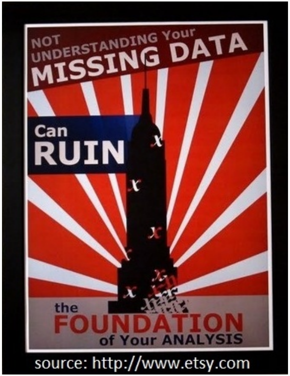
Treatment effect estimation with missing attributes

Julie Josse

École Polytechnique, INRIA
Visiting Researcher, Google Brain



Mathematical Methods of Modern Statistics, June 2020

Collaborators

Methods: Imke Mayer (PhD X, EHESS), Jean-Philippe Vert (Google Brain), Stefan Wager (Stanford)
Assistance Publique Hopitaux de Paris

l'X École Polytechnique | Institut Polytechnique de Paris

Inria — Inventeurs du monde numérique

L'École des Hautes Études en Sciences Sociales

Assistance Publique Hôpitaux de Paris

Google Research

Covid data

- 4780 patients (patients with at least one PCR-documented SARS-CoV-2 RNA from a nasopharyngeal sample)
- 119 continuous and categorical variables: **heterogeneous**
- 34 hospitals: **multilevel data**

Hospital	Treatment	Age	Sex	Weight	DDI	BP	dead28	...
Beaujon	HCQ	54	m	85	NA	180	yes	
Pitie	AZ	76	m	NA	NA	131	no	
Beaujon	HCQ+AZ	63	m	80	270	145	yes	
Pitie	HCQ	80	f	NA	NA	107	no	
HEGP	none	66	m	98	5890	118	no	
⋮								⋱

Covid data

- 4780 patients (patients with at least one PCR-documented SARS-CoV-2 RNA from a nasopharyngeal sample)
- 119 continuous and categorical variables: **heterogeneous**
- 34 hospitals: **multilevel data**

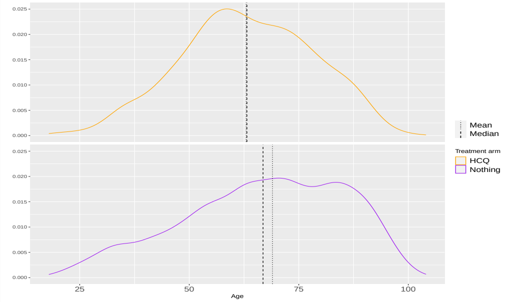
Hospital	Treatment	Age	Sex	Weight	DDI	BP	dead28	...
Beaujon	HCQ	54	m	85	NA	180	yes	
Pitie	AZ	76	m	NA	NA	131	no	
Beaujon	HCQ+AZ	63	m	80	270	145	yes	
Pitie	HCQ	80	f	NA	NA	107	no	
HEGP	none	66	m	98	5890	118	no	
⋮								⋱

$\Rightarrow$ **Estimate causal effect**: Administration of the **treatment** "Hydroxychloroquine" on the **outcome** 28-day mortality.

Observational data: non random assignment

	survived	deceased	Pr(survived | treatment)	Pr(deceased | treatment)
HCQ	497 (11.4%)	111 (2.6%)	0.817	0.183
HCQ+AZI	158 (3.6%)	54 (1.2%)	0.745	0.255
none	2699 (62.1%)	830 (19.1%)	0.765	0.235

Mortality rate 23% - for HCQ 18% - non treated 24%: treatment helps?



Comparison of the distribution of Age between HCQ and non treated.

Severe patients (with higher risk of death) are less likely to be treated.
If control group does not look like treatment group, difference in response may be **confounded** by differences between the groups.

Potential outcome framework (Neyman, 1923, Rubin, 1974)

Causal effect

- n iid samples $(X_i, W_i, Y_i(1), Y_i(0)) \in \mathbb{R}^d \times \{0, 1\} \times \mathbb{R} \times \mathbb{R}$
- Individual causal effect of the treatment: $\Delta_i \triangleq Y_i(1) - Y_i(0)$

Missing problem: Δ_i never observed (only observe one outcome/indiv)

Covariates			Treatment	Outcome(s)	
X_1	X_2	X_3	W	Y(0)	Y(1)
1.1	20	F	1	**?**	Survived
-6	45	F	0	Dead	**?**
0	15	M	1	**?**	Survived
	...		...	...	...
-2	52	M	0	Survived	**?**

Potential outcome framework (Neyman, 1923, Rubin, 1974)

Causal effect

- n iid samples $(X_i, W_i, Y_i(1), Y_i(0)) \in \mathbb{R}^d \times \{0, 1\} \times \mathbb{R} \times \mathbb{R}$
- Individual causal effect of the treatment: $\Delta_i \triangleq Y_i(1) - Y_i(0)$

Missing problem: Δ_i never observed (only observe one outcome/indiv)

Covariates			Treatment	Outcome(s)	
X_1	X_2	X_3	W	Y(0)	Y(1)
1.1	20	F	1	**?**	Survived
-6	45	F	0	Dead	**?**
0	15	M	1	**?**	Survived
	...		...	...	...
-2	52	M	0	Survived	**?**

Average treatment effect (ATE): $\tau \triangleq \mathbb{E}[\Delta_i] = \mathbb{E}[Y_i(1) - Y_i(0)]$
The ATE is the difference of the average outcome had everyone gotten treated and the average outcome had nobody gotten treatment.

ATE=0.05: mortality rate in the treated group is 5% points higher than in the control group. So, on average the treatment increases the risk of dying.

Assumption for ATE identifiability in observational data

Unconfoundedness - selection on observables

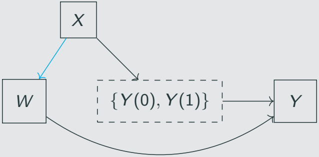
$$\{Y_i(0), Y_i(1)\} \perp\!\!\!\perp W_i \mid X_i$$

Treatment assignment W_i is random conditionally on covariates X_i

Measure enough covariates to capture dependence between W_i and outcomes

- Observed outcome: $Y_i = W_i Y_i(1) + (1 - W_i) Y_i(0)$

Unconfoundedness - graphical model



Unobserved confounders make it impossible to separate correlation and causality when correlated to both the outcome and the treatment.

ATE not identifiable without assumption: it is not a sample size problem!

Assumption for ATE identifiability in observational data

Propensity score: probability of treatment given observed covariates.

Propensity score - overlap assumption

$$e(x) \triangleq \mathbb{P}(W_i = 1 \mid X_i = x) \quad \forall\, x \in \mathcal{X}.$$

We assume overlap, i.e. $\eta < e(x) < 1 - \eta, \quad \forall\, x \in \mathcal{X}$ and some $\eta > 0$

Left: Non smoker and never treated

Right: Smokers and all treated

If proba to be treated when smoker $e(x) = 1$, how to estimate the outcome for smokers when not treated $Y(0)$? How to extrapolate if total confusion?

