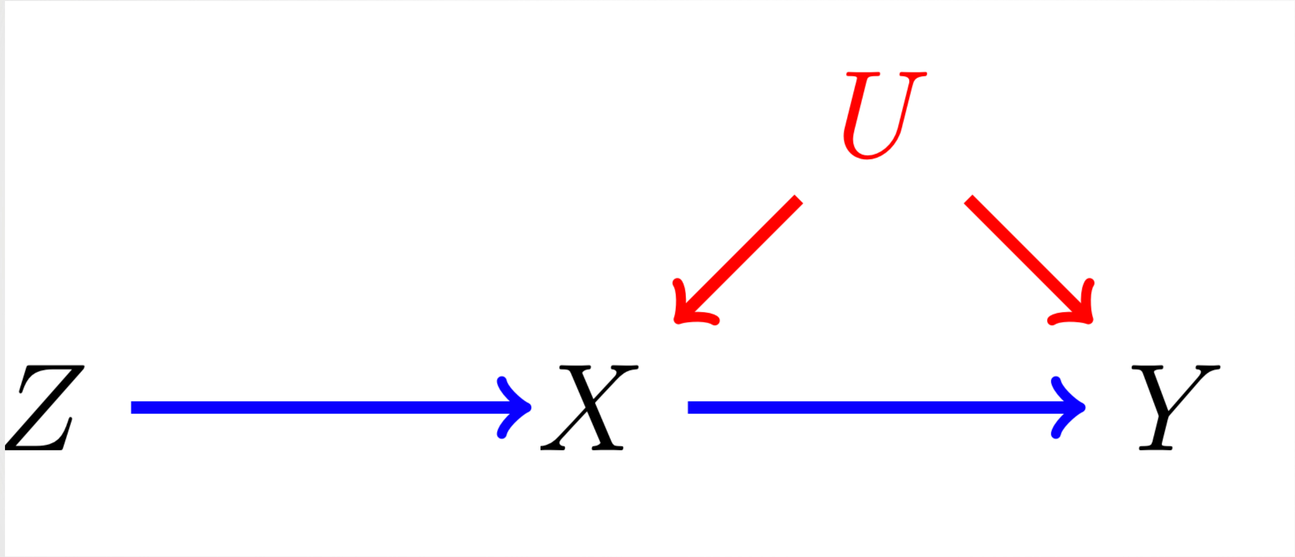
9 – INSTRUMENTAL VARIABLES

THE PLAN

1. Instrumental Variables (IV) regression.
2. Example: "The slave trade and the origins of mistrust in Africa" (Nunn & Wantchekon, 2011)
3. Statistical Inference
4. IV regression with control variables & multiple instruments.
5. Example: "Multiple Experiments for the Causal Link between the Quantity and Quality of Children" (Angrist et al, 2010)

INSTRUMENTAL VARIABLES (IV): OVERVIEW



9.1 INSTRUMENTAL VARIABLES REGRESSION

University *of* Massachusetts Amherst

SOME JARGON

- Endogenous
- Exogenous
- Endogeneity

IV REGRESSION

- $\beta_1^\star$ = the true causal effect of X on Y.
- Linear *causal* regression model:

$$Y_i = \beta_0 + \beta_1^\star X_i + u_i$$

- BUT X is endogenous: $corr(X_i, u_i) \neq 0$
  - $\rightarrow E(\hat{\beta}_1^{OLS}) \neq \beta_1^\star$
- How do we estimate $\beta_1^\star$?
- IV can be a way.

IV REGRESSION

- IV regression breaks X into:
  - an endogenous component
  - an exogenous component
- This is done using an *instrumental variable Z*
- *Z* must be:

1. Relevant: $corr(Z_i, X_i) \neq 0$
2. Exogenous: $corr(Z_i, u_i) = 0$

THE 2SLS ESTIMATOR

- The IV method is implemented through the two-stages-least squares (2SLS) estimator.

1st stage: a OLS regression for the effect of Z on X:

$$X_i = \boxed{\pi_0 + \pi_1 Z_i} + \boxed{v_i}$$

Variation in X driven by Z, and therefore **exogenous** (uncorrelated with u)

Variation in X coming from sources other than Z: **endogenous** (correlated with u)

THE 2SLS ESTIMATOR

- The IV method is implemented through the two-stages-least squares (2SLS) estimator.

