

# Econometrics 1

## Endogeneity and instrumental variables

Christelle Dumas

U. of Cergy-Pontoise

November 17, 2009

# Outline

- 1 Introduction
- 2 The canonical example: returns to education
- 3 A formalization with 1 instrument
- 4 Identification with multiple instruments: the 2SLS

# Correlation vs. causality

- Correlation is not causality
  - E.g. traffic-jams and police officers
- But correlates can sometimes provide pretty good evidence of a causal relation
  - Even when the variable of interest has not been manipulated by experimentation
  - How to do so?

# But what is causality?

- Causality is when we can attribute a change to a cause
- Thought experiment: everything stays the same but one variable
  - Have exactly twice the same individual but give one more year of schooling to one but not the other
  - What happens on their wages?
  - *Other things being equal*

# The effect of a change in price on supply or demand of a good

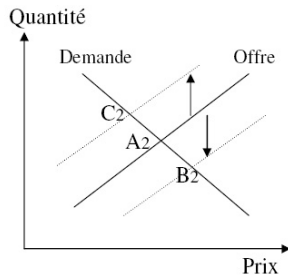
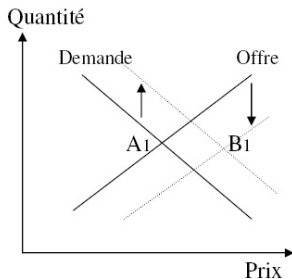
- Subsidize exports, tax cigarettes...
- Price and quantities exchanged such as market is cleared (supply = demand)

# How to estimate the slope of supply and demand curves?

Find something that affects

- Supply curve, leaving demand unchanged
- The reverse

If both move in the same time, we can't say anything on slope



## Example: fish market

Variable that shifts the supply curve but leaves the demand curve unchanged:

- Weather: stormy, clear...
- If the weather is stormy, then less supply, shifts the curve to the right: price is higher for the same quantity
- Angrist, Graddy, Imbens (2000) "The Interpretation of Instrumental Variables Estimators in Simultaneous Equations Models with an Application to the Demand for Fish", Restud.

## So far...

- Omitted variables can create biases
- Simultaneous equations are also a problem
  - Number of police officers depend on whether there is a traffic jam
  - Traffic jams depend on the number of police officers (and no additional controls can solve that)
  - Reverse causality
- Find variables that affect the RHS variable without impacting directly the LHS variable
- Instruments or instrumental variables

## In addition: measurement error

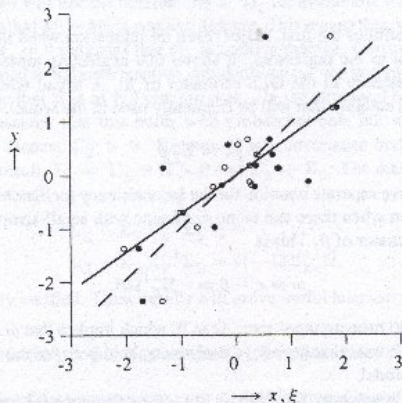
### Bias from measurement error

- $w = bS + cE + v$ ,  $Cov(E, v) = Cov(S, v) = 0$
- But experience is imprecisely measured (due to unemployment spells)
- True variable is  $E$  but what is observed is  $E_m = E + e$ , the estimated eq.:
  - $w = bS + c(E_m - e) + v = bS + cE_m + (v - ce)$
  - $E_m$  is correlated with residual, since correlated with  $e$
  - Hence bias

### Bias toward zero

- Variable with only measurement error
- Find no effect

# Graphical intuition



*The effect of measurement error: regression line based on data without measurement error (dashed line, open circles) and regression line based on data with measurement error (solid line, filled circles).*

# What we will see in this course

## IV and causality:

- Understanding the biases
- What is a good instrument
  - 2 conditions, equally important
- Mechanics of the estimation

## Left for future lectures:

- Does the instrument explain enough of the RHS variable?
- Is it excluded from the interest regression?
- What if the effect is not the same for everybody?
- Properties of the IV estimator?

