Lecture 20: Instrumental Variables Models

ECON 480 - Econometrics - Fall 2018

Ryan Safner

December 5, 2018

Instrumental Variables Models

Two-Stage Least Squares

An Example of 2SLS in R

Instrumental Variables Models

- **Endogeneity** remains the hardest (and most common) econometric challenge

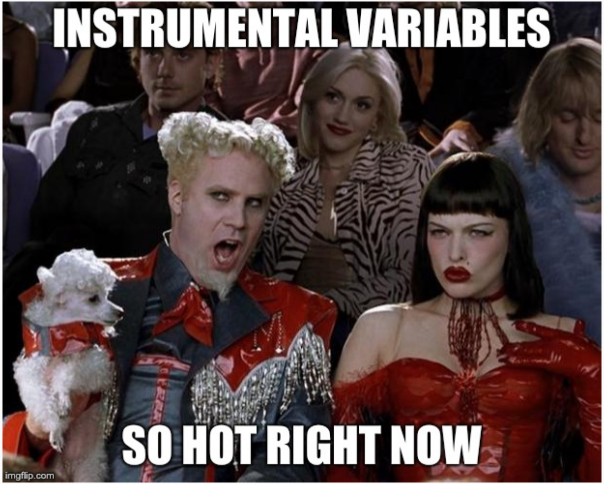
- **Endogeneity** remains the hardest (and most common) econometric challenge
- Additionally, there may be **reverse causation** or **simultaneous causation** between $X \iff Y$

- **Endogeneity** remains the hardest (and most common) econometric challenge
- Additionally, there may be **reverse causation** or **simultaneous causation** between $X \iff Y$
- DND/Fixed Effects are one strategy to remove endogeneity

- **Endogeneity** remains the hardest (and most common) econometric challenge
- Additionally, there may be **reverse causation** or **simultaneous causation** between $X \iff Y$
- DND/Fixed Effects are one strategy to remove endogeneity
    - *Requires* panel data

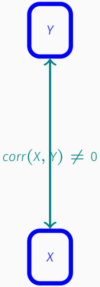
- **Endogeneity** remains the hardest (and most common) econometric challenge
- Additionally, there may be **reverse causation** or **simultaneous causation** between $X \iff Y$
- DND/Fixed Effects are one strategy to remove endogeneity
    - *Requires* panel data
    - Can't use time-varying omitted variables that are correlated with regressors

- **Endogeneity** remains the hardest (and most common) econometric challenge
- Additionally, there may be **reverse causation** or **simultaneous causation** between $X \iff Y$
- DND/Fixed Effects are one strategy to remove endogeneity
    - *Requires* panel data
    - Can't use time-varying omitted variables that are correlated with regressors
- Another strategy to is to find some source of exogenous variation that removes the endogeneity of a variable, using that source as a instrumental variable

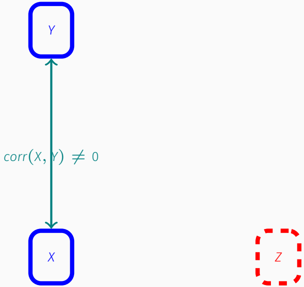

INSTRUMENTAL VARIABLES
SO HOT RIGHT NOW
imgflip.com

Understanding Instrumental Variables

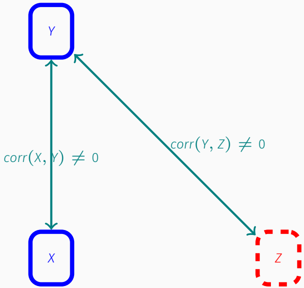
Y



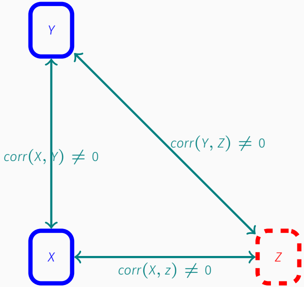
- X and Y correlated



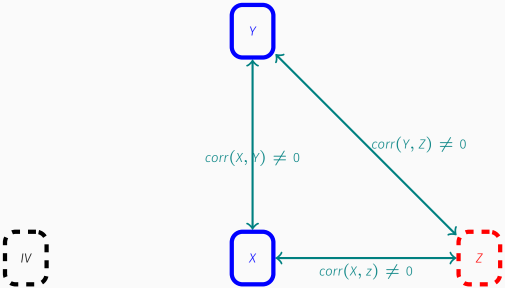
- X and Y correlated
- Omitted Variable, Z



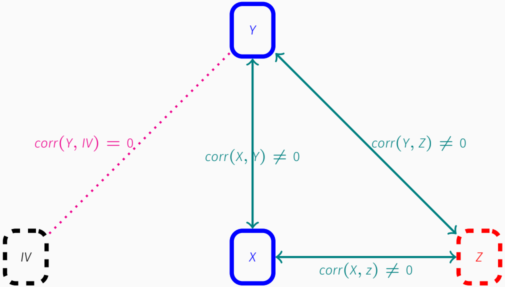
- X and Y correlated
- Omitted Variable, Z
    - Z explains Y (Z in ϵ)



