Instrumental Variables Design

Prof. Tzu-Ting Yang
楊子霆

Institute of Economics, Academia Sinica
中央研究院經濟研究所

May 22, 2023

Instrumental Variables Design: Main Idea

Main Idea of Instrumental Variables

- The Instrumental Variable (IV) is an exogenous sources of variation that drives the treatment D_i but unrelated to other confounding factors X_i that affect outcome Y_i
- Intuitively, IV breaks variation of the treatment D_i into two parts
  1. A part that might be correlated with other confounding factors X_i
  2. A part that is not
- We use the variation in D_i that is not correlated with X_i to estimate causal effect of the treatment

Unobservable Omitted Variable

- Suppose the true model is:

$$Y_i = \delta + \alpha D_i + \beta X_i + \epsilon_i$$

- But now X_i is the **unobserved characteristics**
  - e.g. ability, preference, health
- So we cannot include it into our regression and estimate the following model:

$$Y_i = \delta + \alpha D_i + u_i$$

- where $u_i = \beta X_i + \epsilon_i$

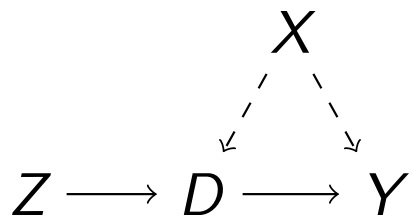
Unobservable Omitted Variable

- As mentioned before, failure to include key covariates will lead to omitted variable bias

$$\hat{\alpha} \xrightarrow{p} \alpha + \frac{Cov(u_i, D_i)}{V(D_i)}$$
$$= \alpha + \beta \frac{Cov(X_i, D_i)}{V(D_i)}$$

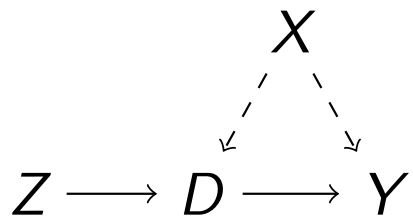
- Remember there is NO omitted variable bias (OVB) if D_i is unrelated to X_i
  1. X_i is unrelated to Y_i: $\beta = 0$
  2. X_i is unrelated to D_i: $Cov(X_i, D_i) = 0$
- To obtain causal effect (eliminate OVB), we need a variation in D_i that is unrelated to unobserved confounding factor X_i

Main Idea of Instrumental Variables



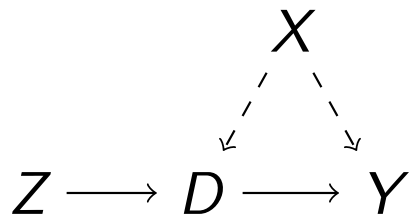
- Y is an outcome (e.g. earnings),
- Z is the instrument
- D is the treatment (e.g. college degree)
- X is the **unobserved** confounding factor (e.g. ability)

Main Idea of Instrumental Variables



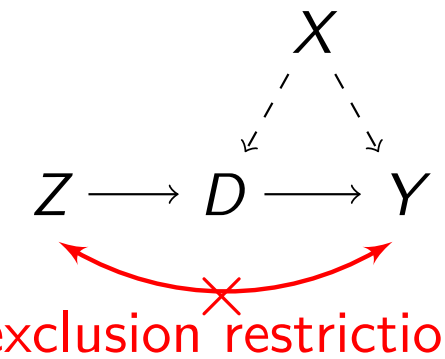
- **Unobserved ability** X might confound with the effect of college degree D
  - Since ability affects people to get college degree D and their earnings after graduation Y
- We need to find an IV that generate a variation in getting college degree D that is unrelated to ability X

Main Idea of Instrumental Variables



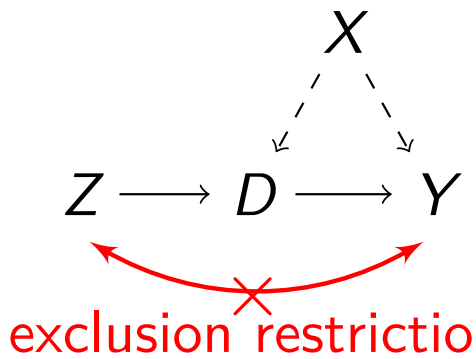
- IV initiates a causal chain: the instrument Z affects D, which in turn affects Y
- A valid IV needs to satisfy the following conditions:
  1. First-stage relationship (Instrument relevance): Z affects D

Main Idea of Instrumental Variables



- A valid IV needs to satisfy the following conditions:
  2. Exclusion restriction (Instrument exogeneity):
     - **No direct or indirect effect** of the instrument Z on the outcome Y NOT through the treatment variable D
     - The instrument Z affects the outcome Y **only through the treatment variable** D

Main Idea of Instrumental Variables



- We can test whether the **instrument relevance** is satisfied
- But the **instrument exogeneity** cannot be tested
- You have to try to convince your audience that it is satisfied

Instrumental Variables Design: Potential Outcome Framework

Example of Instrumental Variables

Joshua D. Angrist (1990) "**Lifetime Earnings and the Vietnam Era Draft Lottery: Evidence from Social Security Administrative Records**" AER

- He wanted to examine the effect of military service on lifetime income.
- We will use Angrist's paper on the effects of military service (D_i) on earnings (Y_i) as an example to go through key concept of IV design

Example of Instrumental Variables

- Joining military service is a personal choice
- Is there any selection bias due to **unobservable confounding factors** in this example ?
  - **Time preference:**
    - Less patient people may voluntarily join military service early
    - This myopic thinking may have negative impact on their earnings (e.g. less human capital investment)
  - **Health condition:**
    - Better health people can join military service
    - Better health condition also have positive impact on their earnings
- We need a IV for the treatment variable of joining military service

Example of Instrumental Variables

- Angrist (1990) uses the **Vietnam draft lottery** (Z_i) as in IV for military service
  - In the 1960s and early 1970s, young American men were draft for military service to serve in Vietnam
  - Concerns about the fairness of the conscription policy lead to the introduction of a **draft lottery** in 1970

Example of Instrumental Variables

- From 1970 to 1972 **random sequence numbers** were assigned to each birth date in cohorts of 19-year-olds
  - Men with lottery numbers below a cutoff were drafted while men with numbers above the cutoff could not be drafted
- The draft did NOT perfectly determinate military service:
  - Many draft-eligible men were exempted for health and other reasons
  - Draft-ineligible men volunteered for service
- Next, we briefly discuss whether draft eligibility induced by lottery is a good IV or not

Example of Instrumental Variables

- First-stage relationship (Instrument relevance): Z_i affects D_i
  - Vietnam veteran status (joining military service) was not completely determined by randomized draft eligibility
  - But draft eligibility is highly correlated with Vietnam veteran status
- Exclusion restriction (Instrument exogeneity):
  - The draft eligibility is determined by random numbers
  - These numbers should not affect one's earnings directly

IV and Potential Outcomes

- Treatment Assignment

$$Z_i = \begin{cases} 1 & \text{if an individual } i \text{ is eligible for a treatment} \\ 0 & \text{if an individual } i \text{ is not eligible for a treatment} \end{cases}$$

  - $Z_i = 1$: those who get draft eligibility
  - $Z_i = 0$: those who do not get draft eligibility
    - Due to lottery results

IV and Potential Outcomes Framework

- Potential Treatments
  - D_i^z: Potential treatment status given the value of Z
    - D_i^1: Potential treatment status if eligible for a treatment
    - D_i^0: Potential treatment status if not eligible for a treatment
- Observed Treatment

$$D_i = \begin{cases} \mathrm{D}_i^1 & \text{if } Z_i = 1 \\ \mathrm{D}_i^0 & \text{if } Z_i = 0 \end{cases}$$

- or, in a more compact notation: $\mathrm{D}_i = Z_i \mathrm{D}_i^1 + (1 - Z_i)\mathrm{D}_i^0$