# Outline

- 1 Introduction
- 2 **The canonical example: returns to education**
- 3 A formalization with 1 instrument
- 4 Identification with multiple instruments: the 2SLS

# Estimating returns to education

- $Y_i = \alpha + \rho S_i + \delta A_i + v_i = \alpha + \rho S_i + \eta_i$
- $\rho$  are returns to education,  $A_i$  are unobserved abilities.
- Let  $Z$  be a variable that affects education but is not correlated with any other determinant of wages:  $Cov(Z_i, \eta_i) = 0$
- "exclusion restriction" =  $Z$  should not enter equation

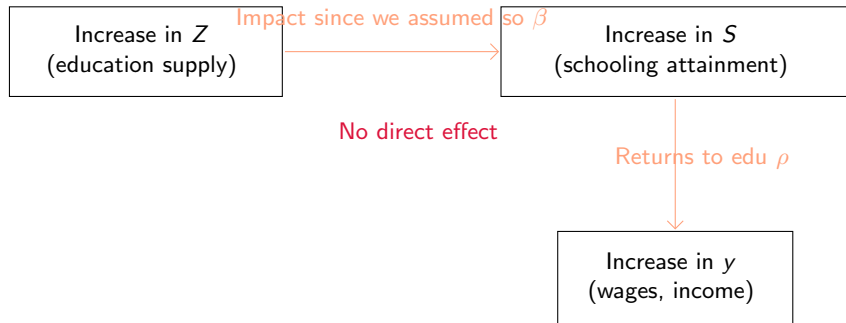
## Application: Duflo on Indonesia

- "Schooling and labor market consequences of school construction in Indonesia: Evidence from an unusual policy experiment", AER, 1991.
- Use of a major program of schools building that improved schooling attainment without affecting wages directly.

TABLE 4—EFFECT OF THE PROGRAM ON EDUCATION AND WAGES: COEFFICIENTS OF THE INTERACTIONS BETWEEN COHORT DUMMIES AND THE NUMBER OF SCHOOLS CONSTRUCTED PER 1,000 CHILDREN IN THE REGION OF BIRTH

	Observations	Dependent variable					
		Years of education			Log(hourly wage)		
		(1)	(2)	(3)	(4)	(5)	(6)
<i>Panel A: Experiment of Interest: Individuals Aged 2 to 6 or 12 to 17 in 1974</i>							
<i>(Youngest cohort: Individuals ages 2 to 6 in 1974)</i>							
Whole sample	78,470	0.124 (0.0250)	0.15 (0.0260)	0.188 (0.0289)			
Sample of wage earners	31,061	0.196 (0.0424)	0.199 (0.0429)	0.259 (0.0499)	0.0147 (0.00729)	0.0172 (0.00737)	0.0270 (0.00850)

# The mechanics



TOTAL EFFECT of  $Z$  on  $y$ :  $\rho \cdot \beta$ ,

- $\beta$  easily measured
- We can compute  $\rho$

Careful: does not work anymore if there is a direct effect or if  $\beta$  close to 0

# The reduced form equation

- Regression of  $S_i$  on  $Z_i$  is called the "first-stage":

$$S_i = K + \beta Z_i + \zeta_i$$

- Plugged into the interest regression:

$$\begin{aligned} Y &= \alpha + \rho(K + \beta Z_i + \zeta_i) + \eta_i \\ &= (\alpha + \rho K) + \rho\beta Z_i + (\rho\zeta_i + \eta_i) \\ &= K + \delta Z_i + \xi_i \end{aligned}$$

where  $Z$  not correlated with  $\zeta_i$  nor  $\eta_i$ .

- Called reduced-form equation
- Careful:  $Z$  has an impact on  $Y$ , but it only transits through  $S_i$ ; and no impact if  $\rho = 0$ .
- $\rho = \delta/\beta =$  estimate of the reduced form/estimate of the 1st-stage

# Recap of assumptions

- $Z$  must have an impact on  $S$  ( $\beta \neq 0$ )
  - Testable assumption
- Exclusion restriction: the only reason for the relationship between  $Y$  and  $Z$  is the 1st stage
  - Can we test this?