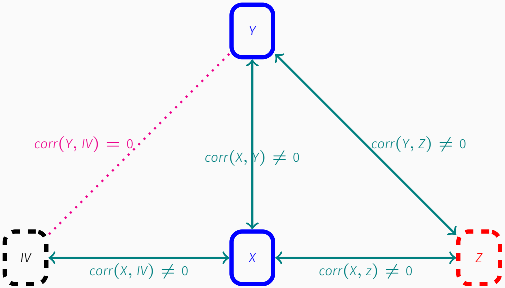
- X and Y correlated
- Omitted Variable, Z
  - Z explains Y (Z in ϵ)
  - Z & X correlated (X endogenous)



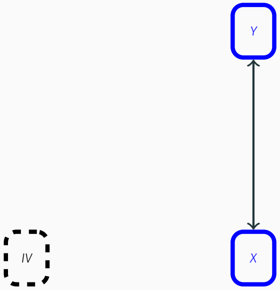
- X and Y correlated
- Omitted Variable, Z
  - Z explains Y (Z in ϵ)
  - Z & X correlated (X endogenous)
- **Instrumental Variable,** IV



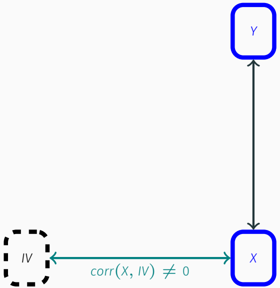
- X and Y correlated
- Omitted Variable, Z
    - Z explains Y (Z in ϵ)
    - Z & X correlated (X endogenous)
- **Instrumental Variable,** IV
    - IV doesn't explain Y (IV *not* ϵ)



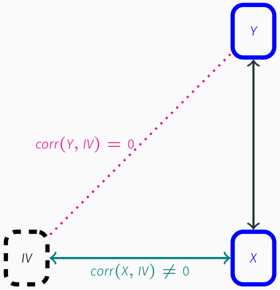
- X and Y correlated
- Omitted Variable, Z
  - Z explains Y (Z in ϵ)
  - Z & X correlated (X endogenous)
- **Instrumental Variable,** IV
  - IV doesn't explain Y (IV *not* ϵ)
  - IV & X correlated (X exogenous)



- **Variable** *IV* is a good **instrument** for *X* if it satisfies two conditions:

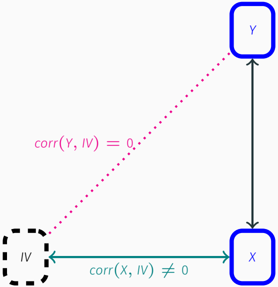


- **Variable** *IV* is a good **instrument** for *X* if it satisfies two conditions:
    1. Inclusion Condition:
       IV statistically significantly explains *X*



- **Variable** *IV* is a good **instrument** for *X* if it satisfies two conditions:
    1. Inclusion Condition:
       IV statistically significantly explains *X*
    2. Exclusion Condition:
       IV uncorrelated with ϵ, so doesn't directly affect *Y*

Understanding Instrumental Variables II



- **Variable** *IV* is a good **instrument** for *X* if it satisfies two conditions:
    1. Inclusion Condition:
       IV statistically significantly explains *X*
    2. Exclusion Condition:
       IV uncorrelated with ϵ, so doesn't directly affect *Y*
- *IV* **only** affects *Y* **through** *X*

Example
How do police affect crime?

$$Crime_{it} = \beta_0 + \beta_1 Police_{it} + \epsilon_{it}$$

- Police$\longrightarrow$crime (more police reduces crime)

Example
How do police affect crime?

$$Crime_{it} = \beta_0 + \beta_1 Police_{it} + \epsilon_{it}$$

- Police$\rightarrow$crime (more police reduces crime)
- Crime$\rightarrow$Police (high crime areas tend to have more police)

Example
How do police affect crime?

$$Crime_{it} = \beta_0 + \beta_1 Police_{it} + \epsilon_{it}$$

- Police$\longrightarrow$crime (more police reduces crime)
- Crime$\longrightarrow$Police (high crime areas tend to have more police)
- $corr(Police, \epsilon) \neq 0$: population, income per capita, drug use, recessions, demography, etc.

