# Instrumental Variables

Han Zhang

Apr 13, 2021

# Outline

# Logistics

- Final presentations in two weeks
  - we have 15 students; split into two weeks
- Exercise 3 due next week

# Recommend Readings

- Joshua D. Angrist and Jorn-Steffen Pischke. *Mostly Harmless Econometrics: An Empiricists Companion* . Princeton University Press, 2009. (Chapter 4)

# Short overview so far

- If we assume selection on observables/ignorability, matching and regression both identifies causal effect
  - But we are making a very strong assumptions: no unobserved confounders!
  - In Swiss naturalization example, it is more likely to be true
  - But in most real-life cases, ignorability is too strong
- Instead, a better approach acknowledge that there are unobserved confounders
  - And try to see eliminate unobserved confounders by adding less strong assumptions
  - Natural experiments
  - Fixed effects/DiD: exploit natural groups; use within-group difference to cancel out group-level unobserved confounders
  - General principle: approximates experiment ideal

# Instrumental Variable

- $Y = \alpha + \rho D + \epsilon$;
- $\rho$ identifies ATT/ATE if selection on observables are true
- When there are unobserved confoundings, $\rho$ will be a biased estimate of $ATT$
- In the linear regression framework, the presence of unobserved funding means that
  - Zero conditional mean assumption does not hold; $E(\epsilon|X) \neq 0$
  - Or, $X$ is an endogenous regressor
- Instrumental variables (IV) recognizes that unobserved confounders indeed exist
- IV exploits exogenous variation that drive the treatment but do not otherwise affect the outcome.

# IV setup

- IV Setup (assuming a linear regression!)
  - Second stage: $Y = \alpha + \rho D + \epsilon$
    - This is what you would normally run without IV
  - First stage: $X = \gamma + \beta Z + \eta$
- $Z$ is an IV
  - It drives treatment assignment $D$
  - But it does not have a direct impact on outcome $Y$
- $Z$ is kind of like a researcher who manipulates $D$ but not outcome
  - Later you will become more clear what this means

# IV example

- Angrist and Krueger, 1991, QJE:
- $Y$: future earnings
- $D$: years of schooling
- $Z$: season of birth.
  - Compulsory schooling laws require children to enter school in the calendar in which they turn 6
    - $Z = 0$: born in 1st quarter; enter school in Sep around 6.5
    - $Z = 1$: born in 4th quarter; enter school in Sep when they were less than 6
  - US compulsory schooling laws are in terms of age (16), not number of years of schooling completed. You can drop out on your 16th birthday (even if in the middle of the school year).
  - So those born in 4th quarter are compelled to take more educations than those born in 1st quarter

# IV assumptions

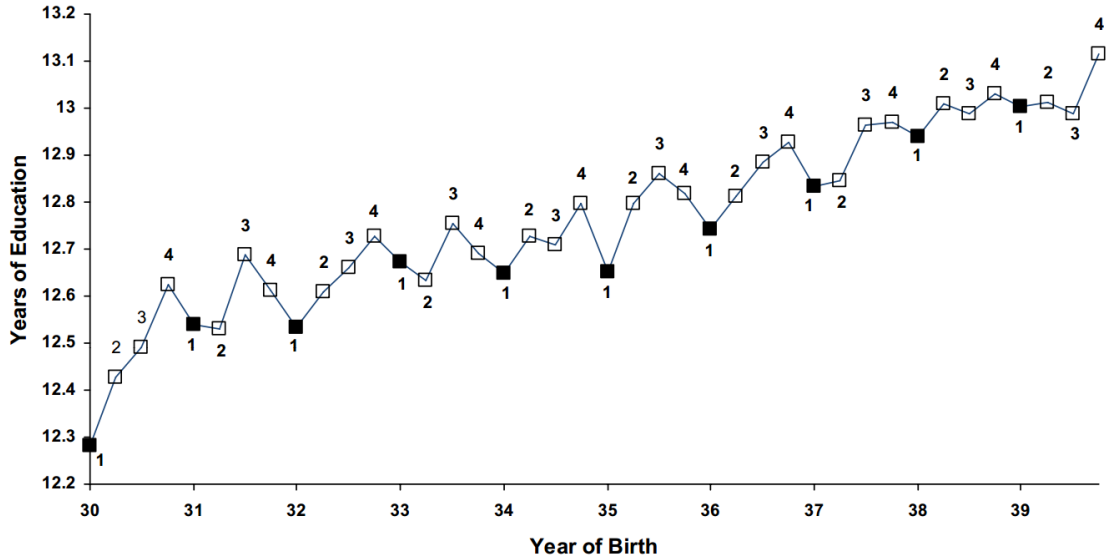
- Second stage: $Y = \alpha + \rho D + \epsilon$
- First stage: $D = \gamma + \beta Z + \eta$
- IV Assumptions:
  1. Exogeneity: $Z$ is an exogenous regressor (non-testable)
     - $E(\eta|Z) = 0$
     - $E(\epsilon|Z) = 0$
  2. Exclusion restriction: (non-testable)
     - $Z$ does not has a direct effect on $Y$
     - That is, in the true (unknown) data generating process, $Z$ does not appear in the second stage model
  3. First stage relevance: (testable)
     - $Z$'s effect on $D$ should not be 0 ($\beta \neq 0$)
  4. Constant effect $\rho$
     - It will be more clear what this assumption entails later

