TSE February 1st, 2017

Applied Econometrics for Development: Instrumental Variables I

Ana GAZMURI

Paul SEABRIGHT

Motivation

- Consider a single equation linear model

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + u$$

- Key conditions for OLS estimation of $\beta's$:
  - $E(u) = 0$
  - $Cov(x_j, u) = 0, j = 1,2, \dots, k$
- What if $Cov(x_k, u) \neq 0$?
  - If we estimate this model by OLS, will we get a consistent estimate of β_k?
- Endogeneity usually arises for 3 reasons:
  - Omitted Variables
  - Measurement Error
  - Simultaneity

Omitted Variables

- Observed association between y and the x_k is likely to be misleading because it partially reflects omitted factors related to both variables

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \underbrace{q + w}_{u}$$

- If q is unobserved and correlated with at least one x, the estimate of $\beta's$ will be biased
- Can you think of examples?
  - Self-selection: if agents are choosing x_k, this decision might depend in unobservable factors

Omitted Variables

- Example: Wage Equation with Unobserved Ability

$$\log(wage) = \beta_0 + \beta_1 exper + \beta_2 exper^2 + \beta_3 educ + \gamma\ ability + u$$

$$E(u\ |exper, educ, ability) = 0$$

- Data on ability is typically unobserved
- The parameter on interest here is β_3
- If $ability$ and $educ$ are correlated, then β_3 is not identified
- If $abil = \delta_0 + \delta_1 educ + r$, with r uncorrelated with $exper$, then $\widehat{\beta_3} = \beta_3 + \gamma\delta_1$

Measurement Error

- We want to measure the effect of x_k^* but we observe only an imperfect measure x_k

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k^* + u$$

- x_k^* and x_k are uncorrelated with u

$$e_K = x_k - x_k^* \text{ and } E(e_k) = 0$$

- e_k is uncorrelated with x_j, $j = 1, \ldots, k-1$ (usual assumption)
- Two possible assumptions:
  - e_k is uncorrelated with the observed measure $\mathrm{Cov}(x_k, e_k) = 0$ and e_k correlated with unobserved variable x_k^*
  - e_k is uncorrelated with the unobserved variable $\mathrm{Cov}(x_k^*, e_k) = 0$ (Classical errors-in-variables assumption)

Measurement Error

- e_k is uncorrelated with the unobserved variable

$$Cov(x_k^*, e_k) = 0$$

$$x_k = x_k^* + e_k$$

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + (\underbrace{u + \beta_k e_k}_{v})$$

$$\begin{aligned} Cov(x_k, v) &= E(x_k v) \\ &= E((x_k^* e_k)(u + \beta_k e_k)) \\ &= \beta_k \sigma_{e_k}^2 \neq 0 \end{aligned}$$

- OLS regression will give inconsistent estimators of all $\beta_j' s$ when x_j is correlated with x_k

Measurement Error

- For variables correlated with x_k

$$\hat{\beta} = \beta \frac{\sigma_{x^*}^2}{\sigma_e^2 + \sigma_{x^*}^2}$$

- $\frac{\sigma_{x^*}^2}{\sigma_e^2+\sigma_{x^*}^2}$ is between 0 and 1
- This type of measurement error is called attenuation bias
- Measurement error shrink estimates towards zero
- What happens if the measurement error is in the dependent variable?

