UNIVERSITY OF OSLO

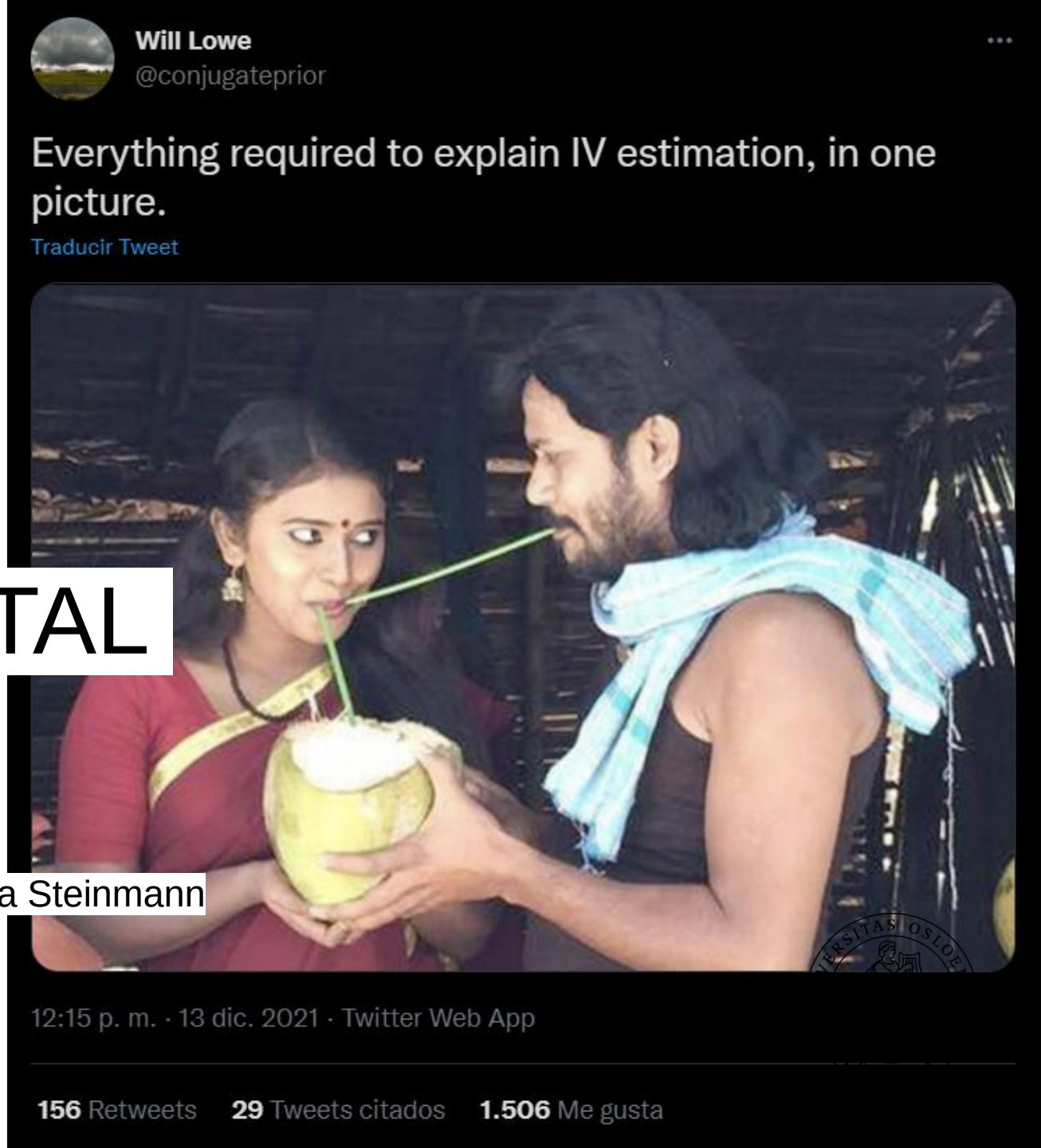
INSTRUMENTAL VARIABLES

José Manuel Arencibia Alemán & Isa Steinmann

Session 7

UiO : Centre for Educational Measurement
Det utdanningsvitenskapelige fakultet

February 7, 2022



5	Regression Models – Time: 31 January 2022, 12:15 – 14:00h – Main Instructor: Isa Steinmann – Required Reading: Angrist & Pischke (2015), Chapter 2
6	Further Control Strategies – Time: 03 February 2022, 12:15 – 14:00h – Main Instructor: Isa Steinmann – Required Reading: -
7	Instrumental Variable Approaches – Time: 07 February 2022, 12:15 – 14:00h – Main Instructor: José Manuel Arencibia Alemán – Required Reading: Angrist & Pischke (2015), Chapter 3
8	Regression Discontinuity Designs I – Time: 10 February 2022, 12:15 – 14:00h – Main Instructor: José Manuel Arencibia Alemán – Required Reading: Angrist & Pischke (2015), Chapter 4

