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 22, 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 $X = \left(X_1, X_2, ..., X_p\right)^\top$.
- 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 $\{(X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n)\}$, where $X_i = \left(X_{1,i}, X_{2,i}, ..., X_{p,i}\right)^\top$.
- $X_{j,i}$: the value of the j-th predictor, or input, for observation i, where $i = 1, 2, ..., n$ and $j = 1, 2, ..., p$.
- $|\Pi|$ denotes the number of elements of Π. A partition $\Pi = \left\{\ell_1, \ell_2, ..., \ell_{|\Pi|}\right\}$ 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 $\mathcal{S} = \{1, ..., n\}$ 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 ℓ to which x belongs.
- A partition estimator of $f(x) = \mathrm{E}[Y \mid X = x]$ using the partition Π is

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

Splitting rule

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

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

- The RT algorithm recursives solves

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

  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

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

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

- In the initial step, we solve

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

where we have

$$\sum_{i\in\mathcal{S}}\left(Y_i-\widehat{f}\left(X_i \mid \Pi_{j,c}\right)\right)^2$$

$$=\sum_{i\in\mathcal{S}}\left\{1\left(X_i\in\ell_{j,c}^{+}\right)\left(Y_i-\frac{1}{\left|\left\{i: X_i\in\ell_{j,c}^{+}\right\}\right|}\sum_{i:X_i\in\ell_{j,c}^{+}} Y_i\right)^2\right.$$

$$\left.+1\left(X_i\in\ell_{j,c}^{-}\right)\left(Y_i-\frac{1}{\left|\left\{i: X_i\in\ell_{j,c}^{-}\right\}\right|}\sum_{i:X_i\in\ell_{j,c}^{-}} Y_i\right)^2\right\}.$$

- The RT method has the distinct feature of implict 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

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

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

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

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

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

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

- And

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

  therefore,

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

- In each step, the RT algorithm solves

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

- 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:

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

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

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

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 $\mathbb{X} = \{L, R\}$, $\overline{Y}_R = |\{i : X_i = R\}|^{-1} \sum_{i:X_i=R} Y_i$ and $\overline{Y}_L = |\{i : X_i = L\}|^{-1} \sum_{i:X_i=L} Y_i$;
  - Consider the estimated split rule:

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

  - Then, $\widehat{f}\left(x \mid \widehat{\Pi}\right)$ is biased for $\mathrm{E}[Y \mid X = x]$, $x \in \mathbb{X}$.

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 $\mathcal{S}^{\text{est}}$, $\mathcal{S} = \mathcal{S}^{\text{est}} \cup \mathcal{S}^{\text{tr}}$ 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

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

  using the estimation sample and Π should be constructed using only the training sample.

Honest splitting

- Denote

$$f\left(x \mid \Pi\right) = \mathrm{E}\left[Y \mid X \in \ell_{\Pi}\left(x\right)\right] = \mathrm{E}\left[f\left(X\right) \mid X \in \ell_{\Pi}\left(x\right)\right].$$

- We have

$$\begin{aligned} Y_i &= \sum_{\ell \in \Pi} 1\left(X_i \in \ell\right) \alpha_{\ell} + \epsilon_i \\ \mathrm{E}\left[\epsilon_i \mid \left\{1\left(X_i \in \ell\right) : \ell \in \Pi\right\}\right] &= 0, \end{aligned}$$

  where

$$\alpha_{\ell} = \frac{\mathrm{E}\left[Y_i 1\left(X_i \in \ell\right)\right]}{\mathrm{E}\left[1\left(X_i \in \ell\right)\right]} = \mathrm{E}\left[Y_i \mid X_i \in \ell\right]$$

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

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

$$\begin{aligned} f\left(x \mid \Pi\right) &= \sum_{\ell \in \Pi} \alpha_{\ell} 1\left(x \in \ell\right) \\ \mathrm{E}\left[Y_i \mid \left\{1\left(X_i \in \ell\right) : \ell \in \Pi\right\}\right] &= f\left(X_i \mid \Pi\right). \end{aligned}$$

- And,

$$\mathrm{E}\left[\widehat{f}\left(x \mid \mathcal{S}^{\mathsf{est}}, \Pi\right)\right] = \mathrm{E}\left[\sum_{\ell \in \Pi} 1\left(x \in \ell\right) \widehat{Y}_{\ell}^{\mathsf{est}}\right]$$

$$= \sum_{\ell \in \Pi} 1\left(x \in \ell\right) \mathrm{E}\left[\frac{\sum_{i \in \mathcal{S}^{\mathsf{est}}} 1\left(X_i \in \ell\right) Y_i}{\sum_{i \in \mathcal{S}^{\mathsf{est}}} 1\left(X_i \in \ell\right)}\right] = \sum_{\ell \in \Pi} 1\left(x \in \ell\right) \alpha_{\ell}$$

