Machine learning techniques for propensity score matching with clustered data. A simulation study.

B. Arpino, F.C. Billari and M. Cannas

Abstract Propensity score method is a classic tool for obtaining causal estimates from non-randomized data. In the applied literature this tool is increasingly applied in contexts where causal inference is complicated by data having a hierarchical structure, a typical case being that of patients clustered within different practitioners and hospitals. This is questionable since several studies suggest that these cluster-level variables can have a considerable effect both on the treatment intake and the outcome, i.e., they are potential confounders which can bias causal estimates in absence of adequate control. Via Monte-Carlo simulations, we assess the performance of several strategies for the estimation of the propensity score with clustered data. We compare classic fixed and random-effects models with machine learning algorithms, which outperformed standard strategies with unclustered data when the link between the treatment and the covariates is not linear and additive. We found that a novel algorithm, Generalized Mixed Effect Regression Trees, gives benefits analogous to those found in the non-hierarchical setting by other authors.

1 Background

In observational studies the absence of randomization is the fundamental difficulty in estimating treatment effects. The idea of propensity score methodology is to summarize a large number of confounding variables in a single variable, i.e., the propensity score, and then to use this summary for balancing covariates across treated and control units.

Until recently, most theoretical works on this topic dealt with unstructured data. However, clustered data are the norm in many fields [4]. Multilevel structures pose challenges for propensity score methodology. First, the treatment assignment may be multilevel in nature, that is, it may not depend not only on individual characteristics but also on characteristics of the cluster the individual belongs. Second, the individual out-

Bruno Arpino
Bocconi University, Via Roentgen 1, 20136 Milan - e-mail: bruno.arpino@unibocconi.it

Francesco Billari
Bocconi University, Via Roentgen 1, 20136 Milan - e-mail: francesco.billari@unibocconi.it

Massimo Cannas
Bocconi University, Via Roentgen 1, 20136 Milan - e-mail: massimo.cannas@phd.unibocconi.it

comes may be affected by cluster level characteristics. Moreover there can be complex interactions between individual and cluster level covariates affecting both treatment assignment and the outcome.

Propensity score methods need to be somehow adapted for applications in a multilevel context. In multisite studies of educational programs or interventions, the use of single- level models for estimating propensity scores followed by the use of the resulting propensity scores as a basis for matching students within each school has been considered an effective strategy [4, 3]. Following Arpino and Mealli, we present a simulation study which investigates the possibility of using alternative methods for estimating the propensity score with clustered data.

2 Methods

Machine learning methods considered here are based on a predictive algorithm known as *tree*. In contrast with traditional models, trees and their derivations are flexible enough to handle automatically non-linearity and non-additivity. Machine learning methods have already been proposed for the estimation of the propensity score but only in the single-level case. Simulation experiments implemented by Setoguchi et al. [8] and Lee et al.[9] showed that machine learning algorithms outperform standard logit models when the relation between treatment and outcome is not linear and additive.

An extension of machine learning algorithms to a multi-level setting can be carried out with the Generalized Mixed-Effect Regression Trees (GMERT), developed by Hajjem et al. [6]. To fix the idea, consider the classic mixed-effect logistic model:

$$logit[\frac{p_{ij}}{1-p_{ij}}] = x_{ij}\beta + z_{ij}b_i$$

$$b_i \sim N(0, D)$$

$$i = 1, \cdots, n \quad ; \quad j = 1, \cdots, n_i$$

where the x_{ij} are the fixed-effect covariates and the z_{ij} are the random-effect covariates. This model is usually estimated using maximum likelihood methods in the framework of the expectation-maximization (EM) algorithm. In the corresponding GMERT model the term x_{ij} is replaced by a more general function $f(x_{ij})$ that can be estimated using a standard regression tree.

The simulation experiment extends the set-up created by Setoguchi et al. [8] to a multilevel context where the treatment is administered at the individual level but cluster-level variables can enter both the assignment mechanism and the outcome variable. For each data set we generated ten basic covariates (four confounders associated with both the treatment and the outcome, three treatment predictors and three outcome predictors) and a cluster-level covariate acting as a confounder. The treatment assignment is generated according to various simulation scenarios obtained by varying the degree of linearity and additivity between the treatment and the individual-level covariates.In each scenario we estimate the propensity score using the following methods:

- Logistic regression with main effects only (LR)
- Logistic regression with dummies for cluster effects (LRD)
- Mixed-effect logistic regression (MELR)
- Mixed-effect regression tree (MERT)
- Boosted Regression Tree (BTREE)

The propensity score we used to produce a balanced data set via nearest neighbor matching. The R package *MatchIt* was used to perform all calculations [11].

3 Results

In Table 1 we show the results for a set of 1000 simulations, obtained from n=1000 observations clustered in nc=50 groups of equal size (W and C indicate an individual and cluster-level covariate, respectively).

Table 1 Simulation results (N=500) based on data sets of 1000 observations clustered in 50 groups of equal size.