Last session’s take-away messages

- Matching methods can be useful complementation of regression methods but they do not solve omitted variable bias
- Pretest variables are usually very valuable control variables
- Fixed-effects can be valuable additions to control strategies

Session’s learning goal

- Understand the intuition behind instrumental variables approach, its potential and its limitations

Content

1. Introduction
   - ✓ Instrument Variables vs. Control Strategies
2. The Charter Conundrum
   - ✓ IV requirements
   - ✓ LATE (vs. TOT)
3. Family Size and Parental Investments in Children
   - ✓ 2SLS
4. R Example

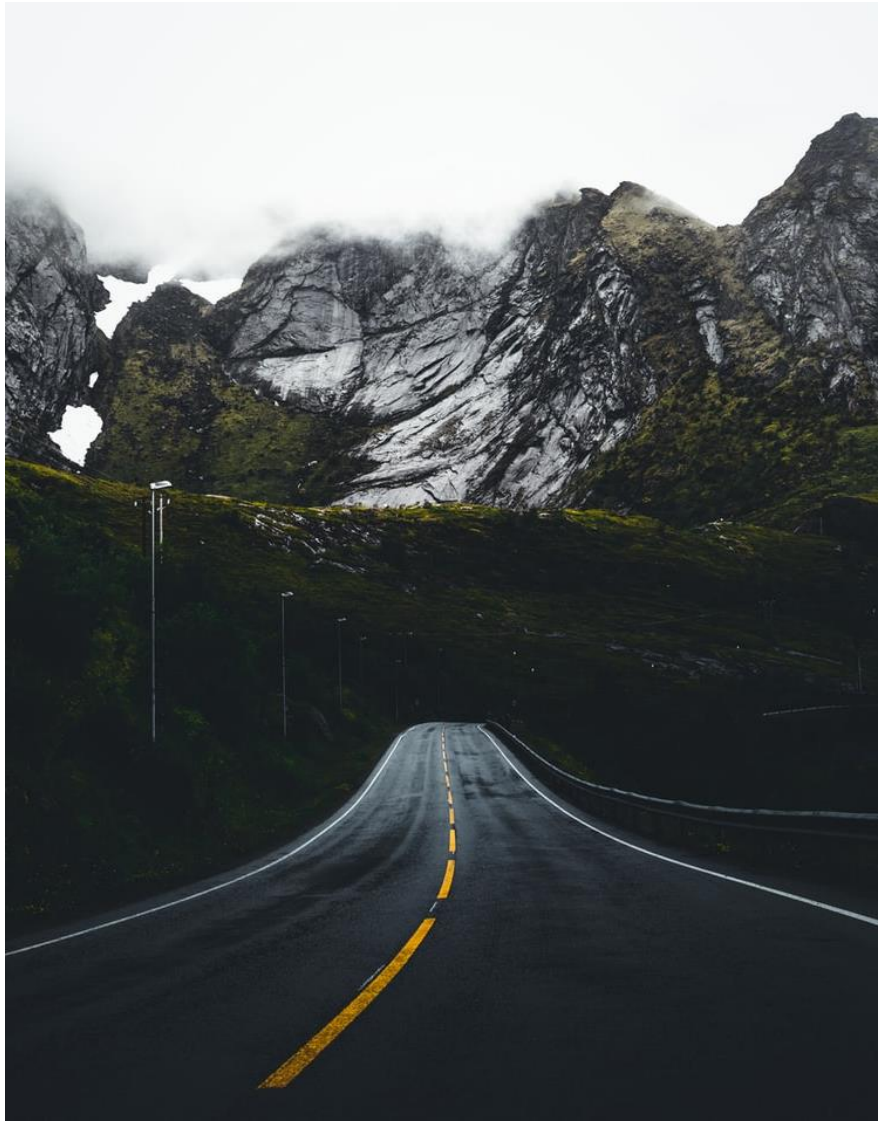


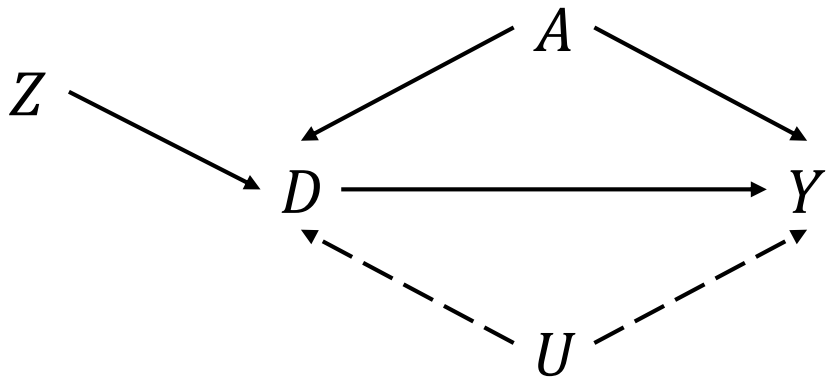
Photo by Taneli Lahtinen on Unsplash