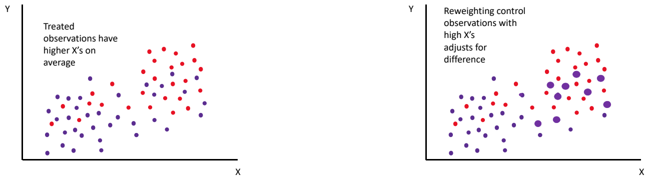
Inverse-propensity weighting estimation of ATE

Average treatment effect (ATE): $\tau \triangleq \mathbb{E}[\Delta_i] = \mathbb{E}[Y_i(1) - Y_i(0)]$
Propensity score: $e(x) \triangleq \mathbb{P}(W_i = 1 \mid X_i = x)$

IPW estimator (Horvitz-Thomson, survey)

$$\hat{\tau}_{IPW} \triangleq \frac{1}{n} \sum_{i=1}^{n} \left(\frac{W_i Y_i}{\hat{e}(X_i)} - \frac{(1 - W_i) Y_i}{1 - \hat{e}(X_i)} \right)$$

$\Rightarrow$ Balance the differences between the two groups
$\Rightarrow$ Consistent estimator of τ as long as $\hat{e}(\cdot)$ is consistent.



Credit: S. Athey

Doubly robust ATE estimation

Model Treatment on Covariates $e(x) \triangleq \mathbb{P}(W_i = 1 \,|\, X_i = x)$
Model Outcome on Covariates $\mu_{(w)}(x) \triangleq \mathbb{E}[Y_i(w) \,|\, X_i = x]$

Augmented IPW - Double Robust (DR)

$$\hat{\tau}_{AIPW} \triangleq \frac{1}{n}\sum_{i=1}^{n} \left(\hat{\mu}_{(1)}(X_i) - \hat{\mu}_{(0)}(X_i) + W_i \frac{Y_i - \hat{\mu}_{(1)}(X_i)}{\hat{e}(X_i)} - (1 - W_i) \frac{Y_i - \hat{\mu}_{(0)}(X_i)}{1 - \hat{e}(X_i)} \right)$$

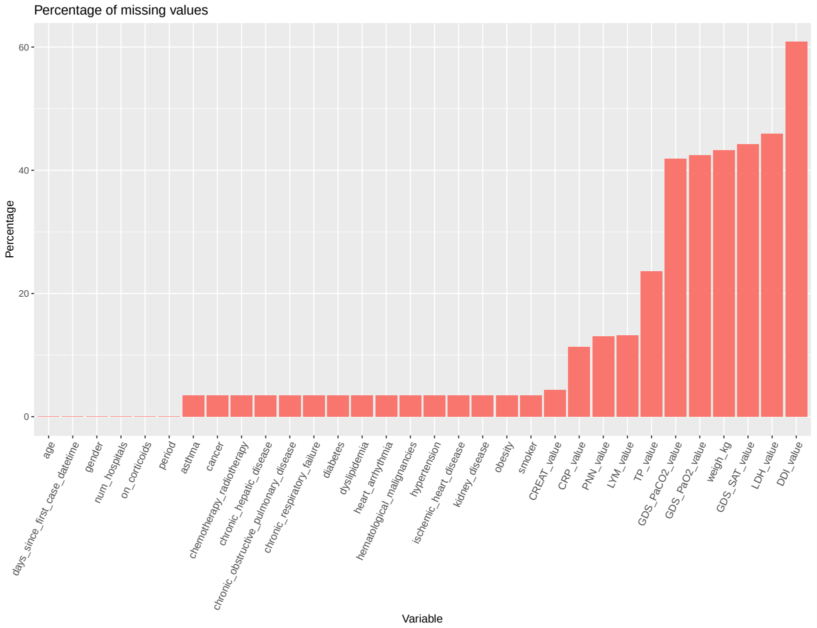
is consistent if either the $\hat{\mu}_{(w)}(x)$ are consistent or $\hat{e}(x)$ is consistent.

Possibility to use **any (machine learning) procedure** such as **random forests**, deep nets, etc. to estimate $\hat{e}(x)$ and $\hat{\mu}_{(w)}(x)$ without harming the interpretability of the causal effect estimation.

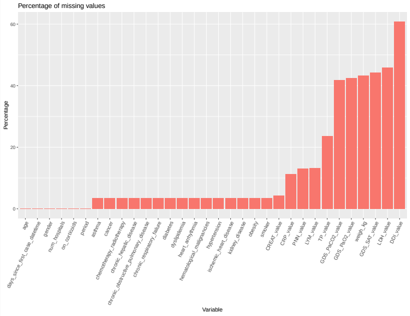
Properties - Double Machine Learning (Chernozhukov et al., 2018)

If $\hat{e}(x)$ and $\hat{\mu}_{(w)}(x)$ converge at the rate $n^{1/4}$ then
$\sqrt{n}\,(\hat{\tau}_{DR} - \tau) \xrightarrow[n \to \infty]{d} \mathcal{N}(0, V^*)$, V^* semiparametric efficient variance.

Missing values



Missing values



Deleting rows with missing values?

"One of the ironies of Big Data is that missing data play an ever more significant role" (R. Samworth, 2019)

An $n \times p$ matrix, each entry is missing with probability 0.01

$p = 5 \quad \implies \approx 95\%$ of rows kept

$p = 300 \implies \approx \ \ 5\%$ of rows kept

Missing (informative) values in the covariates

Straightforward – but often biased – solution is complete-case analysis.

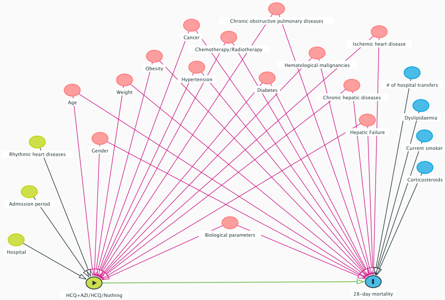
Covariates			Treatment	Outcome(s)	
X_1	X_2	X_3	W	Y(0)	Y(1)
NA	20	F	1	?	Survived
-6	45	NA	0	Dead	?
0	NA	M	1	?	Survived
NA	32	F	1	?	Dead
1	63	M	1	Dead	?
-2	NA	M	0	Survived	?

→ Often not a good idea! What are the alternatives?

