Lecture 2
Causal Inference

CLM: Asymptotic Assumptions

• Last semester, to get asymptotic results for OLS, we presented a new set of assumptions for the CLM:

(**A1**) DGP: $\boldsymbol{y} = \boldsymbol{X}\,\beta + \boldsymbol{\varepsilon}$.

(**A2'**) $\boldsymbol{X}$ stochastic, but $E[\boldsymbol{X}'\boldsymbol{\varepsilon}] = 0$ and $E[\boldsymbol{\varepsilon}] = \mathbf{0}$.

(**A3**) $Var[\boldsymbol{\varepsilon}\,|\,\boldsymbol{X}] = \sigma^2\,\mathbf{I}_T$

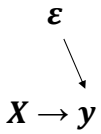
(**A4'**) plim $(\boldsymbol{X}'\boldsymbol{X}/T) = \mathbf{Q}$ (p.d. matrix with finite elements, rank=k)

• We studied the large sample properties of OLS:

- $\boldsymbol{b}$ and s^2 are consistent
- $\boldsymbol{b} \xrightarrow{a} N(\beta, (\sigma^2/T)\,\mathbf{Q}^{-1})$
- *t-tests* asymptotically N(0, 1), Wald tests asymptotically $\chi^2_{rank(S_T)}$ and *F-tests* asymptotically $\chi^2_{rank(Var[\boldsymbol{m}])}$.
- Small sample behavior: Do simulations and/or bootstrapping.

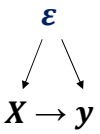
CLM: Causality

• The CLM implicitly models causation:



• $\boldsymbol{b}$ estimates a marginal effect: The change in $\boldsymbol{y}$ when $\boldsymbol{X}$ changes by a small amount (one unit).

• But, what happens when $\boldsymbol{\varepsilon}$ affects both $\boldsymbol{X}$ & $\boldsymbol{y}$? That is,



$\Rightarrow$ We have **endogeneity**: A violation of (**A2'**).

Endogeneity: The IV Problem

• We start with our CLM's DGP:

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}.$$

- Let's pre-multiplying the DGP by $\boldsymbol{X'}$

$$\boldsymbol{X'y} = \boldsymbol{X'X\beta} + \boldsymbol{X'\varepsilon}.$$

- We can interpret $\boldsymbol{b}$ as the solution obtained by first approximating $\boldsymbol{X'\varepsilon}$ by zero, and then solving the k equations in k unknowns

$$\boldsymbol{X'y} = \boldsymbol{X'X\,b} \qquad \text{(normal equations)}.$$

Note: What makes $\boldsymbol{b}$ consistent when $\frac{\boldsymbol{X'\varepsilon}}{T} \xrightarrow{p} \mathbf{0}$ is that approximating $(\frac{\boldsymbol{X'\varepsilon}}{T})$ by $\mathbf{0}$ is reasonably accurate in large samples.

• Now, we challenge this approximation. We relax (**A2'**) –i.e., $\{x_i, \varepsilon_i\}$ is **not** sequence of independent observations. That is,

$$plim(\frac{\boldsymbol{X'\varepsilon}}{T}) \neq \mathbf{0}. \qquad \Rightarrow \text{This is the IV Problem!}$$

Endogeneity: OLS is Inconsistent

• A correlation between $\boldsymbol{X}$ & $\boldsymbol{\varepsilon}$ is not rare in economics, especially in corporate finance, where endogeneity is pervasive.

Endogenous in econometrics: A variable is correlated with the error term, $\boldsymbol{\varepsilon}$.

• Q: What is the implication of the violation of $plim\left(\frac{\boldsymbol{X}'\boldsymbol{\varepsilon}}{T}\right) = 0$?

From the asymptotic CLM version, we keep (**A1**), (**A3**), and (**A4'**):
(**A1**) $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$.
(**A3**) $\text{Var}[\boldsymbol{\varepsilon}\,|\,\boldsymbol{X}] = \sigma^2\,\mathbf{I}_T$
(**A4'**) $plim(\frac{\boldsymbol{X}'\boldsymbol{X}}{T}) = \boldsymbol{Q}$

• Now, we assume (**A2"**) $plim(\frac{\boldsymbol{X}'\boldsymbol{\varepsilon}}{T}) \neq 0$.

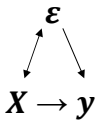
Endogeneity: OLS is Inconsistent

Then, $$\begin{aligned} plim\ \boldsymbol{b} &= plim\ \boldsymbol{\beta} + plim(\tfrac{\boldsymbol{X}'\boldsymbol{X}}{T})^{-1}\ plim\,(\tfrac{\boldsymbol{X}'\boldsymbol{\varepsilon}}{T}) \\ &= \boldsymbol{\beta} + \boldsymbol{Q}^{-1}\,plim\,(\tfrac{\boldsymbol{X}'\boldsymbol{\varepsilon}}{T}) \\ &\neq \boldsymbol{\beta} \end{aligned}$$

Under the new assumption, $\boldsymbol{b}$ is not a consistent estimator of $\boldsymbol{\beta}$.

Note: For finite samples, we could have challenged assumption (**A2**) $\text{E}[\boldsymbol{\varepsilon}|\,\boldsymbol{X}] = 0$. Then, $\text{Cov}(\boldsymbol{X}, \boldsymbol{\varepsilon}) \neq 0 \Rightarrow \text{E}[\boldsymbol{b}|\,\boldsymbol{X}] \neq \boldsymbol{\beta}$.

• Diagram with $\text{Cov}(\boldsymbol{X}, \boldsymbol{\varepsilon}) \neq 0$:



IV: Conditions for Instruments

• The solution to the endogeneity/IV problem, in the "traditional IV" literature, aims to "cure" the problem by finding l instruments $\boldsymbol{Z}$, such that the instruments used are both valid and relevant/informative.

• That is, we look for $\boldsymbol{Z}$ such that

(1) $\text{Cov}(\boldsymbol{X}, \boldsymbol{Z}) \neq \mathbf{0}$ - *relevance condition*

(2) $\text{Cov}(\boldsymbol{Z}, \boldsymbol{\varepsilon}) = \mathbf{0}$ - *valid condition (exclusion restriction)*

(1) ⇒ the $\text{Cov}(\boldsymbol{X}, \boldsymbol{Z})$ should be high enough to produce an $\widehat{\boldsymbol{X}}$ (from the first stage) that has relevant informative about $\boldsymbol{X}$.

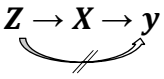
(2) In an omitted variables problem, we can think of (2) as broken into two parts:

(a) $\boldsymbol{Z}$ is uncorrelated to $\boldsymbol{\varepsilon}$

(b) only affects $\boldsymbol{y}$ through $\boldsymbol{X}$

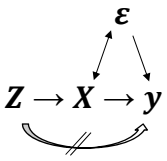
IV: Conditions for Instruments

• (2) ⇒ $\boldsymbol{Z}$ is not only uncorrelated to $\boldsymbol{\varepsilon}$, but only affects $\boldsymbol{y}$ through $\boldsymbol{X}$ – after all, it is excluded from structural equation!



From the 2nd part: Once I know the effect of $\boldsymbol{Z}$ on $\boldsymbol{X}$, I can throw $\boldsymbol{Z}$. Now, we have a framework to study "causality" in the presence of endogenous explanatory variables, $\boldsymbol{X}$.

• Combining assumptions:



Terminology: As seen later, in this lecture, $\boldsymbol{Z}$ will be structured as a binary variable ("treated" or "not-treated") and $\boldsymbol{y}$ as the outcome.

IV Estimation

• To get the IV estimator, we start from the system of equations:

$$\boldsymbol{W'Z'X}\, \boldsymbol{b}_{IV} = \boldsymbol{W'Z'y}$$

where $dim(\boldsymbol{W}) = l \times k$; $dim(\boldsymbol{Z}) = T \times l$; and $dim(\boldsymbol{X}) = T \times k$.

• Usual case in modern IV: # of instruments (l) = # of regressors (k)

$\Rightarrow dim(\boldsymbol{Z}) = dim(\boldsymbol{X}) = T \times k$ $\quad\quad$ $\Rightarrow \boldsymbol{Z'X}$ is a $k \times k$ pd matrix

- In this case, $\boldsymbol{W}$ is irrelevant, say, $\boldsymbol{W} = \mathbf{I}$. Then,

$$\boldsymbol{b}_{IV} = (\boldsymbol{Z'X})^{-1}\boldsymbol{Z'y}$$

Note: Let $\boldsymbol{Z} = \boldsymbol{X}$. Then,

$$\boldsymbol{b}_{IV} = \boldsymbol{b} = (\boldsymbol{X'X})^{-1}\boldsymbol{X'y}$$

That is, under the usual assumptions, $\boldsymbol{b}$ is an IV estimator with $\boldsymbol{X}$ as its own instrument.

IV Estimators: Asymptotic Properties

• Properties of $\boldsymbol{b}_{IV}$ (Under Lecture 8 assumptions):

(1) Consistent

$$plim(\boldsymbol{b}_{IV}) = \boldsymbol{\beta} + \boldsymbol{Q}_{ZX}^{-1}\, plim(\frac{\boldsymbol{Z'\varepsilon}}{T}) = \boldsymbol{\beta}$$

(2) Asymptotic normality (Using the Lindberg-Feller CLT)

$$\sqrt{T}\,(\boldsymbol{b}_{IV} - \boldsymbol{\beta}) \xrightarrow{d} N(\mathbf{0}, \sigma^2\, \boldsymbol{Q}_{ZX}^{-1}\, \boldsymbol{Q}_{ZZ}\, \boldsymbol{Q}_{XZ}^{-1})$$

• Properties of $\hat{\sigma}^2$ $\quad(\hat{\sigma}^2 = \boldsymbol{e}_{IV}'\boldsymbol{e}_{IV}/T)$

Consistent

$$plim(\hat{\sigma}^2) = plim(\boldsymbol{\varepsilon'\varepsilon}/T) - 2\, plim[(\boldsymbol{\varepsilon'X}/T)\,((\boldsymbol{Z'X}/T)^{-1}(\boldsymbol{Z'\varepsilon}/T)] + \\ + plim(\boldsymbol{\varepsilon'Z}\,(\boldsymbol{Z'X})^{-1}\,\boldsymbol{X'X}\,(\boldsymbol{Z'X})^{-1}\boldsymbol{Z'\varepsilon}/T) = \sigma^2.$$

• Then, $\quad$ Est Asy. Var$[\boldsymbol{b}_{IV}] = \hat{\sigma}^2\,(\boldsymbol{Z'X})^{-1}\boldsymbol{Z'Z}(\boldsymbol{Z'X})^{-1}$

IV Estimators: Example

Simplest case: Linear model, two endogenous variables, one IV.

$$\boldsymbol{y}_1 = \boldsymbol{y}_2\beta + \boldsymbol{\varepsilon} \qquad -\ \boldsymbol{\varepsilon} \sim \mathrm{N}(\mathbf{0}, \sigma_{\varepsilon\varepsilon})$$
$$\boldsymbol{y}_2 = \boldsymbol{z}\,\pi + \boldsymbol{v} \qquad -\ \boldsymbol{v} \sim \mathrm{N}(\mathbf{0}, \sigma_{VV})$$

with reduced form:

$$\boldsymbol{y}_1 = \boldsymbol{z}\,\pi\,\beta + \boldsymbol{v}\,\beta + \boldsymbol{\varepsilon}$$
$$= \boldsymbol{z}\,\gamma + \boldsymbol{\xi}.$$

The parameter of interest is β $(= \frac{\gamma}{\pi})$.