$$= f\left(x \mid \Pi\right).$$

- 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 $\widehat{f}\left(X_0 \mid \mathcal{S}^{\mathsf{est}}, \Pi\right)$ to Y_0:

$$\mathrm{E}\left[\left(Y_0 - \widehat{f}\left(X_0 \mid \mathcal{S}^{\mathsf{est}}, \Pi\right)\right)^2\right]$$

$$= \mathrm{E}\left[\left(Y_0 - f\left(X_0 \mid \Pi\right)\right)^2\right] + \mathrm{E}\left[\left(\widehat{f}\left(X_0 \mid \mathcal{S}^{\mathsf{est}}, \Pi\right) - f\left(X_0 \mid \Pi\right)\right)^2\right],$$

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

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

- Then,

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

and by LIE

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

- Then,

$$\mathrm{E}\left[\left(Y_0 - f\left(X_0 \mid \Pi\right)\right)^2\right] = \mathrm{E}\left[Y_0^2\right] - \mathrm{E}\left[f\left(X_0 \mid \Pi\right)^2\right].$$

- The honest population-level criterion takes the form

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

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

Estimate the criterion

- We have

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

- By CLT,

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

and therefore,

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

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

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

where

$$\widehat{f}\left(x \mid \mathcal{S}^{\mathsf{tr}}, \Pi\right) = \frac{1}{\left|\left\{i \in \mathcal{S}^{\mathsf{tr}} : X_i \in \ell_\Pi\left(x\right)\right\}\right|} \sum_{i \in \mathcal{S}^{\mathsf{tr}} : X_i \in \ell_\Pi(x)} Y_i.$$

- An estimator for

$$\mathrm{E}\left[\left(\widehat{f}\left(X_0 \mid \mathcal{S}^{\mathsf{est}}, \Pi\right) - f\left(X_0 \mid \Pi\right)\right)^2\right] = \mathrm{E}\left[\mathrm{Var}\left[\widehat{f}\left(X_0 \mid \mathcal{S}^{\mathsf{est}}, \Pi\right)\right]\right]$$

is

$$\frac{1}{\left|\mathcal{S}^{\mathsf{est}}\right|} \sum_{\ell \in \Pi} \widehat{\sigma}^2\left(\ell\right).$$

- And, since $\mathrm{E}\left[\widehat{f}\left(x \mid \mathcal{S}^{\mathsf{tr}}, \Pi\right)\right] = f\left(x \mid \Pi\right)$,

$$\mathrm{Var}\left[\widehat{f}\left(x \mid \mathcal{S}^{\mathsf{tr}}, \Pi\right)\right] = \mathrm{E}\left[\widehat{f}\left(x \mid \mathcal{S}^{\mathsf{tr}}, \Pi\right)^2\right] - f\left(x \mid \Pi\right)^2.$$

- An estimator for $\mathrm{E}\left[f\left(X_0 \mid \Pi\right)^2\right]$ using the training sample is

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

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

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

- The honest RT algorithm recursively solves

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

- The conventional (adaptive) RT algorithm recursively solves

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

- 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; $X_{1,i}, ..., X_{p,i}$: other observed explanatory variables; ϵ_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.
- $\left(X_{1,i}, ..., X_{p,i}\right)$ 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 $Y_i = g\left(D_i, X_{1,i}, ..., X_{p,i}, \epsilon_i\right)$.
- g is unknown and in the treatment effect model, we do not assume g is linear.

- The outcome variable $Y_i(1) = g\left(1, X_{1,i}, ..., X_{p,i}, \epsilon_i\right)$ 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 $Y_i(0) = g\left(0, X_{1,i}, ..., X_{p,i}, \epsilon_i\right)$ 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)

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

and ATT (average treatment effect on the treated)

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

- 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 $D_i \perp\!\!\!\perp \left(Y_i\left(1\right), Y_i\left(0\right), X_{1,i}, ..., X_{p,i}\right)$. 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 $(Y_i\left(1\right) - Y_i\left(0\right))$ for individual i.

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

$$\mathrm{E}\left[f\left(V, W\right) \mid W = w\right] = \mathrm{E}\left[f\left(V, w\right)\right]. \tag{1}$$

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

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

- By LIE,

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

  where

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

  follows from (1).

- Similarly, we have

$$\mathrm{E}\left[Y_i \mid D_i = 0\right] = \frac{\mathrm{E}\left[Y_i\left(1 - D_i\right)\right]}{\mathrm{E}\left[1 - D_i\right]}.$$

- We can write

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

  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):

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

- 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 $X_{1,i}, ..., X_{p,i}$ 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

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

  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

$$\begin{aligned} Y_i(0) &= \mu_0 + U_i(0) \\ Y_i(1) &= \mu_1 + U_i(1), \end{aligned}$$

  where μ_0 and μ_1 are constants common across individuals and assumed to be nonstochastic and $(U_i(0), U_i(1))$ are stochastic components.

