Instrumental Variables

Alessandro Caggia

June 2025

Abstract

MONOTONICITY IS REQUIRED TO EXTARCT THE LATE UNDER HETEROGENOUS TE, IF YOU HAVE HOMOEGNEOUS TE THE EFFECT YOU ARE FINDING IS ALREADY THE ATE

1. Instrumental Variables

1.1. Introduction

- What if CIA (selection on observables) does not hold? (eg when individuals select themselves into treatment based on **unobservable characteristics** that are correlated with outcomes).
  - IV has been traditionally used to correct for measurement error and to recover structural parameters (remember the demand and supply example from class 1).
  - IV can be used to solve the omitted selection variable bias problem.

For now assume Homogeneous Treatment Effects

- Assume the treatment effect is constant: $Y_{1i} - Y_{0i} = \gamma$.
- Observable: $Y_i = Y_{0i} + D_i \cdot \gamma = \gamma_0 + \gamma D_i + \varepsilon_i$.
  - Where $\gamma_0 = \mathbb{E}(Y_{0i})$, $\varepsilon_i = Y_{0i} - \mathbb{E}(Y_{0i})$.
- This is basically a simple regression model with a constant slope.

1.2. IV Definition

- We want to find a way to recover the required lack of correlation between the error term and D.
- Classic Conditions for a variable Z to be a Valid instrument:
  - **Relevance:** $\text{Cov}(Z, D) \neq 0$
    the instrument has to capture part of the variation of the (endogenous) treatment variable.
  - **Exogeneity** implies two different assumptions:
    - **Randomness:** $\text{Cov}(Z, \varepsilon) = \text{Cov}(Z, Y_0) = 0$
      the instrument should be as good as randomly allocated.
    - **Exclusion Restriction:** $\delta = 0$ in the regression of Y on Z controlling for D:
      $Y_i = \gamma_0 + \gamma D_i + \delta Z_i + \varepsilon_i$
      the instrument can only affect the outcome through its effect on D.
    - We need both: Even if Z is randomly assigned, the exclusion restriction might not hold: Z can affect Y through channels different from D.

1.2.1. Testing the Conditions

- We can test for **Relevance**, since we observe both Z and D.
- **Exclusion Restriction:** This condition cannot be tested directly because it would require knowing the true causal effect γ. Why? So through our instrument Z we estimate:

$$Y_i = \gamma_0 + \gamma D_i + \delta Z_i + \varepsilon_i$$

We might think that if $\delta = 0$, the exclusion restriction is satisfied. But this is incorrect. Nothing says that your are effectively capturing the exogenous component of the endogenous variable D!

- Suppose D is endogenous because it is correlated with some unobserved confounder E st $D \to E \to Y$.
- If Z is still correlated with the same unobserved variable E, which causes the endogeneity of D, then $Z \to E \to Y$ remains an open backdoor path. **BUT** Crucially, controlling for D, we block or mask this effect mechanically (or better, there is this possibility that **makes the whole untestable**). SO if the endogenous channel is the same we see no effect on delta. Key idea: (1) We want to eliminate the confounding path $D \leftarrow E \to Y$, so we use $Z \to D \to Y$. (2) But if $Z \to E \to Y$, the instrument is invalid. it is usign the same exact endogeous channel you wanted to cut!
- That's why we say that we want Z to affect Y though D only! and this is more deep than it seems[1].

1.3. Identification of the IV Estimator

If Z is a valid instrument, we can identify the ATE (which in the homogeneous case is equal to the ATT, and to the individual treatment effect):

1. Remember for OLS, we have:

$$\gamma^{OLS} = \frac{\text{cov}(Y, D)}{\text{var}(D)}$$

The idea is that we replace the endogenous variable by its linear projection on an instrumental variable. For a binary Z, with no covariates, the IV recovers γ as:

$$\gamma^{IV} = \frac{\text{cov}(Y, Z)}{\text{cov}(D, Z)}$$

We only use the variability in D induced by Z. Numerator: how much Y varies with the exogenous component Z. before at denominator: variation in D, now we have exogenous variation in D generated by z (but this clearly shrinks the variation in the z hence the effective sample size)

2. Identify $= \text{cov}(D, Z)/\text{var}(Z) = \delta_1$ from the **First Stage**:
   - For a binary Z, with no covariates.
   - Model: $Y_i = \gamma_0 + \gamma D_i + \varepsilon_i$
   - Let: $D_i = \delta_0 + \delta_1 Z_i + \nu_i \leftarrow$ The first Stage
   - We can identify δ_1 from a regression of D on Z, which we call the First Stage. The relevance assumption is equivalent to assuming $\delta_1 \neq 0$, and if the only channel to y has to be D, then Z must be correlated with D.
3. Identify $\text{cov}(Y, Z)/\text{var}(Z) = \gamma\delta_1$ from the **Reduced form**:

$$\begin{aligned} Y_i &= \gamma_0 + \gamma(\delta_0 + \delta_1 Z_i + \nu_i) + \varepsilon_i \\ &= (\gamma_0 + \gamma\delta_0) + (\gamma\delta_1) Z_i + (\gamma\nu_i + \varepsilon_i) \end{aligned}$$

[1] Nothing prevents us from saying: D is affecting y through observable and unobservable channels. we use Z so that we rule out the unobservable channels (if it is truly exogenous). Now D is affecting y direcly if we control for the observables. Then the instrument becomes truly exogenous (is just capturing the direct effect of D on Y) only if you control for the observables. so that $\mathbb{E}(u_0 \mid X, Z) = 0$, $\mathbb{E}(u_1 \mid X, Z) = 0$

Since $\text{cov}(Z_i, \nu_i) = 0$ by construction, and $\text{cov}(Z_i, \varepsilon_i) = 0$ by assumption, we can identify $\gamma\delta_1$ from the reduced-form regression of Y on Z.

