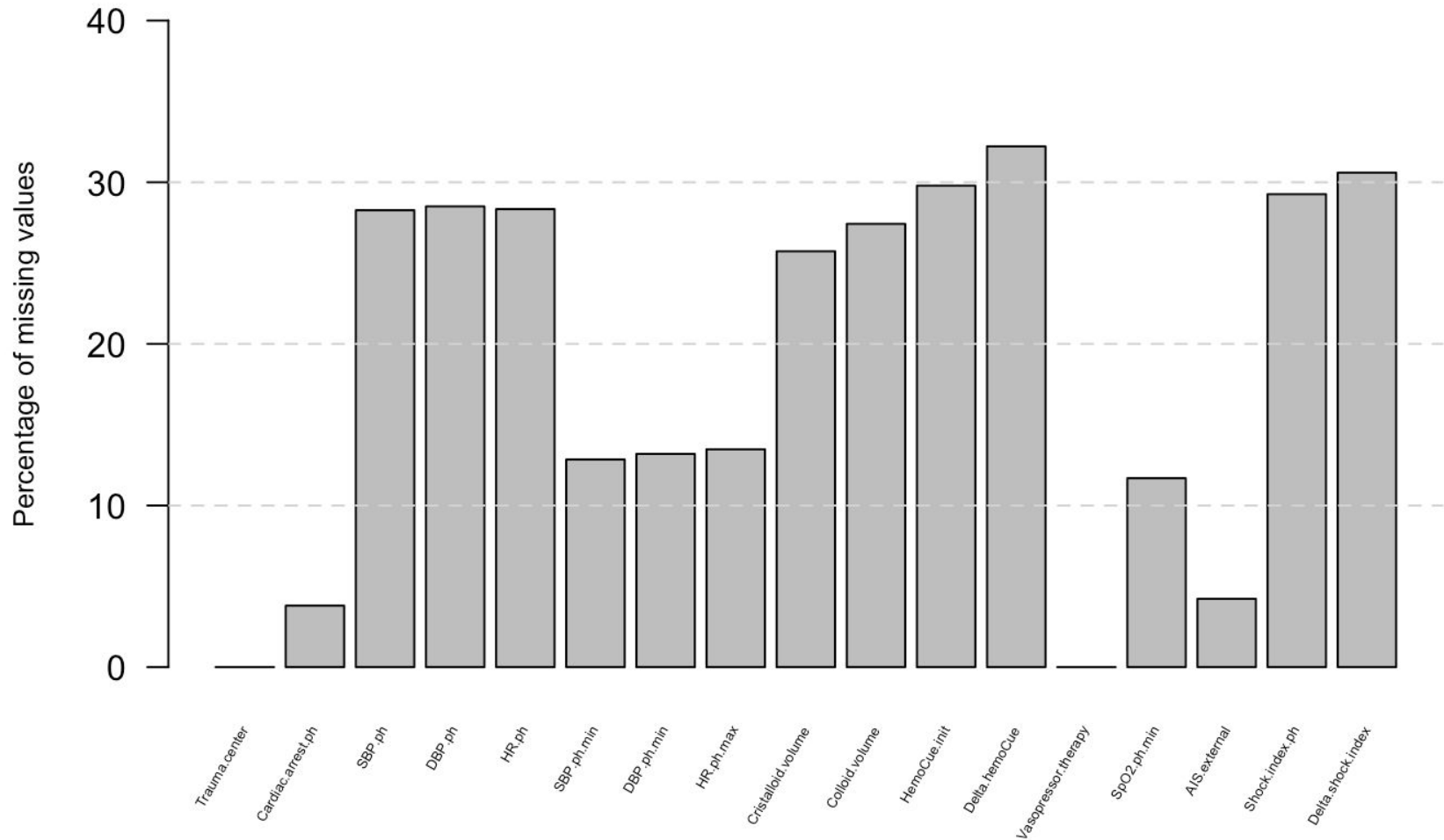
ATT estimation, scenario X - Y



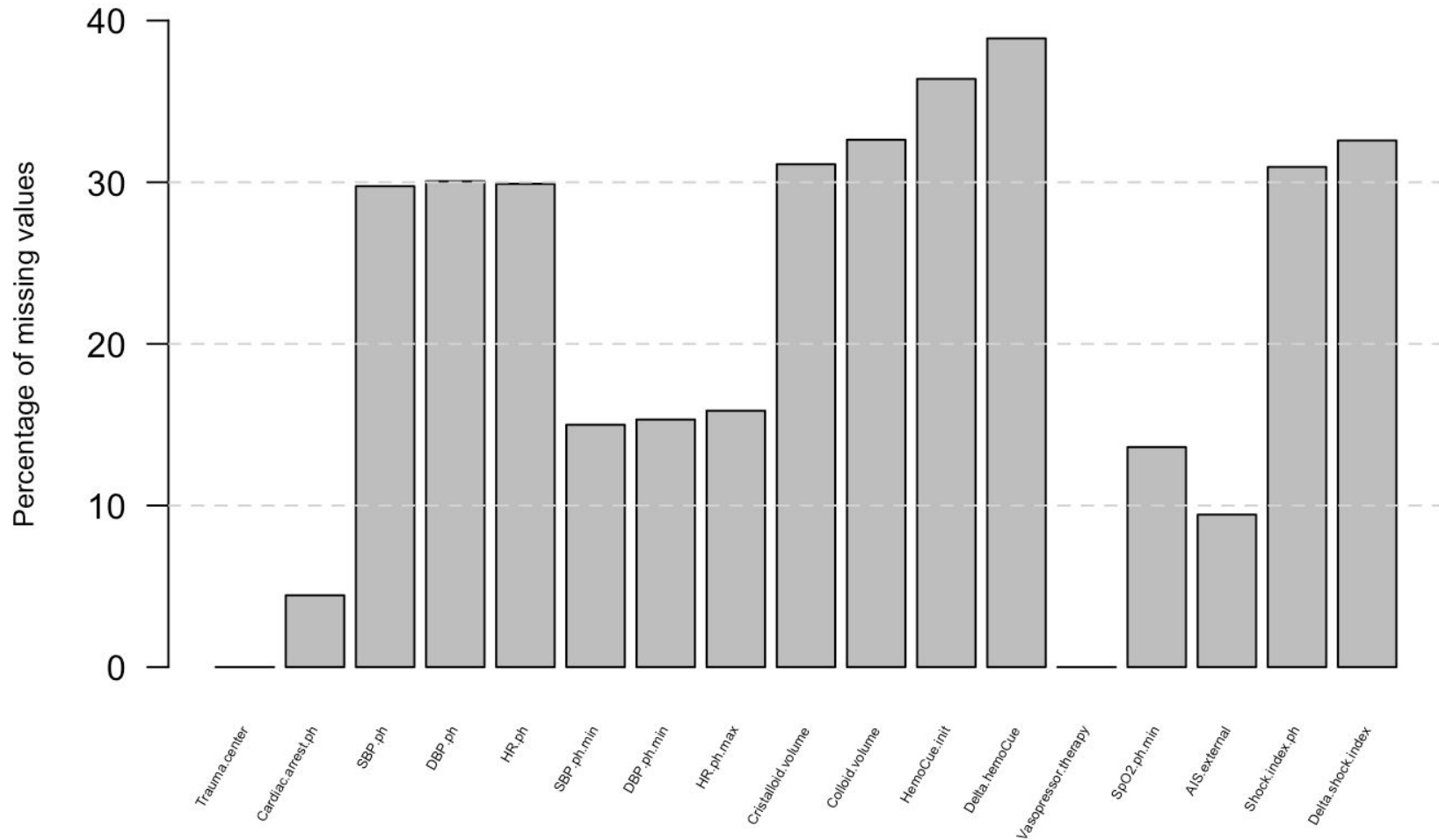
4. Application to Traumabase

- Traumabase database, including 20,037 patients and 272 variables
- Pre-processing to isolate the 17 study variables
- Missing data problem
- Mixed data problem: PCA