		No cluster effect					Random Intercept				
Metric[†]	Method[‡]	S1[♭]	S2	S3	S4	S5	S1	S2	S3	S4	S5
ASAM W	LR	6.3	10.2	10.4	10.3	10.5	6.1	9.7	9.9	10.2	10.2
	LRD	6.1	8.9	10.0	10.2	10.4	7.2	8.8	11.3	11.2	11.2
	MELR	6.2	8.8	9.9	10.2	10.4	6.8	9.1	10.9	10.9	11.0
	MERT	8.4	8.1	8.2	8.4	8.6	8.5	8.7	8.9	8.8	8.8
	BTREES	7.2	11.0	12.1	12.1	12.3	9.2	11.4	12.9	12.7	12.8
ASAM C	LR	7.8	7.0	6.2	5.3	5.4	47.9	48.1	47.2	48.9	40.3
	LRD	5.9	5.4	5.4	4.0	4.4	6.0	5.0	4.1	4.2	4.3
	MELR	8.1	6.0	6.7	5.3	5.4	9.6	9.7	9.7	9.7	10.0
	MERT	7.9	6.7	6.9	6.7	6.8	8.3	8.4	8.5	8.7	8.5
	BTREES	7.8	6.4	6.3	7.5	7.7	9.8	11.4	12.3	12.3	12.5
RBIAS	LR	1.2	13.8	15.9	17.9	29.4	70.1	34.9	39.1	34.8	35.9
	LRD	1.2	18.7	19.6	28.6	32.4	2.2	18.9	14.3	27.2	25.0
	MELR	2.4	18.9	16.9	16.4	26.3	5.6	19.5	8.6	15.4	18.6
	MERT	3.2	11.4	10.8	6.2	12.3	5.4	11.5	6.5	6.8	17.6
	BTREES	10.6	14.8	17.3	16.5	15.5	16.6	23.0	25.8	25.5	57.8
SE	LR	0.019	0.014	0.014	0.021	0.023	0.101	0.030	0.037	0.024	0.013
	LRD	0.019	0.016	0.015	0.018	0.023	0.021	0.023	0.019	0.018	0.018
	MELR	0.015	0.014	0.015	0.016	0.019	0.017	0.019	0.010	0.015	0.013
	MERT	0.013	0.015	0.014	0.015	0.013	0.015	0.016	0.015	0.020	0.014
	BTREES	0.035	0.034	0.034	0.032	0.035	0.04	0.045	0.040	0.039	0.076

[†] ASAM: 100*average standardized absolute mean difference across treated and controls, RBIAS: mean absolute (per cent), SE: mean standard error.

[‡] LR: single level logistic regression, LRD: logistic regression with dummies for clusters, MELR: random-intercept logistic regression, MERT: random-intercept regression tree, BTREE: boosted tree with dummies for clusters.

[♭] S1: additive and linear, S2: moderately non-linear (+10 two-ways interactions), S3: mild non-additivity and non linearity (+3 two-ways interactions and 1 quadratic term), S4: moderate non-additivity (+3 quadratic terms), S5: moderate non-additivity and non-linearity (+10 two-ways interactions and 3 quadratic terms).

Usually it is customary to consider a good result an ASAM lower than 20 [9]. From Table 1 we can see that, in mean, no methods yields an higher ASAM, with the obvious exception of the logistic regression ignoring clustering. Also it must be noted that LRD is the most effective in reducing the imbalance in the cluster variable, even when the random intercept is added. The performance of MERT in reducing imbalance is slightly better than that of MELR in the main effect scenario S1 but the gap clearly becomes more important as we move toward scenarios deviating from the standard model.

4 Concluding remarks

We analyzed the performance of various propensity score estimation methods in a simulation context which extends that of Lee et al. [9] to clustered data. We assume that a cluster-level covariate impacts the treatment assignment and looked at the performance of propensity score estimation methods as the link between the treatment and the covariates departures from standard modeling assumpitions. The simulations show that the performance of all methods degrades when non linearity and non additivity increase, both in terms of higher imbalance and bias of the causal estimates. However, Mixed Effect Regression Trees [6] and classic Mixed Effect Logistic Regression performed better than other methods, with MERT also showing a more stable performance. The study extends to a multilevel setting previous findings of Setoguchi et al.[8] and Stuart et al.[9], which advocated the use of machine learning methods in propensity score estimation with univariate data.

References

1. Rosenbaum P R and Rubin, D B The central role of propensity score in observational studies for causal effects. Biometrika; 70, 41-55 (1983)
2. Kim S.J. and Seltzer, M. Causal Inference in Multilevel Settings in which Selection Process Vary across Schools. Working Paper 708, Center for the Study of Evaluation (CSE): Los Angeles, (2007)
3. Rosenbaum, P. R.Dropping Out of High School in the United States: An Observational Study. Journal of Educational Statistics, 11(3), 207-224. 1986.
4. Hong, G. and Raudenbush, S.W. (2008) Causal inference for time-varying instructional treatments. The Journal of Educational and Behavioral Statistics. Vol. 33, No. 3, pp 333-362.
5. Arpino B and Mealli F: The specification of the propensity score in multilevel studies. Computational Statistics and Data Analysis, 55, pp. 1770-1780 (2011)
6. Ahlem Hajjem, Franois Bellavance and Denis Larocque Mixed effects regression trees for clustered data. Statistics & Probability Letters, vol. 81, issue 4, pages 451-459 (2011)
7. Breiman L, Friedman J, Stone C. Classification and Regression Trees. Wadsworth Belmont; CA (1984)
8. Soko Setoguchi, Sebastian Schneeweiss, M. Alan Brookhart, Robert J. Glynn and E. Francis Cook: Evaluating uses of Data Mining techniques in Propensity Score Estimation: A Simulation Study. Pharmacoepidemiology and Drug Safety; 17: 546555 (2008)
9. Brian K.Lee, Justin Lessler and Elizabeth A. Stuart: Improving propensity score weighting using machine learning. Statistics in Medicine; 29 337-346 (2010)
10. McCaffrey DF, Ridgeway G, Morral AR. Propensity score estimation with boosted regression for evaluating causal effects in observational studies. Psychological Methods; 9(4):403-425 (2004)
11. Jasjeet S. Sekhon. Multivariate and Propensity Score Matching. Software with Automated Balance Optimization: The Matching package for R. Journal of Statistical Software. Forthcoming.