4. Finally, recalling 1. , we can write the Wald ratio estimator γ^{IV}:

$$\gamma^{IV} = \frac{\text{cov}(Y,Z)}{\text{cov}(D,Z)} = \frac{\text{cov}(Y,Z)/\text{var}(Z)}{\text{cov}(D,Z)/\text{var}(Z)} = \frac{\gamma\delta_1}{\delta_1} = \gamma$$

Note $\delta_1 \neq 0$ is essential for identification.

1.4. Estimation 2SLS

- Steps:
  1. Regress D on Z, get $\hat{D} = \hat{\delta}_0 + \hat{\delta}_1 Z_i$
  2. Regress Y on $\hat{D}$, that is: $Y_i = \gamma_0 + \gamma(\hat{\delta}_0 + \hat{\delta}_1 Z_i) + \varepsilon_{2i}$
- From the second step:

$$\gamma = \frac{\text{cov}(Y,\hat{D})}{\text{var}(\hat{D})} = \frac{\text{cov}(Y, \hat{\delta}_0 + \hat{\delta}_1 Z_i)}{\text{var}(\hat{\delta}_0 + \hat{\delta}_1 Z_i)} = \frac{\hat{\delta}_1 \,\text{cov}(Y, Z_i)}{\hat{\delta}_1^2 \,\text{var}(Z_i)} = \frac{\widehat{\gamma\delta_1}}{\hat{\delta}_1}$$

- In practice, we estimate the two steps jointly to correct for the errors in estimating δ_1 and δ_0, which are used as regressors in the second step.

1.5. OLS vs. IV

- In economics, we usually find larger IV estimates and se to be **larger than OLS estimates**. If the instrument is too weak, or if the instrument is slightly endogenous, it might be better not to use it (other issues are heterogeneous treatment effects or measurement errors).
  - The OLS estimator converges in probability to:

$$\frac{\text{cov}(Y,D)}{\text{var}(D)} = \frac{\text{cov}(\gamma_0 + \gamma D + \varepsilon, D)}{\sigma_D^2} = \gamma + \frac{\text{cov}(\varepsilon, D)}{\sigma_D^2}$$

  - Whereas, the IV estimator converges to:

$$\frac{\text{cov}(Y,Z)}{\text{cov}(D,Z)} = \frac{\text{cov}(\gamma_0 + \gamma D + \varepsilon, Z)}{\text{cov}(D,Z)} = \gamma + \frac{\text{cov}(\varepsilon, Z)}{\text{cov}(D,Z)}$$

    * The numerator is exogeneity, with full exogeneity there is no bias. Recall:
      · Inconsistency in OLS is due to nonzero $\text{corr}(\varepsilon, D)$.
      · Valid IV guarantees a consistent estimation since:

$$\text{corr}(\varepsilon, Z) = 0$$

    * the denominator is the first stage, if there is exogeneity, this amplifies asymptotic bias! (makes sense: recall the den is a scale up !). This is WHY even a slightly endogenous instrument with $\text{corr}(\varepsilon, Z) < \text{corr}(\varepsilon, D)$ may be worse than the no instrument scenario.

- Additionally, Weak instruments also lead to **large standard errors**.
  - WHY THIS? The asymptotic variance of the normalized IV estimator in the binary case is:

$$\text{Avar}\left(\sqrt{N}(\hat{\gamma}^{IV} - \gamma)\right) = \frac{\sigma_\varepsilon^2}{\sigma_D^2(\text{corr}(Z,D))^2}$$

  - The weaker the instrument, the higher the asymptotic se.

1.6. Adding Covariates in the General Model

1.6.1. Thoeretical framework

- We augment our models with controls and heterogeneous treatment effects based on that controls. We also allow for different gains based on unobserved factors u_1, u_0. The potential outcomes when treated or when controlled are then:

$$Y_0 = \mu_0 + \beta_0(x - \mu_x) + u_0, \qquad Y_1 = \mu_1 + \beta_1(x - \mu_x) + u_1$$

  - $\mu_0 = \mathbb{E}(Y_0),\ \mu_1 = \mathbb{E}(Y_1),\ \mu_x = \mathbb{E}(x)$

We can re-express the observed outcome:

$$\begin{aligned} Y &= Y_0 + D(Y_1 - Y_0) \\ &= \mu_0 + (x - \mu_x)\beta_0 + u_0 + D(\mu_1 - \mu_0) \\ &\quad + D(x - \mu_x)(\beta_1 - \beta_0) + D(u_1 - u_0) \\ Y &= \alpha_0 + \gamma D + x\beta_0 + u_0 + D(x - \mu_x)\delta + D(u_1 - u_0) \end{aligned}$$

- $D(u_1 - u_0)$ captures the interaction between treatment and unobserables from treatment.

1.6.2. Step by step expansion

- Use an instrument to capture the exogenous component of the endogenous regressor: $\mathbb{E}(u_0 \mid Z) = 0$, $\mathbb{E}(u_1 \mid Z) = 0$.
- To start, we keep the constant treatment effect assumption, then $\delta = \beta_1 - \beta_0 = 0$ and $u_1 = u_0$.
- Then:

$$\begin{aligned} Y &= Y_0 + D_i(Y_1 - Y_0) \\ &= \alpha_0 + \gamma D + x\beta_0 + u_0 + D(x - \mu_x)\delta + D(u_1 - u_0) \\ &= \alpha_0 + \gamma D + x\beta_0 + u_0 \end{aligned}$$

- Just use 2SLS as above with instrument Z)
- Identification and 2SLS as above!