Instrumental Variables vs. Control Strategies

- (Very) simply put, within the framework of causal inference, control strategies aim at…(?)
  - …eliminating the confounding effect of certain variables on the estimation of causal effects.

$$Y_i = \alpha + \beta D_i + \gamma A_i + \varepsilon_i$$

- Instrumental Variables (IV) approach aims at using any source of variation, Z, that would allow you to isolate the causal path of interest. In other words, attempts to separate the causal effect from that of confounding variables.



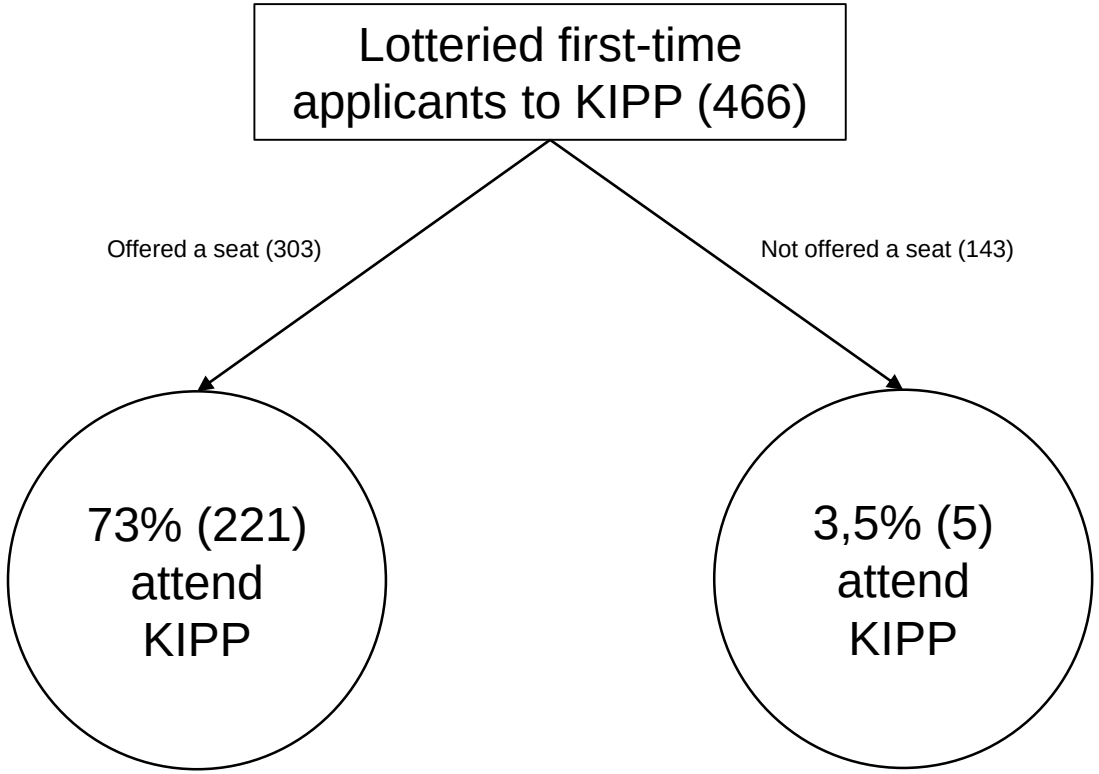
"The instrumental variables (IV) method harnesses partial or incomplete random assignment, whether naturally occurring or generated by researchers." (Angrist and Pischke 2015, p. 98)

