# Causal Inference: Identification Under Unmeasured Confounding (Instrumental Variables)

Hyunseung Kang

2025-02-11

# Concepts Covered Today

- Identification with an instrument
  - Monotonicity-based approach and the local/complier average treatment effect (LATE)
  - No additive interaction approach
- Randomized encouragement designs
- References:
  - Chapter 16 of M. Hernán and Robins (2020)
  - For monotonicity-based approach: Baiocchi, Cheng, and Small (2014)
  - For no additive interaction approach: Wang and Tchetgen Tchetgen (2018)

# Review: Strong Ignorability and Observational Studies I

We identified various causal estimands under the following assumptions:

- (A1, SUTVA): $Y = AY(1) + (1-A)Y(0)$
- (A2, Conditional randomization of $A$): $A \perp Y(1), Y(0) | X$
- (A3, Positivity/Overlap): $0 < \mathbb{P}(A = 1 | X = x) < 1$ for all $x$

Assumptions (A2) and (A3) are referred to as **strong ignorability.**

Under (A1)-(A3), we showed that the ATE can be identified as

$$\text{ATE} = \mathbb{E}[Y(1) - Y(0)] = \mathbb{E}[\mathbb{E}[Y \mid A = 1, X]] - \mathbb{E}[\mathbb{E}[Y \mid A = 0, X]]$$

(A1)-(A3) are plausible in a stratified randomized experiment. But the assumptions, especially (A2), is often implausible in an observational study.

This lecture will focus on the failure of (A2).

# When Does (A2) Fail? Non-compliance in Experiments I

One way for (A2) to fail is due to non-compliance from a randomized experiment.

Consider a randomized experiment to study the causal effect of a new therapy program versus the "standard" program.

1. Participants are randomized to the new therapy program (i.e., $A = 1$) or the standard program (i.e., $A = 0$).
2. But after randomization, some participants either
   - Drop out of the new program for the standard program or
   - Opt into the new program from the standard program

This is referred to as **non-compliance** because participants are not "complying" to the initial randomization of treatment.

Suppose the goal is to study the causal effect of *using* the new program versus not using it.

- Participant's usage/adherence is not randomized. Formally,

# When Does (A2) Fail? Non-compliance in Experiments II

- Let $D \in \{0, 1\}$ denote the *treatment receipt* of an individual
  - $D = 1$: individual uses the new therapy program.
  - $D = 0$: individual uses the standard program.
- Let $Y(d), d \in \{0, 1\}$ denote the counterfactual outcome of the new therapy (i.e.., $Y(1)$) or the standard program (i.e, $Y(0)$).
- We have $D \not\perp Y(1), Y(0), X$

If, however, participant's usage of the program $D$ is as-if random after adjusting for measured $X$, $D$ will satisfy (A2), i.e., $D \perp Y(1), Y(0) \mid X$, and we can use previous lectures to identify the causal effect of using the program.

- For example, if using the new therapy program is effectively random after adjusting for patient's age and gender, then $D$ will satisfy (A2).
- In most cases, investigators rarely believe that $D$ satisfies (A2) with $X$.

# When Does (A2) Fail? Unmeasured Confounders I

Another way for (A2) to fail is due to a **presence of unmeasured confounders** $U$ in an observational study.

- People select themselves into treatment (or control) based on measured covariates $X$ and unmeasured covariates $U$.
- More formally, strong ignorability holds with $X$ and $U$:

$$A \perp Y(1), Y(0) | X, U \quad \text{and } 0 < \mathbb{P}(A = 1 | X = x, U = u) < 1 \text{ for}$$

In both examples, we no longer have the identification result:

$$\mathbb{E}[Y(1) - Y(0)] \neq \mathbb{E}[\mathbb{E}[Y \mid A = 1, X]] - \mathbb{E}[\mathbb{E}[Y \mid A = 0, X]]$$

# Identification Without (A2): Instrumental Variables (IVs)

Instrumental variables (IVs) are a popular approach to identify a causal effect when (A2) does not hold; see M. A. Hernán and Robins (2006) and Baiocchi, Cheng, and Small (2014) for a review. Roughly speaking, an instrument relies on finding a variable $Z$, called an **instrument**, where

- $Z$ is related to the treatment $A$,
- $Z$ is independent from all unmeasured confounders that affect the outcome $Y$ and the treatment $A$, and
- $Z$ is related to the outcome $Y$ via the treatment $A$.

Here, we discuss two approaches to making the above statements about $Z$ precise.

1. Monotonicity-based approach
2. No additive interaction approach

# Randomized Encouragement Designs: Motivation I

Sexton and Hebel (1984) studied the causal effect of maternal smoking on birth weight. Because randomizing pregnant mothers to smoking is unethical, the authors considered an experimental design that randomized the *encouragement* to quit smoking.

1. Randomly assign some mothers to the encouragement intervention (i.e. $Z = 1$) or the usual care (i.e. $Z = 0$). The encouragement intervention encouraged mothers to not smoke through information, support, practical guidance, and the usual care.
2. Observe mothers' smoking status where $A = 1$ denotes that the mother is not smoking during pregnancy and $A = 0$ denotes that the mother is smoking during pregnancy.
3. Observe the birth weight of the newborn, denoted as $Y$.

# Randomized Encouragement Designs: Motivation II

We refer to $Z$ as the treatment assignment variable or **the instrument**. We refer to $A$ as the treatment receipt variable. This type of experimental design is referred to as a **randomized encouragement design** because the encouragement (or lack thereof; $Z$) was randomized. But, the treatment receipt $A$ is not randomized.

