Selection on Observables

Thomas Wiemann
University of Chicago

Econometrics
Econ 21020

Updated: May 11, 2022

Introduction

Lecture 5 discussed the random assignment (RA) assumption:

- Proved point-identification of ATE, ATT, and ATU;
- Discussed estimation of causal parameters with discrete W.

Lecture 6 introduced the BLP and discussed OLS estimation:

- BLP as best linear approximation to the CEF;
- Allowed for approximate causal interpretation under RA.

But RA is not ubiquitously plausible or desired.

- RA suitable for experiments, not when agents optimize;
- RA implies ATE=ATT=ATU, but may be interested in selection.

Today: *Selection on Observables*.

- More general identifying assumption;
- Allows for studying selection on observed characteristics.

Outline

1. Selection
   - ▷ Roy Model
   - ▷ Confounders
2. Selection on Observables
   - ▷ Definition
   - ▷ Identification of Common Causal Parameters
3. Estimation with Discrete Variables
   - ▷ Conditional Average Treatment Effect Estimation
   - ▷ Average Treatment Effect Estimation
4. Evaluating Selection on Observables
5. Case Study: 401(k) Retirement Savings

Outline

1. **Selection**
   - ▷ **Roy Model**
   - ▷ Confounders
2. Selection on Observables
   - ▷ Definition
   - ▷ Identification of Common Causal Parameters
3. Estimation with Discrete Variables
   - ▷ Conditional Average Treatment Effect Estimation
   - ▷ Average Treatment Effect Estimation
4. Evaluating Selection on Observables
5. Case Study: 401(k) Retirement Savings

Roy Model

Consider an extension of the all causes model

$$Y = g(W, U) = \tilde{g}(W, X, \tilde{U}), \tag{1}$$

$$W = \mathbb{1}\left\{E_{\tilde{U}}\left[\tilde{g}(1, X, \tilde{U})|X\right] - c \geq E_{\tilde{U}}\left[\tilde{g}(0, X, \tilde{U})|X\right]\right\}. \tag{2}$$

where $c \in \mathbb{R}$ is a fixed threshold and $(Y, W, X, \tilde{U})$ is a random vector:

- $Y \equiv$ an outcome;
- $W \equiv$ a *binary* policy variable;
- $X \equiv$ all determinants of Y other than W *observed* by the agent;
- $\tilde{U} \equiv$ all determinants of Y other than W *unobserved* by the agent;
- and an economic model $\tilde{g} : \operatorname{supp} W \times \operatorname{supp} X \times \operatorname{supp} \tilde{U} \to \operatorname{supp} Y$.

The model in (1)-(2) is a version of the *Roy model*.

- Introduces *selection equation* to endogenize W;
- Agent decides whether $W = 1$ or $W = 0$ depending on whether the expected pay-off is larger than the threshold c.

Roy Model (Contd.)

Example 1

Recall the returns to education example from Lecture 1. We may have

- $Y \equiv$ lifetime earnings;
- $W \equiv$ an indicator for having obtained a college degree;
- $X \equiv$ grades from high school or perceived cleverness;
- $\tilde{U} \equiv$ ability on the job or future macroeconomic conditions;
- $g \equiv$ a labor production function;
- $c \equiv$ tuition fees.

According to the Roy model in (1)-(2):

- An individual pursues college if her expected lifetime earnings given her perceived cleverness improve by more than the tuition fees.

Roy Model (Contd.)

Example 2

A large literature (in the 90-2000s) studies the effects of 401(k) plans on retirement savings (e.g., Poterba et al., 1994, 1995).

- Tax-deferred savings option w/ employer contribution.

Here, we may have

- $Y \equiv$ retirement savings (in USD);
- $W \equiv$ an indicator for being enrolled in a 401(k) plan;
- $X \equiv$ income, non-401(k) savings, or financial literacy;
- $\tilde{U} \equiv$ future health or macroeconomic conditions;
- $g \equiv$ a savings preference function;
- $c \equiv$ (current-value) cost of 401(k) plan.

According to the Roy model in (1)-(2):

- An individual enrolls in a 401(k) plan if her retirement savings increase by more than the (current-value) cost of enrollment.

Outline

1. **Selection**
    - ▷ Roy Model
    - ▷ **Confounders**
2. Selection on Observables
    - ▷ Definition
    - ▷ Identification of Common Causal Parameters
3. Estimation with Discrete Variables
    - ▷ Conditional Average Treatment Effect Estimation
    - ▷ Average Treatment Effect Estimation
4. Evaluating Selection on Observables
5. Case Study: 401(k) Retirement Savings

Confounders

The Roy model in (1)-(2) economically motivates treatment:

- Optimizing agents choose treatment based on personal info;
- Treatment is *endogenous*.

We differentiate between X and $\tilde{U}$:

- Both X and $\tilde{U}$ are determinants of Y other than W...
- ... but the agent selects into treatment using only X.

A variable that affects *both* Y and W is called a *confounder*.

- A variable that does not affect *either* Y or W is *not a confounder*.

Selection Bias

In the presence of confounders, RA is violated.

Theorem 1

Let (Y, W, U) *be a random vector with joint distribution characterized by Equation* (1) *and* $\operatorname{supp} W = \{0, 1\}$*. Then*

$$E[Y|W = 1] - E[Y|W = 0] = ATE + \gamma_0 P(W = 1) + \gamma_1 P(W = 0),$$

where

$$\gamma_w \equiv E\left[g(w, U)|W = 1\right] - E\left[g(w, U)|W = 0\right], \; w \in \{0, 1\}.$$

The term $\gamma_1 P(W = 1) + \gamma_0 P(W = 0)$ is often dubbed *selection bias*.

- Captures expected difference in potential outcomes for treated and untreated individuals: It's the consequence of ignoring selection!
- Function of (the distribution of) U.

