Propensity score matching with clustered data.

19th European Young Statisticians Meeting

Prague, the Czech Republic, 30 August - 4 September 2015

Massimo Cannas

University of Cagliari, Italy

joint work with

Bruno Arpino, *Pompeu Fabra University, Spain*

Outline

- Motivating case study: the effect of caesarean section on the Apgar score
- Propensity score methods for causal inference
- Existing studies with multilevel data
- Simulations design and results
- Case study results
- Concluding remarks

Motivating case study (1/4)

- Estimation of the causal effect of **caesarean section** (treatment) versus **natural delivery** (control) on a widely employed indicator of the clinical state of newborns, the 5-minute Apgar score (AS).
- The AS is a composite measure of breathing effort, heart rate, muscle tone, reflexes and skin color. Each item is scored 0, 1, or 2, and thus the total score ranges from 0 to 10.
- Infants with a score of ≥ 7 are usually considered normal (American Academy Pediatrics 2006). Low AS is strongly associated with abnormal future development of the child and infant mortality risk.

Motivating case study (2/4)

- Y = 1 if AS < 7 (“low” AS), = 0 otherwise.
- Our dataset covers all hospitalized deliveries Sardinia, years 2010 and 2011. The source is the official form on the birth event (known as CedAP).
- We focus on the subset of **non-complicated pregnancies**: women delivering at 32 or more weeks of gestational age with a singleton and living infant in vertex (head-down) position, without birth anomalies. We further restrict the sample to nulliparous mothers aged between 15 and 44. Our working sample includes 14,757 observations clustered in 20 hospitals.

Motivating case study (3/4)

- Unbalanced structure

- Proportion of treated < than proportion of control units

Hospital	N. births	N. caesarean sections	% caesarean sections
1	2,532	1,166	46.0
2	1,788	623	34.8
3	1,687	540	32.0
4	1,473	632	42.9
5	1,253	410	32.7
6	1,197	428	35.7
7	980	240	24.4
8	875	238	27.2
9	529	190	35.9
10	434	135	31.1
11	403	164	40.6
12	396	117	29.5
13	351	134	38.1
14	266	74	27.8
15	208	99	47.5
16	191	122	63.8
17	103	40	38.8
18	50	9	18.0
19	32	13	40.6
20	9	1	11.1

Motivating case study (4/4)

- Selection mechanism: what are the factors influencing caesarean section?
- Individual-level: maternal age and education, infant weight, gestational age, pathologies during pregnancy.
- Hospital-level: hospital practices and culture, managerial preferences and guidelines, volume, type (teaching/not teaching), etc. (Caceras et al, 2013; Bragg et al, 2010).

Potential outcome framework

- Consider a group of units, indexed by i = 1,…,N.
- Let T_i be a binary treatment indicator: = 1 if mother i delivered with caesarean section (treated), = 0 otherwise.
- Let $Y_i(1)$ and $Y_i(0)$ denote the potential outcomes on the mother's infant (Apgar score).
- Causal estimand of interest: ATT = E[Y(1) - Y(0) | T = 1].
- $Y_i(0)$ is always unobserved for $T_i = 1$.

Propensity score (PS) methods

- Identifying assumptions:
  - **$Y(1), Y(0) \perp T \mid X$ (unconfoundedness)**
  - $0 < P(T=1 \mid X) < 1$ (overlap)

- PS: $e(X) \equiv Pr\{T = 1|X\} = E\{T|X\}$.
- Rosenbaum and Rubin (1983):
  - the propensity score is a balancing score, i.e., $X \perp T \mid e(X)$,
  - if unconfoundedness holds, then $Y(1), Y(0) \perp T \mid e(X)$.

- These results justify matching / stratification / weighting on e(X) instead than on X.

Clustered data structures

- Clustered data structures are very common in many fields (patients into hospitals, individuals into geographical areas, students into schools)
- PS methods have been developed and applied in the context of unstructured data.
- In clustered data bias can arise from omitted individual and/ or cluster-level confounders.
- **How should we apply PS methods to these data?**
- **How can we use knowledge on clusters' memberships?**
- Few methodological and applied works exist in the case of clustered data.

Existing studies with clustered data

- Arpino and Mealli (2011)
  - Show the benefit of using random or fixed effects models for the estimation of the propensity score to reduce the bias due to unmeasured cluster-level variables in PS matching (PSM).
  - Focus on high number of small clusters.

- Thoemmes and West (2011) and Li et al (2013) considered stratification and re-weighting using PS, respectively.

Our contribution

- Unbalanced data structure with both big and small clusters.
- Realistic simulated dataset that mimic real data.
- We compare different approaches:

Strategy	PS model	Matching criteria
A	Single-level logit	Pooled
B	Single-level logit	(pure) Within-cluster
C	Single-level logit	“Preferential” within-cluster
D	Random-effect logit	Pooled
E	Fixed-effect logit	Pooled

Approach A (pooled matching)

- It ignores the clustered structure in both PS estimation:

$$\text{logit}(e_{ij}) = \alpha_0 + X_{ij}\beta \qquad (1)$$

- and matching

$$A_{rj} = \{kj' \in I_0 : \hat{e}_{kj'} = \min_{kj' \in I_0} \left| \hat{e}_{rj} - \hat{e}_{kj'} \right| < 0.2\hat{\sigma}_e\} \qquad (2)$$

- We use one-to-one nearest neighbor matching within a caliper of 0.2 standard deviation of the estimated PS (both with and without replacement).

