INSTRUMENTAL VARIABLES

Technical Track Session IV

Emanuela Galasso
The World Bank

These slides were developed by Christel Vermeersch and modified by Emanuela Galasso for the purpose of this workshop

An example to start off with…

- Say we wish to evaluate a voluntary job training program
  - Any unemployed person is eligible
  - Some people choose to register (Participants)
  - Other people choose not to register (Non-participants)

- Some simple (but not-so-good) ways to evaluate the program:
  - Compare before and after situation in the participant group
  - Compare situation of participants and non-participants after the intervention
  - Compare situation of participants and non-participants before and after.

Voluntary job training program

Say we decide to compare outcomes for those who participate to the outcomes of those who do not participate:

- A simple model to do this:

$$y = \alpha + \beta_1 P + \beta_2 x + \varepsilon$$

$$P = \begin{cases} 1 & \text{If person participates in training} \\ 0 & \text{If person does not participate in training} \end{cases}$$

$x =$ Control variables (exogenous & observed)

- Why is this not working? 2 problems:
  - Variables that we omit (for various reasons) but that are important
  - Decision to participate in training is endogenous.

Problem #2: Endogenous Decision to Participate

- True model is:

$$y = \gamma_0 + \gamma_1 x + \gamma_2 P + \eta$$

with

$$P = \pi_0 + \pi_1 x + \pi_2 M_2 + \xi$$

M_2 = Vector of unobserved / missing characteristics
(i.e. we don't fully know why people decide to participate)

- Since we don't observe M_2, we can only estimate a simplified model:

$$y = \beta_0 + \beta_1 x + \beta_2 P + \varepsilon$$

- Is $\beta_{2,\,OLS}$ an unbiased estimator of γ_2?

Problem #2: Endogenous Decision to Participate

- We estimate: $y = \beta_0 + \beta_1 x + \beta_2 P + \varepsilon$
- But true model is: $y = \gamma_0 + \gamma_1 x + \gamma_2 P + \eta$

  with $P = \pi_0 + \pi_1 x + \pi_2 M_2 + \xi$
- Is $\beta_{2,\ OLS}$ an unbiased estimator of γ_2?

  Corr (ε, P) $= corr\ (\varepsilon, \pi_0 + \pi_1 x + \pi_2 M_2 + \xi)$

  $= \pi_1 \ corr\ (\varepsilon, x) + \pi_2 \ corr\ (\varepsilon, M_2)$

  $= \pi_2 \ corr\ (\varepsilon, M_2)$
- If there is a correlation between the missing variables that determine participation (e.g. Talent) and outcomes not explained by observed characteristics, then the OLS estimator will be biased.

HUMAN DEVELOPMENT NETWORK

What can we do to solve this problem?

- We estimate: $y = \beta_0 + \beta_1 x + \beta_2 \boldsymbol{P} + \boldsymbol{\varepsilon}$
- So the problem is the correlation between ***P*** and ***ε***
- How about we replace ***P*** with "something else", call it ***Z***:
  - ***Z*** needs to be similar to ***P***
  - But is not correlated with ***ε***

HUMAN DEVELOPMENT NETWORK

Back to the job training program

- ***P*** = participation
- ***ε*** = that part of outcomes that is not explained by program participation or by observed characteristics
- I'm looking for a variable ***Z*** that is:
  (1) Closely related to participation *P*
  (2) but doesn't directly affect people's outcomes *Y, other than through its effect on participation.*
- So this variable must be coming from outside.

HUMAN DEVELOPMENT NETWORK

An outside variable for the job training program

- Say that a social worker visits unemployed persons to encourage them to participate.
  - She only visits 50% of persons on her roster, and
  - She randomly chooses whom she will visit
- If she's got at it, many people she visits will enroll.
  There will be a correlation between receiving a visit and enrolling
- But visit does not have direct effect on outcomes (e.g. income) apart from its effect through enrollment in the training program.

