Topics in Econometrics

Estimating the Treatment Effect under Unconfoundedness

Instructor: Ma, Jun

Renmin University of China

May 26, 2022

Causal inference

- Stuart Mill (1843): *If a person eats a particular dish, and dies in consequence, that is, would not have died if he had not eaten of it, people would be apt to say that eating of that dish was the source of his death*.
- Ronald Aylmer Fisher (1918): *If we say, " this boy has grown tall because he has been well fed,"... we are suggesting that he might probably have been worse fed, and that in this case he would have been shorter*.

Examples

- Suppose that a selected set of individuals receive training or education initiated by the government with a view to enhancing their employment prospects. The government has collected the earnings data for the individuals who received the training and for the individuals who did not. We aim to quantify and estimate the effect of the training program.
- 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.

Causality

- Causality is tied to any action (or manipulation, treatment, or intervention) applied to a unit.
- The unit might be a village or city: What is the effect of a measles vaccination campaign implemented at the village level on village-level incidence of measles?
- The unit might be a firm: What is the effect of receipt of management consulting on firm productivity?
- 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.
- The simplest scenario is that we consider two actions. Often one of these actions corresponds to a more active treatment in contrast to a more passive action. In such cases we sometimes refer to the first action as the treatment as opposed to the control.

Fundamental problem of causal inference

"Fundamental Problem of Casual Inference"' (Holland, 1986)

- Among those who received treatment, we observe what happened to them with treatment but we do not observe what would have happened to them without treatment.
- Among those who did not receive treatment, we observe what happened to them without treatment but we do not observe what would have happened to them with treatment.
- Cannot observe same person simultaneously treated and not treated.

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 built on the potential outcome framework of Fisher (1935) and Rubin (1974).
  - The outcome variable $Y_i(1)$ 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 $Y_i(0)$ 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). 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.

- Potential outcome can be built on a structural model.
- 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.
- $X_i = (X_{1,i}, ..., X_{p,i})^\top$ 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 $Y_i = g(D_i, X_i, \epsilon_i)$.
- g is unknown and in the treatment effect model, we do not assume g is linear.

- The outcome variable $Y_i(1) = g(1, X_i, \epsilon_i)$ 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 $Y_i(0) = g(0, X_i, \epsilon_i)$ 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)
$$\mathrm{ATE} = \mathrm{E}\left[Y_i(1) - Y_i(0)\right],$$
and ATT (average treatment effect on the treated)
$$\mathrm{ATT} = \mathrm{E}\left[Y_i(1) - Y_i(0) \mid D_i = 1\right].$$
- 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?

Selection bias

- $Y_i(1) - Y_i(0)$ may be correlated with D_i, when D_i is determined by individuals. $Y_i(1) - Y_i(0)$ and D_i may be determined by the same factors.
- E.g., sickest individuals are the ones who take the medicine.
- High ability students are the ones who attend college.
- The selection bias can be thought of as $(\mathrm{E}[Y_i \mid D_i = 1] - \mathrm{E}[Y_i \mid D_i = 0]) - \mathrm{ATE}$, which is equal to $\mathrm{E}[Y_i(0) \mid D_i = 1] - \mathrm{E}[Y_i(0) \mid D_i = 0]$, when $Y_i(1) - Y_i(0)$ is constant.
- Solutions: 1. Covariates; 2. Randomized experiments; 3. Instrumental 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 $D_i \perp\!\!\!\perp (Y_i(1), Y_i(0), X_i)$. 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,

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

- 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

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

- 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

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

  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):

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

- 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 $Y_i\left(1\right) - Y_i\left(0\right) = \gamma_1$ is constant. This is very unrealistic. We investigate alternative model assumptions.
- We first consider the following model assumption

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

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

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

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

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

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

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

- Let us write

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

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

- By the assumption $\mathrm{E}\left[U_i(0) \mid X_i\right] = \mathrm{E}\left[U_i(1) \mid X_i\right]$, we have

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

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

- Therefore, we have

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

- Define

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

and now we have the following regression model:

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

- 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

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

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

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

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

- 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.

