Development Economics

Slides 9

Debraj Ray, NYU

Testing for the Long Shadow of Institutions

- **Institutions are endogenous to development.**

- So how to establish causality?

Testing for the Long Shadow of Institutions

- **Institutions are endogenous to development.**
- So how to establish causality?
- **Problems of endogeneity:**
- Reverse causality: richer countries can afford better institutions
- Omitted variables: Some other factor moves both outcomes
- Measurement error: Independent variable measured with noise
- Perceptions: perceiving better institutions in richer countries

A Detour: Instrumental Variables

- **Example 1: Schooling and Earnings** (Angrist and Kreuger 1991)

- Imagine we want the effect of schooling on wages, and regress

$$y_i = C + b_1 s_i + \epsilon_i$$

y_i is earnings (say log wages) and s_i is years of schooling.

A Detour: Instrumental Variables

- **Example 1: Schooling and Earnings** (Angrist and Kreuger 1991)

- Imagine we want the effect of schooling on wages, and regress

$$y_i = C + b_1 s_i + \epsilon_i$$

y_i is earnings (say log wages) and s_i is years of schooling.

- **Problem:** Omitted variable we do not measure: "ability."

- The "true" regression is

$$y_i = C + b_1 s_i + b_2 a_i + \epsilon_i$$

but we can't run this regression because we don't see a_i!

- **Example 2: Poverty and Conflict** (Miguel, Satyanath and Sergenti 2004)

- We want to know if low incomes cause conflict, and regress

$$c_i = A + by_i + \epsilon_i$$

  where c_i is conflict incidence and y_i is per-capita income.

- **Problem:** Reverse causality. Conflict can affect income.

- **Example 2: Poverty and Conflict** (Miguel, Satyanath and Sergenti 2004)

- We want to know if low incomes cause conflict, and regress

$$c_i = A + by_i + \epsilon_i$$

  where c_i is conflict incidence and y_i is per-capita income.

- **Problem:** Reverse causality. Conflict can affect income.

- **Example 3: Noisy data from the past**

- We think we're running $y_i = A + bx_i + \epsilon_i$, but we're *really* running is

$$y_i = A + b[x_i + m_i] + \epsilon_i', \text{ where } \epsilon_i' = \epsilon_i - bm_i.$$

- **Problem:** Measurement error m_i when measuring x_i.

- **Example 2: Poverty and Conflict** (Miguel, Satyanath and Sergenti 2004)

- We want to know if low incomes cause conflict, and regress

$$c_i = A + by_i + \epsilon_i$$

  where c_i is conflict incidence and y_i is per-capita income.

- **Problem:** Reverse causality. Conflict can affect income.

- **Example 3: Noisy data from the past**

- We think we're running $y_i = A + bx_i + \epsilon_i$, but we're *really* running is

$$y_i = A + b[x_i + m_i] + \epsilon_i', \text{ where } \epsilon_i' = \epsilon_i - bm_i.$$

- **Problem:** Measurement error m_i when measuring x_i.

- **Common Theme:** x_i correlated with error term ϵ_i ; biases b.

Summary of the Problem

- **Regression coefficient given by**

$$\hat{b} = \frac{\text{SampleCov}(x, y)}{\text{SampleVar}(x)} = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2}$$

Summary of the Problem

- **Regression coefficient given by**

$$\hat{b} = \frac{\text{SampleCov}(x, y)}{\text{SampleVar}(x)} = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2}$$

- At the same time,

$$y_i - \bar{y} = [A + bx_i + \epsilon_i] - [A + b\bar{x} + \bar{\epsilon}] = b[x_i - \bar{x}] + [\epsilon_i - \bar{\epsilon}].$$

Summary of the Problem

- **Regression coefficient given by**

$$\hat{b} = \frac{\text{SampleCov}(x, y)}{\text{SampleVar}(x)} = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2}$$

- At the same time,

$$y_i - \bar{y} = [A + bx_i + \epsilon_i] - [A + b\bar{x} + \bar{\epsilon}] = b[x_i - \bar{x}] + [\epsilon_i - \bar{\epsilon}].$$