1.6.3. Heterogeneous Treatments Effects Based on X

Background:

- Back to our model, where we allow heterogeneity based on X
- Now we allow for δ not to be 0, but we still assume $u_1 = u_0$:

$$Y = \alpha_0 + \gamma D + X\beta_0 + D(X - \mu_x)\delta + u_0$$

- Use as instruments: $(Z, Z(X - \overline{X}))$. Basically we are saying that we need to instrument also the interaction term.
- ATEx: $\hat{\gamma}(X) = \hat{\gamma} + \hat{\gamma}_{XD}(X_i - \overline{X})$
- The ATE (evaluated at the means) is just $\hat{\gamma}$

1.7. Multiple Instruments

- **2SLS strategy:**
  1. Regress D on $Z_1, Z_2, \ldots, Z_K, X$, get:

$$\widehat{D} = \hat{\delta}_0 + \hat{\delta}_1 Z_{1i} + \hat{\delta}_2 Z_{2i} + \cdots + \hat{\delta}_K Z_{Ki} + \hat{\delta}_x X_i$$

  2. Regress Y on $\widehat{D}, X$
- `ivregress 2sls Y X (D = Z1 Z2 ...ZK)`

1.8. Multiple Endogenous Variables

- The intuition is the same: we will have one first stage for each endogenous variable.
- We need as many instruments as endogenous variables. For example, with D_1 and D_2 endogenous, and two instruments.
- **2SLS strategy:**

1a. Regress D_1 on Z_1, Z_2, X, get:

$$\widehat{D}_1 = \hat{\delta}_0 + \hat{\delta}_1 Z_{1i} + \hat{\delta}_2 Z_{2i} + \hat{\delta}_x X_i$$

1b. Regress D_2 on Z_1, Z_2, X, get:

$$\widehat{D}_2 = \hat{\rho}_0 + \hat{\rho}_1 Z_{1i} + \hat{\rho}_2 Z_{2i} + \hat{\rho}_x X_i$$

2. Regress Y on $\widehat{D}_1, \widehat{D}_2, X$

- `ivregress 2sls Y X (D1 D2 = Z1 Z2)`

2. IV with heterogeneous treatement effects

- Will be assuming binary treatment (D) and binary instrument (Z).
- Z is a dummy for being assigned to treatment, an instrument for actual treatment D.

2.1. Potentital outcome frameworks Causality

2.1.1. Objects of Interest

There are three core causal effects in this setup:

1. **The Casual Effect of Z on D** (First Stage): the effect of the instrument on treatment.
2. **The Causal Effect of Z on Y** (Reduced Form): this is the **Intention-to-Treat (ITT)** effect the effect of the instrument on the outcome.
3. **The Causal Effect of D on Y** (Treatment Effect)

2.1.2. Potential Outcomes

We now consider the outcome as a function of both the treatment received and the treatment assignment:

$$Y_i(D_i, Z_i)$$

This means each unit has **four potential outcomes**, not just two.

$Y_{11} = Y(D = 1, Z = 1)$	Assigned and treated
$Y_{01} = Y(D = 0, Z = 1)$	Assigned and not treated
$Y_{10} = Y(D = 1, Z = 0)$	Not assigned but treated
$Y_{00} = Y(D = 0, Z = 0)$	Not assigned and not treated

This generalization:

- **Imposes SUTVA**: each units potential outcome depends only on its own (D_i, Z_i).
- **Does not impose exclusion restriction**: we allow Z to have a direct effect on Y, *beyond its effect through D* but expands the framework to model all combinations of assignment and compliance.

2.1.3. Potenital Treatment statuses

Key Idea: Every person has two potential treatment statuses, what they would do if assigned and if not assigned:

$$D_{1i} \text{ (if assigned, } Z_i = 1), \quad D_{0i} \text{ (if not assigned, } Z_i = 0)$$

But we only observe one; this is a switching equation expanded to account for heterogeneous compliance behavior:

$$D_i = D_{0i}(1 - Z_i) + D_{1i}Z_i = D_{0i} + Z_i(D_{1i} - D_{0i})$$

Note that there are two potential treatment statuses that are a **function of the treatment assignment + individual preferences**.

- The term $D_{1i} - D_{0i} \equiv \delta_{1i}$ has a i because it captures how an individual reacts to assignment: can take values 1, -1 and 0.
- This reaction may vary across individuals we allow **heterogeneous compliance**.
- Thus: $D_i = \delta_0 + \delta_{1i}Z_i$. Eg always takers have d_{0i} equal to 1 so that even if Zi is zero they have treatment.

This allows us to classify individuals as:

- **Compliers**: $D_{1i} = 1$, $D_{0i} = 0$, they switch and they follow the assigment (comply)
- **Always-takers**: $D_{1i} = 1$, $D_{0i} = 1$
- **Never-takers**: $D_{1i} = 0$, $D_{0i} = 0$, even if assigned it may decide not to take up
- **Defiers**: $D_{1i} = 0$, $D_{0i} = 1$, they switch and they always do the opposite than what they are assigned to do

2.1.4. Observed treatment compliance and treatment assignment

Key Idea: In IV designs, we only observe treatment D_i and assignment Z_i, not the underlying type (complier, defier, etc.). This creates a **missing data problem**: we observe only one realization of the potential outcomes and compliance types.

