# Econometrics

## Practical Session 21

### Instrumental Variables: The Single-Instrument Case

Ricardo Gouveia-Mendes
rgouveiamendes@ucp.pt

Spring 2025-26

Católica-Lisbon School of Business and Economics

# Theoretical Wrap-up

# Why IV Regression?

Three sources of $E[u_i \mid x_i] \neq 0$ **(endogeneity)** → OLS biased and inconsistent:

**1. OMITTED VARIABLE BIAS**

An omitted factor affects *y* and correlates with *x*

**2. SIMULTANEOUS CAUSALITY**

*x* causes *y*,
but *y* also causes *x*

**3. ERRORS-IN-VARIABLES**

*x* is measured with error → $\tilde{x} = x + v$ → attenuation bias

- **The solution:** remove from *x* the variability related to *u*
- **How?** Through an **instrumental variable** *z* correlated with *x* but unrelated to *y*
- **IV regression** can **restore OLS** unbiasedness and consistency

# Two Conditions for a Valid Instrument

CATÓLICA LISBON
BUSINESS & ECONOMICS

### CONDITION 1: RELEVANCE

$$\text{corr}(z_i, x_i) \neq 0$$

- $z$ must explain variation in $x$
- **Testable:** run a first-stage regression $x_i = \pi_0 + \pi_1 z_i + v_i$ and check the $F$-statistic

### CONDITION 2: EXOGENEITY

$$\text{corr}(z_i, u_i) = 0$$

- $z$ must affect $y$ **only through** $x$ → not through any other channel
- **Not directly testable:** requires judgment

### THE EXCLUSION RESTRICTION

$z$ is excluded from the structural equation for $y$ → it affects $y$ only via $x$

# Two-Stage Least Squares (2SLS/TSLS)

CATÓLICA LISBON
BUSINESS & ECONOMICS

### STAGE 1

Regress $x_i$ on $z_i$ and save fitted values:

$$x_i = \pi_0 + \pi_1 z_i + v_i \quad \rightarrow \quad \hat{x}_i = \hat{\pi}_0 + \hat{\pi}_1 z_i$$

### STAGE 2

Regress $y_i$ on $\hat{x}_i$ using OLS:

$$y_i = \beta_0 + \beta_1 \hat{x}_i + u_i$$

Because $\hat{x}_i$ is uncorrelated with $u_i$, $\hat{\beta}_j$ are consistent

$$\hat{\beta}_1^{TSLS} = \frac{s_{YZ}}{s_{XZ}} = \frac{\text{cov}(y_i, z_i)}{\text{cov}(x_i, z_i)}$$

### ⚠ STANDARD ERRORS

- The 2nd-stage ignores estimation errors on $\hat{x}_i$
  → SE **too small**
  → **invalid inference**
- Statistical **softwares handle this**

# Paper Analysis: Acemoglu, Johnson & Robinson (2001)

# The AJR (2001) Study

- **Question:** What is the causal effect of economic institutions on long-run income?
- Acemoglu, D., Johnson, S. & Robinson, J.A. (AER, 2001). "The Colonial Origins of Comparative Development: An Empirical Investigation."
  - Sample: **64 former European colonies**, cross-section
  - Long-standing puzzle: **why are some former colonies poor** and others rich?
  - **Key insight:** colonial history shaped institutions, which shaped development

### The colonial channel

- High settler mortality
  → extractive institutions
  → weak property rights
  → lower long-run income
- Low settler mortality
  → "Neo-European" institutions created
  → secure property rights
  → higher long-run income

# Data and Variables

| Variable | Role | Description |
|---|---|---|
| `loggdp` | Outcome $y$ | Log of real GDP per capita (PPP), 1995 |
| `risk` | Endogenous $x$ | Average protection against expropriation risk, 1985–95 (0–10 scale; higher = better institutions) |

**ENDOGENEITY CONCERNS**

- Richer countries can afford better institutions → **simultaneous causality**
- Measurement error in the regressor → **attenuation bias**

# Data and Variables

| Variable | Role | Description |
|---|---|---|
| `logmort0` | Instrument $z$ | Log of 19th-century European settler mortality rate (deaths per 1 000 soldiers per year) |

