Instrumental Variable

Eduard Bukin

Return to schooling and the Selection bias

- Does more years of schooling cause higher wages?
- What other methods can we use to compute the return to schooling?

Short Regression

$$Y_i = \alpha^S + \rho^S s_i + \beta^S X_i + \varepsilon_i^S \qquad (1.1)$$

- annual earning Y_i
- years of education s_i
- X_i vector of other control variables, such as experience.

Is the ceteris paribus fulfilled in Equation 1.1?

- Is control for experience and education sufficient?
- At a given experience/education level, are more- and less-educated workers equally able and diligent? (see Joshua D. Angrist & Pischke, 2014, Ch. 6)

Long Regression

$$Y_i = \alpha + \rho s_i + \beta X_i + \gamma A_i^{'} + \varepsilon_i \qquad (1.2)$$

- where $A_i^{'}$ is the ability variable that we desire to have in order to ensure the unbiased estimates of ρ.
- Omitting $A_i^{'}$ causes a selection bias or endogeneity:
  - $\rho^S = \rho + \underbrace{\delta_{A^{'}s} \times \gamma}_{\text{ability bias}}$

Endogeneity

Is another terminology for a selection bias!

Definition

- Consider following **LONG** and **SHORT** models:

$$Y_i = \alpha + \rho s_i + \beta X_i + \gamma A_i' + \varepsilon_i, \qquad \text{long}$$
$$Y_i = \alpha^S + \rho^S s_i + \beta^S X_i + \varepsilon_i^S, \qquad \text{short}$$

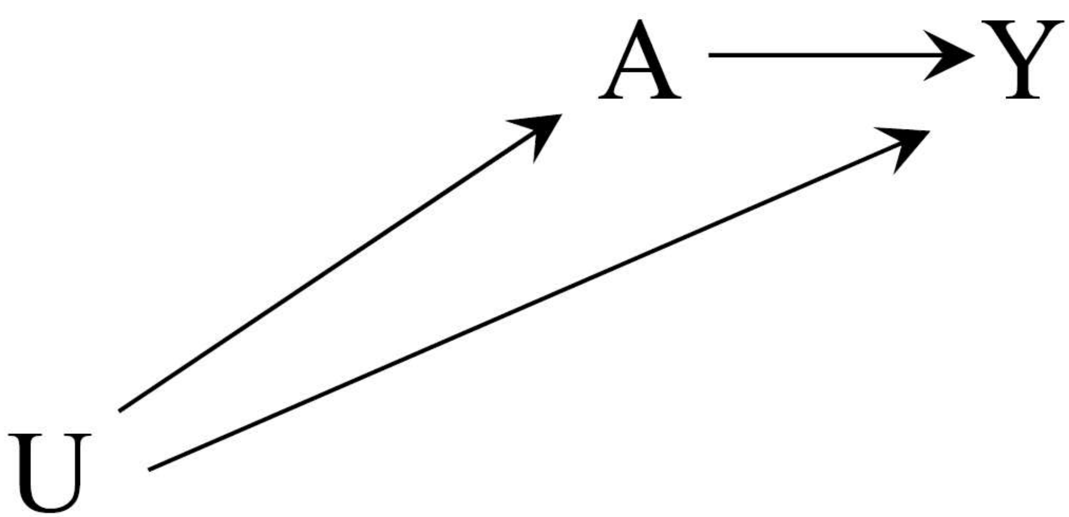
- s_i is the causal variable of interest (education)
- A_i' is the vector of control variables that we desire to have in order to ensure unbiased estimates of ρ;
- Variable s_i is **endogenous** if it correlates with the error terms ε_i^S :

$$Cov(s_i, \varepsilon_i^S) \neq 0$$

Endogeneity in practice:

- variation in the independent variable s_i (education) is not “random” as compared to the variation in the dependent variable Y_i, but
- an external process U affects variation in both s_i and Y_i;
- thus, s_i is endogenous to Y_i;