# Are the assumptions met for Angrist and Krueger (1991)?

1. Exogeneity: $Z$ is as-if randomly assigned.
   - non-testable
   - Reasonable? Will parents selectively give birth in certain months because they think children will be more beneficial?
2. Exclusion restriction: $Z$ does not has a direct effect on $D$
   - non-testable
   - Could birthdays affect earnings directly, beyond their effect on years of schooling?
     - e.g., those who were born in 1st quarter are older so that they will be more mentally developed compared with those born in 4th quarter
3. First stage relevance: instrument $Z$ should has an impact on $D$
   - Plot $D$ against $Z$; visually inspect correlation
   - Or, regress $D$ on $Z$;
     - The traditional rule-of-thumb is that first-stage regression's F-statistics $> 10$
     - Some recent work suggests that F-statistics $> 100$

# First stage relevance example

- MHE, Fig. 4.1.1
- In general, those who were born in the 4th quarter has more schooling than those born in the 1st quarter




- F-statistics $= 24.1 > 10$

# Math of IV

- We are interested in $\rho$, but directly regressing $Y$ on $D$ yields a biased estimate of $\rho$ (because $E(\epsilon|D) \neq 0$)

$$\begin{aligned} Y &= \alpha + \rho D + \epsilon \\ &= \alpha + \rho(\gamma + \beta Z + \eta) + \epsilon \\ &= (\alpha + \rho\gamma) + \rho\beta Z + \rho\eta + \epsilon \end{aligned} \tag{1}$$

- Substituting $D$ by the first-stage equation
- Now the second-stage becomes regressing $Y$ on instrument $Z$, and the errors are mean independent of $Z$: $E(\rho\eta + \epsilon|Z) = 0$)
  - $E(\eta|Z) = 0$)
  - $E(\epsilon|D) = 0$
- $\widehat{\rho\beta}$ is a consistent estimate of $\rho\beta$
- $\hat{\beta}$ is a consistent estimate of $\beta$ (regress $D$ on $Z$)
- $\frac{\widehat{\rho\beta}}{\hat{\beta}}$ yields a consistent estimate of $\rho$: indirect least square estimator

# Indirect Least Square (Wright, 1928)

- Indirect least square suggests that
  - Regress $Y$ on $Z$ and get coefficient estimate $\widehat{\rho\beta}$
  - Regress $D$ on $Z$ and get coefficient estimate $\hat{\beta}$
  - $\frac{\widehat{\rho\beta}}{\hat{\beta}}$ yields the causal effect of interest
- Regressing $Y$ on instrument $Z$ is called reduced form regression
- If we have multiple instruments $Z_1, Z_2$, each of them being a valid instrument, resulting in a different estimate of $\beta$
  - That is, we may have different ways to write the first stage: $D$ on $Z_1$ or $D$ on $Z_2$
  - Treatment effect depends on your instrument. More on this later

# Wald estimator (Wald, 1940)

- In the case of binary instrument, assuming the same linear model:
- $E(D|Z=1) = \gamma + \beta + \eta$ and $E(D|Z=0) = \gamma + \eta$
- $\beta = E(D|Z=1) - E(D|Z=0)$
- Similarly, we can get $\rho\beta = E(Y|Z=1) - E(Y|Z=0)$
- Wald Estimator

$$\hat{\rho} = \frac{\hat{E}(Y|Z=1) - \hat{E}(Y|Z=0)}{\hat{E}(D|Z=1) - \hat{E}(D|Z=0)}, \text{(Wald Estimator)} \quad (2)$$

$$= \frac{\widehat{\rho\beta}}{\hat{\beta}} \text{(Equals to Indirect Least Square)} \quad (3)$$

