ISCB42

Multiple imputation in propensity score matching: obtaining correct confidence intervals

Corentin Ségalas, Clémence Leyrat, James Carpenter and Elizabeth Williamson

LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE

Table of Contents

Statistical context: causal inference

Observational study with:

- a binary outcome Y
- a binary exposure Z (1 if treated, 0 if not)
- a vector of baseline covariates $X^\top$, all potential confounders

Statistical context: causal inference

Observational study with:

- a binary outcome Y
- a binary exposure Z (1 if treated, 0 if not)
- a vector of baseline covariates $X^\top$, all potential confounders

Goal is to estimate the average treatment effect on the treated (ATT) defined as:

$$ATT = E(y_i^1 | z_i = 1) - E(y_i^0 | z_i = 1)$$

where y_i^1 and y_i^0 are the potential outcomes for patient i if treated and not treated respectively and z_i the treatment indicator for patient i.

Propensity score (PS)

Definition (Rosenbaum and Rubin, 1983)

The propensity score for patient i is defined as

$$ps_i = P(z_i = 1|x_i)$$

and can be estimated using a logistic regression or more advanced techniques (Westreich et al., 2010)

Propensity score (PS)

Definition (Rosenbaum and Rubin, 1983)

The propensity score for patient i is defined as

$$ps_i = P(z_i = 1|x_i)$$

and can be estimated using a logistic regression or more advanced techniques (Westreich et al., 2010)

Under technical assumptions, the PS can be used to build an unbiased estimator of the true ATT using different methods: *PS matching, PS stratification, inverse probability of treatment weighting, etc.*

Propensity score matching

The goal is to match each treated patient to at least one untreated patient based on the distance between their PS.

- matching algorithm
- metric used for the distance
- caliper whose size limits the difference between a pair
- number of non-treated patients matched to each treated patient
- sampling with or without replacement

Propensity score matching

The goal is to match each treated patient to at least one untreated patient based on the distance between their PS.

- matching algorithm
- metric used for the distance
- caliper whose size limits the difference between a pair
- number of non-treated patients matched to each treated patient
- sampling with or without replacement

$\Longrightarrow$ direct estimation of the ATT by comparing outcomes between treated and non-treated on the matched dataset

Propensity score matching

The goal is to match each treated patient to at least one untreated patient based on the distance between their PS.

- matching algorithm
- metric used for the distance
- caliper whose size limits the difference between a pair
- number of non-treated patients matched to each treated patient
- sampling with or without replacement

$\Longrightarrow$ direct estimation of the ATT by comparing outcomes between treated and non-treated on the matched dataset
$\Longrightarrow$ loss of power due to the discarding of many patients from final analysis. The method is still widely used and have an intuitive understanding.

Multiple imputation and Rubin's rules

If data is missing in $X^\top$, PS can not be estimated. Under missing at random (MAR) mechanism, multiple imputation (MI) can be used to create m complete datasets.

For each complete dataset k, PS matching is done and an estimate $\hat{\theta}_k$ of the ATT is computed. The $(\hat{\theta}_k)_k$ are aggregated using Rubin's rules (Leyrat et al. 2019, Granger et al. 2019):

$$\hat{\theta} = \frac{1}{m}\sum_{k=1}^{m}\hat{\theta}_k, \quad \widehat{Var}(\hat{\theta}) = W + \left(1 + \frac{1}{m}\right) B$$

where

$$W = \frac{1}{m}\sum_{k=1}^{m}\widehat{Var}(\hat{\theta}_k), \quad B = \frac{1}{m-1}\sum_{k=1}^{m}(\hat{\theta}_k - \hat{\theta})^2.$$

Multiple imputation and Rubin's rules

If data is missing in $X^\top$, PS can not be estimated. Under missing at random (MAR) mechanism, multiple imputation (MI) can be used to create m complete datasets.

For each complete dataset k, PS matching is done and an estimate $\hat{\theta}_k$ of the ATT is computed. The $(\hat{\theta}_k)_k$ are aggregated using Rubin's rules (Leyrat et al. 2019, Granger et al. 2019):

$$\hat{\theta} = \frac{1}{m}\sum_{k=1}^{m}\hat{\theta}_k, \quad \widehat{Var}(\hat{\theta}) = W + \left(1 + \frac{1}{m}\right) B$$

where

$$W = \frac{1}{m}\sum_{k=1}^{m}\widehat{Var}(\hat{\theta}_k), \quad B = \frac{1}{m-1}\sum_{k=1}^{m}(\hat{\theta}_k - \hat{\theta})^2.$$

Motivation behind this project

PS matching is known to reduce sample size: especially true when the outcome is rare. In a simulation study, over-coverage was observed for the independent PS matching

Literature insight (Reiter, 2008)

Reiter states that, in the context of measurement error, using multiple imputation leads to over-coverage because a lot of patients who contributed to the imputation model are discarded prior to the analysis model. He proposed a correction to Rubin's rules.