Simultaneity

- At least one explanatory variable is determined simultaneously along with the dependent variable
- Estimation of supply and demand
- Examples:
    - Murder rate and size of police force: size of police force is partially determined by the murder rate

Remarks Regarding Endogeneity

- The distinctions among the 3 forms of endogeneity are not always sharp
- One model can have more than one source of endogeneity
- Example:
  - Effect of alcohol consumption on worker productivity (measure by wages)
  - We would worry that:
    - Alcohol usage is correlated with unobserved factors that also affect wage (family background)
    - Alcohol demand generally depends on income
    - Alcohol usage may be imprecisely measured

Instrumental Variables

- Instrumental variables provide a general solution to the problem of an endogenous explanatory variable
- We need an observable variable Z, not in the model, that explains variation in the endogenous X
- This instrument Z cannot determine Y in any way except through its effect on X

Intuition

- Back to the linear model

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + u$$

- Think of x_k as having ‘good’ and ‘bad’ variation
  - Good variation is not correlated with u
  - Bad variation is correlated with u
- A good IV is a variable that explains variation in x_k but doesn’t explain y
  - i.e. It only explains the ‘good’ variation in x_k
- We can use the IV to extract the ‘good’ variation and replace x_k with only that component

Required Assumptions

- An IV must satisfy two conditions:
  - Relevance
  - Exclusion
- Which is harder to satisfy? Can we test them?
- Let's start with the simplest case: one problematic regressor and one instrument

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + u$$

- We have an instrument z for the problematic regressor x_k

Relevance Condition

- In the following model:

$$x_k = \alpha_0 + \alpha_1 x_1 + \cdots + \alpha_{k-1} x_{k-1} + \gamma z + v$$

- z satisfies the relevance condition if $\gamma \neq 0$
- Easy to test, just run the regression of x_k on all the other $x's$ and the instrument z
- This is the first stage of the IV estimation
- Important: you need to include all the other regressors in the equation
  - i.e. z is relevant to explaining the problematic regressor after partialling out the effect of all the other regressors in the original model

Exclusion Condition

- In the original model:

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + u$$

- z satisfies the exclusion restriction if $Cov(z, u) = 0$
- z has no explanatory power with respect to y, only through its effect on x_k
- This condition cannot be tested because u is unobservable
- You must find a convincing economic argument to why the exclusion restriction is not violated

Example

- Suppose you want to estimate job training effect on worker's productivity

  x_k: job training hours per worker

  y: measure of average worker productivity

  There exists a government program randomly assigning grants for job training to firms

- Natural possible instruments:
  - A binary variable indicating whether a firm received a job training grant
  - The actual amount of the grant per worker, if the amounts varies by firm

Implementation

- You have a good IV, now what?
- Two steps:
  - First stage: regress x_k on other $x's$ and z
  - Second stage: take predicted $\hat{x}_k$ from the first stage and use it in the original model instead of x_k
  - This is why we call IV estimations two stage least squares (2SLS)

Implementation – First Stage of 2SLS

- Estimate the following

$$x_k = \alpha_0 + \alpha_1 x_1 + \cdots + \alpha_{k-1} x_{k-1} + \gamma z + v$$

- Calculate predicted values $\hat{x}_k$

$$\hat{x}_k = \hat{\alpha}_0 + \hat{\alpha}_1 x_1 + \cdots + \hat{\alpha}_{k-1} x_{k-1} + \hat{y} z$$

- Always report your first stage results and R^2
  - It's a direct test of relevance condition
  - It helps determining whether there might be a weak IV problem

Implementation – Second Stage of 2SLS

- Use predicted values to estimate:

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k \hat{x}_k + u$$

Predicted values replace
problematic regressor

- 2SLS estimation yields consistent estimates of all $\beta's$ when relevance and exclusion conditions are satisfied
- If you do the estimation step by step <u>standard errors</u> from the second stage <u>will be wrong</u> (Use a software package to do 2SLS, don't do it on your own)
  - The second stage uses estimated values that have their own estimation error. This error needs to be taken into account when calculating standard errors.
  - Careful with models with quadratic terms (do not use $\hat{x}_k$ and $\hat{x}_k^2$)

Implementation – Second Stage of 2SLS

- What would you do if you have quadratic terms for the problematic regressor?
- Why can't we use just the other $x's$ in the first stage? Why do we need z?