- If the encouragement is 100% successful so that $Z = A$, we have effectively randomized $A$ via $Z$. In practice, this is rarely the case.
- Nevertheless, the randomization of $Z$ induces some randomization of $A$, which we can exploit to obtain some causal effect of $A$.

# Randomized Encouragement Designs: Counterfactuals I

To define causal effects in a randomized encouragement design, we define the following counterfactual outcomes

- $A(z)$: the counterfactual treatment receipt under instrument $z$
- $Y(a, z)$: the counterfactual outcome under instrument $z$ and treatment receipt $a$.

In the maternal smoking example:

- $A(1)$: counterfactual smoking status if the mother was encouraged to stop smoking (i.e., $z = 1$)
- $A(0)$: counterfactual smoking status if the mother was not encouraged to stop smoking (i.e. $z = 0$)
- $Y(1, 1)$: counterfactual birth weight if the mother was encouraged to stop smoking (i.e., $z = 1$) and the mother stopped smoking (i.e., $a = 1$)

# Randomized Encouragement Designs: Counterfactuals II

- $Y(1,0)$: counterfactual birth weight if the mother was under the usual care (i.e., $z=0$) and the mother stopped smoking (i.e., $a=1$)
- $Y(0,1)$: counterfactual birth weight if the mother was encouraged to stop smoking (i.e., $z=1$) and the mother kept smoking (i.e., $a=0$)
- $Y(0,0)$: counterfactual birth weight if the mother was under the usual care (i.e., $z=0$) and the mother kept smoking (i.e., $a=0$)

It's also useful to study the following counterfactuals derived from above:

- $Y(A(z),z)$: the counterfactual outcome under instrument $z$ and treatment receipt if it takes on the value $A(z)$
  - Given $z$, the potential outcome $Y(A(z),z)$ is determined.
  - This is in contrast to $Y(a,z)$ where we need to specify both $a$ and $z$.

# Randomized Encouragement Designs: Counterfactuals III

- When used in the context of defining $Y(A(z), z)$, $A(z)$ is sometimes referred to as the "natural value" of $A$.

In the maternal smoking example:

- $Y(A(1), 1)$: counterfactual birth weight if the mother was encouraged to stop smoking (i.e., $z = 1$) and the mother's smoking status was set to her counterfactual smoking status under encouragement $A(1)$.
- $Y(A(0), 0)$: counterfactual birth weight if the mother was not encouraged to stop smoking (i.e., $z = 0$) and the mother's smoking status was set to her counterfactual smoking status under no encouragement $A(0)$.

# Randomized Encouragement Designs: Assumptions I

The following assumptions are implied from a randomized encouragement design.

- (IV1, SUTVA): $A = ZA(1) + (1 - Z)A(0)$ and $Y = ZY(A(1), 1) + (1 - Z)Y(A(0), z)$
- (IV2, Ignorable instrument): $Z \perp Y(1,1), Y(1,0), Y(0,1), Y(0,0), A(1), A(0)$
- (IV3, Overlap/positivity on instrument): $0 < P(Z = 1) < 1$

Assumption (IV1) says we get to observe the counterfactuals that correspond to the observed value of the instrument $Z$.

- For the outcome, we only get to observe the counterfactual outcome $Y(a, z)$ that corresponds to the observed instrument $Z = z$, specifically $Y(A(z), z)$.
- This is the case in the randomized encouragement design above where the researcher only has two interventions: the encouragement to quit smoking or the usual care.

# Randomized Encouragement Designs: Assumptions II

- Note that SUTVA also implies two "mini" assumptions about no multiple versions of treatment and no interference.

Assumption (IV2) says that the instrument (i.e. $Z$) was completely randomized.

- This is the case in the randomized encouragement design above where the encouragement intervention (i.e., $Z$) was completely randomized.

Assumption (IV3) says that all values of the instrument have a non-zero probability of being realized.

- This is also the case in the randomized encouragement design above where some mothers were randomized to the encouragement intervention while other mothers were randomized to the usual care.

# Randomized Encouragement Designs: Assumptions III

In short, (IV1)-(IV3) are conceptually identical to (A1)-(A3) where $Z$ is replaced by $A$.

This implies that if our goal is to simply identify the causal effect of $Z$ (i.e., the causal effect of encouragement versus usual care), we can use the prior lectures do this.

- For example, suppose we are interested in the effect of the encouragement (i.e., $Z$) on mother's smoking status (i.e., $A$), say $\mathbb{E}[A(1) - A(0)]$.
- From the previous lectures, under (IV1)-(IV3), we can identify the causal effect
  $\mathbb{E}[A(1) - A(0)] = \mathbb{E}[A \mid Z = 1] - \mathbb{E}[A \mid Z = 0]$

# Intent-to-Treat (ITT) Effects I

We can also identify the causal effect of $Z$ on the outcome $Y$, which is often called the **intent-to-treat (ITT) effect**.

$$\text{ITT} = \mathbb{E}[Y(A(1), 1) - Y(A(0), 0)]$$

The ITT effect is also written as $\mathbb{E}[Y(1) - Y(0)]$ where the counterfactual outcome is re-defined so that $Y(z) = Y(A(z), z)$.