Motivation behind this project

PS matching is known to reduce sample size: especially true when the outcome is rare. In a simulation study, over-coverage was observed for the independent PS matching

Literature insight (Reiter, 2008)

Reiter states that, in the context of measurement error, using multiple imputation leads to over-coverage because a lot of patients who contributed to the imputation model are discarded prior to the analysis model. He proposed a correction to Rubin's rules.

Does the same phenomenon might be happening when combining multiple imputation and propensity score matching?

Combining multiple imputation and PS matching

Objectives:

1. Does the discarding of unmatched individuals lead to over-coverage when combining multiple imputation and propensity score matching using Rubin's rules?

2. Implement the Reiter's correction in the context of propensity score matching and assess its performance

Table of Contents

From Rubin's rules to Reiter's rules

- Reiter has observed that Rubin's rules could lead to inflated variance when some patients contributed to the imputation model but not to the analysis model

- Reiter proposed to create r (instead of 1) complete datasets for each parameter draw leading to a total of $m \times r$ complete datasets

- This additional variability is used to reduce the inflation of the variance

Reiter's rules

$$\hat{\theta} = \frac{1}{mr}\sum_{k=1}^{m}\sum_{j=1}^{r}\hat{\theta}_{k,j} = \frac{1}{m}\sum_{k=1}^{m}\hat{\theta}_k,$$

$$\widehat{Var}(\hat{\theta}) = W + \left(1 + \frac{1}{m}\right)B - \left(1 + \frac{1}{r}\right)U,$$

where

$$W = \frac{1}{mr}\sum_{k=1}^{m}\sum_{j=1}^{r}\widehat{Var}(\hat{\theta}_{k,j}), \quad B = \frac{1}{m-1}\sum_{k=1}^{m}\left(\hat{\theta}_k - \hat{\theta}\right)^2,$$

$$U = \frac{1}{m(r-1)}\sum_{k=1}^{m}\sum_{j=1}^{r}\left(\hat{\theta}_{k,j} - \hat{\theta}_k\right)^2.$$

Reiter's rules

$$\hat{\theta} = \frac{1}{mr}\sum_{k=1}^{m}\sum_{j=1}^{r}\hat{\theta}_{k,j} = \frac{1}{m}\sum_{k=1}^{m}\hat{\theta}_k,$$

$$\widehat{Var}(\hat{\theta}) = \underbrace{W + \left(1+\frac{1}{m}\right)B}_{\text{Rubin's variance}} - \left(1+\frac{1}{r}\right)U,$$

where

$$W = \frac{1}{mr}\sum_{k=1}^{m}\sum_{j=1}^{r}\widehat{Var}(\hat{\theta}_{k,j}), \quad B = \frac{1}{m-1}\sum_{k=1}^{m}\left(\hat{\theta}_k - \hat{\theta}\right)^2,$$

$$U = \frac{1}{m(r-1)}\sum_{k=1}^{m}\sum_{j=1}^{r}\left(\hat{\theta}_{k,j} - \hat{\theta}_k\right)^2.$$

Implementation in `R`

- In `mice`, no easy access to the imputation model
- The argument `trace` was designed to impute a dataset while making the distinction between:
  - training data used to estimate the imputation model
  - validation data not used to estimate the imputation model
- By repeating the original dataset *r* times, and using the `trace` argument, it is easy to impute *r* times using a same parameter draw. This can be done for each *m* parameter draw.

