Instrumental variables I

Session 11

PMAP 8521: Program evaluation
Andrew Young School of Policy Studies

Plan for today

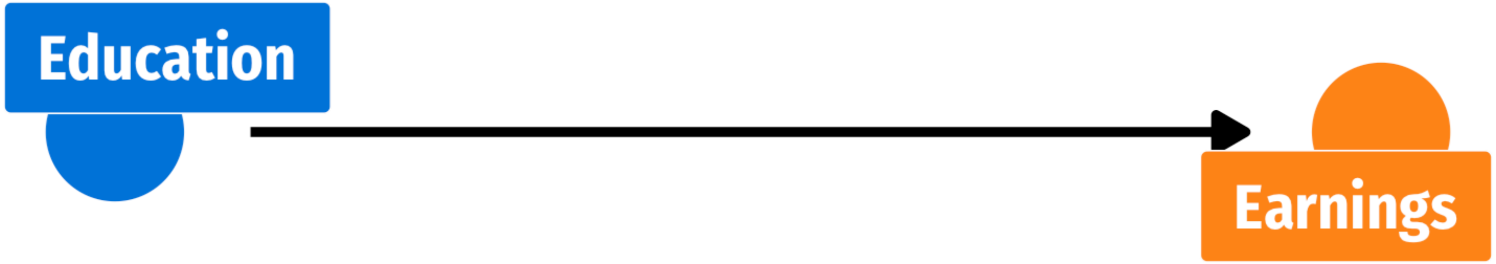
Endogeneity and exogeneity

Instruments

Using instruments

Endogeneity and exoegneity

Does education cause higher earnings?



$$\text{Earnings}_i = \beta_0 + \beta_1 \text{Education}_i + \varepsilon_i$$

If we ran this regression, would β_1 give us the causal effect of education?

$$\text{Earnings}_i = \beta_0 + \beta_1 \text{Education}_i + \varepsilon_i$$

No!

Omitted variable bias!

Unclosed backdoors!

Endogeneity!

Exogeneity and endogeneity

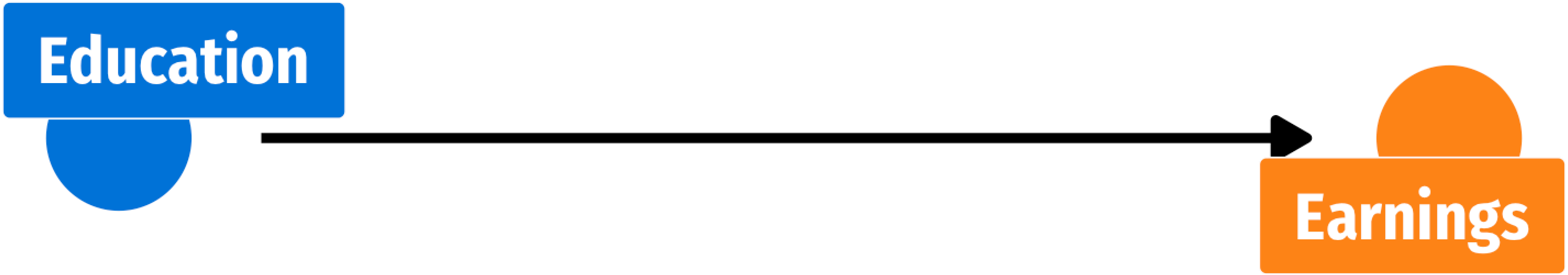
Exogenous variables

Value is not determined by anything else in the model

In a DAG, a node that doesn't have arrows coming into it

Exogeneity

Education is exogenous: no arrows *into* it



Exogeneity and endogeneity

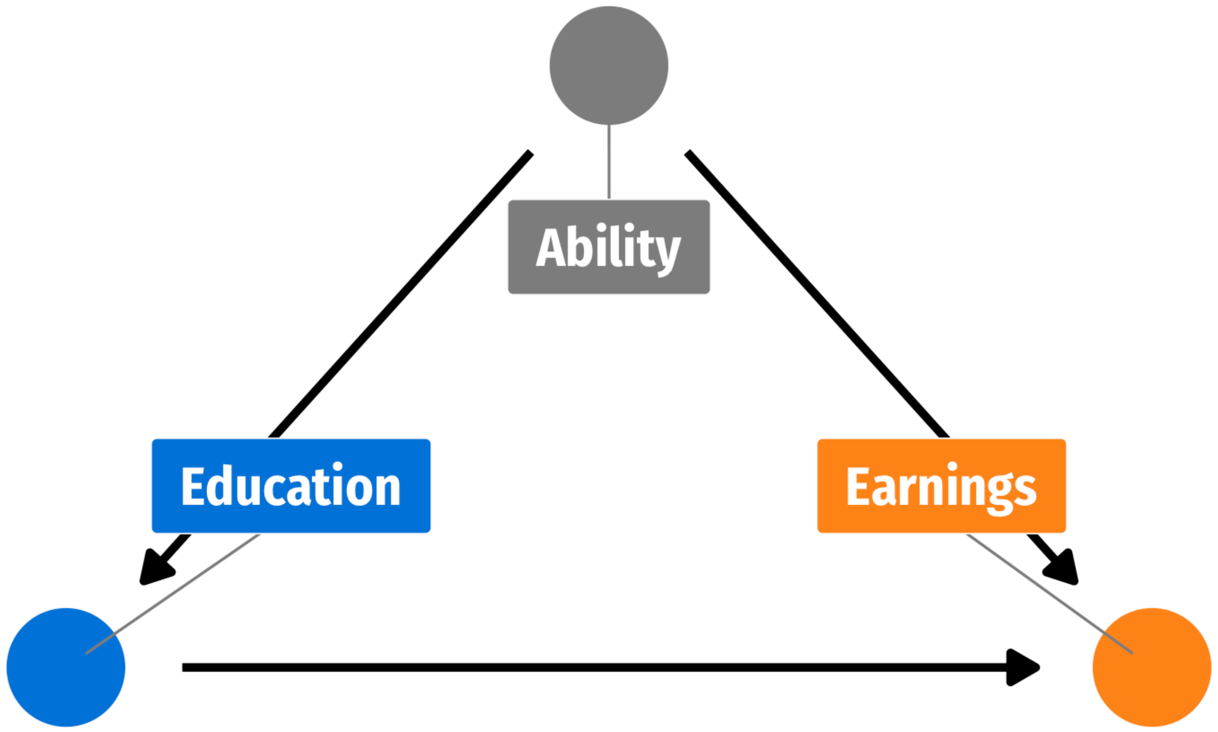
Endogenous variables

Value is determined by something else in the model

In a DAG, a node that has arrows coming into it

Endogeneity

Education is endogenous: Ability → Education

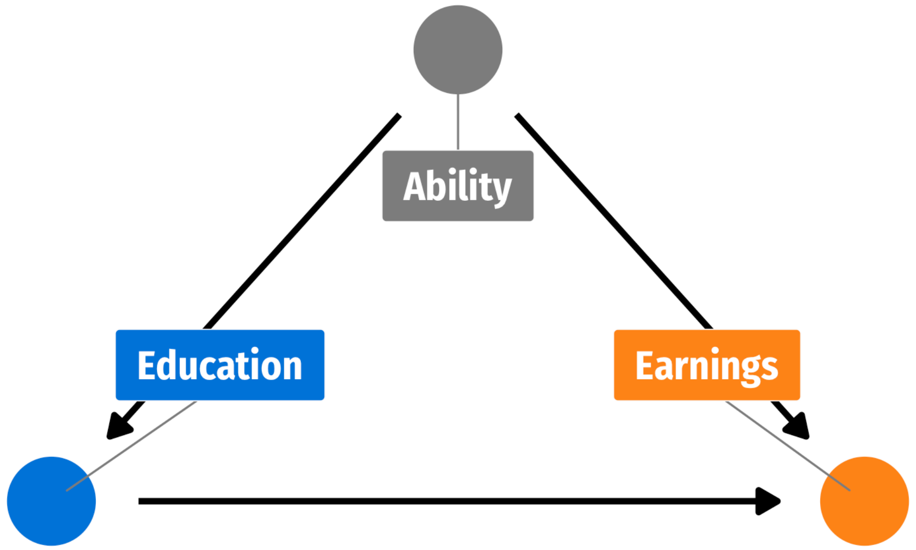


Exgoeneity

What would exogenous variation in education look like?

