

# Introduction to Statistical Machine Learning with Applications in Econometrics

## Lecture 14: Treatment Effect Model and Causal Forests

Instructor: Ma, Jun

Renmin University of China

December 30, 2021

# Causal forests

- ▶ The random forests method is one of the most effective machine learning methods for prediction.
- ▶ A random forest combines a large number of regression trees.
- ▶ The algorithm of the random forests method is very complicated due to its recursive nature and therefore makes it very difficult to study its statistical properties.
- ▶ Athey and Imbens (2016) extended the regression tree algorithm for causal inference.
- ▶ Wager and Athey (2018) extended the random forests method for causal inference. This method is known as causal forests.

## Regression tree

- ▶ Response  $Y$  and  $p$  different predictors  $X = (X_1, X_2, \dots, X_p)^\top$ .
- ▶ Let  $\mathbb{X}$  denote the set of all possible values (support) of  $X$ . This is also called the feature space.
- ▶ Our training data consist of  $\{(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)\}$ , where  $X_i = (X_{1,i}, X_{2,i}, \dots, X_{p,i})^\top$ .
- ▶  $X_{j,i}$ : the value of the  $j$ -th predictor, or input, for observation  $i$ , where  $i = 1, 2, \dots, n$  and  $j = 1, 2, \dots, p$ .
- ▶  $|\Pi|$  denotes the number of elements of  $\Pi$ . A partition  $\Pi = \{\ell_1, \ell_2, \dots, \ell_{|\Pi|}\}$  of  $\mathbb{X}$  is a family of disjoint subsets (leaves) of  $\mathbb{X}$  such that  $\bigcup_{j=1}^{|\Pi|} \ell_j = \mathbb{X}$  and  $\ell_i \cap \ell_j = \emptyset$  if  $i \neq j$ .
- ▶ Let  $\mathcal{S} = \{1, \dots, n\}$  denote the indices of the entire sample. For any  $x \in \mathbb{X}$ ,  $\ell_\Pi(x) = \ell \in \Pi$  such that  $x \in \ell$ .  $\ell_\Pi(x)$  identifies the leaf  $\ell$  to which  $x$  belongs.
- ▶ A partition estimator of  $f(x) = E[Y | X = x]$  using the partition  $\Pi$  is

$$\hat{f}(x | \Pi) = \frac{1}{|\{i \in \mathcal{S} : X_i \in \ell_\Pi(x)\}|} \sum_{i \in \mathcal{S}: X_i \in \ell_\Pi(x)} Y_i.$$

# Splitting rule

- ▶ The regression tree (RT) algorithm determines the final partition and leaves by using the following in-sample criterion:

$$RSS(\Pi) = \frac{1}{|\mathcal{S}|} \sum_{i \in \mathcal{S}} \left( Y_i - \hat{f}(X_i | \Pi) \right)^2.$$

- ▶ The RT algorithm recursively solves

$$\min_{\Pi \text{ is RT-permitted}} RSS(\Pi),$$

where “RT-permitted” means that in each step, the splits are binary with respect to one feature and applied to all remaining nodes that do not satisfy the termination rule.

- ▶ For example, in the initial root step, denote

$$\begin{aligned} \ell_{j,c}^+ &= \left\{ (x_1, \dots, x_p)^\top \in \mathbb{X} : x_j \geq c \right\} \\ \ell_{j,c}^- &= \left\{ (x_1, \dots, x_p)^\top \in \mathbb{X} : x_j < c \right\}, \end{aligned}$$

$$\Pi_{j,c} = \left\{ \ell_{j,c}^-, \ell_{j,c}^+ \right\}.$$

- In the initial step, we solve

$$\min_{j,c} RSS(\Pi_{j,c}),$$

where we have

$$\begin{aligned} & \sum_{i \in \mathcal{S}} \left( Y_i - \hat{f}(X_i | \Pi_{j,c}) \right)^2 \\ &= \sum_{i \in \mathcal{S}} \left\{ 1 \left( X_i \in \ell_{j,c}^+ \right) \left( Y_i - \frac{1}{|\{i : X_i \in \ell_{j,c}^+\}|} \sum_{i: X_i \in \ell_{j,c}^+} Y_i \right)^2 \right. \\ & \quad \left. + 1 \left( X_i \in \ell_{j,c}^- \right) \left( Y_i - \frac{1}{|\{i : X_i \in \ell_{j,c}^-\}|} \sum_{i: X_i \in \ell_{j,c}^-} Y_i \right)^2 \right\}. \end{aligned}$$

- The RT method has the distinct feature of implicit feature selection: variable that is not useful for predicting the response is not selected in the steps of growing the RT.

- ▶ Notice that the criterion function

$$\begin{aligned} \sum_{i \in \mathcal{S}} \left( Y_i - \widehat{f}(X_i | \Pi) \right)^2 \\ = \sum_{i \in \mathcal{S}} Y_i^2 + \sum_{i \in \mathcal{S}} \widehat{f}(X_i | \Pi)^2 - 2 \sum_{i \in \mathcal{S}} Y_i \widehat{f}(X_i | \Pi), \end{aligned}$$

$\sum_{i \in \mathcal{S}} Y_i^2$  does not depend on  $\Pi$  and therefore can be ignored.

- ▶ Let

$$\widehat{Y}_\ell = \frac{1}{|\{i : X_i \in \ell\}|} \sum_{i: X_i \in \ell} Y_i$$

denote the average response in a leaf  $\ell \in \Pi$ . We have

$$\begin{aligned} \sum_{i \in \mathcal{S}} Y_i \widehat{f}(X_i | \Pi) &= \sum_{i \in \mathcal{S}} Y_i \left( \sum_{\ell \in \Pi} 1(X_i \in \ell) \widehat{Y}_\ell \right) \\ &= \sum_{\ell \in \Pi} \left( \sum_{i \in \mathcal{S}} Y_i 1(X_i \in \ell) \right) \widehat{Y}_\ell \\ &= \sum_{\ell \in \Pi} |\{i : X_i \in \ell\}| \widehat{Y}_\ell^2. \end{aligned}$$

► And

$$\begin{aligned}\sum_{i \in \mathcal{S}} \widehat{f}(X_i | \Pi)^2 &= \sum_{i \in \mathcal{S}} \left( \sum_{\ell \in \Pi} 1(X_i \in \ell) \widehat{Y}_\ell \right)^2 \\ &= \sum_{i \in \mathcal{S}} \sum_{\ell \in \Pi} 1(X_i \in \ell) \widehat{Y}_\ell^2 = \sum_{\ell \in \Pi} |\{i : X_i \in \ell\}| \widehat{Y}_\ell^2,\end{aligned}$$

therefore,

$$\sum_{i \in \mathcal{S}} \left( Y_i - \widehat{f}(X_i | \Pi) \right)^2 = \sum_{i \in \mathcal{S}} Y_i^2 - \sum_{i \in \mathcal{S}} \widehat{f}(X_i | \Pi)^2.$$

► In each step, the RT algorithm solves

$$\max_{\Pi \text{ is RT-permitted}} \sum_{i \in \mathcal{S}} \widehat{f}(X_i | \Pi)^2.$$