Three families of methods - different assumptions

- Unconfoundedness with missingness + (no) missing values mechanisms Mayer, J., Wager, Sverdrup, Moyer, Gauss. *AOAS* 2020.
- Classical unconfoundedness + classical missing values mechanisms
- Latent unconfoundedness + classical missing values mechanisms Mayer, J., Raimundo, Vert. 2020.

1. Unconfoundedness with missing + (no) missing hypothesis



Covariates			Treatment	Outcome(s)	
X_1^*	X_2^*	X_3^*	W	Y(0)	Y(1)
NA	20	F	1	?	S
-6	45	NA	0	D	?
0	NA	M	1	?	S
NA	32	F	1	?	D
1	63	M	1	D	?
-2	NA	M	0	S	?

Unconfoundedness: $\{Y_i(1), Y_i(0)\} \perp\!\!\!\perp W_i \,|\, X$ not testable from the data.
$\Rightarrow$ Doctors give us the DAG (covariates relevant for either treatment decision and for predicting the outcome)

Unconfoundedness with missing values: $\{Y_i(1), Y_i(0)\} \perp\!\!\!\perp W_i \,|\, X^*$
$X^* \triangleq (1 - R) \odot X + R \odot NA$; with $R_{ij} = 1$ if X_{ij} is missing, 0 otherwise.

$\Rightarrow$ Doctors decide to treat a patient based on what they observe/record.
We have access to the same information as the doctors.

Under 1: Double Robust with missing values

AIPW with missing values

$$\hat{\tau}^* \triangleq \frac{1}{n}\sum_i \left(\widehat{\mu^*_{(1)}}(X_i) - \widehat{\mu^*_{(0)}}(X_i) + W_i \frac{Y_i - \widehat{\mu^*_{(1)}}(X_i)}{\widehat{e^*}(X_i)} - (1 - W_i)\frac{Y_i - \widehat{\mu^*_{(0)}}(X_i)}{1 - \widehat{e^*}(X_i)} \right)$$

Generalized propensity score **(Rosenbaum and Rubin, 1984)**

$$e^*(x^*) \triangleq \mathbb{P}(W = 1 \,|\, X^* = x^*)$$

One model per pattern: $\sum_{r\in\{0,1\}^d} \mathbb{E}\left[W | X_{obs(r)}, R = r\right] \mathbb{1}_{R=r}$

$\Rightarrow$ Supervised learning with missing values. [1]

- Mean imputation is consistent with a universally consistent learner.
- Missing Incorporate in Attributes (MIA) for trees methods.

[1] consistency of supervised learning with missing values J., Prost, Scornet, Varoquaux. *JMLR 2020*

Under 1: Double Robust with missing values

AIPW with missing values

$$\hat{\tau}^* \triangleq \frac{1}{n}\sum_i \left(\widehat{\mu^*_{(1)}}(X_i) - \widehat{\mu^*_{(0)}}(X_i) + W_i \frac{Y_i - \widehat{\mu^*_{(1)}}(X_i)}{\widehat{e^*}(X_i)} - (1 - W_i)\frac{Y_i - \widehat{\mu^*_{(0)}}(X_i)}{1 - \widehat{e^*}(X_i)} \right)$$

Generalized propensity score (Rosenbaum and Rubin, 1984)

$$e^*(x^*) \triangleq \mathbb{P}(W = 1 \,|\, X^* = x^*)$$

One model per pattern: $\sum_{r\in\{0,1\}^d} \mathbb{E}\left[W | X_{obs(r)}, R = r\right] \mathbb{1}_{R=r}$

$\Rightarrow$ Supervised learning with missing values. [1]

- Mean imputation is consistent with a universally consistent learner.
- Missing Incorporate in Attributes (MIA) for trees methods.

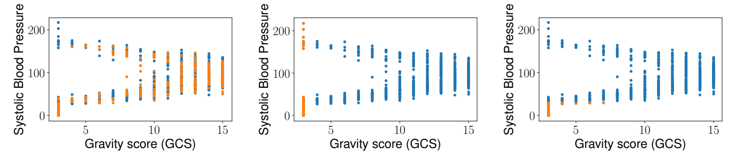
Implemented in `grf` package: combine two non-parametrics models, forests (conditional outcome and treatment assignment) adapted to **any** missing values with MIA.

$\hat{\tau}_{AIPW^*}$ is $\sqrt{n}$-consistent, asymptotically normal given the product of RMSE of the nuisance estimates decay as $o(n^{-1/2})$ Mayer et al. AOAS 2020

[1] consistency of supervised learning with missing values J., Prost, Scornet, Varoquaux. *JMLR 2020*

2. Classical unconfoundedness + missing values mechanism

Aparté on missing values mechanisms taxonomy (Rubin, 1976)



MCAR - MAR - MNAR

Orange: missing values for Systolic Blood Pressure - Gravity index (GCS) is always observed

MCAR (completely at random): Proba to be missing does not depend on SBP neither on gravity
MAR: Proba depends on gravity (we do not measure for too severe patients)
MNAR (not at random): Proba depends on SBP (low SBP not measured)

Under 2: Multiple Imputation

Consistency of IPW with missing values (Seaman and White, 2014)

Assume **Missing At Random (MAR)** mechanism. Multiple imputation (MICE using (X^*, W, Y)) with IPW on each imputed data is consistent when Gaussian covariates and logistic/linear treatment/oucome model

X_1^*	X_2^*	X_3^*	...	W	Y
NA	20	10	...	1	survived
-6	45	NA	...	1	survived
0	NA	30	...	0	died
NA	32	35	...	0	survived
-2	NA	12	...	0	died
1	63	40	...	1	survived

1) Generate M plausible values for each missing value

X_1	X_2	X_3	...	W	Y
3	20	10	...	1	s
-6	45	6	...	1	s
0	4	30	...	0	d
-4	32	35	...	0	s
-2	15	12	...	0	d
1	63	40	...	1	s

X_1	X_2	X_3	...	W	Y
-7	20	10	...	1	s
-6	45	9	...	1	s
0	12	30	...	0	d
13	32	35	...	0	s
-2	10	12	...	0	d
1	63	40	...	1	s

X_1	X_2	X_3	...	W	Y
7	20	10	...	1	s
-6	45	12	...	1	s
0	-5	30	...	0	d
2	32	35	...	0	s
-2	20	12	...	0	d
1	63	40	...	1	s

2) Estimate ATE on each imputed data set: $\hat{\tau}_m, \widehat{Var}\left(\hat{\tau}_m\right)$

3) Combine the results (Rubin's rules): $\hat{\tau} = \frac{1}{M} \sum_{m=1}^{M} \hat{\tau}_m$

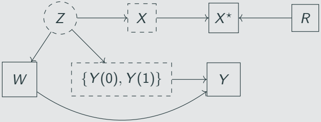
$$\widehat{Var}\left(\hat{\tau}\right) = \frac{1}{M} \sum_{m=1}^{M} \widehat{Var}\left(\hat{\tau}_m\right) + \left(1 + \frac{1}{M}\right) \frac{1}{M-1} \sum_{m=1}^{M} \left(\hat{\tau}_m - \hat{\tau}\right)^2$$

3. Latent unconfoundedness + missing values mechanism

Latent confounding assumption

The covariates X are **noisy (incomplete) proxies** of the true **latent confounders** Z (Kallus et al., 2018; Louizos et al., 2017).