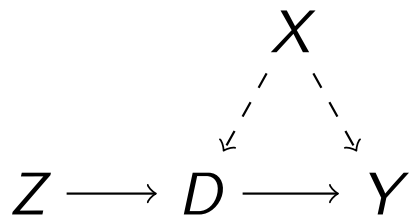
IV and Potential Outcomes Framework

- Potential Outcomes
  - Y_i^1: outcome if an individual *i* get treatment
    - Either $\mathrm{D}_i^1 = 1$ or $\mathrm{D}_i^0 = 1$
  - Y_i^0: outcome if an individual *i* does not get treatment
    - Either $\mathrm{D}_i^0 = 0$ or $\mathrm{D}_i^1 = 0$
- Observed Outcomes

$$Y_i = \begin{cases} \mathrm{Y}_i^1 & \text{if } \mathrm{D}_i^1 = 1 \text{ or } \mathrm{D}_i^0 = 1 \\ \mathrm{Y}_i^0 & \text{if } \mathrm{D}_i^0 = 0 \text{ or } \mathrm{D}_i^1 = 0 \end{cases}$$

- or, in a more compact notation: $\mathrm{Y}_i = D_i^z \mathrm{Y}_i^1 + (1 - D_i^z)\mathrm{Y}_i^0$

Identification Results for IV



- The IV method characterize a causal chain reaction leading from the instrument Z_i (draft eligibility) to outcome Y_i (earnings)
- Intuitively:

$$\begin{aligned}&\text{Effect of instrument on outcome}\\ &= (\text{Effect of instrument on treatment})\\ &\times (\text{Effect of treatment on outcome})\end{aligned}$$

Identification Results for IV

- Rearranging, the causal effect of military service on earnings is:

$$\text{Effect of treatment on outcome} = \frac{\text{Effect of instrument on outcome}}{\text{Effect of instrument on treatment}}$$

- Formal representation:

$$\alpha_{IV} = \frac{\mathrm{E}[Y_i|Z_i = 1] - \mathrm{E}[Y_i|Z_i = 0]}{\mathrm{E}[D_i|Z_i = 1] - \mathrm{E}[D_i|Z_i = 0]}$$

Identification Results for IV

- By using the following identification assumptions, we can prove the causal effect that IV identify is a **local average treatment effect (LATE)**

IV Identify LATE

$$\alpha_{IV} = \frac{\mathrm{E}[Y_i|Z_i = 1] - \mathrm{E}[Y_i|Z_i = 0]}{\mathrm{E}[D_i|Z_i = 1] - \mathrm{E}[D_i|Z_i = 0]} = \mathrm{E}[\mathrm{Y}_i^1 - \mathrm{Y}_i^0|\mathrm{D}_i^1 > \mathrm{D}_i^0]$$

Identification Assumptions for IV

First-Stage Relationship

- **First-Stage Relationship**: Z_i can affect treatment D_i

$$\mathrm{E}[D_i|Z_i = 1] - \mathrm{E}[D_i|Z_i = 0] \neq 0$$

  - Draft eligibility affects the probability of joining military

Identification Assumptions for IV

Independent Assumption

- **Independent Assumption**: Z_i is independent of potential outcomes and potential treatment (i.e. as good as randomly assigned)

$$(\mathrm{Y}_i^1, \mathrm{Y}_i^0, \mathrm{D}_i^1, \mathrm{D}_i^0) \perp\!\!\!\perp Z_i$$

  - Draft eligibility is unrelated to people's potential earnings and potential treatment status
    - $\mathrm{E}[\mathrm{D}_i^0 | Z_i = 1] = \mathrm{E}[\mathrm{D}_i^0 | Z_i = 0]$
    - $\mathrm{E}[\mathrm{D}_i^1 | Z_i = 1] = \mathrm{E}[\mathrm{D}_i^1 | Z_i = 0]$
    - $\mathrm{E}[\mathrm{Y}_i^0 | Z_i = 1] = \mathrm{E}[\mathrm{Y}_i^0 | Z_i = 0]$
    - $\mathrm{E}[\mathrm{Y}_i^1 | Z_i = 1] = \mathrm{E}[\mathrm{Y}_i^1 | Z_i = 0]$

Identification Assumptions for IV

Exclusion Restriction

- **Exclusion Restriction**: Z_i affects outcome Y_i only through changing treatment status D_i
  - The instrument has no direct effect on the outcome, once we fix the value of the treatment

$$Y_i = \begin{cases} \mathrm{Y}_i^1 & \text{if } \mathrm{D}_i^{z=1} = 1 \text{ or } \mathrm{D}_i^{z=0} = 1 \\ \mathrm{Y}_i^0 & \text{if } \mathrm{D}_i^{z=0} = 0 \text{ or } \mathrm{D}_i^{z=1} = 0 \end{cases}$$

Identification Assumptions for IV

Monotonicity Assumption

- **Monotonicity Assumption**: $\mathrm{D}_i^1 \geq \mathrm{D}_i^0$
  - Monotonicity says that the presence of the instrument never dissuades someone from taking the treatment
    - This is sometimes called **no defiers**
  - In the draft lottery example: draft eligibility should encourage people to join military service

IV and Compliers

- The variation in treatment D_i (veteran status) was not entirely from the draft eligibility Z_i but also from individual choice
- Thus, $\mathrm{D}_i^1 = 1$ or $\mathrm{D}_i^1 = 0$
  - $\mathrm{D}_i^1 = 1$: Those who get draft eligibility choose to join military service
  - $\mathrm{D}_i^1 = 0$: Those who get draft eligibility choose not to join military service
- Similarly, $\mathrm{D}_i^0 = 1$ or $\mathrm{D}_i^0 = 0$
  - $\mathrm{D}_i^0 = 1$: Those who did not get draft eligibility choose to join military service
  - $\mathrm{D}_i^0 = 0$: Those who did not get draft eligibility choose not to join military service

IV and Compliers

- We can define four types of individuals based on whether they fellow the draft eligibility results:
  - **Compliers**: $\mathrm{D}_i^1 > \mathrm{D}_i^0$ ($\mathrm{D}_i^0 = 0$ and $\mathrm{D}_i^1 = 1$)
    - David got draft eligibility and joined military service
    - Tim did not get draft eligibility and did not join military service
  - **Always-takers**: $\mathrm{D}_i^1 = \mathrm{D}_i^0 = 1$
    - John always joined military service no matter the lottery results (whether he got draft eligibility)
  - **Never-takers**: $\mathrm{D}_i^1 = \mathrm{D}_i^0 = 0$
    - Trump never joined military service no matter the lottery results (whether he got draft eligibility)
  - **Defiers**: $\mathrm{D}_i^1 < \mathrm{D}_i^0$ ($\mathrm{D}_i^0 = 1$ and $\mathrm{D}_i^1 = 0$)
    - Jimmy got draft eligibility but did NOT join military service
    - Jonson did NOT get draft eligibility but joined military service

Identification Results for IV

IV Identify LATE

$$\alpha_{IV} = \frac{\mathrm{E}[Y_i|Z_i = 1] - \mathrm{E}[Y_i|Z_i = 0]}{\mathrm{E}[D_i|Z_i = 1] - \mathrm{E}[D_i|Z_i = 0]} = \mathrm{E}[\mathrm{Y}_i^1 - \mathrm{Y}_i^0|\mathrm{D}_i^1 > \mathrm{D}_i^0]$$

- IV estimator represents the causal effect for compliers
- Lottery IV can identify the causal effect of military service on lifetime earnings for those who obey the lottery results (e.g. David and Tim)