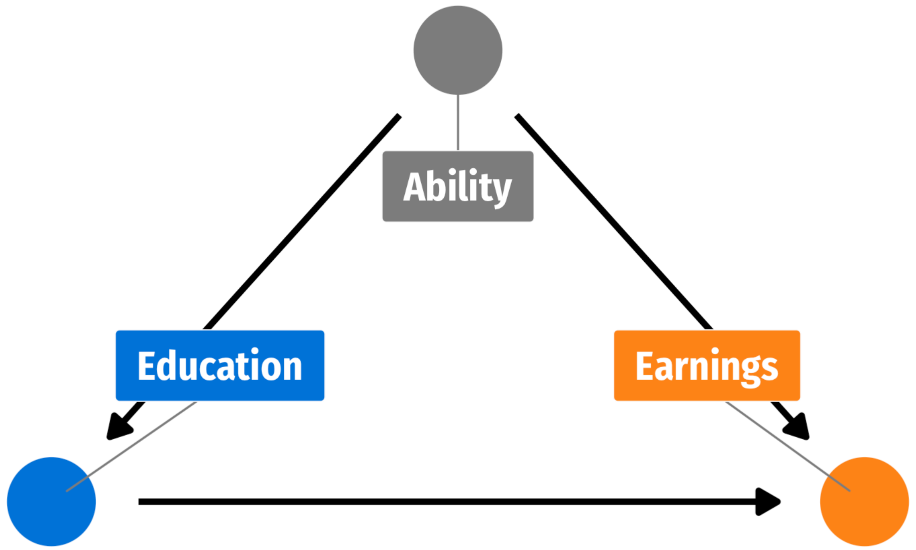
Choices to get more education that are essentially random (or at least uncorrelated with omitted variables)

We'd like education to be exogenous (an outside decision or intervention), but it's not!



Part of it is exogenous, but part of it is caused by ability, which is in the DAG

Fixing endogeneity with DAGs



Close backdoor and adjust for ability

Adjustment filters out the endogenous part of education and leaves us with just the endogenous part

$$\text{Earnings}_i = \beta_0 + \beta_1 \text{Education}_i + \beta_2 \text{Ability}_i + \varepsilon_i$$

	Outome = wage	
	Unadjusted	**Adjusted**
(Intercept)	−59.378***	−85.571***
	(10.376)	(7.198)
educ	13.124***	7.767***
	(0.618)	(0.456)
ability		0.344***
		(0.010)
Num.Obs.	1000	1000
R2	0.311	0.673
RMSE	39.13	26.97

+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001

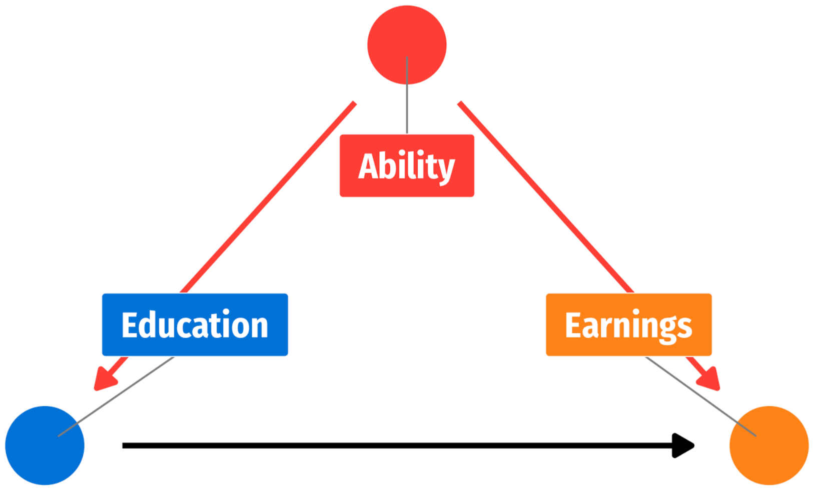
Unadjusted is wrong!

Adjusted is right!

One year of education causes hourly wage to increase by $7.77

(FAKE DATA)

But we can't measure ability!



$$\text{Earnings}_i = \beta_0 + \beta_1 \text{Education}_i + \beta_2 \text{Ability}_i + \varepsilon_i$$

Unmeasurable ability node is in the error term (ε)

$$\text{Earnings}_i = \beta_0 + \beta_1 \text{Education}_i + \varepsilon_i$$

Split exogeneity and endogeneity

What if we could somehow separate education into its endogenous and exogenous parts?

$$
\begin{aligned}
\text{Earnings}_i =& \beta_0 + \beta_1 \text{Education}_i + \varepsilon_i \\
& \beta_0 + \beta_1 (\text{Education}_i^{\text{exog.}} + \text{Education}_i^{\text{endog.}}) + \varepsilon_i \\
& \beta_0 + \beta_1 \text{Education}_i^{\text{exog.}} + \underbrace{\beta_1 \text{Education}_i^{\text{endog.}} + \varepsilon_i}_{\omega_i} \\
& \beta_0 + \beta_1 \text{Education}_i^{\text{exog.}} + \omega_i
\end{aligned}
$$

Find exogeneity with One Weird Trick™

$$\text{Earnings}_i = \beta_0 + \beta_1 \text{Education}_i^{\text{exog.}} + \omega_i$$

How do we find only Education$^{\text{exog.}}$?

Use an instrument!

Instruments

What is an instrument?

Something that is correlated with the policy variable

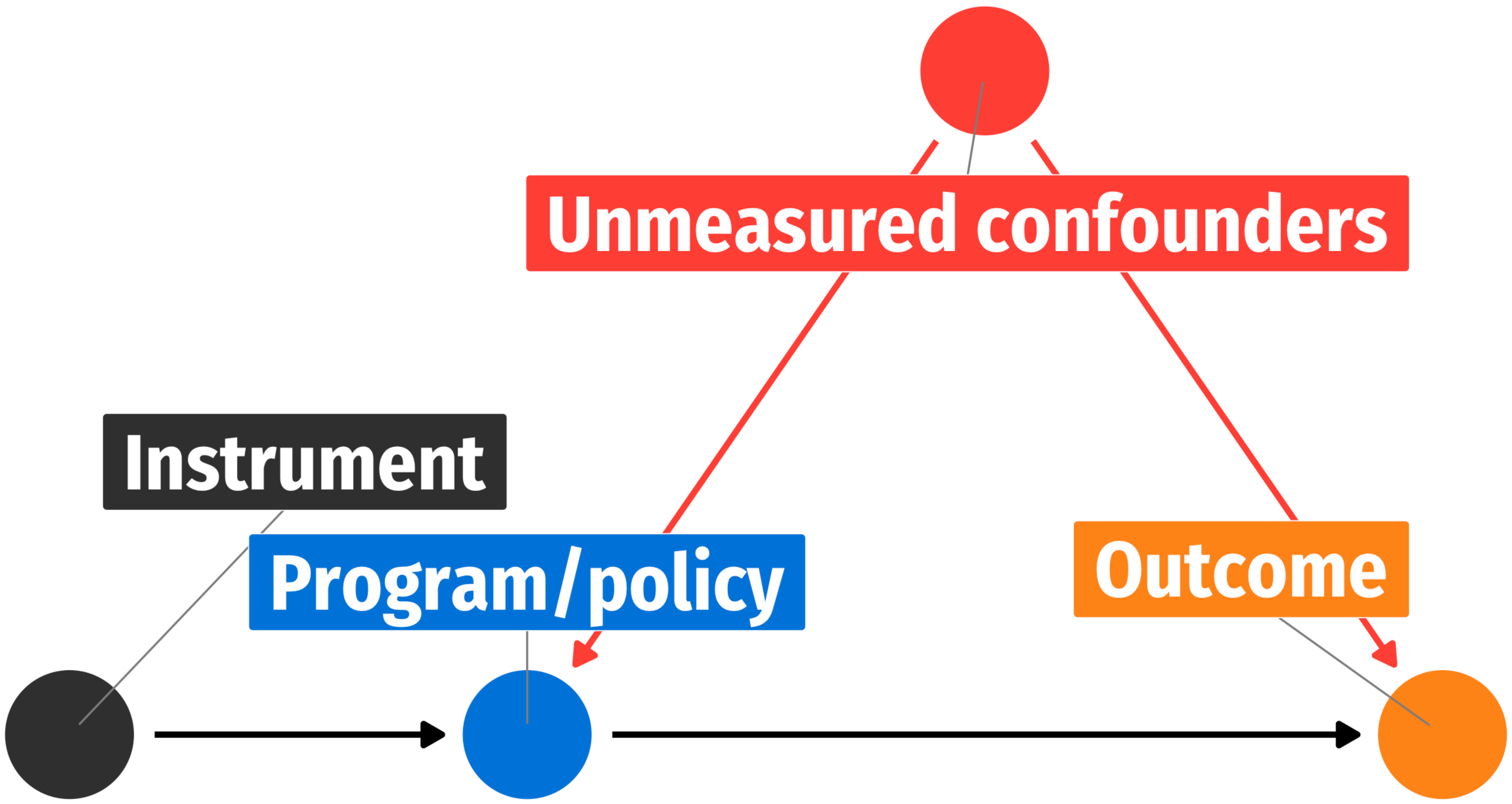
(Relevance)

