Instrumental variables I & II

April 8, 2020

PMAP 8521: Program Evaluation for Public Service
Andrew Young School of Policy Studies
Spring 2020

Plan for today

Endogeneity & exogeneity

Instruments

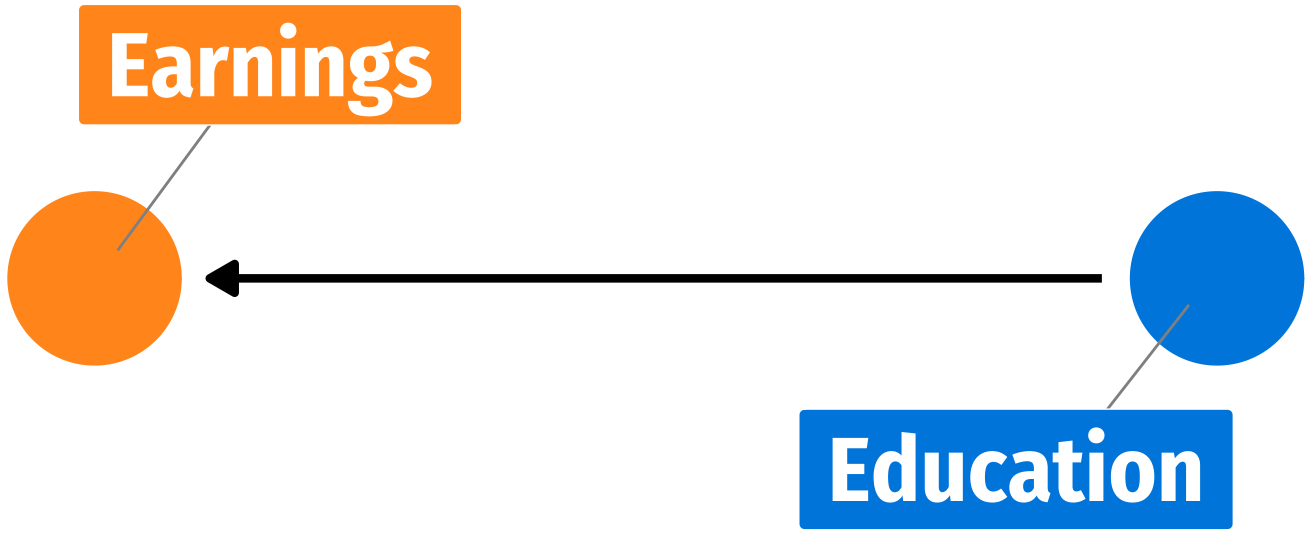
Using instruments

IV with R

Treatment effects & compliance

Endogeneity
& exogeneity

Does education cause higher earnings?



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

Outcome variable

Policy/program variable

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

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

No!

Omitted variable bias!

Unclosed backdoors!

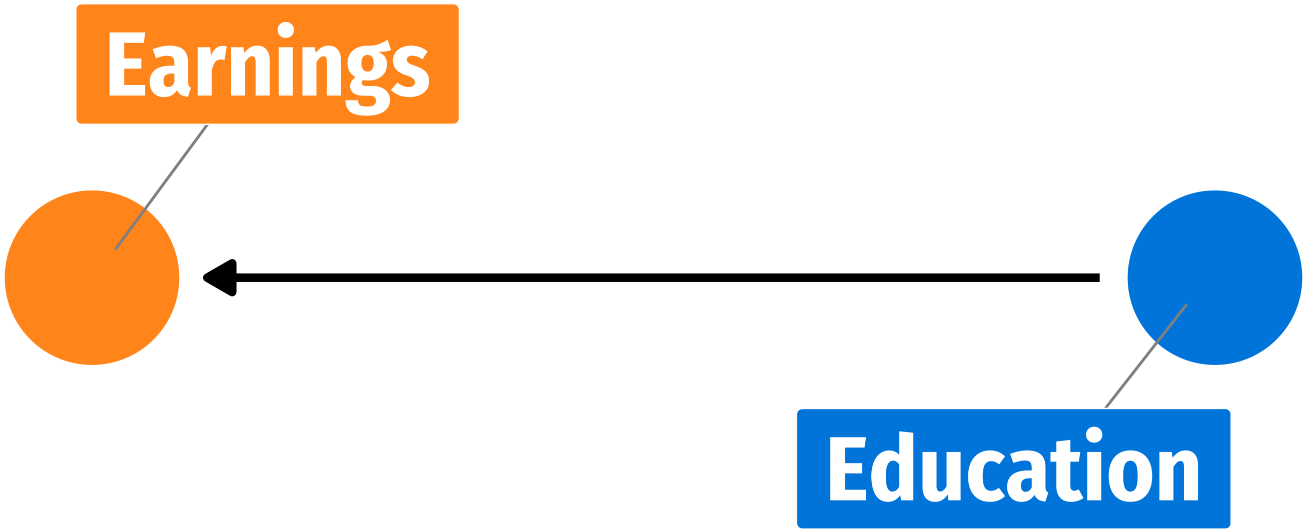
Endogeneity!

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



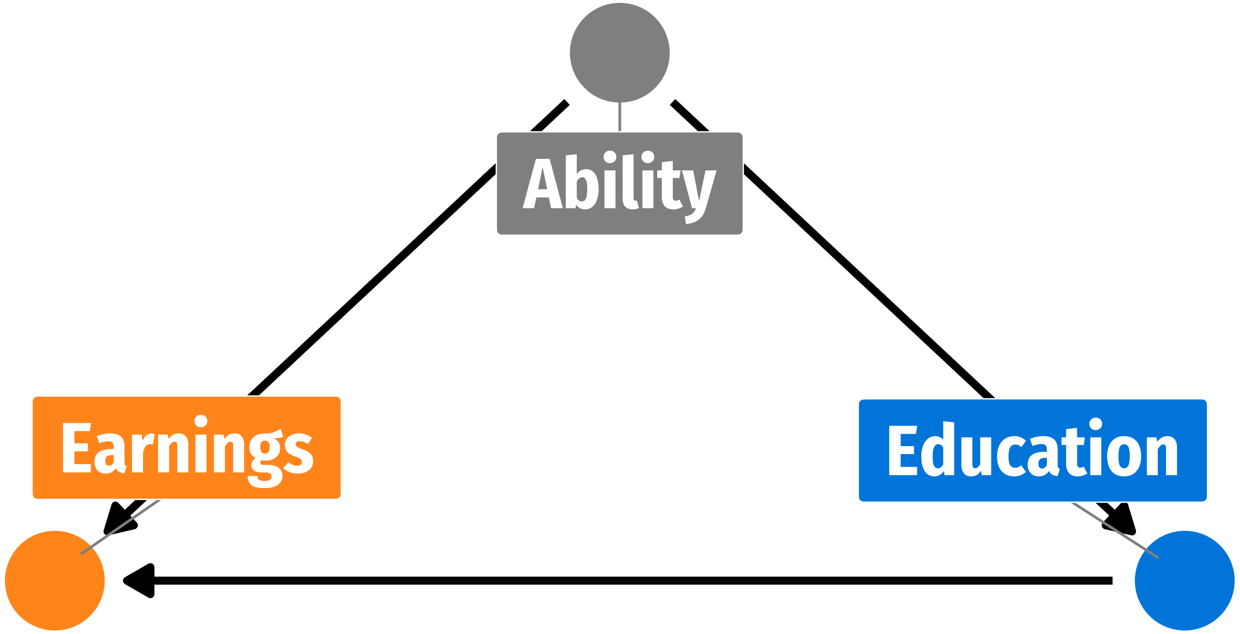
Education is exogenous here

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



Education is endogenous now

Exogeneity and endogeneity

Endogeneity

The error term (ϵ) is related to the explanatory variables

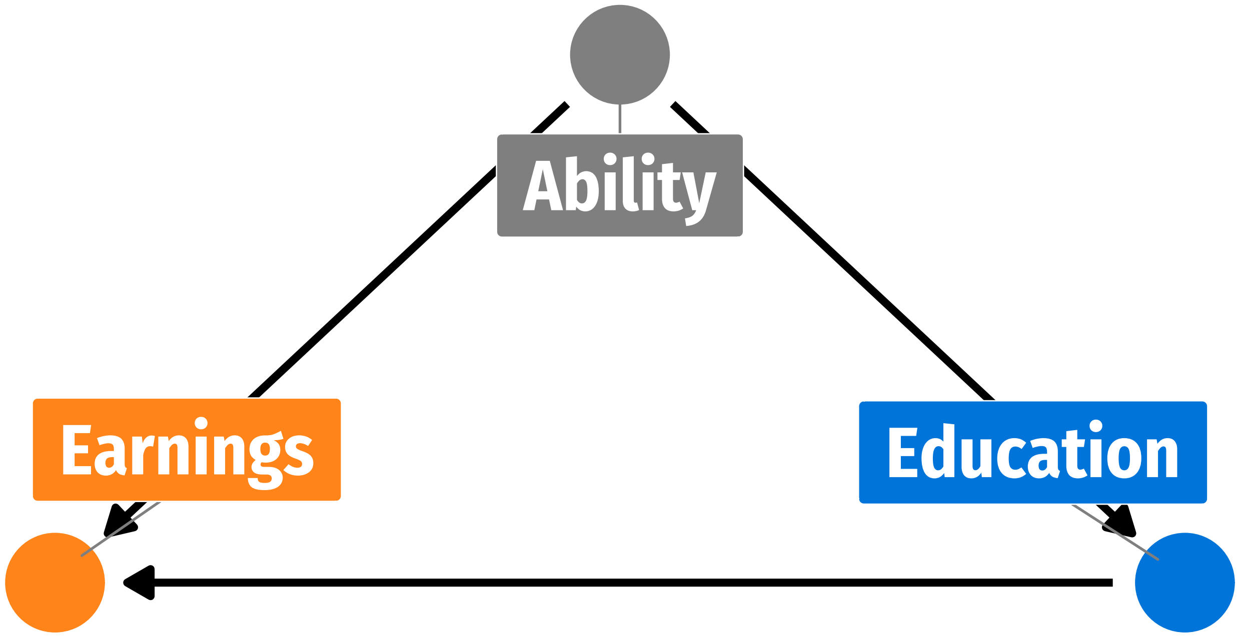
$$\text{Earnings}_i = \beta_0 + \beta_1 \text{Education}_i + \epsilon_i$$

Education is related to some part of this this unobserved stuff ϵ

What would exogenous variation in education look like?

