# INSTRUMENTAL VARIABLES I

PMAP 8521: Program Evaluation for Public Service

November 11, 2019

**Fill out your reading report on iCollege!**

# PLAN FOR TODAY

**Endogeneity and exogeneity**

**Instruments**

**Using instruments**

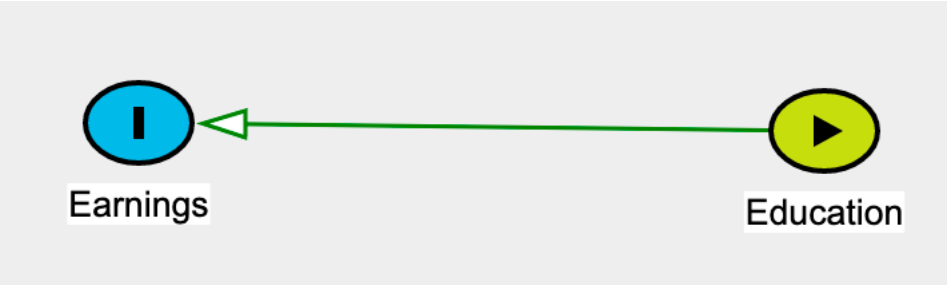
---

**IV regression with R**

# ENDOGENEITY & EXOGENEITY

# OUR FAVORITE QUESTION

**Does education cause higher earnings?**




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

Outcome variable

Policy/program variable

**Would $\beta_1$ in this regression give us the causal effect of the program?**

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

**Omitted variable bias!**

**Selection bias!**

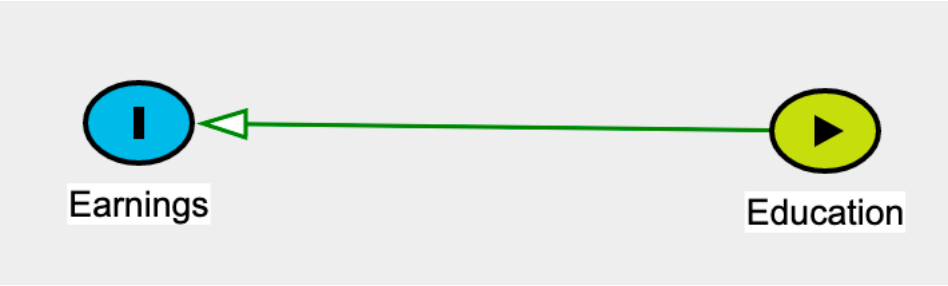
**Endogeneity!**

# TYPES OF VARIATON

## 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

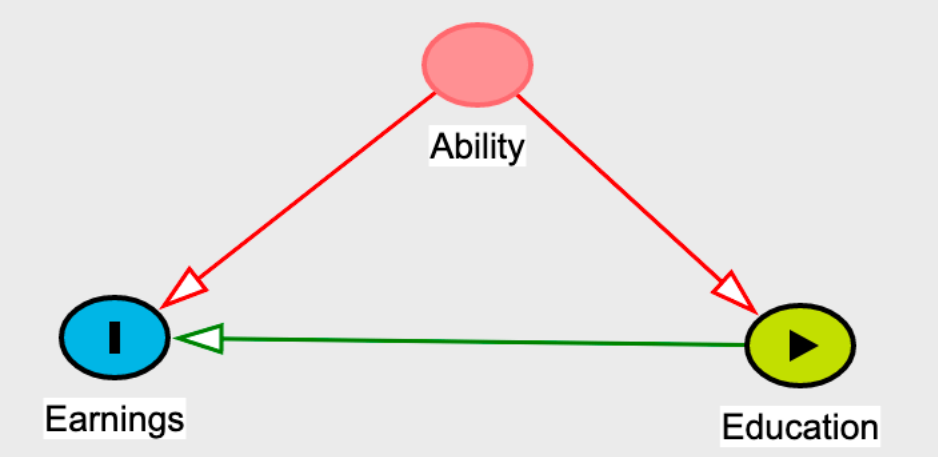


# TYPES OF VARIATON

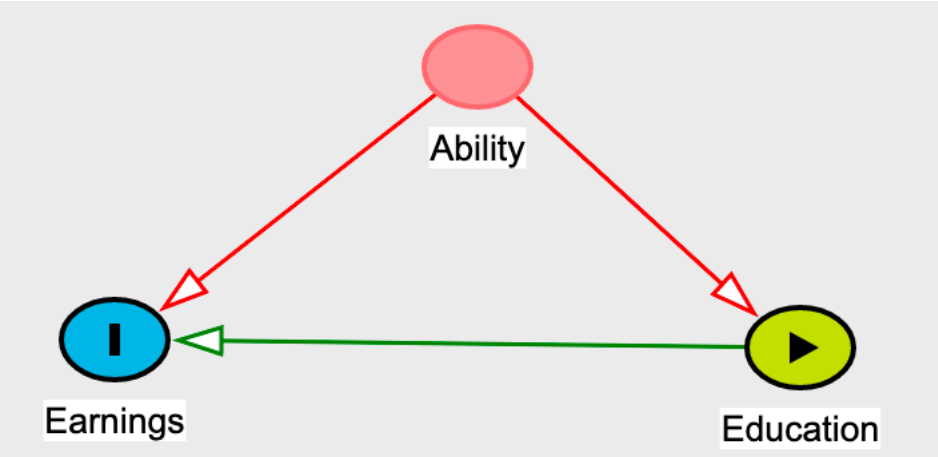
## Endogenous variables

Value is determined by something else in the model