$X^* \triangleq (1 - R) \odot X + R \odot NA$; with $R_{ij} = 1$ if X_{ij} is missing, 0 otherwise
Observed outcome: $Y_i = W_i Y_i(1) + (1 - W_i) Y_i(0)$



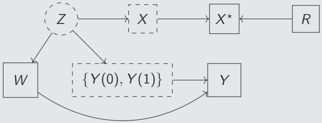
3. Latent unconfoundedness + missing values mechanism

Latent confounding assumption

The covariates X are **noisy (incomplete) proxies** of the true **latent confounders** Z (Kallus et al., 2018; Louizos et al., 2017).

$X^* \triangleq (1-R) \odot X + R \odot NA$; with $R_{ij} = 1$ if X_{ij} is missing, 0 otherwise

Observed outcome: $Y_i = W_i Y_i(1) + (1 - W_i) Y_i(0)$



Matrix Factorization as a pre-processing step

• Assume data are generated as a low-rank structure corrupted by noise. Estimate Z using matrix completion from X^*(softimpute types).

• Plug $\hat{Z}$ in regression model of outcome on treatment and confounders: $Y = \tau W + Z\beta + \varepsilon$, $\varepsilon \sim \mathcal{N}(0, \sigma^2 I)$ (or in the (A)IPW estimators)

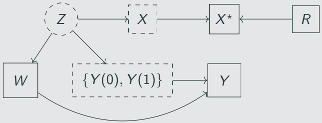
• Kallus et al. (2018) show that $\hat{\tau}$ is a consistent estimator under MCAR of the Average Treatment Effect.

3. Latent unconfoundedness + missing values mechanism

Latent confounding assumption

Covariates $X_{n\times d}$ proxies of the latent confounders $Z_{n\times q}$.

$X^* \triangleq (1-R)\odot X + R \odot NA$; with $R_{ij} = 1$ if X_{ij} is missing, 0 otherwise



MissDeepCausal (MDC) Mayer, J., Raimundo, Vert, 2020.

- Assume a Deep Latent Variable Model instead of linear factor analysis

- Leverage VAE with MAR values (Mattei and Frellsen, 2019). Imputing NA with 0 maximizes an ELBO of the **observed** log-likelihood.

- Draw $(Z^{(j)})_{1\leq j\leq B}$ from the posterior distribution $P(Z|X^\star)$ (using importance sampling with $Q(Z|X^\star)$ for proposal).

MDC-Multiple Imputation: estimate ATE on each $(Z^{(j)})$
MDC-process plug-in $\hat{Z}(x^\star) \triangleq \mathbb{E}[Z|X^\star = x^\star]$ in classical estimators

Flexible with promising empirical results.

Methods to do causal inference with missing values

	Covariates		Missingness		Unconfoundedness			Models for (W, Y)	
	multivariate normal	general	M(C)AR	general	Missing	Latent	Classical	logistic-linear	non-param.
1. *(SA)EM* [2]	✓	✗	✓	✗	✓	✗	✗	✓	✗
1. Mean.GRF	✓	✓	✓	(✓)	✓	✗	✗	✓	✓
1. MIA.GRF	✓	✓	✓	(✓)	✓	✗	✗	✓	✓
2. *Mult. Imp.*	✓	✓	✓	✗	(✗)	✗	✓	✓	(✗)
3. *MatrixFact.*	✓	✗	✓	✗	✗	✓	✗	✓	(✗)
3. MissDeep-Causal	✓	✓	✓	✗	✗	✓	✗	✓	✓

Methods & assumptions on data generating process (models for covariates, outcome, treatment), missing values mechanism and identifiability conditions.

✓: can be handled ✗: not applicable in theory

(✓): empirical results and ongoing work on theoretical guarantees

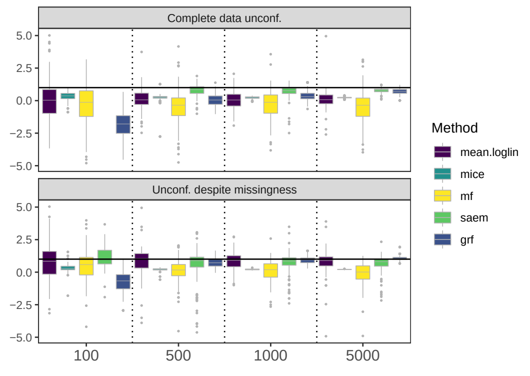
(✗): no theoretical guarantees but heuristics.

[2] Use of EM algorithms for logistic regression with missing values. Jiang et al. (2020)

Simulations: no overall best performing method.

- 10 covariates generated with Gaussian mixture model $X_i \sim \mathcal{N}_d(\mu_{(c_i)}, \Sigma_{(c_i)}) | C_i = c_i$, C from a multinomial distribution with three categories.
- Unconfoundedness on complete/observed covariates, 30% NA
- Logistic-linear for $(W,\ Y)$, $logit(e(X_{i.})) = \alpha^T X_{i.}$, $Y_i \sim \mathcal{N}(\beta^T X_{i.} + \tau W_i, \sigma^2)$

Figure 1: Estimated with AIPW and true ATE $\tau = 1$



→ GRF-MIA is asymptotically unbiased under unconfoundedness despite missingness.
→ Multiple imputation requires many imputations to remove bias.

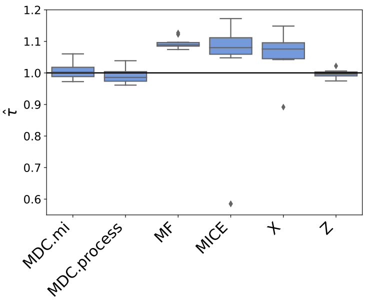
Simulations: no overall best performing method.

- 100 covariates generated with a **DLVM** model, **latent confounding** ($q = 3$):

$Z_i \sim \mathcal{N}_q(0, \sigma_z)$, covariates X_i sampled from $\mathcal{N}_d(\mu_{(Z)}, \Sigma_{(Z)})$, where $(\mu_{(Z)}, \Sigma_{(Z)}) = (V \tanh(UZ + a) + b, \mathrm{diag}\{\exp(\eta^T \tanh(UZ + a) + \delta)\})$ with U, V, a, b, δ, η drawn from standard Gaussian and uniform distributions.