- ▶ Athey and Imbens (2016) refers the conventional RT algorithm as the adaptive RT to distinguish it from the honest RT algorithm proposed in this paper.
- ▶ Such a criterion is also used in the pruning stage:

$$\widehat{\Pi}^* = \operatorname{argmax}_{\Pi: \Pi < \widehat{\Pi}} \sum_{i \in \mathcal{S}} \widehat{f}(X_i | \Pi)^2 + \lambda |\Pi|,$$

where  $\lambda > 0$  denotes the penalty parameter,  $\Pi < \widehat{\Pi}$  means that  $\Pi$  is the leaves of a sub-tree of the tree corresponding to  $\widehat{\Pi}$  and

$$\widehat{\Pi} = \operatorname{argmax}_{\Pi \text{ is RT-permitted}} \sum_{i \in \mathcal{S}} \widehat{f}(X_i | \Pi)^2.$$



## Sample selection problem

- ▶ The conventional RT algorithm suffers from sample selection and due to this problem it produces biased estimates of  $f(x)$ .
- ▶ The selection bias is not an issue for prediction since more biased estimators can be better predictors than less biased ones due to the bias-variance tradeoff.
- ▶ This selection problem is due to the fact that the split point depends on observations on the response variable and we use the same data for estimation.
- ▶ Example from Athey and Imbens (2016):
  - ▶ Suppose that  $\mathbb{X} = \{L, R\}$ ,  $\bar{Y}_R = |\{i : X_i = R\}|^{-1} \sum_{i: X_i=R} Y_i$  and  $\bar{Y}_L = |\{i : X_i = L\}|^{-1} \sum_{i: X_i=L} Y_i$ ;
  - ▶ Consider the estimated split rule:

$$\hat{\Pi} = \begin{cases} \{\{L\}, \{R\}\} & \text{if } \left| \bar{Y}_R - \bar{Y}_L \right| > c \\ \{\{L, R\}\} & \text{if } \left| \bar{Y}_R - \bar{Y}_L \right| \leq c; \end{cases}$$

- ▶ Then,  $\hat{f}(x | \hat{\Pi})$  is biased for  $E[Y | X = x]$ ,  $x \in \mathbb{X}$ .

## The honest approach

- ▶ To address this issue, Athey and Imbens (2016) proposes to split the sample into the estimation sample  $\mathcal{S}^{\text{est}}$  and the training sample  $\mathcal{S}^{\text{tr}}$ ,  $\mathcal{S} = \mathcal{S}^{\text{est}} \cup \mathcal{S}^{\text{tr}}$  and  $\mathcal{S}^{\text{est}} \cap \mathcal{S}^{\text{tr}} = \emptyset$ .
- ▶ The training sample is used for growing the RT and the estimation sample is used to estimate  $f(x)$ . The two samples are mutually independent. This is called honest RT in Athey and Imbens (2016).
- ▶ Denote

$$\begin{aligned}\widehat{f}(x \mid \mathcal{S}^{\text{est}}, \Pi) &= \frac{1}{|\{i \in \mathcal{S}^{\text{est}} : X_i \in \ell_{\Pi}(x)\}|} \sum_{i \in \mathcal{S}^{\text{est}} : X_i \in \ell_{\Pi}(x)} Y_i \\ &= \sum_{\ell \in \Pi} 1(x \in \ell) \widehat{Y}_{\ell}^{\text{est}} \\ \widehat{Y}_{\ell}^{\text{est}} &= \frac{1}{|\{i \in \mathcal{S}^{\text{est}} : X_i \in \ell\}|} \sum_{i \in \mathcal{S}^{\text{est}} : X_i \in \ell} Y_i\end{aligned}$$

using the estimation sample and  $\Pi$  should be constructed using only the training sample.

# Honest splitting

- ▶ Denote

$$f(x | \Pi) = \mathbb{E}[Y | X \in \ell_{\Pi}(x)] = \mathbb{E}[f(X) | X \in \ell_{\Pi}(x)].$$

- ▶ We have

$$Y_i = \sum_{\ell \in \Pi} 1(X_i \in \ell) \alpha_{\ell} + \epsilon_i$$

$$\mathbb{E}[\epsilon_i | \{1(X_i \in \ell) : \ell \in \Pi\}] = 0,$$

where

$$\alpha_{\ell} = \frac{\mathbb{E}[Y_i 1(X_i \in \ell)]}{\mathbb{E}[1(X_i \in \ell)]} = \mathbb{E}[Y_i | X_i \in \ell]$$

denotes the linear projection coefficients of  $Y_i$  on  $\{1(X_i \in \ell) : \ell \in \Pi\}$ .

- ▶ In this case, the linear projection of  $Y_i$  on  $\{1(X_i \in \ell) : \ell \in \Pi\}$  and  $\mathbb{E}[Y_i | \{1(X_i \in \ell) : \ell \in \Pi\}]$  coincide and

$$f(x | \Pi) = \sum_{\ell \in \Pi} \alpha_{\ell} 1(x \in \ell)$$

$$\mathbb{E}[Y_i | \{1(X_i \in \ell) : \ell \in \Pi\}] = f(X_i | \Pi).$$

► And,

$$\begin{aligned} \mathbb{E} \left[ \widehat{f}(x | \mathcal{S}^{\text{est}}, \Pi) \right] &= \mathbb{E} \left[ \sum_{\ell \in \Pi} 1(x \in \ell) \widehat{Y}_\ell^{\text{est}} \right] \\ &= \sum_{\ell \in \Pi} 1(x \in \ell) \mathbb{E} \left[ \frac{\sum_{i \in \mathcal{S}^{\text{est}}} 1(X_i \in \ell) Y_i}{\sum_{i \in \mathcal{S}^{\text{est}}} 1(X_i \in \ell)} \right] = \sum_{\ell \in \Pi} 1(x \in \ell) \alpha_\ell \\ &= f(x | \Pi). \end{aligned}$$

► Let  $(Y_0, X_0)$  be an unseen data point which is independent of the sample  $\mathcal{S}$ . We compute the mean square prediction error of  $\widehat{f}(X_0 | \mathcal{S}^{\text{est}}, \Pi)$  to  $Y_0$ :

$$\begin{aligned} &\mathbb{E} \left[ \left( Y_0 - \widehat{f}(X_0 | \mathcal{S}^{\text{est}}, \Pi) \right)^2 \right] \\ &= \mathbb{E} \left[ (Y_0 - f(X_0 | \Pi))^2 \right] + \mathbb{E} \left[ \left( \widehat{f}(X_0 | \mathcal{S}^{\text{est}}, \Pi) - f(X_0 | \Pi) \right)^2 \right], \end{aligned}$$

since by  $\mathbb{E} \left[ \widehat{f}(x | \mathcal{S}^{\text{est}}, \Pi) \right] = f(x | \Pi)$  and LIE,

$$\mathbb{E} \left[ (Y_0 - f(X_0 | \Pi)) \left( \widehat{f}(X_0 | \mathcal{S}^{\text{est}}, \Pi) - f(X_0 | \Pi) \right) \right] = 0.$$

► Then,

$$\begin{aligned} E \left[ (Y_0 - f(X_0 | \Pi))^2 \right] \\ = E \left[ Y_0^2 \right] + E \left[ f(X_0 | \Pi)^2 \right] - 2 \cdot E \left[ Y_0 f(X_0 | \Pi) \right] \end{aligned}$$

and by LIE

$$\begin{aligned} E \left[ Y_0 f(X_0 | \Pi) \right] &= E \left[ E \left[ Y_0 f(X_0 | \Pi) \mid \{1(X_0 \in \ell) : \ell \in \Pi\} \right] \right] \\ &= E \left[ f(X_0 | \Pi)^2 \right]. \end{aligned}$$

► Then,

$$E \left[ (Y_0 - f(X_0 | \Pi))^2 \right] = E \left[ Y_0^2 \right] - E \left[ f(X_0 | \Pi)^2 \right].$$

► The honest population-level criterion takes the form

$$H(\Pi) = E \left[ f(X_0 | \Pi)^2 \right] - E \left[ \left( \widehat{f}(X_0 | \mathcal{S}^{\text{est}}, \Pi) - f(X_0 | \Pi) \right)^2 \right],$$

and we use the training data to estimate  $H(\Pi)$ .