Table of Contents

Aims, data generation mechanisms & estimands

- Aims: assess the impact of discarding patients between imputation and estimation and evaluate Reiter's rules in this context

- Data generation mechanisms:
  - $N = 1,000$ datasets with $10,000$ patients
  - three confounders $x = (x_1, x_2, x_3) \sim \mathcal{N}(0, I_3)$
  - three levels of confounding: strong, moderate and weak
  - around 15% of missing at random x_2

- Estimands: ATT as an odds-ratio

Methods & performance measures

- Method implemented using `R`:
  - multiple imputation using `mice`
  - PS estimation using `glm`
  - PS matching using `MatchIt`
  - ATT estimation using the function `glm.cluster` from `miceadds` to take into account the matched nature of the data
  - aggregation of the results using either Rubin's or Reiter's rules

- Performance measures: relative bias, empirical standard error (EmpSE), average model standard error (ModSE), 95% confidence intervals coverage rate (CR)

Results: Rubin's rules

Table: Results for the 1,000 replicates of the ATT estimation using Rubin's rules

Confounding	% of treated	Rel. bias	EmpSE	ModSE	CR
Strong	30	-0.010	0.049	0.063	0.985
Strong	20	-0.001	0.051	0.074	0.996
Strong	10	0.005	0.062	0.106	0.999
Moderate	30	-0.001	0.046	0.065	0.994
Moderate	20	0.002	0.051	0.077	0.996
Moderate	10	0.004	0.064	0.112	1.000
Weak	30	-0.001	0.064	0.081	0.989
Weak	20	0.000	0.073	0.097	0.990
Weak	10	0.004	0.102	0.139	0.994

Results: Rubin's rules

Table: Results for the 1,000 replicates of the ATT estimation using Rubin's rules

Confounding	% of treated	Rel. bias	EmpSE	ModSE	CR
Strong	30	-0.010	0.049	0.063	0.985
Strong	20	-0.001	0.051	0.074	0.996
Strong	10	0.005	0.062	0.106	0.999
Moderate	30	-0.001	0.046	0.065	0.994
Moderate	20	0.002	0.051	0.077	0.996
Moderate	10	0.004	0.064	0.112	1.000
Weak	30	-0.001	0.064	0.081	0.989
Weak	20	0.000	0.073	0.097	0.990
Weak	10	0.004	0.102	0.139	0.994

Results: Rubin's rules

Table: Results for the 1,000 replicates of the ATT estimation using Rubin's rules

Confounding	% of treated	Rel. bias	EmpSE	ModSE	CR
Strong	30	-0.010	0.049	0.063	0.985
Strong	20	-0.001	0.051	0.074	0.996
Strong	10	0.005	0.062	0.106	0.999
Moderate	30	-0.001	0.046	0.065	0.994
Moderate	20	0.002	0.051	0.077	0.996
Moderate	10	0.004	0.064	0.112	1.000
Weak	30	-0.001	0.064	0.081	0.989
Weak	20	0.000	0.073	0.097	0.990
Weak	10	0.004	0.102	0.139	0.994

Results: Rubin's rules

Table: Results for the 1,000 replicates of the ATT estimation using Rubin's rules

Confounding	% of treated	Rel. bias	EmpSE	ModSE	CR
Strong	30	-0.010	0.049	0.063	0.985
Strong	20	-0.001	0.051	0.074	0.996
Strong	10	0.005	0.062	0.106	0.999
Moderate	30	-0.001	0.046	0.065	0.994
Moderate	20	0.002	0.051	0.077	0.996
Moderate	10	0.004	0.064	0.112	1.000
Weak	30	-0.001	0.064	0.081	0.989
Weak	20	0.000	0.073	0.097	0.990
Weak	10	0.004	0.102	0.139	0.994

Results: Reiter's rules

Table: Results for the 1,000 replicates of the ATT estimation using Reiter's rules

Confounding	% of treated	Rel. bias	EmpSE	ModSE	CR
Strong	30	0.014	0.046	0.047	0.937
Strong	20	0.000	0.051	0.050	0.950
Strong	10	0.000	0.061	0.061	0.950
Moderate	30	0.001	0.048	0.046	0.933
Moderate	20	0.001	0.049	0.050	0.956
Moderate	10	0.002	0.063	0.064	0.958
Weak	30	0.001	0.066	0.065	0.946
Weak	20	0.000	0.076	0.074	0.940
Weak	10	0.002	0.097	0.100	0.958

Results: Reiter's rules

Table: Results for the 1,000 replicates of the ATT estimation using Reiter's rules

Confounding	% of treated	Rel. bias	EmpSE	ModSE	CR
Strong	30	0.014	0.046	0.047	0.937
Strong	20	0.000	0.051	0.050	0.950
Strong	10	0.000	0.061	0.061	0.950
Moderate	30	0.001	0.048	0.046	0.933
Moderate	20	0.001	0.049	0.050	0.956
Moderate	10	0.002	0.063	0.064	0.958
Weak	30	0.001	0.066	0.065	0.946
Weak	20	0.000	0.076	0.074	0.940
Weak	10	0.002	0.097	0.100	0.958