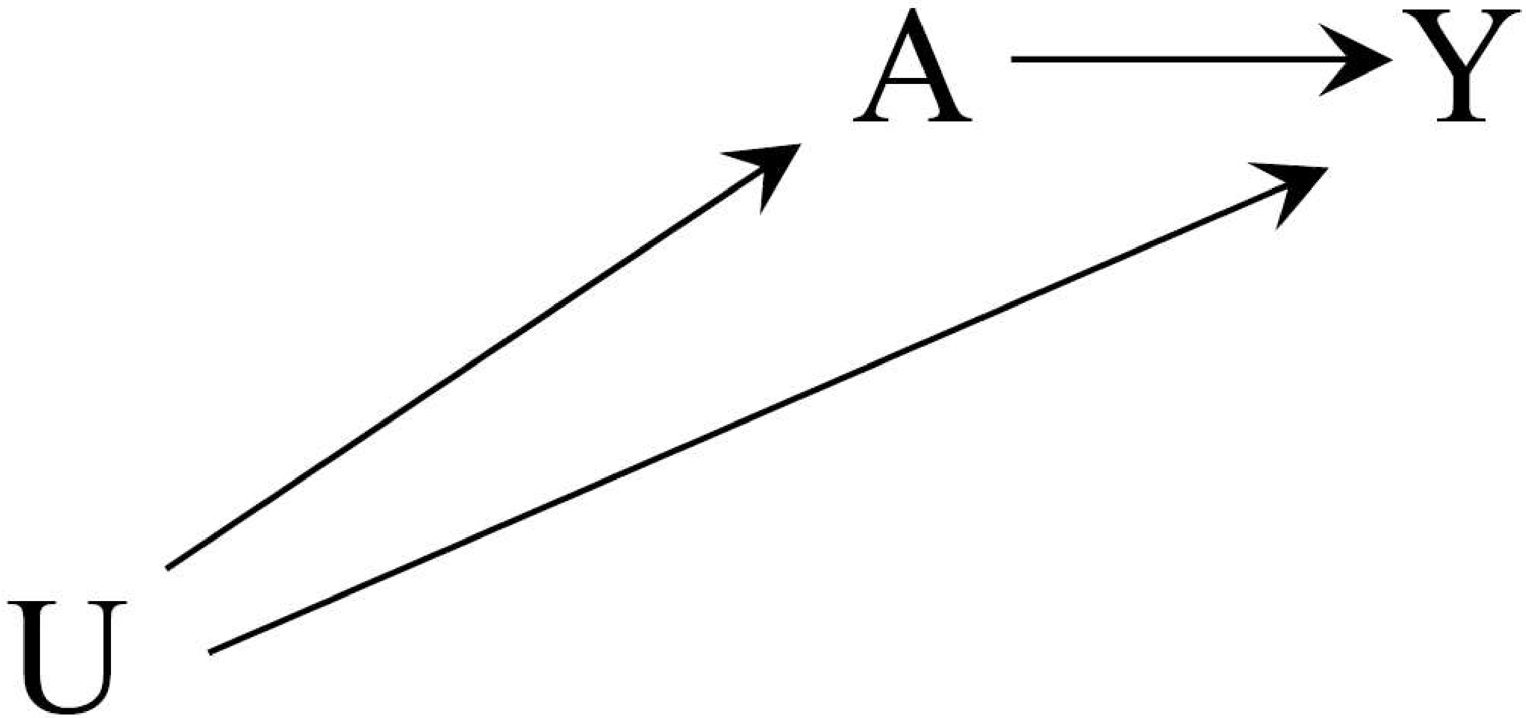
If variance of s_i is truly independent of Y_i, s_i is exogenous.