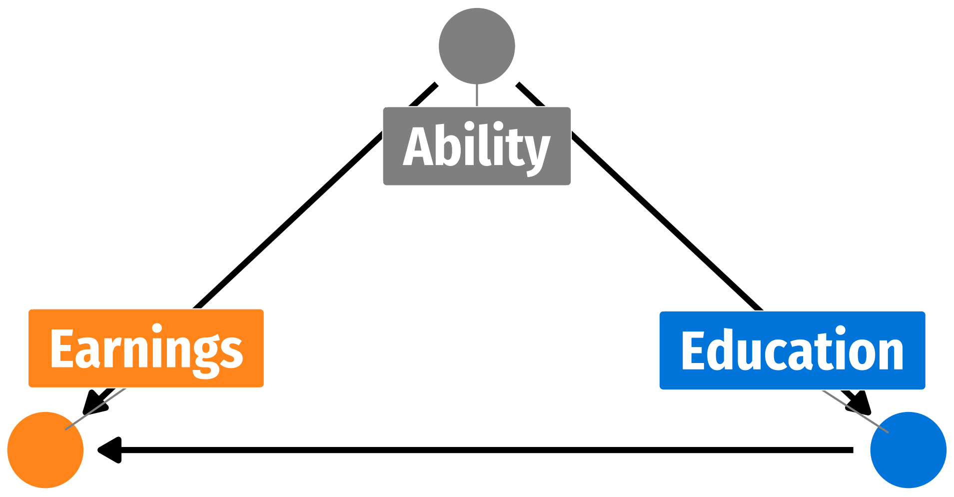
Choices to get more education that are essentially random (or at least uncorrelated with omitted variables)

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 back door and adjust for ability

Filters out the endogenous part of education and leaves us with just the exogenous part

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

	Outcome = Wage	
	Unadjusted	Adjusted
(Intercept)	-53.085***	-80.263***
	(8.492)	(5.659)
educ	12.240***	9.242***
	(0.503)	(0.343)
ability		0.258***
		(0.007)
Num.Obs.	1000	1000
R2	0.372	0.726
Adj.R2	0.371	0.726

* p < 0.1, ** p < 0.05, *** p < 0.01

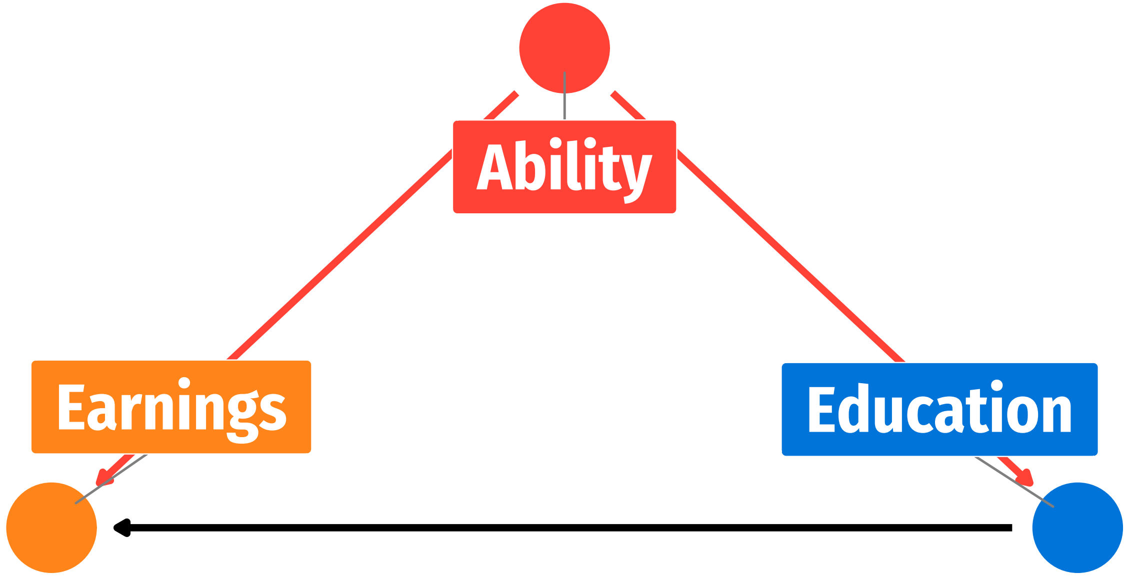
	Outcome = Wage	
	Unadjusted	Adjusted
(Intercept)	-53.085***	-80.263***
	(8.492)	(5.659)
educ	12.240***	9.242***
	(0.503)	(0.343)
ability		0.258***
		(0.007)
Num.Obs.	1000	1000
R2	0.372	0.726
Adj.R2	0.371	0.726

* p < 0.1, ** p < 0.05, *** p < 0.01

Wrong!

Right!

But we can't measure ability!



Unmeasurable!

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

Ability is in here

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

Split exogeneity and endogeneity

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

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

Isolate exogeneity with this One Weird Trick™

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

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

Use an instrument!

Instruments

What is an instrument?

Something that is correlated with the policy variable

Something that does not directly cause the outcome

Something that is not correlated with the omitted variables

Relevance

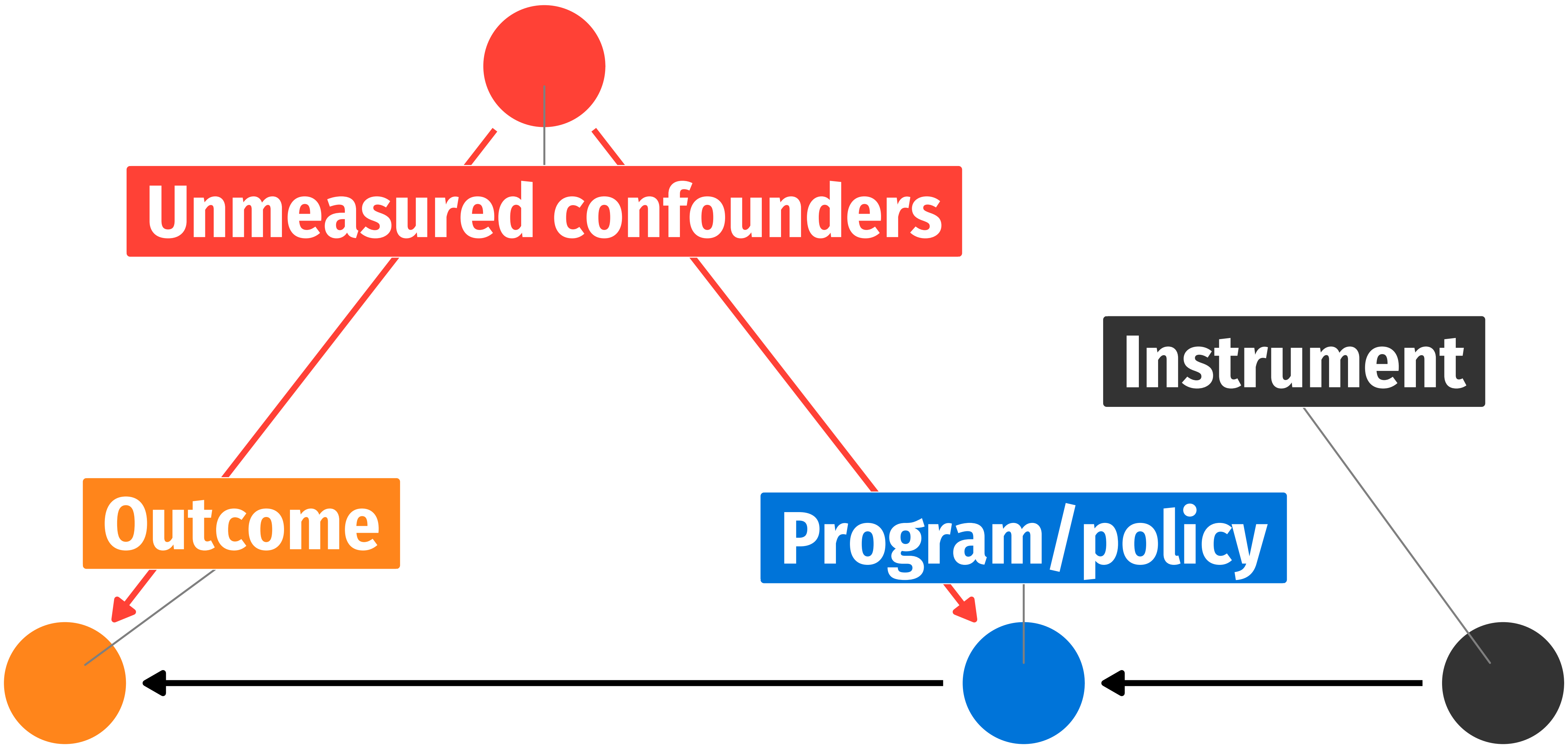
Testable with stats!