- We estimate β with IV: $b_{IV} = \frac{\frac{1}{T}\sum_i^T (y_{1,i} - \overline{y_1})(z_i - \bar{z})}{\frac{1}{T}\sum_i^T (y_{2,i} - \overline{y_2})(z_i - \bar{z})} = \frac{Cov(\boldsymbol{y}_1, \boldsymbol{z})}{Cov(\boldsymbol{y}_2, \boldsymbol{z})}$

- There is a special case: When the instrument is a dummy variable.

IV Estimators: Example – Special Case

- We estimate β with IV:

$$b_{IV} = \frac{\frac{1}{T}\sum_i^T (y_{1,i} - \overline{y_1})(z_i - \bar{z})}{\frac{1}{T}\sum_i^T (y_{2,i} - \overline{y_2})(z_i - \bar{z})} = \frac{Cov(\boldsymbol{y}_1, \boldsymbol{z})}{Cov(\boldsymbol{y}_2, \boldsymbol{z})}$$

- Special case: z_i is a dummy variable. Then,

$$\mathrm{E}[y_{1,i}|Z_i = 1] = \beta\ \mathrm{E}[y_{2,i}|Z_i = 1]$$
$$\mathrm{E}[y_{1,i}|Z_i = 0] = \beta\ \mathrm{E}[y_{2,i}|Z_i = 0]$$
$$\Rightarrow \mathrm{E}[y_{1,i}|Z_i = 1] - \mathrm{E}[y_{1,i}|Z_i = 0] = \beta\ (\mathrm{E}[y_{2,i}|Z_i = 1] - \mathrm{E}[y_{2,i}|Z_i = 1])$$

- Rearranging, we estimate β with:

$$b_{IV} = \frac{\overline{y_1}(Z_i{=}1) - \overline{y_1}(Z_i{=}0)}{\overline{y_2}(Z_i{=}1) - \overline{y_2}(Z_i{=}0)}$$

which is the **Wald estimator**.

IV and Causal Inference

• Finding a good IV is very difficult. The last three decades have seen a different perspective of IV, emphasizing **causal inference**, in the context of randomized experiments. The instrument reflects an "experiment," with **treated** and **non-treated** units.

• The traditional IV setting recognizes that units (individuals, firms) actively influence or select the level of the treatment they receive. Thus, units receiving the treatment are different from those receiving the "non-treatment." This makes the treatment potentially **endogenous**, creating **selection problems** -Heckman (1979).

Example: Individual i optimally selects her level of education (say, H or L). Individuals choosing higher levels of education may have higher skills, which lead to higher wages (w) for given levels of education.

$\Rightarrow$ Difficult to compute returns to education.

Causal Inference: RCT

• It would be easier to measure returns to education if the level of education can be seen as **randomly assigned**.

• In a **Randomized controlled trial** (**RCT**) a researcher randomly assigns participating units to treatment or control groups to yield balanced (unbiased) comparisons.

• Since the mean of a random sample from the population of units is an unbiased estimator for the mean of the population, a researcher can compare the means of the treated and control groups to gauge the effect of treatment in the population.

• A **RCT** is considered the **gold standard** in experimental research because it eliminates selection problems.

Causal Inference: RCT – Fundamental Problem

• The **experimental design approach** we will review is based on Rubin (1980) and Neyman (1923), known as **Rubin causal model** (**RCM**). It is based on the idea of **potential outcomes**.

• We want to study the effect of an MBA (the *treatment*) on the stock trading (*outcome*) of an individual. To study the effect of an MBA on stock trading, we need to compare two outcomes: the stock trading activity of the same individual (unit) with and without an MBA.

• The comparison is impossible to make.

• This is the **fundamental** problem of causal inference. At the unit level, causality is impossible to compute. This is where randomization helps. It will allow a researcher to compute an "**average treatment effect**" over a sample of treated and non-treated units.

Causal Inference: RCT – Randomization

• Randomization is the key to isolate causal effects.

Example: In medical RCTs, a drug and a placebo may be randomly given to patients, who do not know what they are taking. The researcher in charge is also uninformed regarding the drug-placebo assignment. This is called "**double blinding**."

Patients are expected to be alike (and behave similarly). If individuals comply with the random assignment, we can infer the average causal effect of the drug on a health outcome for the population in the experiment.

Note: If patients self-select into treatment or a doctor selects the best candidates into treatment, the sample is **not** a random sample. This is a big problem.

Causal Inference: RCT – ATE

• Under some assumptions –i.e., "an ideal experiment"–, an RCT provides an unbiased estimate of the **average treatment effect** (**ATE**). Also, ideally, the results can be generalized to other populations and settings.

• But, poorly designed RCTs can create biased results –i.e., ATE may not be well estimated. Researchers should convince their audience that the RCT is well-designed:

1) Clear definition of **Hypothesis** and **Experiment**.
2) Understanding **potential outcomes** –i.e., with no "surprises."
3) Understanding of **Assignment Mechanism**.
4) Define **covariates** (which ones to control, which ones to avoid).
5) Parameter(s) to be estimated to compute a **treatment effect**.

Causal Inference: RCT – Terminology

• Basic Concepts/Terminology:

- **Unit**: A physical object, i, in the sample at a particular point in time.

- **Treatment**: The intervention. This creates the dummy variable, Z_i:

$Z_i = 1$ (i treated),

$Z_i = 0$ (i control).

We want to assess the effect on the units relative to no intervention.

- **Potential outcomes**: The values to be observed of a unit's measurements of interest after treatment and non-treatment. For example, y_i(treated, $Z_i = 1$) & y_i(control, $Z_i = 0$).

- **Causal Effect** or **TE:** For each unit, we compare the potential outcome under both scenarios (treatment & non-treatment):

$$TE_i = y_i(Z_i = 1) - y_i(Z_i = 0).$$

Causal Inference: Treatment Effect (TE)

- **Causal Effect** or **TE** (Treatment effect):

$$TE_i = y_i(Z_i = 1) - y_i(Z_i = 0).$$

Remark: We define causality in terms of "potential outcomes" for each unit i. The parameter of interest ("**estimand**") is a "*within*" quantity. Of course, as defined, TE_i cannot be computed: we only observe one outcome for each unit i.

• Note: Different i's may have different TE_i. Heterogeneity is likely.

• Using the definition of Z_i, we decompose observed outcome as:

$$\begin{aligned} y_i &= y_i(1) * Z_i + y_i(0) * (1 - Z_i) \quad (\text{"}\textit{switching equation}\text{"}) \\ &= y_i(0) + [y_i(1) - y_i(0)] * Z_i \\ &= \text{baseline} + TE_i * Z_i \end{aligned}$$

Causal Inference: ATE, ATT & ATU

The **Average Treatment Effect** (**ATE**) is the population average over all i:

$$\textbf{ATE} = \mathrm{E}[\delta] = \mathrm{E}[y(1) - y(0)] = \mathrm{E}[y(1)] - \mathrm{E}[y(0)].$$

Similarly, we can also define an ATE on the Treated (**ATT**) and an ATE on the Untreated (**ATU**):

$$\textbf{ATT} = \mathrm{E}[\delta | Z_i = 1] = \mathrm{E}[y(1) | Z_i = 1] - \mathrm{E}[y(0) | Z_i = 1]$$
$$\textbf{ATU} = \mathrm{E}[\delta | Z_i = 0] = \mathrm{E}[y(1) | Z_i = 0] - \mathrm{E}[y(0) | Z_i = 0].$$

• Fundamental Problem of Causal Inference: We only observe one outcome, either $y_i(1)$ or $y_i(0)$. The "**counterfactuals**," say, $y_i(0 | Z_i = 1)$, are always missing. Thus, ATE, ATT & ATU cannot be computed. We can think of this problem as a "**missing data problem**" (or just as an **imputation** of the counterfactual problem).

Causal Inference: Sample Causal Effects

• We compare (observed) outcomes –in the MBA-stock trading example, y_i(MBA, $Z_i = 1$) & y_i(no-MBA, $Z_i = 0$) on a subset of units, the i's. For example,

$$\text{Average} = \frac{1}{N}\sum_i^N (y_i(1) - y_i(0)).$$

• If we have some information about unit i, the "**covariates**," then, we can look at the effects by covariates. That is, we allow for **heterogeneous effects**. For example, suppose we have x_i = Undergraduate Major (Social Sciences, Business, Humanities, etc.), then

$$\text{Average for Humanities} = \frac{1}{N_{Hum}}\sum_i^{N_{Huml}} (y_i(1) - y_i(0))$$

Causal Inference: ATE, ATT & ATU - Example

Example: A taxi driver may receive a loan ("treatment") from a bank, with the idea that the driver will buy a car and, thus, increase income ("potential outcome") and generate more business with the bank. If the taxi driver receives a loan, the taxi driver becomes the treated unit. Suppose the bank has the following ideal data:

Unit	No loan	Loan	Diff
1	10	8	-2
2	3	3	0
3	6	10	4
4	9	10	1
5	12	6	-6
6	9	11	2
7	6	4	-2
Average	7.8571	7.4286	**-0.375**

The bank computes ATE = **-0.375**. That is, if every taxi drives receives a bank loan, on average, the bank loan reduces income by **USD 375**.

Causal Inference: ATE, ATT & ATU - Example

Example (continuation): Suppose the first 3 taxi drivers receive a loan –i.e., the "treated." Then, the bank can compute ATT & ATU:

ATT = (-2 + 0 + 4)/3 = **0.6667**.

ATU = (1 – 6 + 2 + -2)/4 = **-1.25**.

But, in practice, these computations are impossible: The bank only observes one outcome for each taxi driver i.

• Missing counterfactuals: $y_i(0 \,|\, \text{loan}, Z_i = 1)$ & $y_i(1 \,|\, \text{no loan}, Z_i = 0)$.

Note: Mechanically, ATE is a weighted average of ATT and ATU, where the weights are given by the share of units in the treated (π) and non-treated ($1 - \pi$):

$$\text{ATE} = \pi * \text{ATT} + (1 - \pi) * \text{ATU}$$

Causal Inference: ATE, ATT & ATU - Example

• Interpretation:

ATE: It measures the average effect of treatment across all units –i.e., giving a bank loan to every taxi driver.

ATT: It measures the average effect of treatment on the units already treated –i.e., the effect of a loan on the taxi drivers who received the loan. A useful quantity if the bank plans to remove the loan program.

ATU: It measures the average effect of treatment on units untreated – i.e., the effect of a loan on the taxi drivers who have not received the loan. A useful quantity if the bank plans to expand the loan program.

See Greifer and Stuart (2023) for a discussion and applications of these effects.