% of missing values for cofounders in patients



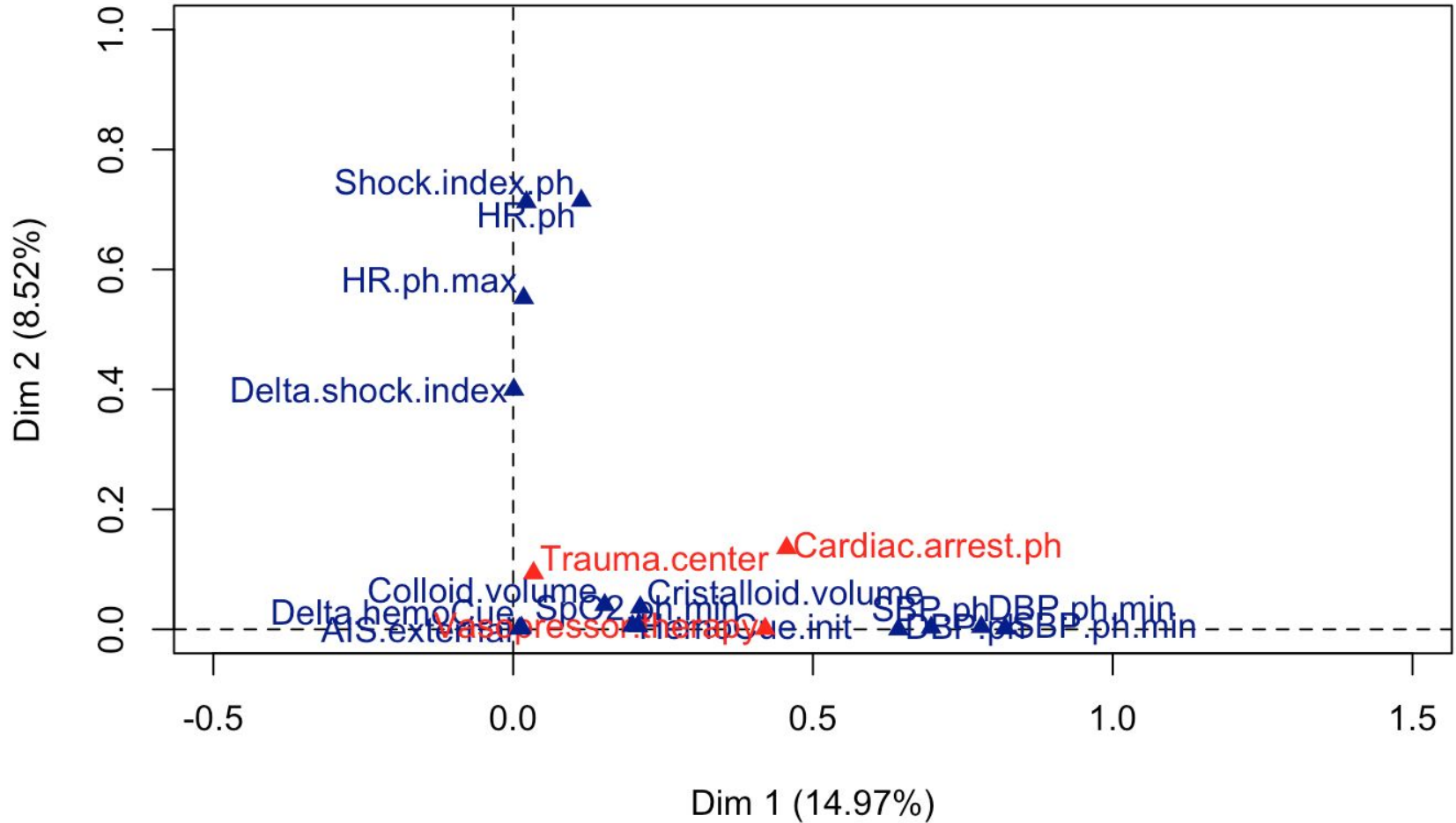
% of missing values for cofounders in patients w/ trauma cranien or AIS.tete >= 2