# Testing the exclusion restriction when only 1 instrument?

- Introduce  $Z$  in the interest regression and test whether its coef is 0.
- $Y_i = \alpha + \rho S_i + \mu Z_i + w_i$
- But cannot be estimated through OLS since  $S$  correlated with  $w$ .
- So, 1st-stage and plugged:

$$\begin{aligned} Y_i &= \alpha + \rho(K + \beta Z_i + \zeta_i) + \mu Z_i + w_i \\ &= (\alpha + \rho K) + (\rho\beta + \mu)Z_i + (\rho\zeta_i + w_i) \end{aligned}$$

You get the total effect of  $Z$  and you cannot disentangle the direct effect from the one that transits through  $S$ .

- = Identifying assumption

# Outline

- 1 Introduction
- 2 The canonical example: returns to education
- 3 A formalization with 1 instrument**
- 4 Identification with multiple instruments: the 2SLS

# The Wald estimator

The model:

- a single binary instrument
- one endogenous regressor
- no covariates = no control variables

$$Y = \alpha + \rho S_i + \eta_i$$

where  $S_i$  and  $\eta_i$  may be correlated. If we assume that  $E(\eta_i|Z_i) = 0$  then,

$$\begin{aligned} E(Y_i|Z_i) &= \alpha + \rho E(S_i|Z_i) \\ \rho &= \frac{E[Y_i|Z_i = 1] - E[Y_i|Z_i = 0]}{E[S_i|Z_i = 1] - E[S_i|Z_i = 0]} \end{aligned}$$

→ What is the Wald estimator?

## Let's think of sensible instruments

- testing (H2): t-test after OLS estimation
- testing (H1): not possible, involves unobservable

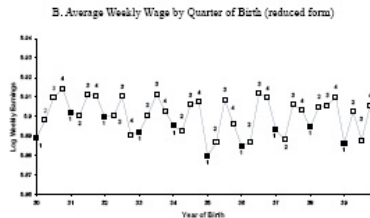
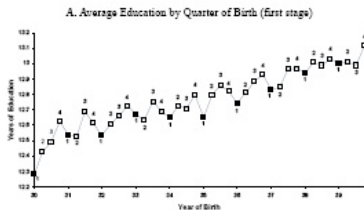
IV for education in a wage equation?

$Corr(educ, u) \neq 0$  due to unobserved abilities

- mother's education?
  - ok, partially correlated
  - but could be correlated with omitted factors in  $u$
- last digit of one's social security number
  - (H1) probably holds
  - (H2) does not
- Angrist and Krueger: quarter of birth (1 if born in 1st quarter of the year)
  - (H1): ok, a priori independent of ability
  - (H2): compulsory school attendance laws affect final level ?

# Angrist Krueger

"Does Compulsory School Attendance Affect Schooling and Earnings", 1991, QJE.



# Making sure the data fit with the story (1)

THE EFFECT OF QUARTER OF BIRTH ON VARIOUS EDUCATIONAL  
 OUTCOME VARIABLES

Outcome variable	Birth cohort	Mean	Quarter-of-birth effect <sup>a</sup>			<i>F</i> -test <sup>b</sup> [ <i>P</i> -value]
			I	II	III	
Total years of education	1930–1939	12.79	-0.124 (0.017)	-0.086 (0.017)	-0.015 (0.016)	24.9 [0.0001]
	1940–1949	13.56	-0.085 (0.012)	-0.035 (0.012)	-0.017 (0.011)	18.6 [0.0001]
High school graduate	1930–1939	0.77	-0.019 (0.002)	-0.020 (0.002)	-0.004 (0.002)	46.4 [0.0001]
	1940–1949	0.86	-0.015 (0.001)	-0.012 (0.001)	-0.002 (0.001)	54.4 [0.0001]
Years of educ. for high school graduates	1930–1939	13.99	-0.004 (0.014)	0.051 (0.014)	0.012 (0.014)	5.9 [0.0006]
	1940–1949	14.28	0.005 (0.011)	0.043 (0.011)	-0.003 (0.010)	7.8 [0.0017]
College graduate	1930–1939	0.24	-0.005 (0.002)	0.003 (0.002)	0.002 (0.002)	5.0 [0.0021]
	1940–1949	0.30	-0.003 (0.002)	0.004 (0.002)	0.000 (0.002)	5.0 [0.0018]
Completed master's degree	1930–1939	0.09	-0.001 (0.001)	0.002 (0.001)	-0.001 (0.001)	1.7 [0.1599]
	1940–1949	0.11	0.000 (0.001)	0.004 (0.001)	0.001 (0.001)	3.9 [0.0091]
Completed doctoral degree	1930–1939	0.03	0.002 (0.001)	0.003 (0.001)	0.000 (0.001)	2.9 [0.0332]
	1940–1949	0.04	-0.002 (0.001)	0.001 (0.001)	-0.001 (0.001)	4.3 [0.0050]

