Lectures 14 & 15

Heterogeneous Treatment Effects

Oscar Volpe

11/15/2021 & 11/17/2021

Notation

Suppose you want to understand the impact of a policy intervention D on some outcome of interest Y. For simplicity, assume D is either 0 and 1.

- D is our *treatment* ($D = 1$ if *treated*, $D = 0$ if *untreated*).
- Y is our *outcome*, which is assumed to depend somehow on D.

We refer to Y_d as the *potential outcome* in the event that $D = d$. Write:

$$Y = \begin{cases} Y_0 & \text{if } D = 0 \\ Y_1 & \text{if } D = 1 \end{cases}$$

For any individual, only one potential outcome is observed. For example, we cannot see Y_0 for someone who was treated, i.e. for whom $D = 1$.

- We seek to draw inference about the *counterfactuals* (Y_d for $d \neq D$).
- To understand the causal effect of treatment, we need to ask about what might have happened in the scenarios that did not occur.

Example: Free Preschool

Suppose the *treatment* D is a free, high-quality preschool education. Let the *outcome* Y be earnings in adulthood. You collect data $\{D_i, Y_i\}_{i=1}^n$.

- *Note:* we observe $Y_i = Y_{D_i,i}$, but we do not observe $Y_{d,i}$ for $d \neq D_i$.
- We generally assume that people will self-select into treatment.

Inutitively, families decide whether to enroll their children in preschool for specific reasons. In this case, the treatment D is not chosen randomly.

- *Selection Bias:* treatment D could depend on Y_d.
  - Maybe children from under-resourced families, i.e. for whom Y_d is already lower, are more likely to be enrolled in the preschool program.
- *Selection on the Gains:* treatment D could depend on $Y_1 - Y_0$.
  - Maybe children with better outside options, i.e. for whom $Y_1 - Y_0$ is lower, would be less likely to be enrolled in the preschool program.

ATE, *ATT*, and *ATU*

For individual i, the *treatment effect* is $Y_{1,i} - Y_{0,i}$. We never observe this. The *average treatment effect* (ATE) is the expectation of $Y_{1,i} - Y_{0,i}$.

$$ATE = E(Y_1 - Y_0)$$

Intuitively, ATE is the average causal effect of the intervention D on everyone in the population. This differs from conditional treatment effects.

- *Avg. Treatment Effect on the Treated*: $ATT = E(Y_1 - Y_0 | D = 1)$.
- *Avg. Treatment Effect on the Untreated*: $ATU = E(Y_1 - Y_0 | D = 0)$.

Which of these effects is of greatest interest to the researcher?

- Policy relevance depends on the context we are considering.
- Perhaps we care most about how a program affects certain subgroups.

Homogenous Treatment Effects

By the Law of Iterated Expectations, we know that:

$$ATE = P(D = 1) \times ATT + P(D = 0) \times ATU,$$

where $P(D = 1) + P(D = 0) = 1$. That is, the ATE is a weighted average of the ATT and the ATU. If $Y_1 - Y_0$ equals some constant c, then:

$$\begin{aligned} ATE &= E(Y_1 - Y_0) = c \\ ATT &= E(Y_1 - Y_0|D = 1) = c \\ ATU &= E(Y_1 - Y_0|D = 0) = c \end{aligned}$$

So, if everybody has the same treatment effect, then $ATE = ATT = ATU$.

- In this lecture, we assume *heterogeneous treatment effects*, i.e. different individuals are affected differently by an intervention.
- Hence, $Y_{1,i} - Y_{0,i}$ need not equal $Y_{1,j} - Y_{0,j}$ for two people i and j.

Modeling Causal Effects

Let Y be an outcome, let D be some treatment, and let U represent all unobserved determinants of Y. A causal model for the outcome is:

$$Y = g(D, U)$$

Under this model, the treatment effect equals $g(1, U) - g(0, U)$.

