Microeconometrics (Causal Inference)
Weeks 7 and 8 - Instrumental variables

Joshua D. Merfeld
KDI School of Public Policy and Management

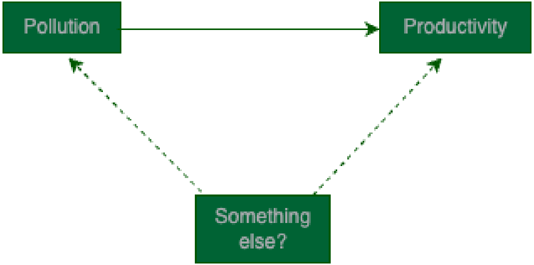
2023-10-19

What are we doing today?

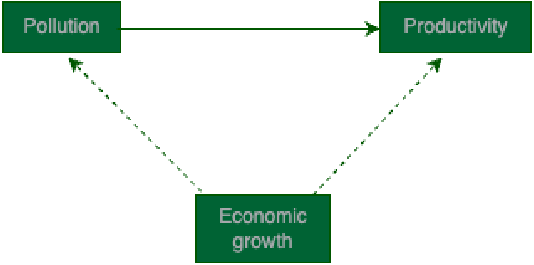
- Introduction to IVs
  - Requirements/assumptions
- IVs and RCTs
- In a world of LATE
- Weak instruments

- Instrumental variables (IVs) are a way to estimate causal effects when we have endogeneity
  - The endogeneity can take many forms: omitted variables, measurement error, simultaneity, etc.

- Consider my paper: effects of pollution on agricultural productivity
  - What's the problem with simply regression productivity on pollution?


Pollution
Productivity
Something else?

Endogeneity in the pollution example



Differences in differences?

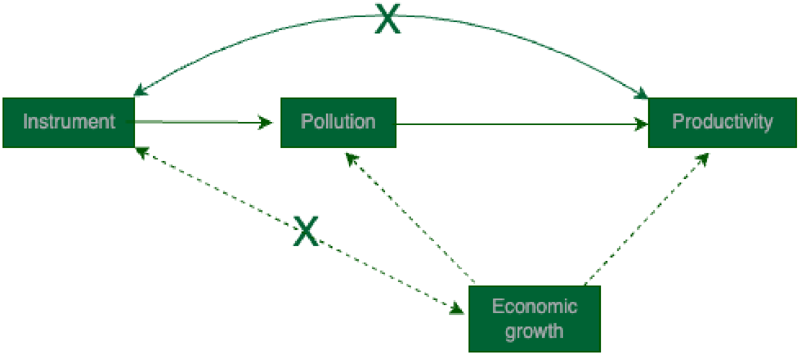
- One solution is to use a differences-in-differences (DiD) approach
- This requires the assumption of parallel trends
  - That is, the trends in the outcome variable would have been the same in the absence of the treatment
- But what if changing economic growth is leading to changes in both pollution and productivity?
  - Then the parallel trends assumption is violated since areas with more pollution are also experiencing faster economic growth

Control for growth?

- If you're willing to make assumptions about what the omitted variables are, maybe you could control for theme
- But this is a strong assumption
    - No matter what we do, we'll have to make assumptions, though

Enter: instruments

- Let's take a different approach
- We'll use an instrument
  - A variable that is correlated with the endogenous variable (pollution) but is not correlated with the error term


X
Instrument
Pollution
Productivity
X
Economic growth

- I very purposefully created the example so that the instrument is correlated with pollution
  - But it's not *directly* correlated with productivity
  - And it's not correlated with the omitted variable (the error term... will show you this in a second)

- Let's look at these more formally

- What we really want to estimate is this:

$$productivity_{it} = \beta_0 + \beta_1 pollution_{it} + \beta_2 X_{it} + \epsilon_{it} \tag{1}$$

- But if we don't properly control for everything, we are really estimating this:

$$productivity_{it} = \tilde{\beta}_0 + \tilde{\beta}_1 pollution_{it} + \eta_{it}, \tag{2}$$

  where $\eta_{it} = \beta_2 X_{it} + \epsilon_{it}$.