# Making sure the data fit with the story (2)

PERCENTAGE OF AGE GROUP ENROLLED IN SCHOOL BY BIRTHDAY AND LEGAL DROPOUT AGE<sup>a</sup>

Date of birth	Type of state law <sup>b</sup>		Column (1) - (2)
	School-leaving age: 16 (1)	School-leaving age: 17 or 18 (2)	
Percent enrolled April 1, 1960			
1. Jan 1-Mar 31, 1944 (age 16)	87.6 (0.6)	91.0 (0.9)	-3.4 (1.1)
2. Apr 1-Dec 31, 1944 (age 15)	92.1 (0.3)	91.6 (0.5)	0.5 (0.6)
3. Within-state diff. (row 1 - row 2)	-4.5 (0.7)	-0.6 (1.0)	-4.0 (1.2)
Percent enrolled April 1, 1970			
4. Jan 1-Mar 31, 1954 (age 16)	94.2 (0.3)	95.8 (0.5)	-1.6 (0.6)
5. Apr 1-Dec 31, 1954 (age 15)	96.1 (0.1)	95.7 (0.3)	0.4 (0.3)
6. Within-state diff. (row 1 - row 2)	-1.9 (0.3)	0.1 (0.6)	-2.0 (0.6)
Percent enrolled April 1, 1980			
7. Jan 1-Mar 31, 1964 (age 16)	95.0 (0.1)	96.2 (0.2)	-1.2 (0.2)
8. Apr 1-Dec 31, 1964 (age 15)	97.0 (0.1)	97.7 (0.1)	-0.7 (0.1)
9. Within-state diff. (row 1 - row 2)	-2.0 (0.1)	-1.5 (0.2)	0.5 (0.3)

# Results

PANEL A: WALD ESTIMATES FOR 1970 CENSUS—MEN BORN 1920–1929<sup>a</sup>

	(1) Born in 1st quarter of year	(2) Born in 2nd, 3rd, or 4th quarter of year	(3) Difference (std. error) (1) – (2)
ln (wkly. wage)	5.1484	5.1574	-0.00898 (0.00301)
Education	11.3996	11.5252	-0.1256 (0.0155)
Wald est. of return to education			0.0715 (0.0219)
OLS return to education <sup>b</sup>			0.0801 (0.0004)

Panel B: Wald Estimates for 1980 Census—Men Born 1930–1939

	(1) Born in 1st quarter of year	(2) Born in 2nd, 3rd, or 4th quarter of year	(3) Difference (std. error) (1) – (2)
ln (wkly. wage)	5.8916	5.9027	-0.01110 (0.00274)
Education	12.6881	12.7969	-0.1088 (0.0132)
Wald est. of return to education			0.1020 (0.0239)
OLS return to education			0.0709 (0.0003)

# Are natural experiments always needed?

Economists also use variation in prices

- issue: need to control for all (?)/lots of prices

Card (95) uses college proximity as an IV for education:

Table 3: Reduced Form and Structural Estimates of Education and Earnings Models

	Reduced Form Models:				Structural Models	
	Education		Earnings		of Earnings	
	(1)	(2)	(3)	(4)	(5)	(6)

A: Treat Experience and Experience Squared as Exogenous

1. Live Near College in 1966	0.320 (0.088)	0.322 (0.083)	0.042 (0.018)	0.045 (0.018)	--	--
2. Education	--	--	--	--	0.132 (0.055)	0.140 (0.055)
3. Family Background	no	yes	no	yes	no	yes

## Other example

Dumas & Lambert (08) use availability of education infrastructures for the parents (and birth order) as an IV to estimate the effect of parental education on schooling achievement of the children in Senegal.