Homogeneous and heterogeneous treatment effects

- The individual treatment effect $\Delta_i = Y_i(1) - Y_i(0)$ is random, in general.
- Four cases (Heckman, Vytlacil and Urzua, 2006): A. Homogeneous; B. Homogeneous conditional on X; C1. Heterogeneous without essential heterogeneity; C2. Heterogeneous with essential heterogeneity.
  - A: Δ_i is constant. All individuals have the same treatment effect. In the structural model framework, this is equivalent to the restriction that $Y_i = g(X_i) + \Delta \cdot D_i + \epsilon_i$.
  - B: $\Delta_i = \Delta(X_i)$, a function of X_i only. In the structural model framework, this is equivalent to $Y_i = g(X_i, D_i) + \epsilon_i$.
  - C1: Δ_i is random, conditional on X_i but $\Delta_i \perp\!\!\!\perp D_i \mid X_i$. In the structural model framework, this holds when $\epsilon_i \perp\!\!\!\perp (D_i, X_i)$. This assumption is known as the unconfoundedness condition. 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.
  - C2: Δ_i is correlated with D_i conditional on X_i. Individuals select into treatment based on Δ_i. In this case, an instrument is needed.

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

$$\begin{aligned} Y_i &= g\left(D_i, X_i, \epsilon_i\right) \\ D_i &= m\left(X_i, \eta_i\right) \end{aligned}$$

and $\epsilon_i \perp\!\!\!\perp \eta_i \mid X_i$.

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_i to be correlated with D_i,
- The unconfounded condition allows D_i to be affected by X_i, while the randomized experiment assumption does not.

Identification of ATE

- By LIE, we have

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

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

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

and

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

- Therefore, we have

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

and similarly,

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

- 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 $x \in \mathbb{R}^p$.

Inverse probability weighting (IPW) estimator

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

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

  This is known as the IPW estimator (Hirano, Imbens and Ridder, 2003).
- It is straightforward to construct $\widehat{p}(x)$ if X_i is discrete:

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

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

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

  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 $\left(\widehat{\beta}_0, ..., \widehat{\beta}_p\right)$). We bootstrap the standard errors.

- The estimated propensity score is

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

- 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 $N_k(x_0)$ denote the identities of the k-nearest neighbors of x_0:
$$N_k(x_0) = \left\{ i : d_i(x_0) \leq d_{(k)}(x_0) \right\}.$$
- The k-NN nonparametric estimator of $p(x_0)$ is
$$\widehat{p}_{kNN}(x_0) = \frac{1}{k} \sum_{i \in N_k(x_0)} D_i.$$
- 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 $\widehat{p}_{kNN}(X_i)$ is consistent and asymptotically normal. It does not require a parametric model for the propensity score.

Overlap (common support) condition

- The overlap condition: for any value x in the support of X_i,

$$0 < \Pr\left[D_i = 1 \mid X_i = x\right] < 1.$$

- Rigorously, the overlap condition was assumed when establishing identification of the parameter of interest.
- This implies that the conditional supports of X_i given $D_i = 1$ and $D_i = 0$ should overlap.
- The overlap condition can be a problem in empirical researches.E.g., if suppose that all the high income individuals are concentrated on the treated group and all the low income individuals are concentrated on the control group, then the overlap condition is violated.
- The reason this can be a problem in practice is because we typically have many variables in X_i and the overlap condition is violated if there exists a particular variable such that the subjects are divided completely into two different groups.
- Note the potential tension between overlap and unconfoundedness.

- For "regular" estimation, we need a stronger overlap condition: for some small $\epsilon > 0$, for any value x in the support of X_i,

$$\epsilon < \Pr\left[D_i = 1 \mid X_i = x\right] < 1 - \epsilon.$$

Regression-based matching

- First step: parametric/nonparametric regression of Y on (D, X) to recover $\widehat{\mathrm{E}}\left[Y \mid D, X\right]$ (estimator of $\mathrm{E}\left[Y \mid D, X\right]$).
- Second step: Estimate ATT by

$$\frac{1}{N_1} \sum_{i=1}^{n} D_i \left(Y_i - \widehat{\mathrm{E}}\left[Y_i \mid X_i, D_i = 0\right] \right)$$

  where $N_1 = \sum_{i=1}^{n} 1\left(D_i = 1\right)$.
- See for Heckman, Ichimura and Todd (1997, 1998) for estimation theory.

Nearest-Neighbor matching

- Choose a positive integer M (often $M = 1$).
- Let $d(\cdot, \cdot)$ be a distance measure.
- For each i let $j_m(i)$ be an index satisfying $D_{j_m(i)} = 1 - D_i$ and

$$\sum_{l:D_l=1-D_i} 1\left\{d\left(X_l, X_i\right) \leq d\left(X_{j_m(i)}, X_i\right)\right\} = m.$$