## Estimate the criterion

- ▶ We have

$$\begin{aligned}\text{Var} \left[ \widehat{f}(x | \mathcal{S}^{\text{est}}, \Pi) \right] &= \mathbb{E} \left[ \left( \widehat{f}(x | \mathcal{S}^{\text{est}}, \Pi) - f(x | \Pi) \right)^2 \right] \\ &= \sum_{\ell \in \Pi} \mathbb{1}(x \in \ell) \mathbb{E} \left[ \left( \widehat{Y}_{\ell}^{\text{est}} - \alpha_{\ell} \right)^2 \right] \\ &= \sum_{\ell \in \Pi} \mathbb{1}(x \in \ell) \mathbb{E} \left[ \left( \frac{\sum_{i \in \mathcal{S}^{\text{est}}} \mathbb{1}(X_i \in \ell) \epsilon_i}{\sum_{i \in \mathcal{S}^{\text{est}}} \mathbb{1}(X_i \in \ell)} \right)^2 \right].\end{aligned}$$

- ▶ By CLT,

$$\frac{\sum_{i \in \mathcal{S}^{\text{est}}} \mathbb{1}(X_i \in \ell) \epsilon_i}{\sum_{i \in \mathcal{S}^{\text{est}}} \mathbb{1}(X_i \in \ell)} \underset{a}{\sim} \mathcal{N} \left( 0, \frac{1}{|\mathcal{S}^{\text{est}}|} \frac{\mathbb{E} [\epsilon_i^2 \mathbb{1}(X_i \in \ell)]}{(\Pr[X_i \in \ell])^2} \right),$$

and therefore,

$$\text{Var} \left[ \widehat{f}(x | \mathcal{S}^{\text{est}}, \Pi) \right] \approx \sum_{\ell \in \Pi} \mathbb{1}(x \in \ell) \frac{1}{|\mathcal{S}^{\text{est}}|} \frac{\mathbb{E} [\epsilon_i^2 \mathbb{1}(X_i \in \ell)]}{(\Pr[X_i \in \ell])^2}.$$

- By using the training sample, an estimator of  $E[\epsilon_i^2 1(X_i \in \ell)] / \Pr[X_i \in \ell]$  is

$$\hat{\sigma}^2(\ell) = \frac{1}{|\{i \in \mathcal{S}^{\text{tr}} : X_i \in \ell\}|} \sum_{i \in \mathcal{S}^{\text{tr}} : X_i \in \ell} \left( Y_i - \hat{f}(X_i | \mathcal{S}^{\text{tr}}, \Pi) \right)^2,$$

where

$$\hat{f}(x | \mathcal{S}^{\text{tr}}, \Pi) = \frac{1}{|\{i \in \mathcal{S}^{\text{tr}} : X_i \in \ell_{\Pi}(x)\}|} \sum_{i \in \mathcal{S}^{\text{tr}} : X_i \in \ell_{\Pi}(x)} Y_i.$$

- An estimator for

$$E \left[ \left( \hat{f}(X_0 | \mathcal{S}^{\text{est}}, \Pi) - f(X_0 | \Pi) \right)^2 \right] = E \left[ \text{Var} \left[ \hat{f}(X_0 | \mathcal{S}^{\text{est}}, \Pi) \right] \right]$$

is

$$\frac{1}{|\mathcal{S}^{\text{est}}|} \sum_{\ell \in \Pi} \hat{\sigma}^2(\ell).$$

- And, since  $E \left[ \hat{f}(x | \mathcal{S}^{\text{tr}}, \Pi) \right] = f(x | \Pi)$ ,

$$\text{Var} \left[ \hat{f}(x | \mathcal{S}^{\text{tr}}, \Pi) \right] = E \left[ \hat{f}(x | \mathcal{S}^{\text{tr}}, \Pi)^2 \right] - f(x | \Pi)^2.$$

- ▶ An estimator for  $E [f (X_0 | \Pi)^2]$  using the training sample is

$$\frac{1}{|\mathcal{S}^{\text{tr}}|} \sum_{i \in \mathcal{S}^{\text{tr}}} \widehat{f} (X_i | \mathcal{S}^{\text{tr}}, \Pi)^2 - \frac{1}{|\mathcal{S}^{\text{tr}}|} \sum_{\ell \in \Pi} \widehat{\sigma}^2 (\ell).$$

- ▶ An estimator for  $H (\Pi)$  using the training sample is

$$\widehat{H} (\Pi) = \frac{1}{|\mathcal{S}^{\text{tr}}|} \sum_{i \in \mathcal{S}^{\text{tr}}} \widehat{f} (X_i | \mathcal{S}^{\text{tr}}, \Pi)^2 - \left( \frac{1}{|\mathcal{S}^{\text{tr}}|} + \frac{1}{|\mathcal{S}^{\text{est}}|} \right) \sum_{\ell \in \Pi} \widehat{\sigma}^2 (\ell).$$

- ▶ The honest RT algorithm recursively solves

$$\max_{\Pi \text{ is RT-permitted}} \widehat{H} (\Pi).$$

- ▶ The conventional (adaptive) RT algorithm recursively solves

$$\max_{\Pi \text{ is RT-permitted}} \sum_{i \in \mathcal{S}} \widehat{f} (X_i | \Pi)^2.$$



- ▶ Disadvantages of the honest RT:
  - ▶ smaller samples for feature space splitting (training sample) and estimation (estimation sample);
  - ▶ the results depend on how the data is split into the training and estimation samples;
  - ▶ the forest approach alleviates this issue by using many random splits.

# The treatment effect model

- ▶ We consider the problem of estimating the causal effect of a binary explanatory variable, which is referred as the treatment effect in the literature. The treatment effect model is different from the linear regression model.
- ▶ In econometrics, the treatment effect model is very often used for evaluating social program/experiment.
- ▶ Example 1: Suppose that a selected set of individuals receive training or education initiated by the government with a view to enhancing their employment prospects. Suppose that the government has collected the earnings data for the individuals who received the training and for the individuals who did not. The main purpose of methods of program evaluations is to quantify and estimate the effect of the training program.

- ▶ Example 2: Suppose that an education program required high schools to agree to assign teachers and students to small (13 to 17 students) or large (22 to 26 students) classes. The government is interested in the effect of class size on student achievement.
- ▶ Such a question can arise in various other situations. A medical experiment studies on the effects of new treatment ask similar questions. One group of patients has received new treatment, and the other group has not.