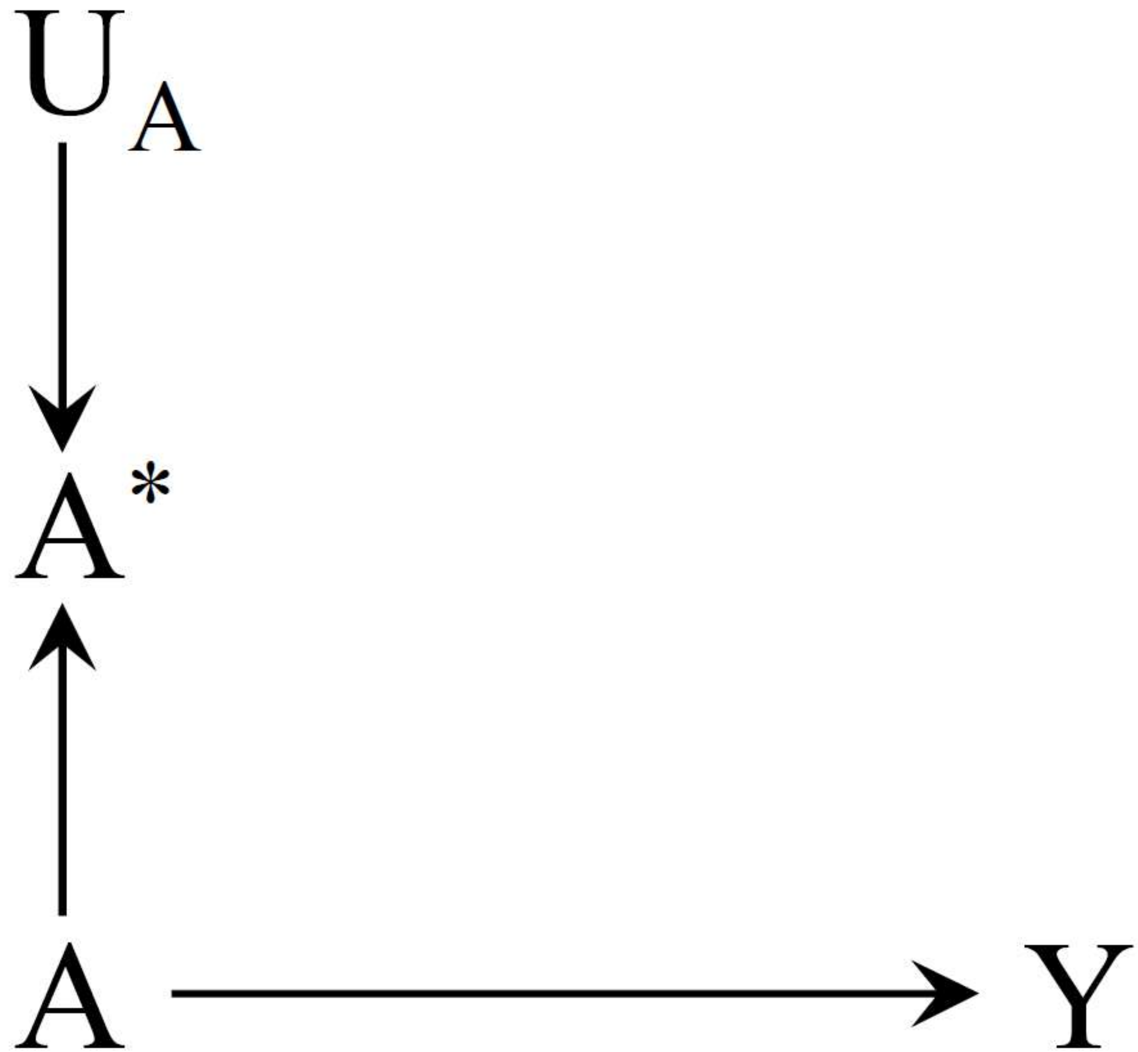
Causes of endogeneity

- Omitted Variable Bias (familiar)
- Measurement Error
- Simultaneity

OVB



Measurement error



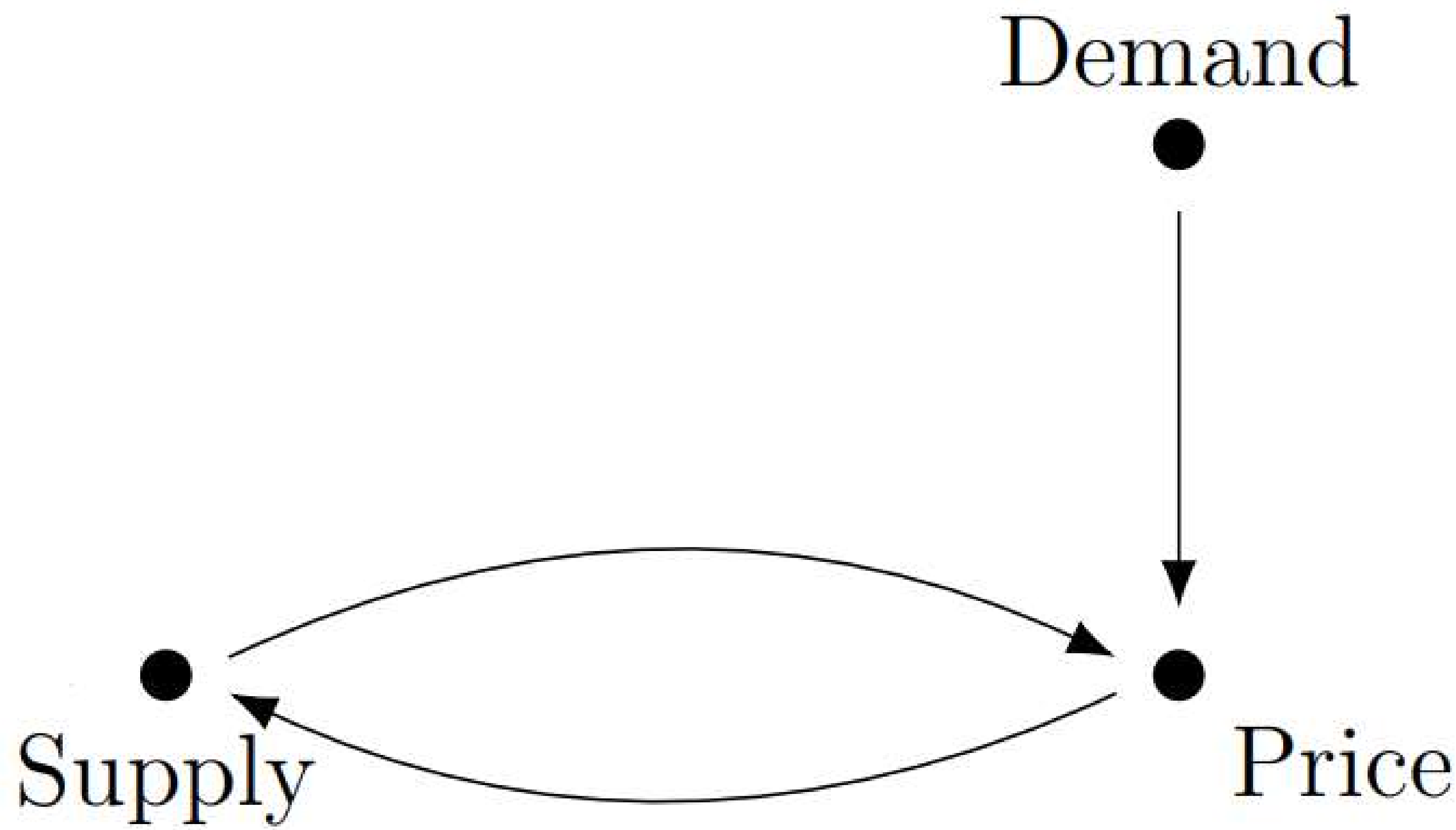
Measurement error

- We estimate a long model: $Y_i = \alpha + \beta s_i^* + e_i$,
  - but s_i^* is unavailable, we only have $s_i = s_i^* + m_i$ instead,
  - m_i is a systematic measurement error,
  - $E[m_i] = 0$ and $Cov(s_i^*, m_i) = Cov(e_i, m_i) = 0$.
- Desired coefficient $\beta = \frac{Cov(Y_i, s_i)}{Var(s_i)}$
- But with the erroneous data, we estimate biased coefficient β_b

$$\beta_b = \frac{Cov(Y_i, s_i)}{Var(s_i)} = \frac{Cov(a + \beta s_i^* + e_i, s_i^* + m_i)}{Var(s_i)}$$

$$= \frac{\beta \cdot Cov(s_i^*, s_i^*)}{Var(s_i)} = \beta \frac{Var(s_i^*)}{Var(s_i)}$$

Simultaneity



Simultaneity

- Simultaneity occurs if at least two variables are jointly determined.