The Charter Conundrum

Setting the Stage

- Knowledge Is Power Program (KIPP):
  - 140 charter schools affiliated
  - Emphasis on discipline, long-school days, focus on math and reading skills
  - Higher average scores than nearby schools
- Teachers: recent graduates of America's most selective colleges and universities
- Students and context:
  - Low-performing school districts. 95% ethnic minorities.
  - >80% poor enough to qualify for the federal government's subsidized lunch program

Application and enrollment data from KIPP Lynn lotteries
(Angrist and Pischke, 2015, p. 103)



Angrist and Pischke (2015)

The Charter Conundrum

Playing the Lottery

- Differences in Pre-treatment outcomes are insignificant
- 78,7% of lottery winners attended KIPP for 4,6% of losers ($= .787 - .741$)
- Lottery winners had an average math score (state standardized) close to 0. Markedly superior to Lynn public 5th graders.
- Since lottery offers are randomly assigned, the difference in math scores between winners and losers is the average causal effect of winning lottery…

But what does this say about the effect of KIPP attendance?

Analysis of KIPP lotteries
(Angrist and Pischke, 2015, p. 104)

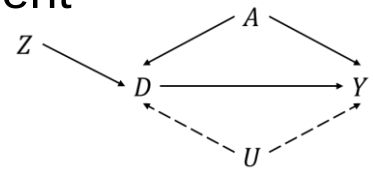
		KIPP applicants			
	Lynn public 5th graders	Lottery winners	Winners vs. losers	Attended KIPP	Attended vs. others
Panel A. (Selected) Baseline Characteristics					
Math score (4th grade)	-.307	-.290	.102 (.120)	-.289	0.69 (.109)
Verbal score (4th grade)	-.356	-.386	.0.63 (.125)	-.368	.088 (.114)
Panel B. Outcomes					
Attended KIPP	.000	.787	.741 (.037)	1	1
Math score	-.363	-.003	.355 (.115)	.095	.467 (.103)
Verbal score	-.417	-.262	.113 (.122)	-.211	.211
Sample size	3,964	253	371	204	371

Angrist and Pischke (2015)

The Charter Conundrum

Instrumental Variables

- The IV approach, here, aims at exploit the partial random assignment to treatment generated by the lottery system.



- Requirements
  1. First stage or relevance: the instrument, Z, has a causal effect on the instrumentalized variable, D.
  2. Independence: the instrument, Z, is as good as randomly assigned (it is independent from (at least) any, unobserved variable, U).
  3. Exclusion: any effect the instrument, Z, may have on the outcome, Y, is through the instrumentalized variable, D.

Analysis of KIPP lotteries
(Angrist and Pischke, 2015, p. 104)

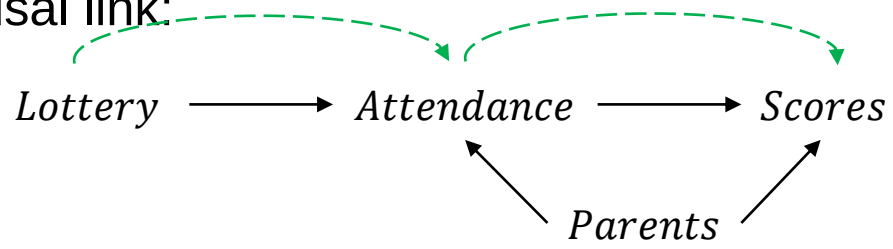
		KIPP applicants			
	Lynn public 5th graders	Lottery winners	Winners vs. losers	Attended KIPP	Attended vs. others
Panel A. (Selected) Baseline Characteristics					
Math score (4th grade)	-.307	-.290	.102 (.120)	-.289	0.69 (.109)
Verbal score (4th grade)	-.356	-.386	.0.63 (.125)	-.368	.088 (.114)
Panel B. Outcomes					
Attended KIPP	.000	.787	.741 (.037)	1	1
Math score	-.363	-.003	.355 (.115)	.095	.467 (.103)
Verbal score	-.417	-.262	.113 (.122)	-.211	.211
Sample size	3,964	253	371	204	371