## Potential outcome variables

- ▶  $Y_i$ : outcome variable;  $D_i \in \{0, 1\}$ : the binary explanatory variable;  $X_{1,i}, \dots, X_{p,i}$ : other observed explanatory variables;  $\epsilon_i$ : unobserved explanatory factors.
- ▶ The variable  $D_i$  is a binary variable taking 1 if the individual has gone through the treatment and 0 otherwise. The treatment here represents the actual treatment. The econometrician usually observes the treatment status for each individual  $D_i$ .
- ▶  $(X_{1,i}, \dots, X_{p,i})$  represents a vector of various demographic characteristics for individual  $i$ . E.g., the variables can be annual income, age, gender, status of marriage, the number of children, education, etc. These represent all the observable characteristics of individual  $i$ .
- ▶ Suppose that  $Y_i$  is generated by  $Y_i = g(D_i, X_{1,i}, \dots, X_{p,i}, \epsilon_i)$ .
- ▶  $g$  is unknown and in the treatment effect model, we do not assume  $g$  is linear.

- ▶ The outcome variable  $Y_i(1) = g(1, X_{1,i}, \dots, X_{p,i}, \epsilon_i)$  represents a potential outcome of an individual  $i$  in the treatment state (e.g. training is received or studying in a reduced-size class). The variable  $Y_i(0) = g(0, X_{1,i}, \dots, X_{p,i}, \epsilon_i)$  represents a potential outcome of the same individual  $i$  in the control state (e.g. training is received or studying in a normal-size class).
- ▶ Thus, each individual has a random vector  $(Y_i(1), Y_i(0))$  that represents potential outcomes depending on the state (treatment or control). Certainly,  $(Y_i(1), Y_i(0))$  are correlated.
- ▶ The econometrician cannot observe the random vector  $(Y_i(1), Y_i(0))$  jointly, because for each individual, only one potential outcome ( $Y_i(1)$  or  $Y_i(0)$ ) is realized, depending on whether the individual  $i$  has gone through the treatment or not.

## The relationship between $D_i$ and $(Y_i(1), Y_i(0))$

- ▶ In a medical experiment, the individual is chosen to be in the treatment group through some randomization device or a lottery. In these cases,  $D_i \perp\!\!\!\perp (Y_i(1), Y_i(0))$  (i.e.,  $D_i$  is independent of  $(Y_i(1), Y_i(0))$ ).
- ▶ For evaluating social experiment/program with observational data, it may not be convincing to assume  $D_i \perp\!\!\!\perp (Y_i(1), Y_i(0))$ .

## Treatment effects

- ▶ The individual treatment effect (ITE) for each individual  $i$  is defined as:

$$Y_i(1) - Y_i(0).$$

- ▶ The ITE is the difference between the potential outcomes in two different states for the same person.
- ▶ The ITE is a counterfactual quantity, in the sense that in the actual world, we cannot observe the vector  $(Y_i(1), Y_i(0))$ .
- ▶ There are mainly two quantities of interest: ATE (average treatment effect)

$$\text{ATE} = E[Y_i(1) - Y_i(0)],$$

and ATT (average treatment effect on the treated)

$$\text{ATT} = E[Y_i(1) - Y_i(0) \mid D_i = 1].$$

- ▶ The average treatment effect on the treated is the treatment effect of the people who have gone through the treatment.

- ▶ Note that the expectation in the definition of ATE involves the joint distribution of  $(Y_i(1), Y_i(0))$ , and the expectation in the definition of ATT involves the joint distribution of  $(Y_i(1), Y_i(0), D_i)$ , which are both unobserved.
- ▶ ATE or ATT can not be estimated accurately merely by collecting a large size of samples.



# The observed information

- ▶ The econometrician observes the treatment status  $D_i$  and covariates  $X_i$ . She also observes the outcome variable:

$$Y_i = D_i Y_i(1) + (1 - D_i) Y_i(0).$$

- ▶ The observed outcome variable  $Y_i$  is not the same as the potential outcomes  $Y_i(1)$  or  $Y_i(0)$ . It is a realized outcome for an individual  $i$  depending on whether she has received treatment ( $Y_i$  is realized to be  $Y_i(1)$ ) or not ( $Y_i$  is realized to be  $Y_i(0)$ ).
- ▶ Identification of these parameters is concerned with the following question: can we uniquely determine the value of these parameters once we know the joint distribution of the observable random variables?

## Randomized experiments

- ▶ In medical experiments, the treatment is performed using a randomization device. More specifically, for patient  $i$ , a lottery is run, and the patient is selected into the treated group with the design probability  $p$ , and stays in the control group with the design probability  $1 - p$ .
- ▶ In these cases, we have  $D_i \perp\!\!\!\perp (Y_i(1), Y_i(0), X_{1,i}, \dots, X_{p,i})$ . Randomized experiment assumption requires that knowing whether patient  $i$  is treated or not gives one no informational advantage in predicting the potential outcomes of  $i$  over another who does not know whether patient  $i$  is treated or not.
- ▶ This assumption is still possibly violated in medical studies if only those patients who have higher potential treatment effect are selected into treatment among all the patients in the study on purpose.
- ▶ In this case, observing  $D_i$  will give information about the treatment effect  $(Y_i(1) - Y_i(0))$  for individual  $i$ .

- ▶ We use the following result from probability theory: if  $V \perp\!\!\!\perp W$ , then for any function  $f$ ,

$$E [f (V, W) | W = w] = E [f (V, w)] . \quad (1)$$

- ▶ By (1) and the randomized experiment assumption,  $D_i \perp\!\!\!\perp (Y_i (1), Y_i (0))$ , we have

$$\begin{aligned} \text{ATE} &= E [Y_i (1) - Y_i (0)] \\ &= E [Y_i (1)] - E [Y_i (0)] \\ &= E [D_i Y_i (1) + (1 - D_i) Y_i (0) | D_i = 1] \\ &\quad - E [D_i Y_i (1) + (1 - D_i) Y_i (0) | D_i = 0] \\ &= E [Y_i | D_i = 1] - E [Y_i | D_i = 0] . \end{aligned}$$

- By LIE,

$$\begin{aligned} E [Y_i D_i] &= E [E [Y_i D_i | D_i]] \\ &= \Pr [D_i = 1] E [Y_i D_i | D_i = 1] \\ &\quad + \Pr [D_i = 0] E [Y_i D_i | D_i = 0] \\ &= E [D_i] E [Y_i | D_i = 1], \end{aligned}$$

where

$$\begin{aligned} E [Y_i D_i | D_i = 0] &= E [(D_i Y_i (1) + (1 - D_i) Y_i (0)) D_i | D_i = 0] \\ &= 0 \end{aligned}$$

follows from (1).

- Similarly, we have

$$E [Y_i | D_i = 0] = \frac{E [Y_i (1 - D_i)]}{E [1 - D_i]}.$$

- ▶ We can write

$$\text{ATE} = \frac{\text{E}[Y_i D_i]}{\text{E}[D_i]} - \frac{\text{E}[Y_i (1 - D_i)]}{\text{E}[1 - D_i]},$$

where the right hand side depends on the joint distribution of the observed random variables.