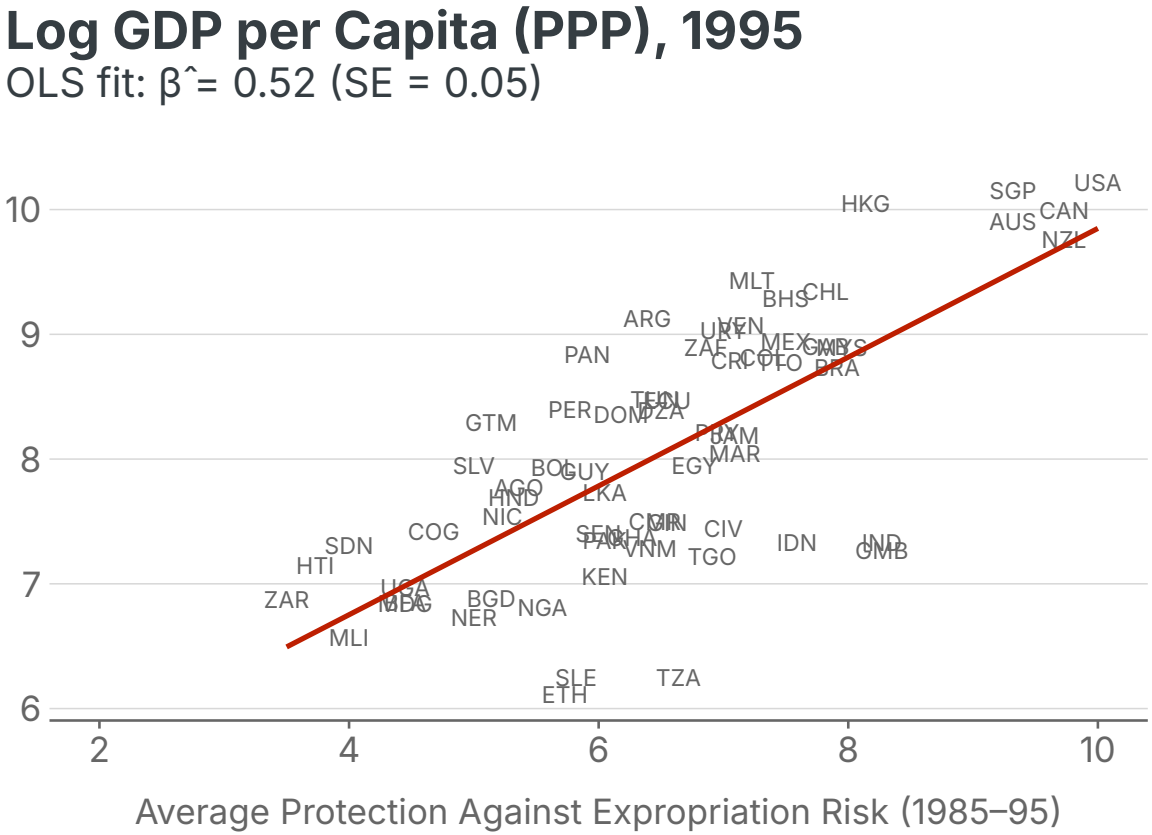
**VALIDITY CONDITIONS TO CHECK**

- Is the private property safety index ($x$) related to settler mortality ($z$)?
  → **relevance**
- Is the log-GDP per capita ($y$) related to settler mortality ($z$)?
  → **exogeneity**

# Step 1: OLS Regression

## Does expropriation risk explain log GDP per capita?

- Model:
  $\text{loggdp}_i = \beta_0 + \beta_1 \text{risk}_i + u_i$
- $\hat{\beta}_1^{OLS} = 0.516$ → large and significant, **but biased**
- The **net direction** of bias is uncertain: both forces operate simultaneously



# The Instrument: 19th-Century Settler Mortality

CATÓLICA LISBON
BUSINESS & ECONOMICS

### RELEVANCE ARGUMENT

- High disease burden (malaria, yellow fever)
  → Europeans avoided settling
  → set up **extractive** institutions
  → low protection against expropriation today
- corr(logmort0, risk) < 0 → **testable**

### EXOGENEITY ARGUMENT

19th-century disease environment and mortality affects 1995 GDP **only through** its effect on institutions