Identification Results for IV

Proof:

$$\alpha_{IV} = \frac{\mathrm{E}[Y_i|Z_i = 1] - \mathrm{E}[Y_i|Z_i = 0]}{\mathrm{E}[D_i|Z_i = 1] - \mathrm{E}[D_i|Z_i = 0]}$$

$$= \frac{E[\mathrm{Y}_i^1\mathrm{D}_i^1 + \mathrm{Y}_i^0(1 - \mathrm{D}_i^1)|Z_i = 1] - E[\mathrm{Y}_i^1\mathrm{D}_i^0 + \mathrm{Y}_i^0(1 - \mathrm{D}_i^0)|Z_i = 0]}{E[\mathrm{D}_i^1|Z_i = 1] - E[\mathrm{D}_i^0|Z_i = 0]}$$

$$= \frac{E[\mathrm{Y}_i^1\mathrm{D}_i^1 + \mathrm{Y}_i^0(1 - \mathrm{D}_i^1)] - E[\mathrm{Y}_i^1\mathrm{D}_i^0 + \mathrm{Y}_i^0(1 - \mathrm{D}_i^0)]}{E[\mathrm{D}_i^1] - E[\mathrm{D}_i^0]}$$

$$= \frac{E[(\mathrm{Y}_i^1 - \mathrm{Y}_i^0)(\mathrm{D}_i^1 - \mathrm{D}_i^0)]}{E[\mathrm{D}_i^1] - E[\mathrm{D}_i^0]}$$

Identification Results for IV

- Note that since D_i^z is a dummy
- IV estimates cannot say anything about causal effect for **always takers** or **never takers**: $\mathrm{D}_i^1 - \mathrm{D}_i^0 = 0$
- $\mathrm{D}_i^1 - \mathrm{D}_i^0 = 1$ (**compliers**) or $\mathrm{D}_i^1 - \mathrm{D}_i^0 = -1$ (**defiers**)
- Using Monotonicity Assumption: only $\mathrm{D}_i^1 - \mathrm{D}_i^0 = 1$ exists
- Therefore, $E[(\mathrm{Y}_i^1 - \mathrm{Y}_i^0)(\mathrm{D}_i^1 - \mathrm{D}_i^0)]$ can become the following terms:

$$
\begin{aligned}
&E[(\mathrm{Y}_i^1 - \mathrm{Y}_i^0)(\mathrm{D}_i^1 - \mathrm{D}_i^0)] \\
&= E[(\mathrm{Y}_i^1 - \mathrm{Y}_i^0) \times (1) | \mathrm{D}_i^1 - \mathrm{D}_i^0 = 1] \Pr(\mathrm{D}_i^1 - \mathrm{D}_i^0 = 1) \\
&+ E[(\mathrm{Y}_i^1 - \mathrm{Y}_i^0) \times (-1) | \mathrm{D}_i^1 - \mathrm{D}_i^0 = -1] \Pr(\mathrm{D}_i^1 - \mathrm{D}_i^0 = -1) \\
&= E[(\mathrm{Y}_i^1 - \mathrm{Y}_i^0) \times (1) | \mathrm{D}_i^1 - \mathrm{D}_i^0 = 1] \Pr(\mathrm{D}_i^1 - \mathrm{D}_i^0 = 1)
\end{aligned}
$$

- Note that $E[\mathrm{D}_i^1] - E[\mathrm{D}_i^0] = \Pr(\mathrm{D}_i^1 - \mathrm{D}_i^0 = 1)$

Identification Results for IV

- Continue the LATE proof

Proof:

$$
\begin{aligned}
\alpha_{IV} &= \frac{\mathrm{E}[Y_i|Z_i=1]-\mathrm{E}[Y_i|Z_i=0]}{\mathrm{E}[D_i|Z_i=1]-\mathrm{E}[D_i|Z_i=0]} \\
&= \frac{E[\mathrm{Y}_i^1\mathrm{D}_i^1+\mathrm{Y}_i^0(1-\mathrm{D}_i^1)|Z_i=1]-E[\mathrm{Y}_i^1\mathrm{D}_i^0+\mathrm{Y}_i^0(1-\mathrm{D}_i^0)|Z_i=0]}{E[\mathrm{D}_i^1|Z_i=1]-E[\mathrm{D}_i^0|Z_i=0]} \\
&= \frac{E[\mathrm{Y}_i^1\mathrm{D}_i^1+\mathrm{Y}_i^0(1-\mathrm{D}_i^1)]-E[\mathrm{Y}_i^1\mathrm{D}_i^0+\mathrm{Y}_i^0(1-\mathrm{D}_i^0)]}{E[\mathrm{D}_i^1]-E[\mathrm{D}_i^0]} \\
&= \frac{E[(\mathrm{Y}_i^1-\mathrm{Y}_i^0)(\mathrm{D}_i^1-\mathrm{D}_i^0)]}{E[\mathrm{D}_i^1]-E[\mathrm{D}_i^0]} \\
&= \frac{E[(\mathrm{Y}_i^1-\mathrm{Y}_i^0)(1)|\mathrm{D}_i^1-\mathrm{D}_i^0=1]\,\mathrm{Pr}(\mathrm{D}_i^1-\mathrm{D}_i^0=1])}{\mathrm{Pr}(\mathrm{D}_i^1-\mathrm{D}_i^0=1)} \\
&= E[\mathrm{Y}_i^1-\mathrm{Y}_i^0|\mathrm{D}_i^1>\mathrm{D}_i^0]=\alpha_{\mathrm{LATE}}
\end{aligned}
$$

Identification Results for IV

- **Never takers** and **always takers** do NOT change their treatment status when the instrument gets switched on and off
  - So only **defiers** and **compliers** contribute to IV estimate
  - IV estimate is the sum of those two effects
- By using monotonicity assumption, we rule out the effect from **defiers**
- Therefore, IV estimates the **average treatment effect for compliers**

Identification Results for IV

LATE

$\alpha_{\text{LATE}} = E[\text{Y}_i^1 - \text{Y}_i^0 | \text{D}_i^1 > \text{D}_i^0]$, the average treatment effect (ATE) for compliers is often called **Local Average Treatment Effect (LATE)**.

- ATE for the individuals whose treatment status (join military service) are changed by the instrument (lottery draft)
  - This is the ATE among the compliers
- LATE (α_{LATE}) is different when using different instruments, Z_i
- Whether LATE is interesting or not depends on the instrument

Identification Results for IV

LATE and ATE

- Without further assumptions (e.g. constant causal effects), LATE is not informative about effects on never-takers or always-takers
  - Because the instrument does not affect their treatment status.
- In most applications we would be mostly interested in estimating the average treatment effect on the whole population (ATE).

$$E[\mathrm{Y}_i^1 - \mathrm{Y}_i^0]$$

- This is usually not possible with IV.

Identification Results for IV

LATE and ATT

Special Cases

- If D_i is randomized (e.g. RCT) and everybody is a complier, then $Z_i = D_i$
- That is, no never taker or always taker
- One-sided noncompliance, $\mathrm{D}_i^0 = 0$, then:

$$\begin{aligned} E[\mathrm{Y}_i^1 - \mathrm{Y}_i^0 | \mathrm{D}_i^1 > \mathrm{D}_i^0] &= E[\mathrm{Y}_i^1 - \mathrm{Y}_i^0 | \mathrm{D}_i^1 = 1] \\ &= E[\mathrm{Y}_i^1 - \mathrm{Y}_i^0 | Z_i = 1, \mathrm{D}_i^1 = 1] \\ &= E[\mathrm{Y}_i^1 - \mathrm{Y}_i^0 | \mathrm{D}_i = 1] \\ &= E[\mathrm{Y}_i^1 - \mathrm{Y}_i^0] \end{aligned}$$

$$\Rightarrow \quad \alpha_{\mathsf{LATE}} = \alpha_{\mathsf{ATT}} = \alpha_{\mathsf{ATE}}$$

Instrumental Variables Design: Estimation

Review: IV Estimation

A intuitive way

- Causal relationship of interest: the effect of military service on earnings

$$Y_i = \delta + \alpha_{IV} D_i + u_i$$

- Remember we just drive:

$$\begin{aligned} \alpha_{IV} &= \text{Effect of treatment on outcome} \\ &= \frac{\text{Effect of instrument on outcome}}{\text{Effect of instrument on treatment}} \end{aligned}$$

Review: IV Estimation

A intuitive way

- We can estimate α_{IV} by running the following two regressions:

- Reduced form regression: the effect of lottery draft on earnings

$$\mathrm{Y}_i = \mu + \alpha_{RF} Z_i + \varepsilon_i$$

$$\alpha_{RF} = \frac{Cov(Y_i, Z_i)}{V(Z_i)}$$

- First-Stage regression: the effect of lottery draft on military service

$$\mathrm{D}_i = \kappa + \alpha_{FS} Z_i + \zeta_i$$

$$\alpha_{FS} = \frac{Cov(D_i, Z_i)}{V(Z_i)}$$

- The IV estimator is:

$$\hat{\alpha}_{IV} = \frac{\hat{\alpha}_{RF}}{\hat{\alpha}_{FS}} = \frac{\hat{Cov}(Y_i, Z_i)}{\hat{Cov}(D_i, Z_i)}$$

Review: IV Estimation

Two Stage Last Squares (TSLS)

- In practice we often estimate IV using Two stage least squares estimation (TSLS)
- If identification assumptions only hold after conditioning on X, covariates are often introduced using TSLS regression
- It is called TSLS because you could estimate it as follows:
  1. Obtain the first stage fitted values:
  $$\hat{D}_i = \hat{\kappa} + \hat{\alpha}_{FS} Z_i + X'\hat{\beta}$$
  2. Plug the first stage fitted values into the "second-stage equation"
  $$Y_i = X'\gamma + \alpha_{TSLS}\hat{D}_i + u_i^*$$

Review: IV Estimation

Two Stage Least Squares (TSLS)

- However, this estimation is usually not done in two steps
- If you would do that the standard errors would be wrong
  - Following the two steps procedure, we calculate the standard errors based on u_i^*

$$u_i^* = Y_i - X'\gamma - \alpha_{TSLS}\hat{D}_i$$

  - But what we really want is the standard errors based on u_i:

$$u_i = Y_i - X'\gamma - \alpha_{TSLS}D_i$$

- STATA or other regression softwares are usually doing the job for you and get the standard errors right

Review: IV Estimation

Two Stage Least Squares (TSLS)

- The intuition of TSLS, however, is very useful:
  - TSLS only retains the variation in D_i that is generated by quasi-experimental variation Z_i (and thus hopefully exogenous)

Review: Inference in TSLS

- In large samples, the sampling distribution of the TSLS estimator is normal
- Inference (hypothesis tests, confidence intervals) proceeds in the usual way

Summary of Hypothesis Testing for TSLS Regression

- We estimate the following regressions and want to test whether there is treatment effect:

$$\mathrm{Y}_i = \delta + \alpha_{TSLS} D_i + X'\beta + u_i$$
$$\mathrm{D}_i = \kappa + \alpha_{FS} Z_i + X'\beta + \zeta_i$$

1. Check IV relevance (first stage):
   - Test whether α_{FS} is statistically significant different from zero
   - If it is not statistically significant (weak IV), you should change your IV
2. Choose a null hypothesis:
   - $H_0 : \alpha_{TSLS} = 0$ or $H_0 : \alpha_{TSLS} = \mu$
   - Claim we would like to reject

Summary of Hypothesis Testing for Regression

3. Choose a test statistic
   - $t = \frac{(\hat{\alpha}_{TSLS} - \alpha_{TSLS})}{\hat{\mathrm{SE}}(\hat{\alpha}_{TSLS})}$
4. Estimate standard error of the estimator
   - $\hat{\mathrm{SE}}(\hat{\alpha}_{TSLS}) = \frac{\sigma_u}{\sqrt{n}} \times \frac{1}{\sigma_{\overline{D}}}$
   - σ_u: standard deviation of u_i
   - $\sigma_{\overline{D}}$: standard deviation of first-stage fitted values

Summary of Hypothesis Testing for Regression

5. Determine the distribution of the test statistic under the null hypothesis
    - If sample size is sufficient large, using CLT, t-statistic will have standard normal distribution
6. Calculate the probability of wrongly reject null hypothesis given null hypothesis is true (p-value)
    - We reject the null hypothesis $H_0 : \alpha_{TSLS} = 0$ against the alternative $H_1 : \alpha_{TSLS} \neq 0$ at the 5% significance level if $|t| > 1.96$

Summary of Findings on Vietnam Draft Lottery

1. **First stage results:**
   - Having a low lottery number (being eligible for the draft) increases veteran status by about 16 percentage points
   - Note that the mean of veteran status is about 27 percent
2. **Second stage results:**
   - Serving in the army lowers earnings by between \$2,050 and \$2,741 per year.
3. **Placebo test:**
   - There is no evidence of an association between draft eligibility (having a low lottery number) and earnings in 1969
   - Note that 1969 earnings are realized before the 1970 draft lottery

Summary of Findings on Vietnam Draft Lottery

	Earnings		Veteran Status		Wald Estimate of Veteran Effect
Earnings year	Mean	Eligibility Effect	Mean	Eligibility Effect	
	(1)	(2)	(3)	(4)	(5)
1981	16,461	-435.8	0.267	0.159	-2,741
		(210.5)		(0.040)	(1,324)
1971	3,338	-325.9			-2050
		(46.6)			(293)
1969	2,299	-2.0			
		(34.5)			

Notes: Adapted from Angrist (1990), Tables 2 and 3. Standard errors are shown in parentheses. Earnings data are from Social Security administrative records. Figures are in nominal dollars. Veteran status data are from the Survey of Program Participation. There are about 13,500 individuals in the sample.

Instrumental Variables Design: STATA Example

STATA Example: Acemoglu, Johnson, and Robinson (2001)

Acemoglu, Johnson, and Robinson (2001) "**The Colonial Origins of Comparative Development: An Empirical Investigation**" AER

- They want to examine the effect of institutions on growth (or level) of income.
- Do countries with better institutions achieve a greater level of income ?
  - Good institutions: more secure property rights, less distortionary policies, invest more in physical and human capital

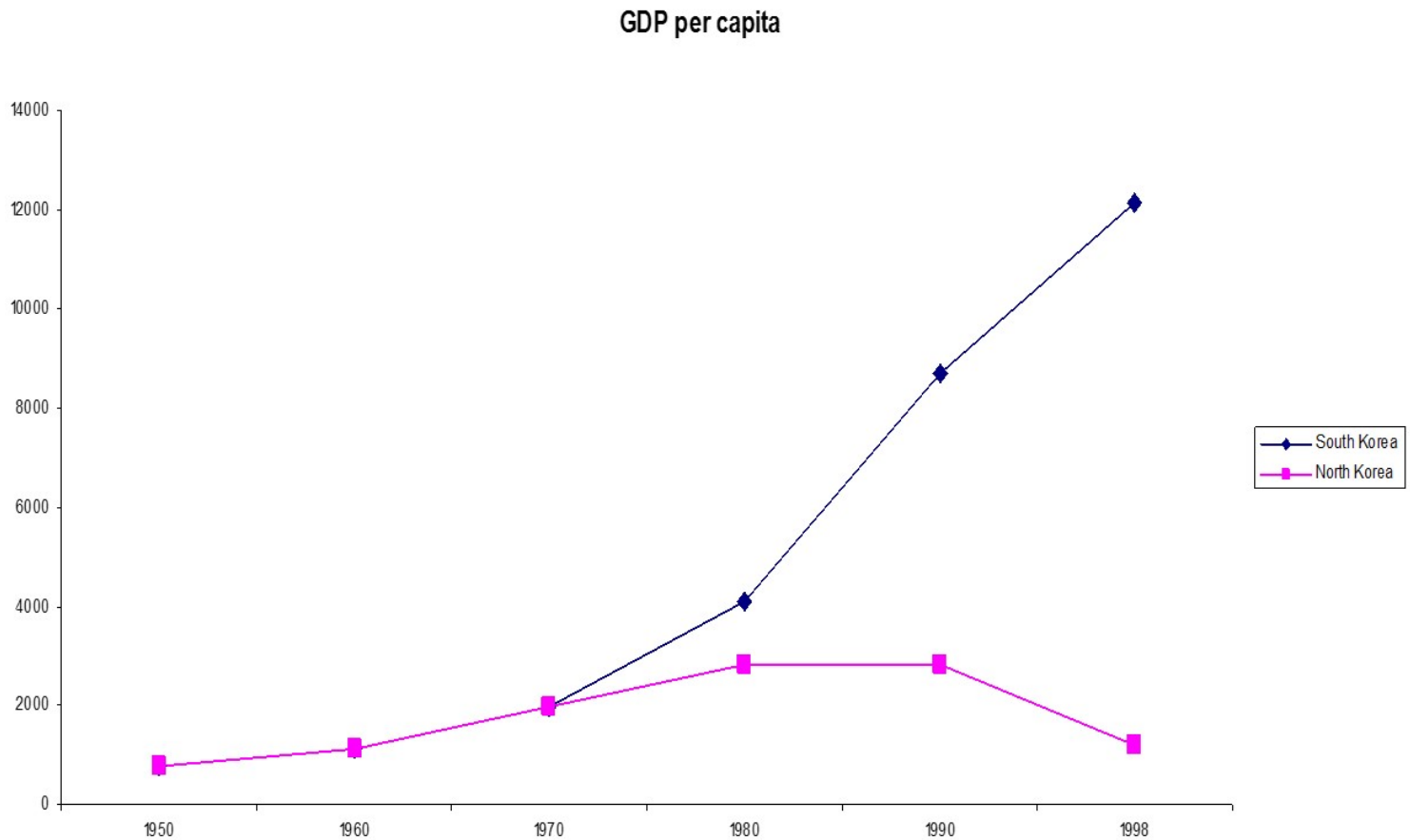
STATA Example: Acemoglu, Johnson, and Robinson (2001)