  - A typical case is when observed outcomes are the result of separate behavioral mechanisms that are coordinated in an equilibrium.
- The prototypical case is a system of demand and supply equations:
  - $D(p)$ = how high would demand be if the price was set to p?
  - $S(p)$ = how high would supply be if the price was set to p?
- Number of police people and the crime rate.
- (see M. J. Wooldridge, 2020, Ch. 17) for more details on the problem and solutions.

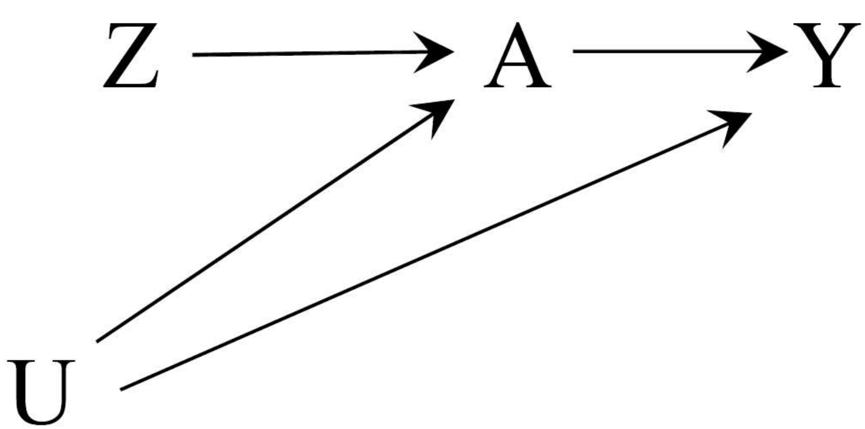
JUSTUS-LIEBIG-UNIVERSITÄT GIESSEN

IV - one of the solutions to endogeneity

IV stands for Instrumental Variable

Instrumental Variable

is another variable Z_i that affects only endogenous regressor s_i and satisfies:



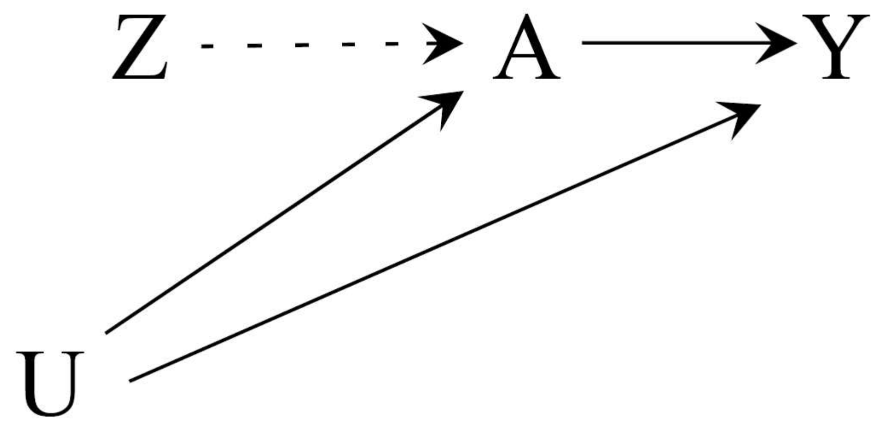
1. **Relevance condition:**
2. **Exclusion restriction:**
3. **Independence assumption:**