Causal Inference: SDO

• We observe mean outcomes for the treated ($E[y(1)\,|Z_i = 1]$) and the untreated $E[y(0)\,|Z_i = 0]$. After some algebra, we compute **SDO** (Simple difference in mean outcomes) as:

$$\begin{aligned}\text{SDO} &= E[y(1)\,|Z_i = 1] - E[y(0)\,|Z_i = 0] \\ &= E[y(1)] - E[y(0)] + \\ &+ E[y(0)\,|Z_i = 1] - E[y(0)\,|Z_i = 0] + \\ &+ (1 - \pi)\,(\text{ATT} - \text{ATU})\end{aligned}$$

where π is the share of units in the treatment group.

• Interpretation of terms:

- ATE = $E[y(1)] - E[y(0)]$
- Selection Bias: $E[y(0)\,|Z_i = 1] - E[y(0)\,|Z_i = 0]$
- Heterogeneous Treatment Effect Bias: $(1 - \pi)\,(\text{ATT} - \text{ATU})$:

Causal Inference: SDO & Selection Bias

• Selection Bias: $E[y(0)\,|Z_i = 1] - E[y(0)\,|Z_i = 0]$

It measures the difference in outcome between the treatment and control units when neither is treated.

Example: We are interested in the effect of an MBA in stock trading. We get a sample of UH graduates with an MBA and without an MBA. A simple comparison of the average trading of the MBA group (the *treatment group*) with the without MBA group (the *control group*) may overestimate the effect.

This bias arises from the different characteristics between the two groups that are confounding the treatment effect. Very likely the MBA group would have traded more than the non-MBA group, even without getting an MBA --i.e., without treatment:

$$E[y(0)\,|Z_i = 1] > E[y(0)\,|Z_i = 0]$$

Causal Inference: SDO & Selection Bias

• Selection bias in causal inference is when one or both mean potential outcomes differ by treatment status.

• The **source of the bias** is caused by the why people get treated, or the **treatment assignment mechanism**.

• Usually, the selection bias is addressed by:

(1) **Modeling it** directly and remove it (Heckman).

(2) **By design**: Experiments, randomization (Rubin).

• We will deal with (1) later on in the semester. In this class, we focus on (2).

Causal Inference: SUTVA & Assignments

• More Concepts:

- **SUTVA** (Stable Unit Treatment Value Assumption): There is no interference among units –i.e., each unit's potential outcomes are independent of what happens to other units. In addition, for each unit, there is only one form of treatment and non-treatment.

Example: A taxi driver buys a car with the loan if another driver does not buy a car with the loan. This is a violation of SUTVA.

- **Assignment Mechanism**: The rules that determine which units receive treatment and which receive control.

Example: A loan officer evaluates a taxi driver personality and gives a loan to the one that she believes will make better use of the loan –i.e., increase income.

Causal Inference: Assignment Mechanism

• To draw conclusions from the causal effects, we need to understand the assignment mechanism of treatment.

Example: Recall the potential outcomes from the loan experiment are (in USD 1,000):

Unit	No loan	Loan	Diff
1	10	8	-2
2	3	3	0
3	6	10	4
4	9	10	1
5	12	6	-6
6	9	11	2
7	6	4	-2
	7.8571	7.4286	**-0.375**

ATE = **-0.375** ⇒ If all taxi drivers receives the loan (the treatment), their income would get reduced by **USD 375**.

Causal Inference: Assignment Mechanism

Example: Suppose the loan officer ("Clairvoyant Loan Officer") gives the loan to the taxi driver that is going to use it in a positive way:

Unit	Z (assignment)	No loan $y_i(0)$	Loan $y_i(1)$
1	0	10	x
2	0	x	x
3	1	x	10
4	1	x	10
5	0	12	x
6	1	x	11
7	0	6	x
		9.3333	10.3333

The observed average causal effect is **1** > **-0.375**.

Based on the observed difference in sample means, we infer that the treatment is very positive! But, this is wrong for 4 units.

Causal Inference: Randomization & ATE

• We assume independence of treatment and potential outcomes.

Independence Assumption

Treatment is orthogonal to the population's potential outcomes. Thus,

$$\{y(1), y(0)\} \perp Z$$

(Or with covariates: $\{y(1), y(0)\} \perp Z|X$.

• This assumption is called **unconfoundedness** or **ignorability**: How a unit is assigned to treatment is irrelevant, given everything we know about that unit. Potential outcomes are **exchangeable**.

• Then, mean y, for the treatment and control group are the same:

$$\text{E}[y(1)\,|Z_i = 1] = \text{E}[y(1)\,|Z_i = 0].$$
$$\text{E}[y(0)\,|Z_i = 1] = \text{E}[y(0)\,|Z_i = 0].$$

Causal Inference: Randomization & ATE

• Implications for ATE. Let's look at the SDO decomposition:

SDO = ATE + Selection Bias + Heterogeneous Treatment Effect Bias

• Now, under independence assumption:

- Selection Bias: $\text{E}[y(0)\,|Z_i = 1] - \text{E}[y(0)\,|Z_i = 0] = 0$.

Heterogeneous Treatment Effect Bias: $(1 - \pi)$ (ATT – ATU) = 0.

Check last term:

$$\text{ATT} - \text{ATU} = (\text{E}[y(1)\,|Z_i = 1] - \text{E}[y(0)\,|Z_i = 1]) - (\text{E}[y(1)\,|Z_i = 0] - \text{E}[y(0)\,|Z_i = 0]) = 0$$

Then, with randomized treatment assignment, the SDO is an unbiased estimator of ATE.

Causal Inference: Randomization & Regression

• We can also estimate ATE using a regression.

• We assume that treatment effects are constant: $\delta = y_i(1) - y_i(0)$.

Then, after some algebra, we get:

$$\begin{aligned} y_i &= y_i(1)\, Z_i + (1 - Z_i)\, y_i(0) \\ &= y_i(0) + (y_i(1) - y_i(0))\, Z_i \\ &= y_i(0) + \delta\, Z_i \\ &= y_i(0) + \delta\, Z_i + E[y_i(0)] - E[y_i(0)] \\ &= E[y_i(0)] + \delta\, Z_i + \varepsilon_i \\ &= \alpha + \delta\, Z_i + \varepsilon_i \end{aligned}$$

where ε_i is the error component of $y_i(0)$. This is what we presented last semester, for example in Card's (1990) paper. This framework allows easy incorporation of covariates ("control variables").

Causal Inference: Randomization & Regression

• Under the usual conditions, we have

- Consistency

$$\hat{\delta} \xrightarrow{p} ATE$$

- Asymptotic Normality

$$\sqrt{n}\,(\hat{\delta} - ATE) \xrightarrow{d} N(0,\ V)$$

where V is the variance, which can be estimated as usual or with a bootstrap (though, in some matching cases a bootstrap will not work).

• C.I. can be estimated as usual. For SE, it is common to use:

$$\text{SE}\,[\hat{\delta}] = \text{sqrt}[\,\frac{s_1^2}{n_1} + \frac{s_0^2}{n_0}\,]$$

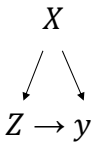
where s_j^2 & n_j are the sample variance & number of units of group j.

Causal Inference: Confounders

• So far, we have only paid attention to y (outcome) & Z (treatment), implicitly, in a directional way:

$$Z \rightarrow y.$$

• Now, we incorporate information about unit i, the covariates, x_i, into our analysis. The relevant covariates are the "**confounders**," the i variables that affect both Z & y. In graphic form:



• The confounders are a source of selection bias. We will use X to get "**stratified**" conditional estimates, say, by gender or income brackets, and then, take a weighted average.

Causal Inference: Matching

• Matching allows us to compute counterfactuals: A treated unit is "matched" to a comparison unit that is identical on known and quantified relevant confounders.

Example: A taxi driver that receives a loan will be matched with another that didn't receive a loan. Matching is based on the k relevant covariates: age, experience, income, family members, and city.

• Q: If we can use a regression to estimate ATE, why matching?

A: Two main reasons:

- Matching does not assume a functional form. If functional form is incorrect, the estimates are inconsistent.

- Matching only uses the untreated units "similar" (in covariates) to the treated. In a regression, all the untreated individuals are used to estimate the counterfactual for a given treated unit.

Causal Inference: Exact Matching

• If we can find an exact counterpart, we have **exact matching**, otherwise, we have **inexact matching**.

• Suppose we can do **exact matching**, that is, in the previous example, we find a non-treated taxi driver with exactly the same covariates as treated taxi driver i. The matched non-treated driver has outcome $y_{j(i)}(0)$. This is the **counterfactual** for unit i.

Under exact matching we can estimate the treatment effect for i:

$$\widehat{\delta_i} = y_i(1) - y_{j(i)}(0)$$

• Now, by matching all treated units in the experiment, we can compute the ATT.

Causal Inference: Exact Matching

• We do the exact matching for every treated taxi driver in the sample (suppose we have N_T treated taxi drivers). Then, the ATT estimate:

$$\widehat{ATT} = \frac{1}{N_T}\sum_i^{N_T}(y_i(1) - y_{j(i)}(0))$$

where $y_{j(i)}$ is the outcome of the j-th unit matched to the i-th unit based on exact matching.

• If we find more than one unit with identical covariates, say M matching units, then we use an average of the matching units to estimate ATT:

$$\widehat{ATT} = \frac{1}{N_T}\sum_i^{N_T}(y_i(1) - [\frac{1}{M}\sum_i^{M} y_{jm(i)}(0)])$$

Note: When matching we do not use all the data, only the relevant data that allows us to estimate ATT.

Causal Inference: Exact Matching Issues

• **Continuous covariates**

In practice, exact matching with continuous covariates, X_i, is almost impossible. We have to, somewhat, discretize the data. Usually, we rely on ranges of the X_i. We call this process *coarsening* or *binning* of X_i.

Example: We match on income if between \$25,000 and \$27,500.

• When we do binning, the matching is called **coarsened exact matching** (**CEM**).

• **Common support**

To do exact matching, we need **common support**, that is, for each value of X_i, there exist an observation in the treated group and in the control group. If there is no common support, there are units without matches, and TE cannot be estimated.

Causal Inference: Inexact Matching

• It is not easy to find exact matches for all treated units. One solution is to use the "**closest**" value among all of the control observations.

• We can use the Euclidean distance or the Manhattan distance to define closeness. It is common to use normalized distances to make the metric invariant to the units of the covariates. For example, the **Mahalanobis distance**:

$$||\boldsymbol{X}_i - \mathbf{X}_j|| = sqrt\left((\boldsymbol{X}_i - \mathbf{X}_j)'\widehat{\boldsymbol{V}}^{-1}(\boldsymbol{X}_i - \mathbf{X}_j)\right)$$

where $\widehat{\boldsymbol{V}}$ is the variance-covariance matrix of $\boldsymbol{X}$.

• As k increases, we run into the "**curse of dimensionality**." Suppose $k = 6$ and each X_k is binary, then, we have 64 ($=2^6$) possible combinations to match. Not easy. Usual advice, be greedy. Match only using the variables that make sense to control for, no more, no less.

Causal Inference: Matching Bias

• Matching biases will appear in the estimates, though asymptotically, matching estimators tend to be consistent.