- Note that $\beta \neq \tilde{\beta}$, so we have a problem of endogeneity.

$$productivity_{it} = \tilde{\beta}_0 + \tilde{\beta}_1 pollution_{it} + \eta_{it} \tag{2}$$

- Can we estimate a version of equation 2 – that is, without controlling for X_{it} – and still get causal effects?
- Maybe, if we can find a valid instrument.
- So what makes an instrument valid?

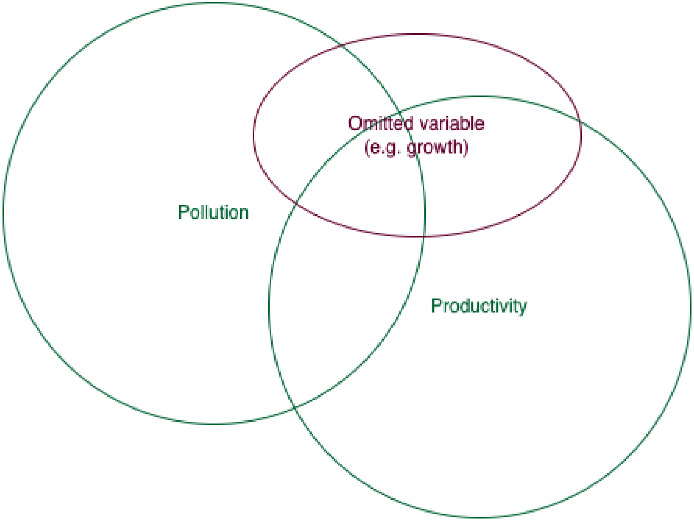
$$productivity_{it} = \tilde{\beta_0} + \tilde{\beta_1} pollution_{it} + \eta_{it} \quad (2)$$

1. The instrument must be correlated with the endogenous variable (pollution)
2. The instrument must not be correlated with the error term (η_{it})
   - Note that this implies two things:
     - The instrument must not be correlated with any omitted variable (here X_{it})
     - The instrument must not directly affect the outcome ($productivity_{it}$)

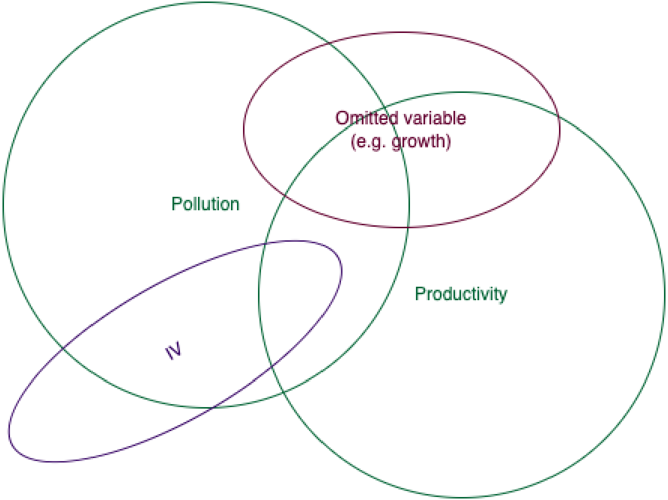
Using an instrument

- If we can find a valid instrument, we can use it to estimate the causal effect of pollution on productivity
- The simplest example uses two stages:
  1. $pollution_{it} = \pi_0 + \pi_1 instrument_{it} + \nu_{it}$
  2. $productivity_{it} = \phi_0 + \phi_1 pollution_{it} + \zeta_{it}$
- We can then estimate ϕ_1 using OLS
  - Note that only under certain circumstances will $\phi_1 = \beta_1$
  - More on this later