Proof.

$$\left\{ \begin{aligned} &E[g(w,U)\mid W=w] \\ &= E[g(W,U)\mid W=w] = E[Y\mid W=w] \end{aligned} \right.$$

$$
\begin{aligned}
ATE &= ATT\,P(W=1) + ATU\,P(W=0) \\
&= E[g(1,u) - g(0,u) \mid W=1]\,P(W=1) + E[g(1,u) - g(0,u) \mid W=0](1-P(W=1)) \\
&\quad + E[Y\mid W=1] - E[Y\mid W=1] \\
&= E[Y\mid W=1] - E[g(0,u)\mid W=0] + E[g(1,u)\mid W=1]\,P(W=1) - E[Y\mid W=1] \\
&\quad + E[g(0,u)\mid W=0]\,P(W=1) - E[g(0,u)\mid W=1]\,P(W=1) + E[g(1,u)\mid W=0](1-P(W=1)) \\
&= E[Y\mid W=1] - E[Y\mid W=0] + \left(E[g(1,u)\mid W=0] - E[g(1,u)\mid W=1]\right)P(W=0) \\
&\quad + \left(E[g(0,u)\mid W=0] - E[g(0,u)\mid W=1]\right)P(W=1)
\end{aligned}
$$

$$\Leftrightarrow\ E[Y\mid W=1] - E[Y\mid W=0] = ATE + \gamma_1 P(W=0) + \gamma_0 P(W=1)$$

□

Confounders (Contd.)

Example 3

Recall the returns to education Example 1. Examples of confounders are:

- Perceived intellect/talent;
- Work discipline;
- Parent's connections in industry/government;
- etc...

Are the following confounders? Why or why not?

- Winning the lottery at age 18;
- Winning the lottery at age 53;
- Chicago's Polar Vortex in 2018.

Confounders (Contd.)

Example 4

Recall the 401(k) Example 2. Examples of confounders are:

- ▷ Income;
- ▷ Financial literacy;
- ▷ Education;
- ▷ etc...

Are the following confounders? Why or why not?

- ▷ Martial status;
- ▷ Personal saving preferences/risk aversion;
- ▷ A public-awareness campaign for old-age poverty.

Outline

1. Selection
   - ▷ Roy Model
   - ▷ Confounders
2. **Selection on Observables**
   - ▷ **Definition**
   - ▷ Identification of Common Causal Parameters
3. Estimation with Discrete Variables
   - ▷ Conditional Average Treatment Effect Estimation
   - ▷ Average Treatment Effect Estimation
4. Evaluating Selection on Observables
5. Case Study: 401(k) Retirement Savings

Selection on Observables

Theorem 1 shows that the ATE is unidentified in the presence of unobserved confounders. Similar results hold for the ATT and ATU.

We thus require a different identifying assumption.

- ▷ Consider observables (Y, W, X) and unobservables U .

Assumption 1 (Selection on Observables; SO)

Let (Y, W, X, U) be a random vector with joint distribution characterized by Equation (1). *Selection on Observables* assumes

$$W \perp\!\!\!\perp U \,|\, X. \tag{3}$$

In words: Conditional on X, the policy W is independent of U.

- ▷ SO violated if conditional on X (parts of) U affect the policy W.
- ▷ Most plausible when the selection mechanism is known exactly.
- ▷ Most problematic when selection mechanism is intransparent.

Note: *SO is a generalization of RA. To see this, simply take $X = 1$.*

Selection on Observables (Contd.)

Example 5

Recall the returns to education Example 1. Suppose X denotes

- high school grades, gender, age, and martial status.

Does SO seem plausible here?

- SO fails if students who obtained a college degree were systematically different from others with identical high school grades, gender, age, and martial status.
- SO is implausible because students likely select into college based on more characteristics, e.g., connections in industry.
- Even among those with identical high school grades, gender, age, and martial status, students are *not* obtaining a college degree as if it was random: We should expect a substantial association between obtaining a college degree and socio-economic backgrounds.

Selection on Observables (Contd.)

Example 6

Recall the 401(k) Example 2. Suppose X denotes

- ▷ income, years of education, gender, age, and martial status.

Does SO seem plausible here?

- ▷ SO fails if those enrolled in a 401(k) were systematically different from others with identical income, years of education, gender, age, and martial status.

Poterba et al. (1994, 1995) argue for plausibility of SO conditional on employee and employer characteristics. Key idea:

- ▷ 401(k) eligibility is employer-determined;
- ▷ Employees working at similar firms are assumed to be similar.

Later studies place more emphasis on heterogeneous saving preferences.

- ▷ E.g., Chernozhukov and Hansen (2004).

Outline

1. Selection
   - ▷ Roy Model
   - ▷ Confounders
2. **Selection on Observables**
   - ▷ Definition
   - ▷ **Identification of Common Causal Parameters**
3. Estimation with Discrete Variables
   - ▷ Conditional Average Treatment Effect Estimation
   - ▷ Average Treatment Effect Estimation
4. Evaluating Selection on Observables
5. Case Study: 401(k) Retirement Savings

Common Support

We now turn to identification of the ATE, ATT, and ATU.

In addition to Assumption SO, we will require that the conditional expectations $E[Y|W = w, X = x]$ are *well-defined*.

Assumption 2 (Common Support; CS)

Let (Y, W, X, U) be a random vector with joint distribution characterized by Equation (1). *Common Support* assumes

$$\operatorname{supp} X|W = \operatorname{supp} X. \tag{4}$$

- ▷ CS ensures that there are both treated/untreated with the same X.

If X and W are...

- ▷ ... discrete, then $P(X = x, W = w) > 0$,...
- ▷ ... continuous, then $f_{X,W}(x, w) > 0$,...

... $\forall(x, w) \in \operatorname{supp} X \times \operatorname{supp} W$ is sufficient for CS.

Theorem 2

Let (Y, W, X, U) be a random vector with joint distribution characterized by Equation (1). Under SO and CS, CATE(x) is point-identified $\forall x \in \operatorname{supp} X$.