- Let $\mathcal{J}_M(i)$ denote the indices of the M closest matches:

$$\mathcal{J}_M(i) = \left\{j_1(i), ..., j_M(i)\right\}.$$

- Then

$$\widehat{Y}_{0i} = \begin{cases} Y_i & \text{if } D_i = 0 \\ \frac{1}{M} \sum_{j \in \mathcal{J}_M(i)} Y_j & \text{if } D_i = 1 \end{cases}$$

$$\widehat{Y}_{1i} = \begin{cases} \frac{1}{M} \sum_{j \in \mathcal{J}_M(i)} Y_j & \text{if } D_i = 0 \\ Y_i & \text{if } D_i = 1 \end{cases}$$

and the Nearest-Neighbor matching estimator for ATE is

$$\frac{1}{n} \sum_{i=1}^{n} \left(\widehat{Y}_{1i} - \widehat{Y}_{0i} \right).$$

Issues with Nearest-Neighbor matching

- Advantages: 1. Very simple to implement; 2. Only need to pick number of nearest neighbor (usually $M = 1$).
- Estimation theory developed by Abadie and Imbens (2006, 2008).
- Disadvantages: 1. Not an efficient estimator; 2. When number of continuous covariates ≥ 2, the estimator is generally not $\sqrt{n}$-consistent: bias term does not vanish at $n^{-1/2}$ rate. 3. Bootstrap is inconsistent.
- Remedies: 1. Bias correction (Abadie and Imbens, 2011); 2. Weighted bootstrap (Otsu and Rai, 2016).

Propensity score matching

- Propensity score: $\pi(x) = \Pr[D = 1 \mid X = x]$.
- Rosenbaum and Rubin (1983) showed the following properties: 1. (balancing) $X \perp D \perp\!\!\!\perp \pi(X)$; 2. (unconfoundedness) $(Y_0, Y_1) \perp\!\!\!\perp D \mid \pi(X)$.
- Balancing: Conditional on the propensity score, the covariate distributions are balanced between the treatment and the control groups. Hence, they become comparable.
- Unconfoundedness: Instead of conditioning on the entire covariate vector X, conditioning solely on $\pi(X)$ suffices for removing the selection biases.
- Propensity score can be used for dimension reduction. Matching (regression or NN) on the propensity score.

What if overlap condition fails?

- Heckman, Ichimura and Todd (1998) propose to trim to common support.
- For a small $q > 0$, let $S = \left\{x : f_{X|D=1}(x) > q, f_{X|D=0}(x) > q\right\}$ and $\widehat{S} = \left\{x : \widehat{f}_{X|D=1}(x) > q, \widehat{f}_{X|D=0}(x) > q\right\}$.
- Estimation of ATT conditional on $X \in S$:

$$\frac{1}{\sum_{i=1}^{n} D_i 1\left(X \in \widehat{S}\right)} \sum_{i=1}^{n} D_i \left(Y_i - \widehat{\mathrm{E}}\left[Y_i \mid X_i, D_i = 0\right]\right) 1\left(X \in \widehat{S}\right).$$

- The estimator is consistent for $\mathrm{E}\left[Y_1 - Y_0 \mid D = 1, X \in S\right]$. Trimming changes definition of the parameter.

Limited overlap

- Strong overlap: for some small $\epsilon > 0$, for any value x in the support of X_i,

$$\epsilon < \Pr\left[D_i = 1 \mid X_i = x\right] < 1 - \epsilon.$$

- Limited overlap: ϵ is very small. In this case, it is found in simulations that performances of standard estimators (inverse probability weighting, matching...) are bad.

- Solution 1: trimming (i.e., dropping observations with estimated propensity scores that are close to 0 or 1). See Crump, Hotz, Imbens and Mitnik (2009).
- Solution 2: Rothe (2014) modifies the standard IPW-based confidence interval so that the modified confidence interval has correct coverage probability for fixed n if the outcomes are conditionally normally distributed, without assuming strong overlap.
- Solution 3: Ma and Wang (2021) does not assume strong overlap and applies asymptotic trimming: drop observations with estimated propensity scores that are in $(0, b_n) \cup (1 - b_n, 1)$ with $b_n \downarrow 0$. In this case, the asymptotic distributions are non-Gaussian. They work out optimal choice of b_n and show validity of "m-out-of-n" bootstrap inference.

More issues

- How do they perform in finite samples? Busso, DiNardo and McCrary (2014) compared IPW (Hirano, Imbens and Ridder, 2003) and propensity score matching. They find "matching may be more effective when overlap is sufficiently poor".
- It is known that IPW is efficient (i.e., smallest possible asymptotic variance, see Hahn, 1998 and Hirano, Imbens and Ridder, 2003). See Chan, Yam and Zhang (2016) for a new efficient estimator called empirical balancing estimator, which does not require estimating the propensity score.