In words, the ITT effect measure the causal effect of the initial random assignment (or the instrument $Z$) on the outcome

- The initial random assign represents the investigator's intent to assign treatment (or control) to participants in an experiment.
- In the maternal smoking example, the ITT effect is the causal effect of the *encouragement intervention* on the newborn's birth weight

# Intent-to-Treat (ITT) Effects II

- The ITT effect does not directly measure the causal effect of *maternal smoking* on the newborn's birth weight.
- The ITT effect is often reported in many randomized experiments and IV studies.

The identification of the ITT effect follows from assumptions (IV1)-(IV3), i.e.,

$$\begin{aligned}\mathbb{E}[Y \mid Z=1] &= \mathbb{E}[ZY(A(1),1) + (1-Z)Y(A(0),0) \mid Z=1] && \text{(IV1)}\\ &= \mathbb{E}[Y(A(1),1) \mid Z=1] \\ &= \mathbb{E}[Y(A(1),1)] && \text{(IV2)}\end{aligned}$$

By a similar argument, we have $\mathbb{E}[Y \mid Z=0] = \mathbb{E}[Y(A(0),0)]$. Combined, we have

$$\mathbb{E}[Y(A(1),1)] - \mathbb{E}[Y(A(0),0)] = \mathbb{E}[Y \mid Z=1] - \mathbb{E}[Y \mid Z=0]$$

# Randomized Encouragement Designs: Connection to Missing Data I

We can also interpret assumptions (IV1), (IV2), and (IV3) using the data table that includes both counterfactuals $Y(a,z), A(z)$ and observed variables $Z, A, Y$:

| | $Y(1,1)$ | $Y(1,0)$ | $Y(0,1)$ | $Y(0,0)$ | $A(1)$ | $A(0)$ | $A$ | $Z$ | $Y$ |
|---|---|---|---|---|---|---|---|---|---|
| Chloe | 15 | NA | NA | NA | 1 | NA | 1 | 1 | 15 |
| Sally | NA | NA | 20 | NA | 0 | NA | 0 | 1 | 20 |
| Kate | NA | NA | NA | 18 | NA | 0 | 0 | 0 | 18 |
| Julie | NA | 25 | NA | NA | NA | 1 | 1 | 0 | 25 |

# Randomized Encouragement Designs: Connection to Missing Data II

The variables $Z$ and $A$ both serve as missingness indicators. But, we only make assumptions about the missingness indicator $Z$ via (IV2) and (IV3); we don't make any assumptions about the missingness indicator $A$.

- In other words, assumptions (IV2) and (IV3) say that the missingness in the columns $A(1)$ and $A(0)$ are completely at random (MCAR) as the missingness in these columns are completely determined by $Z$, which is random by (IV2).
- But, the missingness of the four $Y(\cdot)$ columns may not be MCAR because (IV2) and (IV3) do not imply MCAR for $A$.
  - For these columns, the missingness is determined by $Z$ and $A$.
  - For these columns' missingness to be MCAR, we need $A, Z \perp Y(1,1), Y(1,0), Y(0,1), Y(0,0)$.

# Randomized Encouragement Designs: Connection to Missing Data III

Because the entries of the columns of $A(\cdot)$ are MCAR, we can identify the column means of $A(\cdot)$ by simply taking the mean of the observed entries of $A(\cdot)$, i.e.,

$$\mathbb{E}[A(1)] = \mathbb{E}[A \mid Z = 1]$$

In contrast, we cannot directly identify all four column means of $Y(\cdot)$ with the means of the observed entries as the missingness in these columns are not MCAR.

- But from the ITT slides above, we can identify a particular "mixture" of columns of $Y(\cdot)$ so long as the missingness in this "mixture column" is MCAR (via $Z$).

# Randomized Encouragement Designs: Connection to Missing Data IV

| | $A(1)$ | $A(0)$ | $Y(A(1),1)$ | $Y(A(0),0)$ | $Z$ | $Y$ |
|---|---|---|---|---|---|---|
| Chloe | 1 | NA | 15 | NA | 1 | 15 |
| Sally | 0 | NA | 20 | NA | 1 | 20 |
| Kate | NA | 0 | NA | 18 | 0 | 18 |
| Julie | NA | 1 | NA | 25 | 0 | 25 |

# Randomized Encouragement Designs: Connection to Missing Data V

The two new columns $Y(A(1), 1)$ and $Y(A(0), 0)$ essentially combine the four columns $Y(1, 1)$, $Y(1, 0)$, $Y(0, 1)$, and $Y(0, 0)$ so that the missingness in the two new columns only depend on $Z$

The column $Y(A(1), 1)$ "fuses" the columns $Y(1, 1)$ and $Y(0, 1)$ (i.e., $z = 1$)

- In words, this column represents a mixture of two sub-population of mothers under the encouragement intervention:
  1. mothers who decided to stop smoking after the encouragement (i.e., $A(1) = 1$)
  2. mothers who continued smoking after the encouragement (i.e., $A(1) = 0$)
- The average of this column represents the birth weight of infants from two sub-population of mothers.

# Randomized Encouragement Designs: Connection to Missing Data VI

The column $Y(A(0), 0)$ "fuses" the columns $Y(1, 0)$ and $Y(0, 0)$ (i.e., $z = 0$)

- In words, this column represents a mixture of two sub-population of mothers under the usual care:
  1. mothers who decided to stop smoking after the usual care (i.e., $A(0) = 1$)
  2. mothers who continued smoking after the usual care (i.e., $A(0) = 0$)
- The average of this column represents the birth weight of infants from two sub-population of mothers.