Proof.

$$
\begin{aligned}
CATE(x) &= E[g(1,u) - g(0,u) \mid X = x] \\
&= E[g(1,u) \mid X = x] - E[g(0,u) \mid X = x] \\
&\overset{SO}{=} E[g(1,u) \mid W = 1, X = x] - E[g(0,u) \mid W = 0, X = x] \\
&= E[g(W,u) \mid W = 1, X = x] - E[g(W,u) \mid W = 0, X = x] \\
&= E[Y \mid W = 1, X = x] - E[Y \mid W = 0, X = x]
\end{aligned}
$$

$\square$

Identification (Contd.)

Corollary 1

Let (Y, W, X, U) be a random vector with joint distribution characterized by Equation (1). Under SO and CS, the ATE, ATT, and ATU are point-identified.

Proof.

$$ATT = E[g(1,u) - g(0,u) \mid W=1] = E[E[g(1,u) - g(0,u) \mid W=1, X] \mid W=1]$$

$$\overset{SO}{=} E[E[g(1,u) - g(0,u) \mid X] \mid W=1] = E[CATE(X) \mid W=1]$$

$$ATU = E[g(1,u) - g(0,u) \mid W=0] = E[E[g(1,u) - g(0,u) \mid W=0, X] \mid W=0]$$

$$\overset{SO}{=} E[E[g(1,u) - g(0,u) \mid X] \mid W=0] = E[CATE(X) \mid W=0]$$

$$ATE = E[g(1,u) - g(0,u)] = E[E[g(1,u) - g(0,u) \mid X]] = E[CATE(X)]$$

□

Identification (Contd.)

Notice that under SO, the ATE, ATT, and ATU (potentially) differ!

- ▷ Proof of Corollary 1 showed differences stem from CATE$(X)|W$;
- ▷ Agents select into treatment based on observables only.

SO allows for studying observed selection mechanism.

- ▷ Improvement over RA which prohibits selection;
- ▷ When selection mechanism is known, SO may be plausible.
- ▷ When agents select based on unobservables, SO fails.

Outline

1. Selection
   - ▷ Roy Model
   - ▷ Confounders
2. Selection on Observables
   - ▷ Definition
   - ▷ Identification of Common Causal Parameters
3. **Estimation with Discrete Variables**
   - ▷ **Conditional Average Treatment Effect Estimation**
   - ▷ Average Treatment Effect Estimation
4. Evaluating Selection on Observables
5. Case Study: 401(k) Retirement Savings

CATE Estimation

The identification proofs showed that CATE, ATE, ATT, and ATU can be expressed as known functions of the moments of observables (Y, W, X).

▷ Suggests sample analogue estimator when (W, X) are discrete.

For everything that follows, we consider binary W and discrete X.

Theorem 2 showed that under SO and CS, we have

$$\text{CATE}(x) = E[Y|W = 1, X = x] - E[Y|W = 0, X = x]. \tag{5}$$

Consider a sample $(Y_1, W_1, X_1), \ldots, (Y_n, W_n, X_n) \overset{iid}{\sim} (Y, W, X.)$

For discrete (W, X), we can construct a sample analogue estimator:

$$\widehat{\text{CATE}}_n(x) \equiv \frac{\sum Y_i \mathbb{1}_{(1,x)}(W_i, X_i)}{\sum \mathbb{1}_{(1,x)}(W_i, X_i)} - \frac{\sum Y_i \mathbb{1}_{(0,x)}(W_i, X_i)}{\sum \mathbb{1}_{(0,x)}(W_i, X_i)} \tag{6}$$

CATE Estimation (Contd.)

For discrete (W, X), $\widehat{\text{CATE}}_n(x)$ is a difference in binning estimators:

- Asymp. properties of $\widehat{\text{CATE}}_n(x)$ follow from Theorem 2 & Lecture 5;
- We state consistency, asymptotic distribution, and the standard error for completeness.

Corollary 2

Let (Y, W, X, U) be a random vector with joint distribution characterized by Equation (1). Consider a random sample $(Y_1, W_1, X_1), \ldots, (Y_n, W_n, X_n) \overset{iid}{\sim} (Y, W, X)$, and let $\widehat{CATE}_n(x)$ be the estimator in (6). Under SO and CS, it holds that

$$\widehat{CATE}_n(x) \xrightarrow{p} CATE(x), \tag{7}$$

$\forall x \in \operatorname{supp} X$.

CATE Estimation (Contd.)

Corollary 3

Under the conditions of Corollary 2, it holds that

$$\sqrt{n}\left(\widehat{CATE}_n(x) - CATE(x)\right) \xrightarrow{d} N\left(0, \sigma^2_{CATE}(x)\right), \tag{8}$$

where

$$\sigma^2_{CATE}(x) = \frac{Var(Y|W=1, X=x)}{P(W=1, X=x)} + \frac{Var(Y|W=0, X=x)}{P(W=0, X=x)}.$$

CATE Estimation (Contd.)

Corollary 4

Under the conditions of Corollary 2, it holds that

$$\frac{\widehat{CATE}_n(x) - CATE(x)}{se\left(\widehat{CATE}_n(x)\right)} \xrightarrow{d} N(0,1), \tag{9}$$

where

$$se\left(\widehat{CATE}_n(x)\right) = \frac{1}{\sqrt{n}}\sqrt{\hat{\sigma}^2_{CATE}(x)},$$

$$\hat{\sigma}^2_{CATE}(x) = \frac{\hat{\sigma}^2_{1,n}(x)}{\hat{p}_{1,n}(x)} + \frac{\hat{\sigma}^2_{0,n}(x)}{\hat{p}_{0,n}(x)}, \quad \hat{p}_{w,n}(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{(w,x)}(W_i, X_i),$$

$$\hat{\sigma}^2_{w,n}(x) = \frac{\frac{1}{n}\sum_{i=1}^{n} Y_i^2 \mathbb{1}_{(w,x)}(W_i, X_i)}{\hat{p}_{w,n}(x)} - \left(\frac{\frac{1}{n}\sum_{i=1}^{n} Y_i \mathbb{1}_{(w,x)}(W_i, X_i)}{\hat{p}_{w,n}(x)}\right)^2.$$