Example

- Levitt (2002): use number of firefighters as an instrumental variable

Levitt, Steven D, (2002), "Using Electoral Cycles in Police Hiring to Estimate the Effect of Police on Crime: Reply," *American Economic Review* 92(4): 1244-1250

Example: Levitt (2002)

Example

- Levitt (2002): use number of firefighters as an instrumental variable
- Some police are hired for *endogenous* reasons (respond to crime, changes in economy, demographics, etc)

Levitt, Steven D, (2002), "Using Electoral Cycles in Police Hiring to Estimate the Effect of Police on Crime: Reply," *American Economic Review* 92(4): 1244-1250

Example

- Levitt (2002): use number of firefighters as an instrumental variable
- Some police are hired for *endogenous* reasons (respond to crime, changes in economy, demographics, etc)
- Some police are hired for *exogenous* reasons (city just gains a larger budget and so hires more police)

Levitt, Steven D, (2002), "Using Electoral Cycles in Police Hiring to Estimate the Effect of Police on Crime: Reply," *American Economic Review* 92(4): 1244-1250

Example: Levitt (2002)

Example

- Levitt (2002): use number of firefighters as an instrumental variable
- Some police are hired for *endogenous* reasons (respond to crime, changes in economy, demographics, etc)
- Some police are hired for *exogenous* reasons (city just gains a larger budget and so hires more police)
  - These exogenous dynamics affect the number of firefighters in a city—*not* due to crime, but due to excess budgets, etc.

Levitt, Steven D, (2002), "Using Electoral Cycles in Police Hiring to Estimate the Effect of Police on Crime: Reply," *American Economic Review* 92(4): 1244-1250

Example: Levitt (2002)

Example

- Levitt (2002): use number of firefighters as an instrumental variable
- Some police are hired for *endogenous* reasons (respond to crime, changes in economy, demographics, etc)
- Some police are hired for *exogenous* reasons (city just gains a larger budget and so hires more police)
  - These exogenous dynamics affect the number of firefighters in a city—*not* due to crime, but due to excess budgets, etc.
- Isolate that portion of variation in Police that covaries with Firefighters for those *exogenous* changes (i.e. for reasons *other* than crime or its causes), see how *these* changes in Police affect crime

Levitt, Steven D, (2002), "Using Electoral Cycles in Police Hiring to Estimate the Effect of Police on Crime: Reply," *American Economic Review* 92(4): 1244-1250

Two-Stage Least Squares

- We use an instrument via Two Stage Least Squares (2SLS) method:

[1] Consider for convenience X_{2i} representing other control variables

- We use an instrument via Two Stage Least Squares (2SLS) method:
- Our main equation is[1]:

$$\widehat{Y_i} = \widehat{\beta_0} + \widehat{\beta_1} X_{1i} + \widehat{\beta_2} X_{2i} + \widehat{\epsilon_i}$$

[1] Consider for convenience X_{2i} representing other control variables

- We use an instrument via Two Stage Least Squares (2SLS) method:
- Our main equation is[1]:

$$\widehat{Y}_i = \widehat{\beta_0} + \widehat{\beta}_1 X_{1i} + \widehat{\beta}_2 X_{2i} + \widehat{\epsilon}_i$$

- We proceed with the **First Stage** using the *IV* to explain X_{1i} in an **auxiliary regression**:

$$\widehat{X}_{1i} = \widehat{\gamma}_0 + \widehat{\gamma}_1 IV_i + \widehat{\gamma}_2 X_{2i} + \widehat{\nu}_i$$

[1] Consider for convenience X_{2i} representing other control variables

- We use an instrument via Two Stage Least Squares (2SLS) method:
- Our main equation is[1]:

$$\widehat{Y_i} = \widehat{\beta_0} + \widehat{\beta_1} X_{1i} + \widehat{\beta_2} X_{2i} + \widehat{\epsilon_i}$$

- We proceed with the **First Stage** using the *IV* to explain X_{1i} in an **auxiliary regression**:

$$\widehat{X_{1i}} = \widehat{\gamma_0} + \widehat{\gamma_1} IV_i + \widehat{\gamma_2} X_{2i} + \widehat{\nu_i}$$

- Then substitute in the *predicted* value of $\widehat{X_{1i}}$ for the original regression in the **Second Stage**:

$$\widehat{Y_i} = \widehat{\beta_0} + \widehat{\beta_1} \widehat{X_{1i}} + \widehat{\beta_2} X_{2i} + \widehat{\epsilon_i}$$

[1] Consider for convenience X_{2i} representing other control variables

Conditions for a Valid Instrument

- Recall, an IV needs to pass:

- Recall, an IV needs to pass:
  1. **Inclusion Condition**:*IV* statistically significantly explains X

Conditions for a Valid Instrument

- Recall, an IV needs to pass:
  1. **Inclusion Condition**: IV statistically significantly explains X
  2. **Exclusion Condition**: IV uncorrelated with ϵ, doesn't directly affect Y

Conditions for a Valid Instrument

- Recall, an IV needs to pass:
    1. **Inclusion Condition**:*IV* statistically significantly explains X
    2. **Exclusion Condition**: *IV* uncorrelated with ϵ, doesn't directly affect Y
- We test inclusion condition by examining the First Stage regression:

$$\widehat{X_{1i}} = \widehat{\gamma_0} + \widehat{\gamma_1} IV_i + \widehat{\gamma_2} X_{2i} + \widehat{\nu_i}$$

- Recall, an IV needs to pass:
  1. **Inclusion Condition**: *IV* statistically significantly explains *X*
  2. **Exclusion Condition**: *IV* uncorrelated with ϵ, doesn't directly affect *Y*
- We test inclusion condition by examining the First Stage regression:

$$\widehat{X_{1i}} = \widehat{\gamma}_0 + \widehat{\gamma}_1 IV_i + \widehat{\gamma}_2 X_{2i} + \widehat{\nu}_i$$

