Project Report: Observational studies with multi-level treatments

STAT 6390.001: Introduction to Causal Inference

1 Introduction

Causal inference is the process of determining the independent, effect of a particular treatment that is applied to a selected population. In most of the studies estimation of the average treatment effect is an important goal. In an investigation of the causal effect of treatment, one of the most common issues is that the groups of subjects in the study are not randomly assigned. Then there is noninterference or manipulation of the research subjects and no control over treatment groups. These studies are called observational studies. Confounders are the variables associated with both the treatment and the outcome. When Confounders are present, non-random group assignments could result in biased estimates of the treatment effect. There are many of popular literature to overcome this problem. Among the literature, Propensity score-based methods have been widely improved to adjust for confounders in studies to estimate causal treatment effects for binary treatments. Even though the initial work on propensity scores focused on the case of binary treatments, there are more recent works that generalize the Propensity score-based methods to treatments with more than two levels (Multi-level treatment).

There are many generalized Propensity score-based methods to estimate the treatment effect. Among them, the natural extension is the multinomial logistic regression model for multi-level treatments (multinomial logistic regression, MLR). On the other hand, there are also many machine-learning approaches developed to estimate propensity scores. Random forests (RF), generalized boosted models (GBM), and data-adaptive matching scores (DAMS) are a few popular approaches that can be found in the literature.

2 Existing literature

2.1 Logistic regression model,

In an observational study of the effect of treatment on outcomes, the propensity score is the probability of receiving the treatment of interest conditional on measured baseline covariates $p = Pr(A = 1|X)$, where X denotes the vector of measured covariates and A denotes treatment status. The propensity score is often estimated using a logistic regression model, with the propensity scores being the predicted probabilities generated by that model.

$$logit[Pr(A = 1)] = \beta_0 + \beta^T X.$$

2.2 Random forest

The Random Forests method involves several steps. First, multiple random samples are drawn from the data. Second, a Classification Tree model of the covariates X on treatment assignment Z is estimated using the data from each random sample. Third, the participants propensity scores are estimated from each Classification Tree model. Finally, these propensity scores are averaged across all models to obtain the Random Forests propensity score for each participant.

2.3 Generalized boosted models

GBM estimates the propensity score for the binary treatment indicator using a flexible estimation method that can adjust for a large number of pretreatment covariates. GBM estimation involves an iterative process with multiple regression trees to capture complex and nonlinear relationships between treatment assignment and the pretreatment covariates without over-fitting the data. It works with continuous and discrete pretreatment variables and is invariant to monotonic transformations of them.

3 Theoretical results

Let A_i be the observed treatment status for the ith subject, so $A_i = t$ if subject i was observed under treatment $t \in 1, ..., M$, where there are a total of M treatments. Also, let the $A_i(t) = I(A_i = t)$ be an indicator function. Y_i denote the observed outcome for subject i and the set of potential outcomes be $\{Y_i(1), \ldots, Y_i(M)\}$. Let Xi denote the $p \times 1$ vector of p observed pretreatment covariates for subject i.

3.1 Generalized propensity score

The generalized propensity score is the conditional probability of subject i receiving treatment t given covariates x.

$$\begin{aligned} p_i(t|x) &= P(A_i = t|X_i = x) \\ &= P(A_i(t) = 1|X_i = x). \end{aligned}$$

3.2 Average treatment effect (ATE)

The ATE of treatment t relative to treatment s is the difference of mean outcomes had the entire population been observed under t versus had the entire population been observed under s.

$$\begin{aligned} ATE_{ts} &= E[Y(t)] - E[Y(s)] \\ &= \mu_t - \mu_s. \end{aligned}$$

3.3 Average treatment effect among the treated (ATT)

Average treatment effect among the treated or ATT of t relative to s is the difference in mean outcome among subjects who were treated with t versus the mean outcome they would

have had if they received s.

