ISSUES IN ECONOMICS RESEARCH
ECON5108 - WRITTEN ASSIGNMENT

Duong Trinh - 2494479T

Adam Smith Business School

University of Glasgow

Word Count:

January 2021

Contents

1 Literature Review

Firstly, causal ML methods are powerful tools in using data to recover complex interactions among variables and flexibly estimate the relationship between the outcome, the treatment indicator and covariates.

Secondly, causal ML methods allow for the inclusion of a large number of covariates, even when the sample size is relatively small, by assuming that the model is sparse, (i.e., only a small number of covariates are relevant), and using regularized regressions.

Thirdly, the use of causal ML methods allows to implement systematic model selection.

Finally, causal machine learning methods prove to be very useful when one is interested in estimating heterogeneous treatment effects.

2 Framework and Definition

I employ the Neyman-Rubin potential outcome framework, and assume a superpopulation or distribution $\mathcal{P}$ from which a realization of N independent random variables is given as the training data. That is, $(Y_i^0, Y_i^1, X_i, W_i) \sim \mathcal{P}$, where:

- $X_i \in \mathbb{R}^d$ is a $d-$ dimensional covariate or feature vector
- $W_i \in \{0, 1\}$ is the binary treatment assignment indicator (to be defined precisely later)
- $Y_i^0 \in \mathbb{R}$ is the potential outcome of unit i when i is assigned to the control group, and Y_i^1 is the potential outcome when i is assigned to the treatment group

The fundamental problem of causal inference is that each individual can either receive the treatment or not, thus only one of the two potential outcomes (Y_i^w) is observable:

$$Y_i^{obs} = W_i Y_i^1 + (1 - W_i) Y_i^0$$

To avoid clutter, denote Y_i^{obs} simply by Y_i from now on.

Hence, the *individual treatment effect* (ITE) $\xi_i = Y_i^1 - Y_i^0$ of W_i on Y_i is never observed, and directly training machine learning methods on this difference is not possible as a result. However, the identification of expectations of ξ_i may be possible under plausible assumptions. For example, the identification of the *average treatment effect* (ATE) could be defined as $\tau = E[\xi_i]$. My main focus is the *conditional average treatment effects* $(CATEs)$. $CATEs$ take the expectations of ξ_i conditional on exogenous pre-treatment covariates. Define the finest conditioning level that uses all available covariates X_i as the *individualized average treatment effect* $(IATE)$:

$$\tau(x) = E[\xi_i \mid X_i = x] = \mu^1(x) - \mu^0(x)$$

where $\mu^w(x) = \mathbb{E}[Y_i^w | X_i = x] = \mathbb{E}[Y_i | X_i = x, W_i = w]$ denotes the conditional expectation of the unobserved potential outcomes.

Assumption 1: Unconfoundedness

The unconfoundedness assumption states that, once we condition on observable characteristics, the treatment assignment is independent to how each person would respond to the treatment. In other words, the rule that determines whether or not a person is treated is

determined completely by their observable characteristics. This allows, for example, for experiments where people from different genders get treated with different probabilities, but it rules out experiments where people self-select into treatment due to some characteristic that is not observed in our data.

$$Y_i^1, Y_i^0 \perp W_i \mid X_i$$

Assumption 2: Overlap

In order to estimate the treatment effect for a person with particular characteristics $X_i = x$, we need to ensure that we are able to observe treated and untreated people with those same characteristics so that we can compare their outcomes. The overlap assumption states that at every point of the covariate space we can always find treated and control individuals.

$$\forall x \in \text{supp}(X), \quad 0 < P(W = 1 \mid X = x) < 1$$

2.1 Specific Machine Learning

$$\hat{\tau}(x) = \sum_{i=1}^{N} W_i w_i^1(x) Y_i - \sum_{i=1}^{N} (1 - W_i)\, w_i^0(x) Y_i \tag{1}$$

One estimator-specific approach is a modification of a specific machine learning estimator to move the target from the estimation of outcomes to the estimation of IATEs. In particular, this strand of the literature modifies machine learning algorithms based on regression trees [Breiman et al., 1984] to estimate IATEs, and includes these previous works:

Athey and Imbens [2016]

- A data-driven approach to partition the data into subpopulations that differ in the magnitude of their treatment effects. The approach enables the construction of valid confidence intervals for treatment effects
- made use of the Horvitz-Thompson transformation of the outcome to inform the process of building causal trees, with the main goal of predicting CATE.
- (P): provided a valid inference result on average treatment effects for groups defined by the tree leaves, conditional on the data split in two subsamples: one used to build the tree leaves and the one to estimate the predicted values given the leaves.
- (P): essentially assumption-free

- (C): limited to trees and does not account for splitting uncertainty, which is important in practical settings.

Wager and Athey [2018]

- proposed a subsampling-based construction of a causal random forest
- (P) providing valid pointwise inference for CATE for the case when covariates are very low-dimensional (and essentially uniformly distributed)
- (C) this condition rules out the typical high-dimensional settings that arise in many empirical problems, especially in current RCT where the number of baseline covariates is potentially very large.
- (F) the dimension d is fixed, the analysis relies on the Stone's model with smoothness index $= 1$, in which no consistent estimator exists once $d > log(n)$. It'd be interesting to establish consistency properties and find valid inferential procedures for the random forest in high-dimensional ($d \propto n$ or $d \gg n$) approximately sparse cases, with continuous and categorical covariates.

Athey et al. [2019]

- (I) Generalized Random Forest - the most recent estimator in this line of research.
- (P) The Causal Forest is specialized to maximize heterogeneity in experimental settings but it is not built to explicitly account for selection. Thus, it is prone to choose splits that do not sufficiently remove selection bias. However, Causal Forests with local centring address this problem by partialling out the selection effects in a first step.
- (C)

2.2 Generic Machine Learning

Generic approach splits the causal estimation problem into several standard prediction problems and may be combined with a large variety of supervised machine learning estimators.

Knaus et al. [2021]

- structure approaches that estimate IATEs on a common background - solving a weighted minimization problem with modified outcomes:

$$\min_{\tau} \left\{ \frac{1}{N} \sum_{i=1}^{N} w_i \left[Y_i^* - \tau \left(X_i \right) \right]^2 \right\} \tag{2}$$

- considers only estimators that require at most one additional estimation step on top of the estimation of the nuisance parameters, owing to restrictions in computation power.
- (C) focus on the finite-sample performance of point estimates rather than the investigation of inference procedures

Chernozhukov et al. [2018]

- does not focus directly on the HTEs, but on features of HTEs such as: the best linear predictor of the heterogeneous effects (BLP), the group average treatment effects (GATES) sorted by the groups induced by machine learning proxies, and the average characteristics of the units in the most and least affected groups, or classification analysis (CLAN).
- based on random splitting of the data into an auxiliary and a main sample - two samples are approximately equal in size. Based on the auxiliary sample, a ML estimator, called proxy predictor, is constructed for the conditional average treatment effect (CATE).
- (P) Credibility: Estimation and inference relies on data splitting. The inference quantifies uncertainty coming from both parameter estimation and the data splitting, allows us to avoid overfitting and all kinds of non-regularities.

Kuenzel [2019]

- a new meta-algorithm for CATE estimation, translate any supervised learning or regression algorithm or a combination of such algorithms into a CATE estimator.
- (P) X-learner performs particularly well when one of the treatment groups is much larger than the other or when the separate parts of the X-learner are able to exploit the structural properties of the response and treatment effect functions.

- (P) X-learner can adapt to these different settings: exploits the advantages of both S-learner (pooling the data across treatment and control conditions is beneficial if the treatment effect is simple, or even zero) and T-learner (separating data is beneficial if the treatment effect is strongly heterogeneous).

3 Causal Machine Learning of Effect Heterogeneity

3.1 Post Lasso

3.2 Causal Tree

Reference: Athey and Imbens [2016]

3.3 Causal Forest

Reference: Wager and Athey [2018]; Athey et al. [2019]

3.4 X Learner

Reference: Kuenzel [2019]
In particular, I implement the X Learners with Random Forest (RF) as base learners. In the original paper, the X Learners with BART is considered as well.