Consistent, but Biased

- IV is a consistent, but biased estimator
  - For any finite number of observations N, the IV estimates are biased towards the OLS estimate
  - As N approaches infinity, the IV estimates converge to the true coefficients
- This feature of IV leads to what is called the weak instrument problem

Weak Instruments Problem

- A weak instrument is an IV that doesn't explain very much of the variation in the problematic regressor
  - Small sample bias of estimator is greater when instrument is weak
  - Hahn and Hausman (2005) show that finite sample bias is $\approx \frac{j\rho(1-r^2)}{Nr^2}$

j =number of IV's

ρ =correlation between x_k and u

$r^2 = R^2$ from first-stage regression

$N =$ sample size

Low explanatory power can result in large bias even if N is large

More instruments may help increase r^2 but if they are weak they can increase bias

Weak Instruments Problem

- Detecting weak instruments
  - Large standard errors in IV estimates (you'll get large SE when covariance between instrument and problematic regressor is low)
  - Low F statistic from first stage
    - The higher the F statistic for excluded instruments the better
    - From Stock, Wright, and Yogo (2002), above 10 likely OK

Multiple IVs and Overidentification Tests

- What if we have more than one problematic regressor?
  - IVs can still solve this
  - You need at least one IV for each endogenous regressor
  - Then estimate 2SLS in similar way
- We need at least one exogenous variable that does not appear in the structural equation as an instrument for each endogenous variable
- What if we have more instruments than needed?
  - H endogenous variables
  - M>H instruments
  - The model is overidentified (M-H overidentifying restrictions)

Multiple IVs and Overidentification Tests

- Relevance condition
  - Each first-stage must have at least one IV with non-zero coefficient
  - Of the M instruments, at least H of them must be partially correlated with problematic regressors
  - You can't just have one IV correlated with all the problematic regressors and the other IV's not
- Not obvious that you want more instruments
  - If you have a very good instrument, not clear you want to add some extra less-good IVs (it will increase small sample bias)
  - If your IVs satisfy the relevance conditions, you'll get more efficiency with more IVs
- When model is overidentified you could 'test' the quality of the IVs

Testing Overidentifying Restrictions

- If we have more instruments than we needed to identify the model
  - We can test whether the additional instruments are valid in the sense that they are uncorrelated with u
- Hausman (1978) suggested comparing the 2SLS estimator using all instruments to 2SLS using a subset that just identifies the equation
  - If all the IVs are valid, then you can get consistent estimates using any subset of the IVs
  - So, compare IV estimates from different subsets. Estimates should only differ as a result of sampling error
  - The test implicitly assumes that some subset of instruments is valid (which may not be the case)
  - You need to use economic arguments to motivate that the IV satisfies the exclusion restriction

IV with interactions

- Suppose you want to estimate:

$$y = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \alpha_3 x_1 x_2 + u$$

where $Cov(x_1, u) = 0$ and $Cov(x_2, u) \neq 0$

- Now both x_2 and $x_1 x_2$ are problematic
- Suppose you can only find one IV z , is there a way to get consistent estimates?
- You can construct other instruments from the one IV
  - Use z as IV for x_2
  - Use $x_1 z$ as IV for $x_1 x_2$

Common Sources of Instruments

- Sometimes, convincing instruments arise in the context of program evaluation
  - Individuals randomly selected for a job training program
  - Students randomly assigned a school voucher
  - Actual participation is almost always voluntary and it can be endogenous
  - However, eligibility is exogenous
  - Eligibility can be used as an IV for job training
- Natural experiments are another source of instruments
  - Some feature of the context we are studying, produces exogenous variation in an otherwise endogenous variable
- Regional variation in prices or taxes
  - Local price of alcohol may induce some exogenous variation in alcohol consumption