# Step 2: First Stage

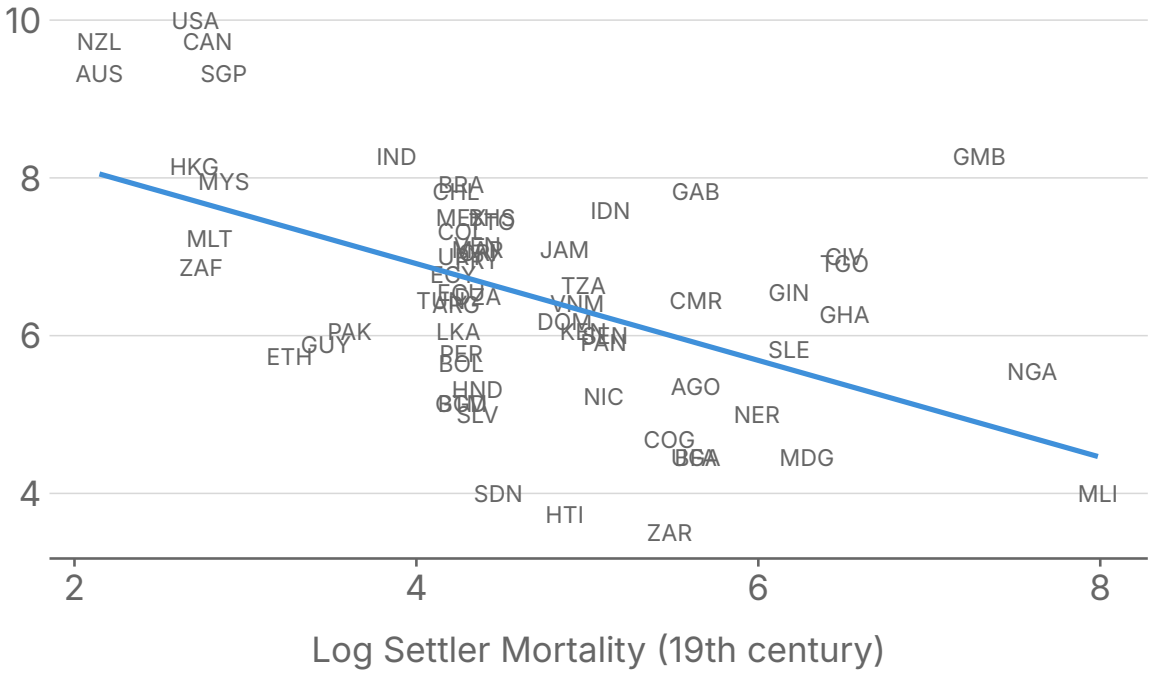
## Does settler mortality explain current institutional quality?

$$\text{risk}_i = \pi_0 + \pi_1 \text{logmort0}_i + v_i$$

| Coefficient | Estimate |
|---|---|
| Intercept | 9.37*** (0.71) |
| `logmort0` | −0.613*** (0.152) |

**First-stage *F*-stat: 23.3** > 10

**→ strong instrument ✓**



# Step 3: Two-Stage Least Squares

| | OLS | TSLS |
|---|---|---|
| Intercept ($\hat{\beta}_0$) | 4.69<br>(0.32) | 1.99<br>(1.15) |
| Risk ($\hat{\beta}_1$) | **0.516**<br>(0.051) | **0.929**<br>(0.173) |

- A 1-unit increase in the expropriation risk index is associated with a 0.929 log-point increase in GDP per capita (TSLS) → roughly **double the OLS estimate**

# Step 4: Hausman Endogeneity Test

- **Method:** add the first-stage residuals $\hat{v}_i = \text{risk}_i - \widehat{\text{risk}}_i$ to the OLS regression:

$$\text{loggdp}_i = \beta_0 + \beta_1 \text{risk}_i + \delta \hat{v}_i + \varepsilon_i$$

- Individual significance test for $\delta$
  - $H_0$: $\delta = 0$ (risk is exogenous, OLS is consistent)
  - $H_1$: $\delta \neq 0$ (risk is endogenous, prefer TSLS)
- **Result**: $\hat{\delta} = -0.569$ ($p < 0.001$) is **highly significant** → reject $H_0$
- `risk` is endogenous → **prefer TSLS over OLS**

# Discussion

**1.** The TSLS estimate for $\beta_{\text{risk}}$ (0.929) is roughly double the OLS estimate (0.516). Which source of bias → simultaneous causality or measurement error → appears to dominate? Why?

**KEY TAKEAWAYS**

- TSLS > OLS → the net OLS bias is **downward → measurement error dominates**
- Attenuation bias from using an imperfect risk index compresses the OLS coefficient more than the upward push from reverse causality expands it
- AJR argue the risk index captures true institutional quality only noisily: **the index is based on surveys of foreign investors and changes year-to-year**, introducing classical measurement error

# Discussion

**2.** AJR argue that settler mortality is exogenous. What is the strongest threat to this argument, and how might it be addressed?

**KEY TAKEAWAYS**

- **Strongest threat:** countries with historically high mortality (tropical climates) may face modern disease burdens (malaria, parasites) → **direct** reduction of productivity today, not just through institutions
- **AJR's response:** include malaria prevalence and latitude as controls → results are robust. They also use multiple instruments (3 versions of the mortality measure) and fail to reject instrument validity.
- **Remaining skepticism:** tropical geography may affect institutions and income through many channels → the exclusion restriction cannot be fully tested and must ultimately rely on institutional knowledge

# Discussion

**3.** The sample has only 64 countries. Does this affect the reliability of the TSLS estimate?

**KEY TAKEAWAYS**

- Small sample → larger standard errors and less precise estimates
- But the TSLS coefficient is still statistically significant (the first-stage $F$ = 23.3)
- More serious concern: **external validity**. The sample is restricted to former European colonies → the results may not generalize to non-colonized countries or to the modern institutional environment