Angrist and Pischke (2015)

The Charter Conundrum

Instrumental Variables

- Causal link:



- If lottery only affect scores through attendance, and you knew ϕ and λ, how would you calculate ρ?
  - ρ: effect of lottery on scores
  - ϕ: effect of lottery on attendance
  - λ: effect of attendance on scores

$$\rho = \phi \times \lambda \rightarrow$$

$\lambda = \frac{\rho}{\phi} \equiv LATE$ (Local Average Treatment Effect)

- First stage:

$$\phi = E[D_i|Z_i = 1] - E[D_i|Z_i = 0]$$

- The reduced Form:

$$\rho = E[Y_i|Z_i = 1] - E[Y_i|Z_i = 0]$$

- The Local Average Treatment (LATE):

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

"The LATE is the causal effect of treatment ***for those whose treatment was solely determined by the instrument****"*

The Charter Conundrum

Instrumental Variables

- What is the IV estimator (LATE) of attendance on math and verbal scores?

$$LATE \equiv \lambda = \frac{\rho}{\phi} = \frac{E[Y_i|Z_i = 1] - E[Y_i|Z_i = 0]}{E[D_i|Z_i = 1] - E[D_i|Z_i = 0]}$$

$$\lambda_{math} = \frac{\rho_{math}}{\phi_{math}} = \frac{.355}{.741} = .479$$

$$\lambda_{verbal} = \frac{\rho_{verbal}}{\phi_{verbal}} = \frac{.113}{.741} = .152$$

Analysis of KIPP lotteries
(Angrist and Pischke, 2015, p. 104)

		KIPP applicants			
	Lynn public 5th graders	Lottery winners	Winners vs. losers	Attended KIPP	Attended vs. others
Panel A. (Selected) Baseline Characteristics					
Math score (4th grade)	-.307	-.290	.102 (.120)	-.289	0.69 (.109)
Verbal score (4th grade)	-.356	-.386	.0.63 (.125)	-.368	.088 (.114)
Panel B. Outcomes					
Attended KIPP	.000	.787	.741 (.037)	1	1
Math score	-.363	-.003	.355 (.115)	.095	.467 (.103)
Verbal score	-.417	-.262	.113 (.122)	-.211	.211
Sample size	3,964	253	371	204	371

WARNING

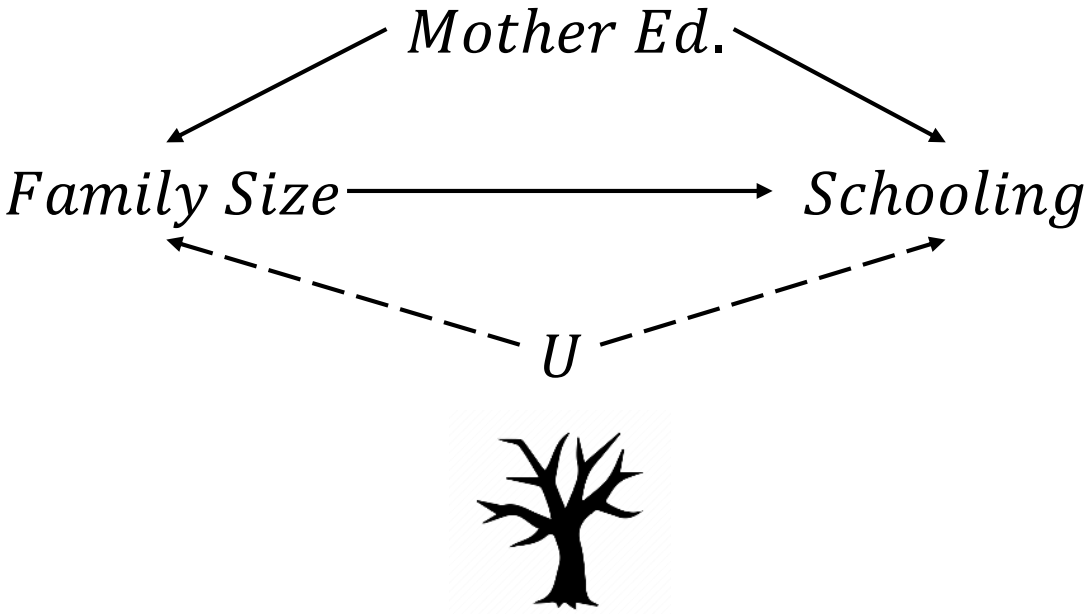
"Because the treated population includes always-takers [*in our example, those who found a way to be treated even though they did not win the lottery (*5/226=.022 *of KIPP attendants*)], LATE and TOT [Treatment-On-the-Treated, *in our example, all attendants*] are usually not the same." (Angrist and Pischke 2015, p. 114).