⇒ need for control variables.

## A model with covariates

Equation of interest:

$$y = \beta_0 + \beta_1 x_1 + \dots + \beta_{K-1} x_{K-1} + \beta_K x_K + u \quad (1)$$

with  $E(u) = 0$ ,  $Cov(x_j, u) = 0$ ,  $j = 1, \dots, K - 1$  ( $H_0$ ). i.e. all the  $x$  are exogenous except  $x_K$ .

We assume:

$$Cov(z_1, u) = 0 \quad (H_1)$$

$x_1, x_2, \dots, x_{K-1}, z_1$  is the set of exogenous variables.  
and

$$x_K = \delta_0 + \delta_1 x_1 + \dots + \delta_{K-1} x_{K-1} + \theta_1 z_1 + r_K$$

where  $\theta_1 \neq 0$  ( $H_2$ )

i.e.  $z_1$  partially correlated with  $x_K$  once the other exogenous variables  $x_1, \dots, x_{K-1}$  have been netted out.

## IV solve the identification problem: light proof

$$y = \beta_0 + \beta_K x_K + u$$

$$x_K = \delta_0 + \theta_1 z_1 + r_K$$

hence:

$$\begin{aligned} \text{Cov}(y, z_1) &= \beta_0 \text{Cov}(z_1, 1) + \beta_K \text{Cov}(x_K, z_1) + \text{Cov}(u, z_1) \\ &= \beta_K \text{Cov}(x_K, z_1) \\ \beta_K &= \frac{\text{Cov}(y, z_1)}{\text{Cov}(x_K, z_1)} \quad \text{if } \text{Cov}(x_K, z_1) \neq 0 \\ &= \frac{\text{Cov}(y, z_1)/V(z_1)}{\text{Cov}(x_K, z_1)/V(z_1)} \quad \text{if } \text{Cov}(x_K, z_1) \neq 0 \\ &= \frac{\text{Reduced-form coeff}}{\text{1st-stage coeff}} \quad \text{if 1st-stage } \neq 0 \end{aligned}$$

## Idem with control variables

$$X = (1, x_1, \dots, x_K); Z = (1, x_1, \dots, x_{K-1}, z_1); \beta = (\beta_0, \beta_1, \dots, \beta_K)'$$

$$y = X\beta + u$$

$$(H) = (H_0) + (H_1) : E(Z'u) = 0$$

hence:

$$Z'y = Z'X\beta + Z'u$$

$$E(Z'y) = E(Z'X)\beta + E(Z'u) = E(Z'X)$$

$$\beta = E(Z'X)^{-1}E(Z'Y) \quad \text{if rank } E(Z'X) = K$$

Given a random sample the IV estimator is:

$$\widehat{\beta}_{IV} = \left( \frac{1}{N} \sum_i z_i' x_i \right)^{-1} \left( \frac{1}{N} \sum_i z_i' y_i \right) = (Z'X)^{-1} Z'y$$

# Outline

- 1 Introduction
- 2 The canonical example: returns to education
- 3 A formalization with 1 instrument
- 4 Identification with multiple instruments: the 2SLS

# Useful to have more than 1 instrument

- Helps to strengthen the 1st stage: better explain the endogenous RHS variable
- Gives the opportunity to test some of the identification assumptions

## More than 1 instrument for $x_K$

Let's assume we  $M$  instruments are available for  $x_K$ :

$$z_1, \dots, z_M / \text{Cov}(z_h, u) = 0 \quad \forall h \in 1, \dots, M$$

- $M$  different IV estimators
- more than that: any combination of  $z_h$  would satisfy the assumption

Which one should be chosen?