- Note: Wald estimator is a more general non-parametric estimator; it equals to Indirect Least Square in binary case only when the linear model assumption is true

# 2SLS estimator

- To solve the multiple instrument problem, we slightly change how we combine the terms

$$\begin{aligned} Y &= \alpha + \rho D + \epsilon \\ &= \alpha + \rho(\gamma + \beta Z + \eta) + \epsilon \\ &= \alpha + \rho(\gamma + \beta Z) + (\eta + \rho\epsilon) \end{aligned}$$

- Errors $\eta + \rho\epsilon$ are mean independent of $\gamma + \beta Z \implies$ regressing $Y$ on $\gamma + \beta Z$ gives consistent estimate of $\rho$
- $\gamma + \beta Z$ can be approximated by the predicted values of the first stage, $\hat{D} = \hat{\gamma} + \hat{\beta} Z$
- Hence, 2SLS estimator (Two-stage least square):
  1. Fit the first-stage regression using all instruments ($Z_1$ and $Z_2$, for example)

$$D = \gamma + \beta_1 Z_1 + \beta_2 Z_2 + \eta \tag{4}$$

  2. Regress $Y$ on predicted $D$

$$Y = \alpha + \rho \hat{D} + \epsilon$$

- Now, coefficient of $D$ from the second stage consistently estimate $\rho$

# Standard Errors

- The standard error estimate of $\rho$ is a little more tricky
- 2SLS suggests that you can regress $Y$ on predicted $D$; coefficient of $D$ is your estimate of $\rho$

$$Y = \alpha + \rho\hat{D} + \epsilon$$

- This gives you a consistent point estimate of $\rho$, but not a consistent estimate of the standard error of $\rho$
- Directly regress $Y$ on $\hat{D}$ assumes that standard error are related to $\hat{D}$

- But correct standard errors are associated with $D$
- We want standard errors related to the real $D$, not the predicted $\hat{D}$
- Software will correct this for you.

# Weak instruments

- Weak instrument ($Z$ is not a good predictor of $D$)
- Weak instrument can lead to biased estimates
- To get the intuition:

$$\hat{\rho} = \frac{\widehat{\rho\beta}}{\hat{\beta}} = \frac{\hat{E}(Y|Z=1) - \hat{E}(Y|Z=0)}{\hat{E}(D|Z=1) - \hat{E}(D|Z=0)}$$

- Weak instrument $Z$ suggests that $Z$ has little power to distinguish between $\hat{E}(D|Z=1)$ from $\hat{E}(D|Z=0)$
- In other words, $\hat{E}(D|Z=1) - \hat{E}(D|Z=0)$ is close to 0
- Alternatively, first stage regression coefficient $\hat{\beta}$ is close to 0
- Small disturbance in the denominator can lead to large bias
- That's why we require first-stage relevance

# Durbin-Wu-Hausman test

- Durbin-Wu-Hausman test (or Hausman's specification test)
- Traditionally this was believed to be a way to "test whether a regressor is exogenous or endogenous"
  - You may still see some of this saying in the old textbook/articles
- Basically, DWH test compares OLS estimate (second-stage) with 2SLS estimate
  - The Null is that the two are the same
  - The Alternative is that the two are different
- If you do not pass DWH test, your IV is bad and probably should not be used.
- But if you pass DWH test (p-value $< 0.05$), it does not mean that you have a valid IV

# Three causal effects

- Second-stage: regress outcome $Y$ on treatment $D$
  - Identifies the causal effect of treatment on outcome
- First-stage: regress treatment $D$ on instrument $Z$
  - Identifies the causal effect of instrument on treatment
- Reduced form: regress outcome $Y$ on instrument $Z$
  - Identifies the causal effect of instrument on outcome
- Most times we care about the effect of treatment on outcome
  - Does the coefficient represents $ATT$? $ATE$? or something else?
- But the effect of instrument on treatment, and the effect of instrument on outcome, can also be meaningful

# Traditional IV assumptions again

- Second stage: $Y = \alpha + \rho D + \epsilon$
- First stage: $D = \gamma + \beta Z + \eta$
- IV Assumptions:
  1. Exogeneity: $Z$ is an exogenous regressor (non-testable)
     - $E(\eta|Z) = 0$
     - $E(\epsilon|Z) = 0$
  2. Exclusion restriction: (non-testable)
     - $Z$ does not has a direct effect on $Y$
     - In linear model, $Z$ does not appear in the second stage model
  3. First stage relevance: (testable)
     - $Z$'s effect on $D$ should not be 0 ($\beta \neq 0$)
  4. Constant effect $\rho$
     - It will be more clear what this assumption entails later