Results: Reiter's rules

Table: Results for the 1,000 replicates of the ATT estimation using Reiter's rules

Confounding	% of treated	Rel. bias	EmpSE	ModSE	CR
Strong	30	0.014	0.046	0.047	0.937
Strong	20	0.000	0.051	0.050	0.950
Strong	10	0.000	0.061	0.061	0.950
Moderate	30	0.001	0.048	0.046	0.933
Moderate	20	0.001	0.049	0.050	0.956
Moderate	10	0.002	0.063	0.064	0.958
Weak	30	0.001	0.066	0.065	0.946
Weak	20	0.000	0.076	0.074	0.940
Weak	10	0.002	0.097	0.100	0.958

Results: Reiter's rules

Table: Results for the 1,000 replicates of the ATT estimation using Reiter's rules

Confounding	% of treated	Rel. bias	EmpSE	ModSE	CR
Strong	30	0.014	0.046	0.047	0.937
Strong	20	0.000	0.051	0.050	0.950
Strong	10	0.000	0.061	0.061	0.950
Moderate	30	0.001	0.048	0.046	0.933
Moderate	20	0.001	0.049	0.050	0.956
Moderate	10	0.002	0.063	0.064	0.958
Weak	30	0.001	0.066	0.065	0.946
Weak	20	0.000	0.076	0.074	0.940
Weak	10	0.002	0.097	0.100	0.958

Table of Contents

Application: context

National Cancer Registry of the Office for National Statistics:

- $31,351$ patients diagnosed with cancer
- covariates: stage of the cancer, sex of the patient, patient's level of deprivation, comorbidity (CharIson score) and the patient's performance status
- 25% of performance and 10% of stage data were missing

We have studied the effect of age at diagnosis as a binary variable (median as the cutoff) on the risk of surgery

- impact of $(m; r) = (20; 10), (20; 30), (30; 10), (30; 30), (50; 10)$ and $(50; 30)$
- impact of random fluctuation: 1604 and 1993 as seeds