$$\begin{aligned} ATT_{t,ts} &= E[Y(t)|A=t] - E[Y(s)|A=t] \\ &= \mu_{t,t} - \mu_{t,s}. \end{aligned}$$

3.4 Assumptions

A1: The stable unit treatment value assumption

This assumption states that the distribution of potential outcomes for one subject/unit is independent of the treatment assignment of any other subject.

$$(Y_i(1), \ldots, Y_i(M)) \perp A_j \text{ for } i \neq j.$$

A2: The weak unconfoundedness assumption

This assumption states that the treatment assignment indicator does not depend on the potential outcome given the observed covariates.

$$A(t) \perp Y(t)|X.$$

This assumption means that we can model the conditional distribution of the treatment assignment given the covariates without having to condition on the outcome. Furthermore, if this assumption holds, then the treatment assignment is weakly unconfounded given the generalized propensity score.

$$A(t) \perp Y(t)|p(t|X).$$

A3: Sufficient overlap or positivity assumption

There is a nonzero probability of being assigned to each treatment.

$$0 < p(t|X) < 1 \quad \text{for all} \quad t, X.$$

3.5 Methods under comparison

3.5.1 MLR

Assume an underlying multinomial distribution instead binomial distribution generalized propensity score for each treatment level follows,

$$p(t|x)_{MLR} = \frac{1}{1 + \sum_{s=2}^{M} e^{\beta_s' x}} \text{ for } t = 1,$$
$$p(t|x)_{MLR} = \frac{e^{\beta_t' x}}{1 + \sum_{s=2}^{M} e^{\beta_s' x}} \text{ for } t = 2, \ldots, M,$$

The β's are estimated by maximizing the likelihood.

3.5.2 GBM

GBM is built on an ensemble of regression trees and each regression tree iteratively fits the residuals from the previous tree to approximate the propensity score function. The number of trees to be generated is determined by achieving the maximum balance in covariate distribution among different treatment groups. The algorithm automatically includes interaction and nonlinear terms as the regression tree allows for multi-level splits.

3.5.3 RF

A key feature of RF lies in that each tree in the tree ensemble is built on a random subset of the original covariates on a bootstrap sample of the original dataset to avoid overfitting. Given a vector of covariates, each tree votes for one class label. Then the generalized propensity score for treatment t can be estimated as the fraction of trees that classify/predict the given subject into treatment group t out of the entire collection of trees.

3.5.4 DAMS

This is a weighted average of the propensity score estimates from a parametric model and a nonparametic model.

$$\hat{p}(t|x)_{DAMS} = \lambda \hat{p}(t|x)_{MLR} + (1-\lambda)\hat{p}(t|x)_{RF},$$

$$\text{where } \lambda = \frac{\hat{p}(t|x)_{MLR}{}^{A(t)}[1-\hat{p}(t|x)_{MLR}]^{1-A(t)}}{\hat{p}(t|x)_{MLR}{}^{A(t)}[1-\hat{p}(t|x)_{MLR}]^{1-A(t)} + \hat{p}(t|x)_{RF}{}^{A(t)}[1-\hat{p}(t|x)_{RF}]^{1-A(t)}}.$$

3.6 Treatment effect

Use inverse probability of treatment weighting (IPTW) to estimate an ATE or an ATT.

The weight for observation i is,

$$w_i(t) = \frac{1}{\hat{p}_i(t|x)}. \quad w_i(t,s) = \frac{\hat{p}_i(t|x)}{\hat{p}_i(s|x)}.$$

Then the estimate for the population average outcome for treatment t is the weighted average of the outcomes in the treatment group t,

$$\hat{\mu}_t = \frac{\sum_{i=1}^{n} A_i(t) w_i(t) Y_i}{\sum_{i=1}^{n} A_i(t) w_i(t)}.$$

To estimate the average outcome for treatment t for those actually treated with t as well as the average outcome for treatment s for those actually treated with t.