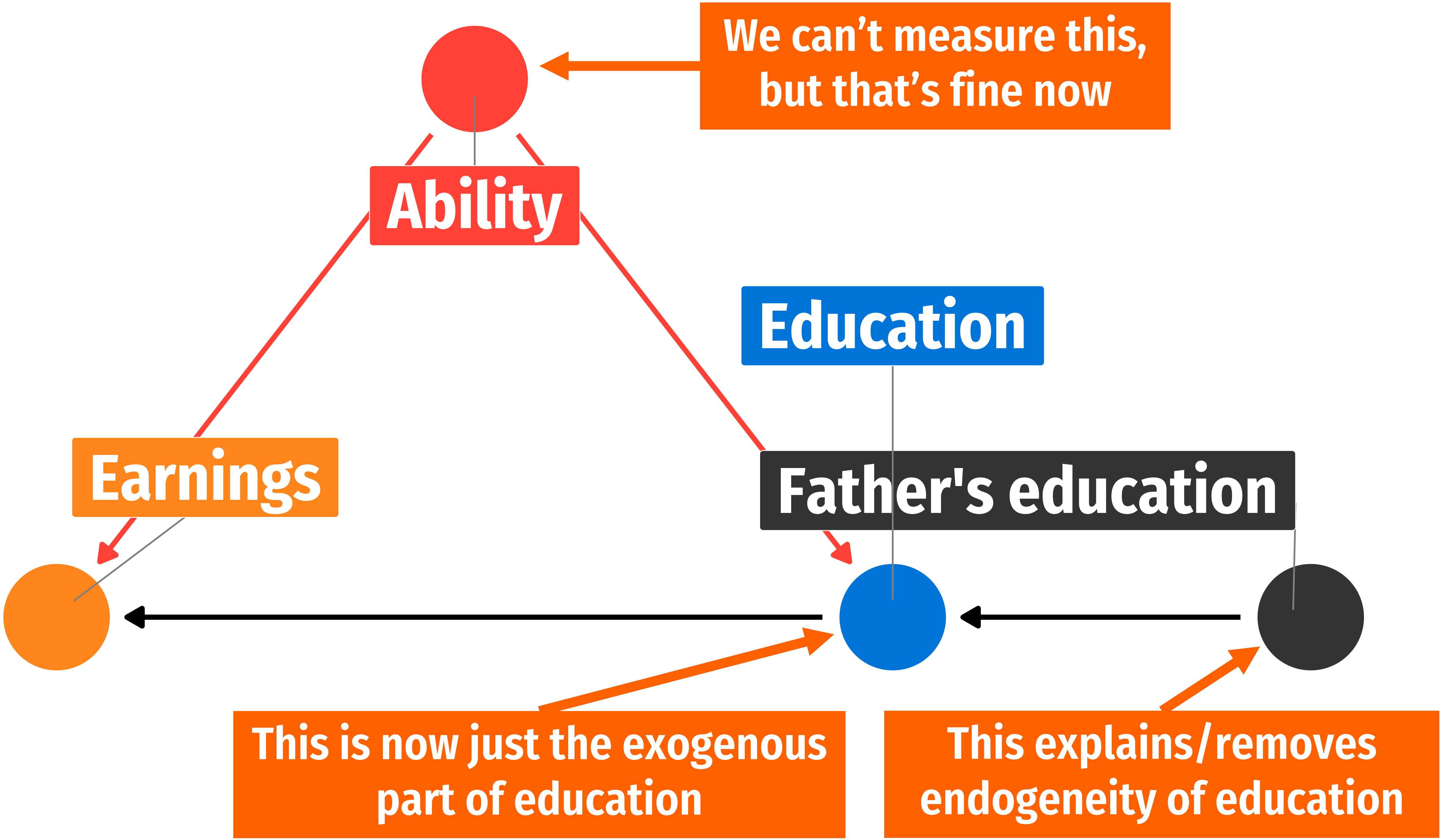
Exclusion

("only through")

Not testable!

Exogenous


Unmeasured confounders
Outcome
Program/policy
Instrument


We can't measure this, but that's fine now
Ability
Education
Earnings
Father's education
This is now just the exogenous part of education
This explains/removes endogeneity of education

What is an instrument?

Something that is correlated with the policy variable

Something that does not directly cause the outcome

Something that is not correlated with the omitted variables

Relevance

Testable with stats!

Exclusion

("only through")

Not testable!

Exogenous

Relevancy

Instrument causes changes in policy

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

Exclusion

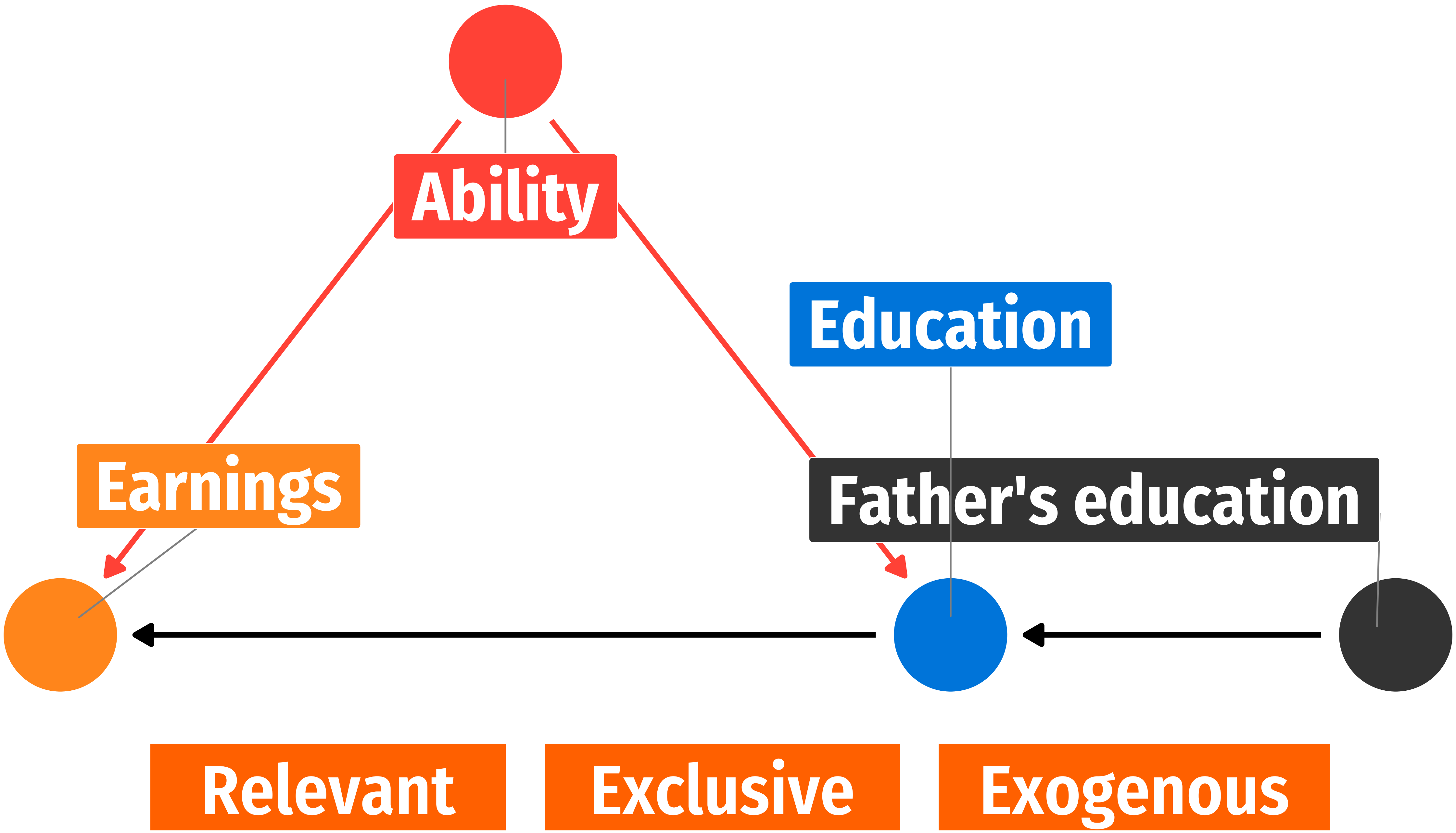
Instrument only causes outcome through the policy/program ("only through" condition)

Social security number	**Exclusive** SSN isn't correlated with hourly wage
3rd grade test scores	**Potentially exclusive** Early grades probably don't cause wages
Father's education	**Exclusive** Parent's education doesn't correlate with your hourly wage

Exogeneity

Instrument independent of all other factors; is randomly assigned

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


Ability
Education
Earnings
Father's education
Relevant
Exclusive
Exogenous

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. 213

Outcome variable	Policy variable	Omitted variable	Instrumental variable
Health	Smoking cigarettes	Other negative health behaviors	Tobacco taxes

Outcome variable	Policy variable	Omitted variable	Instrumental variable
Health	Smoking cigarettes	Other negative health behaviors	Tobacco taxes
Labor market success	Americanization	Ability	Scrabble score of name

Outcome variable	Policy variable	Omitted variable	Instrumental variable
Health	Smoking cigarettes	Other negative health behaviors	Tobacco taxes
Labor market success	Americanization	Ability	Scrabble score of name
Crime rate	Patrol hours	# of criminals	Election cycles

Outcome variable	Policy variable	Omitted variable	Instrumental variable
Health	Smoking cigarettes	Other negative health behaviors	Tobacco taxes
Labor market success	Americanization	Ability	Scrabble score of name
Crime rate	Patrol hours	# of criminals	Election cycles
Income	Education	Ability	Father's education
			Distance to college
			Military draft

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

Outcome variable	Policy variable	Omitted variable	Instrumental variable
Health	Smoking cigarettes	Other negative health behaviors	Tobacco taxes
Labor market success	Americanization	Ability	Scrabble score of name
Crime rate	Patrol hours	# of criminals	Election cycles
Income	Education	Ability	Father’s education
			Distance to college
			Military draft
Crime	Incarceration rate	Simultaneous causality	Overcrowding litigations
Election outcomes	Federal spending in a district	Political vulnerability	Federal spending in the rest of the state

Outcome variable	Policy variable	Omitted variable	Instrumental variable
Health	Smoking cigarettes	Other negative health behaviors	Tobacco taxes
Labor market success	Americanization	Ability	Scrabble score of name
Crime rate	Patrol hours	# of criminals	Election cycles
Income	Education	Ability	Father's education
			Distance to college
			Military draft
Crime	Incarceration rate	Simultaneous causality	Overcrowding litigations
Election outcomes	Federal spending in a district	Political vulnerability	Federal spending in the rest of the state
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