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

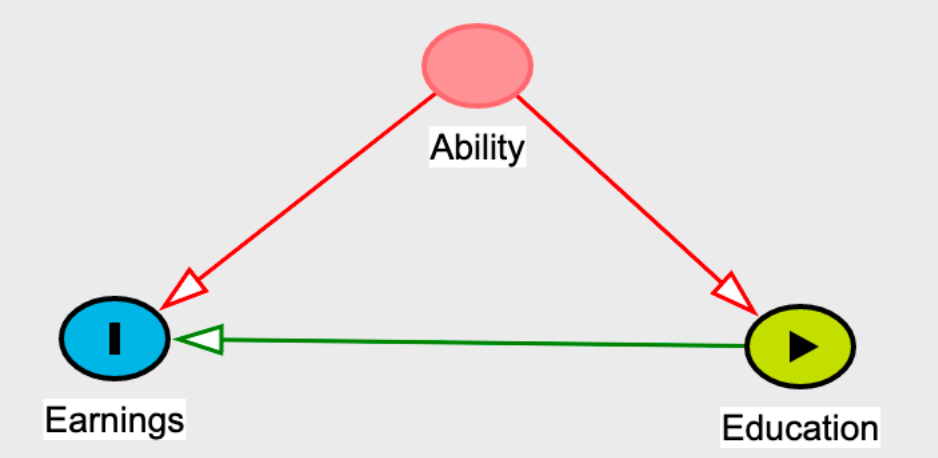


# We'd like education to be exogenous (an outside decision or intervention), but it's not!




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

# How can we fix the endogeneity?

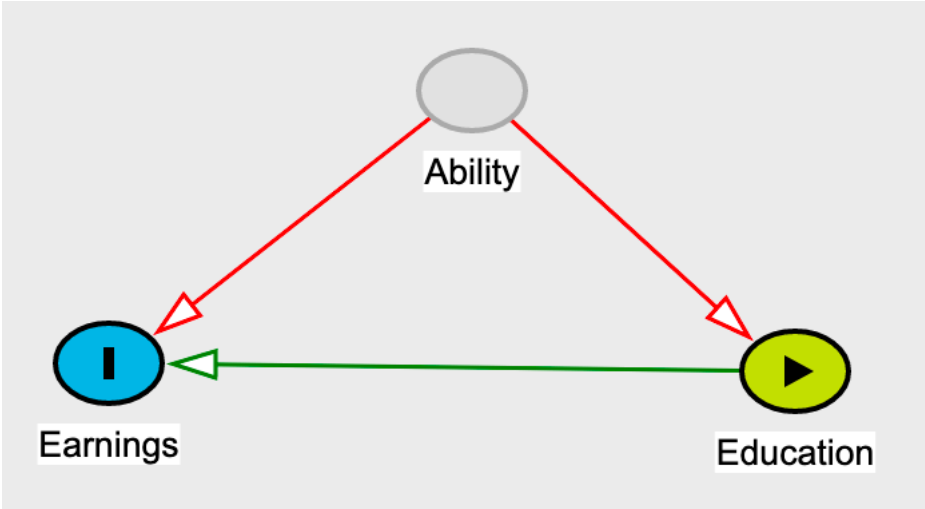



# Close back door and adjust for ability

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

| | *Dependent variable:* | |
|---|---|---|
| | wage | |
| | (1) | (2) |
| educ | 12.240*** | 9.242*** |
| | (0.503) | (0.343) |
| ability | | 0.258*** |
| | | (0.007) |
| Constant | −53.085*** | −80.263*** |
| | (8.492) | (5.659) |
| Observations | 1,000 | 1,000 |
| $R^2$ | 0.372 | 0.726 |
| Adjusted $R^2$ | 0.371 | 0.726 |
| Residual Std. Error | 35.646 (df = 998) | 23.539 (df = 997) |
| F Statistic | 591.469*** (df = 1; 998) | 1,323.969*** (df = 2; 997) |
| *Note:* | | *p<0.1; **p<0.05; ***p<0.01 |

# But what if we can't measure ability?




Unmeasurable!