R Function for CATE Estimation under SO

R Function for CATE Estimation under SO

```
calc_cate <- function(y, w, x, x_val) {
  # Find treated/untreated individuals for x = x_val
  y_w1_x <-y[w == 1 & x == x_val]
  y_w0_x <-y[w == 0 & x == x_val]
  # Estimate conditional means
  mu_w1_x <- mean(y_w1_x)
  mu_w0_x <- mean(y_w0_x)
  # Estimate CATE
  cate_x <- mu_w1_x - mu_w0_x
  # Compute standard error
  n <- length(y)
  p_w1_x <- mean(w == 1 & x == x_val)
  p_w0_x <- mean(w == 0 & x == x_val)
  se_cate_x <- sqrt((var(y_w1_x) / p_w1_x +
                     var(y_w0_x) / p_w0_x) / n)
 # Return CATE and SE
 return(cate_x, se_cate_x)
}#CALC_CATE
```

Outline

1. Selection
   - ▷ Roy Model
   - ▷ Confounders
2. Selection on Observables
   - ▷ Definition
   - ▷ Identification of Common Causal Parameters
3. **Estimation with Discrete Variables**
   - ▷ Conditional Average Treatment Effect Estimation
   - ▷ **Average Treatment Effect Estimation**
4. Evaluating Selection on Observables
5. Case Study: 401(k) Retirement Savings

ATE Estimation

Corollary 1 showed that under SO and CS, we have

$$\text{ATE} = E\left[\text{CATE}(X)\right]. \tag{10}$$

$$= \sum_{x \in \text{supp}X} \text{CATE}(x) P(X = x)$$

For discrete (W, X) a sample analogue estimators for the ATE is

$$\widehat{\text{ATE}}_n \equiv \sum_{x \in \text{supp}X} \widehat{\text{CATE}}_n(x) \left(\frac{1}{n}\sum X_i\right) \tag{11}$$

Estimators for the ATT and ATU are constructed similarly.

Asymptotic properties of the $\widehat{\text{ATE}}_n$ are challenging:

- Average of $\widehat{\text{CATE}}_n(x)$ over *empirical* distribution of X;
- Will prove consistency for discrete X...
- ... but focus on binary X for asymptotic distribution.

ATE Estimation (Contd.)

Theorem 3

Let (Y, W, X, U) be a random vector with joint distribution characterized by Equation (1). Consider $(Y_1, W_1, X_1), \dots, (Y_n, W_n, X_n) \overset{iid}{\sim} (Y, W, X)$, and let $\widehat{ATE}_n$ be the estimators in (11). Under SO and CS, it holds that

$$\widehat{ATE}_n \xrightarrow{p} ATE. \tag{12}$$

Proof. Let $\mathcal{X} \equiv \operatorname{supp} X$. Note $\widehat{ATE}_n = \sum_{x \in \mathcal{X}} \widehat{CATE}_n(x) \left(\frac{1}{n} \sum \mathbb{1}_x(X_i) \right)$

1. $A_n^{(x)} \equiv \frac{1}{n} \sum \mathbb{1}_x(X_i)$, $B_n^{(x)} \equiv \widehat{CATE}_n(x)$, $\forall x \in \mathcal{X}$
2. $g\left((a^{(x)}, b^{(x)})_{x \in \mathcal{X}} \right) = \sum_{x \in \mathcal{X}} a^{(x)} b^{(x)}$
3. By WLLN, $A_n^{(x)} \xrightarrow{p} P(X = x)$, $\forall x \in \mathcal{X}$

   By Corollary 2, $B_n^{(x)} \xrightarrow{p} CATE(x)$, $\forall x \in \mathcal{X}$
4. By CMT, $g\left((A_n^{(x)}, B_n^{(x)})_{x \in \mathcal{X}} \right) \xrightarrow{p} \sum_{x \in \mathcal{X}} CATE(x) P(X = x) = E[CATE(X)] = ATE$. $\square$

ATE Estimation (Contd.)

We state the asymptotic distribution only for binary X.

Theorem 4

Let (Y, W, X, U) *be a random vector with joint distribution characterized by Equation* (1)*. Suppose that* $\operatorname{supp} X = \{0, 1\}$*. Consider* $(Y_1, W_1, X_1), \ldots, (Y_n, W_n, X_n) \overset{iid}{\sim} (Y, W, X)$*, and let* $\widehat{ATE}_n$ *be the estimator in* (11)*. Under SO and CS, it holds that*

$$\sqrt{n}\left(\widehat{ATE}_n - ATE\right) \xrightarrow{d} N(0, \sigma^2_{ATE}), \tag{13}$$

where

$$\begin{aligned}\sigma^2_{ATE} = {} & \sigma^2_{CATE}(1)P(X = 1)^2 + \sigma^2_{CATE}(0)P(X = 0)^2 \\ & + \left(CATE(1) - CATE(0)\right)^2 P(X = 1)P(X = 0).\end{aligned}$$

ATE Estimation (Contd.)