(see Joshua D. Angrist & Pischke, 2014 Ch. 3 and 6; Joshua D. Angrist & Pischke, 2009, Ch. 4.; Hernán & Robins, 2020, Ch. 16; J. M. Wooldridge, 2010, Ch. 8; Söderbom, Teal, & Eberhardt, 2014, Ch. 11; Imbens, 2020)

1. Relevance condition:

- Z_i has a causal effect on s_i;

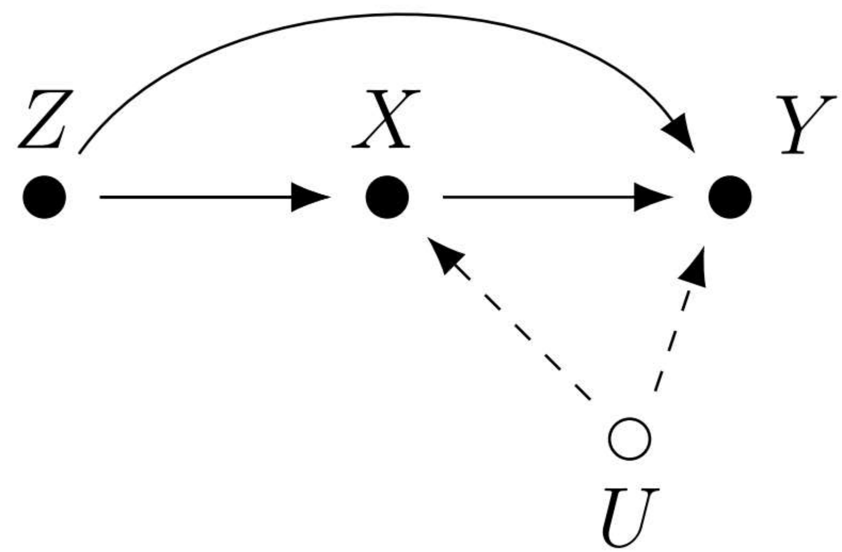
Violation of the relevance condition:



2. Exclusion restriction:

- Z_i does not affect Y_i directly, except through its potential effect on s_i;

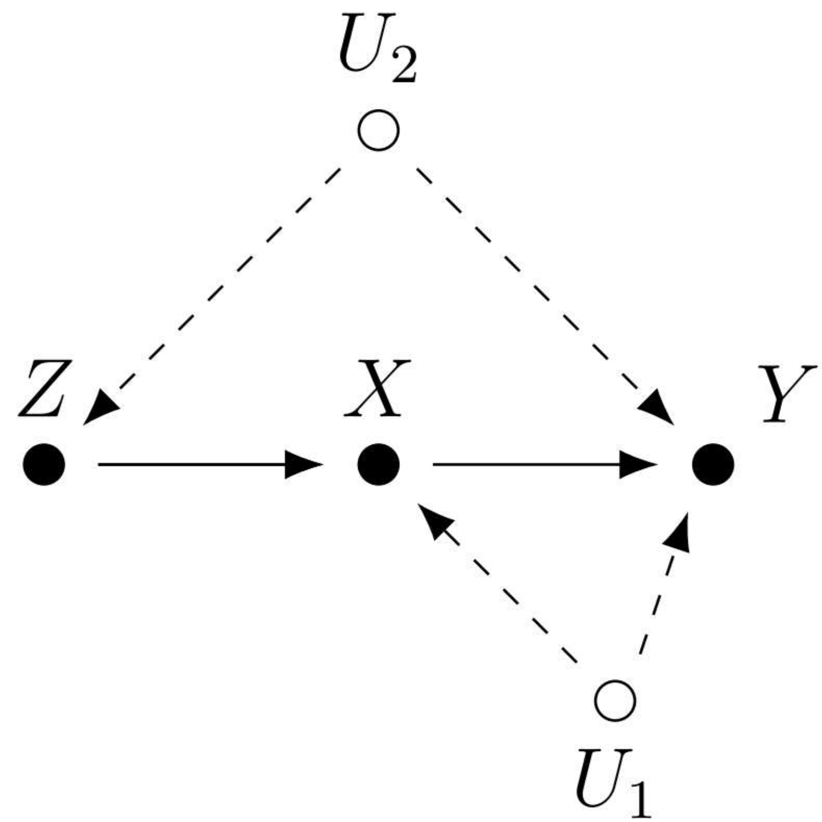
Violation of the exclusion restriction:



3. Independence assumption:

- Z_i is randomly assigned or "as good as randomly assigned", the same as
- Z_i is unrelated to the omitted variables $A_i^{'}$, same as
- Z_i and Y_i do not share any common causes

Violation of the independence assumption:



IV regression algorithm using 2SLS (1)

Stage 1: regress endogenous variable s_i on all X_i plus the instrument Z_i

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

Compute fitted values form the stage 1: $\hat{s}_i = \pi_0 + \pi_1 Z_i + \rho X_i$.