Motivation

- At some level it is obvious that institutions matter
- Witness, for example, the divergent paths of North and South Korea, or East and West German
    - central planning and collective ownership V.S. private property and market economy

STATA Example: Acemoglu, Johnson, and Robinson (2001)

Motivation



STATA Example: Acemoglu, Johnson, and Robinson (2001)

Motivation

- Nevertheless, we lack reliable estimates of the effect of institutions on economic performance
  - **Selection bias 1**: It is quite likely that rich economies choose or can afford better institutions
  - **Selection bias 2**: Economies that are different for a variety of reasons will differ both in their institutions and in their income per capita
- Need to eliminate selection bias

STATA Example: Acemoglu, Johnson, and Robinson (2001)

Identification Strategy

- They propose an IV to generate an exogenous variation in institution based on theory plus history
- They look only among former European colonies

STATA Example: Acemoglu, Johnson, and Robinson (2001)

Identification Strategy

- Their theory is based on the following facts:

1. In some colonies, Europeans had good survival, in others not
2. Where Europeans could survive, they put down roots, established good institutions
    - Replicate European institutions, with strong emphasis on private property and checks against government power

STATA Example: Acemoglu, Johnson, and Robinson (2001)

Identification Strategy

3. Where they were dying like flies, they set up extractive institutions
4. Extractive institutions designed to get resources quickly
    - Extractive institutions did not introduce much protection for private property, nor did they provide checks and balances against government expropriation

- These institutions persisted after decolonialization

STATA Example: Acemoglu, Johnson, and Robinson (2001)

Identification Strategy

- They use **mortality rates expected by the first European settlers** in the colonies as an IV for current institutions in these countries
- Malaria and yellow fever were the major sources of European mortality in the colonies
- Their theory is:

  Health environment $\Rightarrow$ Settler mortality $\Rightarrow$ European settlement $\Rightarrow$ Early institutions $\Rightarrow$ Current institutions $\Rightarrow$ Output today

STATA Example: Acemoglu, Johnson, and Robinson (2001)

Identification Strategy

- **Key assumption – exclusion restriction**: settler mortality can NOT affect output today by any other channel
- Possible threats to identification:
  - Health environment might affect output today directly
  - Having European settlers affects output through some channel other than institutions (language, human capital)

STATA Example: Acemoglu, Johnson, and Robinson (2001)

Identification Strategy

- Yellow fever and malaria had much less effect on native inhabitants, who had acquired and genetic immunity
  - The prevalence of these diseases depend on the microclimate of an area (e.g. temperature and humidity)
- This suggests that mortality rates faced by Europeans are unlikely to be a proxy for some simple geographic or climatic feature of the country

STATA Example: Acemoglu, Johnson, and Robinson (2001)

TSLS estimation

- They conduct a TSLS estimation
- **First stage**: current economic institutions = g(settler mortality)
- **Second stage**: log income per capita = f(current economic institutions)

STATA Example: Acemoglu, Johnson, and Robinson (2001)

Data

- Mortality rates of soldiers stationed in the colonies in the early 19th century
  - They got it from historian Philip Curtin
- **Current economic institutions** proxied by **protection against expropriation risk**
  - Average **protection against expropriation risk** is measured on a scale from 0 to 10
  - A higher score means more protection against expropriation, averaged over 1985 to 1995, from Political Risk Services

STATA Example: Acemoglu, Johnson, and Robinson (2001)