Examples

- Angrist and Krueger (2001)
  - Earliest applications of IVs involved estimation of elasticities of demand and supply
  - Time series data on prices and quantities
  - OLS regression of quantities on prices fails to trace out either the supply or demand relationship
- P.G. Wright(1928) suggests using ‘curve shifters’ to address the problem
  - Demand shifter: price of substitutes
  - Supply shifter: yields per acre, weather
  - He uses 6 different instruments and then averages the 6 estimates
  - 2SLS is a more efficient way to combine multiple instruments

Examples

- IV estimation for education in a wage equation

$$\log(wage) = \beta_0 + \beta_1 exper + \beta_2 exper^2 + \beta_3 educ + \gamma\ ability + u$$

- Going back to the initial model, education is correlated with the error because of omitted ability
- Candidates for IV
  - Mother's education?
- Challenge to come up with convincing instruments
  - Angrist and Krueger (1991) propose using quarter of birth
  - Compulsory school attendance laws induce a relationship between education and quarter of birth. Some people are forced to attend school for longer than they would otherwise do

Examples

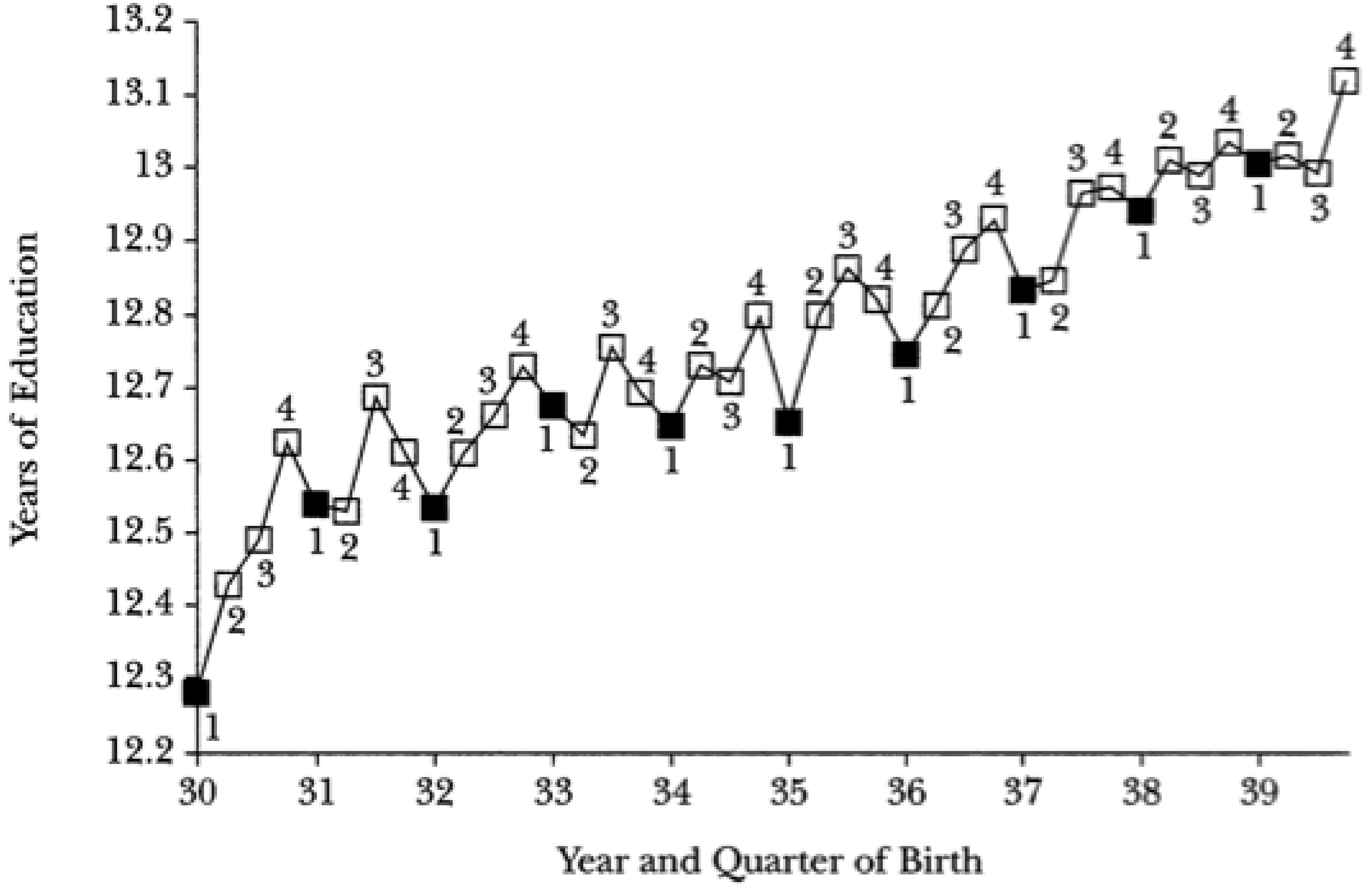
- Do they satisfy the 2 criteria?
  - For mother's education it's hard to argue that $Cov(mothereduc, u) = 0$
  - For quarter of birth, the concern is with relevance, but this can be tested
- Another issue: for who are we estimating the return to education?
  - Even if quarter of birth is a relevant IV
  - If returns to education are not constant across people
  - IV estimates are giving the return to education only for people induced to obtain more schooling because they were born in the first quarter of the year