As mentioned earlier, for the two columns $Y(A(1), 1)$ and $Y(A(0), 0)$, their missingness pattern is MCAR as the missingness only depends on $Z$.

# Randomized Encouragement Designs: Connection to Missing Data VII

- Thus, we can identify the column mean of $Y(A(z), z)$ for $z \in \{0, 1\}$ by simply taking the mean of the observed values, i.e., $\mathbb{E}[Y(A(z), z)] = \mathbb{E}[Y \mid Z = z]$
- This matches the identification result for the intent-to-treat effect.

# Stratified Randomized Encouragement Designs I

Similar to previous lectures where we generalized a completely randomized experiment into a stratified randomized experiment based on covariates $X$, we can generalize a randomized encouragement design into a stratified randomized experiment design.

For example, consider again the smoking and birth weight example above.

- Instead of completely randomizing who gets the encouragement intervention or the usual care, we randomize the encouragement intervention within blocks of mothers.
- Each block is defined by mothers' measurable characteristics (e.g., age)
- Within each block, some mothers get randomized to the encouragement intervention while others get the usual care. Note that the probability of getting the encouragement can differ across blocks.

# Stratified Randomized Encouragement Designs II

- Among mothers who are older than 40, the probability of getting the encouragement intervention is 90%
- Among mothers who are between 25 to 30 years old, the probability of getting the encouragement intervention is 80%

Formally, we can rewrite (IV2) and (IV3) as follows:

- (IV2): $Z \perp Y(1,0), Y(0,1), Y(0,0), A(1), A(0) \mid X$
- (IV3): $0 < \mathbb{P}(Z = 1 \mid X = x) < 1$ for all $x$

Notice that this is nearly identical to a stratified randomized experiment from the previous lecture, except the randomization is done on $Z$ instead of $A$.

# Monotonicity-Based IV Assumptions I

Under a randomized encouragement design, we can formalize the assumptions about the instrument $Z$. This is broadly referred to as "monotonicity-based" IV assumptions.

- (IV4, Instrument relevance): $\mathbb{E}[A(1) - A(0)] \neq 0$
- (IV5, Exclusion restriction): $Y(a, 1) = Y(a, 0) = Y(a)$ for all $a$
- (IV6, Monotonicity/No Defiers): $\mathbb{P}(A(1) - A(0) \geq 0) = 1$

Assumption (IV4) states that the instrument has a non-zero, causal effect on the treatment receipt.

- In the maternal smoking example, (IV4) states that the encouragement intervention caused more mothers to quit smoking during pregnancy.
- Under (IV1)-(IV3), this assumption can be re-written based on the observed data,
  i.e. $\mathbb{E}[A(1) - A(0)] = \mathbb{E}[A \mid Z = 1] - \mathbb{E}[A \mid Z = 0] \neq 0.$

# Monotonicity-Based IV Assumptions II

- This means that we can directly test (IV4) with the observed data by testing whether $\mathbb{E}[A \mid Z = 1] - \mathbb{E}[A \mid Z = 0]$ is zero or not.

Assumption (IV5) states that the counterfactual outcomes are identical between $z = 1$ and $z = 0$ once the treatment receipt status $a$ is fixed.

- In the maternal smoking example, (IV5) states that after fixing the mother's smoking status, whether the mother was encouraged or not does not affect the birth weight of the newborn.
- Unlike (IV4), (IV5) cannot be written as a function of the observed data as it requires observing both $Y(a, 1)$ and $Y(a, 0)$ and from (IV1, SUTVA), this is not possible.
  - In other words, (IV5) cannot be directly tested with the observed data.

# Monotonicity-Based IV Assumptions III

- But, testable implications exist, i.e., if (IV5) holds, the observed data must satisfy certain constraints. See page 1173 in Balke and Pearl (1997) and Theorem 1 of Wang, Robins, and Richardson (2017) for some examples when the instrument is binary.

- (IV5) is the most controversial assumption as the other assumptions (IV1)-(IV4) and (IV6) can be plausibly satisfied by the experimental design (e.g., (IV1)-(IV3), (IV6)) or be directly tested with the observed data (e.g., (IV4)).
- This assumption is referred to as the **exclusion restriction** (Imbens and Angrist (1994), Angrist, Imbens, and Rubin (1996)).

Assumption (IV6) states that the instrument has a non-negative, causal effect on the treatment receipt for everyone.

# Compliance Types (Angrist, Imbens, and Rubin (1996)) I

To interpret assumption (IV6), it's useful to partition individuals based on their counterfactuals $A(0), A(1)$. Because each $A(z)$ takes on two values, there are four possible subgroups of individuals based on the joint values of $A(0), A(1)$:

| $A(0)$ | $A(1)$ | Type |
|---|---|---|
| 1 | 1 | Always-Takers |
| 0 | 1 | Compliers |
| 1 | 0 | Defiers |
| 0 | 0 | Never-Takers |

# Compliance Types (Angrist, Imbens, and Rubin (1996)) II

The names associated with each $A(0), A(1)$ (e.g. always-takers, compliers) come from Table 1 of Angrist, Imbens, and Rubin (1996). In the maternal smoking example,

- Always-takers are mothers who never smoke irrespective of whether they were under the encouragement intervention or the usual care.
- Compliers are mothers who do not smoke when they were under the encouragement intervention, but would smoke if they were under the usual care.
- Defiers are mothers who do not smoke when they are under the usual care, but smokes when they are under the encouragement intervention.
- Never-takers are mothers who always smoke irrespective of whether they were under the encouragement intervention or the usual care.