• A Matching bias arises because of the effect of matching discrepancies between $\mu^0(X_i)$ and $\mu^0(X_{j(i)})$. Lack of **common support** increases as k increases.

• We can minimize matching discrepancies by:

- Using a small M (say, $M = 1$). Larger values of M produce large matching discrepancies.
- Matching with replacement. Because matching with replacement can use untreated units as a match more than once, matching with replacement produces smaller matching discrepancies.
- Trying to match better covariates with a large effect on μ^0.

Causal Inference: Matching Bias & Corrections

• We know the data. Then, we observe matching discrepancies. If discrepancies are large, it is always possible to use bias correction techniques. Abadie and Imbens (2011) propose:

$$\widehat{\delta_{ATT}}^{BC} = \frac{1}{N_T}\sum_i^{N_T}[(y_i(1) - y_{j(i)}(0) - (\widehat{\mu^0}(X_i) - \widehat{\mu^0}(X_{j(i)})]$$

where $\widehat{\mu^0}(X_i)$ is an estimate of $\mathrm{E}[y(0)|X = x,\ Z_i = 0]$, say, using OLS.

• For example, under unconfoundedness:

$$\mathrm{E}[y(0)\ |X = x,\ Z_i = 1] = \mathrm{E}[y(0)\ |X = x,\ Z_i = 0].$$

That is, the relationship between $y(0)$ & X in the control group is the same in the treatment group, even though we do not observe $y(0)$ in the treatment group.

• Then, fit a regression of $y(0)$ against X in the $Z = 0$ group only to predict $y(0)$ in both groups.

Causal Inference: Matching & Propensity Score

• Rosenbaum and Rubin (1983) propose an alternative to the distance methods: Use a "**propensity score (PS)**" to summarize all the covariates into a single scalar and then weight or match on it.

Definition: **Propensity score**

A unit specific probability bounded between 0 and 1 measuring the probability of treatment assignment given a set of covariates:

$$p(X) = P(Z = 1|\, X).$$

• Think of PS as a useful dimension reduction tool. The definition is not specific about how to compute $p(X)$ (logit? Non-parametric?) or what variables are in X. In practice, X should only include variables that affect both treated & non-treated (*confounders*).

Causal Inference: Matching & Propensity Score

• Nice result: Under full unconfoundedness, the PS is a sufficient statistic for X.

• The propensity score is the only covariate needed to achieve (conditional) independence:

$$y(1), y(0) \perp Z|p(X)$$

• As an implication, the distribution of the covariates X should be the same for units with the same PS, regardless of their treatment status:

$$P[X|Z_i = 1, p(X)] = P[X|Z_i = 0, p(X)]$$

• That is, two groups with the same probability of participation will show up in the treated and untreated samples in equal proportions. They can be combined for purposes of comparison. This is the **balancing property** of propensity scores.

Causal Inference: Matching & Propensity Score

• There are many tests of the balancing property. Usually, we look at bins of our data with similar PS and check if there are significant differences in their covariates.

• We still need common support, which should be easier to achieve than in the case with several covariates.

• Usually, we check common support by looking for an overlap of the histograms of PS for treated and non-treated units. For ATE, the distributions should completely overlap. For ATT, we only need a partial overlap (the treated distribution should be a subset of the non-treated distribution).

Causal Inference: Matching & Propensity Score

• A popular use of PS is to use the inverse to weight observations. This is called **inverse probability of treatment weighting** (**IPTW**).

• With IPTW, we give a higher weight to the units treated that had a low PS –i.e., a priori, a unit unlikely to be treated – and, viceversa, a lower weight to the units with higher PS.

• To compute ATE, we compute the difference of the weighted groups, to compute ATT, we only weight the treated group.

Causal Inference: Matching – Remarks

• Matching is a useful tool that avoids the usual assumptions behind regression models, for example, a functional form.

• Matching requires that the researcher observes and uses all the variables that affect both participation and outcomes.

• Matching can be interpreted as reorganizing the data from an observational study in such a way that the assumptions from a randomized experiment hold, at least approximately.

• Matching may be inexact, systematic differences in pre-exposure variables across the matched pairs may remain but can be subsequently be adjusted in the analysis stage.

Natural Experiments

• RCTs cannot always be done; they may be expensive, unfeasible or, just, unethical. In lieu of RCTs, we use events that can produce a random separation of units: some treated and some non-treated (control).

• We call non-RCT methods used to study causal relationships, **quasi-experimental**. A typical example is a **natural experiment**.

• A **natural experiment** is an event or a situation, not under the control of the units under study, which generates variation in the variable of interest that is as if it had been randomly assigned.

• A natural experiment is exogenous to a structural model. Like the previous external examples, the exclusion condition is met.

Natural Experiments & IV Estimation

• "Natural" points out that the researcher did not design the episode to be analyzed, but, the episode can be used to identify causality.

Examples: Weather shocks, policy changes, administrative rules, immigration shocks, birth dates, birth weight, etc.

• **IV framework**: Find a natural experiment, defining $\boldsymbol{Z}$, correlated with $\boldsymbol{X}$, but with no direct effect on $\boldsymbol{y}$ –the impact on $\boldsymbol{y}$ is through $\boldsymbol{X}$.

$\boldsymbol{Z}$ is an exogenous event $\Rightarrow$ resulting values of $\boldsymbol{X}$ induced by $\boldsymbol{Z}$ may be considered **randomized**, like in an RCT.

Key feature: **Randomization of treatment**. We need to show that the two groups are comparable along all dimensions relevant for the outcome variable (age, gender, previous health, etc.) except treatment.

Natural Experiments: RCT Substitution

• Recall that in the CEO compensation model, we want to test the causal effect of networking on compensation, but an omitted variable – the CEO's unobserved skills– creates endogeneity.

• A solution to the omitted variables problem is to assign networking ($\boldsymbol{x}$) randomly: we have two similar groups of CEOs (with similar skills!) & randomly assign them values (say, large network & small network).

• Of course, this randomized experiment is not possible.

• But, we can look for a natural event, $\boldsymbol{Z}$, unrelated to CEO compensation, which randomly assigns networking, $\boldsymbol{x}$, to two groups (immigration? lack of U.S. education?). Then, we can test causality, without the endogeneity problem.

Natural Experiments: Steps

• Steps of a natural experiment:

(1) Experiment defines an IV: $Z_i = 1$ (i treated),
$Z_i = 0$ (i control).

(2) Identify two groups:
– treated (all i with $Z_i = 1$) with observations: $\boldsymbol{y}(1), \boldsymbol{X}(1)$
– control (all i with $Z_i = 0$) with observations: $\boldsymbol{y}(0), \boldsymbol{X}(0)$

(3) We analyze differences between $(\boldsymbol{y}(1), \boldsymbol{X}(1))$ & $(\boldsymbol{y}(0), \boldsymbol{X}(0))$.

Note: Like in RCT, it is impossible to observe $y_i(1)$ & $y_i(0)$ for the same individual i. Thus, we focus on the **ATE** between groups, which under some assumptions can be estimated by SDO. Recall:

$$\textbf{ATE} = \mathrm{E}[\boldsymbol{y}(1)] - \mathrm{E}[\boldsymbol{y}(0)].$$

Natural Experiments: Steps

• Remarks: **Steps 1-3** can be treated like a lab experiment if we show that the treatment is in fact randomly assigned. We need to show that two groups are comparable except for the treatment.

• This is the key for the experiment to be valid. We need to convince the audience that we have a **quasi-random** treatment.

• Heterogeneity. With heterogeneous treatment effects, ATE may vary with different groups: men/women, immigrants/non-immigrants, etc.

• Typical problem: Selection bias

- Individual i selects treatment or not.

- Treatment is assigned to the individual i with the highest chance of being successful (assignment is not independent of potential outcome).

Natural Experiments: Selection Bias

• Treatment should affect the outcome. If treated unit i would have achieved desired outcome without treatment, we have a biased ATE.

• The problem? Treatment is not independent of outcome. That is,

$$\mathrm{E}[\boldsymbol{y}(j)\,|Z_i = 1] \neq \mathrm{E}[\boldsymbol{y}(j)|Z_i = 0] \qquad j = 0, 1$$

Example: Companies pay to show ads after an individual conducts an internet search for a particular product. Consumers who click on the ads ($Z_i = 1$) are likely informed and with high likelihood would have found the product independent of the ad. That is,

$$\mathrm{E}[\boldsymbol{y}(0)|\ Z_i = 1] \neq \mathrm{E}\ [\boldsymbol{y}(0)|\ Z_i = 0]$$

Blake, Nosko and Tadelis (Econometrica, 2015) find that non-experimental measures of returns on paid search ads are huge (over 1600%), but experimental measures of returns are close to zero.

Natural Experiments: Heterogeneity

Example: Angrist (1990) use the Vietnam-era draft lottery, a randomized draw of birth dates, to estimate the effect of military service on earnings later in life.

The draft (natural experiment) is the instrument for military service. But, most of the individuals who served in Vietnam were volunteers. Thus, the draft lottery only affected individuals who would not have served voluntarily in the military.

Implication: ATE is likely not representative of those who volunteered for service in the Vietnam War.

• In most situations, responses are **heterogeneous**. When treatment effects vary across individuals and people make choices, there is likely to be **incomplete compliance** with the (natural) experiment.

Natural Experiments: Compliance

A&K (1991) found a causal return to schooling that is between 8% and 10%, a higher return than the OLS implied return. But, this is not the end of the story.

• The quasi-experimental variation, produced by the Quarter of Birth, mainly affected those with a high probability of dropping out of school as soon as possible. It may well be that the returns to schooling in this part of the population are not representative of the overall population.

• In the language of controlled experiments, those who were unaffected by the natural experiment are "**non-compliers**," and their returns to schooling are potentially different than among the "**compliers**," because of heterogeneous treatment effects..

Natural Experiments: Compliance

• Compliance terminology:

- **Compliers**: The units whose behaviors are affected by instrument as expected –i.e., late birth data ⇒ finish HS.

- **Defiers**: The units whose behaviors are affected by instrument in an unexpected way –i.e., late birth data ⇒ HS drop out.

- **Always takers** and **Never takers**: The units whose behaviors are not affected by instrument (finish/drop-out HS regardless of date of birth).

Example: In Angrist (1990), always-takers serve in the military no matter the lottery number (with, likely, no direct effect on earnings). Never-takers do not serve in the military no matter their lottery number. Very likely, there are no defiers.

Natural Experiments: Identifying ATE

• The combination of treatment heterogeneity and incomplete compliance poses a problem for causal analysis. Under these elements, ATE identification is complicated:

- Heterogeneity: Treatment effects vary by unit.
- Incomplete compliance: *Treatment status* ≠ *Treatment eligibility*.

• In general, treatment status depends on individual treatment effects. Thus, **additional assumptions** are needed to identify ATE:

- Units do not know effect of treatment, Heckman (1997).
- One-sided non-compliance –i.e., the probability of participating is zero for individuals who are not eligible for treatment, Bloom (1984).

• We think of these assumptions as strong assumptions.

Identifying Causal Effects

• Angrist and Imbens (1994) set a general framework to identify causal effects. The key is the link of the assignment mechanism to an instrument (defined by RCT or natural experiment).