Proof.

$$\widehat{ATE}_n - ATE = \sum_{x\in\mathcal{X}} \widehat{CATE}_n(x)\left(\frac{1}{n}\sum 1\!\!1_x(X_i)\right) - \sum_{x\in\mathcal{X}} CATE(x)P(X=x)$$

$$+ \sum_{x\in\mathcal{X}} CATE(x)\left(\frac{1}{n}\sum 1\!\!1_x(X_i)\right) - \sum_{x\in\mathcal{X}} CATE(x)\left(\frac{1}{n}\sum 1\!\!1_x(X_i)\right)$$

Let $\mathcal{X} = \{0,1\}$.

$$\widehat{ATE}_n - ATE = \left(\widehat{CATE}_n(1) - CATE(1)\right)\frac{1}{n}\sum X_i + \left(\widehat{CATE}_n(0) - CATE(0)\right)\frac{1}{n}\sum(1-X_i)$$

$$+ \left(\frac{1}{n}\sum X_i - P(X=1)\right)CATE(1) + \left(\frac{1}{n}\sum(1-X_i) - P(X=0)\right)CATE(0)$$

Define $U_i \equiv Y_i - E[Y_i|W_i, X_i]$. Then by Problem Set 2 + Lecture 5

$$\widehat{ATE}_n - ATE = \left(\frac{\sum U_i W_i X_i}{\sum W_i X_i} - \frac{\sum U_i(1-W_i)X_i}{\sum(1-W_i)X_i}\right)\frac{1}{n}\sum X_i + \left(\frac{\sum U_i W_i(1-X_i)}{\sum W_i(1-X_i)} - \frac{\sum U_i(1-W_i)(1-X_i)}{\sum(1-W_i)(1-X_i)}\right)\frac{1}{n}\sum(1-X_i)$$

$$+ \left(\frac{1}{n}\sum X_i - P(X=1)\right)CATE(1) + \left(\frac{1}{n}\sum(1-X_i) - P(X=0)\right)CATE(0)$$

$$= \frac{\sum X_i}{\sum W_i X_i}\frac{1}{n}\sum U_i W_i X_i - \frac{\sum X_i}{\sum(1-W_i)X_i}\frac{1}{n}\sum U_i(1-W_i)X_i + \frac{\sum(1-X_i)}{\sum W_i(1-X_i)}\frac{1}{n}\sum U_i W_i(1-X_i) - \frac{\sum(1-X_i)}{\sum(1-W_i)(1-X_i)}\frac{1}{n}\sum U_i(1-W_i)(1-X_i)$$

$$+ \left(\frac{1}{n}\sum X_i - P(X=1)\right)CATE(1) + \left(\frac{1}{n}\sum(1-X_i) - P(X=0)\right)CATE(0)$$

ATE Estimation (Contd.)

$$\sqrt{n}(\widehat{ATE}_n - ATE) = \underbrace{\begin{bmatrix} \frac{\sum X_i}{\sum W_i X_i} \\ \frac{-\sum X_i}{\sum (1-W_i) X_i} \\ \frac{\sum (1-X_i)}{\sum W_i (1-X_i)} \\ \frac{-\sum (1-X_i)}{\sum (1-W_i)(1-X_i)} \\ CATE(1) \\ CATE(0) \end{bmatrix}^T}_{\equiv A_n} \underbrace{\sqrt{n} \left(\begin{bmatrix} \frac{1}{n} \sum U_i W_i X_i \\ \frac{1}{n} \sum U_i (1-W_i) X_i \\ \frac{1}{n} \sum U_i W_i (1-X_i) \\ \frac{1}{n} \sum U_i (1-W_i)(1-X_i) \\ \frac{1}{n} \sum X_i \\ \frac{1}{n} \sum (1-X_i) \end{bmatrix} - \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ P(X=1) \\ P(X=0) \end{bmatrix} \right)}_{\equiv B_n}$$

By WLLN + CMT,

$$A_n \xrightarrow{p} \begin{bmatrix} \frac{P(X=1)}{P(W=1, X=1)} & \frac{-P(X=1)}{P(W=0, X=1)} & \frac{P(X=0)}{P(W=1, X=0)} & \frac{-P(X=0)}{P(W=0, X=0)} & CATE(1) & CATE(0) \end{bmatrix} \equiv \mathcal{E}$$

if CS holds!

ATE Estimation (Contd.)

By CLT, $B_n \xrightarrow{d} \mathcal{N}(0, \Sigma)$

w/

$$\Sigma = \begin{bmatrix} Var(UWX) & & & & & \\ \circledast & Var(U(1-W)X) & & & & \\ 0 & 0 & Var(UW(1-X)) & & & \\ 0 & 0 & 0 & Var(U(1-W)(1-X)) & & \\ 0 & 0 & 0 & 0 & Var(X) & \\ 0 & 0 & 0 & 0 & \underbrace{Cov(X, 1-X)}_{\substack{= Cov(X,-X) \\ = -Var(X)}} & Var(1-X) \end{bmatrix}$$

$$\circledast = Cov(UWX, U(1-W)X)$$

$$= E[U^2 \underbrace{W(1-W)}_{=0} X^2] - \underbrace{E[UWX]}_{= E[E[UWX|WX]] = E[\underbrace{E[U|WX]}_{=0} WX] = 0} E[U(1-W)X] = 0$$

ATE Estimation (Contd.)

From Problem Set 2 + Lecture 5:

$$Var(UWX) = Var(Y|W=1, X=1)P(W=1, X=1)$$

$$Var(U(1-W)X) = Var(Y|W=0, X=1)P(W=0, X=1)$$

$$Var(UW(1-X)) = Var(Y|W=1, X=0)P(W=1, X=0)$$

$$Var(U(1-W)(1-X)) = Var(Y|W=0, X=0)P(W=0, X=0)$$

$$Var(X) = P(X=1)(1-P(X=1)) = P(X=1)P(X=0) = Var(1-X)$$

By Slutsky's,

$$\sqrt{n}(\widehat{ATE}_n - ATE) \xrightarrow{d} \epsilon^T N(0, \Sigma) \stackrel{d}{=} N(0, \epsilon^T \Sigma \epsilon)$$

w/

$$\epsilon^T \Sigma \epsilon = Var(Y|W=1, X=1)P(W=1, X=1)\left(\frac{P(X=1)}{P(W=1, X=1)}\right)^2 + Var(Y|W=0, X=1)P(W=0, X=1)\left(\frac{P(X=1)}{P(W=0, X=1)}\right)^2$$

$$+ Var(Y|W=1, X=0)P(W=1, X=0)\left(\frac{P(X=0)}{P(W=1, X=0)}\right)^2 + Var(Y|W=0, X=0)P(W=0, X=0)\left(\frac{P(X=0)}{P(W=0, X=0)}\right)^2$$

$$+ (CATE(1) - CATE(0))^2 P(X=1)P(X=0)$$

$\square$

ATE Estimation (Contd.)