- See if γ_1 on IV_i is statistically significant (ROT: $|t\text{-statistic}| > 3$)

- Recall, an IV needs to pass:
    1. **Inclusion Condition**:*IV* statistically significantly explains *X*
    2. **Exclusion Condition**: *IV* uncorrelated with ϵ, doesn't directly affect *Y*
- We test inclusion condition by examining the First Stage regression:

$$\widehat{X_{1i}} = \widehat{\gamma_0} + \widehat{\gamma_1} IV_i + \widehat{\gamma_2} X_{2i} + \widehat{\nu_i}$$

- See if γ_1 on IV_i is statistically significant (ROT: $|t\text{-statistic}| > 3$)
- Always report the first stage regression results!

Conditions for a Valid Instrument II

- There is no statistical test for the exclusion restriction!

Conditions for a Valid Instrument II

- There is no statistical test for the exclusion restriction!
    - Hardest condition to meet

Conditions for a Valid Instrument II

- **There is no statistical test for the exclusion restriction!**
    - Hardest condition to meet
    - Must give a convincing argument for why $corr(IV, \epsilon) = 0$; $IV \not\Rightarrow Y$; and that *IV only* affects *Y* indirectly *through X*

Conditions for a Valid Instrument II

- **There is no statistical test for the exclusion restriction!**
    - Hardest condition to meet
    - Must give a convincing argument for why $corr(IV, \epsilon) = 0$; $IV \not\Rightarrow Y$; and that IV *only* affects Y indirectly *through* X
    - If $corr(IV, \epsilon) \neq 0$, we would want IV as a control in the regression (omitted variable bias!)

HOOD COLLEGE

- **There is no statistical test for the exclusion restriction!**
    - Hardest condition to meet
    - Must give a convincing argument for why $corr(IV, \epsilon) = 0$; $IV \not\Rightarrow Y$; and that *IV only* affects Y indirectly *through* X
    - If $corr(IV, \epsilon) \neq 0$, we would want IV as a control in the regression (omitted variable bias!)
    - If $corr(IV, \epsilon) = 0$, can't include in the Second Stage regression:

$$\widehat{Y_i} = \widehat{\beta_0} + \widehat{\beta_1} + \widehat{X_{1i}} + \widehat{\beta_2}X_{2i} + \widehat{\beta_3}IV_i + \widehat{\epsilon_i}$$

$\widehat{X_{1i}}$ is a linear function of X_{2i} and IV_i, perfect multicollinearity!

Example

- Levitt's (2002) paper, First Stage:

$$\widehat{ln(Police_{ct})} = \widehat{\gamma}_1 ln(Firefighters_{ct}) + \alpha_c + \tau_t + \widehat{\gamma} Controls_{ct} + \nu_{ct}$$

subscripts for city c at year t, two-way fixed effects: α_c city fixed-effects, τ_t year fixed-effects

Example

- Levitt's (2002) paper, First Stage:

$$ln(\widehat{Police_{ct}}) = \widehat{\gamma}_1 ln(Firefighters_{ct}) + \alpha_c + \tau_t + \widehat{\gamma} Controls_{ct} + \nu_{ct}$$

  subscripts for city c at year t, two-way fixed effects: α_c city fixed-effects, τ_t year fixed-effects

- Second Stage:

$$ln(\widehat{Crime_{ct}}) = \widehat{\beta}_1 ln(\widehat{Police_{ct-1}}) + \alpha_c + \tau_t + \widehat{\beta} Controls_{ct} + \epsilon_{ct}$$

  lag for police (last year's police force determines this year's crime rates)

Instrumental Variables 2SLS Example II

Example

TABLE 2—THE RELATIONSHIP BETWEEN FIREFIGHTE

Variable	First-stage estimates (dependent variable = ln(Police per capita) (i)	(ii)	(iii)
ln(Firefighters per capita)	0.251	0.236	0.206
	(0.050)	(0.054)	(0.050)
ln(Street and highway workers per capita)	—	—	0.014
			(0.014)
ln(State prisoners per capita)	—	−0.101	−0.077
		(0.022)	(0.022)
Unemployment rate	—	0.571	0.265
		(0.276)	(0.314)
State income per capita (×10,000)	—	0.150	0.211
		(0.004)	(0.005)
Effective abortion rate	—	0.033	0.045
(×100)	(0.013)	(0.013)	(0.026)
ln(City population)	—	0.040	−0.014
		(0.040)	(0.047)
Percentage black	—	0.361	0.493
		(0.204)	(0.264)
City-fixed effects and year dummies included?	yes	yes	yes
R^2:	0.947	0.952	0.962
Number of observations:	2,032	2,032	1,445