# IV in randomized experiments

- Traditional IV relies heavily on math tricks; model-based
- How do we understand IV in a design-based framework?
  - making your study more like experiments?
- $Z$ is kind of like a researcher that manipulates $D$ but not outcome

# IV in randomized experiments

- In many randomized experiments, you can assign treatment, but cannot force people to take the treatment
- Sommer and Zenger (1991)
- Goal: Study the effect of vitamin A supplements on infant mortality in Indonesia.
- $Z$: treatment assignment; The vitamin supplements were randomly assigned to villages
- $D$: actual treatment: some of the individuals in villages assigned to the treatment group failed to receive them
- There is a self-selection into taking the treatment, though treatment assignment is random
- Non-compliance: people do not follow the treatment they were assigned to

# IV in randomized experiments

- $Z$ is a valid instrument:
  - It is randomly assigned (exogeneity)
  - It affect outcome only through $D$, actual treatment (exclusion restriction)
  - And it certainly matters for $D$ (first-stage relevance)
- Effect of $Z$ on $Y$ then measures Intent-to-treat (ITT) effect
- Effect of $D$ on $Y$ measures actual treatment effect
- We formally develop the idea using counterfactual framework

# New IV assumptions

1. Exogeneity: $Z$ is as-if randomly assigned
   - Compare with the previous exogenous regressor assumption, which one do you prefer?
   - Under counterfactual framework, this implies that potential outcomes are independent of treatment assignment: $Y_i^0, Y_i^1 \perp\!\!\!\perp D_i$
2. Exclusion restriction: $Z$ does not has a direct effect on $Y$
   - Under counterfactual framework, once we fix the value of the actual treatment $D$, $Z$ does not impact $Y$
3. First-stage relevance

# Potential Outcome Framework under Non-Compliance

- Previously we know potential outcome for outcome $Y$

$$Y_i = \begin{cases} Y_i^0 : D_i = 0 \\ Y_i^1 : D_i = 1 \end{cases}$$

- Now we can define an additional potential outcome for actual treatment $D$ (since it's the "outcome" of treatment assignment)
- Angrist, Imbens, and Rubin, 1996, JASA

$$D_i = \begin{cases} D_i^0 : Z_i = 0 \\ D_i^1 : Z_i = 1 \end{cases}$$

- $D_i^0$: actual treatment if $i$ were not assigned to treatment
- $D_i^1$: actual treatment if $i$ were assigned to treatment
- We only observed $D_i$, not $D_i^0$ and $D_i^1$

# Compliance types

- Under the counterfactual framework:

| | $D_i^0 = 0$ | $D_i^0 = 1$ |
|---|---|---|
| $D_i^1 = 0$ | never-taker | defier |
| $D_i^1 = 1$ | complier | always-taker |

- Never-taker: those who never take the treatment regardless of assignment
- Always-taker: those who always take the treatment regardless of assignment
- Complier: those who would always take the treatment if assigned to, and would not if not assigned to treatment
- Defier: those who would always take the treatment if not assigned to, and vice versa

# Compliance type in observed data

- We only observe $(D_i, Z_i)$, but each time we only observe one of $D_i^0$ and $D_i^1$, never both
- Each observed subgroup is a mix of two types

| | $Z_i = 0$ | $Z_i = 1$ |
|---|---|---|
| $D_i = 0$ | never-taker/complier | defier/never-taker |
| $D_i = 1$ | always-taker/defier | always-taker/complier |

# Assumption 4: Monotonicity/No-Defiers

- Assumption 4: Monotonicity/No-Defiers
- In math:

$$D_i^1 \geq D_i^0$$

- That is, potential treatment if $i$ were assigned to treatment should be no less than the potential treatment if $i$ were not assigned to treatment (hence monotonicity)
- In plain language, no defiers
  - In other words, if you are not assigned to treatment, you cannot take the treatment
- This assumption looks very different from the constant effect assumption (Assumption 4) of traditional IV
- And it is easier to understand

# Compliance type in observed data

- With no-defier assumption:

| | $Z_i = 0$ | $Z_i = 1$ |
|---|---|---|
| $D_i = 0$ | never-taker/complier | never-taker |
| $D_i = 1$ | always-taker | always-taker/complier |

- Actual treated units are a mix of always takers and compliers
- Actual control units are a mix of never takers and compliers