Application: results

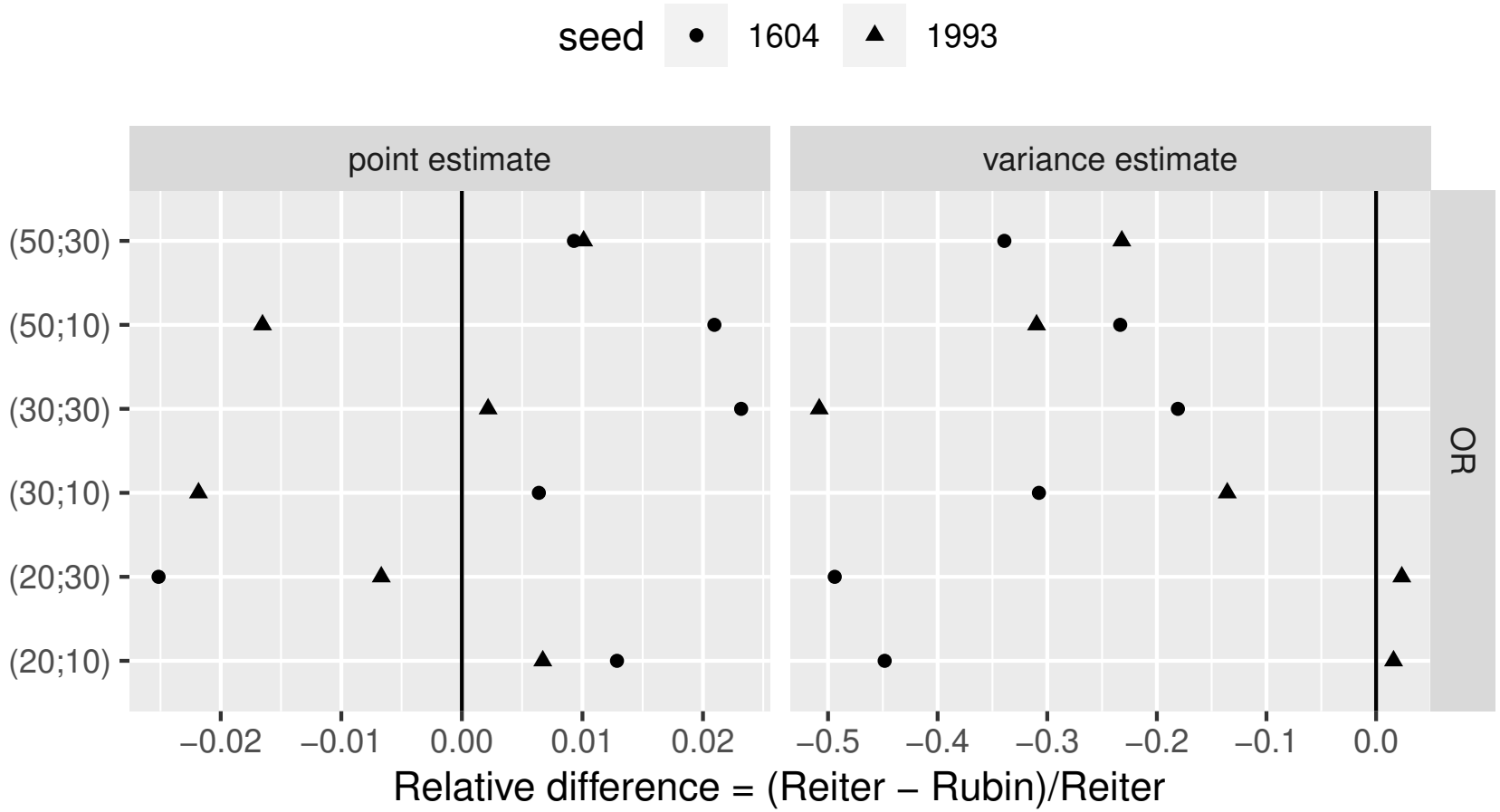

Table of Contents

- Combination of MI and PS matching using Rubin's rules can lead to inflated variance

- Combination of MI and PS matching using Rubin's rules can lead to inflated variance
- Reiter's rules were able to correct the inflation

- Combination of MI and PS matching using Rubin's rules can lead to inflated variance
- Reiter's rules were able to correct the inflation
- Focus on PS matching only

- Combination of MI and PS matching using Rubin's rules can lead to inflated variance
- Reiter's rules were able to correct the inflation
- Focus on PS matching only
- Easy to implement in `R`

- Combination of MI and PS matching using Rubin's rules can lead to inflated variance
- Reiter's rules were able to correct the inflation
- Focus on PS matching only
- Easy to implement in R
- Can become computationally intense ($m \times r$ imputation) with bigger sample sizes

- Combination of MI and PS matching using Rubin's rules can lead to inflated variance
- Reiter's rules were able to correct the inflation
- Focus on PS matching only
- Easy to implement in R
- Can become computationally intense ($m \times r$ imputation) with bigger sample sizes
- What about full matching?

- Combination of MI and PS matching using Rubin's rules can lead to inflated variance
- Reiter's rules were able to correct the inflation
- Focus on PS matching only
- Easy to implement in R
- Can become computationally intense ($m \times r$ imputation) with bigger sample sizes
- What about full matching?

Take home message

Be careful when combining multiple imputation and propensity score matching

References

- Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. Biometrika
- Westreich, D., Lessler, J., and Funk, M. J. (2010). Propensity score estimation: neural networks, support vector machines, decision trees (CART), and meta-classifiers as alternatives to logistic regression. Journal of Clinical Epidemiology
- Granger, E., Sergeant, J. C., and Lunt, M. (2019). Avoiding pitfalls when combining multiple imputation and propensity scores. Statistics in Medicine
- Leyrat, C., Seaman, S. R., White, I. R., Douglas, I., Smeeth, L., Kim, J., Resche-Rigon, M., Carpenter, J. R., and Williamson, E. J. (2019). Propensity score analysis with partially observed covariates: How should multiple imputation be used? Statistical Methods in Medical Research
- Reiter, J. P. (2008). Multiple Imputation When Records Used for Imputation Are Not Used or Disseminated for Analysis. Biometrika

Thanks for your attention!

corentin.segalas@lshtm.ac.uk

Technical assumptions

1. stable unit treatment value assumption (SUTVA)

$$y_i^1, y_i^0 \perp z_j, \quad \forall i \neq j$$

2. consistency

$$y_i = y^{z_i}, \quad \forall i$$

3. positivity

$$0 < P(z_i = 1 | x_i) < 1, \quad \forall i$$

4. strongly ignorable treatment assignment (SITA)

$$y_i^0, y_i^1 \perp z_i | x_i, \quad \forall i$$