Example: If an individual has $Z_i = 0$ and $D_i = 0$, they could be either a: **Never-Taker** ($D_{1i} = 0, D_{0i} = 0$), or a **Complier** ($D_{1i} = 1, D_{0i} = 0$).

Observed Typology Table (based on observed (Z_i, D_i)):

	$Z_i = 0$	$Z_i = 1$
$D_i = 0$	Never-Taker or Complier	Never-Taker or Defier
$D_i = 1$	Always-Taker or Defier	Always-Taker or Complier

2.1.5. Observed Outcome

Rationale: To identify causal effects using an instrumental variable (IV), we must understand how treatment and outcome relate to assignment.

- **Derive the observed outcome Y_i! The outcome depends on the treatment assignment and the take-up type.**
- **Out of 4 potential outcomes for each individual you have just 2 depending on whether he was assigned to treatment or not** (if assigned to treatment, either he complies or not!):

$$Y_i = Y_i(0, Z_i) + D_i\,[Y_i(1, Z_i) - Y_i(0, Z_i)]$$

Substitute $D_i = D_{0i} + Z_i(D_{1i} - D_{0i})$:

$$\begin{aligned} Y_i = {} & Y_i(0, Z_i) + D_{0i} \cdot [Y_i(1, Z_i) - Y_i(0, Z_i)] \\ & + Z_i(D_{1i} - D_{0i}) \cdot [Y_i(1, Z_i) - Y_i(0, Z_i)] \end{aligned} \quad (*)$$

- **Baseline outcome:** potential outcome when untreated, $Y_i(0, Z_i)$. Either the outcome of compliers when unassigned to treatmenr or of never taket when assigned.
- **Baseline selection effect:** applies when $D_{0i} = 1$, i.e. for the always takers
- **Instrument-induced effect:** shift for those whose potenital otucome is truly affected by the treatment assignment **this is where identification comes from**.

Not convinced by the above? see it in practice!

For those with $Z_i = 1$: (from the eq above)

- **A complier** will have:

$$Y_i = Y_i(0, 1) + 0 + \boxed{\text{TE}}$$

the first term is just the baseline outocme of somebody that does not take up the treatment. You have to realize this is a stratified building to identify the three group sin the 6 situations. where $\boxed{\text{TE}}$ is the exogenous treatment effect the instrument activates this.

- **An always-taker (AT)** will have:

$$Y_i = Y_i(0, 1) + \boxed{\text{TE}}$$

where $\boxed{\text{TE}}$ is not exogenously triggered by the instrument (they would take treatment anyway).

- **A never-taker (NT)** will have:

$$Y_i = Y_i(0, 1)$$

- **A defier** will have:

$$Y_i = Y_i(0, 1) + \boxed{\text{TE}} - \boxed{\text{TE}}$$

where:

- TE is taken even when not assigned (like an AT),
- then − TE subtracts the part that is lost when the instrument is activated going against assignment.

Key: As you can clearly understand the only source of variation generate by the exogenous instrument on the treatemtn is the varaition generatedon compliers and defiers TE. so it is just the effect of compliers biased downwards by the effect of defiers!

2.2. Identification of ITT

2.2.1. Assumption 1: SUTVA (Stable Unit Treatment Value Assumption)

- **SUTVA**: allows us to write the causal effect for each unit independently from potential assignments, treatments, and outcomes of **other units**. Without SUTVA, causal inference is not possible because cross-unit spillovers would contaminate the potential outcome definition.
- Implies:

$$D_i(Z) = D_i(Z_i) \quad \text{and} \quad Y_i(D, Z) = Y_i(D_i, Z_i)$$

  - $D_i(Z)$ only depends on unit i's own assignment Z_i.
  - $Y_i(D, Z)$ only depends on unit i's own treatment D_i and assignment Z_i.

2.2.2. Assumption 2: Random Assignment of Instrument

- **Random Assignment:** The instrument Z_i is **independent** of the full vector of potential treatment statuses and potential outcomes:

$$Z_i \perp (D_{0i}, D_{1i}, Y_i(D_{0i}, 0), Y_i(D_{1i}, 1))$$

- In words: the instrument (e.g., treatment assignment) is randomly assigned and thus unrelated to unobserved determinants of treatment take-up or outcomes.
- This is a **stronger assumption** than classical IV exogeneity (which only required $Z \perp Y_0$).
- **Why this matters:** Because individuals self-select into treatment, we cannot assume that treatment is randomly assigned. We only assume the assignment Z_i is.

2.2.3. ITT

Assumption 1 and 2 allow to give causal interpreation to :

- The causal effect of Z on D **First Stage (Effect of Z on D)**:

$$\mathbb{E}[D_i \mid Z_i = 1] - \mathbb{E}[D_i \mid Z_i = 0] = \mathbb{E}[D_{1i} \mid Z_i = 1] - \mathbb{E}[D_{0i} \mid Z_i = 0]$$
$$= \mathbb{E}[D_{1i} - D_{0i}] = \frac{\text{Cov}(D_i, Z_i)}{\text{Var}(Z_i)}$$

- This recovers the causal effect of assignment on treatment take-up (logic of assumption is easy). This foundamentally means that the treatment assignment is affecting the treatment take up!
- The causal effect of Z on D. **Reduced Form (Effect of Z on Y)**: the ITT

$$\mathbb{E}[Y_i \mid Z_i = 1] - \mathbb{E}[Y_i \mid Z_i = 0] = \mathbb{E}[Y_i(D_{1i}, 1) \mid Z_i = 1] - \mathbb{E}[Y_i(D_{0i}, 0) \mid Z_i = 0]$$
$$= \mathbb{E}[Y_i(D_{1i}, 1)] - \mathbb{E}[Y_i(D_{0i}, 0)] = \frac{\text{Cov}(Y_i, Z_i)}{\text{Cov}(D_i, Z_i)}$$

- Combining the two object above you can get the Wald/Iv estimator:

$$\frac{\mathbb{E}[Y_i \mid Z_i = 1] - \mathbb{E}[Y_i \mid Z_i = 0]}{\mathbb{E}[D_i \mid Z_i = 1] - \mathbb{E}[D_i \mid Z_i = 0]} = \frac{\text{Cov}(Y_i, Z_i)}{\text{Cov}(D_i, Z_i)}$$

  - **Numerator:** ITT (Intention-To-Treat effect)
  - **Denominator:** First Stage (Effect of Z on treatment take-up)

But, can we interpret this IV estimand The causal effect of D on Y?