Suppose that the causal model g is linear in D and U, i.e. that:

$$Y = \alpha + \beta D + U$$

As $Y = Y_0 + (Y_1 - Y_0)D$, we can re-arrange this model so that:

$$Y = \underbrace{E(Y_0)}_{\alpha} + \underbrace{(Y_1 - Y_0)}_{\beta} D + \underbrace{Y_0 - E(Y_0)}_{U}$$

Defining terms this way, we ensure that U has mean zero, i.e. $E(U) = 0$.

Random Coefficient Model

Under heterogeneous treatment effects, $Y_{1,i} - Y_{0,i}$ varies across i.

$$\begin{aligned} Y_{0,i} &= \alpha + U_i \\ Y_{0,i} &= \alpha + \beta_i + U_i \end{aligned} \quad \implies \beta_i = Y_{1,i} - Y_{0,i}$$

So, β_i is the treatment effect for individual i. We assume β_i is random.

$$Y_i = \underbrace{E(Y_0)}_{\alpha} + \underbrace{(Y_{1,i} - Y_{0,i})}_{\beta_i} D_i + \underbrace{Y_{0,i} - E(Y_0)}_{U_i}$$

This model is often referred to as the *random coefficient model*.

- $ATE = E(Y_{1,i} - Y_{0,i}) = E(\beta_i)$
- $ATT = E(Y_{1,i} - Y_{0,i} | D_i = 1) = E(\beta_i | D_i = 1)$
- $ATU = E(Y_{1,i} - Y_{0,i} | D_i = 0) = E(\beta_i | D_i = 0)$

Selection Bias

Try comparing the average outcomes between treated/untreated groups:

$$E(Y_i|D_i=1)-E(Y_i|D_i=0) = \underbrace{E(Y_{1,i}-Y_{0,i}|D_i=1)}_{ATT} + \underbrace{E(Y_{0,i}|D_i=1)-E(Y_{0,i}|D_i=0)}_{\text{Selection Bias}}$$
$$= \underbrace{E(Y_{1,i}-Y_{0,i}|D_i=0)}_{ATU} + \underbrace{E(Y_{1,i}|D_i=1)-E(Y_{1,i}|D_i=0)}_{\text{Selection Bias}}$$

This difference does not equal the ATT or ATU when Y_d depends on D.

- If $Y_0 \perp D$, we get the ATT by comparing treated/untreated groups
- If $Y_1 \perp D$, we get the ATU by comparing treated/untreated groups

When $Y_0, Y_1 \perp D$, this difference gives us $ATE = ATT = ATU$. Otherwise, if individuals "select into treatment" based on Y_d, there is selection bias.

- *Note:* selection bias means that D is endogenous in $Y = \alpha + \beta D + U$.
- An OLS regression of Y on D will not estimate the treatment effect.

Example: Homogeneous Treatment Effects

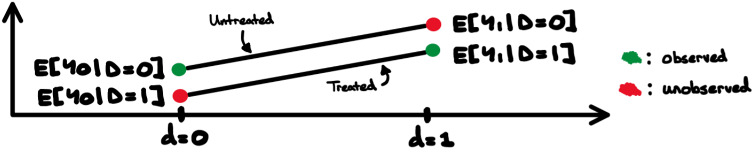
Suppose treatment D is free preschool, and let Y denote future earnings.

- All children receive the same benefits from preschool: $Y_{1,i} - Y_{0,i} = c$.
- There is selection bias, whereby under-resourced children are more likely to enroll in the program: $E(Y_{i,0}|D_i = 1) < E(Y_{i,0}|D_i = 0)$.

Even with constant treatment effects, we cannot recover the value of c.

$$E(Y_{i,1}|D_i = 1) - E(Y_{i,0}|D_i = 0) < E(Y_{i,1} - Y_{i,0})$$

Comparing outcomes between treated/untreated groups underestimates c.



An Experimental Setting

Under random assignment, D is independent of all potential outcomes.

Definition (Random Assignment)

Let Y be an outcome, and let D be some treatment. Let Y_d denote the potential outcome associated with the state $D = d$. The treatment is *randomly assigned* if and only if $Y_d \perp D$, for every d in the support of D.