Reduced-form relationship

- Income and Settler Mortality

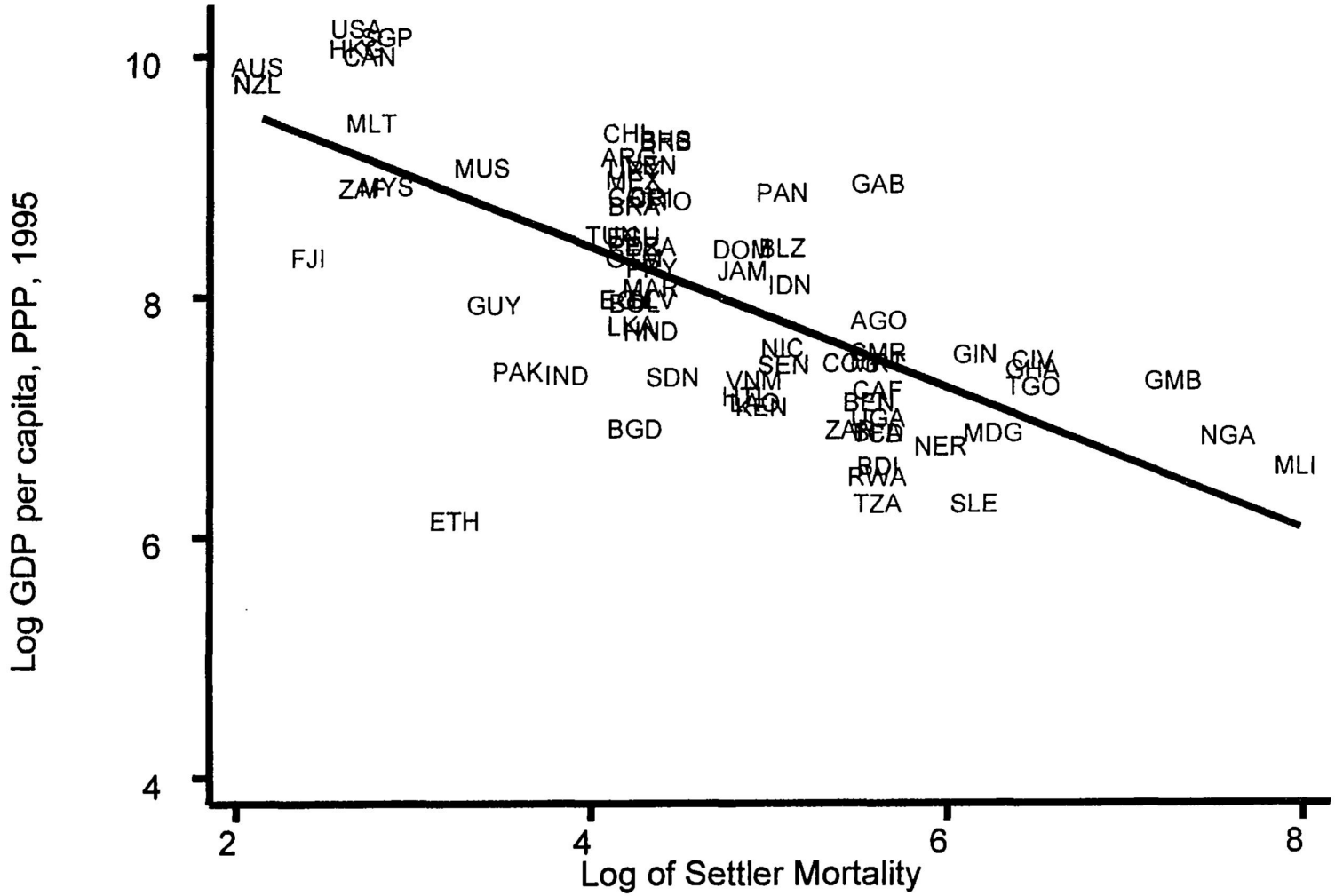


FIGURE 1. REDUCED-FORM RELATIONSHIP BETWEEN INCOME AND SETTLER MORTALITY

STATA Example: Acemoglu, Johnson, and Robinson (2001)

First-stage relationship

- Protection for Expropriation Risk and Settler Mortality

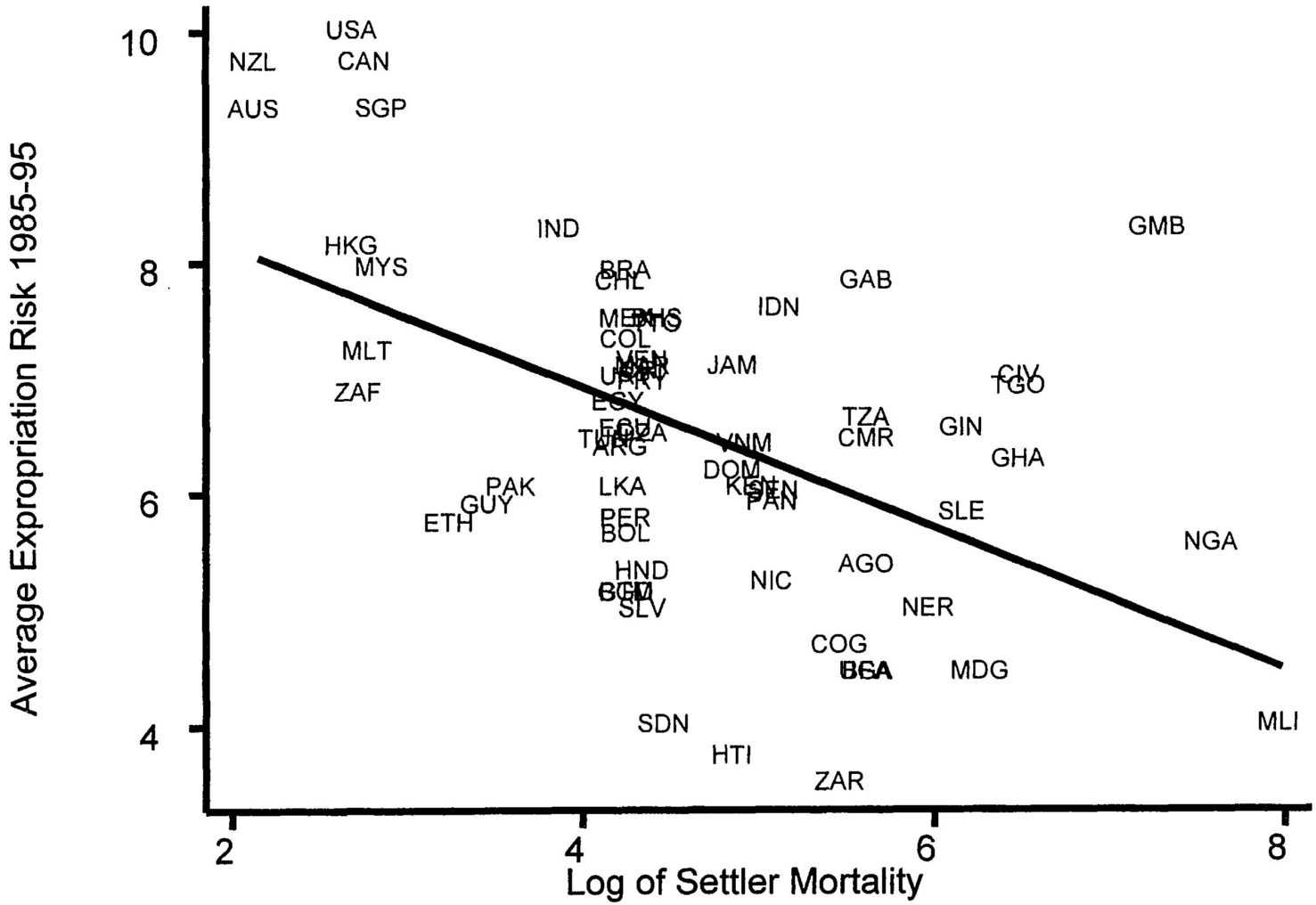


FIGURE 3. FIRST-STAGE RELATIONSHIP BETWEEN SETTLER MORTALITY AND EXPROPRIATION RISK

STATA Example: Acemoglu, Johnson, and Robinson (2001)