2.3. Extra Required Classical Assumptions

Necessary for Causal interpretation (classical)

2.3.1. Assumption 3: Exclusion Restriction

- **Exclusion Restriction:** The instrument Z_i affects the outcome Y_i only through the treatment D_i.
- Formally:

$$Y_i(D_i, 0) = Y_i(D_i, 1) = Y_i(D_i)$$

  Intuition: the channel through which the assignment is affecting the outcome is just the actual take up D.
- In this context, this implies that:

$$Y_{10i} = Y_{11i} = Y_{1i}, \quad Y_{01i} = Y_{00i} = Y_{0i}$$

  Second, we can write the causal effect of D on Y as: $Y_{1i} - Y_{0i}$
- We can rewrite the observed outcome as (finally):

$$\begin{aligned} Y_i &= Y_i(0, Z_i) + D_i\,[Y_i(1, Z_i) - Y_i(0, Z_i)] \\ &= Y_{0i} + D_i(Y_{1i} - Y_{0i}) \\ &= \alpha_0 + \gamma_i D_i + \nu_i \end{aligned}$$

- Where the first equation comes from equation $[*]$ above!
- **Interpretation:** Differently than classical models, here individual-specific (heterogeneous) treatment effect γ_i.

Assumption 3: implies that this causal effect is 0 for always-takers and never-takers (if Z does not affect D, the effect on Y should be 0).

2.3.2. Assumption 4: Relevance

- **Relevance:** The instrument Z_i must have a non-zero causal effect on the treatment D_i.
- Formally:

$$\mathbb{E}(D_{1i}) - \mathbb{E}(D_{0i}) \neq 0 \quad \text{or equiv} \quad \Pr(D_{1i} = 1) \neq \Pr(D_{0i} = 1)$$

- This ensures the instrument moves the probability of treatment it is equivalent to a non-zero first stage in the structural model.
- recall we haev two potenital treatment statuses:

$$D_{1i} \text{ (if assigned, } Z_i = 1), \quad D_{0i} \text{ (if not assigned, } Z_i = 0)$$

- **But:** We never observe both D_{1i} and D_{0i} for the same individual only one is revealed depending on Z_i. This is the classic missing data problem in causal inference.
- Therefore, we estimate this difference at the group level:

$$\mathbb{E}[D_i \mid Z_i = 1] - \mathbb{E}[D_i \mid Z_i = 0] = \mathbb{E}(D_{1i}) - \mathbb{E}(D_{0i})$$

  under the assumption that $Z_i \perp (D_{1i}, D_{0i})$ (random assignment).

Assumption 4: ensures that for compliers and /or defiers there will be a causal effect of Z on D.

2.3.3. IV Estimator, with assumptions 3 and 4

- The Iv estiamtor is written as:

$$\frac{\mathbb{E}[Y_i \mid Z_i = 1] - \mathbb{E}[Y_i \mid Z_i = 0]}{\mathbb{E}[D_i \mid Z_i = 1] - \mathbb{E}[D_i \mid Z_i = 0]} = \frac{\text{Cov}(Y_i, Z_i)}{\text{Cov}(D_i, Z_i)}$$

- **The numerator is:**[2]

$$\text{Cov}(Y_i, Z_i) = \mathbb{E}(Y_i \mid Z_i = 1) - \mathbb{E}(Y_i \mid Z_i = 0) = \mathbb{E}[(D_{1i} - D_{0i})(Y_{1i} - Y_{0i})]$$

- **The Denominator** (by assumption 2):

$$\text{Cov}(D_i, Z_i) = \mathbb{E}(D_i \mid Z_i = 1) - \mathbb{E}(D_i \mid Z_i = 0) = \mathbb{E}(D_{1i}) - \mathbb{E}(D_{0i})$$

[2]

- From equation (*), we can write the observed outcome as:

$$Y_i = Y_i(0, Z_i) + D_{0i} \cdot [Y_i(1, Z_i) - Y_i(0, Z_i)] + Z_i(D_{1i} - D_{0i}) \cdot [Y_i(1, Z_i) - Y_i(0, Z_i)] \quad (1)$$

- By Assumption 3:

$$Y_i = Y_{0i} + D_{0i}(Y_{1i} - Y_{0i}) + Z_i(D_{1i} - D_{0i})(Y_{1i} - Y_{0i})$$

This is the First Stage, nonzero by Assumption A4: Relevance

- **Therefore, the IV Wald estimand is:**

$$\gamma^{IV} = \frac{\text{Cov}(Y_i, Z_i)}{\text{Cov}(D_i, Z_i)} = \frac{\mathbb{E}[(D_{1i} - D_{0i})(Y_{1i} - Y_{0i})]}{\mathbb{E}(D_{1i} - D_{0i})}$$

What's going on? The exogenous instrument captures variability generated by the group of compliers and defiers, $D_{1i} - D_{0i}$ (so it is a weighted average just of compliers and defiers).