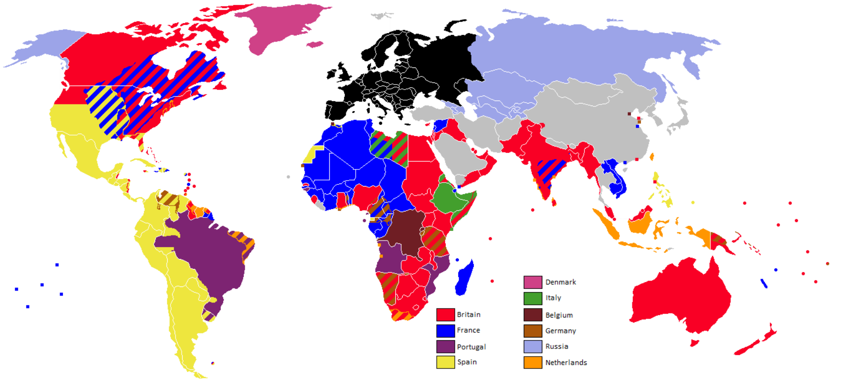
Instrumental Variables 2SLS Example II

Table 3—The Impact of Police on Crime

Variable	Violent crime OLS	Violent crime OLS	Violent crime IV	Property crime OLS	Property crime OLS	Property crime IV
ln(Police per capita)$_{t-1}$	0.562	−0.076	−0.435	0.113	−0.218	−0.501
	(0.056)	(0.061)	(0.231)	(0.038)	(0.052)	(0.235)
ln(State prisoners per capita)$_{t-1}$	0.250	−0.131	−0.171	0.189	−0.273	−0.305
	(0.039)	(0.036)	(0.044)	(0.030)	(0.028)	(0.037)
Unemployment rate	3.573	−0.741	−0.480	1.283	1.023	1.231
	(0.473)	(0.365)	(0.404)	(0.312)	(0.274)	(0.326)
State income per capita (×10,000)	0.050	−0.003	0.003	0.010	0.005	0.009
	(0.005)	(0.006)	(0.007)	(0.003)	(0.004)	(0.006)
Effective abortion rate (×100)	−0.214	−0.150	−0.141	−0.184	−0.118	−0.111
	(0.045)	(0.023)	(0.025)	(0.020)	(0.021)	(0.024)
ln(City population)	0.072	0.203	0.178	−0.064	−0.333	−0.355
	(0.012)	(0.063)	(0.067)	(0.006)	(0.063)	(0.066)
Percentage black	0.627	0.233	0.398	−0.136	0.411	0.517
	(0.074)	(0.334)	(0.345)	(0.057)	(0.271)	(0.291)
City-fixed effects and year dummies included?	only year dummies	yes	yes	only year dummies	yes	yes
R^2:	0.601	0.930	—	0.238	0.819	—
Number of observations:	2,005	2,005	2,005	2,032	2,032	2,032

- A 1% increase in police (last year) leads to a 0.435% decrease in violent crimes, 0.501%

Acemoglu, Johnson, and Robinson (2001): Another Famous IV Model



European Empires at their maximal extents (c.1500-c.1900)

Acemoglu, Daron, Simon Johnson, and James A Robinson, (2001), "The Colonial Origins of Comparative Development: An Empirical Investigation," *American Economic Review* 91(5): 1369-1401

- Acemoglu, Johnson, and Robinson (2001): countries' wealth or poverty today depends strongly on how they were colonized.

- Acemoglu, Johnson, and Robinson (2001): countries' wealth or poverty today depends strongly on how they were colonized.
- Europeans set up one of two types of colonies depending on the disease environment of the country:

- Acemoglu, Johnson, and Robinson (2001): countries' wealth or poverty today depends strongly on how they were colonized.
- Europeans set up one of two types of colonies depending on the disease environment of the country:
    - **"Extractive institutions"**: Europeans facing high mortality rates set up extractive colonies primarily to exploit indigenous population to mine resources to ship back to Europe

Acemoglu, Johnson, and Robinson (2001): Theoretical Argument

- Acemoglu, Johnson, and Robinson (2001): countries' wealth or poverty today depends strongly on how they were colonized.
- Europeans set up one of two types of colonies depending on the disease environment of the country:
    - **"Extractive institutions"**: Europeans facing high mortality rates set up extractive colonies primarily to exploit indigenous population to mine resources to ship back to Europe
    - **"Inclusive institutions"**: Europeans facing low mortality rates set up inclusive colonies primarily for settlement and promoting open access to trade and politics

Acemoglu, Johnson, and Robinson (2001): Theoretical Argument

- Acemoglu, Johnson, and Robinson (2001): countries' wealth or poverty today depends strongly on how they were colonized.
- Europeans set up one of two types of colonies depending on the disease environment of the country:
  - **"Extractive institutions"**: Europeans facing high mortality rates set up extractive colonies primarily to exploit indigenous population to mine resources to ship back to Europe
  - **"Inclusive institutions"**: Europeans facing low mortality rates set up inclusive colonies primarily for settlement and promoting open access to trade and politics
- Those initial colonies carried through to institutions in present countries; inclusive colonies grew wealthy, extractive colonies remain stagnant

Acemoglu, Johnson, and Robinson (2001): Identification Strategy

- Instrument: Settler Mortality in 1500

(potential) settler mortality $\Rightarrow$ settlements

$\Rightarrow$ early institutions $\Rightarrow$ current institutions

$\Rightarrow$ current performance.