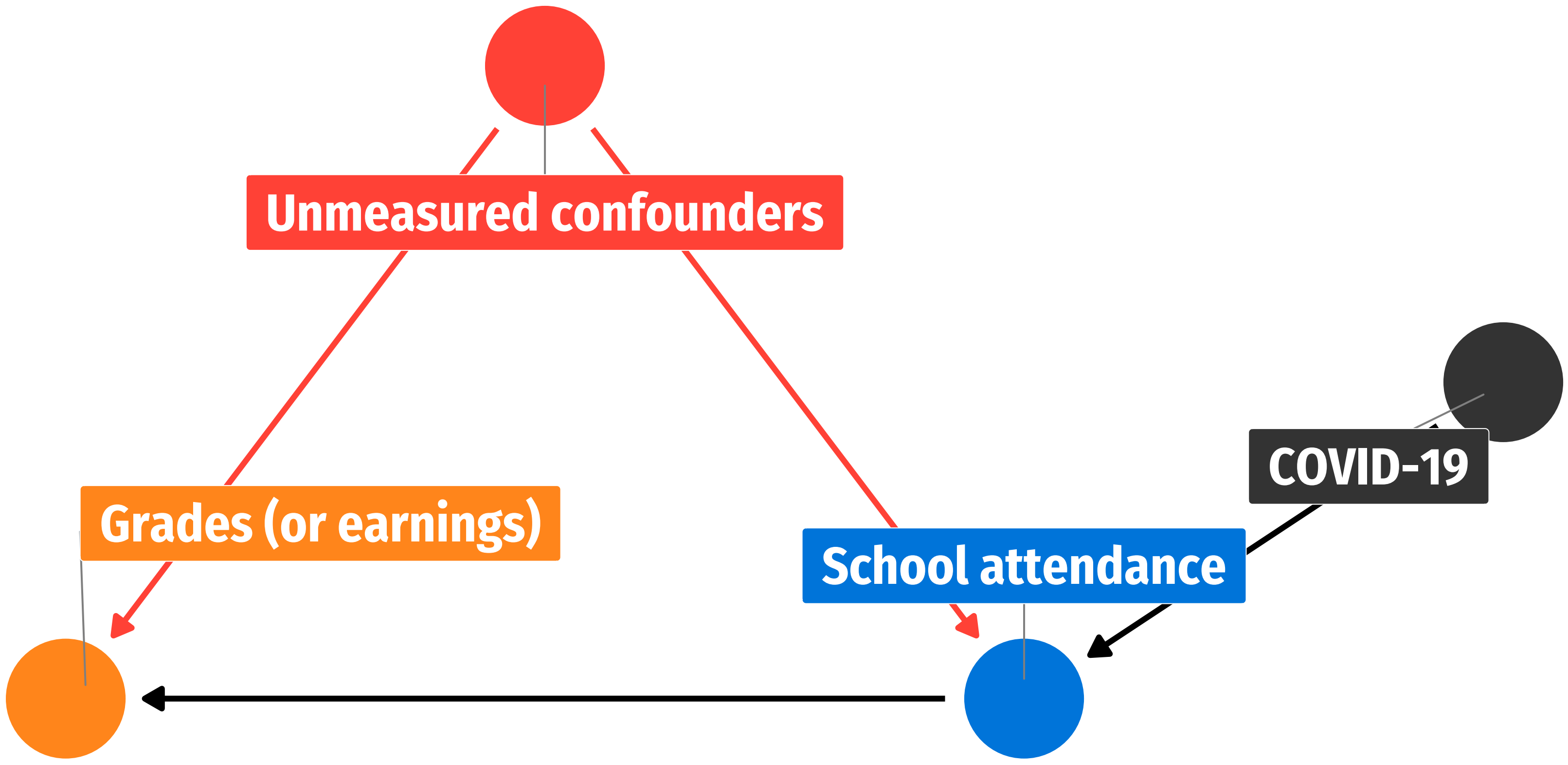
COVID-19 as an instrument

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

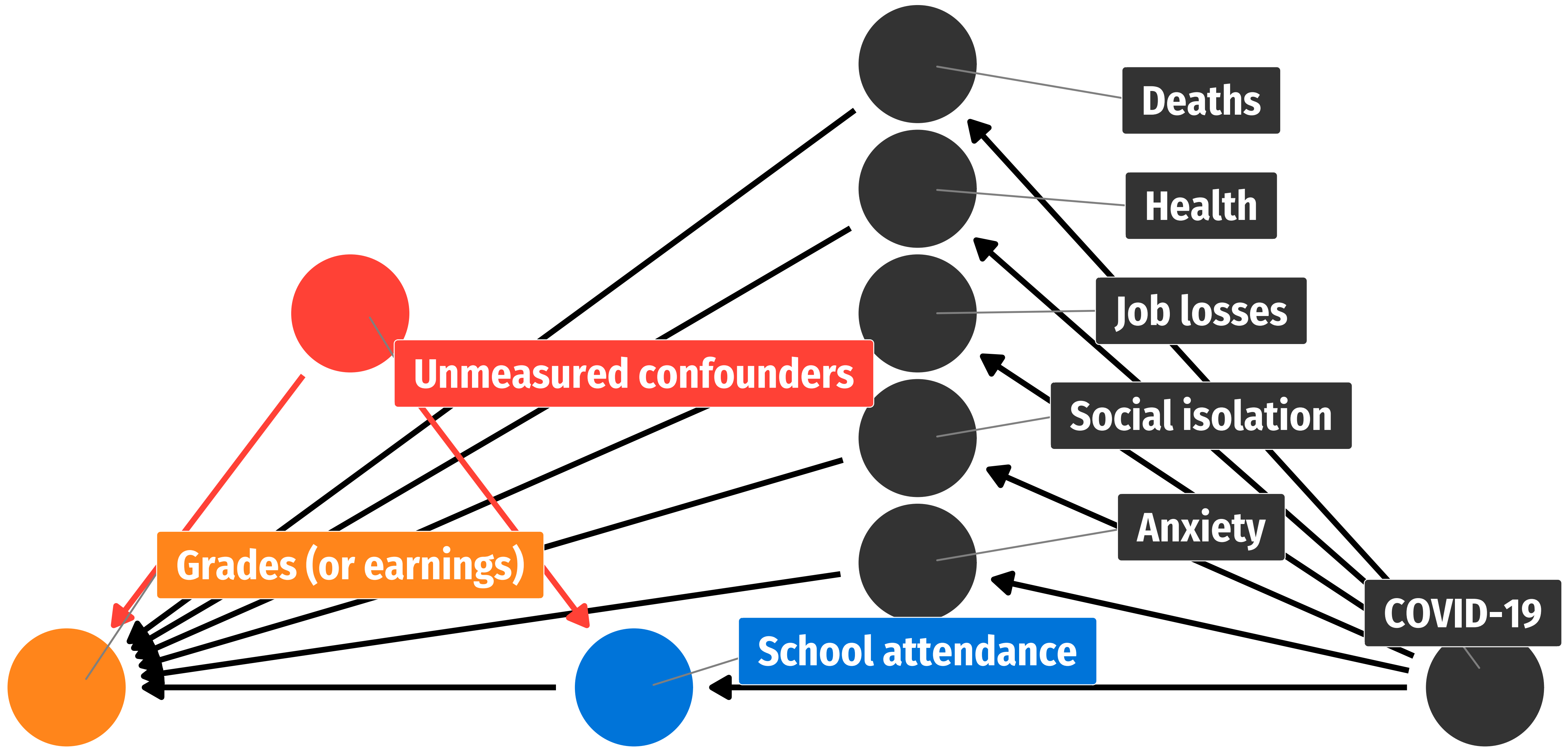
Maybe we can use it as an instrument!

COVID-19 as an instrument

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



lolnope

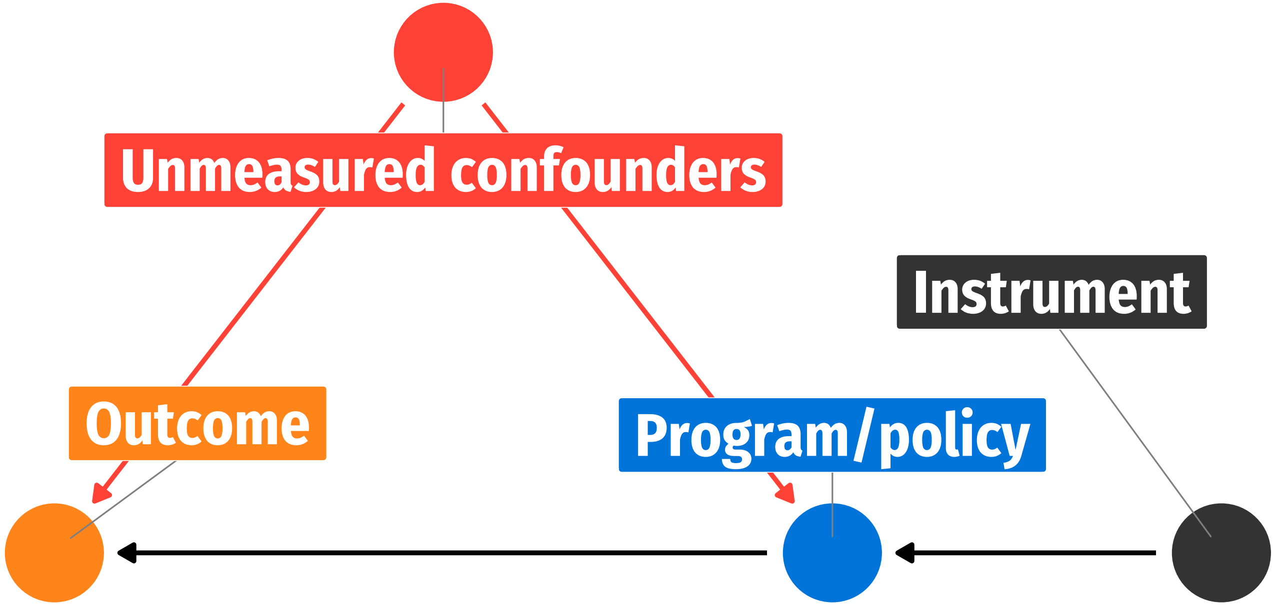


Deaths
Health
Job losses
Unmeasured confounders
Social isolation
Anxiety
Grades (or earnings)
COVID-19
School attendance

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?