Second-stage relationship

- Protection for Expropriation Risk and Income

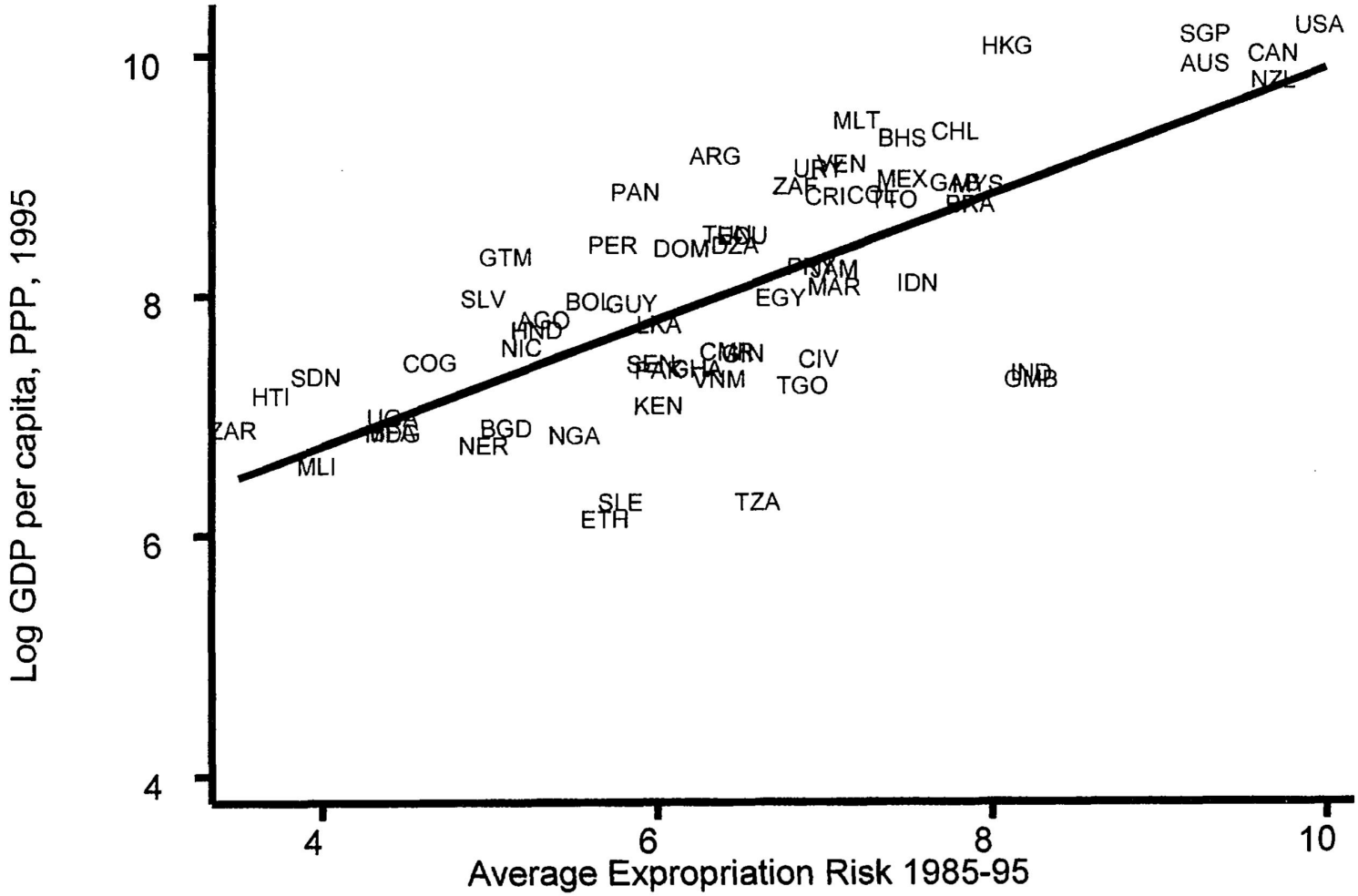


FIGURE 2. OLS RELATIONSHIP BETWEEN EXPROPRIATION RISK AND INCOME

STATA Example: Acemoglu, Johnson, and Robinson (2001)

Descriptive Statistics

TABLE 1—DESCRIPTIVE STATISTICS

			By quartiles of mortality			
	Whole world	Base sample	(1)	(2)	(3)	(4)
Log GDP per capita (PPP) in 1995	8.3	8.05	8.9	8.4	7.73	7.2
	(1.1)	(1.1)				
Log output per worker in 1988	−1.70	−1.93	−1.03	−1.46	−2.20	−3.03
(with level of United States	(1.1)	(1.0)				
normalized to 1)						
Average protection against	7	6.5	7.9	6.5	6	5.9
expropriation risk, 1985–1995	(1.8)	(1.5)				
Constraint on executive in 1990	3.6	4	5.3	5.1	3.3	2.3
	(2.3)	(2.3)				
Constraint on executive in 1900	1.9	2.3	3.7	3.4	1.1	1
	(1.8)	(2.1)				
Constraint on executive in first year	3.6	3.3	4.8	2.4	3.1	3.4
of independence	(2.4)	(2.4)				
Democracy in 1900	1.1	1.6	3.9	2.8	0.19	0
	(2.6)	(3.0)				
European settlements in 1900	0.31	0.16	0.32	0.26	0.08	0.005
	(0.4)	(0.3)				
Log European settler mortality	n.a.	4.7	3.0	4.3	4.9	6.3
		(1.1)				
Number of observations	163	64	14	18	17	15

Notes: Standard deviations are in parentheses. Mortality is potential settler mortality, measured in terms of deaths per annum per 1,000 "mean strength" (raw mortality numbers are adjusted to what they would be if a force of 1,000 living people were kept in place for a whole year, e.g., it is possible for this number to exceed 1,000 in episodes of extreme mortality as those who die are replaced with new arrivals). Sources and methods for mortality are described in Section III, subsection B, and in the unpublished Appendix (available from the authors; or see Acemoglu et al., 2000). Quartiles of mortality are for our base sample of 64 observations. These are: (1) less than 65.4; (2) greater than or equal to 65.4 and less than 78.1; (3) greater than or equal to 78.1 and less than 280; (4) greater than or equal to 280. The number of observations differs by variable; see Appendix Table A1 for details.

STATA Example: Acemoglu, Johnson, and Robinson (2001)

OLS Results

$$log(Y_i) = \mu + \alpha R_i + X_i'\gamma + \varepsilon_i$$

- Y_i is income per capita in country i
- R_i is the protection against expropriation measure
- X_i' is a vector of other covariates
- α represents the **effect of institutions on income per capita**

OLS Results

STATA Implementation

- See AJR-IV.do

```
. regress logpgp95 avexpr lat_abst africa asia other if baseco==1, robust

Linear regression                               Number of obs     =         64
                                                F(5, 58)          =      53.79
                                                Prob > F          =     0.0000
                                                R-squared         =     0.7139
                                                Root MSE          =    .58163
```

logpgp95	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
avexpr	.4012826	.0640653	6.26	0.000	.2730419	.5295233
lat_abst	.8752965	.6142898	1.42	0.160	-.3543381	2.104931
africa	-.8806805	.1555275	-5.66	0.000	-1.192003	-.5693585
asia	-.5767549	.2991853	-1.93	0.059	-1.175639	.0221296
other	.107205	.2229782	0.48	0.632	-.3391344	.5535445
_cons	5.736729	.3893229	14.74	0.000	4.957415	6.516044

```
.
```

OLS Results

TABLE 2—OLS REGRESSIONS

	Whole world (1)	Base sample (2)	Whole world (3)	Whole world (4)	Base sample (5)	Base sample (6)	Whole world (7)	Base sample (8)
	Dependent variable is log GDP per capita in 1995						Dependent variable is log output per worker in 1988	
Average protection against expropriation risk, 1985–1995	0.54 (0.04)	0.52 (0.06)	0.47 (0.06)	0.43 (0.05)	0.47 (0.06)	0.41 (0.06)	0.45 (0.04)	0.46 (0.06)
Latitude			0.89 (0.49)	0.37 (0.51)	1.60 (0.70)	0.92 (0.63)		
Asia dummy				−0.62 (0.19)		−0.60 (0.23)		
Africa dummy				−1.00 (0.15)		−0.90 (0.17)		
"Other" continent dummy				−0.25 (0.20)		−0.04 (0.32)		
R^2	0.62	0.54	0.63	0.73	0.56	0.69	0.55	0.49
Number of observations	110	64	110	110	64	64	108	61

Notes: Dependent variable: columns (1)–(6), log GDP per capita (PPP basis) in 1995, current prices (from the World Bank's World Development Indicators 1999); columns (7)–(8), log output per worker in 1988 from Hall and Jones (1999). Average protection against expropriation risk is measured on a scale from 0 to 10, where a higher score means more protection against expropriation, averaged over 1985 to 1995, from Political Risk Services. Standard errors are in parentheses. In regressions with continent dummies, the dummy for America is omitted. See Appendix Table A1 for more detailed variable definitions and sources. Of the countries in our base sample, Hall and Jones do not report output per worker in the Bahamas, Ethiopia, and Vietnam.

STATA Example: Acemoglu, Johnson, and Robinson (2001)

IV Results

- First stage

$$R_i = \mu + \alpha log(M_i) + X_i'\vartheta + \eta_i$$

- Second stage

$$log(Y_i) = \mu + \alpha R_i + X_i'\gamma + \varepsilon_i$$

- M_i is mortality rates faced by settler at country i

STATA Command: ivregress

- Syntax:

```
ivregress estimator depvar [varlist1] (varlist2 =
    varlistiv) [if] [in] [weight] [, options]
```

- Example:

```
ivregress 2sls logpgp95 lat_abst africa asia
   other_cont (avexpr=logem4), first
estat firststage
```

- **varlist1**is the list of exogenous variables.
- **varlist2** is the list of endogenous variables.
- **varlistiv** is the list of exogenous variables used with varlist1 as instruments for varlist2.
- **2sls**: two-stage least squares

STATA Command: ivregress

- options:
  - **estat firststage**: report first-stage F-statistic
  - **level()**: set confidence level; default is level(95)
  - **first**: requests that the first-stage regression results be displayed

