Session IV
Instrumental Variables

Christel M. J. Vermeersch

January 2008

Human Development Network
Middle East and North Africa Region
Spanish Impact Evaluation Fund

An example to start off with…

- Say we wish to evaluate a voluntary job training program
  - Any unemployed person is elegible
  - Some people choose to register (“Treatments”)
  - Other people choose not to register (“Comparisons”)
- Some simple (but not-so-good) ways to evaluate the program:
  - Compare before and after situation in the treatment group
  - Compare situation of treatments and comparisons after the intervention
  - Compare situation of treatments and comparisons 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:

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

Where $T = 1$ if person participates in training

$T = 0$ if person does not participate in training

x = control variables (exogenous & observable)

Why is this not working properly?

HDN SAR WBI

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:

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

Where $T = 1$ if person participates in training

$T = 0$ if person does not participate in training

x = control variables (exogenous & observable)

Why is this not working properly? 2 problems:

- ▹ Variables that we omit (for various reasons) but that are important
- ▹ Decision to participate in training is endogenous

Problem #1: Omitted Variables

Even in a well-thought model, we'll miss

▹ "forgotten" characteristics: we didn't know they mattered

▹ characteristics that are too complicated to measure

Examples :

▹ Varying talent and levels of motivation

▹ Different levels of information

▹ Varying opportunity cost of participation

▹ Varying level of access to services

The full "correct" model is: $y = \alpha + \gamma_1 T + \gamma_2 x + \gamma_3 D + \eta$

The model we use: $y = \alpha + \beta_1 T + \beta_2 x + \varepsilon$

HDN SAR WBI

Problem #2: Endogenous Decision to Participate

Participation is a decision variable ? It's endogenous!

(I.e. it depends on the participants themselves.)

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

$$T = \pi + \pi_2 D + \xi$$

Problem #2: Endogenous Decision to Participate

Participation is a decision variable ? it's endogenous

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

$$T = \pi + \pi_2 D + \xi$$

$$\Rightarrow y = \alpha + \beta_1(\pi + \pi_2 D + \xi) + \beta_2 x + \varepsilon$$

$$\Rightarrow y = \alpha + \beta_1 \pi + \beta_2 x + \beta_1 \pi_2 D + \beta_1 \xi + \varepsilon$$

Note: in the two cases: we're missing the D term,

which we need to estimate the correct model,

but which we cannot measure properly.

- The "correct model is: $y = \alpha + \gamma_1 T + \gamma_2 x + \gamma_3 D + \eta$
- Simplified model: $y = \alpha + \beta_1 T + \beta_2 x + \varepsilon$

- Say we estimate the treatment effect γ_1 with $\beta_{1,OLS}$
- If D is correlated with T, and we don't include D in the simplified model, then the estimator of the parameter on T will pick up part of the effect of D. This will happen to the extent that D and T are correlated.
- Thus: our OLS estimator $\beta_{1,OLS}$ of the treatment effect γ_1 captures the effect of other characteristics (D) in addition to the treatment effect.
- This means that there is a difference between $E(\beta_{1,OLS})$ y γ_1

  → the expected value of the OLS estimator β_1 isn't γ1, the real treatment effect

  → $\beta_{1,OLS}$ is a biased estimator of the treatment effect γ_1.

CV4

CV4 $\text{corr}(T,\varepsilon)=\text{corr}(T, \gamma3*D+\eta) = \gamma 3*\text{corr}(T, D)$

wb226893, 05/23/2006

- The "correct model is: $y = \alpha + \gamma_1 T + \gamma_2 x + \gamma_3 D + \eta$
- Simplified model: $y = \alpha + \beta_1 T + \beta_2 x + \varepsilon$

- This means that there is a difference between $E(\beta_{1,OLS})$ y γ_1

  → the expected value of the OLS estimator β_1 isn't γ1, the real treatment effect

  → $\beta_{1,OLS}$ is a biased estimator of the treatment effect γ_1.