4. Application to Traumabase

- Traumabase database, including 20,037 patients and 272 variables
- Pre-processing to isolate the 17 study variables
- Missing data problem
- Mixed data problem: PCA

Graph of the variables



4. Application to Traumabase

- Selection of patients who have suffered a head injury
- Patients grouped by type and intensity of trauma

4. Traumabase: methods

Methods used on the traumabase:

- propensity score matching with logistic regression
- propensity score matching with random forests
- propensity score matching with multinomial regression
- coarsened exact matching
- cardinality matching

Estimation of the model quality:

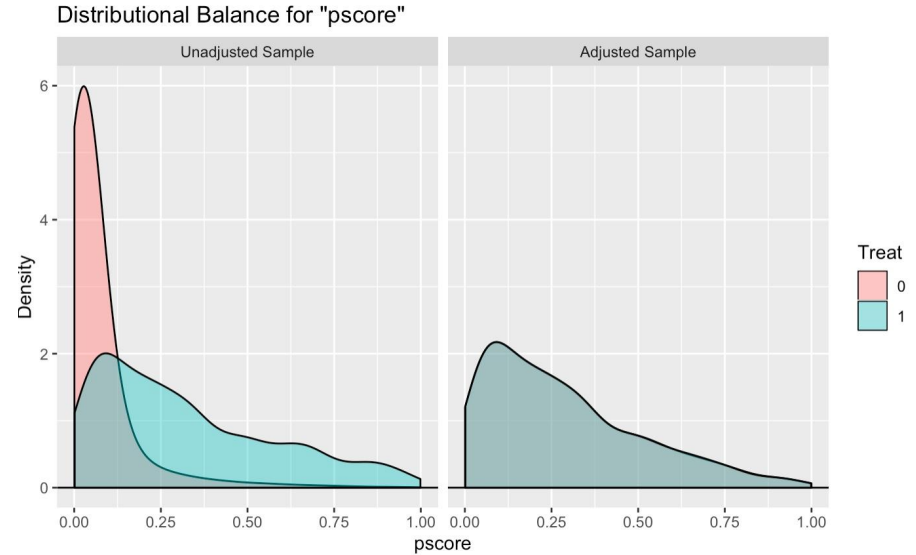
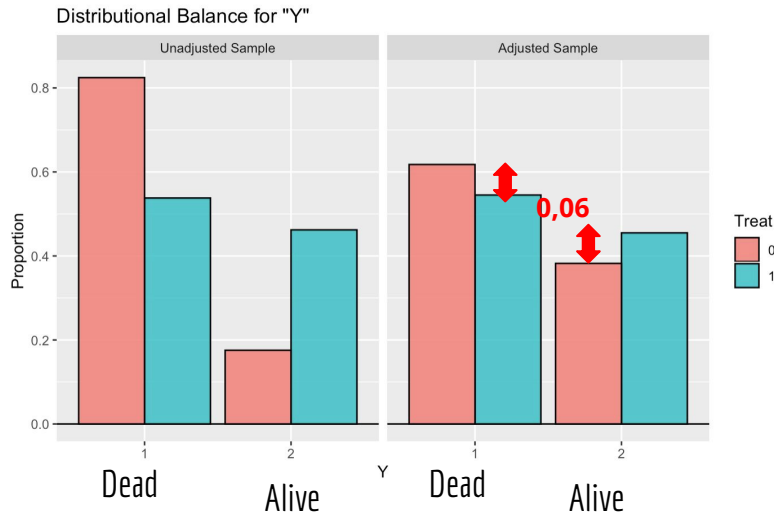
- Bootstrapping
- Residuals analysis
- Covariate balance

4. Traumabase: General results

Methods	Estimated ATT	IC 95%	Nb dropped
Propensity Logistic Reg	0,06	[0.0597, 0.0995]	53
Propensity Random Forest	0,09	[0.0969, 0.1664]	39
Propensity Multinomial	0,08	[0.0590, 0.1016]	53
CEM	0,12	[-0.0277, 0.4400]	3528
Cardinality matching	0,10	[0,0208, 0,16201]	7571