$$\hat{\mu}_{t,t} = \frac{\sum_{i=1}^{n} A_i(t) Y_i}{\sum_{i=1}^{n} A_i(t)}, \quad \hat{\mu}_{t,s} = \frac{\sum_{i=1}^{n} A_i(s) w_i(t,s) Y_i}{\sum_{i=1}^{n} A_i(s) w_i(t,s)}.$$

4 Simulation study

4.1 Simulation setting

- Variables X_1, X_2, X_3 and X_4 are associated with both treatment and outcome. Here, X_1 and X_3 are binary, while X_2 and X_4 are continuous.

- X_5, X_6, and X_7 are associated with the treatment. Here X_5 and X_6 are binary, and X_7 is continuous.
- X_8, X_9, and X_{10} are associated with the outcome only. X_8 and X_9 are binary, while X_{10} is continuous.
- The treatment variable A has three treatment levels.
- The covariates X_1, X_2, X_3, X_4, X_7,and X_{10} are generated through independent $N(0,1)$.

$$P(A=0|X) = \frac{1}{1+e^{f_1(X)}+e^{f_2(X)}},$$
$$P(A=1|X) = \frac{e^{f_1(X)}}{1+e^{f_1(X)}+e^{f_2(X)}},$$
$$P(A=2|X) = \frac{e^{f_2(X)}}{1+e^{f_1(X)}+e^{f_2(X)}},$$

- The rest of the covariates generated from normal distributions and have a correlation as follows,

$$corr(X_1, X_5) = 0.2,\ corr(X_2, X_6) = 0.9,\ corr(X_3, X_8) = 0.2,\ corr(X_4, X_9) = 0.9.$$

Then the true propensity scores,

$$\begin{aligned} E[Y|X] = &-3.85 + 0.3X_1 - 0.36X_2 - 0.73X_3 - 0.2X_4 + 0.71X_8 - 0.19X_9 + 0.26X_{10} \\ &- 0.4I(A=1) - 0.7I(A=2) \end{aligned}$$

The exact forms of $f_1(X)$ and $f2(X)$ in each scenario are listed below and 1000 datasets were generated under each scenario with each dataset containing $n = 1000$ observations.

Consider seven scenarios (Scenario A to Scenario G). The simplest (A) and most complicated (G) scenarios are given below.

Scenario A

$$f_1(X) = 0.8X_1 - 0.25X_2 + 0.6X_3 - 0.4X_4 - 0.8X_5 - 0.5X_6 + 0.7X_7$$
$$f_2(X) = -0.4X_1 - 0.1X_2 + 0.45X_3 + 0.7X_4 + 0.2X_5 - 0.9X_6 - 0.35X_7$$