3.5 Generic ML

Reference: Knaus et al. [2021]
In particular, I implement the MOM IPW with Random Forest - based version.

Note: I specified step-by-step estimation procedure in R code.

4 Simulation Design

4.1 General Simulation for estimating IATEs

In this section, I introduce the general framework of the following simulations. For each simulation, I specify the sample size n, the dimension d of feature space (both low and high dimension setup are considered) as well as the following functions:

- The treatment propensity: $\pi(x) = \mathbb{P}[W = 1 \mid X = x]$
  I consider the constant propensity (balanced case when $\pi(x) = 0.5$ and unbalanced case when $\pi(x)$ is very small) or linear in the sense of logistic function: $\pi(X) = \frac{1}{1+e^{X\beta_w+\epsilon}}$ and $\epsilon \sim N(0,1)$
- The mean effect: $m(x) = 2^{-1}\mathbb{E}\left[Y^{(0)} + Y^{(1)} \mid X = x\right]$
  I consider both linear mean effect $m(x) = X\beta_m$ (when β_m is either dense or sparse)) or non-linear mean effect.
- The treatment effect: $\tau(x) = \mathbb{E}\left[Y^{(1)} - Y^{(0)} \mid X = x\right]$ Since this is main interest, I consider some specific cases of τ.

Then, to simulate an observation, i, in the training set, I simulate its feature vector, X_i, its treatment assignment, W_i, and its observed outcome, Y_i as below:

1. First, I simulate a d-dimensional feature vector X:
   - Independent: $X \sim N(0, I_{d\times d})$ or $X \sim U\left[0,1\right]^d$
   - Dependent: $X \sim N(0, \Sigma)$
2. Next, I we simulate the treatment assignment W according to:
   $$W \sim Bernouli(\pi(X))$$
3. Finally, I create the observed outcome Y:
   $$Y \sim N\left[m(X) + (W - 0.5)\tau(x), \sigma_Y^2\right]$$
   where the conditional variance of Y given X and W is $\sigma_Y^2 = 1$ (noise level).

I train each $IATE$ estimator on a training set of n units, and then evaluate its performance against a test set of n_test units for which the true effect is known. I replicate each experiment $R = 25$ times.

Simulation 1: Low dimensional data; No confounding; Balanced propensity; Linear mean effect; No treatment effect

$$
\begin{aligned}
n = 5000; \quad n_{test} = 1000; \quad d = 5 \\
X \sim N(0, I_{d\times d}) \\
\pi(X) = 0.5 \\
m(X) = \Sigma_{d=1}^{10} \beta_d x_d \quad (\beta \sim U = [1, 30]^d) \\
\tau(X) = 0
\end{aligned}
$$

Simulation 2: Low dimensional data; No confounding; Balanced propensity; Linear, dense/sparse mean effect; Linear treatment effect

$$
\begin{aligned}
n = 5000; \quad n_{test} = 1000; \quad d \in \{10, 20\} \\
X \sim N(0, I_{d\times d}) \\
\pi(X) = 0.5 \\
m(X) = \frac{1}{2}\Sigma_{d=1}^{4} x_d + \Sigma_{d=5}^{10} x_d \\
\tau(X) = \Sigma_{d=1}^{4} x_d
\end{aligned}
$$

Simulation 3: Low dimensional data; No confounding; Balanced propensity; No mean effect; Nonlinear treatment effect

In Wager and Athey [2018]:

$$
\begin{aligned}
n = 5000; \quad n_{test} = 1000; \quad d \in \{2, 4, 6, 8\} \\
X \sim U[0, 1]^d \\
\pi(X) = 0.5 \\
m(X) = 0 \quad (\beta = 0) \\
\tau(X) = \zeta(X_1)\zeta(X_2), \quad \zeta(X) = 1 + \frac{1}{1 + e^{-20(x-1/3)}}
\end{aligned}
$$