- 30% MCAR, $n = 1000$.
- Logistic-linear for $(W,\ Y)$, $logit(e(Z_{i\cdot})) = \alpha^T Z_{i\cdot},\ Y_i \sim \mathcal{N}(\beta^T Z_{i\cdot} + \tau W_i, \sigma^2)$

Figure 1: Estimated with AIPW and true ATE $\tau = 1$.

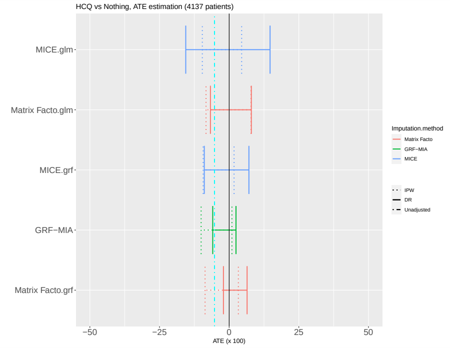


→ MDC empirically unbiased if number of features (d) >> dim of the latent space (q)
Tuning: variance of the prior of Z and $\hat{q}$ chosen by cross-validation using the ELBO

Results for Covid Patients

33 covariates, 26 confounders. 4137 patients.

ATE estimations ($\times$100): effect of Hydroxychloroquine on 28day mortality



(y-axis: estimation approach, solid: **Doubly Robust AIPW**, dotted: **IPW**),
(x-axis: ATE estimation with CI)

The obtained value corresponds to the **difference in percentage points between mortality rates in treatment and control**.

Light Blue: unadjusted (-5.3)

Take-away messages

- **Missing attributes** alter causal analyses. Performance of methods depends on the underlying assumptions

Conclusion and perspectives

Take-away messages

- **Missing attributes** alter causal analyses. Performance of methods depends on the underlying assumptions

Further details in original papers

Mayer, I, J., Wager, S., Sverdrup, E., Moyer, J.D. & Gauss, T. (2020). Doubly robust treatment effect with missing attributes. *Annals of Applied Statistics*

Mayer, I., J., Raimundo, F. & Vert, J.-P. (2020). MissDeepCausal: causal inference from incomplete data using deep latent variable models.

Sbidian, E. *et al.* (2020). Hydroxychloroquine with or without azithromycin and in-hospital mortality or discharge in patients hospitalized for COVID-19 infection: a cohort study of 4,642 in-patients in France.

Future work

- Coupling of observational data and RCT data
- Heterogeneous treatment effects
- Architecture of neural nets with missing values
- More with MNAR data

Missing value website

More information and details on missing values: **R-miss-tastic** platform. (Mayer et al., 2019)

→ Theoretical and practical tutorials, popular datasets, bibliography, workflows (in R and in python), active contributors/researchers in the community, etc.

`rmisstastic.netlify.com`

Interested in contribute to our platform? Feel free to contact us!

MERCI

References

References i

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68.

Jiang, W., Josse, J., Lavielle, M., and Group, T. (2020). Logistic regression with missing covariates?parameter estimation, model selection and prediction within a joint-modeling framework. *Computational Statistics & Data Analysis*, 145:106907.

Kallus, N., Mao, X., and Udell, M. (2018). Causal inference with noisy and missing covariates via matrix factorization. In *Advances in Neural Information Processing Systems*, pages 6921–6932.

Louizos, C., Shalit, U., Mooij, J. M., Sontag, D., Zemel, R., and Welling, M. (2017). Causal effect inference with deep latent-variable models. In *Advances in Neural Information Processing Systems*, pages 6446–6456.

Mattei, P.-A. and Frellsen, J. (2019). MIWAE: Deep generative modelling and imputation of incomplete data sets. In Chaudhuri, K. and Salakhutdinov, R., editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 4413–4423, Long Beach, California, USA. PMLR.

Mayer, I., Josse, J., Tierney, N., and Vialaneix, N. (2019). R-miss-tastic: a unified platform for missing values methods and workflows. *arXiv preprint arXiv:1908.04822*.

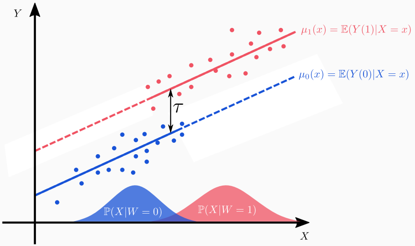
Rosenbaum, P. R. and Rubin, D. B. (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association*, 79(387):516–524.

References ii

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63(3):581–592.

Seaman, S. and White, I. (2014). Inverse probability weighting with missing predictors of treatment assignment or missingness. *Communications in Statistics-Theory and Methods*, 43(16):3499–3515.

Observational data: non random assignment



$\Rightarrow$ Treatment assignment W depends on covariates X.
Distribution of covariates of treated and control are different.

1. Unconfoundedness despite missingness

Adapt the initial assumptions s.t. treatment assignment is unconfounded given only the **observed** covariates and the **response pattern**.

Notations

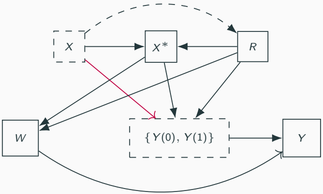
Mask $R \in \{0,1\}^d$, $R_{ij} = 1$ when X_{ij} is missing and 0 otherwise

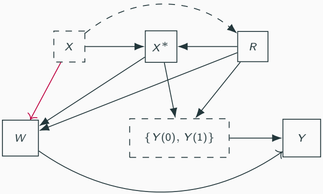
$X^* \triangleq (1-R) \odot X + R \odot NA \in \{\mathbb{R} \cup NA\}^d$

Unconfoundedness despite missingness

$$\{Y_i(1), Y_i(0)\} \perp\!\!\!\perp W_i \mid X^*$$

CIT:$W_i \perp\!\!\!\perp X_i \mid X_i^*, R_i$ **or** CIO:$Y_i(w) \perp\!\!\!\perp X_i \mid X_i^*, R_i$ for $w \in \{0,1\}$

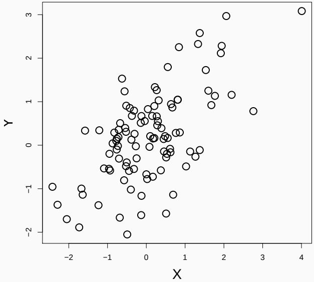




Mean imputation

- $(x_i, y_i) \underset{\text{i.i.d.}}{\sim} \mathcal{N}_2((\mu_x, \mu_y), \Sigma_{xy})$