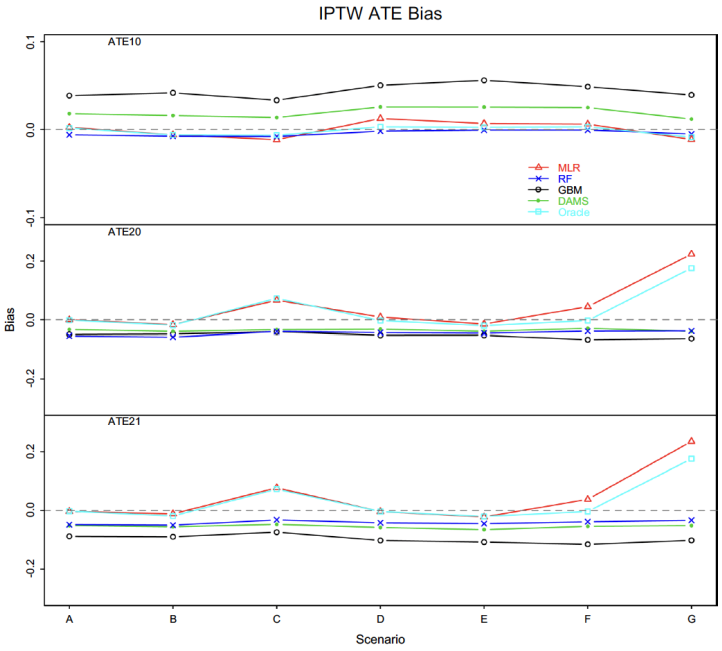
Scenario G

$$
\begin{aligned}
f_1(X) = {} & 0.8X_1 - 0.25X_2 + 0.6X_3 - 0.4X_4 - 0.8X_5 - 0.5X_6 + 0.7X_7 - 0.25X_2^2 \\
& - 0.4X_4^2 + 0.7X_7^2 + 0.4X_1 \times X_3 + 0.4X_1 \times X_6 - 0.175X_2 \times X_4 \\
& + 0.175X_2 \times X_3 + 0.3X_3 \times X_4 + 0.3X_3 \times X_5 - 0.2X_4 \times X_5 \\
& - 0.28X_4 \times X_6 - 0.4X_5 \times X_6 - 0.4X_5 \times X_7 \\
f_2(X) = {} & - 0.4X_1 - 0.1X_2 + 0.45X_3 + 0.7X_4 + 0.2X_5 - 0.9X_6 - 0.35X_7 \\
& - 0.1X_2^2 + 0.7X_4^2 - 0.35X_7^2 - 0.2X_1 \times X_3 - 0.2X_1 \times X_6 - 0.07X_2 \times X_4 \\
& - 0.07X_2 \times X_3 + 0.225X_3 \times X_4 + 0.225X_3 \times X_5 + 0.49X_4 \times X_5 \\
& + 0.49X_4 \times X_6 + 0.1X_5 \times X_6 + 0.1X_5 \times X_7
\end{aligned}
$$

4.2 Simulation comparison (ATE)

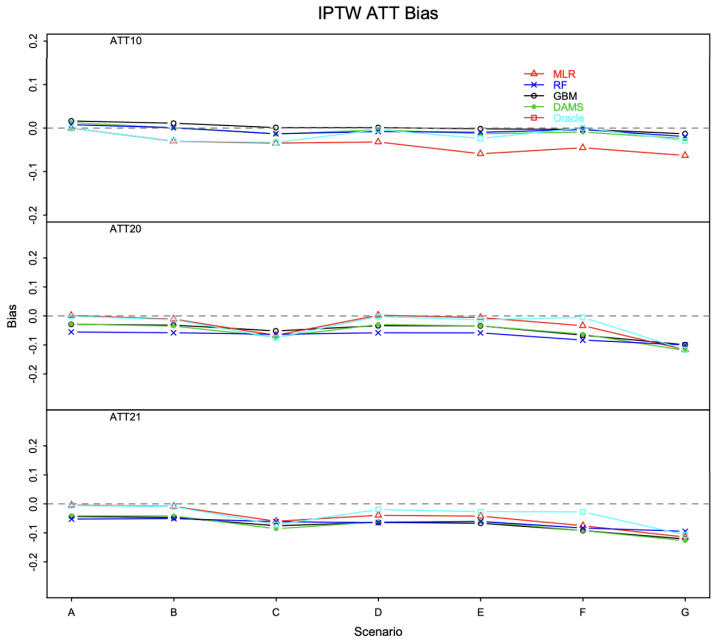
First, consider the bias, which is the difference between the average of the ATE estimates from the 1000 datasets and the respective true ATE. Note that,

$$Bias = \overline{\widehat{ATE}} - ATE \quad \text{where } \overline{\widehat{ATE}} = \frac{1}{1000}\sum_{j=1}^{1000} \widehat{ATE}_j.$$



- Estimating generalized propensity scores by GBM seems to result in the most biased estimated ATE_{10} in all seven scenarios, and this method seems to consistently overestimate the ATE.
- MLR and RF seem to result in lower bias in estimated ATE_{10}.
- The bias of estimated ATE_{20} and ATE_{21} seem to be similar for each propensity score estimation method across all scenarios.
- MLR seems to result in a higher bias whereas the bias of the other three methods remain consistent across the different scenarios.
- When there is considerable non-linearity and non-additivity (like scenario G), the simple MLR model did not do an adequate job in estimating the propensity scores while machine learning methods perform well.