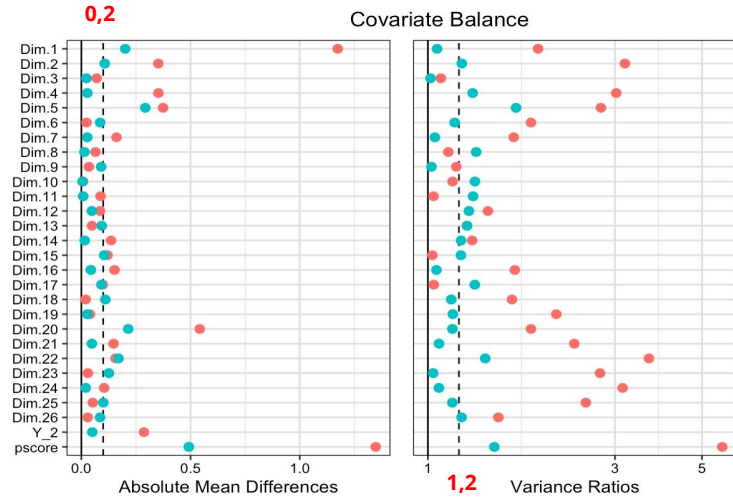
4. Traumabase: PS matching naive model results

ATT: 0,06

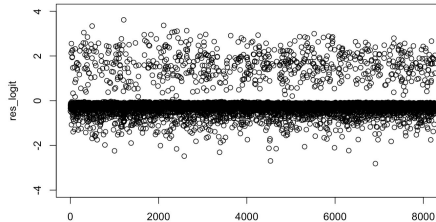
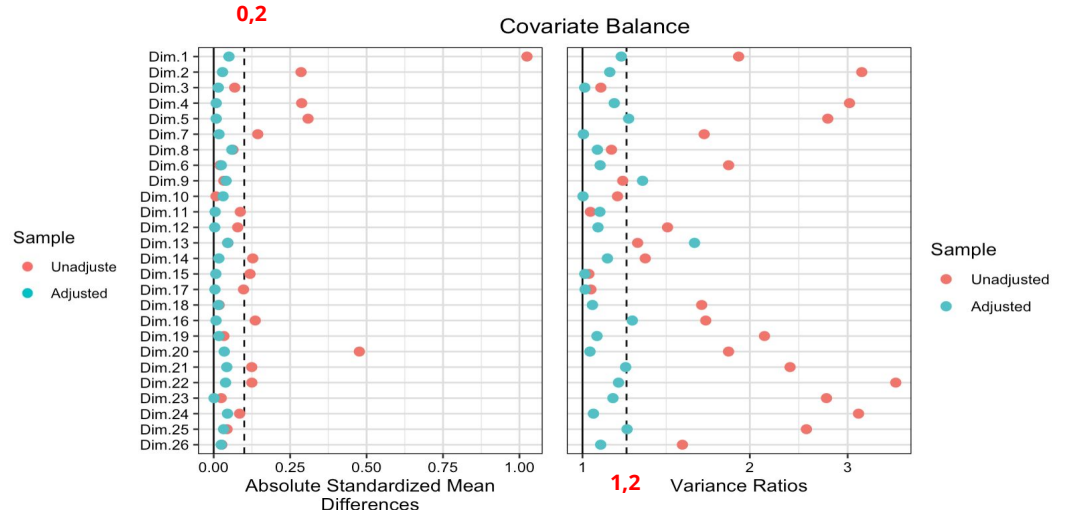