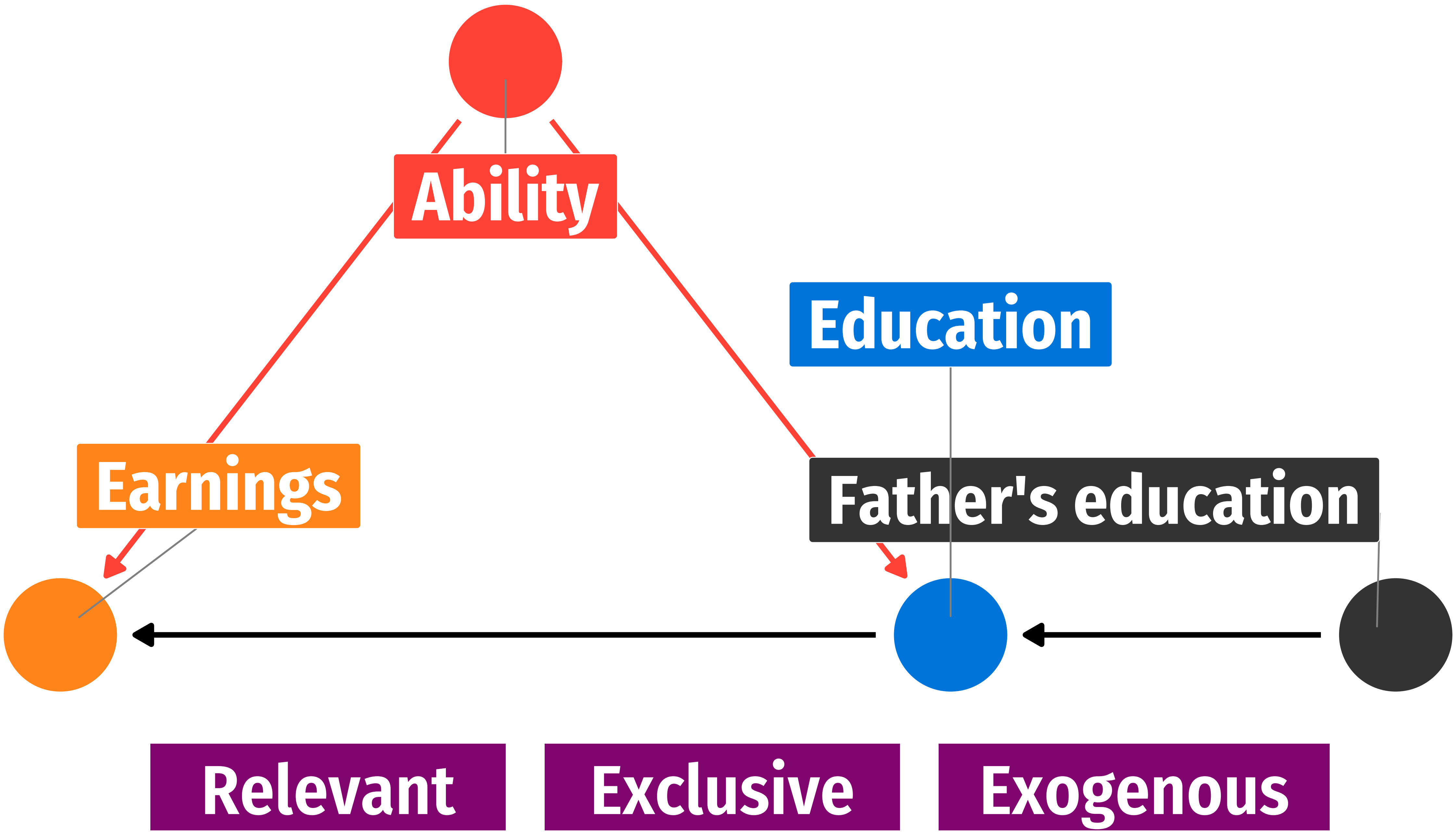
Using instruments

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

	Outcome = Wage	
	Unadjusted	Adjusted
(Intercept)	-53.085***	-80.263***
	(8.492)	(5.659)
educ	12.240***	9.242***
	(0.503)	(0.343)
ability		0.258***
		(0.007)
Num.Obs.	1000	1000
R2	0.372	0.726
Adj.R2	0.371	0.726

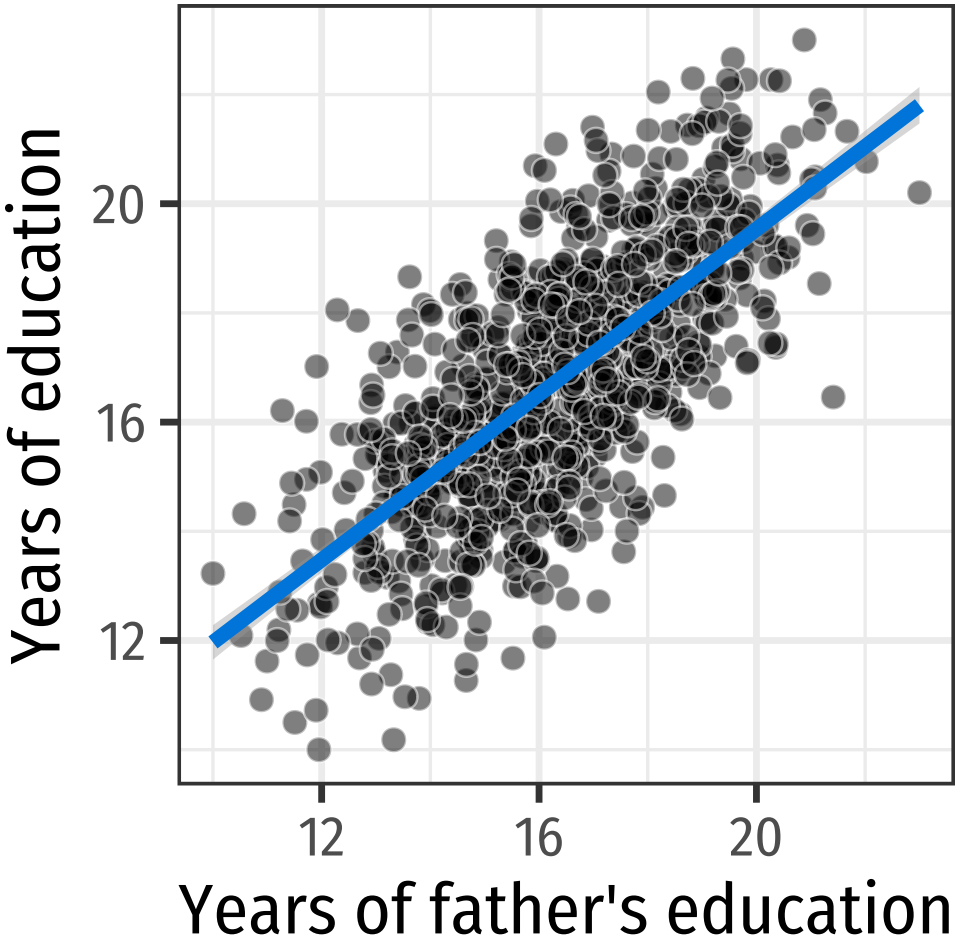
* p < 0.1, ** p < 0.05, *** p < 0.01

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


Ability
Education
Earnings
Father's education
Relevant
Exclusive
Exogenous

Relevancy

Program ~ instrument




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

```r
tidy(first_stage)
```

Clear, significant effect = relevant!

```
## # A tibble: 2 x 5
##   term        estimate std.error statistic   p.value
##   <chr>          <dbl>     <dbl>     <dbl>     <dbl>
## 1 (Intercept)    4.40     0.399       11.0 9.26e- 27
## 2 fathereduc     0.757    0.0243      31.2 1.54e-149
```

```r
glance(first_stage)
```

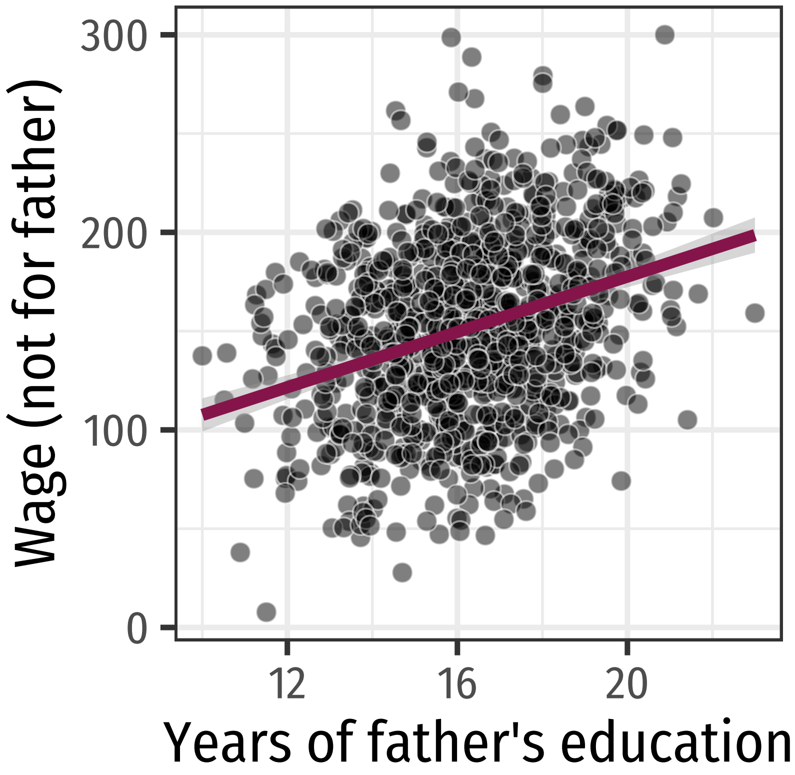
F statistic > 10 = strong instrument

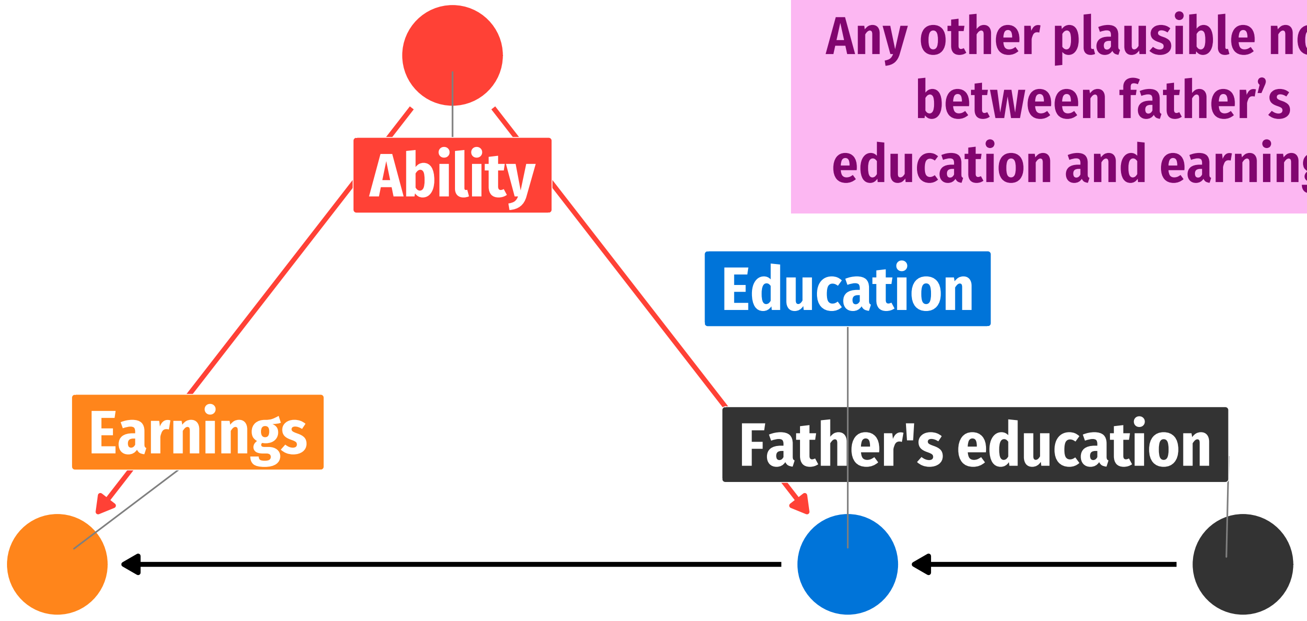
```
## # A tibble: 1 x 11
##   r.squared adj.r.squared sigma statistic   p.value    df logLik   AIC   BIC
##       <dbl>         <dbl> <dbl>     <dbl>     <dbl> <int>  <dbl> <dbl> <dbl>
## 1     0.493         0.493  1.60      972. 1.54e-149     2 -1885. 3777. 3791.
## # ... with 2 more variables: deviance <dbl>, df.residual <int>
```

Exclusion

Does it meet exclusion assumption?