Substitute s_i with the $\hat{s}_i$ from the stage 1.

Stage 2: $Y_i = \alpha^{IV} + \rho^{IV} \hat{s}_i + \beta^{IV} X_i + \varepsilon_i^{IV}$

where

- $\hat{s}_i$ are the fitted values from the first stage
- ρ^{IV} is the causal effect of interest from stage two that is asymptotically equal to ρ , the true effect of interest ($\rho^{IV} \asymp \rho$)

Wage and Education (again)

Wage and Education (again)

$$Y_i = \alpha^S + \rho^S s_i + \beta^S X_i + \varepsilon_i^S$$

- We know that estimate of years of education s_i is biased because of the OVB (ability bias).
- **Think of an RCT experiment that could help to estimate true causal effect of** s_i on income!
- What instrument Z_i can we use for education?

Fantastic IVs and how to find them…

1. Use theory!
    - human capital theory suggests that people make schooling choices by comparing the costs and benefits of alternatives.
2. Think and speculate:
    - What is the ideal experiment that could capture the effect of schooling on education?
    - What are the forces you'd like to manipulate and the factors you'd like to hold constant?
    - What are the other processes that are independent of wage, but may affect schooling?
3. Analyze, what were/are the policies/environments that could mimic the experimental setting?

> Reasoning on how researcher use theory and available observational data to approximate real experiment is called **Identification strategy**!

Fantastic IVs for education

- Loan policies or other subsidies that vary independently of ability or earnings potential
- Region and time variation in school construction (Duflo, 2001)
- Proximity to college(Card, 1994)
- Quarter of birth (Joshua D. Angrist & Krueger, 1991)
- Parents education (Buckles & Hungerman, 2013)
- Number of siblings

Using parents education as the IV for education

```r
library(tidyverse)
library(haven)
library(modelsummary)
dta <-
  read_csv("education_parents.csv") %>%
  mutate(lwagehour = log(wage/hours)) %>%
  mutate(parents_edu = feduc + meduc)
glimpse(dta)
```

```
Rows: 722
Columns: 19
$ wage       <dbl> 769, 808, 825, 650, 562, 600, 1154, 1000, 930, 900, 1318, …
$ hours      <dbl> 40, 50, 40, 40, 40, 40, 45, 40, 43, 45, 38, 40, 50, 45, 40…
$ IQ         <dbl> 93, 119, 108, 96, 74, 91, 111, 95, 132, 125, 119, 118, 105…
$ KWW        <dbl> 35, 41, 46, 32, 27, 24, 37, 44, 44, 40, 24, 47, 37, 39, 36…
$ educ       <dbl> 12, 18, 14, 12, 11, 10, 15, 12, 18, 15, 16, 16, 10, 15, 11…
$ exper      <dbl> 11, 11, 11, 13, 14, 13, 13, 16, 8, 4, 7, 9, 17, 6, 19, 10,…
$ tenure     <dbl> 2, 16, 9, 7, 5, 0, 1, 16, 13, 3, 2, 9, 2, 9, 10, 4, 3, 8, …
$ age        <dbl> 31, 37, 33, 32, 34, 30, 36, 36, 38, 30, 28, 34, 35, 36, 38…
$ married    <dbl> 1, 1, 1, 1, 1, 0, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1…
$ black      <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0…
$ south      <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0…
$ urban      <dbl> 1, 1, 1, 1, 1, 1, 0, 1, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1…
$ sibs       <dbl> 1, 1, 1, 4, 10, 1, 2, 1, 1, 2, 3, 1, 1, 3, 2, 2, 5, 2, 0, …
```

Estimating IV manually

```r
# No IV
ols <-  lm(log(wage) ~ educ + exper + I(exper^2), data = dta)

# No IV but with controls for IQ
ols_iq <-  lm(log(wage) ~ educ + exper + I(exper^2) + IQ, data = dta)

# First stage
first_stage <- lm(educ ~ parents_edu + exper + I(exper^2), data = dta)

# Fitted values of endogenous regressor
dta_fitted <- dta %>% mutate(educ_fit = fitted(first_stage))

# Second stage
second_stage <- lm(log(wage) ~ educ_fit + exper + I(exper^2), data = dta_fitted)
```

	OLS	OLS (with ability proxi)	1 stage (par. educ.)	2 stage (par. educ.)
Education	0.078*** (0.007)	0.058*** (0.008)		0.146*** (0.019)
Parents educ.			0.148*** (0.013)	
Experience	0.009 (0.015)	0.010 (0.015)	−0.021 (0.072)	0.008 (0.016)
Experience sq.	0.001 (0.001)	0.001 (0.001)	−0.007* (0.003)	0.001+ (0.001)
Ability proxi		0.006*** (0.001)		
Num.Obs.	722	722	722	722
R2	0.141	0.169	0.326	0.076
R2 Adj.	0.137	0.164	0.323	0.072
Log.Lik.	−341.745	−329.737	−1462.726	−368.199
F	39.259	36.461	115.813	19.574