Corollary 5

Under the conditions of Theorem 4, it holds that

$$\frac{\widehat{ATE}_n - ATE}{se\left(\widehat{ATE}_n\right)} \xrightarrow{d} N(0,1), \tag{14}$$

where

$$\begin{aligned}
se\left(\widehat{ATE}_n\right) &= \frac{1}{\sqrt{n}}\sqrt{\hat{\sigma}^2_{ATE}}, \\
\hat{\sigma}^2_{ATE} &= \hat{\sigma}^2_{CATE}(1)\hat{p}_n(1)^2 + \hat{\sigma}^2_{CATE}(0)\hat{p}_n(0)^2 \\
&\quad + \left(\widehat{CATE}_n(1) - \widehat{CATE}_n(0)\right)\hat{p}_n(1)\hat{p}_n(0), \\
\hat{p}_n(x) &= \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_x(X_i).
\end{aligned}$$

- Proof: Problem 2 of Problem Set 4.

R Function for ATE Estimation under SO

R Function for ATE Estimation under SO

```
calc_ate <- function(y, w, x) {
  # Estimate CATEs, P(X=1), and P(X=0)
  cate_x1 <- calc_cate(y, w, x, 1)
  cate_x0 <- calc_cate(y, w, x, 0)
  p_x1 <- mean(x == 1)
  p_x0 <- 1 - p_x1
  # Estimate ATE
  ate <- cate_x1[1] * p_x1 + cate_x0[1] * p_x0
  # Compute standard error
  n <- length(y)
  sgm2_ate  <- n * (cate_x1[2] * p_x1)^2 +
    n * (cate_x0[2] * p_x0)^2 +
    (cate_x1[1] - cate_x0[1])^2 * p_x1 * p_x0
  se_ate <- sqrt(sgm2_ate) / sqrt(n)
  # Return ATE and SE
  return(c(ate, se_ate))
}#CALC_ATE
```

Outline

1. Selection
   - ▷ Roy Model
   - ▷ Confounders
2. Selection on Observables
   - ▷ Definition
   - ▷ Identification of Common Causal Parameters
3. Estimation with Discrete Variables
   - ▷ Conditional Average Treatment Effect Estimation
   - ▷ Average Treatment Effect Estimation
4. **Evaluating Selection on Observables**
5. Case Study: 401(k) Retirement Savings

Evaluating Common Support

Identification was based on two assumptions: CS & SO.

Recall CS assumes $\operatorname{supp} X|W = \operatorname{supp} X$.

When W and X are discrete, a sufficient condition for CS is

$$P(W = w,\ X = x) > 0, \quad \forall (w, x) \in \operatorname{supp} W \times \operatorname{supp} X. \tag{15}$$

Notice that condition (15) only involves observables:

- ▷ *Can* verify CS when W and X are discrete!

In practice:

- ▷ Check whether every combination of X and W exists in the data;
- ▷ The more observations per cell, the better (else: small bin problem).

Evaluating Selection on Observables

Suppose now that CS holds. We turn to evaluating SO.

Recall SO assumes $W \perp\!\!\!\perp U | X$.

- Restriction on the joint of (Y, W, X, U);
- Since the sampling process provides no information on the entirety of U, it's impossible to verify SO;
- But SO has implications that we can test;
- Idea: Adapt balance test considered when evaluating RA.

Suppose that we observe *some* additional variables in U not in X, say $\tilde{X}$.

- $\tilde{X}$ assumed not to be necessary for SO;
- Then if SO holds $W \perp\!\!\!\perp U | X \Rightarrow W \perp\!\!\!\perp \tilde{X} | X \Rightarrow E[\tilde{X} | W, X] = E[\tilde{X} | X]$.

Since $(W, X, \tilde{X})$ are observable, we may construct a test!

Evaluating Selection on Observables (Contd.)

Can construct a test based on $E[\tilde{X}|W,X] = E[\tilde{X}|X]$ under SO.

As before, suppose that W is binary and X is discrete. Consider testing

$$H_0: \quad \mu_{\tilde{X}|1}(x) = \mu_{\tilde{X}|0}(x), \quad \forall x \in \operatorname{supp} X$$

$$\Rightarrow \quad 0 = E[\tilde{X}|W=1, X=x] - E[\tilde{X}|W=0, X=x] \equiv \widetilde{CATE}(x), \quad \forall x \in \operatorname{supp} X$$

versus

$$H_1: \quad \exists x \in \operatorname{supp} X \text{ s.t. } \mu_{\tilde{X}|1}(x) \neq \mu_{\tilde{X}|0}(x),$$

where $\mu_{\tilde{X}|w}(x) \equiv E[\tilde{X}|W=w, X=x]$.

Essentially testing whether the CATE(x) of W on $\tilde{X}$ is zero for all X.

- Replace Y with $\tilde{X}$ in previous analysis;
- Then use Corollary 4 to construct a test statistic.

Evaluating Selection on Observables (Contd.)

Suppose we have $(Y_1, W_1, X_1, \tilde{X}_1), \ldots, (Y_n, W_n, X_n, \tilde{X}_n) \overset{iid}{\sim} (Y, W, X, \tilde{X})$.

Our analysis suggests a test statistic given by

$$T_n = n \begin{bmatrix} \widetilde{CATE}_n(x_1) \\ \vdots \\ \widetilde{CATE}_n(x_k) \end{bmatrix}^T \begin{bmatrix} \hat{\sigma}^{-2}_{\widetilde{CATE}}(x_1) & & 0 \\ & \ddots & \\ 0 & & \hat{\sigma}^{-2}_{\widetilde{CATE}}(x_k) \end{bmatrix} \begin{bmatrix} \widetilde{CATE}_n(x_1) \\ \vdots \\ \widetilde{CATE}_n(x_k) \end{bmatrix} \tag{16}$$

We can use the quantiles of a χ^2-distribution as critical values.