4. Traumabase: Model quality and data complexity

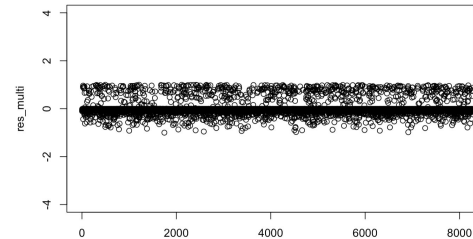
Logistic Regression



Polynomial Model



Residuals analysis



5. Robustness to missing data

- MCAR : Missing Completely At Random

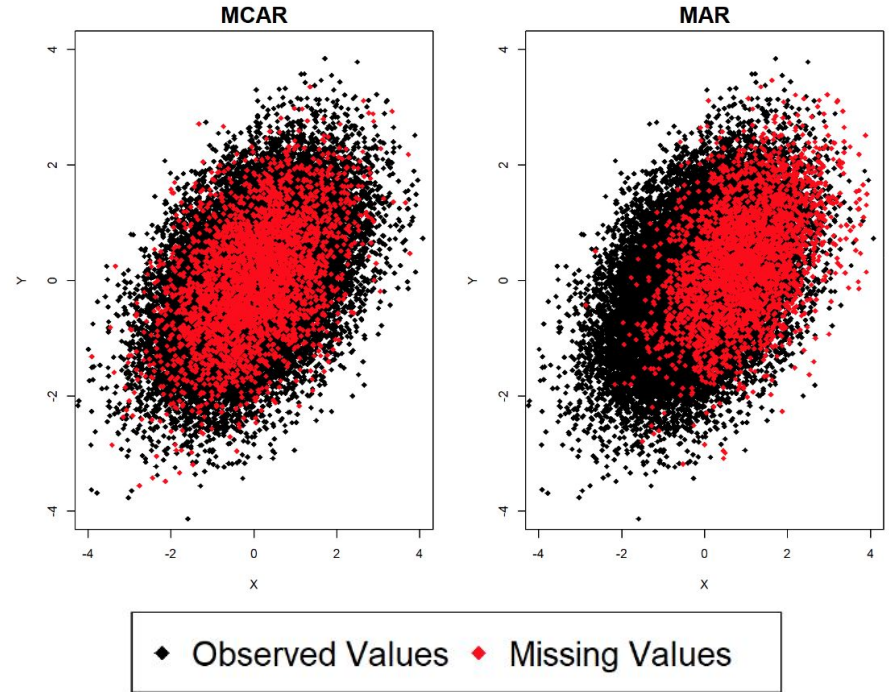
$$\forall \phi$$

$$\mathbb{P}_R(R | X^{obs}, X^{mis}; \phi) = \mathbb{P}_R(R)$$

- MAR : Missing At Random

$$\forall \phi, \forall X^{mis}$$

$$\mathbb{P}_R(R | X^{obs}, X^{mis}; \phi) = \mathbb{P}_R(R | X^{obs}; \phi)$$



5. Robustness to missing data

- MCAR : Missing Completely At Random

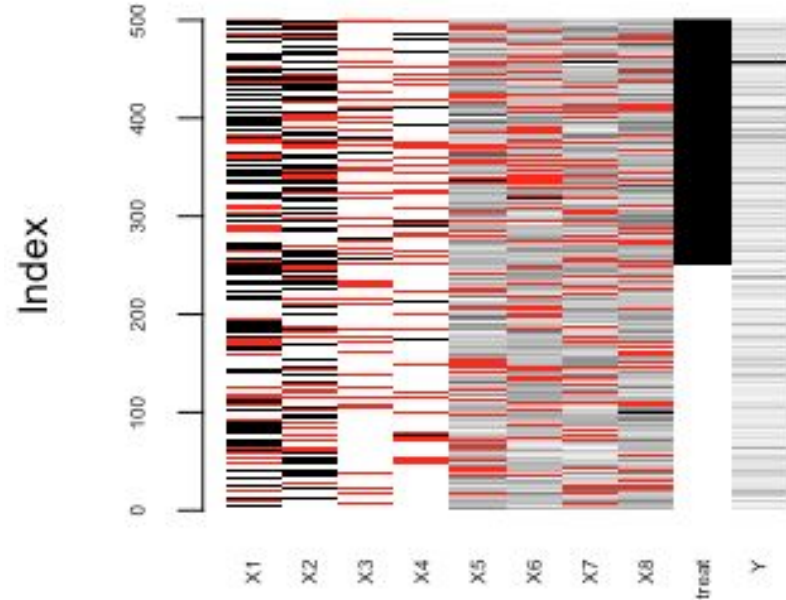