- Combining,

$$\hat{b} = b\frac{\sum_i (x_i - \bar{x})^2}{\sum_i (x_i - \bar{x})^2} + \frac{\sum_i (x_i - \bar{x})(\epsilon_i - \bar{\epsilon})}{\sum_i (x_i - \bar{x})^2}$$

Summary of the Problem

- **Regression coefficient given by**

$$\hat{b} = \frac{\text{SampleCov}(x, y)}{\text{SampleVar}(x)} = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2}$$

- At the same time,

$$y_i - \bar{y} = [A + bx_i + \epsilon_i] - [A + b\bar{x} + \bar{\epsilon}] = b[x_i - \bar{x}] + [\epsilon_i - \bar{\epsilon}].$$

- Combining,

$$\hat{b} = b\frac{\sum_i (x_i - \bar{x})^2}{\sum_i (x_i - \bar{x})^2} + \frac{\sum_i (x_i - \bar{x})(\epsilon_i - \bar{\epsilon})}{\sum_i (x_i - \bar{x})^2}$$

- If ϵ is uncorrelated with x, then $\mathbb{E}(\hat{b}|x_1, \ldots, x_n) = b$.
- But if ϵ is correlated with x, then $\mathbb{E}(\hat{b}|x_1, \ldots, x_n) \neq b$.

More importantly, the problem persists in large samples.

Instruments

- **A magic variable z that satisfies two conditions:**

Instruments

- **A magic variable z that satisfies two conditions:**

(i) z is correlated with the endogenous variable x.

- You can run a separate regression (the "**first stage**") to show this is statistically true.

Instruments

- **A magic variable z that satisfies two conditions:**

(i) z is correlated with the endogenous variable x.

- You can run a separate regression (the "**first stage**") to show this is statistically true.

(ii) z is uncorrelated (except via x) with the dependent variable y.

- That is, z is uncorrelated with the error term ϵ: **exclusion restriction**.
- You are not allowed to show this statistically. You have make the argument by appealing to "theory."

Application: Birthdays and Schooling

Angrist and Krueger 1991, 2001

- In the US, you start school in the year you turn 6.
- Someone born in December 2000 can go to school a year earlier than someone born in January 2001.
- So those born in a later calendar quarter start school early.

Application: Birthdays and Schooling

Angrist and Krueger 1991, 2001

- In the US, you start school in the year you turn 6.
- Someone born in December 2000 can go to school a year earlier than someone born in January 2001.
- So those born in a later calendar quarter start school early.
- On average, this gives late-quarter individuals more years of schooling.
- Angrist-Krueger study men born from 1930 to 1959 (1980 US census).