- External Validity?
  - Depends on the validity of our assumptions
  - Can be tested if we find more than one instrument (that would affect different segments of the population).
  - If that was the case, and estimates were equal, we would have evidences to argue for the external validity of LATE, i.e., that it approaches to TOT.
  - "With no always-takers, all of the treated are compliers, in which case, LATE is TOT." (Angrist and Pischke 2015, p. 121)

Family Size and Parental Investments in Children

Setting the Stage

- There is a negative correlation between family size and development indicators such as schooling
  - Hypothesis (Becker et al.1973, 1976): quantity-quality trade-off; as family size increases parental investment in children decreases
- Is there a causal connection between family size and children's education?
  - Mother's education
    - But also: children characteristics (health problems), family structure, gender, home environment, occupation



"The instrumental variables (IV) method harnesses partial or incomplete random assignment, whether naturally occurring or generated by researchers." (Angrist and Pischke 2015, p. 98)

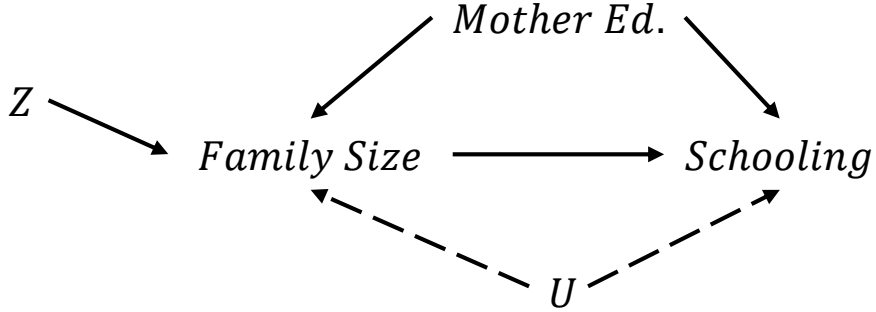
Family Size and Parental Investments in Children

Setting the Stage II

- We need to find some instrument(s)*, Z, that allow us to harness partial or incomplete randomization.
- Potential instruments:
  - **Twin siblings (Z_t)**: The number of siblings of a first-born in a family with at least 2 children will be larger (relevance) if she or he randomly (independence) has twin siblings which, if it affects schooling, does so through family size (exclusion).

    $\phi_t \neq 0; \quad Z_t \sim i.i.d; \quad Cov(Z_t, U)$
  - **Sex (Z_s)**: The number of siblings of a first born who randomly (independence) has a second born sibling of the same sex will be larger (relevance) which, if it affects schooling, does so through family size (exclusion).

    $\phi_s \neq 0; \quad Z_s \sim i.i.d.; \quad Cov(Z_s, U)$



(*) slide 12

Angrist and Pischke (2015)

Family Size and Parental Investments in Children

Two Stages Least Squares (2SLS)

- IV estimates of causal effects (λ) boil down to reduced-form (ρ) comparisons across groups defined by the instrument, scaled by the appropriate first stage (ϕ)

$$\lambda = \frac{\rho}{\phi}$$

- **Two Stages Least Squares (2SLS)** generalize IV in two ways:
  - The efficient use of multiple instruments
  - Allowing control for covariates, mitigating OVB from imperfect variables (e.g., when we suspect that Z correlates with some A)

How to 2SLS

- We know (Ch.2) that the regression coefficient of a variable on an independent dummy variable equals the difference in conditional means of the outcome, i.e.,

$$Y_i = \beta_0 + \beta_1 D_i + e_i \leftrightarrow \beta_1 = E[Y_i|D_i = 1] - E[Y_i|D_i = 0]$$