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

$$\mathrm{E}\left[Y_i\left(1\right) - Y_i\left(0\right) \mid X_i\right] = \mu_1 - \mu_0,$$

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

$$\mathrm{ATE} = \mathrm{E}\left[Y_i\left(1\right) - Y_i\left(0\right)\right] = \mu_1 - \mu_0.$$

- We assume $\mathrm{E}\left[Y_i\left(1\right) \mid D_i, X_i\right] = \mathrm{E}\left[Y_i\left(1\right) \mid X_i\right]$ and $\mathrm{E}\left[Y_i\left(0\right) \mid D_i, X_i\right] = \mathrm{E}\left[Y_i\left(0\right) \mid X_i\right]$, 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 indivdiuals with specific demographic status X_i, $Y_i\left(1\right)$ and $Y_i\left(0\right)$ are both mean independent of D_i.

- Let us write

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

  where the last equality follows from the conditional mean independence assumption.
- By the assumption $\mathrm{E}\left[U_i\left(0\right) \mid X_i\right] = \mathrm{E}\left[U_i\left(1\right) \mid X_i\right]$, we have

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

  where we denote $h\left(X_{1,i}, ..., X_{p,i}\right) = \mathrm{E}\left[U_i\left(0\right) \mid X_i\right]$.

- Therefore, we have

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

- Define

$$V_i = Y_i - \mathrm{E}\left[Y_i \mid D_i, X_i\right]$$

and now we have the following regression model:

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

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

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

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 $X_{1,i}, ..., X_{p,i}$ consistently estimates ATE $= (\mu_1 - \mu_0)$.

- We assume $\mathrm{E}\left[U_i\left(0\right) \mid X_i\right] = \mathrm{E}\left[U_i\left(1\right) \mid X_i\right]$, which implies

$$\mathrm{E}\left[Y_i\left(1\right) - Y_i\left(0\right) \mid X_i\right] = \mu_1 - \mu_0.$$

- 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

$$\begin{aligned} Y_i &= g\left(D_i, X_{1,i}, ..., X_{p,i}, \epsilon_i\right) \\ D_i &= m\left(X_{1,i}, ..., X_{p,i}, \eta_i\right) \end{aligned}$$

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

More on conditional independence

- When V and W are conditionally independent given X, one can easily see that for any function φ,

$$\mathrm{E}\left[\varphi\left(V\right) \mid W, X\right] = \mathrm{E}\left[\varphi\left(V\right) \mid X\right].$$

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

- We notice that

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

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

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

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

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

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

$$\mathrm{E}[\eta(V, W) \mid X, W = w] = \mathrm{E}[\eta(V, w) \mid X]. \tag{2}$$

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 $X_{1,i}, ..., X_{p,i}$ to be correlated with D_i,
- The unconfounded condition allows D_i to be affected by $X_{1,i}, ..., X_{p,i}$, while the randomized experiment assumption does not.

Identification of ATE

- By LIE, we have

$$\begin{aligned} \text{ATE} &= \mathrm{E}\left[Y_i\left(1\right) - Y_i\left(0\right)\right] \\ &= \mathrm{E}\left[\mathrm{E}\left[Y_i\left(1\right) \mid X_i\right]\right] - \mathrm{E}\left[\mathrm{E}\left[Y_i\left(0\right) \mid X_i\right]\right], \end{aligned} \tag{3}$$

and

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

- 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

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

  and

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

- Therefore, we have

$$\mathrm{E}[Y_i D_i \mid X_i] = \Pr[D_i = 1 \mid X_i] \mathrm{E}[Y_i(1) \mid X_i] \qquad (4)$$

  and similarly,

$$\mathrm{E}[Y_i(1 - D_i) \mid X_i] = \Pr[D_i = 0 \mid X_i] \mathrm{E}[Y_i(0) \mid X_i]. \qquad (5)$$

- Now (3), (4), (5) and LIE imply

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

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

- Denote

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

  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:

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

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

- The estimated propensity score is

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

- 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, $p\left(x\right) = \mathrm{E}\left[D_i \mid X_i = x\right]$ 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
$$d_i(x_0) = \|X_i - x_0\| = \sqrt{(X_i - x_0)^\top (X_i - x_0)}$$
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
$$d_{(1)}(x_0) \leq d_{(2)}(x_0) \leq \cdots \leq d_{(n)}(x_0).$$

- Let $N_k\left(x_0\right)$ denote the identities of the k-nearest neighbors of x_0:

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

- The k-NN nonparametric estimator of $p\left(x_0\right)$ is

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

- 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 $\hat{p}_{kNN}\left(X_i\right)$ is consistent and asymptotically normal. It does not require a parametric model for the propensity score.