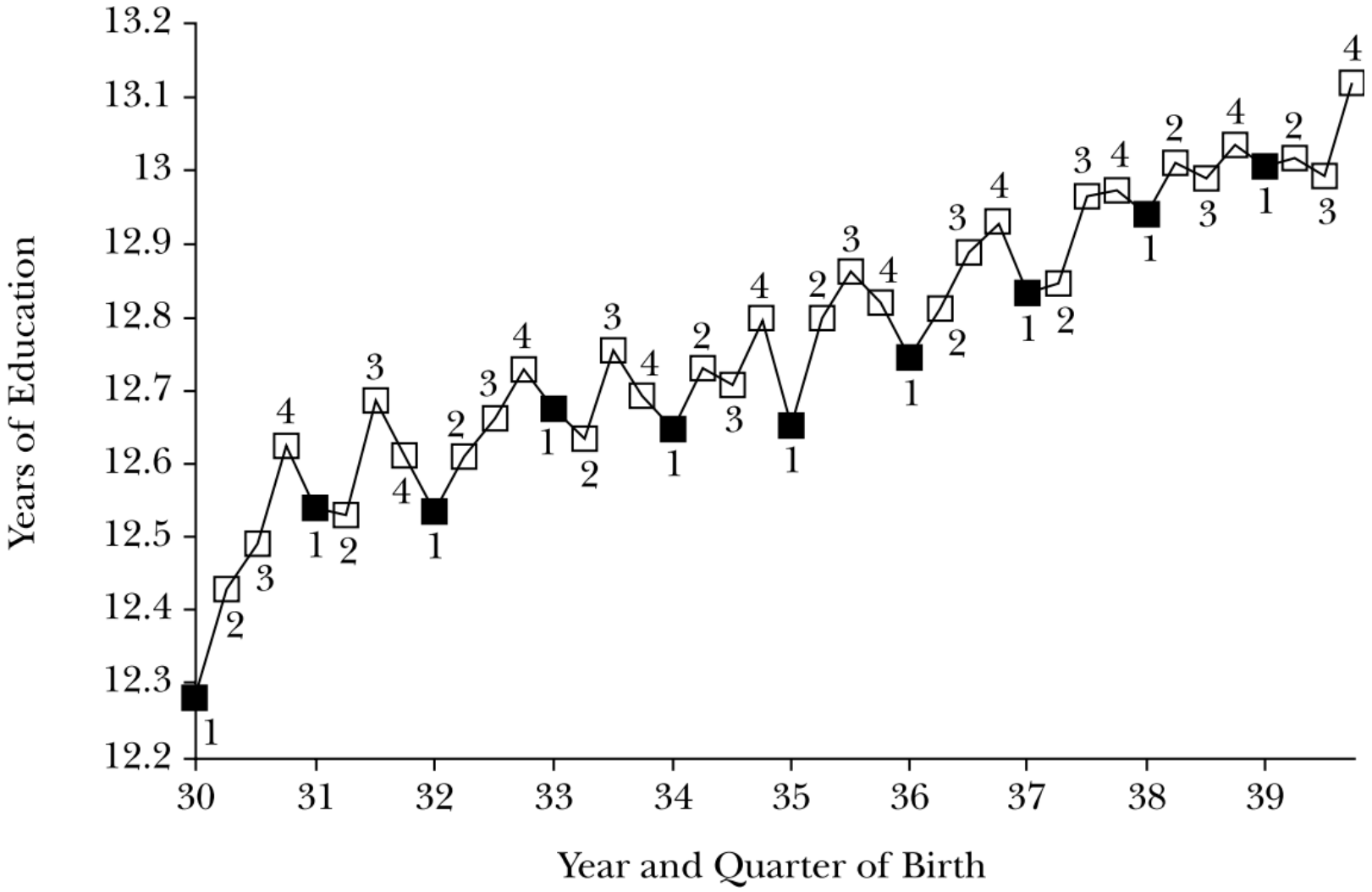
Application: Birthdays and Schooling

Angrist and Krueger 1991, 2001

- In the US, you start school in the year you turn 6.
- Someone born in December 2000 can go to school a year earlier than someone born in January 2001.
- So those born in a later calendar quarter start school early.
- On average, this gives late-quarter individuals more years of schooling.
- Angrist-Krueger study men born from 1930 to 1959 (1980 US census).
- **Is quarter of birth usable as an instrument for education?**