Something that does not directly cause the outcome

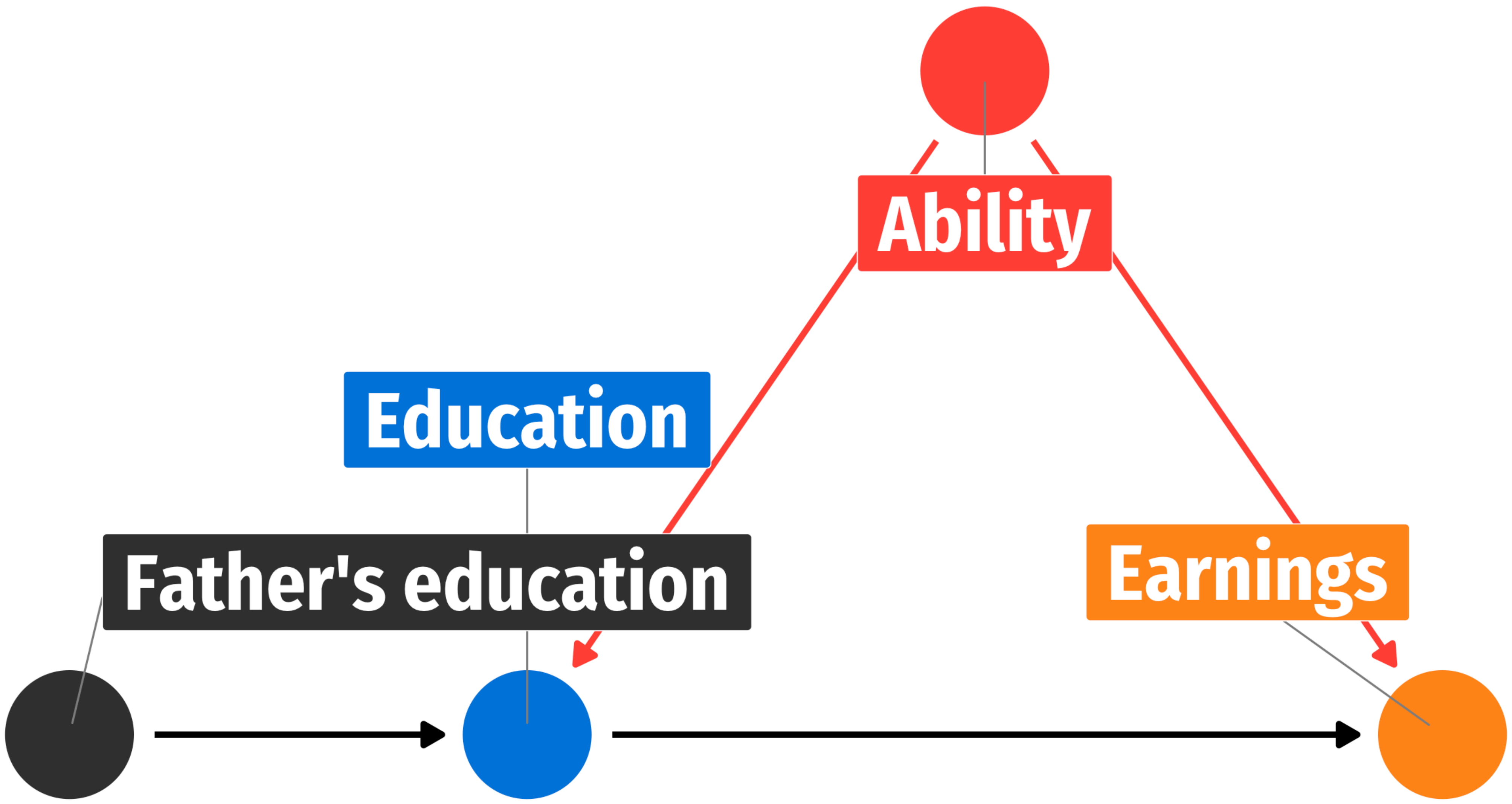
(Exclusion)

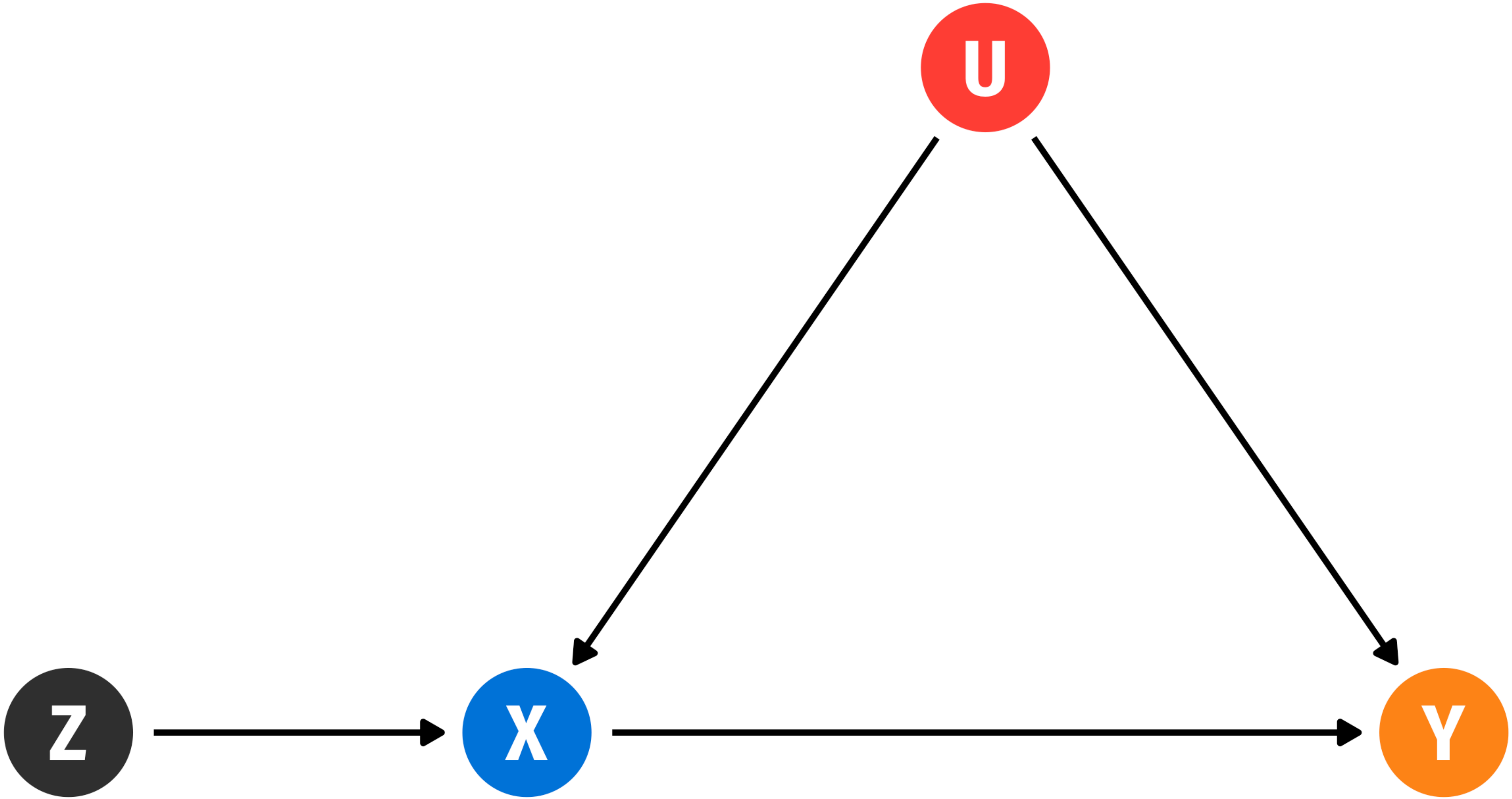
Something that is not correlated with the omitted variables

(Exogenity)



