Greedy (nearest neighbour) matching

04.05.21 13:14

Set-up:

- X ... pre-treatment covariate
- d_{ij} ... distance btw. treated subject i and control subject j
- many more controls than treated subjects
  ↳ in real-world this is often the case
- focus on pair-matching (1-to-1 matching) for now → less bias bc the matches are closer

Greedy matching:

- steps: (1) randomly order list of treated subjects and control subjects
  (2) start with treated subject i=1. Match it to the control j with the smallest distance d_{1j}. (this is greedy property)
  (3) remove the matched control subject j from the list of available matches
  (4) move on to treated subject i=2. Match it to the control j with the smallest distance d_{2j}.
  (5) repeat steps (3) and (4) until all treated subjects are matched with one control subject.

Example: Greedy matching:

Treated Subjects

i	Age	COPD	Female
1	78.17	0	1
2	67.91	0	0

Available Controls

j	Age	COPD	Female
1	70.25	1	0
2	75.33	0	1
3	86.08	1	1
4	54.97	0	0
5	43.63	0	0
6	18.04	0	1

distance d_{1j}: 4.23, 0.17 ←, 3.72, 2.45, 2.89, 3.60

distance d_{2j}: 3.70, 2.06 already matched, 4.34, 0.78 ←, 1.45, 3.60

1st match: i=1 & j=2 ⇒ bc $\min(X_1, X_{j1}) = 0.17$ for j=2. (match id 1)

2nd match: i=2 & j=4 ⇒ bc $\min(X_2, X_{j+2}) = 0.78$ for j=4. (match id 2)

Matched dataset:

Match ID	treated	age	COPD	female
1	1	78.17	0	1
1	0	75.33	0	1
2	1	67.91	0	0
2	0	54.97	0	0

Greedy Matching: (order of the initial list of subjects matters)

- comp. very fast, even for large datasets
  ↳ R package: MatchIt
- not invariant to order of list of treated subjects
- not optimal
  ↳ always choosing smallest distance does not minimize total distance (globally)
  ↳ can lead to bad matches

Many-to-one matching: (greedy method)

- k:1 matching: after every treated subject has 1 control match, go through the list of treated subjects again and find 2nd closest matches.
  ↳ repeat until k-matches assigned to each treated subject i.
  - ↳ advantage: larger sample size of the matched dataset.
  - ↳ disadvantage: increased bias bc the combined matching is less close / accurate.
  - ↳ advantage: reduced variance bc we have more data to compute the causal effect estimates.

Caliper: (maximum distance tolerance limit)

idea: remove treated subjects for whom we do not have a good control match
⇒ a "bad" match can be defined using a caliper (maximum acceptable distance)

- only match two subjects (treated & control) if the best match between treated subject i and control subject j ($d_{ij}^{\min}$ < caliper-value)

⇒ if no matches within cliper, it is a sign that positivity assumption would be violated! ⇒ therefore: excluding these treated subjects using the caliper would make the positivity assumption more realistic