If D is randomly administered at random, then selection bias disappears.

$$\begin{aligned} E(Y_i|D_i = 1) - E(Y_i|D_i = 0) &= E(Y_{1,i}|D_i = 1) - E(Y_{0,i}|D_i = 0) \\ &= \underbrace{E(Y_{1,i}) - E(Y_{0,i})}_{ATE}, \text{ since } Y_{0,i}, Y_{1,i} \perp D_i \end{aligned}$$

- random assign. $\implies ATE = E(Y_{1,i} - Y_{0,i}) = E(Y_{1,i} - Y_{0,i}|D_i = 1) = ATT$.
- random assign. $\implies ATE = E(Y_{1,i} - Y_{0,i}) = E(Y_{1,i} - Y_{0,i}|D_i = 0) = ATU$.

OLS Estimation with Random Assignment

Consider a simple linear regression model $Y_i = \beta_0 + \beta_1 D_i + U_i$. Under a best linear predictor (BLP) interpretation, we derive β_1 to be:

$$\beta_1 = \frac{\text{Cov}(Y_i, D_i)}{\text{Var}(D_i)} = E(Y_{i1}|D_i = 1) - E(Y_{i0}|D_i = 0)$$

- Since D_i is binary, we know that $E(Y_i|D_i) = \text{BLP}(Y_i|D_i) = \beta_0 + \beta_1 D_i$.
- If $Y_{0,i}, Y_{1,i} \perp D_i$, then β_1 equals $ATE = ATT = ATU$.

Given an *i.i.d.* sample $\{Y_i, D_i\}_{i=1}^n$, write down the OLS estimator for β_1.

$$\hat{\beta}_1 = \frac{\frac{1}{n}\sum_{i=1}^n (D_i - \bar{D}_n)(Y_i - \bar{Y}_n)}{\frac{1}{n}\sum_{i=1}^n (D_i - \bar{D}_n)^2}$$

If random assignment holds, then $\hat{\beta}_1$ is both a consistent and unbiased estimator for the ATE, which is also equal to the ATT and the ATU.

Confounding Variables

Recovering average treatment effects in an experiment is straightforward.

- In practice, it is often too costly/infeasible to run an experiment.
- Agents sort into/out of treatment based on observed and unobserved determinants of Y. We refer to these factors *confounding variables*.

Definition (Confounding Variable)

Let Y be an outcome, and let D be some treatment. We say X is a *confounding variable* for the causal effect of D on Y if $X \not\perp Y$, $X \not\perp D$.

When there are confounding variables, treatment is not randomly assigned.

- Suppose we observe all confounding variables and control for them.
- In this case, we can still estimate the average treatment effects.

Selection on Observable Features

Definition (Selection on Observables)

Let Y be an outcome, and let D be some treatment. Let X be a random vector of observable features, and let Y_d denote the potential outcome associated with state $D = d$. *Selection on observables* is the assumption that $Y_d \perp D | X = x$ for all x in the support of X, d in the support of D.

Intuitively, conditional on X, treatment D is as-good-as randomly assigned.

$$E(Y_i | D_i = 1, X_i = x) - E(Y_i | D_i = 0, X_i = x) = \underbrace{E(Y_{1,i} | X_i = x) - E(Y_{0,i} | X_i = x)}_{ATE(x)}$$

- $ATE(x) = E(Y_{1,i} - Y_{0,i} | X_i = x) = E(Y_{1,i} - Y_{0,i} | D_i = 1, X_i = x) = ATT(x)$
- $ATE(x) = E(Y_{1,i} - Y_{0,i} | X_i = x) = E(Y_{1,i} - Y_{0,i} | D_i = 0, X_i = x) = ATU(x)$

Conditioning on $X_i = x$, we can recover the ATE, ATT, and ATU.