Acemoglu, Johnson, and Robinson (2001): Identification Strategy

- Instrument: Settler Mortality in 1500
- **Inclusion Restriction**: Settler mortality in 1500 determines risk of expropriation today

(potential) settler mortality $\Rightarrow$ settlements

$\Rightarrow$ early institutions $\Rightarrow$ current institutions

$\Rightarrow$ current performance.

Acemoglu, Johnson, and Robinson (2001): Identification Strategy

- Instrument: Settler Mortality in 1500
- **Inclusion Restriction**: Settler mortality in 1500 determines risk of expropriation today
- {**Exclusion Restriction**: Settler mortality in 1500 **does not** affect Present GDP

(potential) settler mortality $\Rightarrow$ settlements $\Rightarrow$ early institutions $\Rightarrow$ current institutions $\Rightarrow$ current performance.

Acemoglu, Johnson, and Robinson (2001): Identification Strategy

- Instrument: Settler Mortality in 1500
- **Inclusion Restriction**: Settler mortality in 1500 determines risk of expropriation today
- {**Exclusion Restriction**: Settler mortality in 1500 **does not** affect Present GDP
    - Settler mortality in 1500 **only** affects Present GDP **through** institutions determined by historical path set by settler mortality rates

(potential) settler mortality $\Rightarrow$ settlements $\Rightarrow$ early institutions $\Rightarrow$ current institutions $\Rightarrow$ current performance.

- First Stage:

$$\widehat{\text{Expropriation}}\ \text{Risk}_i = \widehat{\gamma}_0 + \widehat{\gamma}_1 ln\left(\text{Settler Mortality in 1500}_i\right) + \widehat{\gamma} Controls + \nu_i$$

- First Stage:

$$\widehat{\text{Expropriation Risk}}_i = \widehat{\gamma}_0 + \widehat{\gamma}_1 ln\big(\text{Settler Mortality in 1500}_i\big) + \widehat{\gamma} Controls + \nu_i$$

- Second Stage:

$$\ln(\text{Present } \widehat{\text{GDP}} \text{ per capita}) = \hat{\beta}_0 + \hat{\beta}_1 \widehat{\text{Expropriation Risk}}_i + \cdots + \hat{\beta}_k \text{Controls} + \epsilon_i$$