Notation: We use A&K (1991) to set it:

y_i (potential outcomes) = Earnings

d_i (treatment indicator, in this case **binary**) = Finishing high school:

$d_i = 1$ if i finished HS

$d_i = 0$ if i did not finish HS

The instrument, Z_i, is the birth date:

$Z_i = 1$ if i was born July-Dec ("**late birth**")

$Z_i = 0$ if i was born Jan-Jun ("**early birth**").

Recall: We expect more education from late birthers.

Identifying Causal Effects

• Since the treatment indicator (completing HS or not) is endogenous, think of it in terms of potential outcomes, $y_i(d_i, Z_i)$, which are four:

$y_i(1,1)$: born late, finished HS

$y_i(1,0)$: born late, did not finish HS

$y_i(0,1)$: born early, finished HS

$y_i(1,0)$: born early, did not finish HS

• The instrument, Z, should affect y through the treatment, d:

$$Z \rightarrow d \rightarrow y$$

To make sure this happens, we impose restrictions on Z.

Identifying Causal Effects: Assumptions

• A good instrument, Z, should satisfy:

(1) Randomly assigned: $[\{y_i(d_i, Z_i)\ \forall\ d,z\},\ d_i(1), d_i(0)] \perp Z_i$

In the A&K example, earnings are independent of the dates on which individuals are born. Under this assumption, the regression of Y on Z (reduced form) identifies the causal effect of instrument on outcome:

$$E[y_i|\ Z_i{=}1] - E[y_i|\ Z_i{=}0] = E[y_i(1, d_i(1))|Z_i{=}1] - E[y_i(0, d_i(0))|Z_i{=}0]$$
$$= E[y_i(1, d_i(1))] - E[y_i(0, d_i(0))]$$
$$= \text{Causal effect of being born late on earnings}$$

(last step follows from random assignment). Similarly,

$$E[d_i|\ Z_i{=}1] - E[d_i|\ Z_i{=}0] = E[d_i(1)|\ Z_i = 1] - [d_i(0)|\ Z_i = 0]$$
$$= E[d_i(1)] - [d_i(0)]$$
$$= \text{Causal effect of being born late on likelihood of finishing HS}$$

Identifying Causal Effects: Assumptions

(2) Relevant: $E[d_i(1)] - E[d_i(0)] \neq 0$

We require a stronger response to the possibility of dropping out of high school among those who are born early relative to those born late.

(3) Exclusion: $y_i(1, d_i) = y_i(0, d_i) = y_i(d_i)$

We need this assumption since $Z \neq d$. The only way the instrument (birth date) affects the outcome is through the treatment.

Example: In Angrist's (1990) draft lottery paper, the exclusion restriction requires that potential earnings with and without military service be independent of the lottery number.

• Now, we can identify the effect of the treatment (finishing HS) on the outcome (earnings).

Identifying Causal Effects: Assumptions

• Notice that $\{d_i(1) - d_i(0)\}$ can take three values: **1**, **0**, **-1**. Then,

$E[y_i|\, Z_i{=}1] - E[y_i|\, Z_i{=}0] =$

$= \sum_k k P[d_i(1) - d_i(0) = k] * E[y_i(1) - y_i(0) \mid \{d_i(1) - d_i(0)\} = k]$

$= 1\, P[(d_i(1) - d_i(0) = 1] * E[y_i(1) - y_i(0)|\, \{d_i(1) - d_i(0)\} = 1]$

$-1\, P[(d_i(1) - d_i(0) = -1] * E[y_i(1) - y_i(0)|\{d_i(1) - d_i(0)\} = -1]$

$\Rightarrow$ units who do not respond to the instrument –with $d_i(1) = d_i(0)$– do not contribute to identification. Impossible to estimate a causal effect for the units that do not change behavior.

Note: The reduced form can be negative! Why? The treatment effect for those who shift from nonparticipation to participation (or from "0" to "1") when Z is switched from 0 to 1 ("*compliers*") can be cancelled out by the treatment effect of those who shift from 1 to 0 ("*defiers*").

$\Rightarrow$ we need more assumptions.

Identifying Causal Effects: Monotonicity

(4) Monotonicity: $d_i(1) - d_i(0) \geq 0$ (or ≤ 0)

Monotonicity requires that the instrument operates in the same direction on all individual units. Anyone affected by the instrument is affected *in the same direction* (positively or negatively, but not both).

Examples: In A&K, we have 4 cases:

- Compliers: $d_i(1) = 1,\ d_i(0) = 0 \Rightarrow d_i(1) > d_i(0)$
- **Defiers**: $d_i(1) = 0,\ d_i(0) = 1 \Rightarrow d_i(1) < d_i(0)$
- Always takers: $d_i(1) = 1,\ d_i(0) = 1 \Rightarrow d_i(1) = d_i(0)$
- Never takers: $d_i(1) = 0,\ d_i(0) = 0 \Rightarrow d_i(1) = d_i(0)$

Monotonicity requires no defiers. Recall that always takers and never-takers play no role in the computation of $E[y_i|\ Z_i{=}1] - E[y_i|\ Z_i{=}0]$.

Identifying Causal Effects: Monotonicity

Examples (continuation): In Angrist (1990) monotonicity requires that someone who would serve in the military with lottery number n would also serve in the military with lottery number m $(m > n)$, which is plausible.

Dobbie et al. (2018) used the detention tendencies of (quasi)-randomly assigned bail judges to estimate the causal effect of pretrial detention on subsequent outcomes. Monotonicity "requires that individuals released by a strict judge would also be released by a more lenient judge, and that individuals detained by a lenient judge would also be detained by a stricter judge," which may not be true.

• Monotonicity implies $P[(d_i(1) - d_i(0)) = -1] = 0$. Then, the causal effect of instrument on treatment equals:

$$E[d_i(1) - d_i(0)] = P[(d_i(1) - d_i(0)) = 1]$$

Identifying Causal Effects: IVE & ITT

• With assumptions **(1)-(4)**, the causal effect of instrument on outcome (reduced form estimation, also called **intention-to-treat** or **ITT**):

$$E[y_i|\, Z_i = 1] - E[y_i|\, Z_i = 0] =$$
$$= P[(d_i(1) - d_i(0) = 1] * E[y_i(1) - y_i(0)|\, \{d_i(1) - d_i(0)\} = 1]$$

$\Rightarrow$ LHS is the causal effect of Z on y (reduced form). First component of the RHS is the causal effect of Z on d (1st-stage).

• The ratio is the IV (**Wald**) estimator:

$$\frac{E[y_i|\, Z_i{=}1] - E[y_i|\, Z_i{=}0]}{E[d_i(1) - d_i(0)]} = E[y_i(1) - y_i(0)|\{d_i(1) - d_i(0)\} = 1]$$

This represents the average causal effect for the population that changed their treatment status in accordance with the change in the value of the instrument. This is an ATE for a "*local*" group: compliers.

Identifying Causal Effects: LATE

• Angrist and Imbens (1994) called this effect **local average treatment effect** (**LATE**). Since this estimator involves compliers, LATE is also called the **complier average causal effect** (**CACE**).

Example: In Angrist (1990), IV estimates the average effect of military service on earnings for the subpopulation who enrolled in military service because of the draft but would not have served otherwise.

LATE does not tell us what the causal effect of military service was for volunteers or those who were exempted from military service for medical reasons.

Remark: IV estimates the average causal effect for those units affected by the instrument (i.e., complier causal effects only).

$\Rightarrow$ LATE is not ATE.

Identifying Causal Effects: LATE

• For never-takers, the exclusion restriction might be problematic: going to College or moving abroad can have a direct impact on earnings.

• The existence of defiers in the sample biases the results. The bias can be big if we have a significant proportion of defiers. In Angrist (1990), very likely there are very few defiers.

Causal Effects & Experiments: Inference Issues

• Clustering

Following Moulton (1986), researchers cluster (or "group") structures in the data to analyze results. Recall that if units in the same group are exposed to the same variation, one should take the correlation across individuals within group into account.

This is a big concern for Difference-in-differences studies, which use variation across groups over time for identification. See Bertrand, et al. (2004), Donald and Lang (2007), and Hansen (2007).

• Weak Instruments

IV studies rely on strong instruments -Nelson & Startz (1990); Staiger & Stock (1997). But, as seen last semester, it is not easy to find them. If the instruments are weak, IV estimates can be severely biased. See Andrews et al. (2019), Keane and Neal (2021), and Young (2020).

Causal Effects & Experiments: Inference Issues

• **Heteroscedasticity**

Recall that Stock, Wright and Yogo (2002) suggest that an F-stat > 10 from the 1st-stage regression indicates that the instruments are not weak. This F-stat is called **non-robust F**, since it is not robust in the presence of heteroscedasticity.

Kleibergen and Paap (2006) use a **robust F-stat**, based on clustering SE and/or HAC SE.

Andrews, et al. (2019) suggest the Olea and Pflueger (2013) **"effective" 1st-stage F-stat**, which is equal to the *non-robust F* times a correction factor for non-homoscedasticity. With this effective F-stat, the rule of thumb of "effective F-stat" > 10 applies.

Causal Effects & Experiments: Inference Issues

• **Publication bias** (also called **p-hacking**)

Card and Krueger (1995) conducted a meta-analysis of the prior literature on the minimum wage and concluded that it suffered from publication bias.

• More recently, Brodeur, Lé, Sangnier, and Zylberberg (2016) suggested that there is publication bias because there is excess mass of estimates having a p-value just below 0.05 than just above 0.05.

• Brodeur, Cook, and Heyes (2020) focused in particular on the methods associated with the design-based approach. They concluded that p-hacking is more common for DiD and IV methods than for RCTs and RD designs. The last two methods are more "structured" and, thus, more difficult to "cheat."

Identifying Causal Effects: Generalizations

• We know our data. We can use it to gauge heterogeneity and analyze if there is a biased LATE. In many situations, heterogeneity may work in our favor. For example, if the policy-target is low-income individuals, and low-income individuals are overrepresented among compliers, then LATE should be highly relevant for policy purposes.

• With more than one instrument, we can test the heterogeneous responses to the instrument -Angrist, Lavy, and Schlosser, (2010). If heterogeneity is not substantial, LATE estimates may generalize.

• Alternatively, we can fill in the gaps in the data using auxiliary assumptions. For example, Heckman et al. (2001, 2003) and Angrist (2004) used parametric latent-index models to identify causal effects. Brinch et al. (2017) used a structural model. Chamberlain (2010) used Bayesian techniques.

Identifying Causal Effects: MTE

• To evaluate policy effects, many researchers move from estimating LATE to "**marginal treatment effects**" (*MTE*).

• The literature started with Heckman and Vytlacil (1999, 2001, 2005). The MTE is the expected effect of treatment conditional on covariates (observed and unobserved) at a margin (a **small change**).

• We compute MTE in the context of a structural (switching) model, with individuals shifting into (or out of) treatment by a *marginal change* in the instrument (cost of the treatment, propensity score, etc.).

Note: MTE can be done only with a continuous treatment, d_i. MTE is a *marginal effect*, we need a derivative! Thus, we model y_i as function of covariates, X_i, instrument, Z_i, and treatment intensity, d_i.