1st stage: OLS regression of X on Z

$$X_i = \pi_0 + \pi_1 Z_i + v_i$$

  - *Compute predicted values*: $\hat{X}_i = \hat{\pi}_0 + \hat{\pi}_1 Z_i$

2nd stage: OLS regression of *Y* on *predicted X*

$$Y_i = \beta_0 + \beta_1 \hat{X}_i + u_i$$

THE 2SLS ESTIMATOR

2nd stage:

$$Y_i = \beta_0 + \beta_1 \hat{X}_i + u_i$$

- $corr(Z, u) = 0 \rightarrow corr\left(\hat{X} = \hat{\pi}_0 + \hat{\pi}_1 Z_i, u\right) = 0 \rightarrow E\left(\hat{\beta}_1^{TSLS}\right) = \beta_1^{\star}$
- If the instrumental variable Z is relevant & exogenous, TSLS is an unbiased estimate of the *average causal effect* of X.

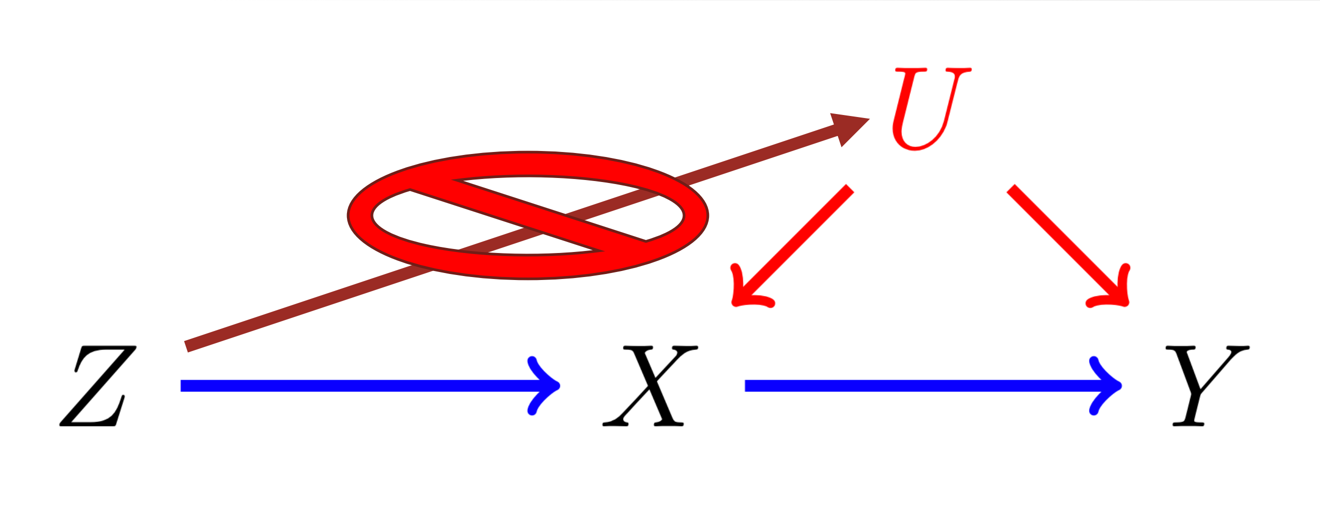
THE "EXCLUSION RESTRICTION"

- $corr(Z, u) = 0$ is not true if the instrumental variable has also an independent effect on Y, that does not involve X.

"Exclusion Restriction":

Z affects Y only through X.

- If that's not true, $corr(Z, u) \neq 0$ and the IV is not valid!



THE "EXCLUSION RESTRICTION"

- "Exclusion Restriction": *Z affects Y only through X.*
  - If that's not true, $corr(Z, u) \neq 0$ and the IV is not valid!

Does the "earthquake" instrument for school size (mentioned in your textbook) satisfy the exclusion restriction?