# Compliance Types (Angrist, Imbens, and Rubin (1996)) III

Assumption (IV6) rules out the existence of defiers in the study population, i.e. individuals who would not take the treatment if randomly assigned to the treatment, but take the treatment if randomly assigned to the control.

Also, **we cannot classify everyone in the study population as always-takers, compliers, and never-takers from the observed data**

- Why? Because this requires observing both $A(1)$ and $A(0)$, which is not possible from (IV1, SUTVA).
- But, as discussed above, we can identify the means $\mathbb{E}[A(1)]$ and $\mathbb{E}[A(0)]$ from (IV1)-(IV3):

$$\mathbb{E}[A(1)] = \mathbb{E}[A \mid Z = 1], \quad \mathbb{E}[A(0)] = \mathbb{E}[A \mid Z = 0]$$

Under the compliance type framework, these means can be interpreted as follows.

# Compliance Types (Angrist, Imbens, and Rubin (1996)) IV

- $\mathbb{E}[A(1)] = \mathbb{P}(A(1) = 1)$ represents the proportion of always-takers and compliers as they both have $A(1) = 1$.
- $\mathbb{E}[A(0)] = \mathbb{P}(A(0) = 1)$ represents the proportion of always-takers and defiers as they both have $A(0) = 0$
- With (IV1)-(IV3), we can identify the proportion of mixtures of subgroups.

With (IV6) where defiers do not exist, we can

- identify the proportion of always-takers via $\mathbb{E}[A(0)] = \mathbb{E}[A \mid Z = 0]$.
- identify the proportion of compliers via $\mathbb{E}[A(1) - A(0)] = \mathbb{E}[A \mid Z = 1] - \mathbb{E}[A \mid Z = 0]$.
- identify the proportion of never-takers via $1 - \mathbb{E}[A \mid Z = 1]$
- Note that the proportion of always-takers, compliers, and never-takers have to sum to $1$.

# Compliance Types (Angrist, Imbens, and Rubin (1996)) V

This concept of dividing up the population into sub-types based on the joint distribution of the post-treatment variables (e.g., $A(1)$ and $A(0)$) is referred to as principal stratification (Frangakis and Rubin (2002)).

# One-Sided, Randomized Encouragement Designs

In some experimental designs, we can enforce (IV6) by blocking access to treatment for all individuals who are randomized to the control $Z = 0$, i.e.,

- (IV6.One, One-Sided Noncompliance): $A(0) = 0$

One-sided non-compliance is plausible when $Z$ represents a new program under evaluation and $A$ represents the actual enrollment into the new program.

- In these settings, those who are not randomized into the new program (i.e., $Z = 0$) usually cannot enroll into the new program (i.e., $A = 0$).
- But, those who are randomized into the new program (i.e., $Z = 1$) can choose to enroll (i.e $A = 1$) or not enroll (i.e., $A = 0$) into the program.

Note that (IV6.One) implies (IV6).

# Causal Estimand: The Local Average Treatment Effect (LATE) I

Under (IV1)-(IV6), we can identify the average treatment effect among the compliers.

- This quantity is sometimes referred to the **local average treatment effect (LATE)** (Imbens and Angrist (1994),Angrist, Imbens, and Rubin (1996)).

$$\mathrm{LATE} = \mathbb{E}[Y(1) - Y(0) \mid \underbrace{A(1) = 1, A(0) = 0}_{\text{Compliers}}]$$

- In the maternal smoking example, the $\mathrm{LATE}$ is the average causal effect of smoking during pregnancy on newborn's birth weight among complying mothers (i.e. mothers who stop smoking if they were under the encouragement intervention, but smoke if they were under the usual care intervention).

# Causal Estimand: The Local Average Treatment Effect (LATE) II

The LATE is not the same as the ATE $\mathbb{E}[Y(1) - Y(0)]$, which represents the average causal effect of smoking during pregnancy on newborn's birth weight among all mothers.

The complier effect also differs from other "local" effects, such as the average causal effect of smoking on newborn's birth weight among never-takers: $\mathbb{E}[Y(1) - Y(0) \mid \underbrace{A(1) = 0, A(0) = 0}_{\text{Never-takers}}]$

# Causal Estimand: The Local Average Treatment Effect (LATE) III

- However, suppose the local effects are identical across the four subgroups, i.e.,

$$
\begin{aligned}
&\mathbb{E}[Y(1) - Y(0) \mid \underbrace{A(1) = 0, A(0) = 0}_{\text{Never-takers}}] \\
=&\mathbb{E}[Y(1) - Y(0) \mid \underbrace{A(1) = 1, A(0) = 0}_{\text{Compliers}}] \\
=&\mathbb{E}[Y(1) - Y(0) \mid \underbrace{A(1) = 1, A(0) = 1}_{\text{Always-takers}}] \\
=&\mathbb{E}[Y(1) - Y(0) \mid \underbrace{A(1) = 0, A(0) = 1}_{\text{Defiers}}],
\end{aligned}
$$

# Causal Estimand: The Local Average Treatment Effect (LATE) IV

- Then, we can identify the ATE with one of the local effects:

$$
\begin{aligned}
&\mathbb{E}[Y(1) - Y(0)] \\
=&\sum_{a_1 \in \{0,1\}, a_0 \in \{0,1\}} \mathbb{E}[\mathbb{E}[Y(1) - Y(0) \mid A(1) = a_1, A(0) = a_0]]\mathbb{P}(A(1) \\
=&\mathbb{E}[Y(1) - Y(0) \mid \underbrace{A(1) = 1, A(0) = 0}_{\text{Compliers}}] \sum_{a_1 \in \{0,1\}, a_0 \in \{0,1\}} \mathbb{P}(A(1) = a \\
=&\mathbb{E}[Y(1) - Y(0) \mid \underbrace{A(1) = 1, A(0) = 0}_{\text{Compliers}}].
\end{aligned}
$$

- In other words, if the causal effect is *homogeneous* across the four subgroups, the local effect equals the (global) effect for the entire population.