Simulation 4: Low dimensional data; No confounding; Unbalanced propensity propensity; Linear, dense/sparse mean effect; Linear treatment effect

$$\begin{aligned} n = 5000; \quad n_{test} = 1000; \quad d \in \{2, 8\} \\ X \sim N(0, I_{d\times d}) \\ \pi(X) = 0.01 \\ m(X) = X^T\beta \quad (\beta \sim U = [-5, 5]^d) \\ \tau(X) = 6.\mathbb{I}(X_1 > 0) + 8.\mathbb{I}(X_2 > 0) \end{aligned}$$

Simulation 5: Low dimensional data; Confounding; Balanced propensity propensity; Linear, sparse mean effect; Nonlinear treatment effect

$$\begin{aligned} n = 5000; \quad n_{test} = 1000; \quad d \in \{2, 8\} \\ X \sim U[0, 1]^d \\ \pi(X) = \frac{1}{4}(1 + \beta_{2,4}(X_1)) \\ m(X) = 2X_1 - 1 \\ \tau(X) = \zeta(X_1)\zeta(X_2), \quad \zeta(X) = 1 + \frac{1}{1 + e^{-20(x-1/3)}} \end{aligned}$$

Simulation 6: High dimensional data; No Confounding; Balanced propensity propensity; No mean effect; Nonlinear treatment effect

$$\begin{aligned} n = 1000; \quad n_{test} = 200; \quad d = 200 \\ X \sim U[0, 1]^d \\ \pi(X) = 0.5 \\ m(X) = 0 \\ \tau(X) = \zeta(X_1)\zeta(X_2), \quad \zeta(X) = 1 + \frac{1}{1 + e^{-20(x-1/3)}} \end{aligned}$$

4.2 Comparison Metrics

I consider three major performance measures: Mean Squared Error (MSE), Absolute Bias ($|Bias|$) and Coverage Rate for 95% confidence interval ($Coverage$) for the prediction of each observation j in the testing sample:

$$MSE_j = \frac{1}{R}\sum_{r=1}^{R}\left[\xi\left(x_j, y_j^0\right) - \hat{\tau}\left(x_j\right)_r\right]^2$$

$$|Bias_j| = |\underbrace{\frac{1}{R}\sum_{r=1}^{R}\hat{\tau}\left(x_j\right)_r}_{\overline{\hat{\tau}}(x_j)_r} - \xi\left(x_j, y_j^0\right)|$$

$$Coverage = \frac{1}{R}\sum_{r=1}^{R}\mathbb{I}\{\overline{\hat{\tau}}\left(x_j\right)_r - 1.96 * SD_j \leq \hat{\tau}\left(x_j\right)_r \leq \overline{\hat{\tau}}\left(x_j\right)_r + 1.96 * SD_j\}$$

where $SD_j = \sqrt{\frac{1}{R}\sum_{r=1}^{R}\left[\hat{\tau}\left(x_j\right)_r - \overline{\hat{\tau}}\left(x_j\right)_r\right]^2}$

Since there are 1000 parameters corresponding to 1000 observations in the testing sample, I summarize the performance over the whole testing sample by taking the averages of $\overline{MSE}$, $\overline{|Bias|}$ and $\overline{Coverage}$.

5 Results

Table 1: Simulation Setup I

Low dimensional data; No confounding; Balanced propensity;
Linear mean effect; No treatment effect

iters = 30

MSE					
d	Post Lasso	Causal Tree	Causal Forest	X Learner	General ML
5	0.0000*	30.7481	0.1501*	5.1510	21.1132
Absolute Bias					
d	Post Lasso	Causal Tree	Causal Forest	X Learner	General ML
5	0.0000*	0.8306	0.0444*	0.2431	0.5558
Coverage Rate					
d	Post Lasso	Causal Tree	Causal Forest	X Learner	General ML
5	1.0000*	0.9422	0.9507	0.9535*	0.9494