- Then we can rewrite our expression for the First-Stage as:

$$D_i = \alpha_1 + \phi Z_i + e_{1i}$$

- And the 2SLS second-stage regression by which LATE can be obtained regression the outcome on the first-stage fit, i.e., the treatment variation due to the random variation in the instrument:

$$Y_i = \alpha_2 + \lambda_{2SLS} \widehat{D}_i + e_{2i}$$

Family Size and Parental Investments in Children

How to 2SLS

- The aforementioned expression extends to non-dummy instruments and treatment variable (footnote, Angrist and Pischke 2015, p. 134).
- Furthermore, the expression is equivalent to the one we have been working on during the first part of this session (Angrist and Pischke 2015, p. 143):

$$\lambda_{2SLS} = \frac{Cov(Y_i, \widehat{D}_i)}{Var(\widehat{D}_i)} \xrightarrow[fitted\ FS]{Plugging\ the}$$

$$\lambda_{2SLS} = \frac{Cov(Y_i, \alpha_1 + \phi Z_i)}{Var(\alpha_1 + \phi Z_i)} = \frac{\phi Cov(Y_i, Z_i)}{\phi^2 Var(Z_i)} = \frac{\phi}{\phi^2}\rho = \frac{\rho}{\phi} \rightarrow$$

$$\lambda_{2SLS} = \lambda$$

- However, as already mentioned, 2SLS has two important advantages (and one more):
  - Allow to the efficient use of multiple instruments to increase statistic precision
  - Allow to introduce control variables to mitigate the OVB from imperfect instruments
  - Software allows for easy estimation of heteroscedasticity-robust standard errors

Family Size and Parental Investments in Children

How to 2SLS

- With two instruments
  - a dummy variable for sex of second born child been the same of the first born, Z_{si},
  - a dummy variable for whether or not the first born has twin siblings, Z_{ti},
- And control variables
  - a dummy variable indicating the sex of the first born, B_i,
  - maternal age, A_i,
- Our 2SLS set up looks like this:
  - FS: $D_i = \alpha_1 + \phi_s Z_{si} + \phi_t Z_{ti} + \delta_1 B_i + \gamma_1 A_i + e_{1i}$
  - SS: $Y_i = \alpha_2 + \lambda_{2SLS} \widehat{D}_i + \delta_2 B_i + \gamma_2 A_i + e_{2i}$

OLS and 2SLS estimates of the quantity-quality trade off (Angrist and Pischke, 2015, p. 137)

		2SLS estimates		
Dependent variable	OLS estimates	Twins estimates	Same-sex estimates	Twins and same-sex estimates
Years of Schooling	-.145 (.005)	.174 (.166)	.318 (.210)	.237 (.128)

- Notice that, as predicted, IV estimates have higher standard errors but, when combining both, statistical precision increases (standard errors fall).

