

# Instrumental Variables

Gov 51 Section — Week 12 | Card (1995) Application



George

Harvard University

April 15, 2026



**Act I: The Problem**  
**Why Does OLS Fail Here?**

# Maria grew up one mile from a community college — did it change her life?

## The story:

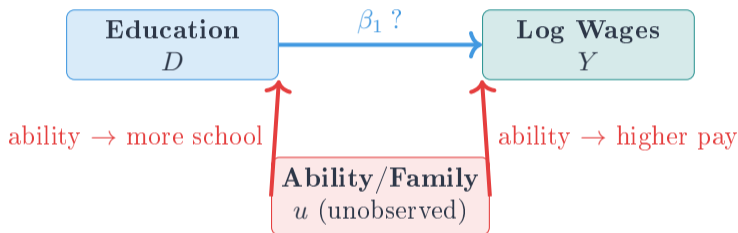
- ▶ Maria's family couldn't afford to move
- ▶ A community college opened near her town
- ▶ She enrolled — not because she was ambitious, but because it was there
- ▶ Ten years later her wages are 15% higher than her neighbor's

**The question:** Did the extra schooling cause those higher wages?

**Card (1995):** *Geographic Variation in College Proximity to Estimate the Return to Schooling*  
Journal of Political Economy

**The problem:** smarter people both go to college *and* earn more — OLS conflates the two

# OLS is biased: unobserved ability drives both education and wages



$$\hat{\beta}_1^{\text{OLS}} = \beta_1 + \underbrace{\text{Cov}(D, u) / \text{Var}(D)}_{\text{ability bias} > 0} \quad \text{— OLS overstates the return}$$



**Act II: The Instrument**  
**A Natural Coin Flip**

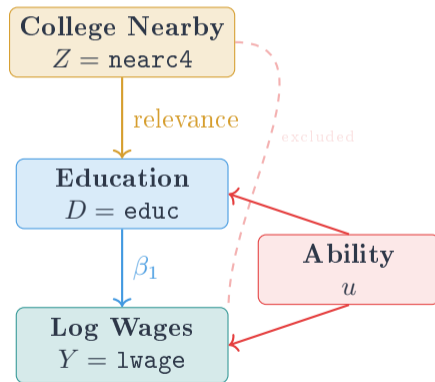
# College proximity is a near-random assignment of educational opportunity

The instrument: `nearc4`

- ▶ 1 if you grew up in a county with a 4-year college
- ▶ 0 otherwise

Why is this a coin flip?

- ▶ College locations were set decades ago
- ▶ Mostly driven by land availability, state charters, railroads
- ▶ You didn't choose where to be born



## Three assumptions — only one is testable

Assumption	Plain English	Testable?
Relevance	Proximity raises education	Yes — run first stage, check $F$
Exclusion	Proximity only affects wages <i>through</i> education	No — requires subject-matter knowledge
Independence	Proximity is as-good-as-random	No — requires thinking about the world

We can check relevance. We must *ar-*  
*gue* for exclusion and independence.

# Exclusion is the hard one: what would violate it?

## Exclusion holds if:

- ▶ College proximity affects wages *only* by raising education
- ▶ No other pathway from  $Z \rightarrow Y$

Card's argument: proximity mainly matters for kids on the margin of attending — not for wages in the local county overall

## What would violate it?

- ▶ Counties with colleges might be richer/more urban
- ▶ Better local labor markets  $\rightarrow$  higher wages directly
- ▶ Card controls for `smsa` and `south` to address this

This is why we include controls. Not to “improve” OLS — but to isolate the instrument's exogenous variation.



**Act III: The First Stage  
Is the Instrument Strong?**

## First stage: does growing up near a college predict years of education?

```
library(estimatr)
library(tidyverse)
card <- read_csv("data/card1995.csv")
# First stage: does college proximity predict education?
fs <- lm_robust(educ ~ nearc4 + exper + expersq + black + south +
               smsa,
               data = card, se_type = "HC2")
tidy(fs) |> filter(term == "nearc4")
```

**Expected result:**  $\text{nearc4} \approx +0.34$  years,  $F \approx 17$ . Growing up near a college adds about 4 months of schooling.