OLS Estimation with Controls

Consider a linear regression model $Y_i = \beta_0 + \beta_1 D_i + X_i'\gamma + U_i$. Under a best linear predictor (BLP) interpretation, we know that β_1 satisfies:

$$\beta_1 \approx E(Y_i|D_i = 1, X_i = x) - E(Y_i|D_i = 0, X_i = x)$$

If $Y_{i,0}, Y_{i,1} \perp D_i|X_i$, then β_1 will approximate the conditional average treatment effect, i.e. $ATE(x)$. We can estimate it with the OLS estimator.

- Run an OLS regression of Y_i on D_i and X_i. Then $\hat{\beta}_1$ estimates ATE as a weighted average of the $ATE(x) = E(Y_{1,i} - Y_{0,i}|X_i = x)$ across values of x.

Another common approach is to "match" on different values of X.

- *Intuition:* we have an experiment for a given value of $X = x$.
  (1) Fix some $X = x$.
  (2) Estimate $ATE(x) = E(Y_i|D_i = 1, X_i = x) - E(Y_i|D_i = 0, X_i = x)$.
  (3) Take ATE as the weighted average of $ATE(x)$ across values of $X = x$.
- This approach can suffer from the "curse of dimensionality".

Example: A Matching Estimator for the *ATE*

Suppose treatment D is free preschool and Y represents future earnings. Let X indicate whether family income is below the federal poverty line.

- Assume low-income children are more likely to enroll in the program, but that selection bias depends only on X, so that $Y_0, Y_1 \perp D|X$.
- Hence, we are assuming there is *selection on observables*.

To estimate the *ATE* from a sample $\{Y_i, D_i, X_i\}_{i=1}^n$, compute:

$$\widehat{\text{ATE}} = \left(\bar{Y}_n^{(D_i=1,X_i=1)} - \bar{Y}_n^{(D_i=0,X_i=1)}\right)\bar{X}_n + \left(\bar{Y}_n^{(D_i=1,X_i=0)} - \bar{Y}_n^{(D_i=0,X_i=0)}\right)\left(1 - \bar{X}_n\right)$$

To show $\widehat{\text{ATE}}$ is a consistent estimator for *ATE*, note that:

$$\begin{aligned} \widehat{\text{ATE}} &\overset{p}{\to} \sum_{x\in\{0,1\}} \left[E(Y_i|D_i = 1, X_i = x) - E(Y_i|D_i = 0, X_i = x)\right] \times P(X_i = x) \\ &= \sum_{x\in\{0,1\}} ATE(x) \times P(X_i = x) = ATE \end{aligned}$$

Notation

Given outcome Y_i and treatment D_i, the random coefficients model is:

$$Y_i = \underbrace{E(Y_0)}_{\alpha} + \underbrace{(Y_{1,i} - Y_{0,i})}_{\beta_i} D_i + \underbrace{Y_{0,i} - E(Y_0)}_{U_i}$$

Suppose that D_i is not randomly assigned, so that D_i is endogenous.

- How do our IV assumptions transfer to the random coefficients model?
- What does linear IV identify when treatment effects are heterogeneous?

Suppose Z_i is an instrument for D_i. For simplicity, assume Z_i is binary.

- Let $D_{z,i}$ be the treatment status at instrument value $Z_i = z$.
- Let $Y_{d,z,i}$ be the outcome at $D_i = d$ and $Z_i = z$.

IV Assumptions

For a binary treatment D and instrument Z, our IV assumptions are:

(1) **Random Assignment:** $Y_{d,z}, D_z \perp Z$ for all values of d and z
- In particular, the instrument Z must be assigned at random.

(2) **Exclusion:** $Y_{d,1} = Y_{d,0}$
- The instrument Z only affects the outcome Y via the treatment D.
- Thus, Z is not an omitted variable in $Y = \alpha + \beta D + U$.

(3) **Relevance/First Stage:** $E(D_1 - D_0) \neq 0$
- Z should have some effect on the average probability of treatment.

(4) **Monotonicity:** $D_1 \geq D_0$ for all individuals i (or vice versa)
- Everyone affected by the instrument Z is affected in the same direction.
- Put simply, the impact of Z on D is uniform across all participants i.