But, can we interpret this IV estimand as the causal eff of D on Y? FOR COMPLIERS

2.3.4. Assumption 5 Monotonicity

- **Need monotonicity assumption:** $D_{1i} \geq D_{0i} \quad \forall i \Rightarrow$ no defiers $\Rightarrow D_{1i} - D_{0i} \in \{0, 1\}$
- Without monotonicity, the IV estimator mixes effects with opposite signs, possibly leading to cancellation or undefined ratios.
- Monotonicity is untestable (obv)

2.3.5. Finally, the LATE

$$\mathbb{E}[(D_{1i} - D_{0i})(Y_{1i} - Y_{0i})] = \mathbb{E}[Y_{1i} - Y_{0i} \mid \text{Compliers}] \cdot \Pr(\text{Compliers}) \quad (\star)$$

and

$$\mathbb{E}(D_{1i} - D_{0i}) = Pr(D_{1i} - D_{0i} = 1) \Pr(\text{Compliers})$$

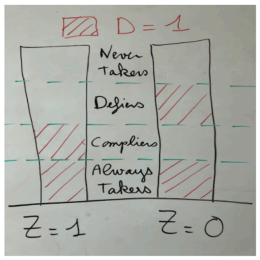
So our initial expression (lhs):

$$\gamma^{IV} = \frac{\mathbb{E}[(D_{1i} - D_{0i})(Y_{1i} - Y_{0i})]}{\mathbb{E}(D_{1i} - D_{0i})} = \mathbb{E}[Y_{1i} - Y_{0i} \mid \text{Compliers}]$$

Conclusion:

- With monotonicity, IV identifies the **Local Average Treatment Effect (LATE)**: defined for *compliers*units whose treatment status changes because of the instrument.
  - Vietnam draft example: a man who serves if drafted and does not serve otherwise.
- **Problem**: LATE may lack external validity because the complier group is instrument-specific.
  - LATE can still be policy-relevant (eg for scale up).
  - If different instruments produce similar IV estimates despite reaching different complier subpopulations, treatment effects are likely homogeneous.

2.4. How to Identify Compliers:



- You cannot simply see who among the z = 1 has a d = 1 bc there are always takers inside!!!
- We identify **Always-Takers (AT)** in $Z = 0$ and **Never-Takers (NT)** in $Z = 1$.
- From these two groups, we identify **Compliers** in $Z = 1$, and then also in $Z = 0$, as well as **AT** in $Z = 1$.
- As we identify **Compliers** in $Z = 1$, we can identify **NT** in $Z = 0$.

3. Back to ATE, ATT and ATNT

- Under homogeneous treatment effects, IV identifies the ATE, which would be equal to the LATE since all groups would have the same effect.
- Under heterogeneous treatment effects, the ATE is a weighted average of the treatment effects for the four groups in the table.
- The ATT is a weighted average of the effect for compliers and always takers (st $D = 1$).
- The ATNT is a weighted average of the effect for compliers and never takers (st $D = 0$).

3.1. Example, and LATE and external validity: When LATE becomes more than LATE!

3.1.1. Angrist & Evans (1998): Fertility → Mothers Income

Will see two methodolgoies, Both are exogenous, but they are different: multiple second births are a shock. Same sex of first two children increases the likelihood.

- **Objective & Setup** Goal: Estimate the causal effect of the number of children on the mother's labor market outcome (e.g., hours worked, employment).
  - Treatment D: Having a third child (#kids = 3 vs. 2).
  - Outcome Y: Mothers labor income or labormarket participation.
  - Endogeneity: the endogeous variable is the total numebr of children. Mothers selfselect based on unobserved preferences or abilities (poorer have mroe children).
- **Instrument (Z)**
  1. **Twin at Second Birth**
     - $Z = 1$ if second birth yields twins; $Z = 0$ if singleton.
     - *First Stage*: Twin induces +1 child, (multiple second birth leads to more kids than planned.) so

$$Z = 1 \implies D = 3 \quad \text{with high p;} \quad Z = 0 \implies D = 2$$

T group: Mothers who had twins at second birth → 3 kids
C group: Mothers who had singleton at second birth → 2 kids

     - *Exclusion Threats*: So the idea is that having twins does not affect female labor force participation through channels different than the number of children itself. Having twins does not directly affect mother's labor (BUT is not just having a child more: i) this +1 is coupled with another +1, economics of scale and it costs less (bring toghetr to school, same nap time), ii) the plus one arrives with another +1! a double birth can have health stress consequencees (that a plus one simply does not have)), except via increasing the number of children.
       - Health stress from twin pregnancy may directly reduce labor supply.
       - Simultaneous infant care differs from sequential births beyond one more child.
       - Twins share resources, lowering perchild cost, possibly affecting work decisions.
     - *Randomness*: Twin occurrence is quasirandom conditional on maternal age and observables.
     - *Compliance Types*:

- Take expectations conditional on Z_i and by Assumption 2:

$$\begin{aligned}\mathbb{E}(Y_i \mid Z_i) &= \mathbb{E}(Y_{0i} \mid Z_i) + \mathbb{E}[D_{0i}(Y_{1i} - Y_{0i}) \mid Z_i] \\ &\quad + Z_i \cdot \mathbb{E}[(D_{1i} - D_{0i})(Y_{1i} - Y_{0i}) \mid Z_i] \\ &= \mathbb{E}(Y_{0i}) + \mathbb{E}[D_{0i}(Y_{1i} - Y_{0i})] + Z_i \cdot \mathbb{E}[(D_{1i} - D_{0i})(Y_{1i} - Y_{0i})]\end{aligned}$$