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

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

Ability is in here

# What would exogenous variation in education look like?

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

# What if we could split education into exogenous and endogenous parts?

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

**How do we isolate the exogenous part of education?**

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

**Use an instrument!**

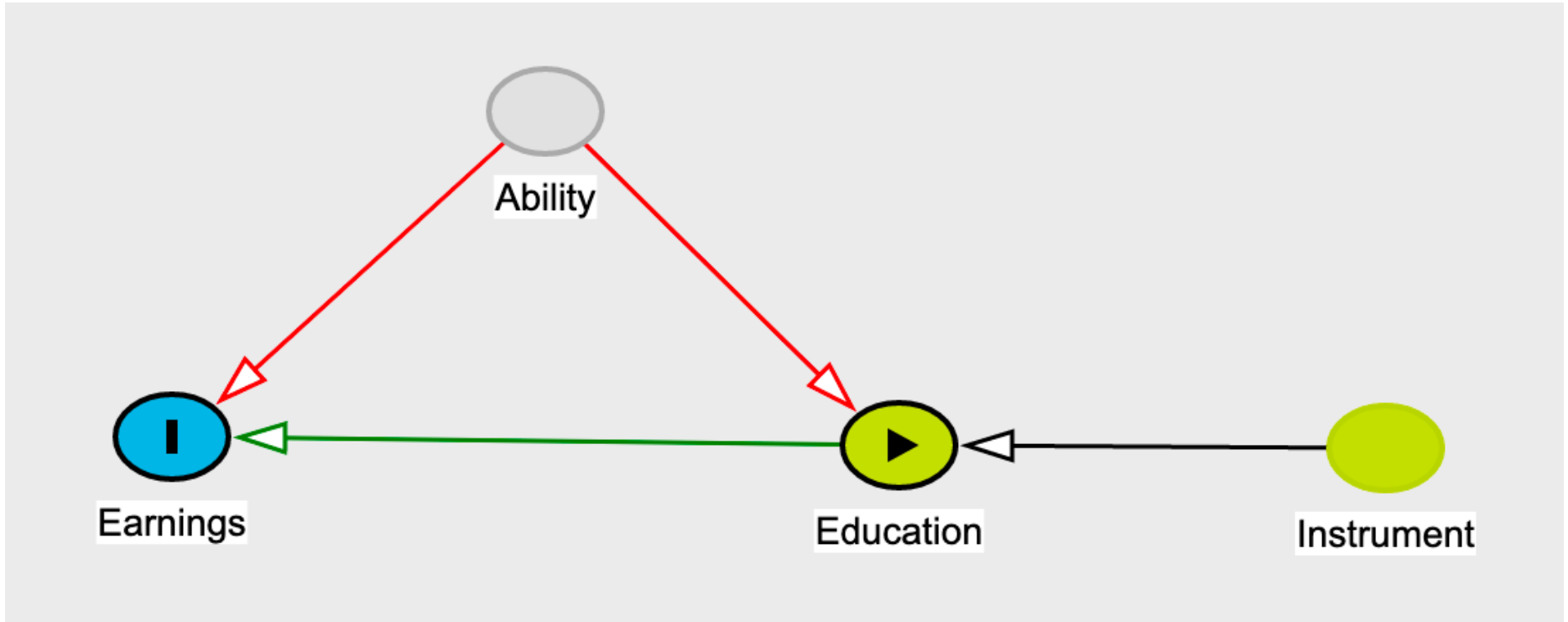
# INSTRUMENTS

# 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**


Ability
Earnings
Education
Instrument

# RELEVANCY

## Instrument causes changes in policy

| | |
|---|---|
| Social security number | **Probably not relevant**<br>Uncorrelated with education |
| 3rd grade test scores | **Potentially relevant**<br>Early grades cause more education |
| Father’s education | **Relevant**<br>Educated parents cause more education |