Unmeasured confounders
Instrument
Program/policy
Outcome


Ability
Education
Father's education
Earnings


U
Z
X
Y

Relevance
Correlated with policy

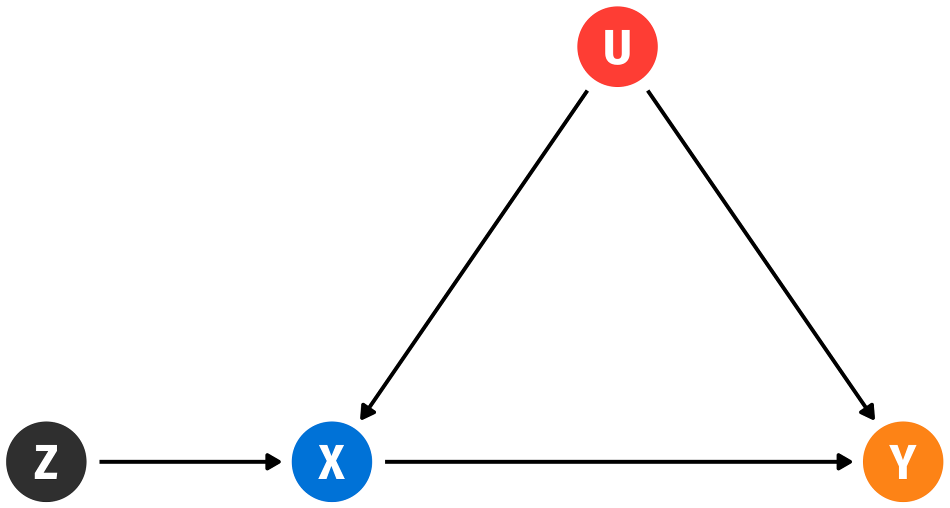
$Z \rightarrow X$ $\quad$ $Cor(Z, X) \neq 0$

Excludability
Correlated with outcome
only through policy

$Z \rightarrow X \rightarrow Y$ $\quad$ $Z \not\rightarrow Y$ $\quad$ $Cor(Z, Y \mid X) = 0$

Exogeneity
Not correlated
with omitted variables

$U \nleftrightarrow Z$ $\quad$ $Cor(Z, U) = 0$



Relevance testable with stats

Excludability testable with stats + story

Exogeneity requires story, no stats

Relevance

Instrument causes change in policy

$Z \rightarrow X \quad \text{Cor}(Z, X) \neq 0$

Social security number — Probably not relevant (uncorrelated with education)

3rd grade test scores — Potentially relevant (early grades cause more education)

Father's education — Relevant (Educated parents cause more education)

Excludability

Instrument causes outcome *only through* policy

$Z \rightarrow X \rightarrow Y \quad Z \not\rightarrow Y \quad Cor(Z, Y \mid X) = 0$

Social security number — Exclusive (SSN isn't correlated with hourly wages)

3rd grade test scores — Potentially exclusive (early grades probably don't cause wages)

Father's education — Exclusive (Parent's education doesn't cause your wages (lol))

Exogeneity

Instrument not correlated with omitted variables

$U \nrightarrow Z \qquad Cor(Z, U) = 0$

Social security number — Exogenous (Unrelated to anything related to education)

3rd grade test scores — Not exogenous (Grades correlated with other education factors)

Father's education — Exogenous (Birth to parents is random)

The huh? factor

"A necessary but not a sufficient condition for having an instrument that can satisfy the exclusion restriction is **if people are confused when you tell them about the instrument's relationship to the outcome.**"

Scott Cunningham, *Causal Inference: The Mixtape*, p. 123

Outcome	Policy	Unobserved stuff	Instrument
Income	Education	Ability	Father's education

Outcome	Policy	Unobserved stuff	Instrument
Income	Education	Ability	Father's education
Income	Education	Ability	Distance to college

Outcome	Policy	Unobserved stuff	Instrument
Income	Education	Ability	Father's education
Income	Education	Ability	Distance to college
Income	Education	Ability	Military draft

Outcome	Policy	Unobserved stuff	Instrument
Income	Education	Ability	Father's education
Income	Education	Ability	Distance to college
Income	Education	Ability	Military draft
Health	Smoking cigarettes	Other negative health behaviors	Tobacco taxes

Outcome	Policy	Unobserved stuff	Instrument
Income	Education	Ability	Father's education
Income	Education	Ability	Distance to college
Income	Education	Ability	Military draft
Health	Smoking cigarettes	Other negative health behaviors	Tobacco taxes
Crime rate	Patrol hours	# of criminals	Election cycles

Outcome	Policy	Unobserved stuff	Instrument
Income	Education	Ability	Father's education
Income	Education	Ability	Distance to college
Income	Education	Ability	Military draft
Health	Smoking cigarettes	Other negative health behaviors	Tobacco taxes
Crime rate	Patrol hours	# of criminals	Election cycles
Crime	Incarceration rate	Simultaneous causality	Overcrowding litigations

Outcome	Policy	Unobserved stuff	Instrument
Income	Education	Ability	Father's education
Income	Education	Ability	Distance to college
Income	Education	Ability	Military draft
Health	Smoking cigarettes	Other negative health behaviors	Tobacco taxes
Crime rate	Patrol hours	# of criminals	Election cycles
Crime	Incarceration rate	Simultaneous causality	Overcrowding litigations
Labor market success	Americanization	Ability	Scrabble score of name

Outcome	Policy	Unobserved stuff	Instrument
Income	Education	Ability	Father's education
Income	Education	Ability	Distance to college
Income	Education	Ability	Military draft
Health	Smoking cigarettes	Other negative health behaviors	Tobacco taxes
Crime rate	Patrol hours	# of criminals	Election cycles
Crime	Incarceration rate	Simultaneous causality	Overcrowding litigations
Labor market success	Americanization	Ability	Scrabble score of name
Conflicts	Economic growth	Simultaneous causality	Rainfall

Instruments are hard to find!

The trickiest thing to prove is the exclusion restriction

Instrument causes the outcome *only through* the policy

Most proposed instruments fail this!

Rainfall as an instrument

People love using weather as an instrument... buuuuut...

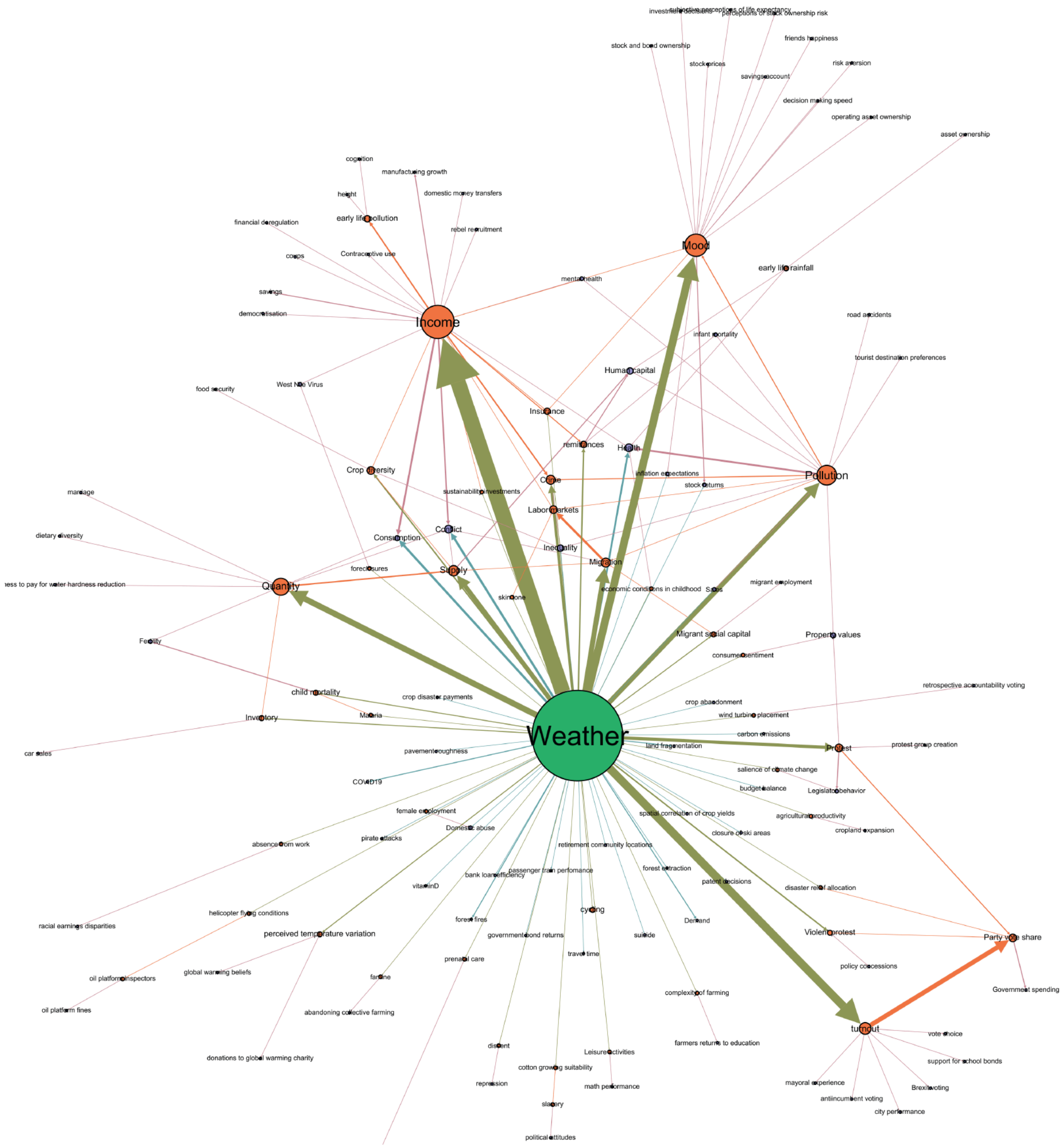
Rain, Rain, Go away: 137 potential exclusion-restriction violations for studies using weather as an instrumental variable