- ▶ For estimation, we replace the population mean by the sample mean (this is sometimes called the analogue principle):

$$\widehat{\text{ATE}} = \frac{\frac{1}{n} \sum_{i=1}^n Y_i D_i}{\frac{1}{n} \sum_{i=1}^n D_i} - \frac{\frac{1}{n} \sum_{i=1}^n Y_i (1 - D_i)}{\frac{1}{n} \sum_{i=1}^n (1 - D_i)}.$$

- ▶ We can check its consistency by using LLN and Slutsky's lemma.
- ▶ This randomization assumption is not convincing when the individuals in the social experiments are people who may select into the treatment or not.

## Comparison with the linear regression

- ▶ It seems that  $D_i$  is nothing but a dummy variable. Can we run a regression of  $Y_i$  on  $D_i$  and  $X_{1,i}, \dots, X_{p,i}$  to estimate the ATE? Can the parameter of interest, the ATE, be formulated as a coefficient in a regression model.
- ▶ One possible assumption is that

$$Y_i = g(D_i, X_{1,i}, \dots, X_{p,i}, \epsilon_i) = \gamma_0 + \gamma_1 D_i + \sum_{j=1}^p \beta_j X_{j,i} + \epsilon_i.$$

In this case, the ITE  $Y_i(1) - Y_i(0) = \gamma_1$  is constant. This is very unrealistic. We investigate alternative model assumptions.

- ▶ We first consider the following model assumption

$$Y_i(0) = \mu_0 + U_i(0)$$

$$Y_i(1) = \mu_1 + U_i(1),$$

where  $\mu_0$  and  $\mu_1$  are constants common across individuals and assumed to be nonstochastic and  $(U_i(0), U_i(1))$  are stochastic components.

- ▶ We denote  $X_i = (X_{1,i}, \dots, X_{p,i})^\top$  for the vector of observed covariates.
- ▶ We assume  $E[U_i(0) | X_i] = E[U_i(1) | X_i]$ , which implies

$$E[Y_i(1) - Y_i(0) | X_i] = \mu_1 - \mu_0,$$

i.e., the ITE is mean independent of  $X_i$  but it can be random.  
And by LIE,

$$ATE = E[Y_i(1) - Y_i(0)] = \mu_1 - \mu_0.$$

- ▶ We assume  $E[Y_i(1) | D_i, X_i] = E[Y_i(1) | X_i]$  and  $E[Y_i(0) | D_i, X_i] = E[Y_i(0) | X_i]$ , i.e., the conditional mean independence of potential outcomes with treatment status, conditional on demographic status  $X_i$ .
- ▶ When we focus on a sub-population of individuals with specific demographic status  $X_i$ ,  $Y_i(1)$  and  $Y_i(0)$  are both mean independent of  $D_i$ .

- ▶ Let us write

$$\begin{aligned} E [Y_i | D_i, X_i] &= D_i E [Y_i (1) | D_i, X_i] + (1 - D_i) E [Y_i (0) | D_i, X_i] \\ &= D_i E [Y_i (1) - Y_i (0) | D_i, X_i] + E [Y_i (0) | D_i, X_i] \\ &= D_i E [Y_i (1) - Y_i (0) | X_i] + E [Y_i (0) | X_i], \end{aligned}$$

where the last equality follows from the conditional mean independence assumption.

- ▶ By the assumption  $E [U_i (0) | X_i] = E [U_i (1) | X_i]$ , we have

$$\begin{aligned} D_i E [Y_i (1) - Y_i (0) | X_i] + E [Y_i (0) | X_i] \\ &= D_i (\mu_1 - \mu_0) + E [Y_i (0) | X_i] \\ &= \mu_0 + D_i (\mu_1 - \mu_0) + h (X_{1,i}, \dots, X_{p,i}), \end{aligned}$$

where we denote  $h (X_{1,i}, \dots, X_{p,i}) = E [U_i (0) | X_i]$ .



- ▶ Therefore, we have

$$E [Y_i | D_i, X_i] = \mu_0 + (\mu_1 - \mu_0) D_i + h (X_{1,i}, \dots, X_{p,i}) .$$

- ▶ Define

$$V_i = Y_i - E [Y_i | D_i, X_i]$$

and now we have the following regression model:

$$Y_i = \mu_0 + (\mu_1 - \mu_0) D_i + h (X_{1,i}, \dots, X_{p,i}) + V_i .$$

- ▶ We have  $E [V_i | D_i, X_i] = 0$  by definition.
- ▶ We assume  $h$  is linear in  $X_{1,i}, \dots, X_{p,i}$ :

$$h (X_{1,i}, \dots, X_{p,i}) = \sum_{j=1}^p \beta_j X_{j,i} ,$$

and then

$$Y_i = \mu_0 + (\mu_1 - \mu_0) D_i + \sum_{j=1}^p \beta_j X_{j,i} + V_i .$$

- ▶ A multiple linear regression of  $Y_i$  on  $D_i$  and  $X_{1,i}, \dots, X_{p,i}$  consistently estimates  $ATE = (\mu_1 - \mu_0)$ .

- ▶ We assume  $E[U_i(0) | X_i] = E[U_i(1) | X_i]$ , which implies

$$E[Y_i(1) - Y_i(0) | X_i] = \mu_1 - \mu_0.$$

- ▶ This assumption implies that the conditional average treatment effect given  $X_i$  does not depend on  $X_i$ , the characteristics of individual  $i$ .
- ▶ This assumption can be unrealistic. E.g., Average treatment of the class-size is the same between students from high-income family and students from low-income family.

# Unconfoundedness assumption

- ▶ Unconfoundedness is the key assumption of the basic treatment effect model.
- ▶ Unconfoundedness assumption:  $(Y_i(1), Y_i(0)) \perp\!\!\!\perp D_i \mid X_i$ , i.e.,  $(Y_i(1), Y_i(0))$  and  $D_i$  are conditionally independent given  $X_i$ .
- ▶ Unconfoundedness can be thought of as an assumption that the decision to take the treatment is purely random for individuals with similar values of the covariates.
- ▶ Suppose that we have three random vectors  $V$ ,  $W$  and  $X$ , where  $(V, W)$  is a continuous random vector. Then we say  $V$  and  $W$  are conditionally independent given  $X$ , if for all possible values of  $v$ ,  $w$  and  $x$ ,

$$f_{(V,W)|X}(v, w \mid x) = f_{V|X}(v \mid x) f_{W|X}(w \mid x).$$

- Unconfoundedness is satisfied if  $(Y_i, D_i)$  are generated by the model

$$Y_i = g(D_i, X_{1,i}, \dots, X_{p,i}, \epsilon_i)$$
$$D_i = m(X_{1,i}, \dots, X_{p,i}, \eta_i)$$

and  $\epsilon_i \perp\!\!\!\perp \eta_i \mid X_{1,i}, \dots, X_{p,i}$ .

## More on conditional independence

- ▶ When  $V$  and  $W$  are conditionally independent given  $X$ , one can easily see that for any function  $\varphi$ ,

$$E[\varphi(V) \mid W, X] = E[\varphi(V) \mid X].$$

I.e., once we observe  $X$ , knowledge of  $W$  does not give us any further advantage in predicting the value of  $\varphi(V)$ .