# Causal Estimand: The Local Average Treatment Effect (LATE) V

From the discussion above, we cannot use the observed data to classify all individuals into the four sub-types (i.e. compliers, always-takers, and never-takers).

- In other words, LATE identifies the average treatment effect among a subgroup of individuals that are defined by *latent classes*.
- This is in contrast to the conditional average treatment effect (CATE, $\mathbb{E}[Y(1) - Y(0) \mid X = x]$), which identifies the average treatment effect among a subgroup of individuals that are defined by observed $X$.

Because the subgroup of individuals are impossible to identify from the data, there is a healthy debate about whether the LATE is a useful estimand:

# Causal Estimand: The Local Average Treatment Effect (LATE) VI

- Some references: M. A. Hernán and Robins (2006),Deaton (2010),Imbens (2010), Imbens (2014),Baiocchi, Cheng, and Small (2014),Swanson and Hernán (2014)
- I personally think the identification of the LATE provides one clear illustration about the difficulty of studying the average treatment effect when strong ignorability fails to hold.

# Proof of Identification of the LATE with (IV1)-(IV6) I

We will show that under (IV1)-(IV6), we have

$$\begin{aligned}\text{LATE} &= \mathbb{E}[Y(1) - Y(0) | A(1) - A(0) = 1] \\ &= \frac{\mathbb{E}[Y \mid Z = 1] - \mathbb{E}[Y \mid Z = 0]}{\mathbb{E}[A \mid Z = 1] - \mathbb{E}[A \mid Z = 0]}\end{aligned}$$

We first begin with the numerator of the above ratio.

$$\begin{aligned}&\mathbb{E}[Y \mid Z = 1] \\ =&\mathbb{E}[ZY(A(1), 1) + (1 - Z)Y(A(0), 0) \mid Z = 1] && \text{(IV1, SUTVA)} \\ =&\mathbb{E}[Y(A(1), 1) \mid Z = 1] \\ =&\mathbb{E}[Y(1, 1)A(1) + Y(0, 1)(1 - A(1)) \mid Z = 1] \\ =&\mathbb{E}[Y(1, 1)A(1) + Y(0, 1)(1 - A(1))] && \text{(IV2, Ignorable } Z\text{)} \\ =&\mathbb{E}[Y(1)A(1) + Y(0)(1 - A(1))] && \text{(IV5, Exclusion restrict}\end{aligned}$$

# Proof of Identification of the LATE with (IV1)-(IV6) II

Note that (IV3, Positivity of $Z$) is needed to ensure that the conditional expectation that conditions on $\{Z = 1\}$ is well-defined.
By a similar argument, we have
$\mathbb{E}[Y \mid Z = 0] = \mathbb{E}[Y(1)A(0) + Y(0)(1 - A(0))]$.

Second, we take the difference between the two expectations of $\mathbb{E}[Y \mid Z = 1]$ and $\mathbb{E}[Y \mid Z = 0]$, we get

$$
\begin{aligned}
&\mathbb{E}[Y \mid Z = 1] - \mathbb{E}[Y \mid Z = 0] \\
=&\mathbb{E}[\{Y(1)A(1) + Y(0)(1 - A(1))\} - \{Y(1)A(0) + Y(0)(1 - A(0))\}] \\
=&\mathbb{E}[Y(1)\{A(1) - A(0)\} - Y(0)\{A(1) - A(0)\}] \\
=&\mathbb{E}[\{Y(1) - Y(0)\}\{A(1) - A(0)\}] \\
=&\mathbb{E}[\{Y(1) - Y(0)\}I(A(1) - A(0) = 1) + \{Y(1) - Y(0)\}I(A(1) - A(0 \\
=&\mathbb{E}[Y(1) - Y(0)|A(1) - A(0) = 1]\mathbb{P}(A(1) - A(0) = 1)
\end{aligned}
$$

The last equality also uses the definition of conditional expectation.

# Proof of Identification of the LATE with (IV1)-(IV6) III

Third, we can rewrite the denominator of the ratio above as follows:

$$\begin{aligned}
&\mathbb{E}[A \mid Z=1] - \mathbb{E}[A \mid Z=0] \\
=&\mathbb{E}[A(1) - A(0)] && \text{(IV1)-(IV3)} \\
=&\mathbb{P}(A(1) - A(0) = 1) && \text{(IV6).}
\end{aligned}$$

Finally, under (IV4, Instrument relevance), we can take the ratio of the two differences and the denominator of this ratio is non-zero:

$$\begin{aligned}
&\frac{\mathbb{E}[Y \mid Z=1] - \mathbb{E}[Y \mid Z=0]}{\mathbb{E}[A \mid Z=1] - \mathbb{E}[A \mid Z=0]} \\
=&\frac{\mathbb{E}[Y(1) - Y(0) | A(1) - A(0) = 1]\mathbb{P}(A(1) - A(0) = 1)}{\mathbb{P}(A(1) - A(0) = 1)} \\
=&\mathbb{E}[Y(1) - Y(0) | A(1) - A(0) = 1]
\end{aligned}$$

# Instrument Under the No-Additive Interaction Assumption I

As seen above, one way to identify causal effects of treatment $A$ when it does not satisfy (A2) is by restricting the heterogeneity of the treatment effect across latent/unobservable variables.