Corollary 6

Let (Y, W, X, U) be a random vector with joint distribution characterized by Equation (1) and let $U = (\tilde{X}, U_2)$. Consider a random sample $(Y_1, W_1, X_1, \tilde{X}_1), \ldots, (Y_n, W_n, X_n, \tilde{X}_n) \overset{iid}{\sim} (Y, W, X, \tilde{X})$, and let T_n be the test statistic given in Equation (16). Under SO and CS, it holds that

$$T_n \xrightarrow{d} \chi^2(d_X),$$

where $\chi^2(d_X)$ is a χ^2-distribution with $d_X \equiv |\operatorname{supp} X|$ degrees of freedom.

Continuous Mapping Theorem for Convergence in Distribution

We require the CMT for convergence in *distribution* for the proof.

Theorem 5 (Continuous Mapping Theorem; CMT)

Let $X_n, n \geq 1$, be a sequence of random vectors, and let and X be another random vector. If $X_n \xrightarrow{d} X$, then

$$g\left(X_n\right) \xrightarrow{d} g\left(X\right), \tag{17}$$

for any function g that is continuous at $g\left(x\right), \forall x \in \operatorname{supp} X$.

Example 7

Let $A_n \xrightarrow{d} A \sim N(0,1)$. Consider $g(a) = a^2$. Then

$$g\left(A_n\right) \xrightarrow{d} A^2 \sim \chi^2(1), \tag{18}$$

by the CMT and Theorem 4 of Lecture 2A.

Proof of Corollary 6

Proof.

1. $A_n = \sqrt{n} \begin{bmatrix} \hat{\sigma}^{-1}_{\widetilde{CATE}}(x_1) & & 0 \\ & \ddots & \\ 0 & & \hat{\sigma}^{-1}_{\widetilde{CATE}}(x_{d_x}) \end{bmatrix} \begin{bmatrix} \widetilde{CATE}_n(x_1) \\ \widetilde{CATE}_n(x_{d_x}) \end{bmatrix}$

2. $g(a) = a^T a$, $g(A_n) = A_n^T A_n = T_n$

3. By Corollary 4 + Proof of Theorem 4,

$$A_n \xrightarrow{d} \mathcal{N}(0, I_{d_x}) \equiv Z$$

4. By CMT, $g(A_n) \xrightarrow{d} g(Z) = Z^T Z$

$$= \sum_{j=1}^{d_x} Z_j^2 \sim \chi^2(d_x)$$

Lecture 2A Theorem 4

R Function for Evaluating SO

The code snippet below implements the balance test for binary X.

R Function for Evaluating SO

```
test_SO <- function(x_tld, w, x) {
  # Calculate CATEs of w on x_tld
  cate_x1 <- calc_cate(x_tld, w, x, 1)
  cate_x0 <- calc_cate(x_tld, w, x, 0)
  # Calculate test statistic
  cates <- c(cate_x1[1], cate_x0[1])
  vars <- c(cate_x1[2], cate_x0[2])^2
  Tn <- cates %*% diag(1 / vars) %*% cates
  # Compute p-value
  pval <- pchisq(Tn, 2, lower.tail = FALSE)
  # Return output
  return(c(Tn, pval))
}#TEST_SO
```

Outline

1. Selection
   - ▷ Roy Model
   - ▷ Confounders
2. Selection on Observables
   - ▷ Definition
   - ▷ Identification of Common Causal Parameters
3. Estimation with Discrete Variables
   - ▷ Conditional Average Treatment Effect Estimation
   - ▷ Average Treatment Effect Estimation
4. Evaluating Selection on Observables
5. **Case Study: 401(k) Retirement Savings**

Case Study: 401(k) Retirement Savings

A large literature in the 90-2000s studies the effect of 401(k) participation on savings: 401(k) plans introduced in 70-80s to incentivize savings.

- ▷ Tax-deferred savings option w/ employer contribution.

Prominent examples are Poterba et al. (1994, 1995).

- ▷ Analysis based on selection on observables assumption;
- ▷ Condition on employee and employer characteristics;
- ▷ Idea: Similar workers at similar firms randomly enroll in 401(k)s.

Data:

- ▷ 9915 households from the 1991 PSID;
- ▷ Net total financial wealth;
- ▷ 401(k) participation;
- ▷ Employee characteristics: e.g., yrs of education, income.

Note: *The specific data used for our analysis is taken from Chernozhukov and Hansen (2004). You can find the data file on Canvas:* `psid91.csv`*. The R script used for estimation can be found on GitHub:* `example_psid91.R`.

Case Study: 401(k) Retirement Savings (Contd.)

Suppose we are interested in assessing the effect of 401(k) participation on net total financial wealth. For this purpose, let (Y, W, X, U) be random variables, where $Y = g(W, U)$ and

- $\triangleright$ $Y \equiv$ net total financial wealth;
- $\triangleright$ $W \equiv$ an indicator for participation in a 401(k) plan;
- $\triangleright$ $X \equiv$ an indicator for at least 16 yrs of education;
- $\triangleright$ $U \equiv$ all determinants of Y other than W.

We assume that the PSID data is the realization of the sample $(Y_1, W_1, X_1), \ldots, (Y_{9915}, W_{9915}, X_{9915}) \overset{iid}{\sim} (Y, W, X)$.

We now proceed with the three distinct tasks of causal analysis!

Task 1: Definition

The conventional parameter of interest is often the ATE:

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

Economic interpretation in the 401(k)-setting:

- ▷ The ATE is the expected causal effect of participation in a 401(k) plan on net total fin. assets for a randomly selected individual.