X	Y
-0.56	-1.93
-0.86	-1.50
.....	...
2.16	0.7
0.16	0.74



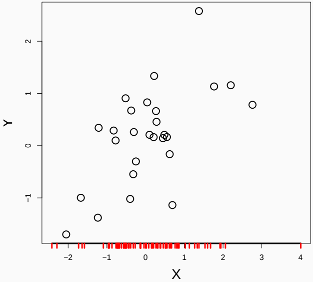
$\mu_y = 0$
$\sigma_y = 1$
$\rho_{xy} = 0.6$

$\hat{\mu}_y = -0.01$
$\hat{\sigma}_y = 1.01$
$\hat{\rho} = 0.66$

Mean imputation

- $(x_i, y_i) \underset{\text{i.i.d.}}{\sim} \mathcal{N}_2((\mu_x, \mu_y), \Sigma_{xy})$
- 70 % of missing entries completely at random on Y

X	Y
-0.56	NA
-0.86	NA
.....	...
2.16	0.7
0.16	NA



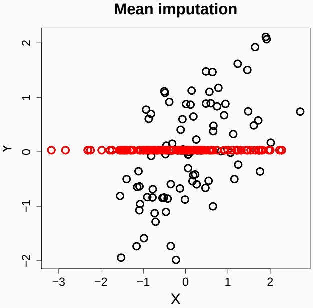
$\mu_y = 0$
$\sigma_y = 1$
$\rho_{xy} = 0.6$

$\hat{\mu}_y = 0.18$
$\hat{\sigma}_y = 0.9$
$\hat{\rho}_{xy} = 0.6$

Mean imputation

- $(x_i, y_i) \underset{\text{i.i.d.}}{\sim} \mathcal{N}_2((\mu_x, \mu_y), \Sigma_{xy})$
- 70 % of missing entries completely at random on Y
- Estimate parameters on the mean imputed data

X	Y
-0.56	**0.01**
-0.86	**0.01**
.....	...
2.16	0.7
0.16	**0.01**

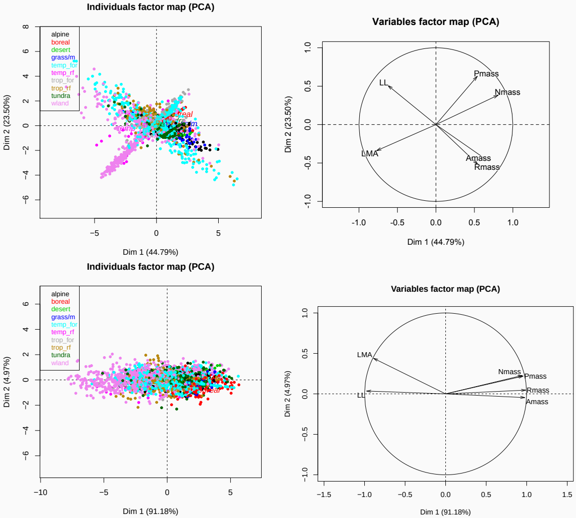


$\mu_y = 0$
$\sigma_y = 1$
$\rho_{xy} = 0.6$

$\hat{\mu}_y = 0.01$
$\hat{\sigma}_y = 0.5$
$\hat{\rho} = 0.30$

Mean imputation deforms joint and marginal distributions

Mean imputation is bad for estimation




```
library(FactoMineR)
PCA(ecolo)
Warning message: Missing
are imputed by the mean
of the variable:
You should use imputePCA
from missMDA
```

```
library(missMDA)
imp <- imputePCA(ecolo)
PCA(imp$comp)
```

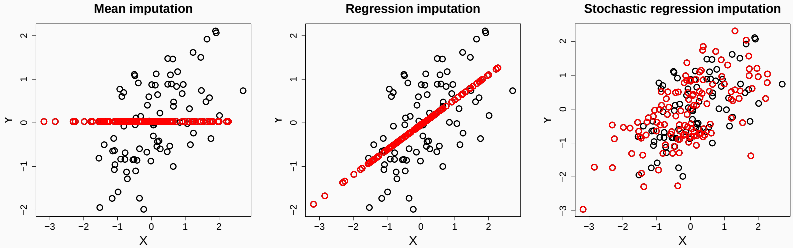
Ecological data: [3] $n = 69000$ species - 6 traits. Estimated correlation between Pmass & Rmass ≈ 0 (mean imputation) or ≈ 1 (EM PCA)

[3]Wright, I. et al. (2004). The worldwide leaf economics spectrum. *Nature*.

Imputation methods

- by regression takes into account the relationship: Estimate β - impute $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i \Rightarrow$ variance underestimated and correlation overestimated
- by stochastic reg: Estimate β and σ - impute from the predictive $y_i \sim \mathcal{N}\left(x_i \hat{\beta}, \hat{\sigma}^2\right) \Rightarrow$ preserve distributions

Here $\hat{\beta}, \hat{\sigma}^2$ estimated with complete data, but MLE can be obtained with EM



	Mean imputation	Regression imputation	Stochastic regression imputation
$\mu_y = 0$	0.01	0.01	0.01
$\sigma_y = 1$	0.5	0.72	0.99
$\rho_{xy} = 0.6$	0.30	0.78	0.59

Imputation methods for multivariate data

Assuming a joint model

- Gaussian distribution: $x_{i.} \sim \mathcal{N}(\mu, \Sigma)$ (Amelia Honaker, King, Blackwell)
- low rank: $X_{n\times d} = \mu_{n\times d} + \varepsilon \; \varepsilon_{ij} \overset{\text{iid}}{\sim} \mathcal{N}\left(0, \sigma^2\right)$ with μ of low rank k (softimpute Hastie & Mazuder; missMDA J. & Husson)
- latent class - nonparametric Bayesian (dpmpm Reiter)
- deep learning using variational autoencoders (MIWAE, Mattei, 2018)

Using conditional models (joint implicitly defined)

- with logistic, multinomial, poisson regressions (mice van Buuren)
- iterative impute each variable by random forests (missForest Stekhoven)

Imputation for categorical, mixed, blocks/multilevel data [4], etc.
$\Rightarrow$ Missing values plateform[5] J., Mayer., Tierney, Vialaneix

[4] J., Husson, Robin & Narasimhan. (2018). Imputation of mixed data with multilevel SVD.
[5] https://rmisstastic.netlify.com/

Mean imputation consistent

Learn on the mean-imputed training data, impute the test set with the **same means** and predict is optimal if the missing data are MAR and the **learning algorithm is universally consistent**