Estimating the ATT

- After the matching algorithm has been applied on each treated unit, the matched dataset is built:

$$M = \{rj : A_{rj} \neq \varnothing\} \cup \left\{ \bigcup_{rj} A_{rj} \right\} \quad (3)$$

- and the ATT is estimated on this set using:

$$\hat{ATT} = \frac{1}{card(M)} \left\{ \sum_{rj \in I_1 \cap M} \left(Y_{rj} - \sum_{kj'} Y_{kj'} w(rj, kj') \right) \right\} \quad (4)$$

Approach B (match within)

- Uses the same PS model than method A (2) but adjusts for clustering in the implementation of the matching that is forced to be within-cluster:

$$A_{rj} = \{kj \in I_0 : \hat{e}_{kj} = \min_{kj \in I_0} \left| \hat{e}_{rj} - \hat{e}_{kj} \right| < 0.2\hat{\sigma}_e \} \quad (5)$$

- Within-cluster matching automatically guarantees that all cluster-level variables are perfectly balanced. But balance of individual-level variables could be worse than with approach A. Also the no. of unmatched units will be higher.

Approach C (preferential-within)

- Tries to combine the benefits of approaches A and B.
- Starts by searching control units within-cluster (according to (5)). If none is found, control units are searched in other clusters (according to (2)).
- It is expected to improve the balancing of cluster-level variables with respect to approach A and reduces the loss of units compared to approach B.

Approaches D and E

- They keep clustering into account in the estimation of the propensity score:

$$\text{logit}(e_{ij}) = \alpha_j + X_{ij}\beta \qquad (6)$$

by estimating cluster-specific random (D) or fixed (E) intercepts.

Simulation studies

- Mimic the real dataset in: no. of clusters, sample sizes, X variables, association between X, T and Y.
- We introduce an unobserved hospital-level confounder, H and consider different effect sizes in the treatment equation (small, medium, high).
- We apply all methods A-E (with and without replacement) omitting the hospital-level simulated variable.
- 500 replicates.

Simulation results: w/o replacement, small H effect

METRIC	PSM Method				
	Raw	*A*	*B*	*C*	*D*	*E*
% of Unmatched	0	8.53	10.45	2.50	9.21	9.17
Balance X	13.01	1.01	1.37	1.22	1.11	0.93
Balance H	17.94	18.09	0	4.22	1.21	0.92
Bias ATT (%)	57.42	18.33	23.41	10.93	16.41	16.51
MSE	6.51	2.98	3.28	2.73	2.87	2.90

Simulation results: w/o replacement, high H effect

METRIC	PSM Method				
	Raw	*A*	*B*	*C*	*D*	*E*
% of Unmatched	0	9.19	18.19	2.74	16.90	16.88
Balance X	12.75	1.33	2.61	1.92	2.36	2.33
Balance H	53.03	53.91	0	20.37	2.10	1.81
Bias ATT (%)	65.88	2.31	23.71	0.15	6.02	6.92
MSE	7.50	2.30	3.40	2.54	2.55	2.57

Simulation results: with replacement, small H effect

METRIC	PSM Method				
	Raw	*A*	*B*	*C*	*D*	*E*
% of Unmatched	0	0.01	0.90	0.01	0.01	0.01
Balance X	13.01	0.95	1.64	1.63	0.93	0.94
Balance H	17.90	18.49	0	0.25	0.88	1.23
Bias ATT (%)	57.42	9.05	3.67	0.61	8.36	8.80
*MSE*1000*	6.52	3.52	3.52	3.33	3.45	2.50

Simulation results: with replacement, high H effect

METRIC	PSM Method					
	Raw	*A*	*B*	*C*	*D*	*E*
% of Unmatched	0	0.01	0.10	0.01	0.01	0.01
Balance X	12.75	1.15	1.93	1.90	1.08	1.09
Balance H	53.03	53.47	0	0.62	0.78	0.79
Bias ATT (%)	65.88	24.24	2.28	3.78	8.72	7.78
*MSE*1000*	7.56	4.22	3.65	3.82	3.70	3.70

Case study

METRIC	PSM Method				
	Raw	*A*	*B*	*C*	*D*	*E*
% of Unmatched	0	0	0.1	0	0	0
Balance X	14.8	0.91	1.97	1.98	1.39	1.38
LOW APGAR (‰)						
Caesarean section	10.9	10.9	11.0	10.9	10.9	10.9
Natural delivery	5.2	9.1	9.6	9.7	9.9	9.9
*Difference (ATT*1000)*	5.75	2.80	1.40	1.23	1.02	1.07
(Estimated) SE of ATT	1.46	1.80	0.47	0.47	1.84	1.96

Concluding remarks

- In general, methods using information on clusters have better performance:
  - matching without replacement: C,D,E have similar performance
  - matching with replacement: B,C better than D,E
  - B better performance than C when confounding is very strong
- How to choose among them? Data structure:
  - B and C perform well here: (majority of) large clusters of unequal size
  - D, E perform well here but also with small clusters (Arpino and Mealli, 2011)
- Method C seems attractive when some clusters are small as it greatly reduces the number of unmatched units.

Bruno Arpino
Pompeu Fabra University
bruno.arpino@upf.edu

Massimo Cannas
University of Cagliari
massimo.cannas@unica.it

A note on estimated SE

- For unclustered data:

- if treatment randomized:

  classic se of difference in means

- Otherwise **corrections** are needed for:

  a) uncertainty in ps estimation

  b) uncertainty due to matching (Abadie, Imbens, 2006)

- For clustered data no theoretical results are available

A note on number of dropped units

- Matching with replacement:

drops B > drops A = drops C

- No longer true without replacement:

Unit	H	T	ps	sd(ps)=0.18
[1,]	1	0	0.10	
[2,]	1	1	0.10	
[3,]	1	0	0.20	
[4,]	1	1	0.30	
[5,]	2	0	0.39	
[6,]	2	1	0.40	
[7,]	2	0	0.60	