ACEMOGLU, JOHNSON, AND ROBINSON (2001): RELATIONSHIP BETWEEN Y AND IV

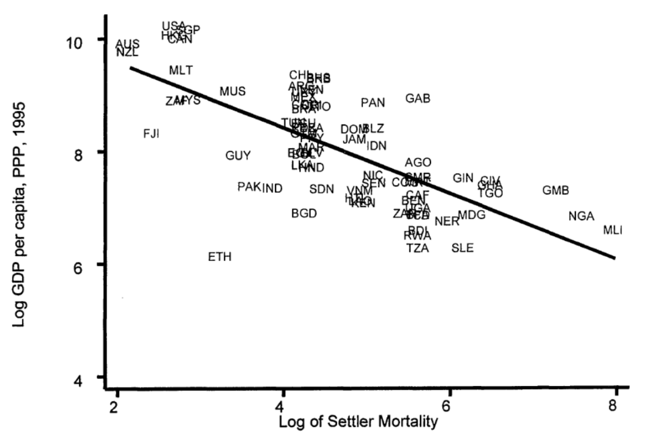


FIGURE 1. REDUCED-FORM RELATIONSHIP BETWEEN INCOME AND SETTLER MORTALITY

Acemoglu, Johnson, and Robinson (2001): Relationship Between X and Y

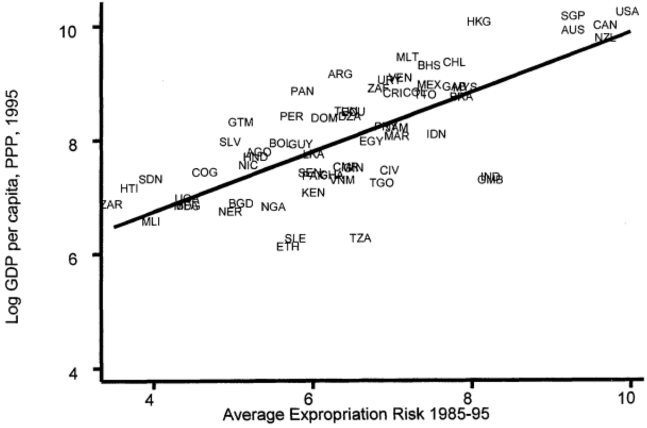


Figure 2. OLS Relationship Between Expropriation Risk and Income

Acemoglu, Johnson, and Robinson (2001): Relationship Between X and IV

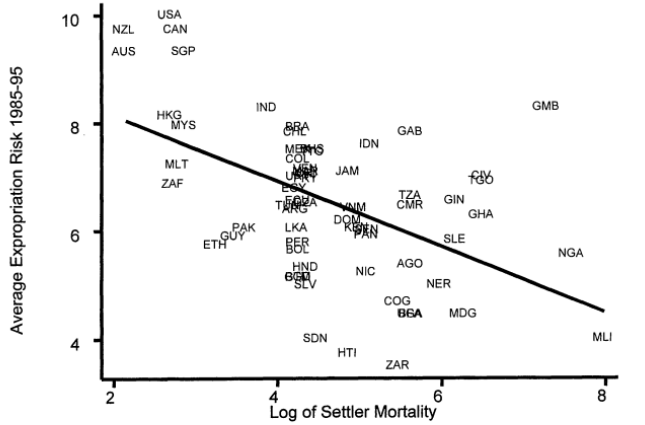


FIGURE 3. FIRST-STAGE RELATIONSHIP BETWEEN SETTLER MORTALITY AND EXPROPRIATION RISK

TABLE 4—IV REGRESSIONS OF LOG GDP PER CAPITA

	Base sample (1)	Base sample (2)	Base sample without Neo-Europes (3)	Base sample without Neo-Europes (4)	Base sample without Africa (5)	Base sample without Africa (6)	Base sample with continent dummies (7)	Base sample with continent dummies (8)	Base sample, dependent variable is log output per worker (9)
Panel A: Two-Stage Least Squares									
Average protection against expropriation risk 1985–1995	0.94	1.00	1.28	1.21	0.58	0.58	0.98	1.10	0.98
	(0.16)	(0.22)	(0.36)	(0.35)	(0.10)	(0.12)	(0.30)	(0.46)	(0.17)
Latitude		−0.65		0.94		0.04		−1.20	
		(1.34)		(1.46)		(0.84)		(1.8)	
Asia dummy							−0.92	−1.10	
							(0.40)	(0.52)	
Africa dummy							−0.46	−0.44	
							(0.36)	(0.42)	
"Other" continent dummy							−0.94	−0.99	
							(0.85)	(1.0)	
Panel B: First Stage for Average Protection Against Expropriation Risk in 1985–1995									
Log European settler mortality	−0.61	−0.51	−0.39	−0.39	−1.20	−1.10	−0.43	−0.34	−0.63
	(0.13)	(0.14)	(0.13)	(0.14)	(0.22)	(0.24)	(0.17)	(0.18)	(0.13)
Latitude		2.00		−0.11		0.99		2.00	
		(1.34)		(1.50)		(1.43)		(1.40)	
Asia dummy							0.33	0.47	
							(0.49)	(0.50)	
Africa dummy							−0.27	−0.26	
							(0.41)	(0.41)	
"Other" continent dummy							1.24	1.1	
							(0.84)	(0.84)	
R^2	0.27	0.30	0.13	0.13	0.47	0.47	0.30	0.33	0.28
Panel C: Ordinary Least Squares									
Average protection against expropriation risk 1985–1995	0.52	0.47	0.49	0.47	0.48	0.47	0.42	0.40	0.46
	(0.06)	(0.06)	(0.08)	(0.07)	(0.07)	(0.07)	(0.06)	(0.06)	(0.06)
Number of observations	64	64	60	60	37	37	64	64	61

An Example of 2SLS in R

Example

Does economic growth reduce the odds of civil conflict? Data ('RainIV.dta' on Blackboard) on 41 African countries between 1981-1999.

- `Internalconflict`: =1 if civil war ($<$ 25 deaths), else =0
- `LaggedGDPGrowth`: 1st lag of GDP growth
- `LagggedRainfallGrowth`: 1st lag of change in mm of rain from previous year
- Other controls

```r
# normal regression
reg1<-lm(InternalConflict~LaggedGDPGrowth, data=rain)
summary(reg1)
```

```
##
## Call:
## lm(formula = InternalConflict ~ LaggedGDPGrowth, data = rain)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.2999 -0.2689 -0.2660  0.7228  0.7876
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      0.26738    0.01631  16.389   <2e-16 ***
## LaggedGDPGrowth -0.08206    0.22485  -0.365    0.715
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4434 on 741 degrees of freedom
## Multiple R-squared:  0.0001797,  Adjusted R-squared:  -0.00117
## F-statistic: 0.1332 on 1 and 741 DF,  p-value: 0.7152
```

```r
# normal regression with controls
reg2<-lm(InternalConflict~LaggedGDPGrowth+InitialGDP+Democracy+Mountains+EthnicFrac+ReligiousFrac, data=rain)
summary(reg2)
```

```
##
## Call:
## lm(formula = InternalConflict ~ LaggedGDPGrowth + InitialGDP +
##     Democracy + Mountains + EthnicFrac + ReligiousFrac, data = rain)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.5654 -0.2811 -0.2221  0.4570  0.9459
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      0.0703555  0.0731012   0.962  0.33614
## LaggedGDPGrowth -0.1087977  0.2200999  -0.494  0.62123
## InitialGDP      -0.0569091  0.0182258  -3.122  0.00186 **
## Democracy        0.0012242  0.0028894   0.424  0.67193
## Mountains        0.0038654  0.0009527   4.057 5.49e-05 ***
## EthnicFrac       0.3247931  0.0918181   3.537  0.00043 ***
## ReligiousFrac    0.0105162  0.0958907   0.110  0.91270
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4334 on 736 degrees of freedom
## Multiple R-squared:  0.05106,    Adjusted R-squared:  0.04332
```

