Instrumental Variables

POST 8000 – Foundations of Social Science Research for Public Policy

Steven V. Miller

Department of Political Science

Goal for Today

Introduce students to instrumental variable analysis.

The Problem, in a Simple Question

Does more education cause higher earnings?

- Of interest to policymakers who (well, should) incentivize more access to better education.

Simply, the intuition:

$$\text{Earnings}_i = \alpha + \beta_1 * \text{Education}_i + \epsilon_i$$

What could be the problem here?

The Problem

- Omitted variable bias
- Selection bias

We contend education is exogenous to earnings, but it's endogenous to a variety of factors.

What is Endogeneity?

Endogeneity is when a covariate in the regression is correlated with the error term (ϵ).
Causes:

- Omitted variable (easy fix, if you have it)
- Measurement error (a “use your head” problem)
- Simultaneity (a bit trickier, but there are solutions)

The consequence of endogeneity is bias, either:

- Rejecting a (null) hypothesis that in fact is true (Type 1)
- Failing to reject a (null) hypothesis that you in fact could (Type 2)

On Education and Earnings

There's surely some exogenous part of education on earnings, but:

- Both education and earnings are a function of "instrinsic ability."
- "Intrinsic ability" determines motivation to pursue more education.
- "Intrinsic ability" influences the wages you obtain.

What Can You Do About Endogeneity?

- Nothing? (Don't...)
- Panel data? (If it's obtainable/feasible/appropriate?)
- Make the unobserved observable? (but what if that's not the problem?)
- Apply instrumental variable (IV) regression? (why not?)

Assumptions

IV regression is ideal for a particular kind of endogeneity. Assumptions:

1. Relevance
2. Exclusion
3. Exogeneity

Relevance

In a simple x (predictor), y (outcome), and z (instrument) setup, z must be correlated x.

- If it didn't, nothing would be gained from considering it in this setup.
- If it were uncorrelated with x, but correlated with y, we have something closer to the ideal setup.

Exclusion

z cannot be correlated with y.

- This is the "only through" language.
- z affects y "only through" its correlation with x.

Absent this exclusion restriction, the x, y, and z setup reduces to an omitted variable problem.

Exogeneity

z is independent of all other factors and is randomly assigned.

The “Huh?” Factor

IVs in the social sciences range from the intuitive to the weird:

- Card (1990): the draft (z) fixes the relationship between military service (x) and earnings (y).
- Levitt (1985): election cycles (z) fix relationship between police patrol hours (x) and the crime rate (y).
- Various: cigarette taxes (z) fix relationship betwen smoking (x) and various health outcomes (y).
- Miguel et al. (2004): rainfall (z) fixes relationship between economic shocks (x) and civil conflict (y).

The “Huh?” Factor

If it’s obvious that it should matter...

- isn’t it already in the model?
- isn’t it also correlated with y (violation of exclusion restriction)?

And if it’s not obvious...

- how the blue hell did you think of it/find it?
- how do you defend the exclusion restriction?

Via Scott Cunningham (p. 213 of *The Mixtape*)

But, let's say you think you do have a good instrument. How might you defend it as such to someone else? A necessary but not a sufficient condition for having an instrument that can satisfy the exclusion restriction is if people are confused when you tell them about the instrument's relationship to the outcome. Let me explain. No one is going to be confused when you tell them that you think family size will reduce female labor supply. They don't need a Becker model to convince them that women who have more children probably work less than those with fewer children. It's common sense. But, what would they think if you told them that mothers whose first two children were the same gender worked less than those whose children had a balanced sex ratio? They would probably give you a confused look. What does the gender composition of your children have to do with whether a woman works?

A Twitter Discussion

scott cunningham
@causalinf

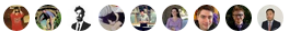
Follow

The reason I think this is because an instrument doesn't belong in the structural error term and the structural error term is all the intuitive things that determine your outcome. So it *must* be weird, otherwise it's probably in the error term.

> **Erik Thulin** @EThulin
> I had never heard of @causalinf's "weirdness" condition for instrumental variables, but I am certainly going to use it when I explain them from here on! twitter.com /andrewheiss/st…

8:48 PM - 11 Nov 2019 from Waco, TX

4 Retweets **42** Likes



A Twitter Discussion (continued)

Andrew Heiss @andrewheiss · 11 Nov 2019

Replying to @causalinf

Like parent education as an instrument for education. None of my students bought that as a valid instrument bc it obviously has a connection to, like, education and earnings. Too easy of an instrument

2

Alton B.H. Worthington @abhworthington · 11 Nov 2019

Replying to @causalinf

so, and I cannot believe I am suggesting this: is this a suggestion for actual use of the spurious correlations examples here?
tylervigen.com/spurious-corre…

2 5

Andrew Heiss @andrewheiss · 11 Nov 2019

omg

1 2

Using Instrumental Variables

IV regression is a lot simpler than it lets on.

- First: regress x on z as a "first-stage" regression.
- Extract fitted values from that regression.
  - These fitted values are effectively "decontaminated" of the source of endogeneity.
- Second: regress y on those fitted values of x from the first-stage.

You'll see this described as **two-stage least squares (2SLS)** regression.

Thus, y_2 in X should be expressed as a linear projection, and other independent variables in X should be expressed by itself. $P_Z = Z(Z'Z)^{-1}Z'$ is a n-by-n symmetric matrix and idempotent (i.e., $P_Z'P_Z = P_Z$). We use $\hat{X}$ as instruments for X and apply the IV estimation as in

$$\begin{aligned}\hat{\beta}_{2SLS} &= (\hat{X}'\hat{X})^{-1}\hat{X}'Y \\ &= (X'P_Z X)^{-1}X'P_Z Y \\ &= (X'Z(Z'Z)^{-1}Z'X)^{-1}X'Z(Z'Z)^{-1}Z'Y \end{aligned} \quad (2)$$

This can be also written as

$$\hat{\beta}_{2SLS} = (\hat{X}'\hat{X})^{-1}\hat{X}'Y$$

```r
# First-stage model...
FSM <- lm(treat ~ control + instr, data=Data)

# Generate treat_hat variable
Data$treat_hat <- fitted(FSM)

# Second-stage model...
SSM <- lm(y  ~ control + treat_hat, data=Data
```

Table of Contents