9.2 INSTRUMENTAL VARIABLES REGRESSION: EXAMPLES

University *of* Massachusetts Amherst

The Slave Trade and the Origins of Mistrust in Africa

Nathan Nunn

Leonard Wantchekon

AMERICAN ECONOMIC REVIEW
VOL. 101, NO. 7, DECEMBER 2011
(pp. 3221-52)

THE SLAVE TRADE AND MISTRUST IN AFRICA

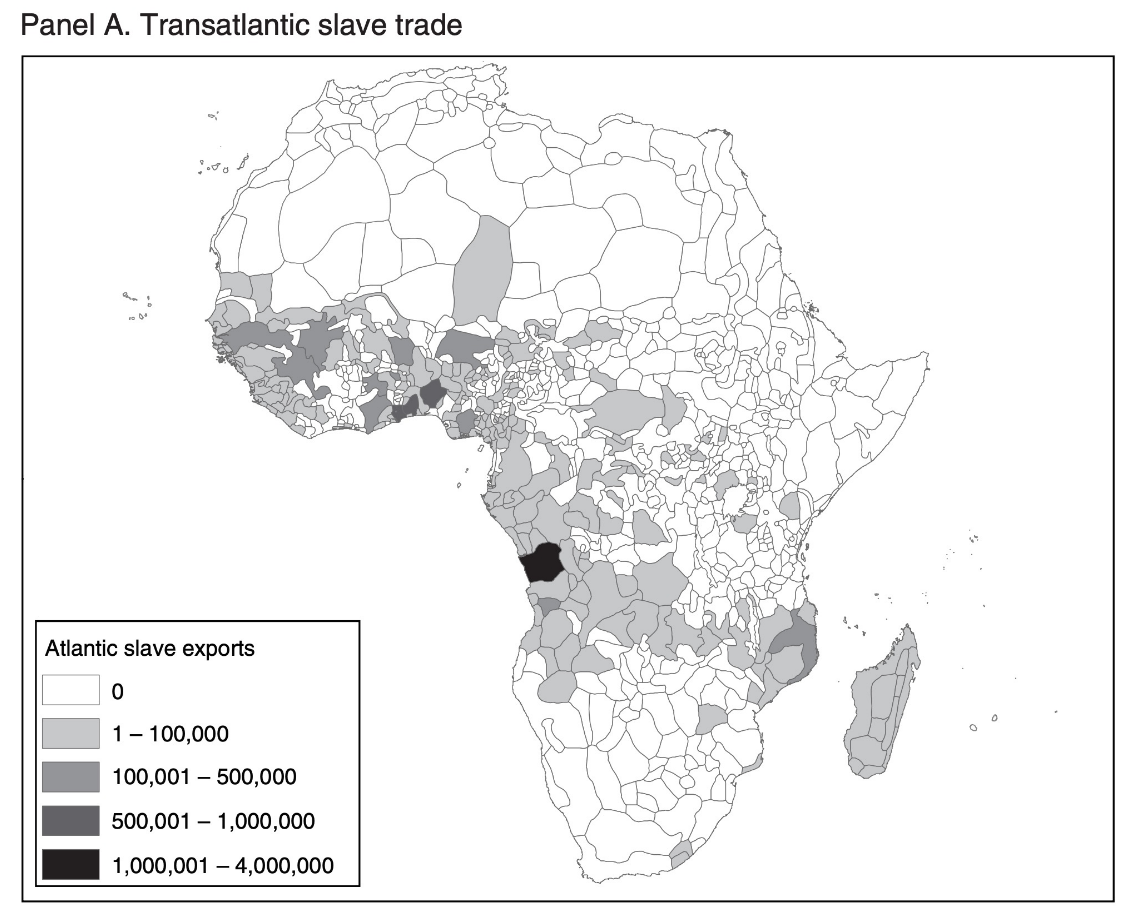
- *Hypothesis*: 500 years of slave trade caused a culture of mistrust to develop within Africa.
- *Empirical test*: do individuals belonging to ethnic groups that were most heavily raided by slave traders exhibit lower levels of trust today?
- *(Simplified) Regression*:

$$Trust_{i,e} = \beta_0 + \beta_1 Slave\ Exports_e + u_{i,e}$$

- *Endogeneity problem*:
  - possible omitted factors causing both low trust and more slave trade.
  - lower trust communities more easily raided (reverse causality)
  - → $\text{corr}(Slave\ Exports_e, u_{i,e}) \neq 0$

THE SLAVE TRADE AND MISTRUST IN AFRICA

- *Instrumental Variable:* the distance of an individual's ethnic group from the coast during the slave trade.
- *Relevant*: Traders purchased slaves on the coast before sailing overseas.
- *Exogenous*: distance from the coast is uncorrelated with other factors affecting trust.