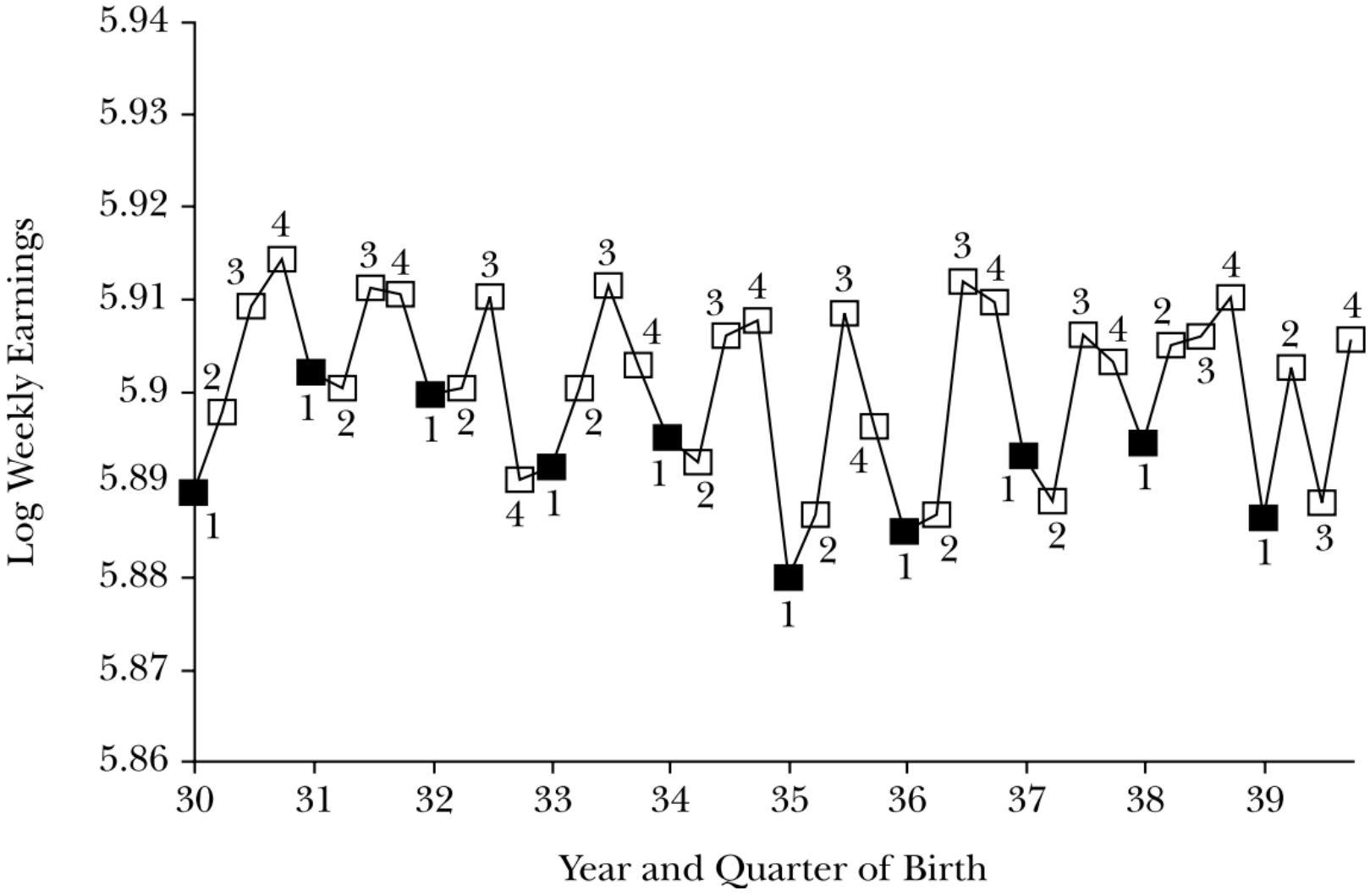
Birthdays and Schooling

Mean years of completed education, by quarter of birth:



Birthdays and Schooling

Mean log weekly earnings, by quarter of birth:



From Angrist and Krueger 2001:

"Differences in earnings by quarter of birth are assumed to be accounted for solely by differences in schooling by quarter of birth, so that the estimated return to schooling is simply the appropriately rescaled difference in average earnings by quarter of birth. Only a small part of the variability in schooling — the part associated with quarter of birth — is used to identify the return to education."

From Angrist and Krueger 2001:

"Differences in earnings by quarter of birth are assumed to be accounted for solely by differences in schooling by quarter of birth, so that the estimated return to schooling is simply the appropriately rescaled difference in average earnings by quarter of birth. Only a small part of the variability in schooling — the part associated with quarter of birth — is used to identify the return to education."

- **Other possible instruments in this context:**
  - School availability (Duflo 1998 Indonesia, Bedi-Gaston 1999 Honduras)
  - Distance to the nearest high school (Maluccio 1997 Philippines)
  - Change in compulsory schooling age (Harmon and Walker 1995 UK)
- Do you think these are good instruments? Why or why not?

IV Procedure: Two-Stage Least Squares (2SLS)

- **First Stage**: regress x on the instrument(s) z.
- Make sure to include any controls to be used in predicting y.
- **Second Stage**: run the regression you originally wanted, except ...
- ...use the *predicted* or fitted values $\hat{x}$ from the First Stage.

IV Procedure: Two-Stage Least Squares (2SLS)

- **First Stage**: regress x on the instrument(s) z.
- Make sure to include any controls to be used in predicting y.
- **Second Stage**: run the regression you originally wanted, except ...
- ...use the *predicted* or fitted values $\hat{x}$ from the First Stage.
- **In effect**: make use of the variation of x that comes from the instruments, using all available right-hand-side variables. That collapses everything neatly into a single object $\hat{x}$, which is used in the second stage.