- Then:

$$\begin{aligned}\mathbb{E}(Y_i \mid Z_i = 1) &= \mathbb{E}(Y_{0i}) + \mathbb{E}[D_{0i}(Y_{1i} - Y_{0i})] + \mathbb{E}[(D_{1i} - D_{0i})(Y_{1i} - Y_{0i})] \\ \mathbb{E}(Y_i \mid Z_i = 0) &= \mathbb{E}(Y_{0i}) + \mathbb{E}[D_{0i}(Y_{1i} - Y_{0i})]\end{aligned}$$

- Subtracting:

$$\mathbb{E}(Y_i \mid Z_i = 1) - \mathbb{E}(Y_i \mid Z_i = 0) = \mathbb{E}[(D_{1i} - D_{0i})(Y_{1i} - Y_{0i})]$$

  - ∗ *Alwaystakers*: Mothers who would have ≥ 3 children regardless.
  - ∗ *Nevertakers*: None, since any twin birth yields ≥ 3 children.
  - ∗ *Compliers*: Would have stopped at two if singleton, but have three when twins occur.
- *LATE*: Remark, you will estiamte LATE: Compliers = Women who have a third child only because they had twins at second birth (i.e., would have stopped at two otherwise).

$$\mathbb{E}[\, Y(3) - Y(2) \mid \text{Compliers} \,].$$

No nevertakers $\implies$ ATNT = LATE; IV estimates the average effect for mothers who would stop at two absent twins.

2. **SameSex First Two Children**
   - $Z = 1$ if first two births are same sex; $Z = 0$ if mixed.
   - *First Stage*: Parents preferring mixedsex set more likely to have a third when first two are samesex:

$$Z = 1 \implies \Pr(D \geq 3) \uparrow; \quad Z = 0 \implies \Pr(D \geq 3) \downarrow .$$

   - *Exclusion Threats*: So the idea is that having two boys or two girls rather than already mixed does not affect female labor force participation through channels different than the number of children itself (twins is random, gender if child is random). BUT what if two girls are more demanding one girl and one boy? 2. less costly to dress two girls (buy same stuff)). The channel would not be prompting a third child.
     - ∗ Parental preferences or SES may correlate with both sex-composition and labor outcomes.
     - ∗ Cost differences of raising two boys vs. two girls may affect labor supply separately.
   - *Randomness*: Child sex is random at conception.
   - *Compliance Types*:
     - ∗ *Compliers*: Have a third only if first two are samesex.
     - ∗ *Defiers*: Negligibleparents who prefer samesex sets and would stop if first two match.
   - *LATE*:

$$\mathbb{E}[\, Y(3) - Y(2) \mid \text{Compliers} \,].$$

A note on external validity: the instrument strength varies significantly across countries. Controlling for maternal age, education, religion, etc. aligns compliance strata to generalize estimates.

3.1.2. Oreopoulos (2006): Schooling → Income

- **Objective & Setup**
  - Treatment D: Completing an extra year of schooling (stay until age 15 vs. leave at 14).
  - Outcome Y: Later earnings.
  - Endogeneity: Clearly, ability and family background confound observational estimates.
- **Instruments (Z)**

1. **QuarterofBirth (QOB)**
   - Entry rule: Children must be age 6 by January 1 to start grade 1 that September.
   - Earlyquarter births (JanMar) enter younger, reach dropout age (16) earlier than latequarter (OctDec). The earliest you can tur 16 the earliest you can drop.
   - *First Stage*: Earlyquarter $\implies$ fewer compulsory years; latequarter $\implies$ more years.
   - *Exclusion*: Instruments affect earnings only via additional schooling. Changes in compulsory schooling laws or quarter of birth: no direct effect of the laws or the quarter of birth on earnings. 1. Law changign when schoos sichaneg affects only time at school and the enring of poeple just through time at school. 2. le being born in Q1 vs Q4 should not directly affect earnings, except by changing how much school you attend. (BUT. Seasonal variation in (besides time spent in school...) prenatal: nutrition, exposure to infections, and quality of prenatal care all vary across seasons, potentially affecting child development. As a result, birth date can correlate with various individual traits. For instance, it may relate to mental health (e.g., seasonal affective disorders or prenatal stress), personality (e.g., age-at-school-entry effects on confidence or leadership), and parental socioeconomic status (e.g., lower-income families may have higher birth rates in certain seasons).
   - *Monotonicity*: No student reduces schooling when instrument induces staying longer.
   - *Randomness*: Birth quarter or policycutoff timing is exogenous. Exact birth dates are random events. (BUT. parental planning may influence timing: wealthier or more educated parents may schedule births to avoid winter (seasonal avriation in parents). Then this would also vilate exlusion restriciotn (richer stay more).
   - *Compliance Types*:
     - ∗ *Compliers*: Stay one extra year only because birth quarter forces it; would drop out earlier otherwise.
     - ∗ *Alwaystakers*: Would stay regardless of QOB.
     - ∗ *Nevertakers*: None; monotonicity holds (no one drops out earlier if forced to stay).
   - *LATE*:

$$\mathbb{E}[\, Y(s+1) - Y(s) \mid \text{Compliers} \,],$$

return to an extra year for those induced by QOB.
   - *LATE & ATNT*: With no nevertakers, ATNT = LATE. IV estimates return to one extra year for would-be dropouts at 14. COMPLIERS REPRESENT THE WHOLE SET OF NON-TREATED! SO THE ATNT! (RECALL IN RUBIN MODEL FORMULA FOR ATT)

- **External Validity**
  - In countries with weaker CSL enforcement, nevertakers exist, so LATE ≠ ATNT.