THE SLAVE TRADE AND MISTRUST IN AFRICA

1st stage regression:

$$Slave\ Exports_e = \pi_0 + \pi_1 DistanceFromCoast_e + v_i$$

2nd stage regression:

$$Trust_{i,e} = \beta_0 + \beta_1 \widehat{Slave\ Exports}_e + u_{i,e}$$

- $\widehat{Slave\ Exports}_e$ is the predicted value from the 1st stage regression
- $\widehat{Slave\ Exports}_e = \hat{\pi}_0 + \hat{\pi}_1 DistanceFromCoast_e$

TABLE 5—IV ESTIMATES OF THE EFFECT OF THE SLAVE TRADE ON TRUST

	Trust of relatives (1)	Trust of neighbors (2)	Trust of local council (3)	Intragroup trust (4)	Intergroup trust (5)
Second stage: Dependent variable is an individual's trust					
ln (1 + exports/area)	−0.190***	−0.245***	−0.221***	−0.251***	−0.174**
	(0.067)	(0.070)	(0.060)	(0.088)	(0.080)
Hausman test (*p*-value)	0.88	0.53	0.09	0.44	0.41
R^2	0.13	0.16	0.20	0.15	0.12
First stage: Dependent variable is ln (1 + exports/area)					
Historical distance of ethnic group from coast	−0.0014***	−0.0014***	−0.0014***	−0.0014***	−0.0014***
	(0.0003)	(0.0003)	(0.0003)	(0.0003)	(0.0003)
Colonial population density	Yes	Yes	Yes	Yes	Yes
Ethnicity-level colonial controls	Yes	Yes	Yes	Yes	Yes
Individual controls	Yes	Yes	Yes	Yes	Yes
District controls	Yes	Yes	Yes	Yes	Yes
Country fixed effects	Yes	Yes	Yes	Yes	Yes
Number of observations	16,709	16,679	15,905	16,636	16,473
Number of clusters	147 / 1,187	147 / 1,187	146 / 1,194	147 / 1,186	147 / 1,184
F-stat of excl. instrument	26.9	26.8	27.4	27.1	27.0
R^2	0.81	0.81	0.81	0.81	0.81

Notes: The table reports IV estimates. The top panel reports the second-stage estimates, and the bottom panel reports first-stage estimates. Standard errors are adjusted for two-way clustering at the ethnicity and district levels.

Estimated $\hat{\beta}_1$ from 2nd stage regression

Estimated $\hat{\pi}_1$ from 1st stage regression

TABLE 8—REDUCED FORM RELATIONSHIP BETWEEN THE DISTANCE FROM THE COAST AND TRUST WITHIN AND OUTSIDE OF AFRICA

	Intergroup trust				
	Afrobarometer sample		WVS non-Africa sample		WVS Nigeria
	(1)	(2)	(3)	(4)	(5)
Distance from the coast	0.00039***	0.00037***	−0.00020	−0.00019	0.00054***
	(0.00013)	(0.00012)	(0.00014)	(0.00012)	(0.00010)
Country fixed effects	Yes	Yes	Yes	Yes	n/a
Individual controls	No	Yes	No	Yes	Yes
Number of observations	19,970	19,970	10,308	10,308	974
Number of clusters	185	185	107	107	16
R^2	0.09	0.10	0.09	0.11	0.06

Notes : The table reports OLS estimates. The unit of observation is an individual. The dependent variable in the WVS sample is the respondent's answer to the question: "How much do you trust <nationality> people in general?" The categories for the respondent's answers are: "not at all," "not very much," "neither trust nor distrust," "a little," and "completely." The responses take on the values 0, 1, 1.5, 2, and 3. Standard errors are clustered at the ethnicity level in the Afrobarometer regressions and at the location (city) level in the Asiabarometer and the WVS samples. The individual controls are for age, age squared, a gender indicator, an indicator for living in an urban location, and occupation fixed effects.