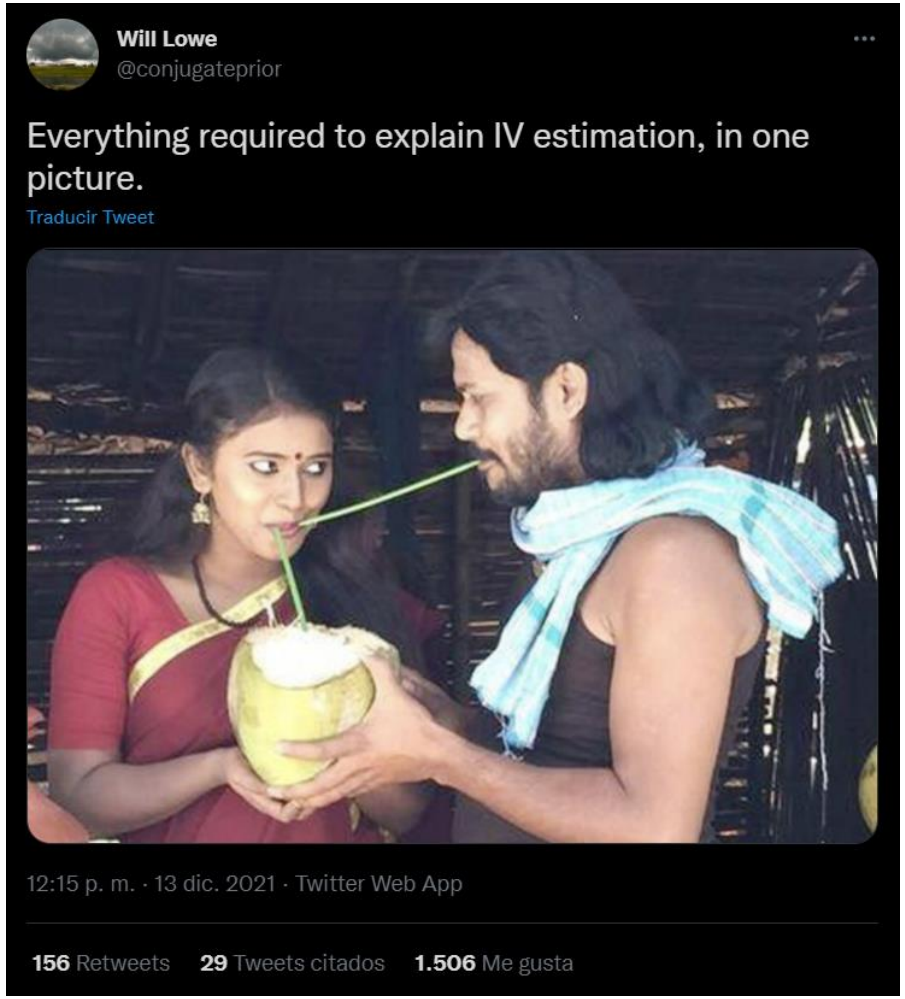
WARNING

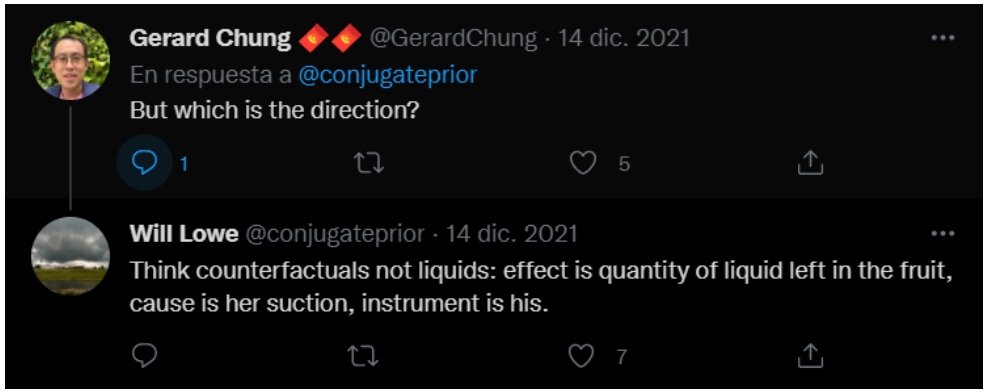
“A researcher blessed with many instruments knows that some produce a stronger first stage than others. The temptation is to use them all anyway […]. The risk here is that 2SLS estimates with many weak instruments can be misleading. A weak instrument is one that isn’t highly correlated with the repressor being instrumented, so the first-stage coefficient associated with this instrument is small or imprecisely estimated.” (Angrist and Pischke 2015, p. 114).

R example

“iv_example.pdf”

Piña collider?





Take away messages

- IV allow us to draw causal inferences when treatment is endogenous
- Good instruments (relevant, independent and satisfying the exclusion restriction) are hard to find.
- External validity of LATEs should be well argued for, specially if instruments are “naturally occurring”.

5	Regression Models – Time: 31 January 2022, 12:15 – 14:00h – Main Instructor: Isa Steinmann – Required Reading: Angrist & Pischke (2015), Chapter 2
6	Further Control Strategies – Time: 03 February 2022, 12:15 – 14:00h – Main Instructor: Isa Steinmann – Required Reading: -
7	Instrumental Variable Approaches – Time: 07 February 2022, 12:15 – 14:00h – Main Instructor: José Manuel Arencibia Alemán – Required Reading: Angrist & Pischke (2015), Chapter 3
8	Regression Discontinuity Designs I – Time: 10 February 2022, 12:15 – 14:00h – Main Instructor: José Manuel Arencibia Alemán – Required Reading: Angrist & Pischke (2015), Chapter 4

Recommended Literature

- Cunningham, S. (2021). Causal inference : the mixtape. New Haven, Connecticut, Yale University Press. (Chapter 5)
- N. Huntington-Klein (2022). The effect : an introduction to research design and causality. Boca Raton, Chapman and Hall/CRC Press. (Chapter 19)

References

- Angrist, J. D. and J.-S. Pischke (2015). Mastering 'metrics : the path from cause to effect. Princeton, N.J, Princeton University Press.