Identifying Causal Effects: MTE – Details

• Consider the selection rule for treatment, Z_i:

$$Z_i = 1 \quad \text{if } v_i \leq W_i'\gamma$$
$$= 0 \quad \text{otherwise}$$

where v_i is the unobservable variable that makes unit i jump into/out of treatment; a function of W_i, a vector of observable variables.

• The marginal unit i is the one with $v_i = W_i'\gamma$. Then,

$$MTE = E[y_i(1) - y_i(0) \mid v_i = W_i'\gamma]$$

• MTE: Gain from treatment for people who were shifted into (or out of) treatment status by a marginal change in the treatment intensity.

• When $y_i(1)$ & $y_i(0)$ are value outcomes, MTE is a willingness to pay measure –see Heckman and Vytlacil (2005).

Identifying Causal Effects: MTE – Details

• It is convenient to rewrite the selection rule as

$$Z_i = 1 \quad \text{if } u_i \leq F(W_i'\gamma)$$
$$= 0 \quad \text{otherwise}$$

where $u_i \sim U(0,1)$ and $F(.)$ is the distribution variable of V. Then, we write the probability of assignment to treatment (PS) as:

$$F(W_i'\gamma) = P(Z_i = 1|W_i) = P(W_i)$$ -propensity score (PS)

• We treat $P(W_i)$ as the (continuous) instrument intensity, d_i, to identify causality. We will take derivatives of the outcome equation with respect to $P(W_i)$, conditioned on covariates, X_i:

$$MTE(p,X) = \frac{\delta}{\delta p} E[y_i(0)| X_i, P(W_i) = p]$$

• When we average over all marginal individuals (integrating over p), we get the Average treatment effect, $ATE(X)$.

Identifying Causal Effects: MTE – Estimation

• We want to estimate:

$$MTE(p, X) = \frac{\delta}{\delta p} E[y_i(0) | X_i, P(W_i) = p]$$

• Usual steps:

- Estimate $P(W_i)$, for example, with a logit model.

$$P(W_i) = P(Z_i = 1 | W_i) = \frac{1}{1 + exp(W_i'\boldsymbol{\beta})}$$

- Estimate the regression of y_i as function of covariates, X_i, and $P(W_i)$. Usually, the estimation is done non-parametrically, with local linear regressions or polynomials.

- Take the first derivative with respect to $P(W_i)$.

Identifying Causal Effects: MTE – LATE

• Under some assumptions (conditional independence, monotonicity, additive separability of effects, etc.), ATE, ATT, ATU & LATE can be recovered using weighted averages over the MTE curve.

Example: We want to analyze the effect of a new program. Suppose that, for a given X, we have two values for the propensity score, $P(W)$: $p_0(X)$ & $p_1(X)$ (with $p_1 > p_0$).

The difference, $p_1 - p_0$, represents the policy induced change in the treatment probability for someone with characteristics X. Now, all the units with a PS $p_0 \leq u_i \leq p_1$ with switch into treatment. Then,

$$LATE(p, X) = \frac{1}{p_1 - p_0} \int_{p_0(X)}^{p_{1(X)}} MTE(p) dp$$

Remark: To get LATE, we integrate over the units that will switch into treatment because of the new program –i.e., the compliers.

Identifying Causal Effects: MTE – LATE

Example (continuation): Interpretation.

When an instrument increases the probability of treatment from $p_0(X) - p_1(X)$

- Individuals with $u \leq p_0(X)$ were already treated: always-takers.
- Individuals with $u \geq p_1(X)$ are still not treated: never-takers.
- Individuals with $p_0 \leq u_i \leq p_1$ switch treatment status: **compliers**.

• LATE is the average height of the MTE curve over this complier interval.

• Cornelissen et al. (2016) provides a review of the literature.

Causal Effects & Experiments: Criticism

• Deaton (2010) and Heckman and Urzua (2010) see an excessive (and inappropriate?) use of experimental methods in empirical work. Both focus their criticism on LATE. Heckman and Urzua (2010) say:

"Problems of identification and interpretation are swept under the rug and replaced by 'an effect' identified by IV that is often very difficult to interpret as an answer to an interesting economic question."

• Experimental studies emphasize "too much" credible identification, or on "**internal validity**" as opposed to "**external validity**."

• Imbens (2010) argues that it is useful to separate the assumptions needed to identify a causal effect in the sample studied from the assumptions needed to generalize an internally valid estimate to other populations. (See also DiNardo and Lee (2011).

Causal Effects & Experiments: Criticism

• It is possible to combine quasi-experimental variation and structural models. For example, Card and Hyslop (2005) used experimental variation to aid the identification of a structural model for welfare participation.

• Design-based estimates can also be used to validate structural models, as done by Blundell (2013). Related to this, Kline and Walters (2019) showed that IV and selection-correction type of estimates (Heckman (1979)) of LATE are numerically equivalent.

• Under some conditions, the choice between these two estimators is unimportant for estimating treatment effects that are identified in the data.

Causal Effects: Remarks

• RCTs are not easily conducted in finance and economics. We usually rely on observational data outside of our control.

• With a natural experiment, treatment or instrument assignment is as good as random, almost like an RCT.

• We would like to compute **ATE** = E[$y(1)$] – E[$y(0)$]. But, the effect identified is **LATE**, the average causal effect among compliers:

The causal effect for the subset of the population that changed behavior because of the value of the instrument.

• Angrist and Imbens (1994) set a general framework where LATE is identified. This framework has become the dominant one for both quasi-experimental and experimental work.

Causal Effects: Other Methods

• There are other quasi-experimental methods for causal inference. We use events that place units into a treated group and non-treated group:

- **Difference-in-differences design** (DiD) (Roth et al. (2023)). Similar to design framework, but we do not assume that units are the same.

- **Regression discontinuity design** (RD) (Hahn et al. (2001)). RD takes advantage of a jump in the likelihood of being treated, generated by an arbitrary threshold.

Regression kink design (RKD) (Nielsen et al. (2010) and Card, Lee, Pei, and Weber (2015)). RKD takes advantage of a change in the slope at the likelihood of being treated at a (kink) point.

Synthetic control method (Abadie (2021)). We replicate unobserved outcomes, $y_i(0)$, using observed outcomes from *donor* units $y_{j(i)}(0)$.

Causal Effects: Difference-in-Difference (DiD)

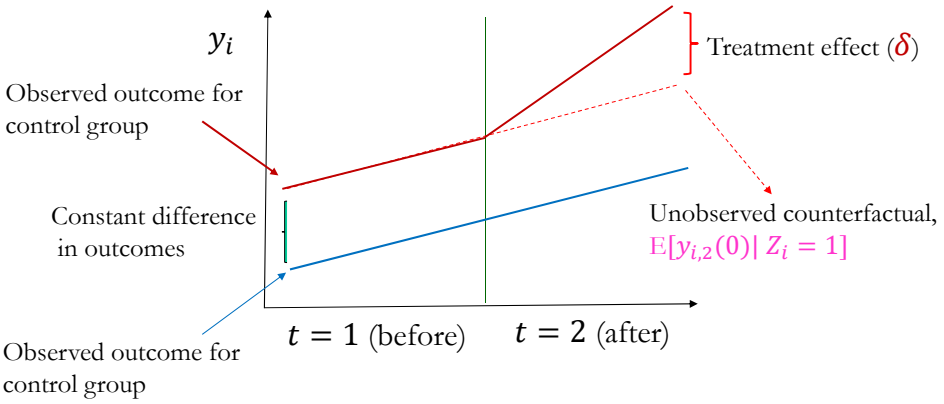
• **Difference-in-differences design** (DiD) (de Chaisemartin and D'Haultfoeuille (2020), Roth et al. (2023)). The canonical DiD model assumes two groups (treated & non-treated, or $Z_i = 1$ & $Z_i = 0$) and two periods (before & after treatment, or $t = 1$ & $t = 2$).

• In the RCM, we assume **exchangeability of outcomes** between treated and control groups. We use DiD when we believe there are differences between both groups. That is,

$$\mathrm{E}[y(1) \,|Z_i = 1] \neq \mathrm{E}[y(1) \,|Z_i = 0].$$
$$\mathrm{E}[y(0) \,|Z_i = 1] \neq \mathrm{E}[y(0) \,|Z_i = 0].$$

• Under DiD, we assume that in absence of treatment, the unobserved differences between treatment and control groups are the same overtime –i.e., under the two periods: **Trends are assumed parallel**.

Causal Effects: DiD – Parallel Trends (D1)



Remark: We formally state the parallel trend assumption as:

$$E[y_{i,2}(0) - y_{i,1}(0)|\ Z_i = 1] = E[y_{i,2}(0) - y_{i,1}(0)|\ Z_i = 0].$$

Any differences in outcomes observed before the treatment between the two groups are due to pre-existing differences.

Causal Effects: DiD – Assumptions

- The key identifying assumption is **parallel trends (D1)**:

$$E[y_{i,2}(0) - y_{i,1}(0)|\ Z_i = 1] = E[y_{i,2}(0) - y_{i,1}(0)|\ Z_i = 0].$$

- We also assume **no anticipation effects (D2)**, that is,

$$y_{i,1}(0) = y_{i,2}(0) \qquad \text{for all } i \text{ with } Z_i = 1.$$

- Then, we can estimate the treatment effect (δ):

$$\delta = E[y_{i,2}(1) - y_{i,1}(0)|\ Z_i = 1] - E[y_{i,2}(0) - y_{i,1}(0)|\ Z_i = 0],$$

using the observed sample means of $y_{i,2}$ & $y_{i,1}$.

Remarks: Different from RCM, there is only one unobserved counterfactual: $E[y_{i,2}(0)|\ Z_i = 1]$. We use assumption **(D1)** to estimate it.

Causal Effects: DiD – Regression

• We can also estimate δ with a regression, including covariates, $\boldsymbol{x}_i$:

$$\Delta y_{i,t} = y_{i,2} - y_{i,1} = \delta_0 + \delta\, Z_i + (\boldsymbol{x}_{i,2} - \boldsymbol{x}_{i,1})'\,\beta + \varepsilon_{i,t}$$

Note: This is a Pooled Model (actually, the CLM, since we work with differences, t plays no role): δ is consistent and asymptotically normal distributed (with N large & T fixed). We use Clustered SE for inference (the usual problems, discussed last semester, apply).

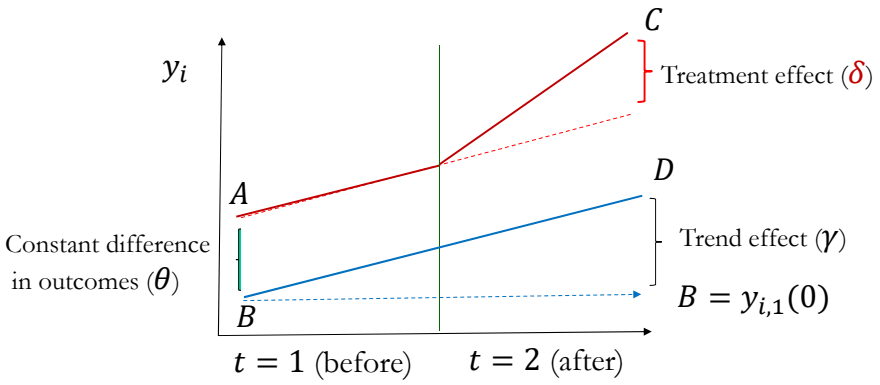
• Or, using levels and including individual, α_i, & time effects, $\theta_t = \gamma * Post_i$, (two-way FEM or **TWFE**), we estimate δ with:

$$y_{i,t} = \alpha_i + \gamma\, Post_i + \theta\, Z_i + \delta\, Z_i * Post_i + \boldsymbol{x}_{i,t}'\beta + \varepsilon_{i,t},$$

where $Post_i$ is a $t = 2$ dummy variable ($= 1$ if $t = 2$).