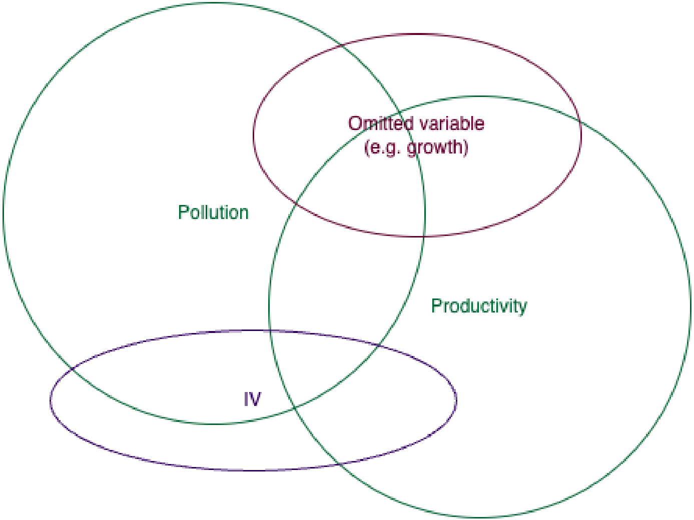
The intuition with venn diagrams



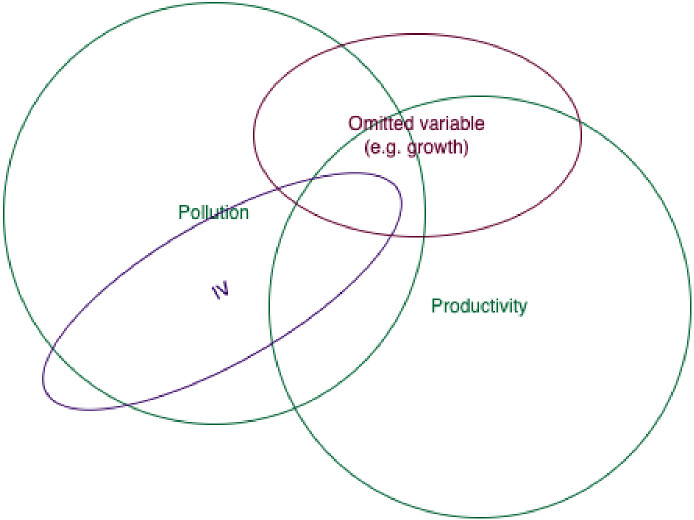
The IV only affects productivity through pollution



This doesn't work. Direct effects on productivity!



This doesn't work. Correlated with growth!



Back to our "two stages", redefining names

$$\text{Stage 1}: \quad T_{it} = \pi_0 + \pi_1 Z_{it} + \nu_{it}$$
$$\text{Stage 2}: \quad Y_{it} = \phi_0 + \phi_1 T_{it} + \zeta_{it}$$

- Requirements:
  - $cov(Z_{it}, T_{it}) \neq 0$
  - $cov(Z_{it}, \zeta_{it}) = 0$

- We first regress T on the instrument to get $\hat{T}_{it}$
- Then, we use the predicted values of T to estimate the effects on Y
  - If the IV is valid, these predicted values *are unrelated to the omitted variables!*

$$\text{Stage 1}: \quad T_{it} = \pi_0 + \pi_1 Z_{it} + \nu_{it}$$

$$cov(Z_{it}, T_{it}) \neq 0 \tag{3}$$

- This is the first requirement
- We can test this!
  - F-test of all *excluded instruments* in the first stages
  - I say all excluded instruments because you can technically have more than one

Some comments

$$\text{Stage 1}: \quad T_{it} = \pi_0 + \pi_1 Z_{it} + \nu_{it}$$
$$\text{Stage 2}: \quad Y_{it} = \phi_0 + \phi_1 T_{it} + \zeta_{it}$$

$$cov(Z_{it}, \zeta_{it}) = 0 \tag{4}$$

- This is the second requirement
- We cannot explicitly test this
  - This is an identifying *assumption*
  - We need this to be true to attribute causality to the second stage