- ▶ We notice that

$$\begin{aligned} f_{(V,W)|X}(v, w \mid x) &= \frac{f_{(V,W,X)}(v, w, x)}{f_X(x)} \\ &= \frac{f_{(V,W,X)}(v, w, x)}{f_{(W,X)}(w, x)} \frac{f_{(W,X)}(w, x)}{f_X(x)} \\ &= f_{V|(W,X)}(v \mid w, x) f_{W|X}(w, x). \end{aligned}$$

- Therefore, we have  $f_{V|X}(v|x) = f_{V|(W,X)}(v|w,x)$ , if  $(V, W)$  are conditionally independent given  $X$ . Hence,

$$\begin{aligned} E[\varphi(V) | W = w, X = x] &= \int \varphi(v) f_{V|(W,X)}(v|w,x) dv \\ &= \int \varphi(v) f_{V|X}(v|x) dv \\ &= E[\varphi(V) | X = x]. \end{aligned}$$

- Therefore, the unconfoundedness assumption  $(Y_i(1), Y_i(0)) \perp\!\!\!\perp D_i | X_i$  implies the conditional mean independence assumption:

$$\begin{aligned} E[Y_i(1) | D_i, X_i] &= E[Y_i(1) | X_i] \\ E[Y_i(0) | D_i, X_i] &= E[Y_i(0) | X_i]. \end{aligned}$$

- We can also show: if  $V \perp\!\!\!\perp W | X$ ,

$$E[\eta(V, W) | X, W = w] = E[\eta(V, w) | X]. \quad (2)$$

# The unconfoundedness and randomization assumptions

- ▶ It can be shown that the randomization assumption  $(Y_i(1), Y_i(0), X_i) \perp\!\!\!\perp D_i$  implies the unconfoundedness assumption  $(Y_i(1), Y_i(0)) \perp\!\!\!\perp D_i \mid X_i$ .
- ▶ The randomized experiment assumption does not allow  $X_{1,i}, \dots, X_{p,i}$  to be correlated with  $D_i$ ,
- ▶ The unconfounded condition allows  $D_i$  to be affected by  $X_{1,i}, \dots, X_{p,i}$ , while the randomized experiment assumption does not.

# Identification of ATE

- By LIE, we have

$$\begin{aligned} \text{ATE} &= E [Y_i (1) - Y_i (0)] \\ &= E [E [Y_i (1) | X_i]] - E [E [Y_i (0) | X_i]], \end{aligned} \quad (3)$$

and

$$\begin{aligned} E [Y_i D_i | X_i] &= E [E [Y_i D_i | X_i, D_i] | X_i] \\ &= \Pr [D_i = 1 | X_i] E [Y_i D_i | X_i, D_i = 1] \\ &\quad + \Pr [D_i = 0 | X_i] E [Y_i D_i | X_i, D_i = 0]. \end{aligned}$$



- By the unconfoundedness assumption:  
 $(Y_i(1), Y_i(0)) \perp\!\!\!\perp D_i \mid X_i$ , the result (2) and the relation  
 $Y_i = D_i Y_i(1) + (1 - D_i) Y_i(0)$ , we have

$$\begin{aligned} & E[Y_i D_i \mid X_i, D_i = 1] \\ &= E[(D_i Y_i(1) + (1 - D_i) Y_i(0)) D_i \mid X_i, D_i = 1] = E[Y_i(1) \mid X_i] \end{aligned}$$

and

$$E[Y_i D_i \mid X_i, D_i = 0] = 0.$$

- Therefore, we have

$$E[Y_i D_i \mid X_i] = \Pr[D_i = 1 \mid X_i] E[Y_i(1) \mid X_i] \quad (4)$$

and similarly,

$$E[Y_i(1 - D_i) \mid X_i] = \Pr[D_i = 0 \mid X_i] E[Y_i(0) \mid X_i]. \quad (5)$$

- ▶ Now (3), (4), (5) and LIE imply

$$\begin{aligned} \text{ATE} &= \mathbb{E} \left[ \frac{\mathbb{E}[Y_i D_i \mid X_i]}{\Pr[D_i = 1 \mid X_i]} \right] - \mathbb{E} \left[ \frac{\mathbb{E}[Y_i (1 - D_i) \mid X_i]}{\Pr[D_i = 0 \mid X_i]} \right] \\ &= \mathbb{E} \left[ \frac{Y_i D_i}{\Pr[D_i = 1 \mid X_i]} - \frac{Y_i (1 - D_i)}{\Pr[D_i = 0 \mid X_i]} \right]. \end{aligned}$$

Now the right hand side depends only on the joint distribution of observed random variables.

- ▶ Denote

$$p(x) = \Pr[D_i = 1 \mid X_i = x].$$

This function is called propensity score. It is the probability of the event that the individual belongs to the treatment group, given that the observed characteristics are  $x \in \mathbb{R}^p$ .

## Baseline estimation of ATE

- ▶ Let  $\hat{p}(x)$  be an estimator of the propensity score, then we can estimate the ATE:

$$\widehat{\text{ATE}} = \frac{1}{n} \sum_{i=1}^n \left\{ \frac{Y_i D_i}{\hat{p}(X_i)} - \frac{Y_i (1 - D_i)}{1 - \hat{p}(X_i)} \right\}.$$

This is known as the inverse probability weighting (IPW) estimator.

- ▶ It is straightforward to construct  $\hat{p}(x)$  if  $X_i$  is discrete:

$$\hat{p}(x) = \frac{\sum_{i=1}^n 1(D_i = 1, X_i = x)}{\sum_{i=1}^n 1(X_i = x)}.$$

- ▶ If  $X_i$  is continuous, we specify a parametric model for the propensity score:

$$\Pr [D_i = 1 \mid X_i] = \Phi(\beta_0 + \beta_1 X_{1,i} + \dots + \beta_p X_{p,i})$$

as what we did for the Probit model. This gives a parametric model for the propensity score.  $(\beta_0, \dots, \beta_p)$  can be estimated by MLE (denoted by  $(\hat{\beta}_0, \dots, \hat{\beta}_p)$ ). We bootstrap the standard errors.

- ▶ The estimated propensity score is

$$\hat{p}(X_i) = \Phi(\hat{\beta}_0 + \hat{\beta}_1 X_{1,i} + \dots + \hat{\beta}_p X_{p,i}).$$

- ▶ This estimator is known to be consistent and asymptotically normally distributed, if our propensity score model is correct.
- ▶ This approach has the drawback that if our model for the propensity score is wrong, the ATE estimator is inconsistent.
- ▶ Actually,  $p(x) = E[D_i | X_i = x]$  can be estimated without specifying a parametric model for it.

## k-NN estimator

- ▶ The  $k$ -nearest neighbor ( $k$ -NN) estimator is the simplest nonparametric estimator of  $p(x)$ .
- ▶ Fix  $x_0$  and suppose that we want to estimate  $p(x_0)$  at this point. Assume that  $p$  is a smooth function, which means that its graph does not change too much.
- ▶  $p(x)$  should be close to  $p(x_0)$  when  $x$  is close enough to  $x_0$ .  $p(X_i)$  would be close to  $p(x_0)$  for observations  $X_i$  close to  $x_0$ .
- ▶ We simply average these  $p(X_i)$  for observations  $X_i$  close to  $x_0$ . We do not observe  $p(X_i)$  but use  $D_i$  instead.
- ▶ Let

$$d_i(x_0) = \|X_i - x_0\| = \sqrt{(X_i - x_0)^T (X_i - x_0)}$$

denote the distance of  $X_i$  to  $x_0$ .