Deriving the IV Estimator

Consider a simple linear regression model $Y_i = \beta_0 + \beta_1 D_i + U_i$, where D_i is endogenous. Given an *i.i.d.* sample $\{Y_i, D_i, Z_i\}_{i=1}^n$, the IV estimator is:

$$\hat{\beta}_1^{IV} = \frac{\frac{1}{n}\sum_{i=1}^n (Z_i - \bar{Z}_n)(Y_i - \bar{Y}_n)}{\frac{1}{n}\sum_{i=1}^n (Z_i - \bar{Z}_n)(D_i - \bar{D}_n)}$$

If the IV assumptions hold, then $\hat{\beta}_1^{IV} \xrightarrow{p} \beta_1^{IV}$, where β_1^{IV} equals:

$$\beta_1^{IV} = \frac{\text{Cov}(Y_i, Z_i)}{\text{Cov}(D_i, Z_i)} = \underbrace{E(Y_{1,i} - Y_{0,i} | D_{1,i} > D_{0,i})}_{LATE}$$

Note that $\hat{\beta}_1^{IV}$ measures the *local average treatment effect* ($LATE$).

- $LATE$ equals ATE among individuals for whom $D_{1,i} = 1$ and $D_{0,i} = 0$.
- The average effect among those who enter treatment because $Z = 1$.

Partitioning the Population

To better understand the setting, we distinguish between four groups:

- *Compliers:* accept treatment *iff* $Z = 1$ ($D_1 = 1$ and $D_0 = 0$)
- *Always-Takers:* always accept treatment ($D_1 = 1$ and $D_0 = 1$)
- *Never-Takers:* never accept treatment ($D_1 = 0$ and $D_0 = 0$)
- *Defiers:* accept treatment *iff* $Z = 0$ ($D_1 = 0$ and $D_0 = 1$)

Note: monotonicity assumes that there are no defiers in the population.

Table: Subgroups Observed in Data (Assuming Monotonicity)

	$Z = 0$	$Z = 1$
$D = 0$	Never Takers & Compliers	Never Takers
$D = 1$	Always Takers	Always Takers & Compliers

ATE among Compliers

The *LATE* measures the average treatment effect among the compliers.

$$\beta^{IV} = E(Y_1 - Y_0 | \underbrace{D_1 > D_0}_{\text{compliers}})$$

If there is *noncompliance*, then this effect may not be especially relevant.

- In general, *LATE* does not equal the *ATE*, *ATT*, or *ATU*.
  - If there are no always takers, then $LATE = ATT$.
  - If there are no never takers, then $LATE = ATU$.
  - If there is full compliance, then $LATE = ATE$.
- *Note:* if there are defiers, then we cannot even measure the *LATE*.

If Z is randomly assigned, we can measure the *intention to treat* (ITT).

$$ITT = E(Y|Z = 1) - E(Y|Z = 0)$$

This tells us the average effect of Z on Y, which may be very relevant.

Example: Estimating the *LATE*

Suppose treatment D is free preschool and Y is future earnings. Let Z indicate whether households receive an informational brochure.

- Suppose D is endogenous because there is selection bias.
- Assume Z satisfies all of the instrumental variables assumptions.

Given an *i.i.d.* sample $\{Y_i, D_i, Z_i\}_{i=1}^n$, you estimate $\hat{\beta}_1^{IV}$, where:

$$\hat{\beta}_1^{IV} = \frac{\frac{1}{n}\sum_{i=1}^n (Z_i - \bar{Z}_n)(Y_i - \bar{Y}_n)}{\frac{1}{n}\sum_{i=1}^n (Z_i - \bar{Z}_n)(D_i - \bar{D}_n)}$$

Under the IV assumptions, $\hat{\beta}_1^{IV} \xrightarrow{p} E(Y_{1,i} - Y_{0,i} | D_{1,i} > D_{0,i}) = LATE$.

- $\hat{\beta}_1^{IV}$ estimates the average effect of preschool on future earnings for those who entered the program *because* their family got a brochure.
- It is often unclear whether the *LATE* is actually relevant. For example, are compliers representative of the wider population?