Jonathan Mellon (University of Manchester)

20-10-2020

Abstract

Instrumental variable (IV) analysis assumes that the instrument only affects the dependent variable via its relationship with the independent variable. Other possible causal routes from the IV to the dependent variable are exclusion-restriction violations and make the instrument invalid. Weather has been widely used as an instrumental variable in social science to predict many different variables. The use of weather to instrument different independent variables represents strong prima facie evidence of exclusion violations for all studies using weather as an IV. A review of 185 social science studies reveals 137 variables which have been linked to weather, all of which represent potential exclusion violations. I conclude with practical steps for systematically reviewing existing literature to identify possible exclusion violations when using IV designs.


Weather
Income
Mood
Pollution
Quantity
Supply
Consumption
Conflict
Crop diversity
Insurance
remittances
Health
Human capital
Crime
Labor markets
Inequality
Migration
Inventory
Protest
turnout
Party vote share
Violent protest
Migrant social capital
Property values
early life pollution
early life rainfall
mental health
infant mortality
road accidents
tourist destination preferences
cognition
manufacturing growth
height
domestic money transfers
financial deregulation
rebel recruitment
coups
Contraceptive use
savings
democratisation
food security
West Nile Virus
stock and bond ownership
friends happiness
stock prices
risk aversion
savings account
decision making speed
operating asset ownership
asset ownership
marriage
dietary diversity
Fertility
child mortality
Malaria
car sales
crop disaster payments
pavement roughness
COVID19
female employment
Domestic abuse
absence from work
pirate attacks
racial earnings disparities
helicopter flying conditions
perceived temperature variation
oil platform inspectors
oil platform fines
global warming beliefs
donations to global warming charity
abandoning collective farming
famine
prenatal care
government bond returns
forest fires
bank loan efficiency
passenger train perfomance
vitamin D
retirement community locations
cycling
travel time
suicide
Demand
forest extraction
patent decisions
dissent
repression
cotton growing suitability
slavery
political attitudes
Leisure activities
math performance
complexity of farming
farmers returns to education
sustainability investments
foreclosures
skin tone
economic conditions in childhood
Status
migrant employment
consumer sentiment
retrospective accountability voting
crop abandonment
wind turbine placement
carbon emissions
land fragmentation
protest group creation
salience of climate change
budget balance
Legislator behavior
spatial correlation of crop yields
agricultural productivity
cropland expansion
closure of ski areas
disaster relief allocation
policy concessions
Government spending
vote choice
support for school bonds
Brexit voting
mayoral experience
antiincumbent voting
city performance
inflation expectations
stock returns