Table 2: Simulation Setup II

Low dimensional data; No confounding; Balanced propensity;
Linear, dense/sparse mean effect; Linear treatment effect

iters = 25

MSE					
d	Post Lasso	Causal Tree	Causal Forest	X Learner	General ML
10	0.0044*	0.7370	0.3830	0.4297	0.3642*
20	0.0049*	0.7983	0.4111*	0.4743	0.4779
Absolute Bias					
d	Post Lasso	Causal Tree	Causal Forest	X Learner	General ML
10	0.0164*	0.4354	0.4422	0.4760	0.4334*
20	0.0155*	0.4461*	0.4603	0.4976	0.5167
Coverage Rate					
d	Post Lasso	Causal Tree	Causal Forest	X Learner	General ML
10	0.9324*	0.8475*	0.4594	0.5017	0.5418
20	0.9430*	0.8574*	0.4758	0.5156	0.4276

Table 3: Simulation Setup III

Low dimensional data; No confounding; Balanced propensity;
No mean effect; Nonlinear treatment effect

iters = 25

MSE					
d	Post Lasso	Causal Tree	Causal Forest	X Learner	General ML
2	0.2308	0.0680	0.0233*	0.0286*	0.0453
4	0.2321	0.0814	0.0298	0.0240*	0.0288*
6	0.2321	0.0950	0.0289	0.0253*	0.0269*
8	0.2324	0.0954	0.0279	0.0260*	0.0254*
Absolute Bias					
d	Post Lasso	Causal Tree	Causal Forest	X Learner	General ML
2	0.3775	0.0614	0.0974	0.0353*	0.0545*
4	0.3781	0.0854	0.1216	0.0757*	0.0664*
6	0.3774	0.0878	0.1128	0.0842*	0.0751*
8	0.3775	0.1025	0.1143	0.0953*	0.0868*
Coverage Rate					
d	Post Lasso	Causal Tree	Causal Forest	X Learner	General ML
2	0.1884	0.9432	0.7731	0.9520*	0.9470*
4	0.2057	0.9341*	0.6662	0.8898	0.9154*
6	0.2186	0.9332*	0.6574	0.8505	0.8860*
8	0.2209	0.9236*	0.6724	0.8155	0.8597*

Table 4: Simulation Setup IV

Low dimensional data; No confounding; Unbalanced propensity propensity; Linear, dense/sparse mean effect; Linear treatment effect

iters = 25

			MSE		
d	Post Lasso	Causal Tree	Causal Forest	X Learner	General ML
2	19.0454*	65.1891	65.3514	27.7156*	36.4505
8	18.5581*	68.4367	65.0075	27.8163*	36.6139
			Absolute Bias		
d	Post Lasso	Causal Tree	Causal Forest	X Learner	General ML
2	2.8977	5.6884	5.6497	2.6281*	0.6852*
8	2.9235	5.6836	5.6447	2.5822*	1.8489*
			Coverage Rate		
d	Post Lasso	Causal Tree	Causal Forest	X Learner	General ML
2	0.3295	0.1771	0.2157	0.6297*	0.9330*
8	0.2298	0.4936	0.1437	0.6496*	0.9035*

Table 5: Simulation Setup V

Low dimensional data; Confounding; Balanced propensity propensity; Linear, sparse mean effect; Nonlinear treatment effect

iters = 25

MSE					
d	Post Lasso	Causal Tree	Causal Forest	X Learner	General ML
2	0.2317	0.1045	0.0301*	0.0299*	0.0622
8	0.2331	0.1130	0.0403*	0.0314*	0.0349
Absolute Bias					
d	Post Lasso	Causal Tree	Causal Forest	X Learner	General ML
2	0.3778	0.0999	0.1103	0.0403*	0.0573*
8	0.3783	0.1132	0.1324	0.1043*	0.1056*
Coverage Rate					
d	Post Lasso	Causal Tree	Causal Forest	X Learner	General ML
2	0.1778	0.9313	0.6959	0.9451*	0.9454*
8	0.2260	0.9269*	0.5834	0.7710	0.8249*