We may also be interested in the conditional causal effects. Here,

$$\text{CATE}(1) = E\left[g(1, U) - g(0, U) | X = 1\right], \tag{19}$$

$$\text{CATE}(0) = E\left[g(1, U) - g(0, U) | X = 0\right]. \tag{20}$$

Economic interpretation in the 401(k)-setting:

- ▷ The CATE(1) (CATE(0)) is the expected causal effect of 401(k) participation on net total fin. assets for a randomly selected individual *with at least (less than) 16 yrs of educ*.

Task 2: Identification

ATE, CATE(1), and CATE(0) are functions (of the distribution of) U.

- ▷ Cannot learn about causal parameters using data alone;
- ▷ An identifying assumption is *necessary*.

We consider a selection on observables assumption: Assume $W \perp\!\!\!\perp U | X$.

- ▷ Assumes that conditional on being having at least/less than 16 yrs of education, 401(k) participation is independent of all other determinants of net total financial assets.

Assumption motivated by arguments in Poterba et al. (1994, 1995):

- ▷ Argue that conditional on employee and employer characteristics, 401(k) participation is reasonably random.

***Note**: If you have concerns regarding the plausibility of the SO assumption here... excellent! You're thinking critically about assumptions underlying causal analysis.*

Task 2: Identification (Contd.)

But ATE, CATE(1), and CATE(0) remain unidentified...

We also need to assume common support: $\operatorname{supp} X|W = \operatorname{supp} X$.

- ▷ Assumes that there exists individuals with at least/less than 16 yrs of education who participate/do not participate in a 401(k) plan.
- ▷ Fails if, e.g., all 401(k) participants are college grads.

Assuming SO and CS, the ATE, CATE(1), and CATE(0) are identified.

- ▷ Follows immediately from Theorem 2 and Corollary 1.

Task 3: Estimation

We can now turn to estimation of the CATEs.

- ▷ Note that W and X are discrete $\Rightarrow$ use sample analogue estimators.

Estimates for the CATE(1) using the 1991 PSID data are:

$$\widehat{\text{CATE}}_n(1) \approx 29{,}955 \quad , \text{ and } se\left(\widehat{\text{CATE}}_n(1)\right) \approx 3{,}936$$

$$c_n^{\text{CATE}(1)} \approx [22240, \ 37669]$$

Similarly, for the CATE(0), we have:

$$\widehat{\text{CATE}}_n(0) \approx 23{,}694 \quad , \text{ and } se\left(\widehat{\text{CATE}}_n(0)\right) \approx 1{,}684$$

$$c_n^{\text{CATE}(0)} \approx [20393, \ 26996]$$

Interpretation:

- ▷ Assuming SO, the expected causal effect of 401(k) participation on net total fin. assets for a randomly selected individual with at least (less than) 16yrs of education is estimated to be 29,955 (23,694).

Task 3: Estimation (Contd.)

For the (unconditional) average effect of 401(k) participation, we have

$$\widehat{\text{ATE}}_n \approx 25{,}262 \quad , \text{ and } se\left(\widehat{\text{ATE}}_n\right) \approx 1{,}602$$

$$c_n^{\text{ATE}} \approx [22122,\ 28402]$$

Interpretation:

- Assuming SO, the expected causal effect of 401(k) participation on net total fin. assets for a randomly selected individual is estimated to be 25262 .

Discussion

We made two key assumptions for identification:

- ▷ Common Support & Selection on Observables.

Since W and X are discrete, we can verify CS straightforwardly:

- ▷ Check $\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{(w,x)}(W_i, X_i) > 0, \forall (w,x) \in \operatorname{supp} W \times \operatorname{supp} X$.
- ▷ We have $\min_{(w,x)} \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{(w,x)}(W_i, X_i) = 0.086 > 0$,
- ▷ Hence, common support holds!

Discussion (Contd.)

Can we verify SO as well?

- No! SO is a restriction on the joint of (Y, W, X, U)...
- ... and the sampling process does not reveal anything about U.

But as discussed, we can conduct a balance test to assess plausibility.

- Let $\tilde{X}$ denote households' income which is included in the PSID;
- Under SO, 401(k) participation has no association with income conditional on having at least/less than 16 years of education.

Discussion (Contd.)

Assume the PSID data is a realization of an iid sample of $(Y, W, X, \tilde{X})$.

Computing the test statistic T_n given in Equation (16) results in

$$T_n \approx 695$$

Using Corollary 6, we can calculate the associated p-value to be

$$p\text{-value} \approx 0$$

On a 5% sgn. level, we reject H_0 of 0-valued CEF differences!

- ▷ On a 5% sgn. level, there is sufficient evidence to reject that for households with at least/less than 16 yrs of education, expected income is not associated with 401(k) participation.

Type I errors exist, but the evidence from the test seems convincing...

- ▷ ... and is of little surprise given the many remaining confounders!

Natural response: Condition on yrs of education & income.

- ▷ But we don't have estimators for continuous X... yet!

Summary

Today:

- ▷ Discussed the Roy model to understand selection;
- ▷ Introduced SO and CS assumptions;
- ▷ Proved identification of common causal parameters under SO & CS.

We're equipped for causal analysis under SO when W and X are discrete:

- ▷ Constructed and analyzed estimators for the CATE and ATE.

Many settings when binning estimators are ill-suited:

- ▷ Continuous policy variables W (e.g., student-teacher ratio);
- ▷ Continuous covariates X (e.g., income);
- ▷ Multiple covariates such that X is a vector;

In the next lecture, we introduce *multiple* linear regression to construct estimates of causal parameters under SO for non-discrete X.

References

Chernozhukov, V. and Hansen, C. (2004). The effects of 401 (k) participation on the wealth distribution: an instrumental quantile regression analysis. *Review of Economics and statistics*, 86(3):735–751.

Poterba, J. M., Venti, S. F., and Wise, D. A. (1994). 401 (k) plans and tax-deferred saving. *Studies in the Economics of Aging*, pages 105–142.

Poterba, J. M., Venti, S. F., and Wise, D. A. (1995). Do 401 (k) contributions crowd out other personal saving? *Journal of Public Economics*, 58(1):1–32.