$$\forall \phi$$

$$\mathbb{P}_R(R | X^{obs}, X^{mis}; \phi) = \mathbb{P}_R(R)$$

- MAR : Missing At Random

$$\forall \phi, \forall X^{mis}$$

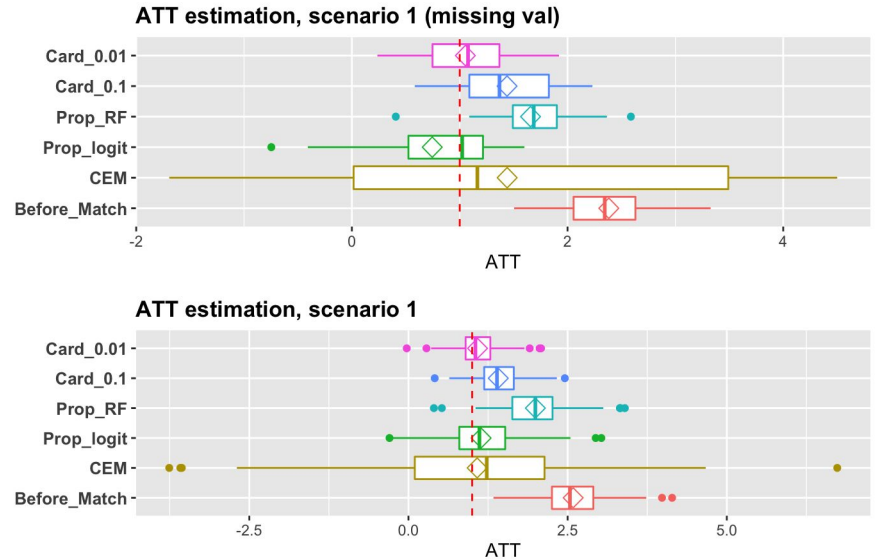
$$\mathbb{P}_R(R | X^{obs}, X^{mis}; \phi) = \mathbb{P}_R(R | X^{obs}; \phi)$$



5. Robustness to missing data

Imputation methods :

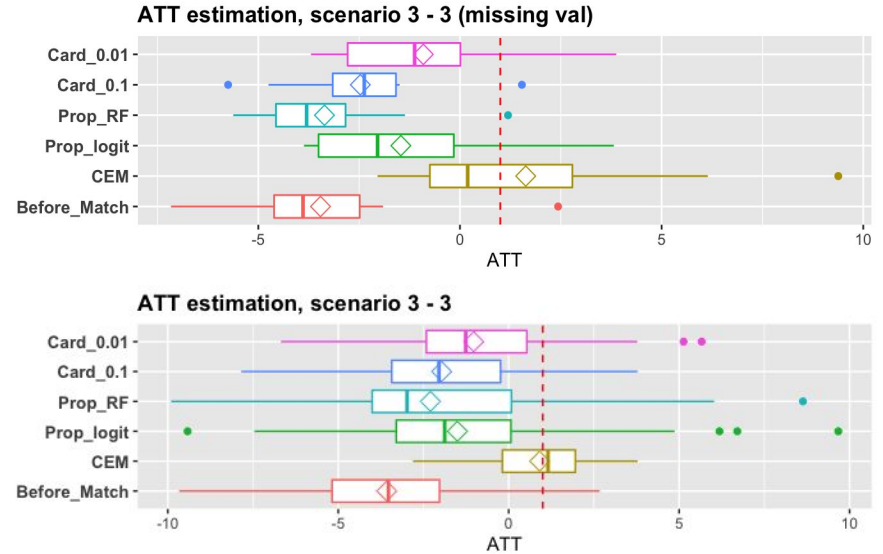
- deletion
- mean
- Amelia
- **impute FAMD**



5. Robustness to missing data

Imputation methods :

- deletion
- mean
- Amelia
- **impute FAMD**



Potential next steps

- Towards study of heterogeneous effect
- Further use of the missing value framework