- ▶ After computing the distance for all  $n$  observations in the sample, we sort them in the increasing order

$$d_{(1)}(x_0) \leq d_{(2)}(x_0) \leq \dots \leq d_{(n)}(x_0).$$

- ▶ Let  $N_k(x_0)$  denote the identities of the  $k$ -nearest neighbors of  $x_0$ :

$$N_k(x_0) = \{i : d_i(x_0) \leq d_{(k)}(x_0)\}.$$

- ▶ The  $k$ -NN nonparametric estimator of  $p(x_0)$  is

$$\hat{p}_{kNN}(x_0) = \frac{1}{k} \sum_{i \in N_k(x_0)} D_i.$$

- ▶ The  $k$ -NN estimator is simply an average of the values of  $D_i$  across the  $k$  closest observations in terms of  $X_i$ .
- ▶ There is a data-driven procedure to select  $k$  in practical applications.
- ▶ The nonparametric ATE estimator using  $\hat{p}_{kNN}(X_i)$  is consistent and asymptotically normal. It does not require a parametric model for the propensity score.

## Conditional average treatment effect

- ▶ In the baseline treatment effect model, the ITE  $Y_i(1) - Y_i(0)$  is not constant. The ATE  $ATE = E[Y_i(1) - Y_i(0)]$  is an average of the individual treatment effects over the entire population.
- ▶ Define the conditional average treatment effect (CATE)

$$CATE(A) = E[Y_i(1) - Y_i(0) \mid X_i \in A], A \subseteq \mathbb{X},$$

where  $A$  is a set of potential values of  $X_i$ . CATE( $A$ ) captures the treatment effect for a sub-population of individuals defined by their covariates. For example,

$$A = \{x \in \mathbb{R}^P : x_1 < c_1, x_2 > c_2\}$$

for some constants  $c_1, c_2$ .

- ▶ When there is a lot of heterogeneity in  $Y_i(1) - Y_i(0)$  across the observations, it is possible that some sub-populations have large effects and some sub-populations have small effects.

- ▶ It is possible that  $ATE = E [Y_i (1) - Y_i (0)] = 0$  but looking at

$$ATE = \Pr [X_i \in A] CATE (A) + \Pr [X_i \in A^c] CATE (A^c) ,$$

we find that  $CATE (A)$  is positive and  $CATE (A^c)$  is negative.

- ▶ Even when treatment is randomized and the covariates are not even needed for the estimation of the ATE, they can be used for learning  $CATE (A)$ , which can be very different in different sub-populations defined by  $A \subseteq \mathbb{X}$ .



# CATE function

- ▶ Define the CATE function

$$\tau(x) = E[Y_i(1) - Y_i(0) \mid X_i = x], x \in \mathbb{X}.$$

Then,

$$\text{CATE}(A) = \frac{E[\tau(X_i) 1(X_i \in A)]}{\Pr[X_i \in A]}.$$

- ▶ If we have an estimator  $\hat{\tau}(x)$  of  $\tau(x)$ , we can easily construct an estimator of  $\text{CATE}(A)$ :

$$\widehat{\text{CATE}}(A) = \frac{\sum_{i=1}^n \hat{\tau}(X_i) 1(X_i \in A)}{\sum_{i=1}^n 1(X_i \in A)}.$$

- ▶ Estimation of CATE reduces to estimation of  $\tau(x)$ .
- ▶ Another motivation for estimation of  $\tau(x)$  is inference on optimal policy.

# Optimal policy

- ▶ Denote  $\mu_1(x) = E[Y_i(1) | X_i = x]$  and  $\mu_0(x) = E[Y_i(0) | X_i = x]$ . Then  $\tau(x) = \mu_1(x) - \mu_0(x)$ .
- ▶ Consider a social planner's problem. Let  $p : \mathbb{X} \rightarrow [0, 1]$  be a function and  $p(x)$  denotes the probability that an individual with characteristic  $x \in \mathbb{X}$  is assigned to treatment.
- ▶ The social welfare corresponding to an assignment rule  $p$  is

$$\begin{aligned} W(p) &= E[\mu_1(X_i)p(X_i) + \mu_0(X_i)(1 - p(X_i))] \\ &= E[\mu_0(X_i)] + E[\tau(X_i)p(X_i)]. \end{aligned}$$

- ▶ The social planner's problem is

$$\max_p W(p) \text{ subject to } E[p(X_i)] \leq c,$$

where  $c \in (0, 1)$  is fraction of individuals that can be treated and  $E[p(X_i)] \leq c$  is the budget constraint.

- ▶ The solution must take the form of  $p_\gamma(x) = 1(\tau(x) \geq \gamma)$ . The solution to the planner's problem is  $p_{\gamma^*}$ , where  $\gamma^*$  is determined by the budget constraint  $E[p_{\gamma^*}(X_i)] = c$ . Therefore,  $\Pr[\tau(X_i) \geq \gamma^*] = c$ .  $\gamma^*$  is the  $(1 - c)$  quantile of the random variable  $\tau(X_i)$ .
- ▶ Estimation of the optimal treatment assignment rule reduces to estimation of  $\tau(x)$ .
- ▶ In a simpler problem which does not allow for probabilistic assignment, if  $X_i \in A$  is assigned treatment, the social welfare is

$$\begin{aligned} W(A) &= E[\mu_1(X_i) 1(X_i \in A) + \mu_0(X_i) 1(X_i \in A^c)] \\ &= E[\mu_0(X_i)] + E[\tau(X_i) 1(X_i \in A)]. \end{aligned}$$

It can be shown that

$$\{x \in \mathbb{X} : \tau(x) \geq 0\} = \underset{A \subseteq \mathbb{X}}{\operatorname{argmax}} W(A).$$

# Tree-based estimation of CATE

- ▶ Since the ITE  $\tau_i = Y_i(1) - Y_i(0)$  is unobserved, the regression tree method is not easy to be extended for estimation of  $\tau(x)$ .
- ▶ Under unconfoundedness, we have

$$\begin{aligned}\tau(x) &= E[Y_i(1) | X_i = x] - E[Y_i(0) | X_i = x] \\ &= E[Y_i | X_i = x, D_i = 1] - E[Y_i | X_i = x, D_i = 0].\end{aligned}$$

Conditional on  $X_i = x$ , the expectation of  $Y_i(1)$  is equal to the expectation of  $Y_i$  for the treated group.

- ▶ This motivates the tree estimator of  $\tau(x)$ :

$$\hat{\tau}(x | \Pi) = \frac{\sum_{i \in \mathcal{S}: X_i \in \ell_{\Pi}(x), D_i=1} Y_i}{|\{i \in \mathcal{S} : X_i \in \ell_{\Pi}(x), D_i = 1\}|} - \frac{\sum_{i \in \mathcal{S}: X_i \in \ell_{\Pi}(x), D_i=0} Y_i}{|\{i \in \mathcal{S} : X_i \in \ell_{\Pi}(x), D_i = 0\}|}.$$

- ▶  $\mathcal{X}$  is partitioned into leaves in  $\Pi$ . When the resulting leaves are small enough so that the observation in the same leaf can be viewed as if they come from a randomized experiment.