Single Instrument for Single Variable, No Controls

- IV estimator given by

$$\hat{b}_{\text{IV}} = \frac{\text{SampleCov}(z, y)}{\text{SampleCov}(z, x)} = \frac{\sum_i (z_i - \bar{z})(y_i - \bar{y})}{\sum_i (z_i - \bar{z})(x_i - \bar{x})},$$

  where z is the instrument for x.

Single Instrument for Single Variable, No Controls

- IV estimator given by

$$\hat{b}_{\text{IV}} = \frac{\text{SampleCov}(z, y)}{\text{SampleCov}(z, x)} = \frac{\sum_i (z_i - \bar{z})(y_i - \bar{y})}{\sum_i (z_i - \bar{z})(x_i - \bar{x})},$$

where z is the instrument for x. As before,

$$y_i - \bar{y} = [A + bx_i + \epsilon_i] - [A + b\bar{x} + \bar{\epsilon}] = b[x_i - \bar{x}] + [\epsilon_i - \bar{\epsilon}].$$

Single Instrument for Single Variable, No Controls

- IV estimator given by

$$\hat{b}_{\text{IV}} = \frac{\text{SampleCov}(z, y)}{\text{SampleCov}(z, x)} = \frac{\sum_i (z_i - \bar{z})(y_i - \bar{y})}{\sum_i (z_i - \bar{z})(x_i - \bar{x})},$$

  where z is the instrument for x. As before,

$$y_i - \bar{y} = [A + bx_i + \epsilon_i] - [A + b\bar{x} + \bar{\epsilon}] = b[x_i - \bar{x}] + [\epsilon_i - \bar{\epsilon}].$$

- Combining,

$$\hat{b}_{\text{IV}} = b\frac{\sum_i (z_i - \bar{z})(x_i - \bar{x})}{\sum_i (z_i - \bar{z})(x_i - \bar{x})} + \frac{\sum_i (z_i - \bar{z})(\epsilon_i - \bar{\epsilon})}{\sum_i (z_i - \bar{z})(x_i - \bar{x})} = b + \frac{\sum_i (z_i - \bar{z})(\epsilon_i - \bar{\epsilon})}{\sum_i (z_i - \bar{z})(x_i - \bar{x})}$$

Single Instrument for Single Variable, No Controls

- IV estimator given by

$$\hat{b}_{\text{IV}} = \frac{\text{SampleCov}(z, y)}{\text{SampleCov}(z, x)} = \frac{\sum_i (z_i - \bar{z})(y_i - \bar{y})}{\sum_i (z_i - \bar{z})(x_i - \bar{x})},$$

  where z is the instrument for x. As before,

$$y_i - \bar{y} = [A + bx_i + \epsilon_i] - [A + b\bar{x} + \bar{\epsilon}] = b[x_i - \bar{x}] + [\epsilon_i - \bar{\epsilon}].$$

- Combining,

$$\hat{b}_{\text{IV}} = b\frac{\sum_i (z_i - \bar{z})(x_i - \bar{x})}{\sum_i (z_i - \bar{z})(x_i - \bar{x})} + \frac{\sum_i (z_i - \bar{z})(\epsilon_i - \bar{\epsilon})}{\sum_i (z_i - \bar{z})(x_i - \bar{x})} = b + \frac{\sum_i (z_i - \bar{z})(\epsilon_i - \bar{\epsilon})}{\sum_i (z_i - \bar{z})(x_i - \bar{x})}$$

- $\Rightarrow \mathbb{E}(\hat{b}_{\text{IV}} | \mathbf{x}, \mathbf{z}) = b + \mathbb{E}\left(\frac{\sum_i (z_i - \bar{z})(\epsilon_i - \bar{\epsilon})}{\sum_i (z_i - \bar{z})(x_i - \bar{x})} | \mathbf{x}, \mathbf{z} \right)$.
- Could still be biased for small samples because x and ϵ are correlated, but ...
- **For large samples,** $\hat{b}_{\text{IV}} \simeq b$, because sample covariances are consistent.