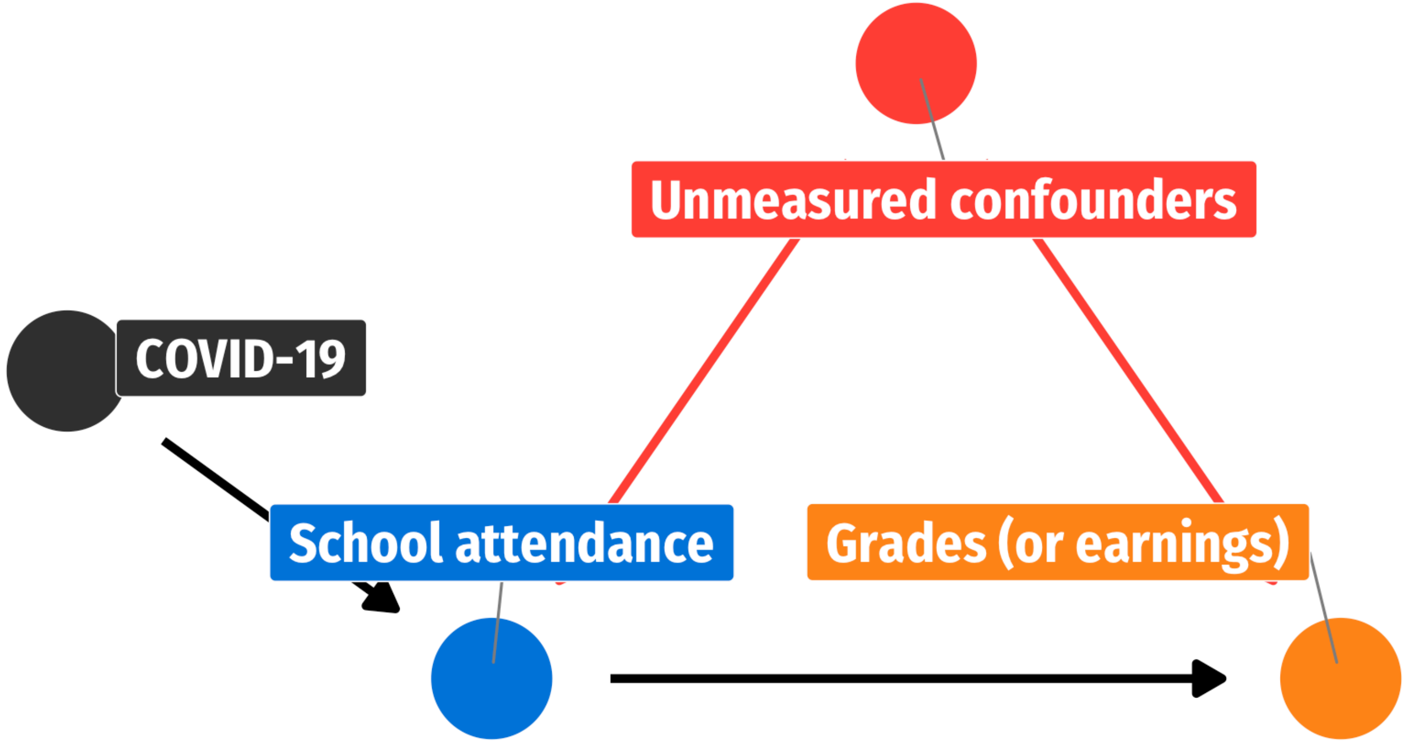
COVID-19 as an instrument

A global pandemic is a huge exogenous shock to social systems everywhere

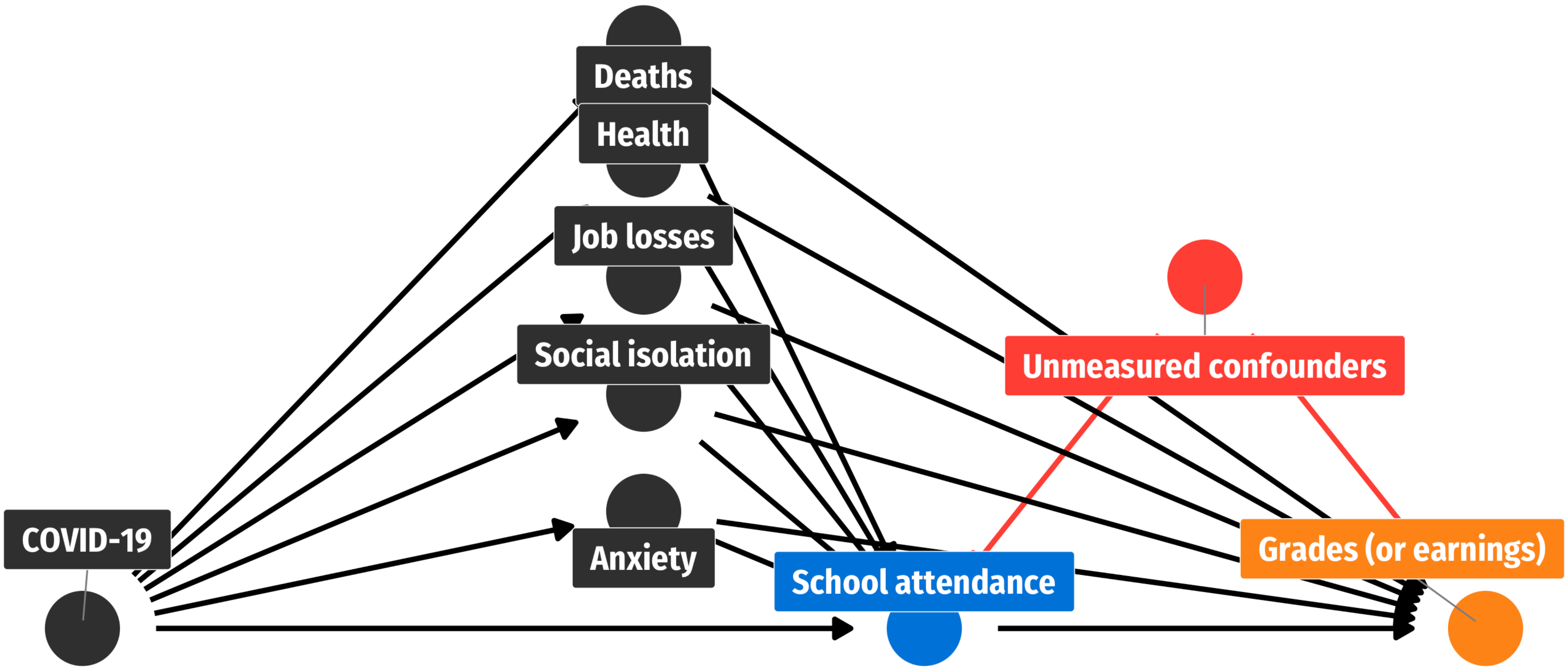
Maybe we can use it as an instrument!

COVID-19 as an instrument

What effect does closing schools have on student performance or lifetime earnings?



lolnope

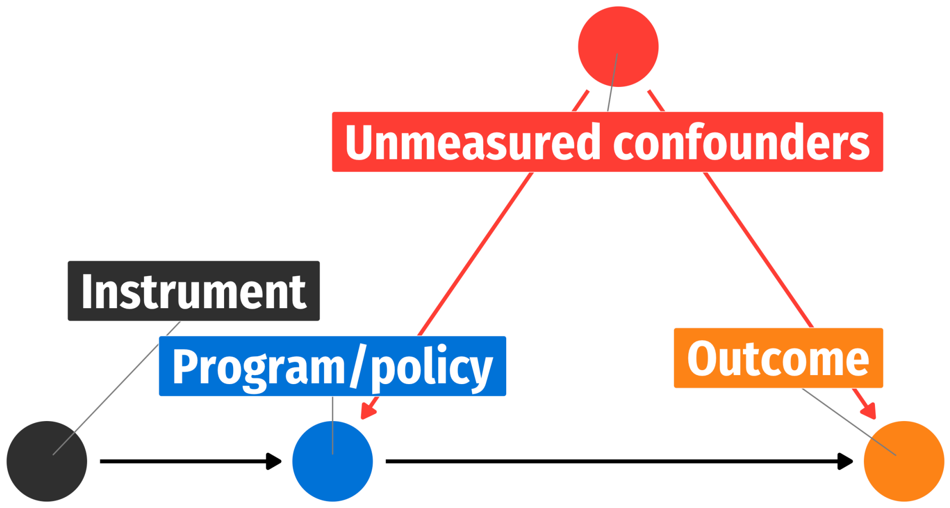


Deaths
Health
Job losses
Social isolation
Anxiety
COVID-19
School attendance
Unmeasured confounders
Grades (or earnings)

Falsifying exclusion assumptions

Can you think of some other way that the instrument can cause the outcome outside of the policy?

If so, the instrument doesn't meet exclusion restriction



Instrument → ?? → outcome?

Rainfall → ?? → civil war?

Tobacco taxes → ?? → health?

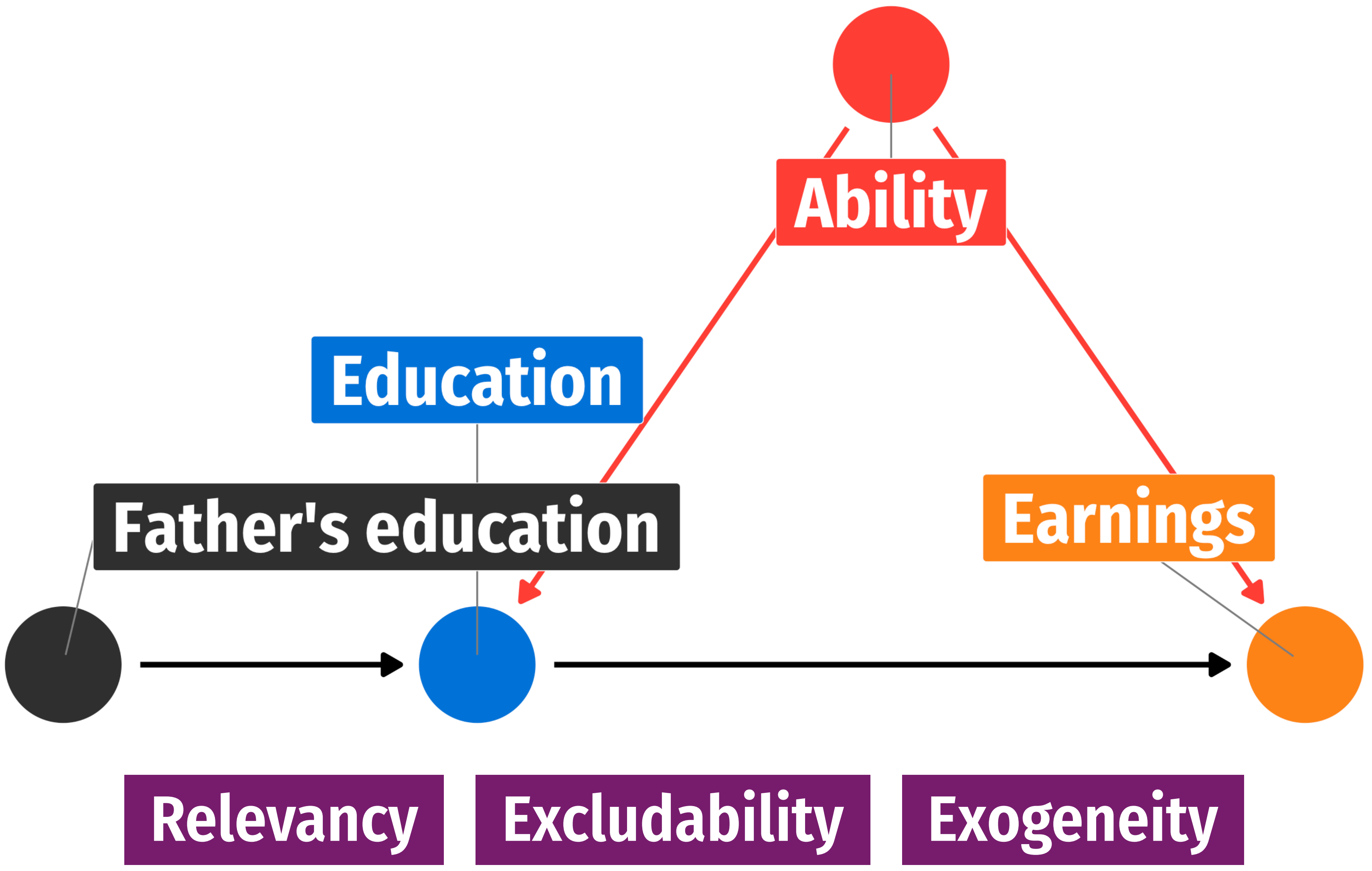
Scrabble score → ?? → Labor market success?