• We covered this topic (and the next example) last semester.

Causal Effects: DiD – Graphical Interpretation



Note: We impute the unobserved counterfactual as

$$\mathrm{E}[y_{i,2}(0) | Z_i = 1] = D + (\mathrm{A} - B)$$

In the above graph, the treatment effect (ATT) is given by:

$$\delta = C - \mathrm{E}[y_{i,2}(0) | Z_i = 1] = (C - A) - (D - B)$$

Causal Effects: DiD – Issues & Extensions

Example: Card (1990) studies the effect of a supply labor shock (the Mariel Liftboat influx of immigrants) in Miami, where we observe outcome $y_i(1)$. Card compares $y_{i=Miami}$ with observed outcomes, y_j, in four similar markets (j = Atlanta, Houston, LA & Tampa).

• Several extensions and issues:

- **Multiple periods** (**Staggered treatment**): Units receive treatment at different times. We have different treatment cohorts over time.

- **Variation in treatment timing**. There are heterogeneous effects over time across the cohort receiving treatment.

- **Non-parallel trends**: Assumption **D1** is violated.

- **Small treated clusters**. As discussed last semester, the CLT may be difficult to invoke.

Causal Effects: DiD – Staggered Treatment

- **Staggered treatment**. Now, taxi drivers can receive the bank loan at different times. The (static) TWFE works well when there is not heterogeneity in treatment effects across either time (this case) or units (next case).

• In many experiments the treatment effect fluctuates over time or among different groups due to learning curves, adaptation, or changing external conditions. In some cases, units may exhibit immediate responses to treatment, while others may need more time to react.

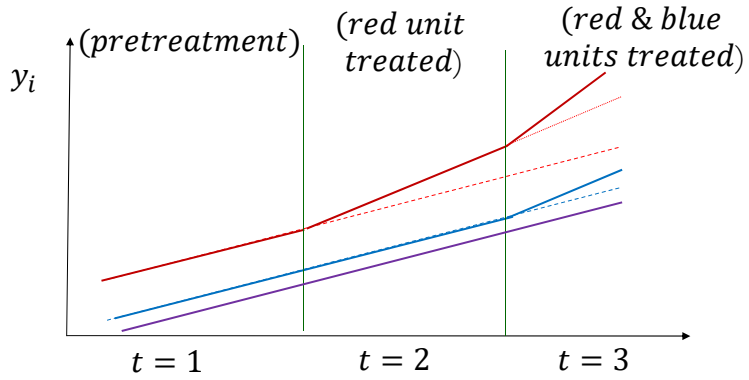
• It turns out the "traditional DiD" TWFE estimator is a weighted average of the heterogeneous time effects (δ_s):

$$\delta = \sum_{s=1}^{T} w_s \delta_s$$

where the weights can be negative, which can lead to a negative δ!

Causal Effects: DiD – Staggered Treatment

• Assuming that a treated unit continues to be treated until the end of the experiment, now, we have 3 groups at time $t = g$: **treated** (T), **not-yet treated** (NYT), & **never treated** (NT).



• At $t = 2$, NYT & NT units can serve as control units.

Causal Effects: DiD – Staggered Treatment

• Note that in the previous figure, the red unit effect increases substantially from $t = 2$ to $t = 3$, more than the blue unit effect (no constant TE across units and time). In this situation, a TWFE can be negative (due to a "*forbidden comparison*").

• Valid comparisons are between T & NYT and/or NT.

• If we assume that the dynamic effect of the bank loan expansion after r years is the same (on average) regardless of what year the taxi driver received the loan, a **dynamic TWFE** can identify ATT (with adaptations of assumptions **D1**-**D2**):

$$y_{i,t} = \alpha_i + \theta_t + \delta \sum_{r \neq 0} I[R_{i,t} = r] + \boldsymbol{x}_{i,t}'\boldsymbol{\beta} + \varepsilon_{i,t},$$

where $R_{i,t}(= t - g_i + 1)$ is the time relative to treatment and g_i is the time of first treatment. (See Borusyak and Jaravel (2018)).

Causal Effects: DiD – Treatment Timing

- **Variation in treatment timing**. We have heterogeneous effects over time across the cohort receiving treatment. For example, the effect of the bank loan on the cohort that was treated –i.e., received a loan– at t may be different from the cohort treated at t'. The effect varies with the period treatment was received. Dynamic TWFE can produce very biased results.

Callaway and Sant'Anna (2021) propose a weighted average of ATT per period and Borusyak, Jaravel and Spiess (2021) propose an estimation using units that where never treated over time.

For example, BJS use averages over the untreated period $(0 - t)$ to estimate $y_i(0)$ for $i = T$ & NT:

$$\hat{\delta}_{BJS} = (\hat{y}_{T,t}(1) - \hat{y}_{T,0-t}(0)) - (\hat{y}_{NT,t}(0) - \hat{y}_{NT,0-t}(0))$$

using the observed sample means.

Causal Effects: DiD – Non-parallel Trends

- **Non-parallel trends**: Assumption **D1** can fail if the confounding factors are time-varying. For example, in the bank loan example, experience may have a time-varying effect on outcomes. If we condition on confounders, we can recover the parallel trend:

$$\mathrm{E}[y_{i,2}(0) - y_{i,1}(0) | Z_i = 1, X_i] = \mathrm{E}[y_{i,2}(0) - y_{i,1}(0) | Z_i = 0, X_i].$$

With this assumption, along a "*strong overlap*" one, we estimate ATT:

$$\delta_x = \mathrm{E}[y_{i,2} - y_{i,1} | Z_i = 1, X_i = x] - \mathrm{E}[y_{i,2} - y_{i,1} | Z_i = 0, X_i = x]$$

The unconditional ATT can then be identified by averaging δ_x over the distribution of X_i in the treated population.

• Abadie (2005) proposes a two-step strategy to deal with this issue:

1) Estimate the PS based on observed covariates and compute $\widehat{PS}$

2) Estimate ATT, using IPW.

Causal Effects: DiD – Small Clusters

- **Small treated clusters**. As discussed last semester, it is a difficult theoretical problem, the CLT does not apply. Model-based solutions (Donald and Lang (2007) and bootstrapping have been proposed (Canay et al. (2021)).

• Model-based solutions: We start with a structural equation:

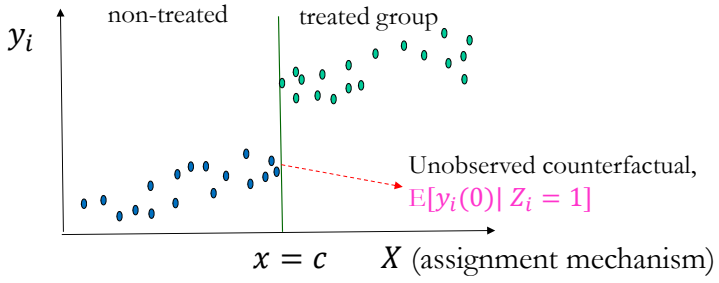
$$y_{i,j,t} = \alpha_j + \theta_t + \delta\, Z_{i,j} * Post_i + v_{j,t} + \varepsilon_{i,t},$$

where j is a cluster and $v_{j,t}$ is the cluster-level error term. With few clusters, the averages of the cluster level shocks $\Delta v_{j,t}$ among treated & untreated clusters will tend not to be approximately by a normal.

We need assumptions: For example, the $v_{j,t}$'s are normally distributed (Donald and Lang (2007)), or that treated and non-treated cluster (a larger cluster) follow the same distribution (Conley and Taber (2011) and Ferman and Pinto (2019), allowing for hereteroscedasticity).

Causal Effects: Regression Discontinuity (RD)

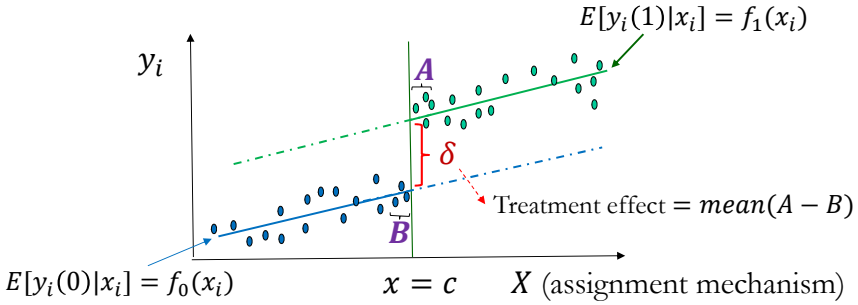
• **Regression discontinuity design** (RD/RDD) (Thistlethwaite and Campbell (1960), Hahn et al. (2001). There is a jump in the likelihood of being treated, generated by an **arbitrary threshold**, c.



Key feature: If units, even while having some influence, are unable to **precisely** manipulate the assignment mechanism, then, the variation in treatment near the threshold is randomized as though from a **randomized experiment**.

Causal Effects: RDD − Continuity

• If we assume that all factors (other than treatment) evolve **smoothly** with respect to X, then, observations in B would be a reasonable guess for the counterfactual for observations in A.



• As long as the unit's control over X is imprecise -the ex ante density of the assignment variable is continuous-, we have a **local randomization** of treatment. Thus, locally, all observable covariates, W, will have the same distribution on either side of $x = c$.

Causal Effects: RDD & RCT

• We can use the idea of "**balanced covariates**" to determine A & B.

• Why don't we use just the closest observations in each group to c? In practice, one **cannot** "only" use data close to c. The narrower the area that is examined, the less data there are.

• RDD has a strong theoretical appeal: Lee (2008) shows that RD does not need to *assume* that treatment variation is "as good as randomized"; RDD's randomized variation is a *consequence* of agents' inability to precisely control the assignment variable near the known cutoff.

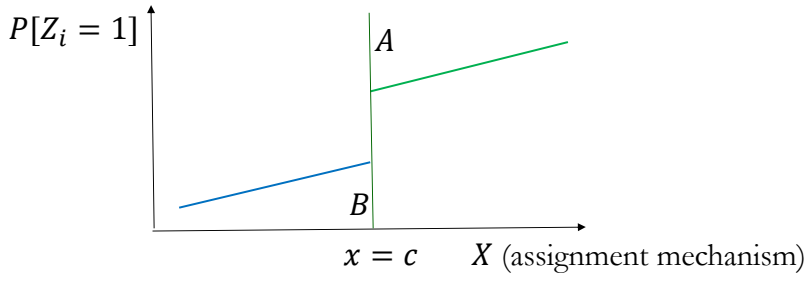
• In a review paper, Lee and Lemieux (JEL, 2010) consider RDD very close to an RCT: "*a much closer cousin of randomized experiments than other competing methods.*"

Causal Effects: RDD – Sharp & Fuzzy

• There are two practical cases:

- **Sharp RDD**: There is a deterministic rule that sets the jump into treatment. The jump at the threshold puts the probability of treatment equal to 1! –like in the previous graphs, under a jump, you are treated.