Using siblings number as the IV for education

```r
library(ivreg)
iv_fit2 <- ivreg(
  log(wage) ~ educ + exper + I(exper ^ 2) | sibs + exper + I(exper ^ 2) ,
  data = dta )
```

	OLS	OLS (with ability proxi)	1 stage (par. educ.)	2 stage (par. educ.)	2 stage (siblings)
Education	0.078*** (0.007)	0.058*** (0.008)		0.146*** (0.019)	0.128*** (0.033)
Parents educ.			0.148*** (0.013)		
Experience	0.009 (0.015)	0.010 (0.015)	−0.021 (0.072)	0.008 (0.016)	0.008 (0.016)
Experience sq.	0.001 (0.001)	0.001 (0.001)	−0.007* (0.003)	0.001+ (0.001)	0.001 (0.001)
Ability proxi		0.006*** (0.001)			
Num.Obs.	722	722	722	722	722
R2	0.141	0.169	0.326	0.076	0.085
R2 Adj.	0.137	0.164	0.323	0.072	0.081
Log.Lik.	−341.745	−329.737	−1462.726	−368.199	
F	39.259	36.461	115.813	19.574	

Pitfalls of the IV

Consistency and unbiasedness

- IV estimates **are not unbiased**, but they **are consistent** (Joshua D. Angrist & Krueger, 2001).
  - **Unbiasedness** means the estimator has a sampling distribution centered on the parameter of interest in a sample of any size, while
  - **Consistency** only means that the estimator converges to the population parameter as the sample size grows.

> ⓘ **Note**
>
> Researchers that use IV should aspire to work with **large samples**.

- No statistical tests is available for checking consistency

Bad instruments (1)

1. Z_i that does not satisfy any of the Relevance condition, Exclusion restriction and Independence assumption;

Bad instruments (2)

2. Z_i that correlate with omitted variable (OV) but **do not cause** changes in it or inflict simultaneity:

- They result into much greater upwards shifting bias compare to the OLS;
- For example the weather in Brazil and supply price and demand quantity of coffee:
  - weather shifts the supply curve, it is random, thus it seems as a plausible instrument for price in the demand model
  - the weather in Brazil determines supply expectations on futures exchange, thus, it also shifts the demand for coffee before the supply price is affected;

Bad instruments (3)

3. Weak instrument Z_i:

- When the instrument Z_i is only weakly correlates with endogenous regressor s_i;
- Find a better one!

Weak instrument test:

- Run the first stage regression with and without the IV;
- Compare the F-statistics
  - If F-statistics with instrument is greater than that without **by 5 of more**,
  - this is a sign of a strong instrument (Staiger & Stock, 1997);
- This test does not ensure that our instruments are independent of omitted variable A_i' or Y_i;
- Staiger & Stock (1997)

Overidentification (1)

- number of instruments G exceeds the number of endogenous variables K.
  - when the IV is overidentified, estimates are biased;
  - bias is proportional to $K - G$;
  - using fewer instruments therefore reduces bias;
- If you have few candidates for IV and one endogenous regressor:
  - select one IV for the first stage, and
  - put the remaining instruments as controls into the second stage

Overidentification (2)

Sargan's overidentification test:

- $H_0 : Cov(Z_i^{'}, \varepsilon_i^{IV}) = 0$ - the covariance between the instrument and the error term is zero
- $H_1 : Cov(Z_i^{'}, \varepsilon_i^{IV}) \neq 0$
- Thus, by rejecting the H_0, we conclude that at least one of the instruments is not valid.

Wu-Hausman test for endogeneity

Wu-Hausman test for endogeneity tests if the variable that we are worried about is indeed endogenous.

- $H_0 : Cov(s_i, \varepsilon_i) = 0$ - the covariance between potentially endogenous variable and the error term is zero
- $H_1 : Cov(s_i, \varepsilon_i) \neq 0$
- Thus, by rejecting the H_0, we conclude that there is endogeneity and there might be a need for IV.

Example 1. The colonial origins of comparative development: An empirical investigation

(Acemoglu, Johnson, & Robinson, 2001). The colonial origins of comparative development: An empirical investigation. American economic review, 91(5), 1369-1401.

Research question and the problem

- **What are the fundamental causes of the large differences in income per capita across countries?**
- with better “institutions,” more secure property rights, and less distortionary policies,
  - countries invest more in physical and human capital, and
  - use these factors more efficiently to
  - achieve a greater level of income.
- Institutions are a likely cause of income growth.

Endogeneity problem

What could be the ideal experiment to find the effect of institutions on income?