# LATE Theorem

- Angrist, Imbens, and Rubin, 1996, JASA
- With Assumptions 1 - 4, Wald estimator equals to average treatment effect for compliers

$$\frac{\hat{E}(Y|Z=1)-\hat{E}(Y|Z=0)}{\hat{E}(D|Z=1)-\hat{E}(D|Z=0)} = E(Y^1-Y^0|\mathit{complier}) = \text{LATE} \quad (6)$$

- IV estimates average complier treatment effect (or LATE, local average treatment effect)
- Treated units are a mix of compliers and always-takers
- In general, $LATE \neq ATT = E(Y^1 - Y^0|D=1))$
- Traditional IV "cheats" by assuming constant treatment effect, thus forcing the effect to be the same for compliers and always-takers

# Compilers, always-takers and never-takers in randomized experiments

- In the Vitamin A example, always taker will be people who always take Vitamin A regradless of whether they were assigned to treatment; and never-taker are those who will never take; both can exist
- Treated units are a mix of compliers and always-takers
- If there is no always-taker, $LATE = ATT$
  - How can you reduce the number of always-takers, if you are implementing the vitamin A experiment?
- Further, if there also no never-takers, $LATE = ATE$
  - How can you reduce the number of never-takers?

# Compilers, always-takers and never-takers in randomized experiments

- In medical trials for new drugs, there may be no always-taker
  - Only people who were assigned to treatment were given new drugs
  - But never-takers may exist: patients refuse to take new drugs

| | $Z_i = 0$ | $Z_i = 1$ |
|---|---|---|
| $D_i = 0$ | never-taker/complier | never-taker |
| $D_i = 1$ | | complier |

- $LATE = ATT$ when there is no always-taker

# Compliers, always-takers and never-takers in observational studies

- In observational studies, you have much less control over the data generating process
- Compliers may be only a small subset of the population
- Back to Angrist and Krueger (1991); instrument is quarter of birth, and actual treatment is school length
- Let us have a thought exercise
- Who are compliers?
  - people would be actually utilize the "advantage" that they were born in the 1st quarter to dropout after 16

# ITT vs LATE

- To understand why LATE may not generalize to the whole population, there is an alternative view
- ITT (intent-to-treat effect; effect on treatment assignment; it's the reduced-form using traditional IV's language)

$$\begin{aligned} \text{LATE} &= \frac{\hat{E}(Y|Z=1) - \hat{E}(Y|Z=0)}{\hat{E}(D|Z=1) - \hat{E}(D|Z=0)} \\ &= \frac{ITT}{\hat{E}(D|Z=1) - \hat{E}(D|Z=0)} \end{aligned} \tag{7}$$

- $\hat{E}(D|Z=1) - \hat{E}(D|Z=0)$ = (% treated among those assigned are assigned to treatment) - (% treated among those assigned to the control group)
- Estimate of LATE is almost always larger than ITT

# ITT vs LATE

- If there is no always-taker (e.g., the new drug example), $E(D|Z=0)=0$
- This is known as one-sided non-compliance
- Theorem 4.4.2 in MHE:

$$ATT = LATE = \frac{ITT}{\hat{E}(D|Z=1)} = \frac{ITT}{\text{compliance rate}} \tag{8}$$

- Estimate of ATT is almost always larger than ITT
- And if compliance rate is low, ATT will be a lot bigger than ITT
  - That is, even among the group that were assigned to treatment, due to self-selection into taking the treatment, ATT will be a lot bigger than ITT

# ITT vs LATE

- For some applications, especially policy evaluation for real field experiments
- ITT might be more important than LATE
- E.g., the Vitamin A experiment
- LATE: whether there is an effect for those who want to take it. A pure scientific quantity
- ITT: if gov want to implement the policy (e.g., providing Vitamin A for everyone), they must take into consideration that some people may not follow the instruction. So ITT provides a more faithful evaluation of the effect on the population affected by the policy

- So be careful how compliers differ from the general population if you care more about ITT, instead of LATE

# Continuous treatment

- The counterfactual framework of IV is built upon binary treatment
- It can be generalized to continuous treatment case
- Continuous treatment example: school length
  - LATE: the effect for compliers at point $s$ (those driven by instrument from just below $s$ to at least $s$)
  - And 2SLS estimate is a weighted mean of all LATE at different $s$ values
- Essentially what you get from 2SLS estimate is some extension of LATE
- Read MHE 4.5 for more extensions
  - Also continuous instrument/multiple instruments/adding covariates/HTE with IV