Doubly robust estimation

- The two identification results for the ATE

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

- Two different approaches to estimation: regression adjustment (RA), estimate $\mathrm{E}\left[Y_i \mid X_i = x, D_i = d\right]$; inverse probability weighting (IPW), estimate $\mathrm{P}\left[D_i = 1 \mid X_i = x\right]$.
- We focus on parametric estimation now. We specify parametric models:

$$\mathrm{E}\left[Y_i \mid X_i = x, D_i = d\right] = m\left(x, d \mid \alpha\right),$$

$$\mathrm{P}\left[D_i = 1 \mid X_i = x\right] = \pi\left(x \mid \beta\right).$$

- Model misspecification leads to inconsistency. A doubly robust estimator is consistent when at least one of the models is correctly specified.

Augmented IPW

- The augmented inverse probability weighting estimator (AIPW) is the most popular doubly robust estimator.
- The AIPW estimator of $\mu_1 = \mathrm{E}\left[Y_{i1}\right]$ is based on the following moment condition:

$$\mu_1 = \mathrm{E}\left[\frac{DY}{\pi\left(X \mid \beta\right)} - \frac{D - \pi\left(X \mid \beta\right)}{\pi\left(X \mid \beta\right)} m\left(X, 1 \mid \alpha\right)\right].$$

- The ATE can be identified by

$$\begin{aligned} &\mathrm{E}\left[\frac{DY}{\pi\left(X \mid \beta\right)} - \frac{D - \pi\left(X \mid \beta\right)}{\pi\left(X \mid \beta\right)} m\left(X, 1 \mid \alpha\right)\right] \\ -&\mathrm{E}\left[\frac{\left(1 - D\right) Y}{1 - \pi\left(X \mid \beta\right)} - \frac{D - \pi\left(X \mid \beta\right)}{1 - \pi\left(X \mid \beta\right)} m\left(X, 0 \mid \alpha\right)\right]. \end{aligned}$$

- Note that the identification of the ATE requires one of the models is correct.

- The AIPW estimator is obtained by the following multiple step procedure:
  1. Estimate α, e.g., by OLS;
  2. Estimate β, e.g., by MLE;
  3. Estimate the sample analog of the doubly robust moment condition:

$$\frac{1}{n}\sum_{i=1}^{n}\left\{\frac{D_i Y_i}{\pi\left(X_i \mid \widehat{\beta}\right)} - \frac{D_i - \pi\left(X_i \mid \widehat{\beta}\right)}{\pi\left(X_i \mid \widehat{\beta}\right)} m\left(X_i, 1 \mid \widehat{\alpha}\right)\right\}$$
$$-\frac{1}{n}\sum_{i=1}^{n}\left\{\frac{\left(1 - D_i\right) Y_i}{1 - \pi\left(X_i \mid \widehat{\beta}\right)} - \frac{D_i - \pi\left(X_i \mid \widehat{\beta}\right)}{1 - \pi\left(X_i \mid \widehat{\beta}\right)} m\left(X_i, 0 \mid \widehat{\alpha}\right)\right\}.$$

Distributional treatment parameters

- Parameters of interest may depend on the joint distribution of $(Y(0), Y(1))$.
- $\Pr[Y(1) > Y(0)]$: probability of benefiting from treatment; $\Pr[Y(1) > c \mid Y(0) < c]$: probability of being above the poverty line with treatment among those who would be below the poverty line without treatment.
- $\mathrm{Median}[Y(1) - Y(0)] \neq \mathrm{Median}[Y(1)] - \mathrm{Median}[Y(0)]$, in comparison with $\mathrm{ATE} = \mathrm{E}[Y(1)] - \mathrm{E}[Y(0)]$, which depends on only marginal distributions.
- It is possible to have $\mathrm{Median}[Y(1) - Y(0)] < 0$ (most people are hurt by the treatment) but $\mathrm{E}[Y(1) - Y(0)] > 0$, if a small fraction of people benefit a lot from the treatment.

- Even with a randomized experiment, one cannot identify distributional treatment parameters without more assumptions/structure.
- Random experiments allow one to identify distribution of $Y(0)$ from those randomly denied treatment, and to identify the distribution of $Y(1)$ from those randomly assigned to treatment, and thus to identify parameters that depend only on the marginal distributions.
- One never observes $Y(0)$ and $Y(1)$ for the same individual, and thus randomized experiments do not identify the dependence between $Y(0)$ and $Y(1)$.

Deconvolution

- The earliest method is to impose the assumption that $Y(1) - Y(0)$ is independent from $Y(0)$. See Heckman, Smith with Clements (1997) for details.
- But this assumption is hard to justify and there are certain issues with the estimation method.