Father's education causes wages only through 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 program/policy variable based on instrument; use *that* to predict outcome

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

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

1st stage

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

2nd stage

Stage 1: Policy ~ instrument

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

```r
tidy(first_stage)
```

```
## # A tibble: 2 x 5
##   term        estimate std.error statistic   p.value
##   <chr>          <dbl>     <dbl>     <dbl>     <dbl>
## 1 (Intercept)    4.40     0.399       11.0 9.26e- 27
## 2 fathereduc     0.757    0.0243      31.2 1.54e-149
```

Use first stage to predict policy

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

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

```r
head(data_with_predictions)
```

```
## # A tibble: 6 x 5
##    wage ability fathereduc  educ educ_hat
##   <dbl>   <dbl>      <dbl> <dbl>    <dbl>
## 1  146.    348.       17.2  18.1     17.4
## 2  148.    181.       14.0  15.8     15.0
## 3  162.    337.       16.0  15.1     16.5
## 4  105.    106.       21.4  16.5     20.6
## 5  168.    302.       16.5  18.8     16.9
## 6  173.    284.       15.4  16.0     16.1
```

educ_hat = 4.4 + (0.757 × 17.2) = 17.4

educ_hat = 4.4 + (0.757 × 16.5) = 16.9

Stage 2: Outcome ~ predicted policy

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

```
## # A tibble: 2 x 5
##   term        estimate std.error statistic  p.value
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)    -3.11    14.4      -0.216 8.29e- 1
## 2 educ_hat        9.25     0.856    10.8   7.49e-26
```

	Outcome = Wage		
	Unadjusted naive	Forbidden model	2SLS IV
(Intercept)	-53.085***	-80.263***	-3.108
	(8.492)	(5.659)	(14.370)
educ	12.240***	9.242***	
	(0.503)	(0.343)	
ability		0.258***	
		(0.007)	
educ_hat			9.252***
			(0.856)
Num.Obs.	1000	1000	1000
R2	0.372	0.726	0.105
Adj.R2	0.371	0.726	0.104

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

	Outcome = Wage		
	Unadjusted naive	Forbidden model	2SLS IV
(Intercept)	-53.085***	-80.263***	-3.108
	(8.492)	(5.659)	
educ	$\boxed{12.240}$***	$\boxed{9.242}$***	
	(0.503)	(0.343)	
ability		0.258***	
		(0.007)	
educ_hat			$\boxed{9.252}$***
			(0.856)
Num.Obs.	1000	1000	1000
R2	0.372	0.726	0.105
Adj.R2	0.371	0.726	0.104

* p < 0.1, ** p < 0.05, *** p < 0.01

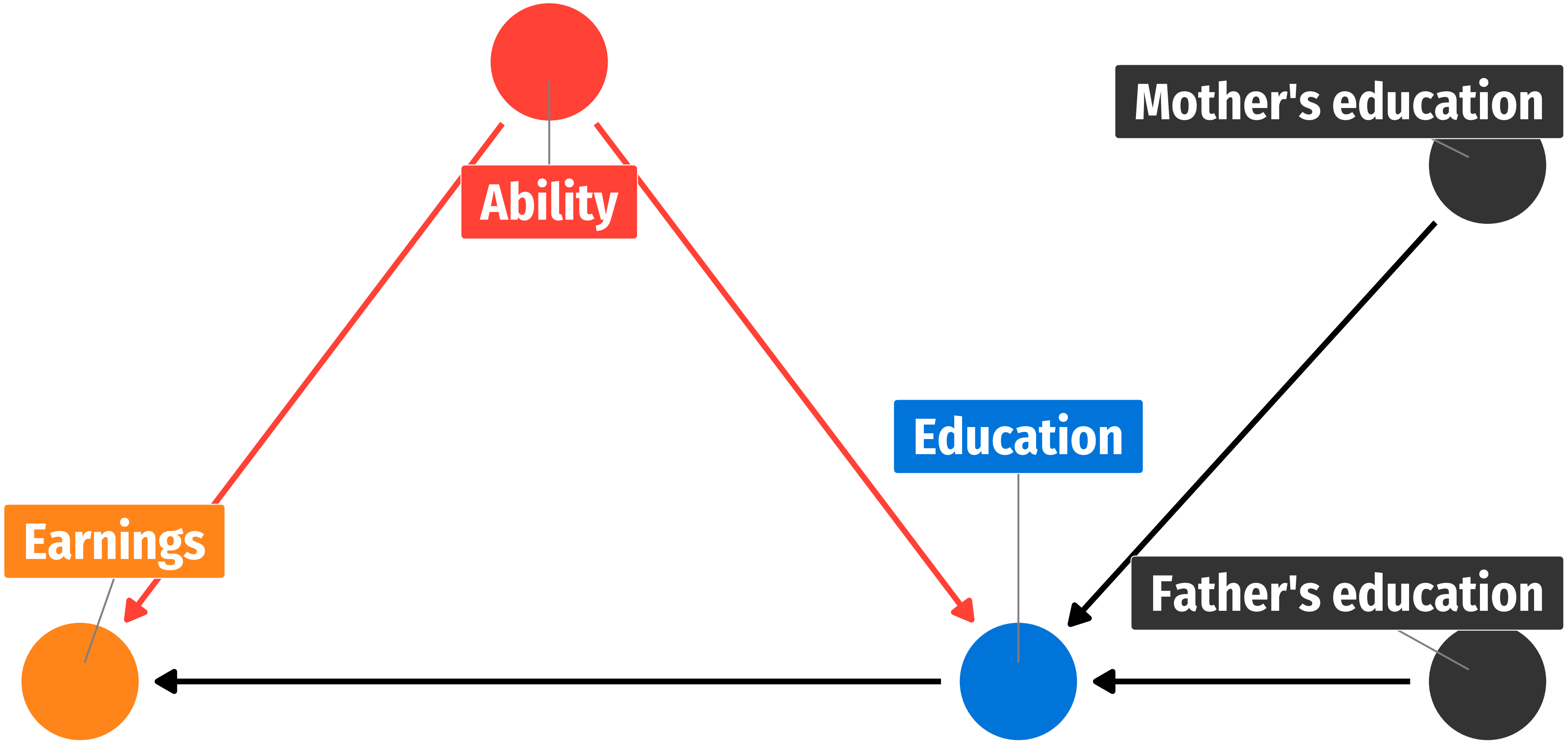
Wrong!

Right, but not measurable

Right!

Multiple instruments

You can use multiple instruments to explain more endogeneity in policy



Multiple instruments

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

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

Other control variables

You can use control variables too!

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

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

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

Faster, more accurate ways to run 2SLS

Running the first stage, getting policy/program 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

`iv_robust()` from the **estimatr** package

Also `ivreg()` in AER and `felm()` in lfe

```r
library(estimatr)

model_estimatr <- iv_robust(wage ~ educ | fathereduc,
                            data = father_education)
```

```r
tidy(model_estimatr)
```

```
##          term  estimate  std.error statistic      p.value   conf.low conf.high
## 1 (Intercept) -3.107953 11.8850068 -0.261502 7.937593e-01 -26.430423  20.21452
## 2        educ  9.251863  0.7078646 13.070102 3.767277e-36   7.862789  10.64094
```

Outcome ~ 2nd stage stuff | 1st stage stuff

	Outcome = Wage			
	Unadjusted naive	Forbidden model	2SLS IV (by hand)	2SLS IV (iv_robust)
(Intercept)	-53.085***	-80.263***	-3.108	-3.108
	(8.492)	(5.659)	(14.370)	(11.885)
educ	12.240***	9.242***		9.252***
	(0.503)	(0.343)		(0.708)
ability		0.258***		
		(0.007)		
educ_hat			9.252***	
			(0.856)	
Num.Obs.	1000	1000	1000	1000
R2	0.372	0.726	0.105	0.350
Adj.R2	0.371	0.726	0.104	0.349

* p < 0.1, ** p < 0.05, *** p < 0.01

	Outcome = Wage			
	Unadjusted naive	Forbidden model	2SLS IV (by hand)	2SLS IV (iv_robust)
(Intercept)	-53.085***	-80.263***	-3.108	-3.108
	(8.492)	(5.659)	(14.370)	(11.885)
educ	12.240***	9.242***		9.252***
	(0.503)	(0.343)		(0.708)
ability		0.258***		
		(0.007		
educ_hat			9.252***	
			(0.856)	
Num.Obs.	1000	1000	1000	1000
R2	0.372	0.726	0.105	0.350
Adj.R2	0.371	0.726	0.104	0.349

* p < 0.1, ** p < 0.05, *** p < 0.01

Wrong!

Right!

IV with R

1: Is the instrument relevant?

Instrument correlated with policy/program; F-statistic in 1st stage is > 10.

2: Does the instrument meet exclusion assumption?

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

3: Is the instrument exogenous?

No arrows going into instrument node in DAG.

4: Run 1st stage

Policy/program ~ instrument

5: Find predicted policy/program values

"Program hat"; plug your data into the first stage model.

6: Run 2nd stage

Outcome ~ program hat

R time!

Treatment effects & compliance

Potential outcomes

$$\delta = (Y|P = 1) - (Y|P = 0)$$

δ = Causal impact of program

P = Program

Y = Outcome

$$\delta = Y_1 - Y_0$$

Fundamental problem of causal inference

$$\delta_i = Y_i^1 - Y_i^0$$

Individual-level effects are impossible to observe

Average treatment effect

Difference between expected value when program is on vs. expected value when program is off