Examples

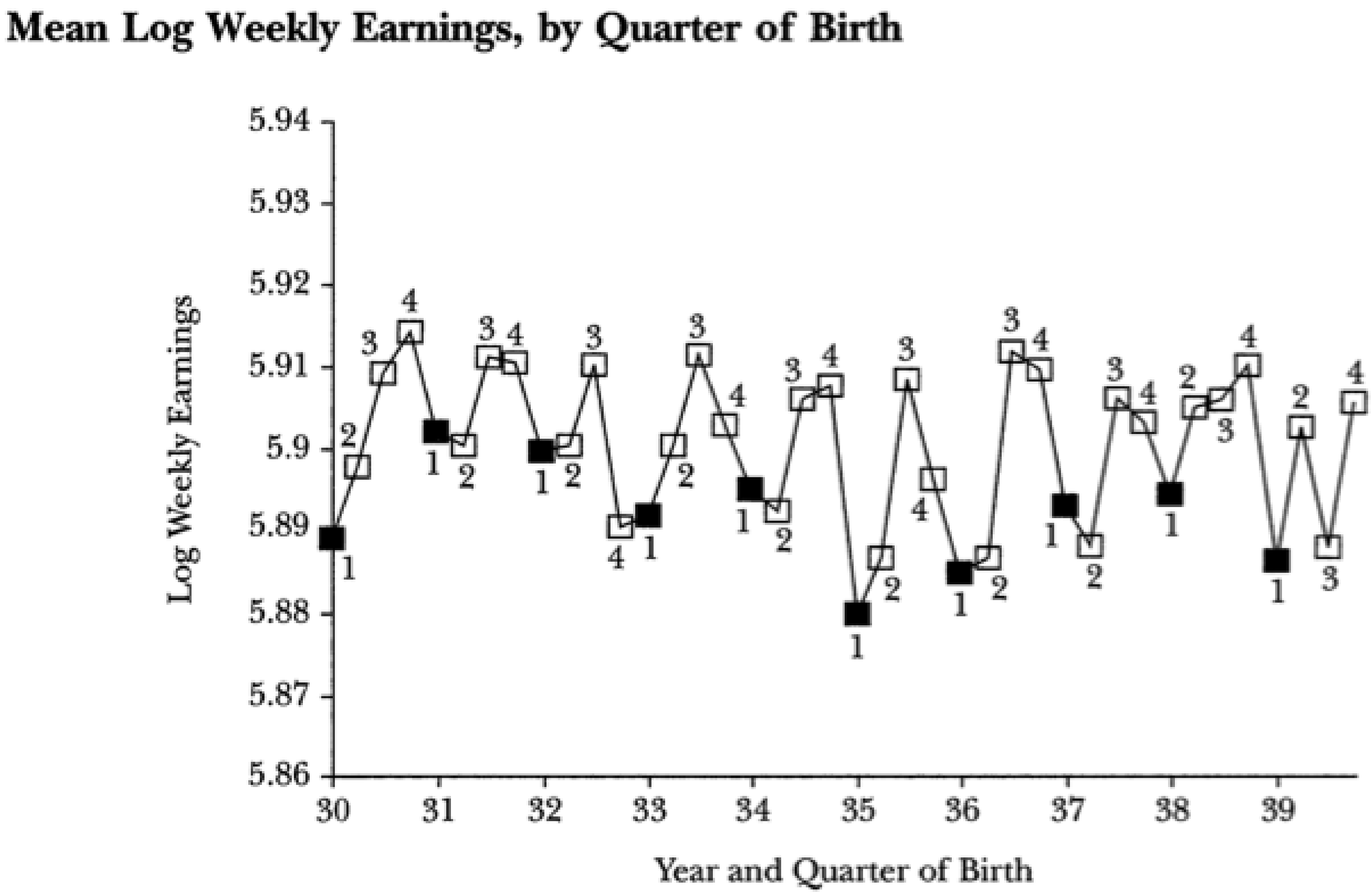
- Angrist and Krueger (1991)
  - Most states require students to enter school in the calendar year in wich they turn six (school start age is a function of date of birth)
  - A kid born in the fourth quarter enters school at 5 3/4 , while those born in the fourth quarter enter school at age 6 ¾
  - Typically, compulsory schooling laws require students to remain in school until their 16th birthday
  - Therefore, students will be in different grades when they reach legal drop out rate
  - This creates a natural experiment in which children are compelled to attend school for different lengths of time