- **Fuzzy RDD**: There is an encouragement/incentive to get treatment. We think of the probability of treatment increasing at threshold.



Causal Effects: RDD – Sharp & Fuzzy

Example: In the U.S., to be a bank regulated as a *systemic risk bank* you need to have more than a certain amount of assets (USD 50B). This creates a Sharp RDD. In many states, some "small" business have access to subsidized loans. This creates a Fuzzy RDD.

In some situations, a Sharp RDD can create endogeneity issues ("*manipulation*"): A bank can precisely manage the amount of assets not to be treated (regulated). In these cases, the RDD design is not valid.

• Manipulation can be tested (McCrary test).

• Both RDD designs assumes a **continuity assumption**: At $x = c$, potential outcomes $\{y_i(1), y_i(0)\}$ are continuous (see dashed lines in slide 90). This assumption ensures that the "discontinuity" in outcomes around $x = c$ is due to the treatment effect, enabling causal inference.

Causal Effects: Sharp RD – Estimation

• In the Sharp RD case, we estimate $E[y_i(1) - y_i(0) | x_i = c]$, where c is the arbitrary cutoff point for receiving treatment. We compare units very close (& on both sides) of c. We compute the estimand by:

1) **Two potential regressions**:

(Treated) $$E[y_i(1) \mid x_i = c] = \lim_{x \downarrow c} E[y_i(1) | x_i = x]$$

(Control) $$E[y_i(0) \mid x_i = c] = \lim_{x \uparrow c} E[y_i(0) \mid x_i = x]$$

$$\Rightarrow \delta = E[y_i(1) \mid x_i = c] - E[y_i(0) \mid x_i = c]$$

2) **A linear regression** (using a window around c):

$$y_i = \alpha + \delta (x_i - c) + \varepsilon_i$$

the bandwidth $(x_i - c)$ can be selected in some optimal way, say, by minimizing the MSE, as suggested by Imbens and Kalyanaraman (2012). Non-linear functional forms are also OK.

Causal Effects: Fuzzy RDD – Estimation

• In the Fuzzy RDD case, we are back to **LATE**, we estimate an effect only for compliers. We have a Wald estimator:

$$\delta = \frac{\lim_{x \uparrow c} E[y_i(0) | x_i = x] - \lim_{x \downarrow c} E[y_i(1) | x_i = x]}{\lim_{x \uparrow c} E[d_i(0) | x_i = x] - \lim_{x \downarrow c} E[d_i(1) | x_i = x]}$$

The estimation involves two estimates around c: A reduced form regression & a first stage regression.

• Fuzzy RDD estimation is **doubly local**: Around the cutoff, c, and on the subsample of compliers.

• In general, determining the local **bandwidth** –i.e., the area around c that determines *A* & *B*– is not easy. Usually, an "optimal bandwidth" can be computed, but trying different bandwidth is common.

Causal Effects: Regression Kink Design (RKD)

• **Regression kink design** (RKD) (Nielsen et al. (2010) and Card, Lee, Pei, and Weber, 2015). A similar method to RDD, but RKD focus is on kinks (or change in slope) in the relation between the assignment variable, X, and the outcome variable, y_i.

• That is, RKD takes advantage of a change in the slope of the likelihood of being treated at a (kink) point c, resulting in a discontinuity in the first-derivative of the assignment function, X.

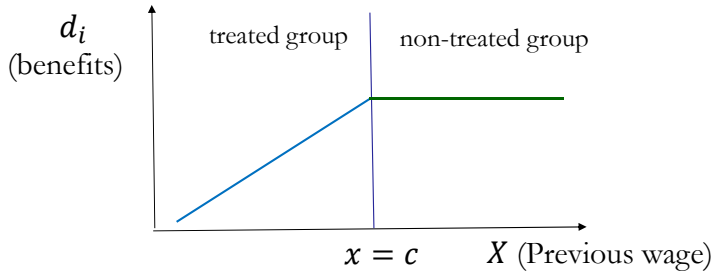
• If individuals on either side of the kink threshold, c, are "similar," any kink in the outcome can be attributed to the effect of treatment.

• In RKD we are interested in estimating a slope change at c for two relations: between X & y_i and between X & d_i -the treatment intensity.

Causal Effects: RKD – Similar to RDD

Example: In the usual government unemployment insurance scheme, insurance payments cover up to some percentage of the previous wage (55% in Austria), up to a maximum amount (usually, with a floor).

Then, we can use RKD to study how long people spend in unemployment, y_i, as function of the amount of benefits they receive.



Remark: RDD takes advantage of a change (jump) in the level of benefits, RKD takes advantage of a change in the slope of benefits.

Causal Effects: RKD – Estimation

• Like RDD, we have two cases:

- **Sharp RKD**, when the kink is deterministic –i.e., the assignment function, X, changes the slope exactly at point $x = c$.

- **Fuzzy RKD**, when the assignment is probabilistic.

• The causal impact, δ_{RKD}, is found by dividing the change in slope for y_i by the change in slope for the treatment intensity, d_i:

$$\delta_{RKD} = \frac{y\text{-}slope[c+h\,] - y\text{-}slope[c-h\,]}{d\text{-}slope[c+h\,] - d\text{-}slope[c-h\,]}$$

where h (bandwidth) is a small change around c. Under the Sharp case, we do not need to estimate the slopes in the denominator, the assignment function is known (in the previous graph, $d - slope[c + h\,] = 0$).

Causal Effects: Sharp RKD – Estimation

• Formally, for the Sharp case, we estimate the causal effect by:

$$\delta_{RKD} = \frac{\lim_{x \uparrow c} \frac{d}{dx} E[y_i(0)|\, x_i{=}x] - \lim_{x \downarrow c} \frac{d}{dx} E[y_i(1)|x_i{=}x]}{\lim_{x \uparrow c} \frac{d}{dx} d_i(x) - \lim_{x \downarrow c} \frac{d}{dx} d_i(x)}$$

where $d_i(x)$ is a known function determining treatment intensity.

Example: For the previous unemployment benefits example, we interpret the two components in δ_{RKD}:

- **Numerator:** It captures the discontinuous change in the slope of the unemployment spell as a function of unemployment benefits at the earning threshold $x = c$.

- **Denominator:** It captures the change in the slope of the unemployment benefits at the earning threshold $x = c$ (the slope of $d_i(x)$ must change –i.e., the kink exists!

Causal Effects: Fuzzy RKD – Estimation

• For the Fuzzy case, we replace in the denominator d_i by an expectation:

$$\delta_{RKD} = \frac{\lim_{x \uparrow c} \frac{d}{dx} E[y_i(0) \mid x_i = x] - \lim_{x \downarrow c} \frac{d}{dx} E[y_i(1) \mid x_i = x]}{\lim_{x \uparrow c} \frac{d}{dx} E[d_i(x) \mid x_i = x] - \lim_{x \downarrow c} \frac{d}{dx} E[d_i(x) \mid x_i = x]}$$

• For both cases, estimation requires two inputs:

- Estimation of the numerator derivatives, usually with a non-parametric **Local linear regression** or with a **higher-order polynomial regression**.

- Estimation of the denominator derivatives (for the Fuzzy case, only; for the Sharp case we know the $d_i(x)$ function). Similar to above.

• In the estimation, it is common to incorporate covariates.

Causal Effects: Synthetic Control Method

- **Synthetic control method** (Abadie and Gardeazabal (2003) and Abadie, Diamond, and Hainemuller, 2010). We use this method for aggregated units (countries, states, cities, etc.) to evaluate the effect of **large-scale interventions**.

• We replicate unobserved outcomes, usually $y_i(0)$, using weighted observed outcomes, from similar (**donor**) units, $y_{j(i)}(0)$:

$$\hat{y}_{i=N}(0) = \sum_{j=1}^{N-1} w_i \, y_{j(i)}(0),$$

where we have $N-1$ donor units (N is the total sample & index for unit under study).

Ideally, we pick the non-negative weights optimally. For example:

$$w_j = \underset{w_j}{\operatorname{argmin}} \sum_{i=1}^{T} \{ y_i(1) - \sum_{j=1}^{N-1} w_j y_{j(i)}(0) \}^2.$$

Causal Effects: Synthetic Control – Matching

• Note: The argmin is $\hat{\delta}_{SC}$! Also, the weights can be computed using a constrained regression ($w_j \geq 0$, $\sum_{j=1}^{N-1} w_j = 1$, and no constant).

• Behind this idea, there is an underling model, with observable ($\boldsymbol{x}$) and unobservable (u) factors:

$$y_{i,t}(0) = \alpha_i + \gamma_t + \delta_t u_i + \boldsymbol{x}_{i,t}'\boldsymbol{\beta} + \varepsilon_{i,t}.$$

• Under some assumptions, the weights reconstruct the counterfactuals for unit N: $u_{i=N}$ & $\boldsymbol{x}_{i=N,t}$ just using donor variables:

$$\sum_{j=1}^{N-1} w_j \boldsymbol{x}_{j,t} = \boldsymbol{x}_{i=N,t} \;\; \& \;\; \sum_{j=1}^{N-1} w_j u_j = u_{i=N}.$$

• Q: How do we select the donors and $\boldsymbol{x}_i$'s? Use donors' $\boldsymbol{x}_{j,t}$ to produce a $\boldsymbol{x}_{i=N,t}$, matching $\boldsymbol{x}_{i=N,t}$ pre-intervention. Pick the donors that produce the best match. We pick $\boldsymbol{x}_{i=N}$'s that commove with $y_{i=N}$.

Causal Effects: Synthetic Control – Example

Example: Abadie et al. (2015) estimate the effect of German reunification on GDP per capita using data from five donor countries (weight in parenthesis): Austria (.42), US (.22), Japan (.16), Switzerland (.11), and Netherlands (.09). They use as observables, $\boldsymbol{x}$: trade openness, inflation rate, industry share, schooling & investment rate.

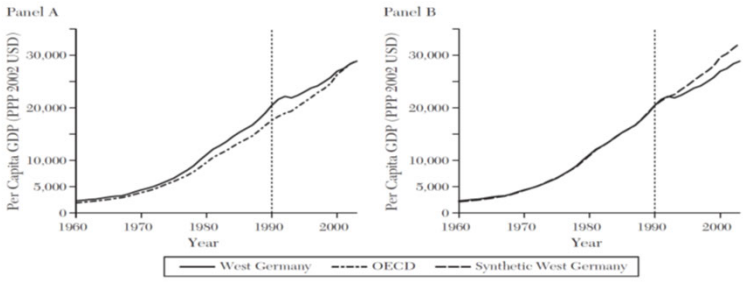


Figure 1. Synthetic Control Estimation in the German Reunification Example

• Abadie (2021) has a review article on JEL.