$$ATE = E(Y_1 - Y_0) = E(Y_1) - E(Y_0)$$

Can be found for a whole population, on average

$$\delta = (\bar{Y} | P = 1) - (\bar{Y} | P = 0)$$

Every individual has a treatment/causal effect

ATE = average of all unit-level causal effects

ATE = average effect for the whole population

Other versions of causal effects

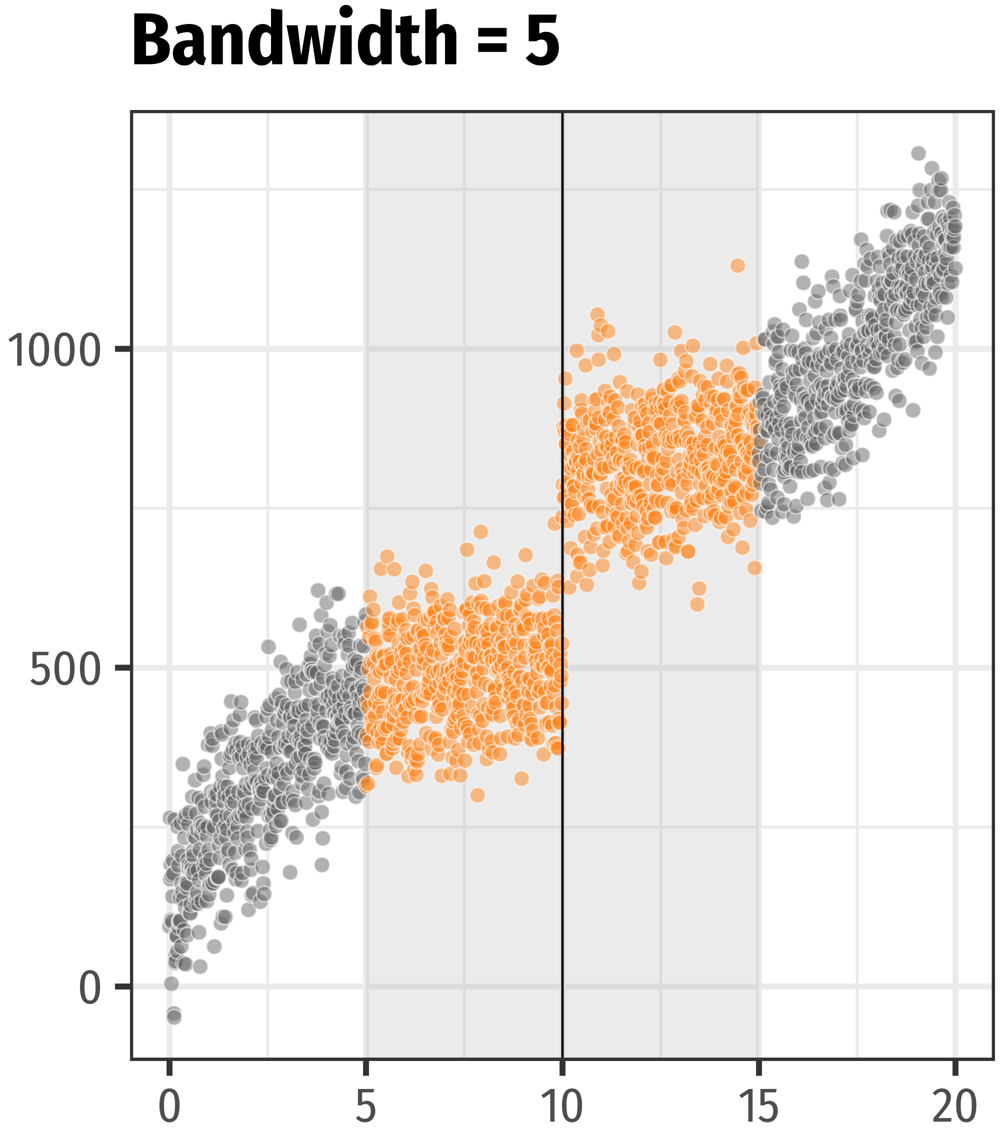
Average treatment on the treated

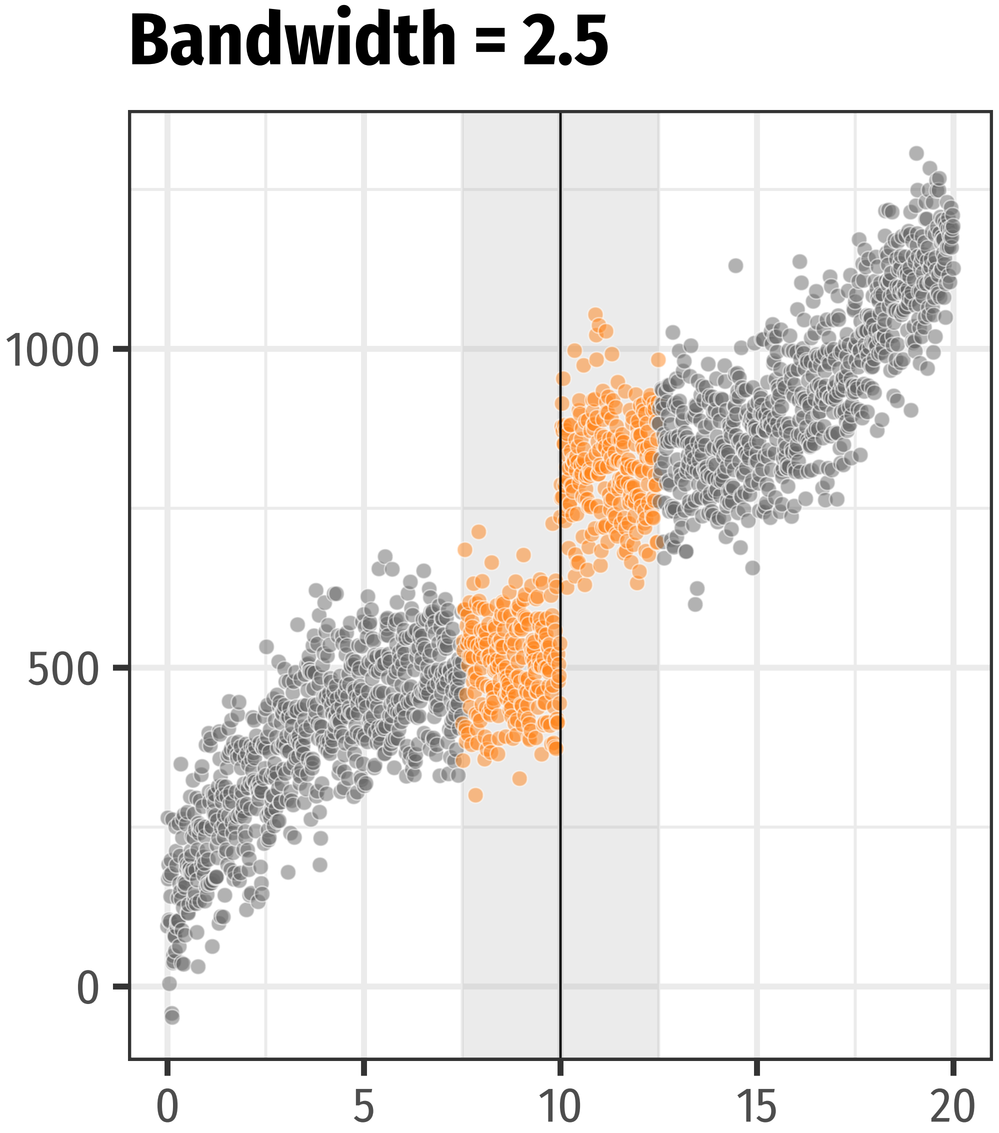
ATT / TOT

Conditional average treatment effect

CATE

Local effects





LATE

Local average treatment effect (LATE) = weighted ATE

Narrower effect; only includes some of the population

Can't make population-level claims with LATE

(But that can be okay)

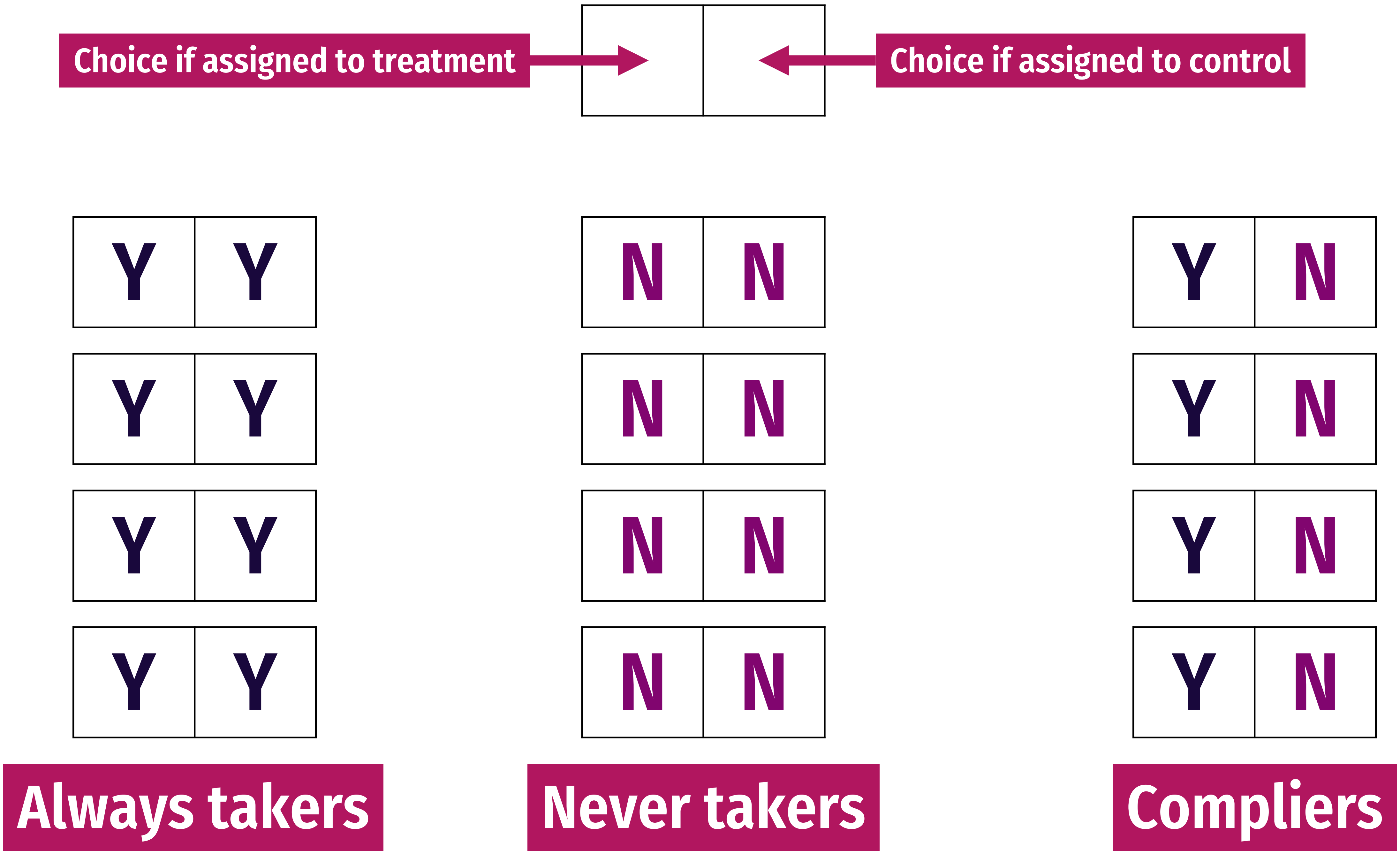
LATE

In RDD, LATE = people in the bandwidth

In RCTs, IVs, etc., LATE = **compliers**

Compliance

Complier	Treatment follows assignment
Always taker	Gets treatment regardless of assignment
Never taker	Rejects treatment regardless of assignment
Defier	Does opposite treatment from assignment



Choice if assigned to treatment
Choice if assigned to control
Y Y
Y Y
Y Y
Y Y
N N
N N
N N
N N
Y N
Y N
Y N
Y N
Always takers
Never takers
Compliers