***Significant at the 1 percent level.
**Significant at the 5 percent level.
*Significant at the 10 percent level.

- Does the "exclusion restriction" hold here?
- Does distance from coast affect trust only through exposure to the slave trade?
- You can never directly test the exclusion restriction (bummer!)
- BUT it is informative to check if Z affects Y in a sample in which there is no link from Z to X.
- *"If distance from the coast affects trust only through the slave trade (ie, if our exclusion restriction is satisfied), then there should be no relationship between distance from coast and trust outside of Africa, where there was no slave trade". (Nunn & Wantchekon, 2011, p.3223)*

PREVIOUSLY ON ECON 452…

- Instrumental variables (IV): key ideas & assumptions.
- Implementation: The Two-Stages-Least-Squares (2SLS) estimator.
- The exclusion restriction.
- Application: The slave trade and mistrust in Africa.

9.3 INSTRUMENTAL VARIABLES REGRESSION: STATISTICAL INFERENCE

University *of* Massachusetts Amherst

2SLS: AN IMPORTANT EQUIVALENCE

- The 2SLS estimator is also equal to:

$$\hat{\beta}_1^{TSLS} = \frac{s_{\hat{X}Y}}{s_{\hat{X}}^2} = \frac{s_{ZY}}{s_{ZX}} = \frac{\beta_1^{ZY}}{\pi_1^{ZX}}$$

Slope of a OLS regression of Y on Z ("reduced form" regression)

Slope of a OLS regression of X on Z ("1st stage" regression)

- By the CLT, $\hat{\beta}_1^{TSLS} = \frac{s_{ZY}}{s_{ZX}}$ is normal in large samples.
  - $\hat{\beta}_1^{TSLS} \sim N(0, \sigma^2_{\hat{\beta}_1^{TSLS}})$
- $var\left(\hat{\beta}_1^{TSLS}\right) = var\left(\frac{s_{ZY}}{s_{ZX}}\right)$

2SLS: STATISTICAL INFERENCE

- $var\left(\hat{\beta}_1^{TSLS}\right) = var\left(\frac{s_{ZY}}{s_{ZX}}\right)$
- $SE\left(\hat{\beta}_1^{TSLS}\right)$ can thus be estimated using sample variances and covariances.
- Hypothesis tests, t-stats & p-values computed as usual.
- Statistical software will do it for you
  - '*ivregress*' command in STATA

APPLICATION: DEMAND FOR CIGARETTES

$$\ln\left(Q_i^{cigarettes}\right) = \beta_0 + \beta_1 \ln\left(P_i^{cigarettes}\right) + u_i$$

- *Data*: Cross-section of 48 US States in 1995
- Why is the OLS estimate of β_1 likely biased?
- Proposed instrumental variable:
  - Z_i = general sales tax per pack = $SalesTax_i$
  - Relevant?
  - Exogenous?

FIRST STAGE REGRESSION

```
. reg log_price sales_tax, robust

Linear regression                               Number of obs     =         48
                                                F(1, 46)          =      40.39
                                                Prob > F          =     0.0000
                                                R-squared         =     0.4710
                                                Root MSE          =     .09394
```

log_price	Coefficient	Robust std. err.	t	P>|t|	[95% conf.	interval]
sales_tax	.0201633	.0031729	6.35	0.000	.0137767	.0265499
_cons	5.037885	.0289177	174.21	0.000	4.979676	5.096093

2SLS ESTIMATE

```
. ivregress 2sls log_cigarettes (log_price = sales_tax), vce(robust)

Instrumental variables 2SLS regression          Number of obs   =        48
                                                Wald chi2(1)    =     12.05
                                                Prob > chi2     =    0.0005
                                                R-squared       =    0.4011
                                                Root MSE        =    .18635

------------------------------------------------------------------------------
             |               Robust
log_cigare~s | Coefficient  std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
   log_price |  -1.083587   .3122036    -3.47   0.001    -1.695494   -.4716788
       _cons |   10.17643   1.627667     6.25   0.000     6.986264     13.3666
------------------------------------------------------------------------------
Instrumented: log_price
 Instruments: sales_tax
```

- Credible?
- Is the instrument really exogenous?
- Are States with high vs low sales tax comparable?
- Do we need control variables? (states FEs? income?)