- Why did this happen?
  - One of the basic conditions for OLS to be BLUE was violated:

- In other words $E(\beta_{1,OLS}) \neq \gamma_1$ (biased estimator)
- Even worse…..$plim(\beta_{1,OLS}) \neq \gamma_1$ (inconsistent estimator)

CV6

CV6 $\text{corr}(T,\varepsilon)=\text{corr}(T, \gamma 3*D+\eta) = \gamma 3*\text{corr}(T, D)$

wb226893, 05/23/2006

What can we do to solve this problem?

$$y = \alpha + \gamma_1 T + \gamma_2 x + \gamma_3 D + \eta$$

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

- Try to clean the correlation between T and **ε:**
- Isolate the variation in T that is not correlated with **ε** through the omitted variable D
- We can do this using an instrumental variable (IV)

Basic idea behind IV

$$y = \alpha + \gamma_1 T + \gamma_2 x + \gamma_3 D + \eta$$

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

- The basic problem is that corr (T,D) ≠ 0
- Find a variable Z that satisfies two conditions:
  1. Correlated with T: corr $(Z, T) \neq 0$

     ---- Z and T are correlated, or Z predicts part of T
  2. Z is not correlated with **ε**: corr $(Z, \varepsilon) = 0$

     ----By itself, Z has no influence on y. The only way it can influence y is because it influences *T*. All of the effect of *Z* on *y* passes through *T*.
- Examples of Z in the case of voluntary job training program?

Two-stage least squares (2SLS)

Remember the original model with endogenous T:

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

Step 1: Regress the endogenous variable T on the instrumental variable(s) Z and other exogenous variables

$$T = \delta_0 + \delta_1 x + \theta_1 Z + \tau$$

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

Two-stage least squares (2SLS)

Step 2: Regress y on the predicted variable $\hat{T}$ and the other exogenous variables

$$y = \alpha + \beta_1 \hat{T} + \beta_2 x + \varepsilon$$

- Note: the standard errors of the second stage OLS need to be corrected because $\hat{T}$ 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 T, we cleaned T of its correlation with ε.
- It can be shown that (under certain conditions) IV yields a consistent estimator of γ_1 (large sample theory)

Uses of instrumental variables

- Simultaneity: X and Y cause each other
  - instrument X

- Omitted Variables: X is picking up the effect of other variables which are omitted
  - instrument X with a variable that is not correlated with the omitted variable(s)

- Measurement Error: X is not measured precisely
  - instrument X

Examples

- Hoxby (2000) and Angrist (1999)
  - The effect of class size on learning achievements
- Chaudhury, Gertler, Vermeersch (work in progress)
  - Estimating the effect of school autonomy on learning in Nepal

School autonomy in Nepal

- Goal is to evaluate
  - A. Autonomous school management by communities
  - B. School report cards
  - C. School information networks
- Data
  - We could integrate about 1500 schools in the evaluation
  - Each community freely chooses to participate or not.
  - School report cards and school networks done by NGOs
    - School network can only be done in autonomous schools
    - School report cards can be done in any school.

School autonomy in Nepal

- Goal is to evaluate
  - A. Autonomous school management by communities
  - B. School report cards
  - C. School information networks
- Data
  - We could integrate about 1500 schools in the evaluation
  - Each community freely chooses to participate or not.
  - School report cards and school networks done by NGOs
    - School network can only be done in autonomous schools
    - School report cards can be done in any school.
  - Assume each community has exactly one schools
- Task: design the implementation of the program so it can be evaluated – propose method of evaluation.

School autonomy in Nepal

Interventions: A. Autonomous school management by communities
B. School report cards
C. School information networks

	Creation of demand for devolution to the community (A)	Feedback of performance indicators (B)	Network support after devolution to the community (C)	Number of schools in the group
Group O (Control)	No	No	No	200
Group A	Yes	No	No	300
Group B	No	Yes	No	100
Group AB	Yes	Yes	No	300
Group AC	Yes	No	Yes	300
Group ABC	Yes	Yes	Yes	300
Total				**1500**

Reminder and a word of caution….

- corr (Z,ε) =0
  - if corr (Z, ε) ≠ 0 "Bad instrument" ¡Problem!
  - ¡Finding a good instrument is hard!
  - ¡Use both theory and common sense to find one!
  - We can think of designs that yield good instruments.
- corr(Z,T) ≠0
  - "Weak instruments": the correlation between Z and T needs to be sufficiently strong.
  - If not, the bias stays large even for large sample sizes.