# EXCLUSION

## Instrument doesn't directly cause outcome

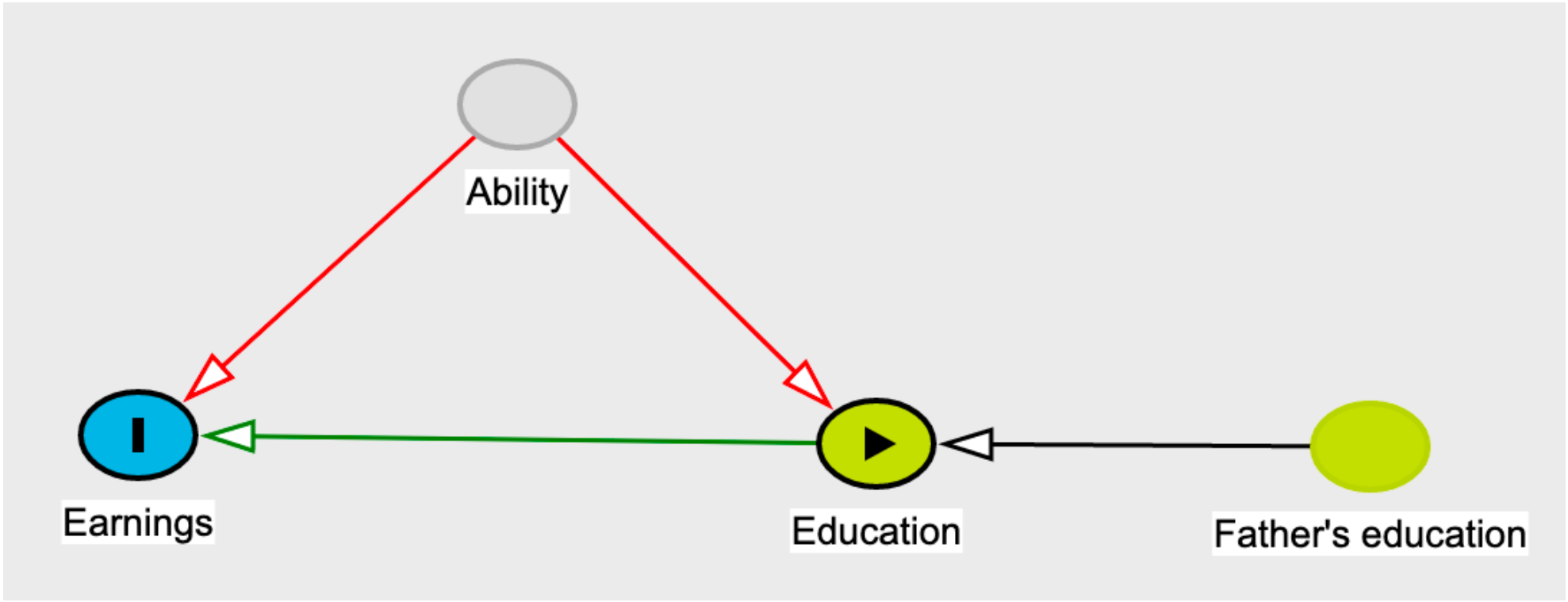
("only through")

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

# EXOGENEITY

**Instrument independent of all other factors; is randomly assigned**

| | |
|---|---|
| **Social security number** | **Exogenous**<br>Unrelated to anything related to education |
| **3rd grade test scores** | **Not exogenous**<br>Grades correlated with other education factors |
| **Father's education** | **Exogenous**<br>Birth to parents is random |