Framework - assumptions

- $Y = f(X) + \varepsilon$
- $X = (X_1, \ldots, X_d)$ has a continuous density $g > 0$ on $[0, 1]^d$
- $\|f\|_\infty < \infty$
- Missing data MAR on X_1 with $R_1 \perp\!\!\!\perp X_1 | X_2, \ldots, X_d$.
- $(x_2, \ldots, x_d) \mapsto P[R_1 = 1 | X_2 = x_2, \ldots, X_d = x_d]$ is continuous
- ε is a centered noise independent of (X, R_1)

(remains valid when missing values occur for variables $X_1, \ldots, X_j$)

Mean imputation consistent

Learn on the mean-imputed training data, impute the test set with the **same means** and predict is optimal if the missing data are MAR and the **learning algorithm is universally consistent**

Mean imputed entry $x' = (x_1', x_2, \dots, x_d)$: $x_1' = x_1 \mathbb{1}_{R_1=0} + \mathbb{E}[X_1]\mathbb{1}_{R_1=1}$

$\tilde{X} = X \odot (1 - R) + \texttt{NA} \odot R$ (takes value in $\mathbb{R} \cup \{\texttt{NA}\}$)

Theorem

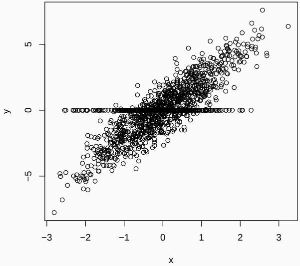
Prediction with mean is equal to the Bayes function almost everywhere

$$f^{\star}_{impute}(x') = \mathbb{E}[Y|X^* = x^*]$$

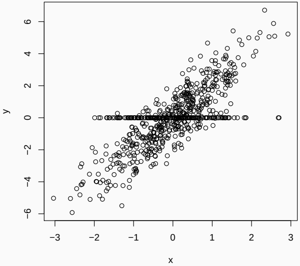
Other values than the mean are OK but use the same value for the train and test sets, otherwise the algorithm may fail as the distributions differ

Consistency of supervised learning with NA: Rationale

- Specific value, systematic like a code for missing
- The learner detects the code and recognizes it at the test time
- With categorical data, just code "Missing"
- With continuous data, any constant:
- Need a lot of data (asymptotic result) and a super powerful learner



Train

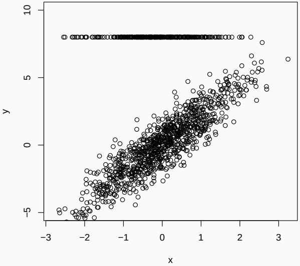


Test

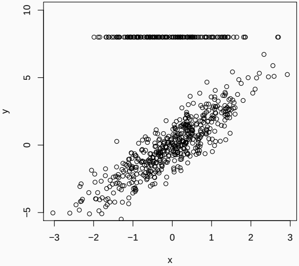
Mean imputation not bad for prediction; it is consistent; despite its drawbacks for estimation - Useful in practice!

Consistency of supervised learning with NA: Rationale

- Specific value, systematic like a code for missing
- The learner detects the code and recognizes it at the test time
- With categorical data, just code "Missing"
- With continuous data, any constant: out of range
- Need a lot of data (asymptotic result) and a super powerful learner



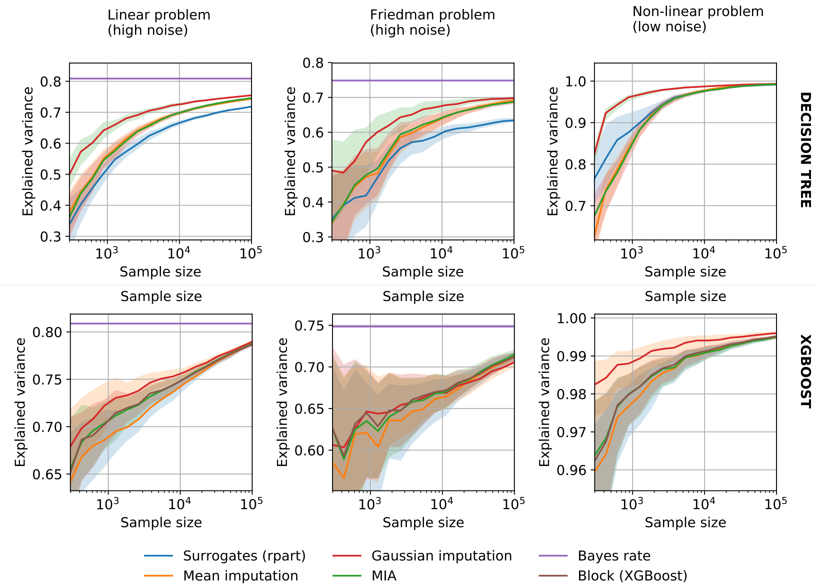
Train



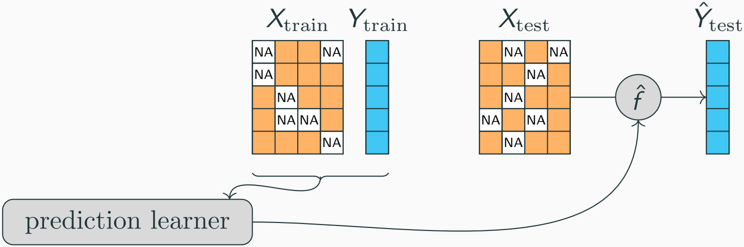
Test

Mean imputation not bad for prediction; it is consistent; despite its drawbacks for estimation - Useful in practice!

Consistency: 40% missing values MCAR



End-to-end learning with missing values

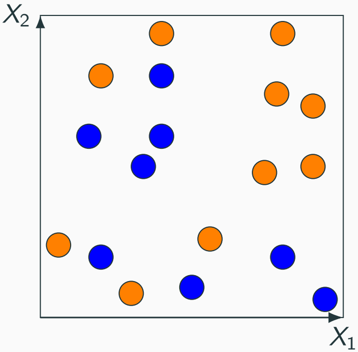


- Random forests powerful learner
- Trees well suited for empirical risk minimization with missing values: Handle half discrete data X^* that takes values in $\mathbb{R} \cup \{\texttt{NA}\}$

CART (Breiman, 1984)

Built recursively by splitting the current cell into two children: Find the feature $j^\star$, the threshold $z^\star$ which minimises the (quadratic) loss

$$(j^\star, z^\star) \in \underset{(j,z)\in\mathcal{S}}{\arg\min}\ \mathbb{E}\Big[\big(Y - \mathbb{E}[Y|X_j \leq z]\big)^2 \cdot \mathbb{1}_{X_j \leq z} + \big(Y - \mathbb{E}[Y|X_j > z]\big)^2 \cdot \mathbb{1}_{X_j > z}\Big].$$