9.4 IV REGRESSION WITH CONTROL VARIABLES & MULTIPLE INSTRUMENTS

University *of* Massachusetts Amherst

A MORE GENERAL IV-2SLS ESTIMATOR

- Regression of interest:

$$Y_i = \beta_0 + \beta_1 X_i + \gamma_1 W_{1,i} + \cdots + \gamma_r W_{r,i} + u_i$$

  - 1 endogenous regressor of interest X
  - *r* control variables W

- We have *m* instruments Z for the endogenous regressor.

- **1st stage:** $X_i = \pi_0 + \pi_1 Z_{1,i} + \cdots + \pi_m Z_{m,i} + \delta_1 W_{1,i} + \cdots + \delta_r W_{r,i} + v_i$

- **2nd stage:** $Y_i = \beta_0 + \beta_1 \hat{X}_i + \gamma_1 W_{1,i} + \cdots + \gamma_r W_{r,i} + u_i$

KEY ASSUMPTIONS FOR CAUSAL INFERENCE

- The instruments are relevant:

$corr(Z_i, X_i) \neq 0$ *for at least one* of the instruments.

  - 1st stage F-statistics (>10 should be good enough)

- The instruments are exogenous, *after controlling for W:*

$corr(Z_i, u_i) = 0$ *for all* instruments.

  - *Now* u_i *does not include the W variables!*
  - J-Statistics

THE J-STATISTICS

- With multiple IVs, can perform *test of overidentifying restrictions*.
- Idea: if all instruments were valid, using each alone would produce similar estimates (estimates would differ only because of sampling variation).
- If estimates are very different, something is off…
  - At least some IV is picking up something different from the effect of X!
- *If J-Statistics rejects the null* ($p<0.05$), at least one of the instruments is probably invalid.
- *If J-Statistics does not reject the null* ($p>0.05$), instruments might or might not be valid.

9.5 MULTIPLE INSTRUMENTS: EXAMPLE

University *of* Massachusetts Amherst

THE EFFECT OF FAMILY SIZE ON CHILDREN'S EDUCATION

Multiple Experiments for the Causal Link between the Quantity and Quality of Children

Joshua Angrist, *MIT and NBER*

Victor Lavy, *Hebrew University, Royal Holloway University of London, and NBER*

Analia Schlosser, *Tel Aviv University*

- Does smaller family size allow parents to invest more in children's education?
- OLS bias: families with many vs. few kids are different in many respects.
- Random sample of families with ≥ 2 kids.
- Two (binary) instrumental variables:
    1. *Second-born twins*
    2. *Same-sex sibships*

TABLE 3.4
Quantity-quality first stages

	Twins instruments		Same-sex instruments		Twins and same-sex instruments
	(1)	(2)	(3)	(4)	(5)
Second-born twins	.320 (.052)	.437 (.050)			.449 (.050)
Same-sex sibships			.079 (.012)	.073 (.010)	.076 (.010)
Male		−.018 (.010)		−.020 (.010)	−.020 (.010)
Controls	No	Yes	No	Yes	Yes

Notes: This table reports coefficients from a regression of the number of children on instruments and covariates. The sample size is 89,445. Standard errors are reported in parentheses.

First stage regression:

Effect of second-born twin and same-sex sibship on family size (number of children).

TABLE 3.5

OLS and 2SLS estimates of the quantity-quality trade-off

		2SLS estimates		
Dependent variable	OLS estimates (1)	Twins instruments (2)	Same-sex instruments (3)	Twins and same-sex instruments (4)
Years of schooling	−.145 (.005)	.174 (.166)	.318 (.210)	.237 (.128)
High school graduate	−.029 (.001)	.030 (.028)	.001 (.033)	.017 (.021)
Some college (for age ≥ 24)	−.023 (.001)	.017 (.052)	.078 (.054)	.048 (.037)
College graduate (for age ≥ 24)	−.015 (.001)	−.021 (.045)	.125 (.053)	.052 (.032)

Notes: This table reports OLS and 2SLS estimates of the effect of family size on schooling. OLS estimates appear in column (1). Columns (2), (3), and (4) show 2SLS estimates constructed using the instruments indicated in column headings. Sample sizes are 89,445 for rows (1) and (2); 50,561 for row (3); and 50,535 for row (4). Standard errors are reported in parentheses.

2SLS estimates:

Effect of family size on the education level of the first-born child.

No negative effect in this sample!