The First Stage

- Use `LaggedRainfallGrowth` as an instrument for `LaggedGDPGrowth`

```
rain.corr<-rain %>%
  select(InternalConflict, LaggedGDPGrowth, LaggedRainfallGrowth)

cor(rain.corr)
```

```
##                      InternalConflict LaggedGDPGrowth LaggedRainfallGrowth
## InternalConflict            1.0000000      -0.0134064           -0.0506118
## LaggedGDPGrowth            -0.0134064       1.0000000            0.1263399
## LaggedRainfallGrowth       -0.0506118       0.1263399            1.0000000
```

- Use `LaggedRainfallGrowth` as an instrument for `LaggedGDPGrowth`
  - **Inclusion restriction**: $corr(LaggedGDPGrowth, LaggedRainfallGrowth) \neq 0$

```
rain.corr<-rain %>%
  select(InternalConflict, LaggedGDPGrowth, LaggedRainfallGrowth)

cor(rain.corr)
```

```
##                      InternalConflict LaggedGDPGrowth LaggedRainfallGrowth
## InternalConflict            1.0000000      -0.0134064           -0.0506118
## LaggedGDPGrowth            -0.0134064       1.0000000            0.1263399
## LaggedRainfallGrowth       -0.0506118       0.1263399            1.0000000
```

- Use `LaggedRainfallGrowth` as an instrument for `LaggedGDPGrowth`
  - **Inclusion restriction**: $corr(LaggedGDPGrowth, LaggedRainfallGrowth) \neq 0$
  - **Exclusion restriction**: $corr(LaggedGDPGrowth, \epsilon) = 0$

```
rain.corr<-rain %>%
  select(InternalConflict, LaggedGDPGrowth, LaggedRainfallGrowth)

cor(rain.corr)
```

```
##                       InternalConflict LaggedGDPGrowth LaggedRainfallGrowth
## InternalConflict             1.0000000      -0.0134064           -0.0506118
## LaggedGDPGrowth             -0.0134064       1.0000000            0.1263399
## LaggedRainfallGrowth        -0.0506118       0.1263399            1.0000000
```

The First Stage II

```r
# normal regression with controls
firststage<-lm(LaggedGDPGrowth~LaggedRainfallGrowth+InitialGDP+Democracy+Mountains+EthnicFrac+ReligiousFrac, data=rain)
summary(firststage)
```

```
##
## Call:
## lm(formula = LaggedGDPGrowth ~ LaggedRainfallGrowth + InitialGDP +
##     Democracy + Mountains + EthnicFrac + ReligiousFrac, data = rain)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.46563 -0.03051  0.00459  0.03159  0.66704
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)          -0.0058041  0.0121558  -0.477 0.633168
## LaggedRainfallGrowth  0.0439770  0.0128127   3.432 0.000632 ***
## InitialGDP           -0.0008056  0.0030287  -0.266 0.790327
## Democracy             0.0005373  0.0004798   1.120 0.263080
## Mountains             0.0001087  0.0001582   0.687 0.492435
## EthnicFrac            0.0031603  0.0152584   0.207 0.835976
## ReligiousFrac        -0.0017176  0.0159357  -0.108 0.914197
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.07201 on 736 degrees of freedom
## Multiple R-squared:  0.01844,    Adjusted R-squared:  0.01044
```

From First to Second Stage in R

```r
# save predicted values of first stage
gdp.hat<-fitted.values(firststage)

# run second stage regression using gdp.hat instead of gdp
secondstage<-lm(InternalConflict~gdp.hat+InitialGDP+Democracy+Mountains+EthnicFrac+ReligiousFrac, data=rain)
summary(secondstage)
```

```
##
## Call:
## lm(formula = InternalConflict ~ gdp.hat + InitialGDP + Democracy +
##     Mountains + EthnicFrac + ReligiousFrac, data = rain)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.5653 -0.2830 -0.2237  0.4748  0.9952
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    0.0625061  0.0733775   0.852 0.394579
## gdp.hat       -2.0631528  1.7522115  -1.177 0.239394
## InitialGDP    -0.0580798  0.0182415  -3.184 0.001514 **
## Democracy      0.0023613  0.0030592   0.772 0.440438
## Mountains      0.0040694  0.0009691   4.199 3.01e-05 ***
## EthnicFrac     0.3288508  0.0918180   3.582 0.000364 ***
## ReligiousFrac  0.0047242  0.0959548   0.049 0.960746
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```