- Under the monotonicity framework, we "restricted" treatment effect heterogeneity in the latent space by simply removing defiers (i.e., (IV6)).
- In a separate line of work by Robins (1994) (see M. A. Hernán and Robins (2006) for a more refined version), an instrument was defined to restrict treatment effect heterogeneity through the **no additive interaction assumption.**
- As you'll see below, the same ratio that identified the LATE also identifies the average treatment effect on the treated (ATT) if an instrument is defined in another way that restricts treatment effect heterogeneity.

Roughly speaking, the no additive interaction framework assumes the following conditions:

# Instrument Under the No-Additive Interaction Assumption II

- (JV1, Causal consistency): $Y = Y(A, Z)$
- (JV2, Exhcangeable instrument): $Z \perp Y(1,1), Y(1,0), Y(0,1), Y(0,0)$
- (JV3, Positivity): $0 < \mathbb{P}(Z = 1) < 1$
- (JV4, Instrument relevance): $\mathbb{E}[A \mid Z = 1] \neq \mathbb{E}[A \mid Z = 0]$
- (JV5, Exclusion restriction) $Y(a, 1) = Y(a, 0) = Y(a)$ for all $a$
- (JV6, No additive interaction) Suppose (JV5) holds. We have $\mathbb{E}[Y(1) - Y(0)|Z = 1, A = 1] = \mathbb{E}[Y(1) - Y(0)|Z = 0, A = 1]$. Note that there is an implicit assumption that $0 < \mathbb{P}(Z = z, A = 1) < 1$ for all $z$.

Some remarks about the assumptions.

- The framework does not assume the existence of counterfactuals $A(1), A(0)$.

# Instrument Under the No-Additive Interaction Assumption III

- Assumption (JV1) and (JV2) are similar to assumptions (IV1) and (IV2), except that assumptions about the counterfactuals $A(1), A(0)$ are no longer present.
- Assumption (JV3) and (IV3) are identical.
- Similar to (IV2) and (IV3) above, we can create conditional versions of (JV2) and (JV3) that conditions on $X$:

$$Z \perp Y(1,1), Y(1,0), Y(0,1), Y(0,0) \mid X, \quad 0 < \mathbb{P}(Z = 1 \mid X = x)$$

- Assumption (JV4) states that the instrument is associated with $A$. In contrast to assumption (IV4), we do not necessarily need to have a causal effect of $Z$ on $A$.
- Assumption (JV5) is identical to (IV5).

# Interpreting the No-Additive Interaction Assumption (JV6) I

Assumption (JV6) can be interpreted by writing out a *saturated* model of the conditional expectation in (JV6).

$$\mathbb{E}[Y(1) - Y(0) \mid Z = z, A = 1] = \beta_0 + \beta_1 z$$

- A saturated model simply means that all of the variations on the left-hand side of the equality (i.e. the conditional expectation) can be explained by the model on the right-hand side of the equality.
- The term $\beta_0$ represents the ATT among individuals with $Z = 0$ and the term $\beta_0 + \beta_1$ represents the ATT among individuals with $Z = 1$.

Then, assumption (JV6) implies $\beta_1 = 0$.

# Interpreting the No-Additive Interaction Assumption (JV6) II

- In other words, the no additive interaction effect says that the "ATT effect" (i.e., the average difference of $Y(1) - Y(0)$ conditional on $A = 1$) is the same among individuals with $Z = 0$ and $Z = 1$.
- Note that (JV6) only restricts the effect of $Z$ on the outcome *conditional* on $A = 1$.
- For example, even under (JV6), it's possible that $\mathbb{E}[Y(1) - Y(0) \mid Z = 1] \neq \mathbb{E}[Y(1) - Y(0) \mid Z = 0]$

In the context of the maternal smoking example, (JV6) states that:

- The effect of smoking on birth weight among mothers that smoked during pregnancy is the same between mother under the encouragement intervention and mothers under the usual care.

# Proof of Identification of the ATT with (JV1)-(JV6) I

Now, we are ready to show that the ratio that was used to identify the LATE can also identify the ATT under (JV1-JV6):

$$\text{ATT} = \mathbb{E}[Y(1) - Y(0) \mid A = 1] = \frac{\mathbb{E}[Y \mid Z = 1] - \mathbb{E}[Y \mid Z = 0]}{\mathbb{E}[A \mid Z = 1] - \mathbb{E}[A \mid Z = 0]}$$

We begin with the numerator of this ratio.