**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 |
| 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 |

# USING INSTRUMENTS

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

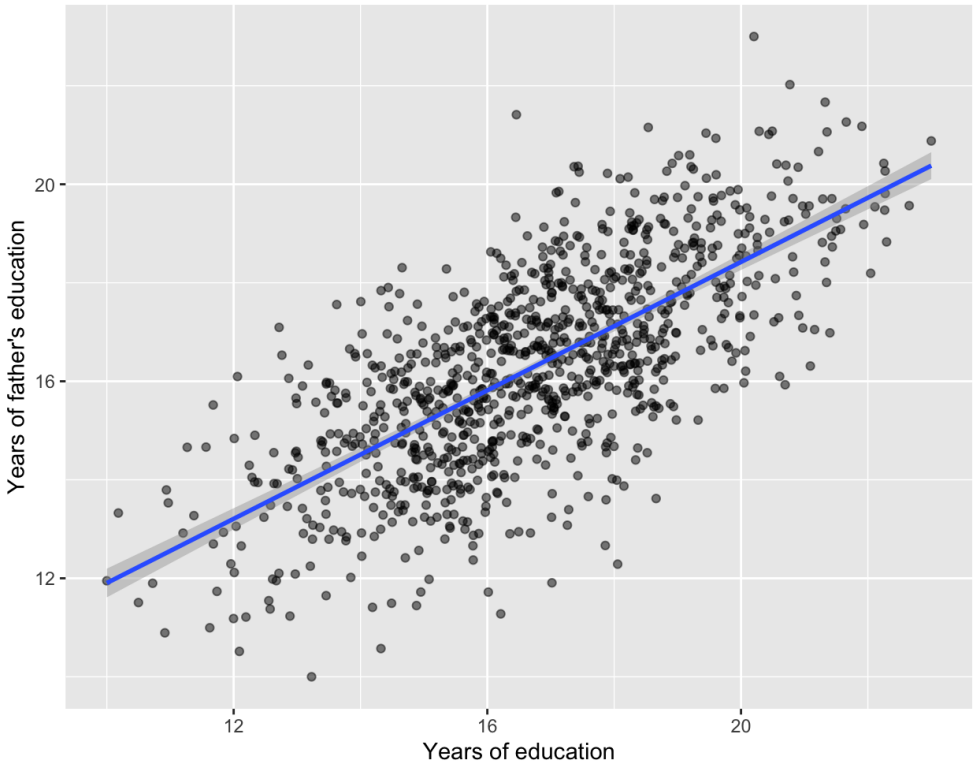
| | *Dependent variable:* | |
|---|---|---|
| | wage | |
| | (1) | (2) |
| educ | 12.240*** | 9.242*** |
| | (0.503) | (0.343) |
| ability | | 0.258*** |
| | | (0.007) |
| Constant | −53.085*** | −80.263*** |
| | (8.492) | (5.659) |
| Observations | 1,000 | 1,000 |
| $R^2$ | 0.372 | 0.726 |
| Adjusted $R^2$ | 0.371 | 0.726 |
| Residual Std. Error | 35.646 (df = 998) | 23.539 (df = 997) |
| F Statistic | 591.469*** (df = 1; 998) | 1,323.969*** (df = 2; 997) |
| *Note:* | | *p<0.1; **p<0.05; ***p<0.01 |

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

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

# RELEVANCY

## Policy ~ instrument




```
model_first <- lm(educ ~ fathereduc, data = dat)
tidy(model_first)
```

| term | estimate | std.error | statistic | p.value |
|---|---|---|---|---|
| (Intercept) | | | | |
| fathereduc | 0.757 | 0.0243 | 31.2 | 1.54e-149 |