HUMAN DEVELOPMENT NETWORK

An outside or instrumental variable

- Define a new variable *Z*

$$Z = \begin{cases} 1 & \text{If person was randomly chosen to receive the encouragement visit from the social worker} \\ 0 & \text{If person was randomly chosen not to receive the encouragement visit from the social worker} \end{cases}$$

- $Corr\ (Z, P) > 0$
  People who receive the encouragement visit are more likely to participate than those who don't
- $Corr\ (Z, \varepsilon) = 0$
  No correlation between receiving a visit and benefit to the program apart from the effect of the visit on participation.
- *Z* is called an instrumental variable

Two-stage least squares *(2SLS)*

Remember the original model with endogenous *P*:

$$y = \beta_0 + \beta_1 x + \beta_2 P + \varepsilon$$

Step 1

Regress the endogenous variable *P* on the instrumental variable(s) *Z* and other exogenous variables

$$P = \delta_0 + \delta_1 x + \delta_2 Z + \tau$$

- Calculate the predicted value of *P* for each observation: *P*
- Since *Z* and *x* are not correlated with ε, neither will be *P*.
- You will need one instrumental variable for each potentially endogenous regressor.

Two-stage least squares *(2SLS)*

Step 2

Regress *y* on the predicted variable *P* and the other exogenous variables

$$y = \beta_0 + \beta_1 x + \beta_2 P + \varepsilon$$

- Note: The standard errors of the second stage *OLS* need to be corrected because *P* is not a fixed regressor.
- In Practice: Use *STATA* ivreg command, which does the two steps at once and reports correct standard errors.
- Intuition: By using *Z* for *P*, we cleaned *P* of its correlation with η
- It can be shown that (under certain conditions) $\beta_{2,IV}$ yields a consistent estimator of γ_2 (large sample theory)

Where do we find instrumental variables?

- Searching for one… hard!
- Creating one with information
  - If everyone is eligible to participate in treatment
  - But some have more information than others (Who has more information will be more likely to participate)
  - Provision of "additional information" on a random basis

Example 1: voluntary job training program

Population eligible for job training program

↓

Random Sample

Randomized assignment

Standard Information Package only	**Standard Information Package + Additional visit**
Monthly income 1 year later = 700	Monthly income 1 year later = 850
30% take-up	90% take-up

Question: what is the impact of the job training program?

Standard Information Package only	**Standard + Additional Information Package**
Monthly income 1 year later = 700	Monthly income 1 year later = 850
30% take-up	90% take-up

Question: what is the impact of the job training program?

Difference between the “well informed” and “not well informed” group:

……………………………………………………………………………………

Corrected for the differential take-up rate:

……………………………………………………………………………………

Practically:
Impact = ………………………………………………………………………

Link back to the estimation formula

Stage 1

- Regress the participation on training on a dummy for whether person received additional visit (linear model)
- Compute predicted value of participation

Stage 2

Regress wages on the predicted value of participation

Example 2: School autonomy in Nepal

Goal

To Evaluate:

A. Autonomous school management by communities
B. School report cards

Data

- You can include 1000 schools in the evaluation
- Each community freely chooses to participate or not
- School report cards done by NGOs
- Each community has exactly one school

Task

Design the implementation of the program so it can be evaluated –propose method of evaluation.

School autonomy in Nepal

		Intervention B: School report card intervention by NGO.		
		Yes	No	Total
Instrumental variable for Intervention A: NGO visits community to inform on procedures for transfer of the school to community management.	Yes	300	300	600
	No	200	200	400
	Total	**500**	**500**	**1000**

Reminder and a word of caution…

- *corr (Z,ε) =0*
  - If *corr (Z , ε) ≠ 0*, "Bad instrument"
  - "Finding" a good instrument is hard!
  - But you can build one yourself with a randomized encouragement design
- *corr (Z,P) ≠0*
  - "Weak instruments": the correlation between *Z* and *P* needs to be sufficiently strong.
  - If not, the bias stays large even for large sample sizes.

Back to *ATE* and *TOT*

- Sometimes eligible units are selected randomly into the treatment group, are offered treatment, but not all of them accept it.

- Computing the Average Treatment Effect (ATE)
  Straight difference in average outcomes between the group to whom you offered treatment, and the group to whom you did not offer treatment

- Computing the Effect of Treatment on the Treated (TOT)
  Use the randomized offering as an instrumental variable *(Z)* for whether people accepted the treatment *(P)*

HUMAN DEVELOPMENT NETWORK

Note: IV is a 'local' effect

- IVE identifies the average gains to persons induced to change their choice by a change of the instrument (referred to as compliers)
  - … however we cannot identify who these people are ("local average treatment effect" or LATE)
  - … different instruments will identify different parameters and answer different questions
- Care in extrapolating to the whole population

HUMAN DEVELOPMENT NETWORK

HUMAN DEVELOPMENT NETWORK

SIEF

Thank You

HUMAN DEVELOPMENT NETWORK

SIEF

Q & A