Using instruments

$$\text{Earnings}_i = \beta_0 + \beta_1 \text{Education}_i + \varepsilon_i$$

	Unadjusted	**Forbidden**
(Intercept)	−59.378***	−85.571***
	(10.376)	(7.198)
educ	13.124***	7.767***
	(0.618)	(0.456)
ability		0.344***
		(0.010)
Num.Obs.	1000	1000
R2	0.311	0.673
RMSE	39.13	26.97
+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001		

$$
\begin{aligned}
\text{Earnings}_i =& \beta_0 + \beta_1 \text{Education}_i + \varepsilon_i \\
& \beta_0 + \beta_1 \left(\text{Education}_i^{\text{exog.}} + \text{Education}_i^{\text{endog.}}\right) + \varepsilon_i \\
& \beta_0 + \beta_1 \text{Education}_i^{\text{exog.}} + \underbrace{\beta_1 \text{Education}_i^{\text{endog.}} + \varepsilon_i}_{\omega_i} \\
& \beta_0 + \beta_1 \text{Education}_i^{\text{exog.}} + \omega_i
\end{aligned}
$$


Ability
Education
Father's education
Earnings
Relevancy
Excludability
Exogeneity

Relevancy

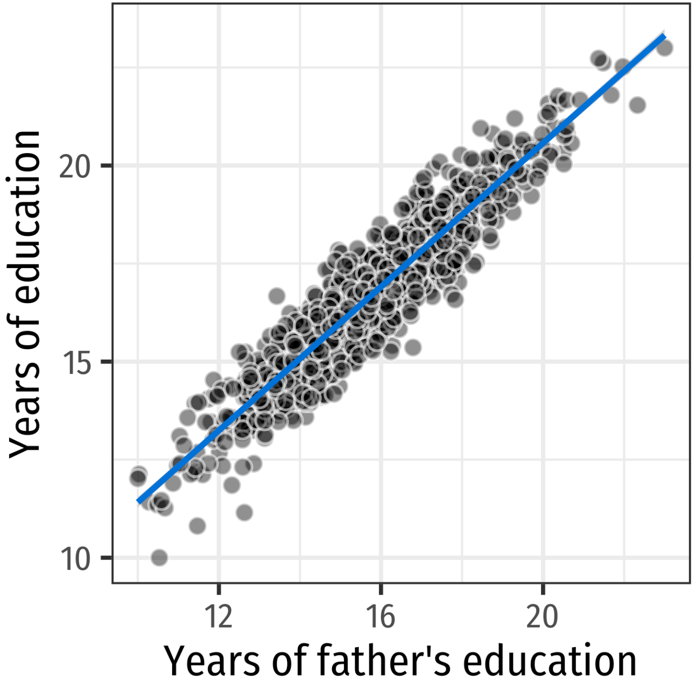
Program ~ instrument

.pull-right-wide[.small-code.smaller[

Clear, significant effect = relevant!

```
first_stage <- lm(educ ~ fathereduc, data
tidy(first_stage)
```

```
##  [38;5;246m# A tibble: 2 × 5 [39m
##   term        estimate std.error statistic
##    [3m [38;5;246m<chr> [39m [23m                    [
##  [38;5;250m1 [39m (Intercept)    2.25     0.172       13.1 3.67
```

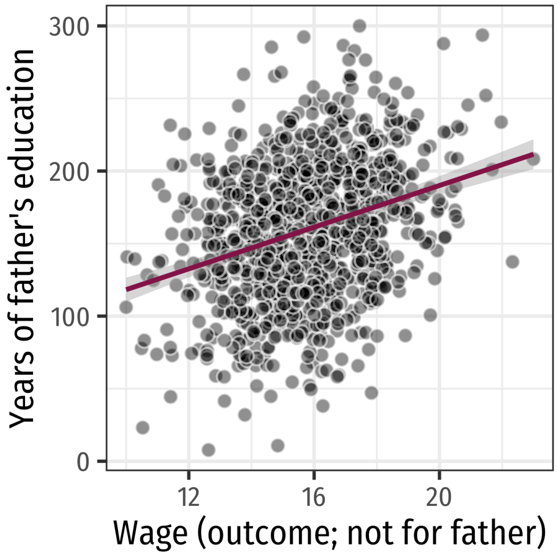


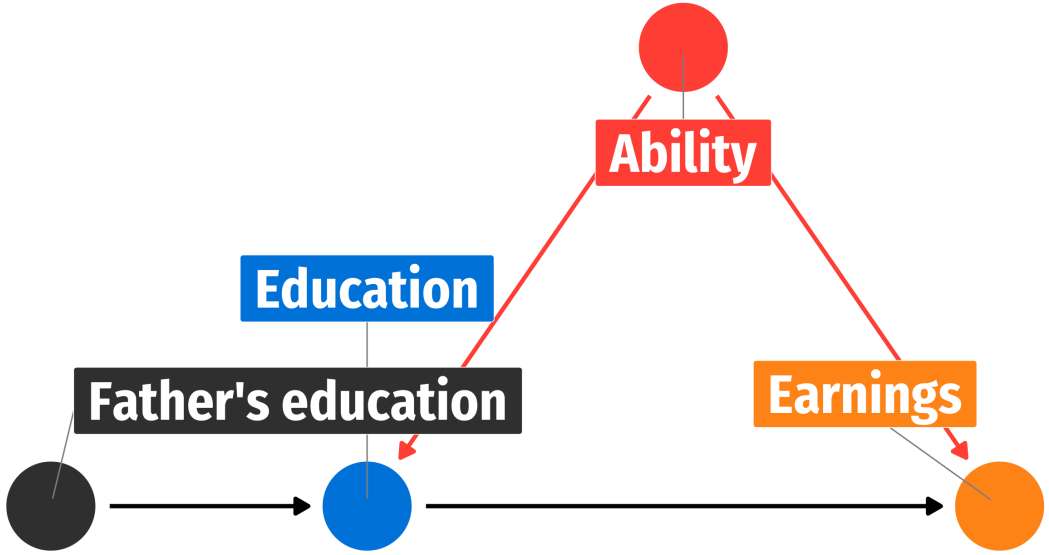
Exclusion

Does it meet exclusion assumption?

Father's education causes your wages *only through* your education?

Any other plausible node between father's education and earnings?





Exogeneity

Is assignment to your parents random?

Sure.

Is your parents' choice to gain education random?

lolz.

Two-stage least squares (2SLS)

Find exogenous part of policy variable based on instrument; use *that* to predict outcome

First stage

$$\widehat{\text{Education}}_i = \gamma_0 + \gamma_1 \text{Father's education}_i + \upsilon_i$$

"Education hat": fitted/predicted values; exogenous part of education

Second stage

$$\text{Earnings}_i = \beta_0 + \beta_1 \widehat{\text{Education}}_i + \varepsilon_i$$

Stage 1: Policy ~ instrument

```
first_stage <- lm(educ ~ fathereduc, data = father_education)

tidy(first_stage)
```

```
##  [38;5;246m# A tibble: 2 × 5 [39m
##   term        estimate std.error statistic  p.value
##    [3m [38;5;246m<chr> [39m [23m           [3m [38;5;246m<dbl> [3
##  [38;5;250m1 [39m (Intercept)    2.25     0.172       13.1 3.67
##  [38;5;250m2 [39m fathereduc     0.916    0.010 [4m8 [24m       8
```

Stage 1: Check instrument strength

Model's F-statistic (`statistic` here) should be > 104
(though most books say > 10)

```
glance(first_stage)
```

```
##  [38;5;246m# A tibble: 1 × 5 [39m
##   r.squared adj.r.squared sigma statistic p.value
##        [3m [38;5;246m<dbl> [39m [23m         [3m [38;5;246m<dbl>
##  [38;5;250m1 [39m     0.877         0.877 0.703      [4m7 [24m13
```

Stage 1: Use first stage to predict policy

$$\widehat{\text{Education}}_i = 2.251 + (0.916 \times \text{Father's education}_i) + v_i$$

```
data_with_predictions <- augment_columns(first_stage, data = father_education) |>
  rename(educ_hat = .fitted)
head(data_with_predictions)
```

.pull-left.small-code[

```
##  [38;5;246m# A tibble: 6 × 5 [39m
##     wage  educ ability fathereduc educ_hat
##     [3m [38;5;246m<dbl> [39m [23m  [3m [38;5;246m<dbl> [39m [23m
##  [38;5;250m1 [39m  180.   18.5    408.        17.2      18.0
##  [38;5;250m2 [39m  100.   16.2    310.        15.5      16.4
```

Stage 2: Outcome ~ predicted policy

```
second_stage <- lm(wage ~ educ_hat,
                   data = data_with_predictions)

tidy(second_stage)
```

```
##  [38;5;246m# A tibble: 2 × 5 [39m
##   term         estimate std.error statistic  p.value
##    [3m [38;5;246m<chr> [39m [23m          [3m [38;5;246m<dbl> [3
##  [38;5;250m1 [39m (Intercept)    28.8      12.7        2.27 2.32
##  [38;5;250m2 [39m educ_hat        7.83      0.755     10.4  5.10
```

	Unadjusted	Forbidden	2SLS IV
(Intercept)	−59.378***	−85.571***	28.819*
	(10.376)	(7.198)	(12.672)
educ	13.124***	7.767***	
	(0.618)	(0.456)	
ability		0.344***	
		(0.010)	
educ_hat			7.835***
			(0.755)
Num.Obs.	1000	1000	1000
R2	0.311	0.673	0.097
RMSE	39.13	26.97	44.80

+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001

Unadjusted is wrong!