Examples

Mean Years of Completed Education, by Quarter of Birth



Examples


Mean Log Weekly Earnings, by Quarter of Birth
Log Weekly Earnings
5.94
5.93
5.92
5.91
5.9
5.89
5.88
5.87
5.86
30
31
32
33
34
35
36
37
38
39
Year and Quarter of Birth

Examples

- College proximity as an IV for education
  - Card(1995): use a dummy variable for whether a man grew up in the vicinity of a four-year college as an instrument for schooling
  - Also includes controls for experience, race, indicators for south, region, and urban
  - IV estimate of return to schooling: 13.2% (vs 7.5% with OLS)
  - Counterintuitive result, we would expect upward bias
  - Some explanations
    - Measurement error gives attenuation bias
    - Instrument is not exogenous in the wage equation

Examples

- Angrist and Lavy (1999) estimate the effects of class size on student achievement
  - They use a bureaucratic ceiling law on class size that induces sharp differences in average class size in Israel
  - OLS estimates show either no effect or positive effect of larger classes
  - IV estimates reveal a statistically significant benefit of smaller classes
- Angrist (1990) estimate the effect of military service on earnings later in life
  - They use Vietnam-era draft lottery numbers as an IV
  - The lottery numbers were randomly assigned to young men in the early 1970s were highly correlated with the probability of being drafted into the military

Interpreting Estimates with Heterogeneous Responses

- Not every observation is affected by the instrument
- The instrument operates by using only part of the variation in the explanatory variable
  - Angrist and Krueger (1991) case, the IV is most relevant for people with a high probability of leaving school as soon as possible with no effect on people going to college
  - Angrist(1990) Vietnam estimates are based only on the experience of those serving in the military because of the draft (not of volunteers)
- Instrumental variables provide an estimate for a specific group, the people manipulated by the instrument
- Extrapolation to other populations is speculative and relies on theory and common sense

TSE February 1st, 2017

Applied Econometrics for Development: Instrumental Variables I

Ana GAZMURI

Paul SEABRIGHT