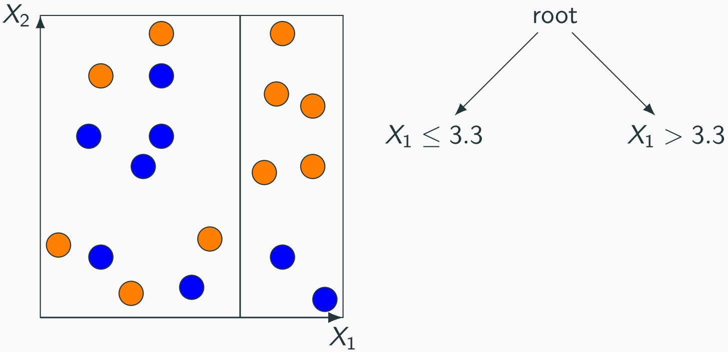


root

CART (Breiman, 1984)

Built recursively by splitting the current cell into two children: Find the feature $j^\star$, the threshold $z^\star$ which minimises the (quadratic) loss

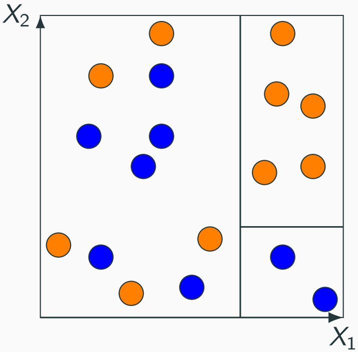
$$(j^\star, z^\star) \in \underset{(j,z)\in\mathcal{S}}{\arg\min}\ \mathbb{E}\Big[\big(Y - \mathbb{E}[Y|X_j \leq z]\big)^2 \cdot \mathbb{1}_{X_j \leq z} \\ + \big(Y - \mathbb{E}[Y|X_j > z]\big)^2 \cdot \mathbb{1}_{X_j > z}\Big].$$

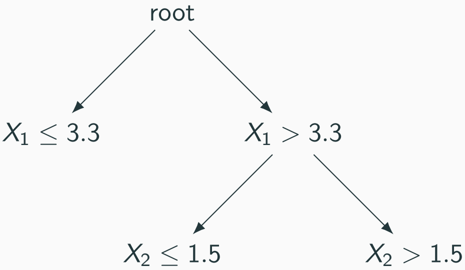


CART (Breiman, 1984)

Built recursively by splitting the current cell into two children: Find the feature $j^\star$, the threshold $z^\star$ which minimises the (quadratic) loss

$$(j^\star, z^\star) \in \mathop{\arg\min}_{(j,z)\in\mathcal{S}} \mathbb{E}\Big[\big(Y - \mathbb{E}[Y|X_j \le z]\big)^2 \cdot \mathbb{1}_{X_j \le z} \\ + \big(Y - \mathbb{E}[Y|X_j > z]\big)^2 \cdot \mathbb{1}_{X_j > z}\Big].$$





CART with missing values

root

	X_1	X_2	Y
1			
2	NA		
3	NA		
4			

CART with missing values

root

$X_1 \leq s_1$ $\qquad$ $X_1 > s_1$

	X_1	X_2	Y
1			
~~2~~	~~NA~~		
~~3~~	~~NA~~		
4			

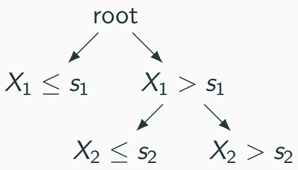
1) Select variable and threshold on observed data [6]

$$\mathbb{E}\Big[\big(Y - \mathbb{E}[Y|X_j \leq z, R_j = 0]\big)^2 \cdot \mathbb{1}_{X_j \leq z, R_j = 0} + \big(Y - \mathbb{E}[Y|X_j > z, R_j = NA]\big)^2 \cdot \mathbb{1}_{X_j > z, R_j = 0}\Big].$$

[6] Variable selection bias (not a problem to predict): `partykit` package, Hothorn, et al.

CART with missing values

	X_1	X_2	Y
1			
2	~~NA~~		
3	~~NA~~		
4			



1) Select variable and threshold on observed data [6]

$$\mathbb{E}\Big[\big(Y - \mathbb{E}[Y|X_j \leq z, R_j = 0]\big)^2 \cdot \mathbb{1}_{X_j \leq z, R_j = 0} + \big(Y - \mathbb{E}[Y|X_j > z, R_j = NA]\big)^2 \cdot \mathbb{1}_{X_j > z, R_j = 0}\Big].$$

2) Propagate observations (2 & 3) with missing values?

- Probabilistic split: $\mathcal{B}ernoulli(\frac{\#L}{\#L+\#R})$ (`Rweeka`)
- Block: Send all to a side by minimizing the error (`xgboost`, `lightgbm`)
- Surrogate split: Search another variable that gives a close partition (rpart)

[6] Variable selection bias (not a problem to predict): `partykit` package, Hothorn, et al.

Missing incorporated in attribute (Twala et al. 2008)

One step: select the variable, the threshold and propagate missing values.
Use missingness to make the best possible splits.

$$f^\star \in \underset{f \in \mathcal{P}_{c,miss}}{\arg\min} \mathbb{E}\Big[\big(Y - f(X^*)\big)^2\Big],$$

where $\mathcal{P}_{c,miss} = \mathcal{P}_{c,miss,L} \cup \mathcal{P}_{c,miss,R} \cup \mathcal{P}_{c,miss,sep}$ with

1. $\mathcal{P}_{c,miss,L} \ \rightarrow \ \{\{X_j^* \leq z \vee X_j^* = \mathtt{NA}\}, \{X_j^* > z\}\}$
2. $\mathcal{P}_{c,miss,R} \ \rightarrow \ \{\{X_j^* \leq z\}, \{X_j^* > z \vee X_j^* = \mathtt{NA}\}\}$
3. $\mathcal{P}_{c,miss,sep} \ \rightarrow \ \{\{X_j^* \neq \mathtt{NA}\}, \{X_j^* = \mathtt{NA}\}\}$.

- Missing values treated like a category (well to handle $\mathbb{R} \cup \mathtt{NA}$)
- Good for informative pattern (R explains Y)
- Implementation trick: duplicate the incomplete columns, and replace the missing entries once by $+\infty$ and once by $-\infty$ (J. Tibshirani) [7]

Target model/pattern: $\mathbb{E}\left[Y|X^*\right] = \sum_{r \in \{0,1\}^d} \mathbb{E}\left[Y|X_{obs(r)}, R = r\right] \mathbb{1}_{R=r}$

Does not require the missing data to be MAR.

[7] Implemented for conditional forests `partykit`, generalized random forest `grf`, `scikitlearn`