## Plot the first stage: see the relevance assumption visually

```
# First stage visualization
card |>
  group_by(nearc4) |>
  summarise(mean_educ = mean(educ),
            se = sd(educ) / sqrt(n())) |>
  ggplot(aes(x = factor(nearc4,
                      labels = c("No college nearby", "College nearby")),
            y = mean_educ, fill = factor(nearc4))) +
  geom_col(width = 0.5) +
  geom_errorbar(aes(ymin = mean_educ - 1.96*se,
                   ymax = mean_educ + 1.96*se), width = 0.2) +
  labs(x = NULL, y = "Mean years of education") +
  scale_fill_manual(values = c("#CBD3DA", "#F6AD55")) +
  theme_minimal(base_size = 14) +
  theme(legend.position = "none")
```

## The F-test measures joint instrument strength — not individual significance

$$F = \frac{(SSR_R - SSR_U) / q}{SSR_U / (n - k - 1)} \quad \text{where } q = \text{number of instruments}$$

### Three benchmarks:

- ▶  $F > 10$ : Stock-Yogo (2005) rule of thumb (homoskedastic)
- ▶ **Olea-Pflüger effective  $F$** : robust to heteroskedasticity
- ▶ `iv_robust()` reports Olea-Pflüger automatically

**Why does weak instrument matter?**

$$\hat{\beta}^{IV} \text{ bias} \approx \text{OLS bias} \times \frac{1}{F + 1}$$

When  $F \approx 3$ , IV is almost as biased as OLS

# Many weak instruments overfit Stage 1 — garbage in, garbage out

## What goes wrong:

- ▶ Suppose you add 20 instruments, each  $F \approx 2$
- ▶ Stage 1 overfits:  $\hat{D}$  absorbs noise from all 20
- ▶ Stage 2 regresses  $Y$  on noisy  $\hat{D}$
- ▶ Result: estimates close to OLS (not IV)

## One strong instrument beats ten weak ones

- ▶ Card uses a single binary instrument (`nearc4`)
- ▶ Clean, interpretable,  $F > 10$

### Rule of thumb:

$F = 50$ , 1 instrument  $\gg$

$F = 3$ , 20 instruments

More instruments  
 $\neq$  more credibility.  
Better instruments  
= more credibility.



**Act IV: Wald and 2SLS**  
**Harvesting the Exogenous Variation**

## Reduced form: does proximity predict wages directly?

```
# Reduced form: instrument on outcome  
rf <- lm_robust(lwage ~ nearc4 + exper + expersq + black + south +  
  smsa,  
               data = card, se_type = "HC2")  
  
tidy(rf) |> filter(term == "nearc4")
```

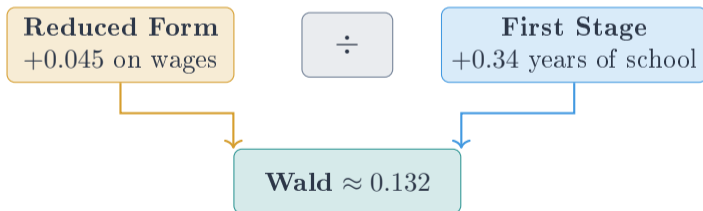
**Reduced form coefficient:**  $\approx +0.045$

Being near a college raises wages by 4.5%

Reduced form = what IV is credibly estimating. If RF is zero, 2SLS will be zero too.

## Wald by hand: reduced form divided by first stage

$$\hat{\beta}_1^{\text{Wald}} = \frac{\text{Cov}(Y, Z)}{\text{Cov}(D, Z)} = \frac{\text{Reduced form}}{\text{First stage}} \approx \frac{0.045}{0.34} \approx \mathbf{0.132}$$



## 2SLS with `iv_robust()`: the formula is one line

```
# 2SLS: education instrumented by college proximity
fit_iv <- iv_robust(
  lwage ~ educ + exper + expersq + black + south + smsa |
        nearc4 + exper + expersq + black + south + smsa,
  data = card, se_type = "HC2"
)

tidy(fit_iv, conf.int = TRUE)
summary(fit_iv) # reports Olea-Pfluger effective F
```

**Formula syntax:** `outcome ~ endogenous  
+ controls | instrument + controls`  
Controls appear on BOTH sides of the pipe.