## The Two Stage Least Square (2SLS) estimator

The 2SLS estimator consists in choosing the one that is the most correlated with  $x_K$ .

$$x_K = \delta_0 + \delta_1 x_1 + \dots + \delta_{K-1} x_{K-1} + \theta_1 z_1 + \dots + \theta_M z_M + r_K$$

$$x_K^* = x_K - r_K$$

is generally called the part of  $x_K$  that is uncorrelated with  $u$

- $x_K$  endogenous because  $r_K$  correlated with  $u$ .

and  $\hat{x}_K$  can be used as an instrument.

$$\widehat{x}_K = \widehat{\delta}_0 + \widehat{\delta}_1 x_1 + \dots + \widehat{\theta}_M z_M$$

$$\widehat{X} = (1, x_1, \dots, x_{K-1}, \widehat{x}_K)$$

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

looks like the OLS estimator of  $y$  onto  $\widehat{X}$ .

# The procedure

- 1 Obtain fitted values  $\widehat{x}_K$  from regression

$$x_K \text{ on } 1, x_1, \dots, x_{K-1}, z_1, \dots, z_M$$

- 2 Run OLS regression

$$y \text{ on } 1, x_1, \dots, x_{K-1}, \widehat{x}_K$$

In practice: use a software package rather than carry out 2 steps procedure. Because:

- standard-error correction
- very often mistakes: forget to include  $x_1, \dots, x_{K-1}$  in 1st-stage

"ivregress 2sls ..." in Stata

# Identify returns to education based on who benefitted from the program

Issues:

- trends in education
- endogenous program placement

→ need to control for that and hence identify on intra-variation.

Assumption: trends would be the same in the 2 regions if there had not been the school building program.

→ additional controls.

# 2SLS estimates of returns to education in Indonesia

TABLE 7—EFFECT OF EDUCATION ON LABOR MARKET OUTCOMES: OLS AND 2SLS ESTIMATES

Method	Instrument	(1)	(2)	(3)	(4)
<i>Panel A: Sample of Wage Earners</i>					
<i>Panel A1: Dependent variable: log(hourly wage)</i>					
OLS		0.0776 (0.000620)	0.0777 (0.000621)	0.0767 (0.000646)	
2SLS	Year of birth dummies*program intensity in region of birth	0.0675 (0.0280) [0.96]	0.0809 (0.0272) [0.9]	0.106 (0.0222) [0.93]	0.0908 (0.0541) [0.9]
2SLS	(Aged 2–6 in 1974)*program intensity in region of birth	0.0752 (0.0338) (0.0338)	0.0862 (0.0336) (0.0336)	0.104 (0.0304) (0.0304)	
<i>Panel A2: Dependent variable: log(monthly earnings)</i>					
OLS		0.0698 (0.000601)	0.0698 (0.000602)	0.0689 (0.000628)	
2SLS	Year of birth dummies*program intensity in region of birth	0.0756 (0.0280) [0.73]	0.0925 (0.0278) [0.63]	0.0913 (0.0219) [0.58]	0.134 (0.0631) [0.7]
Control variables:					
Year of birth*enrollment rate in 1971		No	Yes	Yes	Yes
Year of birth*water and sanitation program		No	No	Yes	No
Propensity score, propensity score squared		No	No	No	Yes

*Notes:* Year of birth dummies, region of birth dummies, and the interactions between year of birth dummies and the number of children in the region of birth in 1971 are included in the regressions. Standard errors are in parentheses. *F*-statistics of the test of overidentification restrictions are in square brackets.

# Q&A

- is the OLS estimate significantly different from 0?
- what are the controls for?
- are the 2SLS estimates significantly different from 0?
- which estimate is the most precise? what is it due to?
- what is the direction of the bias?

## 2sls with more than 1 endogenous variable

- 1 instrument for 1 endogenous variable
- "order condition": as many instruments than endogenous variables (not counting exogenous variables that serve as instruments for themselves)

Example: Card (95): differential of returns to education for blacks:

$$\ln(\text{wage}) = \alpha_1 \text{educ} + \alpha_2 \text{black} \cdot \text{educ} + \alpha_3 \text{black} + \alpha_4 \text{exp} + \alpha_5 \text{exp}^2 + u$$

- Distance to college (*coll*) is an instrument for *educ*,
- *black* · *coll* can serve as an instrument for *black* · *educ*
- both variables are used to predict endogenous variables *educ* and *black* · *educ*