## Conventional (adaptive) causal trees

- ▶ To grow the tree, we should recursively solve

$$\min_{\Pi \text{ is RT-permitted}} \text{RSS}(\Pi),$$

where the in-sample criterion is given by

$$\text{RSS}(\Pi) = \frac{1}{|\mathcal{S}|} \sum_{i \in \mathcal{S}} (\tau_i - \hat{\tau}(X_i | \Pi))^2.$$

- ▶  $\tau_i$  is unobserved, but it is shown that it is equivalent to solve

$$\max_{\Pi \text{ is RT-permitted}} \sum_{i \in \mathcal{S}} \hat{\tau}(X_i | \Pi)^2.$$

- ▶ This is the adaptive causal tree algorithm. It is a modification of the conventional (adaptive) regression tree method.

# Honest causal trees

- ▶ Honest causal trees proposed in Athey and Imbens (2016) split  $\mathcal{S}$  into the training sample  $\mathcal{S}^{\text{tr}}$  and the estimation sample  $\mathcal{S}^{\text{est}}$ . For any partition  $\Pi$ , denote

$$\widehat{\tau}(x \mid \mathcal{S}^{\text{est}}, \Pi) = \frac{\sum_{i \in \mathcal{S}^{\text{est}}: X_i \in \ell_{\Pi}(x), D_i=1} Y_i}{|\{i \in \mathcal{S}^{\text{est}} : X_i \in \ell_{\Pi}(x), D_i = 1\}|} - \frac{\sum_{i \in \mathcal{S}^{\text{est}}: X_i \in \ell_{\Pi}(x), D_i=0} Y_i}{|\{i \in \mathcal{S}^{\text{est}} : X_i \in \ell_{\Pi}(x), D_i = 0\}|}.$$

- ▶ We choose  $\Pi$  to minimize the mean square prediction error of  $\widehat{\tau}(X_0 \mid \mathcal{S}^{\text{est}}, \Pi)$  to  $\tau_0$ :

$$\min_{\Pi \text{ is RT-permitted}} \mathbb{E} \left[ (\tau_0 - \widehat{\tau}(X_0 \mid \mathcal{S}^{\text{est}}, \Pi))^2 \right].$$

- ▶ Similarly, it is shown that  $\mathbb{E} \left[ (\tau_0 - \widehat{\tau}(X_0 \mid \mathcal{S}^{\text{est}}, \Pi))^2 \right]$  is equal to  $H(\Pi)$ , up to a constant, where the honest population-level criterion takes the form

$$H(\Pi) = \mathbb{E} \left[ \tau(X_0 \mid \Pi)^2 \right] - \mathbb{E} \left[ (\widehat{\tau}(X_0 \mid \mathcal{S}^{\text{est}}, \Pi) - \tau(X_0 \mid \Pi))^2 \right].$$

- ▶ We estimate both components of this expectation using only the training sample.
- ▶ The honest causal tree is chosen by solving

$$\max_{\Pi \text{ is RT-permitted}} \widehat{H}(\Pi),$$

where

$$\widehat{H}(\Pi) = \frac{1}{|\mathcal{S}^{\text{tr}}|} \sum_{i \in \mathcal{S}^{\text{tr}}} \widehat{\tau}(X_i | \mathcal{S}^{\text{tr}}, \Pi)^2 - \left( \frac{1}{|\mathcal{S}^{\text{tr}}|} + \frac{1}{|\mathcal{S}^{\text{est}}|} \right) \sum_{\ell \in \Pi} \left( \frac{\widehat{\sigma}_1^2(\ell)}{\widehat{p}_1(\ell)} + \frac{\widehat{\sigma}_0^2(\ell)}{\widehat{p}_0(\ell)} \right),$$

$\widehat{\sigma}_1^2(\ell)$  and  $\widehat{\sigma}_0^2(\ell)$  are the estimated variances of the outcome variable for treated and untreated observations respectively in leaf  $\ell$ ,  $\widehat{p}_1(\ell)$  and  $\widehat{p}_0(\ell)$  are the estimated probabilities of treatment and control in leaf  $\ell$ .

- ▶  $(\widehat{\sigma}_1^2(\ell), \widehat{\sigma}_0^2(\ell))$  and  $(\widehat{p}_1(\ell), \widehat{p}_0(\ell))$  are constructed using the training sample.

## Transformed outcome trees

- ▶ Another method proposed in Athey and Imbens (2016) is the transformed outcome tree (TOT).

- ▶ Denote

$$Y_i^* = \frac{Y_i D_i}{p(X_i)} - \frac{Y_i (1 - D_i)}{1 - p(X_i)}.$$

- ▶ Note that  $Y_i D_i = Y_i(1) D_i$  and  $Y_i (1 - D_i) = Y_i(0) (1 - D_i)$ .
- ▶ Under unconfoundedness  $(Y_i(1), Y_i(0)) \perp\!\!\!\perp D_i \mid X_i$ ,

$$\begin{aligned} E[Y_i^* \mid X_i] &= E\left[\frac{Y_i(1) D_i}{p(X_i)} \mid X_i\right] - E\left[\frac{Y_i(0) (1 - D_i)}{1 - p(X_i)} \mid X_i\right] \\ &= \frac{E[Y_i(1) \mid X_i] E[D_i \mid X_i]}{p(X_i)} - \frac{E[Y_i(0) \mid X_i] E[1 - D_i \mid X_i]}{1 - p(X_i)} \\ &= E[Y_i(1) \mid X_i] - E[Y_i(0) \mid X_i] = E[Y_i(1) - Y_i(0) \mid X_i]. \end{aligned}$$

- ▶ Since  $E[Y_i^* \mid X_i = x] = \tau(x)$ , we can simply apply adaptive and honest regression tree algorithms to  $Y_i^*$ . In the case of observational data,  $p(X_i)$  has to be estimated. In the case of experimental data,  $p(X_i)$  is a known constant.



# Causal forests

- ▶ Causal trees suffer from the same instability as regression trees. Small changes in the data can result in a very different tree
- ▶ One can improve their performance by averaging over many independent causal trees. A forest estimator is an average of trees, where each tree is grown on a bootstrap sample.
- ▶ Wager and Athey (2018) shows that the forest estimator  $\widehat{\tau}^{\text{cf}}(x)$  is asymptotically normal:

$$\frac{\widehat{\tau}^{\text{cf}}(x) - \tau(x)}{\sqrt{\text{Var}[\widehat{\tau}^{\text{cf}}(x)]}} \rightarrow_d \text{N}(0, 1).$$

- ▶ The asymptotic normality result allows us to construct confidence interval for  $\tau(x)$  or CATE ( $A$ ). It relies on two modifications of the conventional random forest algorithm:
  - ▶ the honest approach to growing the tree and estimation;
  - ▶ subsampling: the sample size of the bootstrap subsample  $s_n$  should satisfy  $s_n/n \rightarrow 0$ .

- ▶ The causal forest approach overcomes two drawbacks of the honest causal tree:
  - ▶ When the honest causal tree approach is applied, half of the data is wasted, since the training sample  $\mathcal{S}^{\text{tr}}$  is not used in estimation;
  - ▶ the result you get depends on how  $\mathcal{S}$  is split into  $\mathcal{S}^{\text{est}}$  and  $\mathcal{S}^{\text{tr}}$ .
- ▶ Each bootstrap subsample is split into training and estimation subsamples. However, since subsamples are randomly re-drawn with each tree, the honest causal forest approach does not waste observations: each observation is used for partitioning in some trees and estimation in other trees.
- ▶ Causal forests can be used to compute importance measures of variables in  $X_i$ .