Ignoring defiers

We can generally assume defiers don't exist

In drug trials this makes sense; can't get access to medicine without being in treatment

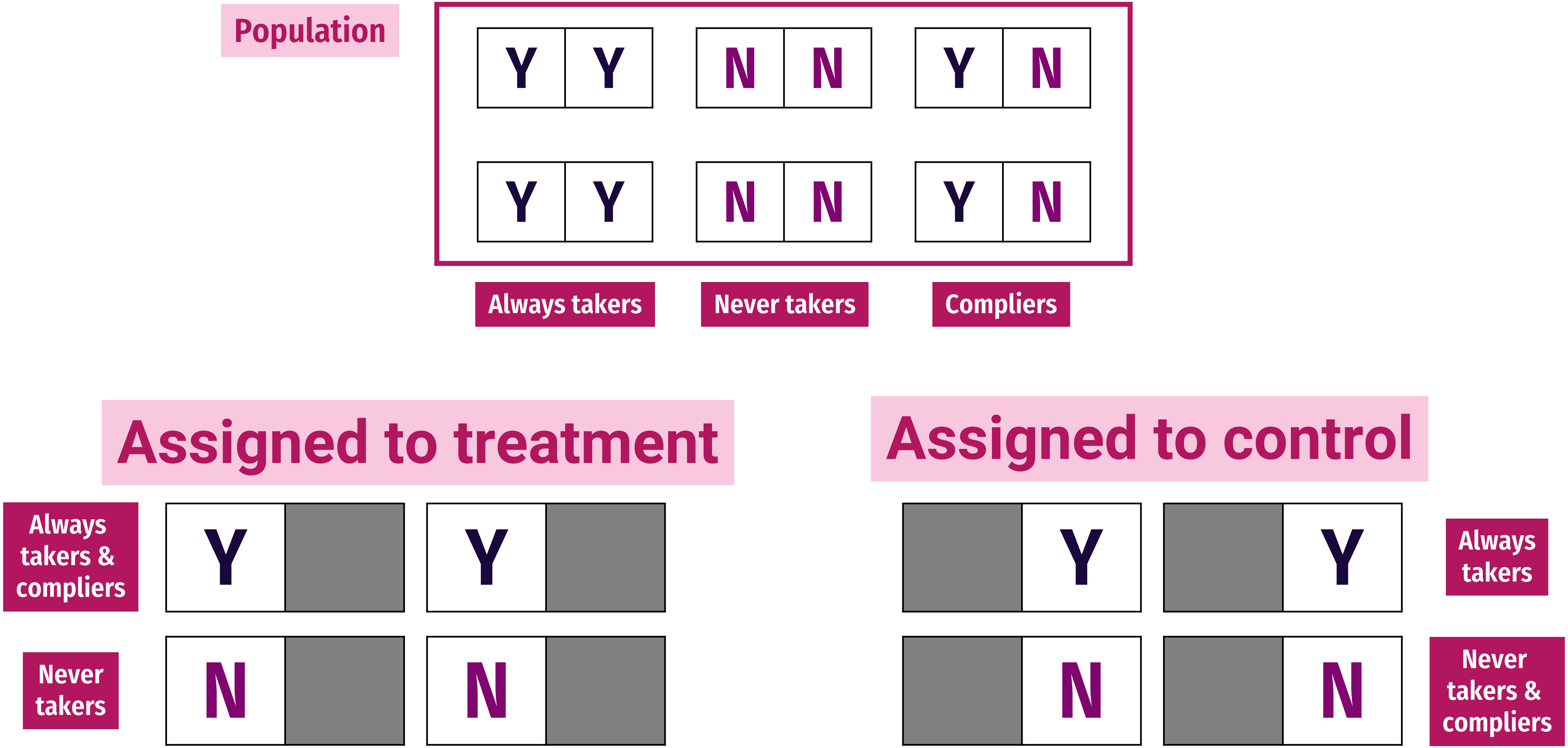
In development, it can make sense; in a bed net RCT, a defier assigned to treatment would have to tear down all existing bed nets out of spite

Ignoring defiers

Monotonicity assumption

Assignment to treatment only has an effect in one direction

Assignment to treatment can only increase—not decrease—your actual chance of treatment


Population
Y Y
N N
Y N
Y Y
N N
Y N
Always takers
Never takers
Compliers
Assigned to treatment
Always takers & compliers
Y
Y
Never takers
N
N
Assigned to control
Y
Y
Always takers
N
N
Never takers & compliers

More causal effects

Intent to treat (ITT)

Effect of assignment (not actual treatment!)

Assigned to treatment

Always takers & compliers	Y	Y
Never takers	N	N

Assigned to control

Y	Y	Always takers
N	N	Never takers & compliers

More causal effects

Complier Average Causal Effect (CACE)

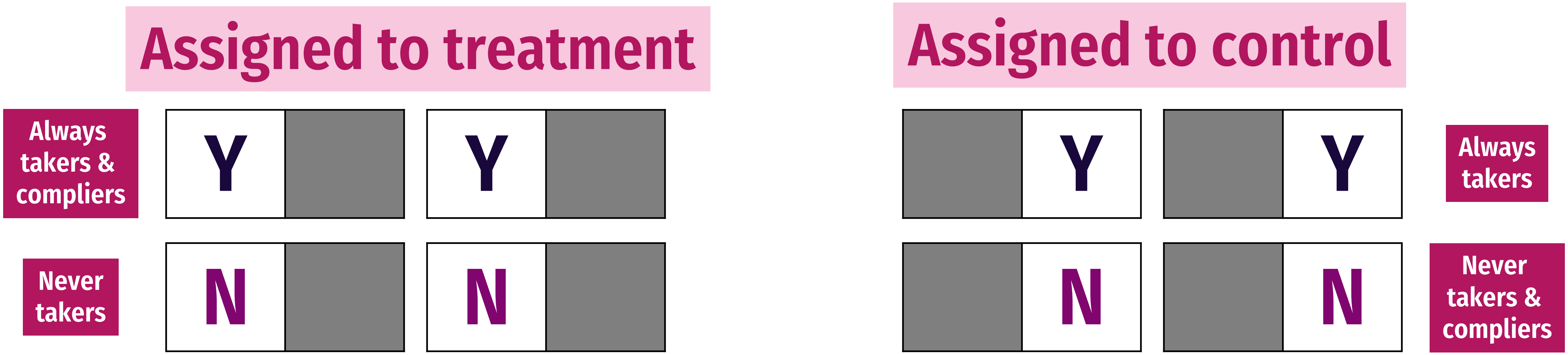
LATE for the compliers

Assigned to treatment

Always takers & compliers	Y		Y	
Never takers	N		N	

Assigned to control

	Y		Y	Always takers
	N		N	Never takers & compliers



$$
\begin{aligned}
\text{ITT} = \; & \pi_{\text{compliers}} \times (\text{T} - \text{C})_{\text{compliers}} + \\
& \pi_{\text{always takers}} \times (\text{T} - \text{C})_{\text{always takers}} + \\
& \pi_{\text{never takers}} \times (\text{T} - \text{C})_{\text{never takers}}
\end{aligned}
$$

$$
\text{ITT} = \pi_{\text{C}}\text{CACE} + \pi_{\text{A}}\text{ATACE} + \pi_{\text{N}}\text{NTACE}
$$

$$\text{ITT} = \pi_{\text{C}}\text{CACE} + \pi_{\text{A}}\text{ATACE} + \pi_{\text{N}}\text{NTACE}$$

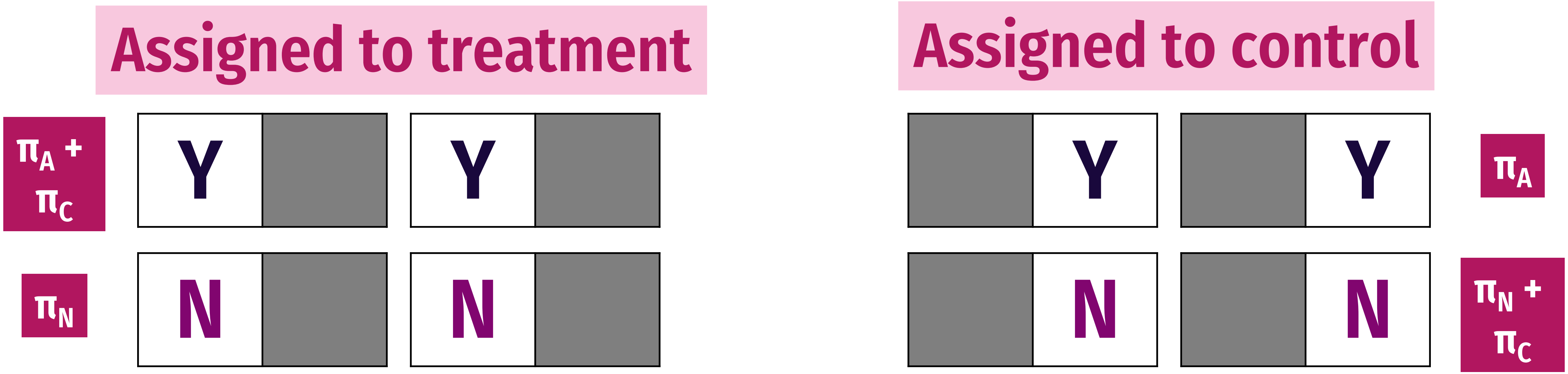
$$\text{ITT} = \pi_{\text{C}}\text{CACE} + \pi_{\text{A}}0 + \pi_{\text{N}}0$$

Exclusion restriction; treatment received is same regardless of assignment

$$\text{ITT} = \pi_{\text{C}}\text{CACE}$$

$$\text{CACE} = \frac{\text{ITT}}{\pi_C}$$

$$\text{CACE} = \frac{\text{ITT}}{\pi_C} \qquad \text{ITT} = (\bar{y}|\text{Treatment}) - (\bar{y}|\text{Control})$$



$$\pi_{\text{A}} + \pi_{\text{C}} = \%\ \text{in treatment and yes}$$

$$\pi_{\text{C}} = \%\ \text{in treatment and yes} - \pi_{\text{A}}$$

$$\text{CACE} = \frac{\text{ITT}}{\pi_C}$$

$$\text{ITT} = (\bar{y}|\text{Treatment}) - (\bar{y}|\text{Control})$$

$$\pi_{\text{C}} = \%\text{ yes in treatment}-$$
$$\%\text{ yes in control}$$

A faster way with 2SLS

LATE for the compliers

If you use assignment to treatment as an instrument, you can find the effect for just compliers

Instrumental variables in general give you the CACE

Example with R