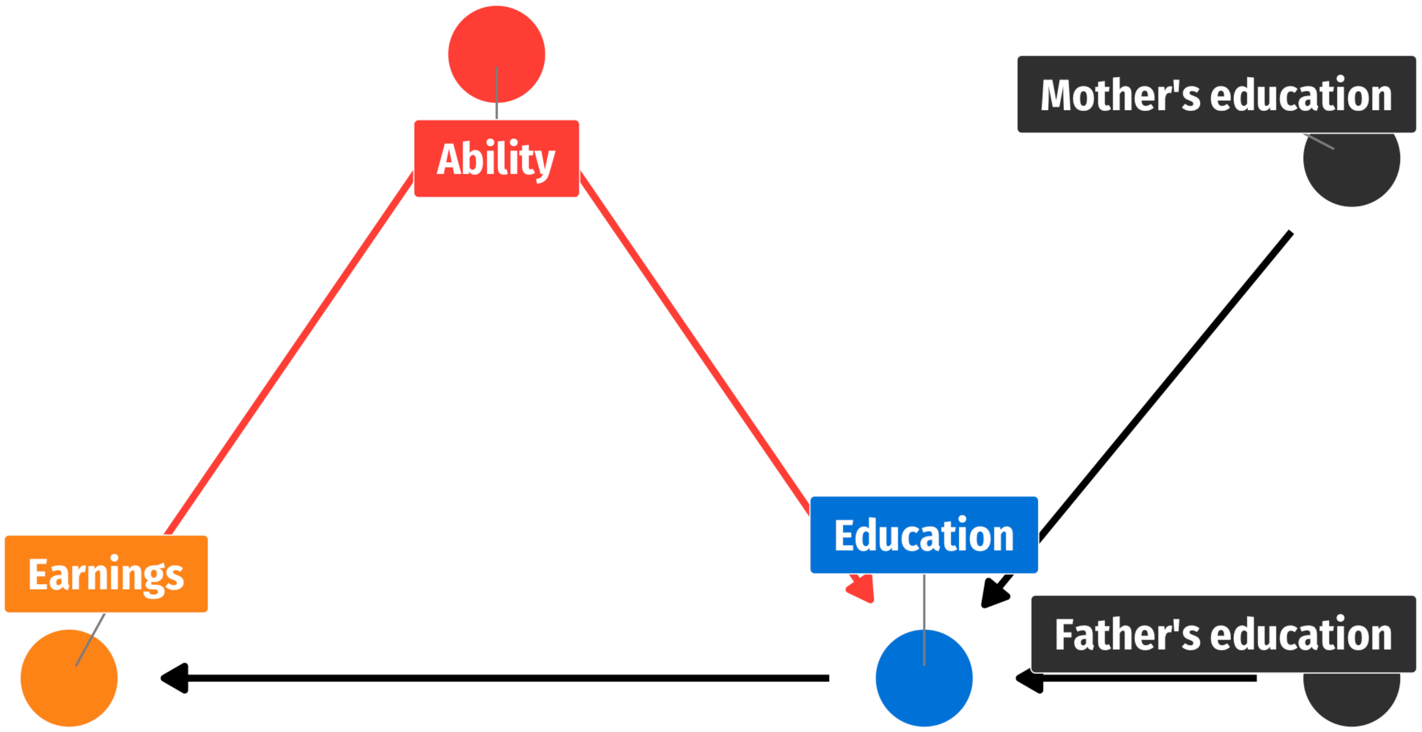
Forbidden is right, but not actually measurable!

2SLS is close *and* measurable!

One year of education causes hourly wage to increase by $7.84

Multiple instruments

You can use multiple instruments to explain more of the endogeneity in the policy node



Multiple instruments

$$\widehat{\text{Education}}_i = \gamma_0 + \gamma_1 \text{Father's education}_i + \\ \gamma_2 \text{Mother's education}_i + v_i$$

$$\text{Earnings}_i = \beta_0 + \beta_1 \widehat{\text{Education}}_i + \varepsilon_i$$

Other control variables

You can use control variables too!

**For mathy reasons,
all exogenous controls need to go in both stages**

$$\begin{aligned}\widehat{\text{Education}}_i = \gamma_0 + \gamma_1 \text{Father's education}_i + \gamma_2 \text{Mother's education}_i + \\ \gamma_3 \text{SES}_i + \gamma_4 \text{State}_i + \gamma_5 \text{Year}_i + \upsilon_i\end{aligned}$$

$$\begin{aligned}\text{Earnings}_i = \beta_0 + \beta_1 \widehat{\text{Education}}_i + \\ \beta_2 \text{SES}_i + \beta_3 \text{State}_i + \beta_4 \text{Year}_i + \varepsilon_i\end{aligned}$$

Faster, more accurate ways to run 2SLS

Running the first stage, calculating policy-hat, then running second stage is neat, but time consuming!

```
first_stage <- lm(educ ~ fathereduc, data = father_education)

data_with_predictions <- augment_columns(first_stage, data = father_education) |>
  rename(educ_hat = .fitted)

second_stage <- lm(wage ~ educ_hat, data = data_with_predictions)
```

Your standard errors will be wrong unless you adjust them with fancy math by hand

Use R packages that do all that work for you instead!

Faster, more accurate ways to run 2SLS

`ivreg()` **from the ivreg package**

Outcome ~ 2nd stage stuff | 1st stage stuff

.pull-left.code-small.tiny[

```
library(ivreg)
model_ivreg <- ivreg(wage ~ educ | fathereduc,
                     data = father_education)
tidy(model_ivreg)
```

```
##  [38;5;246m# A tibble: 2 × 5 [39m
##   term        estimate std.error statistic  p.value
```

Faster, more accurate ways to run 2SLS

`iv_robust()` from the estimatr package

Outcome ~ 2nd stage stuff | 1st stage stuff

```
library(estimatr)
model_iv_robust <- iv_robust(wage ~ educ | fathereduc,
                             data = father_education)
tidy(model_iv_robust)
```

```
##          term  estimate  std.error statistic      p.value conf.low conf.high
## 1 (Intercept) 28.818695 11.1645893  2.581259 9.985789e-03 6.909932 50.727459
## 2        educ  7.834935  0.6635423 11.807739 3.281862e-30 6.532837  9.137033
##    df outcome
## 1 998    wage
## 2 998    wage
```

(See also `lfe()` from the felm package for IV with fancy fixed effects)

	Unadjusted	Forbidden	2SLS IV (by hand)	2SLS IV (ivreg())	2SLS IV (iv_robust())
(Intercept)	−59.378***	−85.571***	28.819*	28.819*	28.819**
	(10.376)	(7.198)	(12.672)	(11.468)	(11.165)
educ	13.124***	7.767***		7.835***	7.835***
	(0.618)	(0.456)		(0.683)	(0.664)
ability		0.344***			
		(0.010)			
educ_hat			7.835***		
			(0.755)		
Num.Obs.	1000	1000	1000	1000	1000
R2	0.311	0.673	0.097	0.261	0.261
R2 Adj.	0.311	0.672	0.096	0.260	0.260

+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001

General IV process

1: Is the instrument relevant?

Instrument correlated with policy/program; F-statistic in 1st stage > 104

2: Does the instrument meet exclusion assumption?

Instrument causes outcome *only through* policy/program. **Good luck.**

3: Is the instrument exogenous?

No arrows going into instrument node in DAG

4: 2-stage least squares (2SLS)

program ~ instrument; outcome ~ program_hat **OR** `iv_robust()`