4.3 Simulation comparison (ATT)



- The nonparametric algorithms, RF, DAMS and GBM lead to less biased estimates compared to MLR, when the MLR model is misspecified.

5 A real example

Apply the proposed methodology to the Taobao dataset collected from Taobao.com, China's largest e-commerce platform. Taobao offers a feedback system to reduce the likelihood of fraud and encourage trust based on reputation. The website adopts a similar reputation rating system in which the cumulative rating score is then categorized into twenty grades from 0 to 20. Based on this, the authors define the reputation variable as follows,

0-5 Low, 6-10 Medium, and 11-20 High.

The dataset is a simple random sample from the original database with a sample size of ten thousand.

Score	Grade	Reputation	Number of sellers	Mean sales
<=250	0–5	Low (0)	5587	2.64×10^4
251–10,000	6–10	Medium (1)	3881	1.47×10^5
>= 10,001	11–20	High (2)	532	4.75×10^5

Here for each seller from June 2011 to December 2011 data were collected. In this analysis, the authors are interested in examining the causal effect of a seller's rating (reputation) on sales (gross revenue). There are 13 potential confounders related to the seller's characteristics like the seller's age, the seller's gender, etc.

5.1 Summary of the results

GPS method	MLR	RF	GBM	DAMS
$\widehat{ATE}_{10,IPTW}$ ("medium" vs. "low")	9.56×10^4	3.76×10^4	8.58×10^4	8.68×10^4
$\widehat{ATE}_{20,IPTW}$ ("high" vs. "low")	3.95×10^5	2.90×10^5	3.41×10^5	3.93×10^5
$\widehat{ATE}_{21,IPTW}$ ("high" vs. "medium")	3.00×10^5	2.52×10^5	2.55×10^5	3.06×10^5

- RF seems to work quite differently and the rest of the three methods yield similar results.

- We can conclude that sellers with "medium" reputation has an increase of around 9×10^4 Yuan in half-year sales, compared to "low" reputation.

6 Conclusion and discussion

From the simulation studies, we can say that using MLR to estimate the generalized propensity scores can result in extreme weights that in turn result in more bias. On the other hand, GBM, DAMS, and RF tend to be more stable across different levels of complexity in the relationship between the treatment assignment and the covariates. DAMS performed the best out of the four propensity score estimation methods in combination with IPTW. In conclusion, we can recommend machine learning methods for propensity score estimation in multi-level treatment settings. In real applications, the true propensity score model is never known for observational studies. Researchers may apply different algorithms and see how different the results.

References

[1] Lin Lin, et al. (2019). "Causal inference for multi-level treatments with machine-learned propensity scores". Health Services and Outcomes Research Methodology 19:106–126.

[2] Peter C Austin et al. (2017). "Estimating the effect of treatment on binary outcomes using full matching on the propensity score". Stat Methods Med Res doi: 10.1177/0962280215601134.

[3] Peng Zhao et al. (2017). "Propensity Score and Proximity Matching Using Random Forest". Contemp Clin Trials doi: 10.1016/j.cct.2015.12.012.

[4] A. Abadie and G. W. Imbens (2016). "Matching on the estimated propensity score". Econometrica, 84:781-807.

[5] Shu Yang, et al. (2016). "Propensity Score Matching and Subclassification in Observational Studies with Multi-level Treatments". Biometrics, 72(4), 1055-1065.

[6] B. B. Hansen (2008). "The prognostic analogue of the propensity score". Biometrika, 95:481-488.