IV Results

STATA Implementation

```
 ivregress 2sls logpgp95 (avexpr=logem4) f_brit f_french, first

First-stage regressions
-----------------------

                                                Number of obs     =         64
                                                F(   3,     60)   =       8.91
                                                Prob > F          =     0.0001
                                                R-squared         =     0.3081
                                                Adj R-squared     =     0.2736
                                                Root MSE          =     1.2518

------------------------------------------------------------------------------
      avexpr |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      f_brit |    .629348   .3664792     1.72   0.091    -.1037196    1.362416
    f_french |   .0474048   .4295458     0.11   0.912    -.8118147    .9066243
      logem4 |  -.5343989    .139576    -3.83   0.000    -.8135924   -.2552053
       _cons |   8.746647   .6904157    12.67   0.000      7.36561    10.12768
------------------------------------------------------------------------------


Instrumental variables (2SLS) regression        Number of obs     =         64
                                                Wald chi2(3)      =      32.21
                                                Prob > chi2       =     0.0000
                                                R-squared         =     0.0483
                                                Root MSE          =     1.0099

------------------------------------------------------------------------------
    logpgp95 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      avexpr |    1.07785    .210709     5.12   0.000     .6648682    1.490832
      f_brit |  -.7777037    .343026    -2.27   0.023    -1.450022   -.1053851
    f_french |  -.1169738   .3435071    -0.34   0.733    -.7902354    .5562878
       _cons |   1.372403    1.34394     1.02   0.307    -1.261672    4.006477
------------------------------------------------------------------------------
Instrumented:  avexpr
Instruments:   f_brit f_french logem4

. estat firststage

  First-stage regression summary statistics
  -----------------------------------------------------------------------------
               |             Adjusted      Partial
      Variable |   R-sq.       R-sq.       R-sq.       F(1,60)   Prob > F
  -------------+---------------------------------------------------------------
        avexpr |  0.3081      0.2736      0.1963       14.6592     0.0003
  -----------------------------------------------------------------------------
```

Robustness Checks

- Control for latitude
- Maybe settler mortality just means bad disease environment
  - Can't control **life expectancy** because clearly this is endogneous (affected by treatment): **bad control**
  - Include measures of temperature and humidity meant to capture disease environment
  - Also put in measures of soil quality
- IV result survives all of these robustness checks

Test Exclusion Restriction

- According to their theory, settler mortality (M) affected settlements (S), settlements affected early institutions (C) and early institutions affected current institutions (R)
- Test whether any of these variables, C, S, and M, has a direct effect on income per capita, log y,
- Using measures of C and S as additional instruments

Test Exclusion Restriction

- The TSLS estimates of the effect of protection against expropriation on GDP per capita using a variety of instruments other than mortality rates
- These estimates are always quite close to those using settler mortality as IV

Test Exclusion Restriction

- Panel D adds the log of mortality as an exogenous regressor of instruments other than mortality rates
- If mortality rates faced by settlers had a direct effect on income per capita, we would expect this variable to come in negative and significant

Test Exclusion Restriction

TABLE 8—OVERIDENTIFICATION TESTS

	Base sample (1)	Base sample (2)	Base sample (3)	Base sample (4)	Base sample (5)	Base sample (6)	Base sample (7)	Base sample (8)	Base sample (9)	Base sample (10)
Panel A: Two-Stage Least Squares										
Average protection against expropriation risk, 1985–1995	0.87	0.92	0.71	0.68	0.72	0.69	0.60	0.61	0.55	0.56
	(0.14)	(0.20)	(0.15)	(0.20)	(0.14)	(0.19)	(0.14)	(0.17)	(0.12)	(0.14)
Latitude		−0.47		−0.34		0.31		−0.41		−0.16
		(1.20)		(1.10)		(1.05)		(0.92)		(0.81)
Panel B: First Stage for Average Protection Against Expropriation Risk										
European settlements in 1900	3.20	2.90								
	(0.62)	(0.83)								
Constraint on executive in 1900			0.32	0.26						
			(0.08)	(0.09)						
Democracy in 1900					0.24	0.20				
					(0.06)	(0.07)				
Constraint on executive in first year of independence							0.25	0.22		
							(0.08)	(0.08)		
Democracy in first year of independence									0.19	0.17
									(0.05)	(0.05)
R^2	0.30	0.30	0.20	0.24	0.24	0.26	0.19	0.25	0.26	0.30
Panel C: Results from Overidentification Test										
p-value (from chi-squared test)	[0.67]	[0.96]	[0.09]	[0.20]	[0.11]	[0.28]	[0.67]	[0.79]	[0.22]	[0.26]
Panel D: Second Stage with Log Mortality as Exogenous Variable										
Average protection against expropriation risk, 1985–1995	0.81	0.88	0.45	0.42	0.52	0.48	0.49	0.49	0.4	0.41
	(0.23)	(0.30)	(0.25)	(0.30)	(0.23)	(0.28)	(0.23)	(0.25)	(0.18)	(0.19)
Log European settler mortality	−0.07	−0.05	−0.25	−0.26	−0.21	−0.22	−0.14	−0.14	−0.19	−0.19
	(0.17)	(0.18)	(0.16)	(0.17)	(0.15)	(0.16)	(0.16)	(0.15)	(0.13)	(0.12)
Latitude		−0.52		0.38		0.28		−0.38		−0.17
		(1.15)		(0.89)		(0.86)		(0.84)		(0.73)

Test Exclusion Restriction

- So these results also show NO evidence that:
  - Mortality rates faced by settlers have a direct effect on income per capita
  - Mortality rates have an effect working through a variable other than institutions on income per capita

Instrumental Variables Design: Practical Tips

Practical Tips For IV Design

Weak IV

1. Check IV relevance
   - Does this IV make sense?
   - Do the coefficients have the right magnitude and sign ?
   - Report the F-statistic in the first stage regression
     - If $F > 10$, instruments are strong - use TSLS
     - If $F < 10$, weak instruments - find better IV
   - If instruments are weak, then the TSLS estimator is biased and the t-statistic has a non-normal distribution

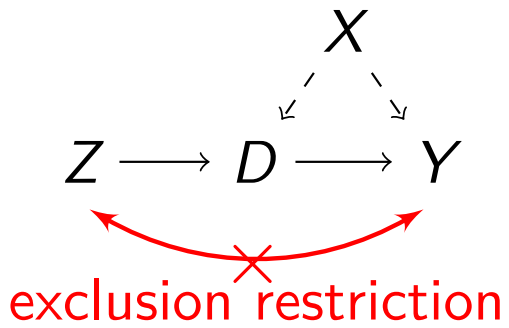
Practical Tips For IV Design

What are the consequences of weak instruments?

$$\hat{\alpha}_{TSLS} = \frac{Cov(Y_i, Z_i)}{Cov(D_i, Z_i)}$$

- If $Cov(D_i, Z_i)$ is small, the small changes in $Cov(Y_i, Z_i)$ can induce big changes in $\hat{\alpha}_{TSLS}$
- Suppose in one sample you calculate $Cov(D_i, Z_i) = 0.00001$!
- Therefore, from one sample to the next, $\hat{\alpha}_{TSLS}$ can change dramatically
- Under this situation, the normal distribution is a poor approximation to the sampling distribution of $\hat{\alpha}_{TSLS}$
- If instruments are weak, the usual methods of inference are unreliable - potentially very unreliable.

Practical Tips For IV Papers



2. Check exclusion restriction

- The exclusion restriction cannot be tested directly, but it can be falsified
- **Placebo test**
  - Test the reduced form effect of Z_i on Y_i in situations where it is impossible or extremely unlikely that Z_i could affect D_i
  - Because Z_i can't affect D_i, then the exclusion restriction implies that this placebo test should have zero effect.

Practical Tips For IV Papers

3. If you have many IVs pick your best instrument and report the just identified model
4. Look at the reduced form
   - Directly estimate the effect of instrument Z on outcome Y
   - If you can't see the causal relationship of interest in the reduced form it is probably not there

Suggested Readings

- Chapter 3, Mastering Metrics: The Path from Cause to Effect
- Chapter 4, Mostly Harmless Econometrics
- Chapter 7, Causal Inference: The Mixtape