**Clear, significant effect = relevant!**

```
glance(model_first)
```

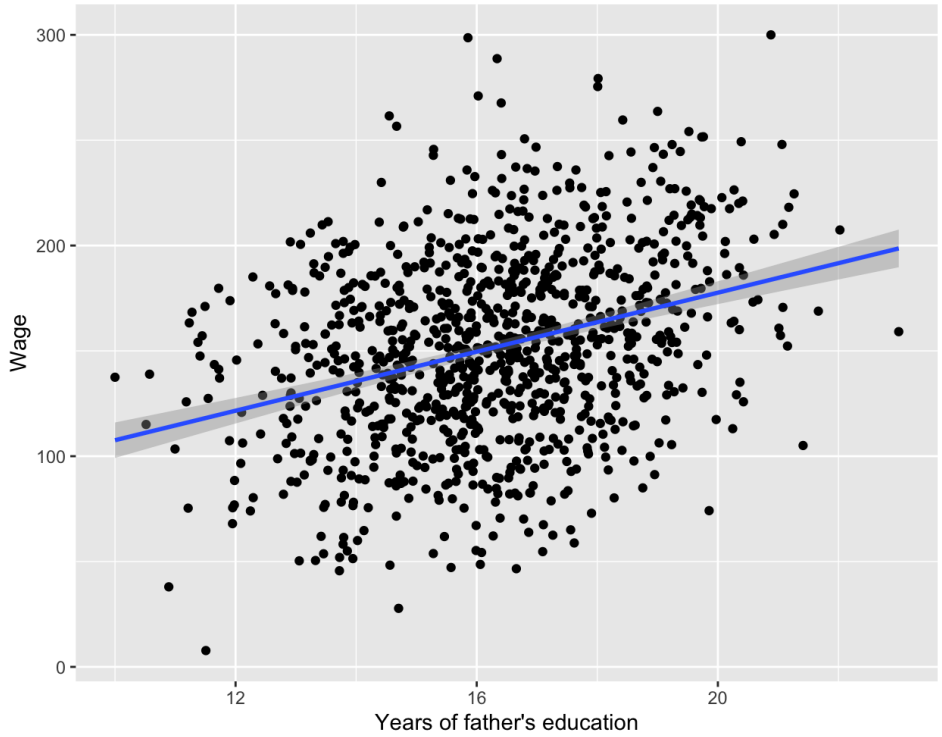
**F statistic > 10 = strong instrument**

| r.squared | adj.r.squared | sigma | statistic | p.value | df | logLik | AIC |
|---|---|---|---|---|---|---|---|
| 0.493 | 0.493 | 1.6 | 972 | 1.54e-149 | 2 | -1.89e+03 | 3.78e+03 |

# EXCLUSION

## Does it meet exclusion assumption?

Father’s education causes wages only through education



# EXOGENEITY

## What would exogeneity of father's education look like?

Compare person A and person B and claim that the differences between them are solely because of their fathers' years of education

# TWO-STAGE LEAST SQUARES (2SLS)

**Find exogenous part of 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

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

| term | estimate | std.error | statistic | p.value |
|---|---|---|---|---|
| (Intercept) | 4.4 | 0.399 | 11 | 9.26e-27 |
| fathereduc | 0.757 | 0.0243 | 31.2 | 1.54e-149 |

# Add predicted education

```
dat_with_predictions <- augment_columns(first_stage, dat)
head(dat_with_predictions)
```

| wage | educ | fathereduc | .fitted | .se.fit | .resid | .hat | .sigma | .cooksd | .std.resid |
|---|---|---|---|---|---|---|---|---|---|
| 146 | 18.1 | 17.2 | 17.4 | 0.0547 | 0.67 | 0.00118 | 1.6 | 0.000104 | 0.42 |
| 148 | 15.8 | 14 | 15 | 0.0752 | 0.862 | 0.00222 | 1.6 | 0.000326 | 0.541 |
| 162 | 15.1 | 16 | 16.5 | 0.051 | -1.4 | 0.00102 | 1.6 | 0.000391 | -0.876 |
| 105 | 16.5 | 21.4 | 20.6 | 0.134 | -4.15 | 0.00708 | 1.59 | 0.0242 | -2.61 |
| 168 | 18.8 | 16.5 | 16.9 | 0.0506 | 1.94 | 0.00101 | 1.6 | 0.000746 | 1.22 |
| 173 | 16 | 15.4 | 16.1 | 0.0546 | -0.0553 | 0.00117 | 1.6 | 7.05e-07 | -0.0347 |

# Stage 2: Outcome ~ predicted policy

```
second_stage <- lm(wage ~ .fitted, data = dat_with_predictions)
tidy(second_stage)
```

| term | estimate | std.error | statistic | p.value |
|---|---|---|---|---|
| (Intercept) | -3.11 | 14.4 | -0.216 | 0.829 |
| .fitted | 9.25 | 0.856 | 10.8 | 7.49e-26 |

| | (1) | (2) | (3) |
|---|---|---|---|
| (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) | |
| .fitted | | | 9.252 *** |
| | | | (0.856) |
| N | 1000 | 1000 | 1000 |
| R2 | 0.372 | 0.726 | 0.105 |
| logLik | -4991.572 | -4576.101 | -5168.868 |
| AIC | 9989.144 | 9160.202 | 10343.735 |

*** p < 0.001; ** p < 0.01; * p < 0.05.

# IV REGRESSION WITH R