3.1.3. Draft lottery as an instrument for serving in the military

First stage: Those drafted are more likely to serve in the military.

Exclusion restriction: Draft is random, no direct effect of being drafted on earnings regardless of the channel of going to war. BUT people can choose to get more education or training to avoid serving if drafted (and this affects earnings!) thus violating the exclusion restriction. Randomness: Draft lottery is random by design, if there is no manipulation (possibly conditional on observable covariates X). Some groups may have a higher likelihood of being drafted. For example, year of birth may influence likelihood important to consider these X's.

3.1.4. Dupas et al. (2018): Bank Account → Savings

- **Objective & Setup**
  - Treatment D: Having a formal bank account.
  - Outcome Y: Total savings.
  - Endogeneity: Selfselection into accounts based on unobserved saving propensity, access, or trust.
- **Instrument (Z)**
  - Randomized offer of a free bank account (no fees, assisted opening).
  - *First Stage*: $Z = 1$ greatly increases $\Pr(D = 1)$; $Z = 0$ leaves $\Pr(D = 1) \approx 0$ due to fees/barriers.
  - *Exclusion Threats*:
    - ∗ Offer is randomized, uncorrelated with baseline savings preferences.
    - ∗ Offer itself does not include any encouragement to save beyond account access.

- *Randomness*: Random assignment ensures exogeneity.
- *Compliance Types*:
  - *Alwaystakers*: Nonebefore offer, nearly no account opening due to costs.
  - *Nevertakers*: Refuse free account despite offer.
  - *Compliers*: Open account if and only if offered.
- *LATE & ATT*: No always-takers $\implies$ LATE = ATT. IV identifies the average effect of an account on savings for all treated (compliers). COMPLIERS REPRESENT THE WHOLE SET OF TREATED! SO THE ATT! (RECALL IN RUBIN MODEL FORMULA FOR ATT)

- **Assumptions**

1. *Relevance*: Freeaccount offer sharply increases takeup.
2. *Exogeneity*: Random assignment of offer ensures no correlation with unobserved determinants of savings.
3. *Exclusion*: Offer influences savings only through account opening.

4. Weak Instruments in Practice

- **Simple rule of thumb:** instruments are strong if the F-statistic for joint significance of the IVs in the first stage is larger than 10
- Procedure: Regress the treatment variable D on all instruments Z and controls X, and run an F-test on *only* the coefficients of Z.
- Stock and Yogo provide critical values for weak IV tests, which depend on sample size, number of instruments, and covariates (available via `ivreg2` output).

5. No Defiers Assumption

- **Definition:** A defier is someone who acts in the opposite direction of the instruments encouragement. Example: does not want a third child if the sex of the first two is the same, but wants one if they differ.
- Suppose some parents want exactly two boys or two girls:
  - If they have two daughters (or two sons), they stop.
  - If they have one boy and one girl, they try for a third.
  - In this case, they are defiers relative to the same-sex instrument.

6. Identifying Compliers

6.1. Complier Definition and Size

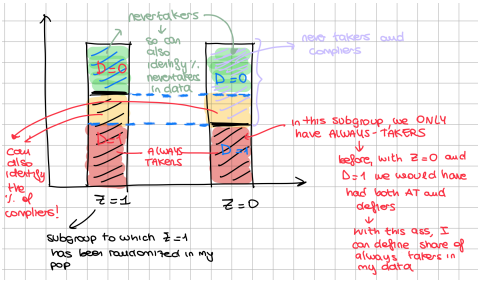
- It is not possible to identify which individual is a complier.
- But we can estimate the **size of the complier group**:

$$\Pr(D_{1i} - D_{0i} = 1) = \mathbb{E}[D_{1i} - D_{0i}] = \mathbb{E}[D_i \mid Z_i = 1] - \mathbb{E}[D_i \mid Z_i = 0]$$

6.2. Control Complier Mean (CCM)

- Always report the outcome mean in the control group in OLS to have perception of the magnitude.
- LATE = mean difference between treated and control compliers.
- For IV, the relevant comparison is the **Control Complier Mean (CCM)**: expected outcome for compliers in the control group.

BUT I can 1) estimate the share of always takers by looking at the ctonrols who took up, 2) estimate never takers by looking at those assigned to treatment that did not take up. Then, I am left just with the share of compliers.



6.3. Defiers and Monotonicity

skip rec unusefull

7. Fuzzy DID

7.1. Theoretical Problem Identified

This is an IV with inside two DIDs:

$$\text{Wald-DID} = \frac{\Delta Y^{\text{treated}} - \Delta Y^{\text{control}}}{\Delta D^{\text{treated}} - \Delta D^{\text{control}}}$$

Causality.

- Under constant treatment effects, refer to the (5) assumptions listed above (coming from the IV) + Parallel trends (coming from DiD) (actually parallel trends is equivalent to randomness).
- However, under heterogeneous treatment effects, LATE is identified only if BOTH:
  - Stable percentage of treatment units in control group (i.e., $\Delta D^{\text{control}} \neq 0$), or treatment effect on treatment switchers should be the same across groups.
  - Stable treatment effect over time (common trend in $Y(1)$).

Rationale of those assumptions: if TEs are heterogeneous we need the conditions above or we confound heterogeneity with the effects the assumptions are blocking.

7.2. How They Solve the Problem

Wald-TC (Time-Corrected): Constructs a counterfactual trend in outcomes using units whose treatment status does not change (always treated or always untreated).