$$\begin{aligned}
&\mathbb{E}[Y \mid Z = z] \\
=&\mathbb{E}[Y(A, Z) \mid Z = z] && \text{(JV1, C} \\
=&\mathbb{E}[Y(A) \mid Z = z] && \text{(JV5, E} \\
=&\mathbb{E}[Y(1)A + Y(0)(1 - A) \mid Z = z] \\
=&\mathbb{E}[(Y(1) - Y(0))A \mid Z = z] + \mathbb{E}[Y(0) \mid Z = z] \\
=&\mathbb{E}[(Y(1) - Y(0))A \mid Z = z] + \mathbb{E}[Y(0)] && \text{(JV2, E} \\
=&\mathbb{E}[Y(1) - Y(0) \mid Z = z, A = 1]\mathbb{P}(A = 1 \mid Z = z) + \mathbb{E}[Y(0)]
\end{aligned}$$

# Proof of Identification of the ATT with (JV1)-(JV6) II

Note that assumption (JV3, Positivity on $Z$) is used to have a well-defined conditional event $\{| \ Z = z\}$. Taking the difference $\mathbb{E}[Y \mid Z = 1] - \mathbb{E}[Y \mid Z = 0]$ yields

$$
\begin{aligned}
&\mathbb{E}[Y \mid Z = 1] - \mathbb{E}[Y \mid Z = 0] \\
=&\mathbb{E}[Y(1) - Y(0) \mid Z = 1, A = 1]\mathbb{P}(A = 1 \mid Z = 1) \\
&\quad - \mathbb{E}[Y(1) - Y(0) \mid Z = 0, A = 1]\mathbb{P}(A = 1 \mid Z = 0) \\
=&\mathbb{E}[Y(1) - Y(0) \mid A = 1] \left(\mathbb{P}(A = 1 \mid Z = 1) - \mathbb{P}(A = 1 \mid Z = 0)\right) \quad \text{(JV}
\end{aligned}
$$

The last equality utilizes the fact that (JV6) implies $\mathbb{E}[Y(1) - Y(0) \mid Z = 1, A = 1] = \mathbb{E}[Y(1) - Y(0) \mid Z = 0, A = 1] = \mathbb{E}[Y(1) - Y(0) \mid A = 1]$.

Dividing the above expression by $\mathbb{P}(A = 1 \mid Z = 1) - \mathbb{P}(A = 1 \mid Z = 0)$, which must be non-zero by assumption (JV4, Instrument relevance) gives us the desired result.

# Proof of Identification of the ATT with (JV1)-(JV6) III

Recent works have relaxed (JV6) to allow identification of the ATT (or the ATE); see Wang and Tchetgen Tchetgen (2018) and Cui and Tchetgen Tchetgen (2021).

# References I


Angrist, Joshua D, Guido W Imbens, and Donald B Rubin. 1996. "Identification of Causal Effects Using Instrumental Variables." *Journal of the American Statistical Association* 91 (434): 444–55.

Baiocchi, Michael, Jing Cheng, and Dylan S Small. 2014. "Instrumental Variable Methods for Causal Inference." *Statistics in Medicine* 33 (13): 2297–2340.

Balke, Alexander, and Judea Pearl. 1997. "Bounds on Treatment Effects from Studies with Imperfect Compliance." *Journal of the American Statistical Association* 92 (439): 1171–76.

Cui, Yifan, and Eric Tchetgen Tchetgen. 2021. "A Semiparametric Instrumental Variable Approach to Optimal Treatment Regimes Under Endogeneity." *Journal of the American Statistical Association* 116 (533): 162–73.

# References II


Deaton, Angus. 2010. "Instruments, Randomization, and Learning about Development." *Journal of Economic Literature* 48 (2): 424–55.

Frangakis, Constantine E, and Donald B Rubin. 2002. "Principal Stratification in Causal Inference." *Biometrics* 58 (1): 21–29.

Hernán, Miguel A, and James M Robins. 2006. "Instruments for Causal Inference: An Epidemiologist's Dream?" *Epidemiology* 17 (4): 360–72.

Hernán, Miguel, and James Robins. 2020. *Causal Inference: What If.* Boca Raton: Chapman & Hall/CRC.

Imbens, Guido W. 2010. "Better LATE Than Nothing: Some Comments on Deaton (2009) and Heckman and Urzua (2009)." *Journal of Economic Literature* 48 (2): 399–423.

———. 2014. "Instrumental Variables: An Econometrician's Perspective." *Statistical Science* 29 (3): 323–58.

# References III


Imbens, Guido W, and Joshua D Angrist. 1994. "Identification and Estimation of Local Average Treatment Effects." *Econometrica* 62 (2): 467–75.

Robins, James M. 1994. "Correcting for Non-Compliance in Randomized Trials Using Structural Nested Mean Models." *Communications in Statistics-Theory and Methods* 23 (8): 2379–2412.

Sexton, Mary, and J Richard Hebel. 1984. "A Clinical Trial of Change in Maternal Smoking and Its Effect on Birth Weight." *Jama* 251 (7): 911–15.

Swanson, Sonja A, and Miguel A Hernán. 2014. "Think Globally, Act Globally: An Epidemiologist's Perspective on Instrumental Variable Estimation." *Statistical Science* 29 (3): 371–74.

# References IV


Wang, Linbo, James M Robins, and Thomas S Richardson. 2017. "On Falsification of the Binary Instrumental Variable Model." *Biometrika* 104 (1): 229–36.

Wang, Linbo, and Eric Tchetgen Tchetgen. 2018. "Bounded, Efficient and Multiply Robust Estimation of Average Treatment Effects Using Instrumental Variables." *Journal of the Royal Statistical Society Series B: Statistical Methodology* 80 (3): 531–50.