- Rich economies choose or can afford better institutions.
- Economies that are different for a variety of reasons
  - will differ both in their institutions and in their income per capita.
- To estimate the impact of institutions on income,
  - we need a **source of exogenous variation in institutions.**

Identification strategy

is the manner in which a researcher uses observational data (i.e., data not generated by a randomized trial) to approximate a real experiment (Joshua D. Angrist & Krueger, 1991)

1. Current performance is caused by
2. Current **institutions**, which are caused by
3. **Early institutions**, which are caused by
4. **Settlements types** during colonization, which are caused by
5. Settlers' (potential) **mortality or colonization risks**.

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

Empirical model (OLS estimator)

$$
\begin{aligned}
\log(\text{GDP per capita}_i) = \beta_0 & \\
& + \beta_1 \text{Proxy for institutions} \\
& + \gamma \text{Control variables} + \epsilon_i
\end{aligned}
$$

- i is the country;
- Dependent variable is the GDP per capita in 1995;
- As the proxy of the institutional quality, authors used **average protection against expropriation risk in 1985-1990** (index/country ranking);
- Controls include latitude of the country and continent-specific dummy variables;

OLS estimation

	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	0.52	0.47	0.43	0.47	0.41	0.45	0.46
	(0.04)	(0.06)	(0.06)	(0.05)	(0.06)	(0.06)	(0.04)	(0.06)
Latitude			0.89	0.37	1.60	0.92		
			(0.49)	(0.51)	(0.70)	(0.63)		
Asia dummy				−0.62		−0.60		
				(0.19)		(0.23)		
Africa dummy				−1.00		−0.90		
				(0.15)		(0.17)		
"Other" continent dummy				−0.25		−0.04		
				(0.20)		(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.

Empirical model (IV estimator)

First stage:

$$
\begin{aligned}
\text{Proxy for institutions} = \beta_0 & \\
& + \beta_1 \log(\text{Settlers mortality in 16-18th cent.}) \\
& + \gamma \text{Control variables} + e_i,
\end{aligned}
$$

- European settlers mortality in the 16-18th centuries is the precise number of how many settlers died in the country that they tried to colonize.

Second stage:

$$
\begin{aligned}
\log(\text{GDP per capita}_i) = \beta_0^{IV} & \\
& + \beta_1^{IV} \widehat{\text{Proxy for institutions}} \\
& + \gamma^{IV} \text{Control variables} + \epsilon_i^{IV},
\end{aligned}
$$

IV results

	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

References

Acemoglu, D., Johnson, S., & Robinson, J. A. (2001). The colonial origins of comparative development: An empirical investigation. *American Economic Review*, *91*(5), 1369–1401. http://doi.org/10.1257/aer.91.5.1369

Angrist, Joshua D., & Krueger, A. B. (1991). Does Compulsory School Attendance Affect Schooling and Earnings? *The Quarterly Journal of Economics*, *106*(4), 979–1014. http://doi.org/10.2307/2937954

Angrist, Joshua D., & Krueger, A. B. (2001). Instrumental variables and the search for identification: From supply and demand to natural experiments. *Journal of Economic Perspectives*, *15*(4), 69–85. http://doi.org/10.1257/jep.15.4.69

Angrist, Joshua D., & Pischke, J.-S. (2009). *Mostly harmless econometrics*. Princeton University Press. http://doi.org/10.1515/9781400829828

Angrist, Joshua D., & Pischke, J.-S. (2014). *Mastering'metrics: The path from cause to effect*. Princeton University Press.

Buckles, K. S., & Hungerman, D. M. (2013). Season of birth and later outcomes: Old questions, new answers. *Review of Economics and Statistics*, *95*(3), 711–724. http://doi.org/10.1162/rest_a_00314

Card, D. (1994). *Earnings, schooling, and ability revisited*. National Bureau of Economic Research.

Duflo, E. (2001). Schooling and labor market consequences of school construction in indonesia: Evidence from an unusual policy experiment. *American Economic Review*, *91*(4), 795–813. http://doi.org/10.1257/aer.91.4.795

Hernán, M. A., & Robins, J. M. (2020). *Causal Inference: What If*. Chapman & Hall/CRC.

Imbens, G. W. (2020). Potential outcome and directed acyclic graph approaches to causality: Relevance for empirical practice in economics. *Journal of Economic Literature*, *58*(4), 1129–1179. http://doi.org/10.1257/jel.20191597

Söderbom, M., Teal, F., & Eberhardt, M. (2014). *Empirical development economics*. ROUTLEDGE. Retrieved from https://www.ebook.de/de/product/21466458/mans_soederbom_francis_teal_markus_eberhardt_empirical

Staiger, D., & Stock, J. H. (1997). Instrumental variables regression with weak instruments. *Econometrica*, *65*(3), 557. http://doi.org/10.2307/2171753

Wooldridge, J. M. (2010). *Econometric analysis of cross section and panel data*. MIT press.

Wooldridge, M. J. (2020). *Introductory econometrics: A modern approach*. South-Western. Retrieved from https://www.cengage.uk/shop/isbn/9781337558860