Table 6: Simulation Setup VI

High dimensional data; No Confounding; Balanced propensity propensity; No mean effect; Nonlinear treatment effect

iters = 25

MSE					
d	Post Lasso	Causal Tree	Causal Forest	X Learner	General ML
200	0.3817*	0.2351*	0.7838	0.6935	0.6242
Absolute Bias					
d	Post Lasso	Causal Tree	Causal Forest	X Learner	General ML
200	0.3819*	0.2118*	0.8129	0.7573	0.7208
Coverage Rate					
d	Post Lasso	Causal Tree	Causal Forest	X Learner	General ML
200	0.7560*	0.8298*	0.0610	0.0860	0.0884

Overall comparison (first thoughts):

- No estimator is uniformly superior for all settings.
- Post Lasso tends to outperform in the settings with no treatment effect/ linear treatment effect, sparse model, high dimensional data...
- X Learner and General ML perform pretty well in the settings with nonlinear treatment effect, confounding, unbalanced propensity...
- Causal Tree are more likely to achieve good coverage rate in many settings, while Causal Forest almost performs poorly in terms of this criterion.

  Note: Causal Forests are expected to perform well in low-dimensional settings, where the regression functions can be well approximated by trees [Wager and Athey, 2018]. The generic approaches that combine standard regression Random Forests could theoretically work also in higher dimensions [Wager and Walther, 2015].

6 Applications

In this section, I re-investigate the gender wage gap during transition school to work period in Vietnam:

In our previous study, we use traditional regression model and Oaxaca decomposition method. Taking advantage of the unusual data of youth, Vietnam School-to-Work Transition Surveys 2015 (VSWTS 2015), we find out some meaningful results on youngsters' wage in early career: Firstly, some determinants of early career wage are confirmed. Consistent with human capital theory, actual experience, highest level of education and job tenure significantly affect the hourly wage. Regarding job shopping theory, the number of job mobility has negative effect on wage in several cases. Secondly, there exists a wage gap between young male and female. Although young female workers tend to own more advantage endowments that positively impact on earning, their average wage is lower than the male counterparts. The wage gap even increase when work characteristics are taken in consideration. This unexplained wage gap is usually implied as gender discrimination from literature.

Furthermore, identifying the detailed contributions of the single predictors or sets of predictors is also attractive. For example, one might want to evaluate how much of the gender wage gap is due to sex different in experiences and how much is due to differences in education. Similarly, it might be informative to determine how much of explained gap is related to differing returns to experience or how much to differing returns to education. However, there are limitations of interpretation arising when we apply the detailed decomposition method, especially to analyze contributors to the explained wage gap. Therefore, I aim to quantify the heterogeneity in the gender wage gap by employing causal machine learning estimators which have been discussed in previous sections.

Some related papers are: Bach et al. [2018], Briel and Töpfer [2020].

References

S. Athey and G. Imbens. Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113(27):7353–7360, 2016.

S. Athey, J. Tibshirani, S. Wager, et al. Generalized random forests. *Annals of Statistics*, 47(2):1148–1178, 2019.

P. Bach, V. Chernozhukov, and M. Spindler. Closing the us gender wage gap requires understanding its heterogeneity. *arXiv preprint arXiv:1812.04345*, 2018.

L. Breiman, J. Friedman, C. J. Stone, and R. A. Olshen. *Classification and regression trees*. CRC press, 1984.

S. Briel and M. Töpfer. The gender pay gap revisited: Does machine learning offer new insights? *University of Erlangen-Nürnberg discus-sion paper*, 111, 2020.

V. Chernozhukov, M. Demirer, E. Duflo, and I. Fernandez-Val. Generic machine learning inference on heterogenous treatment effects in randomized experiments. Technical report, National Bureau of Economic Research, 2018.

M. C. Knaus, M. Lechner, and A. Strittmatter. Machine learning estimation of heterogeneous causal effects: Empirical monte carlo evidence. *The Econometrics Journal*, 24(1):134–161, 2021.

S. R. Kuenzel. *Heterogeneous Treatment Effect Estimation Using Machine Learning*. PhD thesis, UC Berkeley, 2019.

S. Wager and S. Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242, 2018.

S. Wager and G. Walther. Adaptive concentration of regression trees, with application to random forests. *arXiv preprint arXiv:1503.06388*, 2015.