## OLS vs. 2SLS: the IV estimate is larger, not smaller

```
# OLS for comparison
fit_ols <- lm_robust(lwage ~ educ + exper + expersq + black + south
  + smsa,
                    data = card, se_type = "HC2")

# Compare coefficients on education
bind_rows(
  tidy(fit_ols) |> filter(term=="educ") |> mutate(model="OLS"),
  tidy(fit_iv)  |> filter(term=="educ") |> mutate(model="2SLS")
) |> select(model, estimate, std.error, p.value)
```

Model	Return to schooling	SE
OLS	$\approx 0.074$	0.004
2SLS	$\approx 0.132$	0.049

# The OVB formula predicted $IV < OLS$ — Card's result surprised labor economists

What OVB predicts:

$$\hat{\beta}^{OLS} = \beta_1 + \underbrace{\frac{\text{Cov}(\text{ability}, D)}{\text{Var}(D)}}_{> 0} \cdot \lambda$$

- ▶ Ability  $\uparrow \Rightarrow$  more school ( $Cov > 0$ )
- ▶ Ability  $\uparrow \Rightarrow$  higher wages ( $\lambda > 0$ )
- ▶ OLS biased *up*  $\Rightarrow$  IV should be *smaller*

What Card found: 2SLS (0.132)  $>$  OLS (0.073)

Two explanations:

- ▶ *Measurement error*: noisy self-reports  $\Rightarrow$  attenuation  $\Rightarrow$  OLS too small
- ▶ **Card's preferred**: instrument selects *credit-constrained* students who may have unusually *high* returns to schooling

The instrument doesn't just identify  $\beta$  — it identifies  $\beta$  for the compliers. If they have higher returns than average, LATE  $>$  ATE.



**Act V: What Did We Just Estimate?**  
**LATE vs. the Population ATE**

## Normal estimands need no identification; causal estimands do

	Estimand	Estimator	ID needed?
Statistics world	$\mu = E[Y_i]$	$\bar{Y}$	No
Causal (ATE)	$E[Y_{1i} - Y_{0i}]$	$\bar{Y}_{D=1} - \bar{Y}_{D=0}$	Yes
Causal (LATE)	$E[Y_1 - Y_0 \mid \text{complier}]$	Wald / 2SLS	Yes

Without identification assumptions, SDO  $\neq$  ATE — even in infinite data. LLN doesn't help if the estimand is wrong.

## The 2SLS estimate is a LATE — not an ATE

**LATE** = Local Average Treatment Effect =  $E[Y_1 - Y_0 \mid \text{complier}]$

### Three types of people:

- ▶ **Always-takers:** got education regardless of proximity
- ▶ **Never-takers:** skipped college regardless of proximity
- ▶ **Compliers:** went to college *because* of proximity

2SLS estimates the effect *only* for compliers.

### Who are Card's compliers?

Kids who *would not have attended* if the college wasn't nearby — but *did attend* because it was close.

Likely: lower-income, first-generation students.

# The LATE may be bigger or smaller than the ATE — it depends on the compliers

## Card (1995) interpretation:

- ▶ Compliers are marginal students — on the fence about college
- ▶ For them, the return to schooling ( $\approx 13\%$ ) may be *higher* than the population ATE
- ▶ High returns for marginal students is consistent with heterogeneous treatment effects

### **Policy implication:**

LATE answers: "*What does more schooling do for the students who wouldn't otherwise have gone?*"

That's exactly the right question for college expansion policy.



**Act VI: Code Workshop**  
**Your Turn**

## Full analysis in R — step by step

```
library(estimatr); library(tidyverse)
card <- read_csv("data/card1995.csv")

# Step 1: First stage
fs <- lm_robust(educ ~ nearc4 + exper + expersq + black + south +
  smsa,
  data = card, se_type = "HC2")

# Step 2: 2SLS (formula: outcome ~ endog + controls | instrument +
  controls)
fit_iv <- iv_robust(
  lwage ~ educ + exper + expersq + black + south + smsa |
  nearc4 + exper + expersq + black + south + smsa,
  data = card, se_type = "HC2")
summary(fit_iv) # reports Olea-Pfluger effective F
```

# Anderson-Rubin CI: a better confidence interval when the instrument is borderline

## The problem with standard CIs:

- ▶ Standard CI =  $\hat{\beta}_{IV} \pm 1.96 \times \widehat{SE}$
- ▶ SE assumes a strong first stage
- ▶ When  $F$  is small, the CI is too narrow — a false sense of precision

## Anderson-Rubin approach:

- ▶ For each candidate value  $\beta_0$ , form  $\tilde{e}_i = Y_i - \beta_0 D_i$
- ▶ If  $\beta = \beta_0$  truly, then  $\tilde{e}$  should be uncorrelated with  $Z$
- ▶ AR CI = all  $\beta_0$  we cannot reject

### Card (1995) result:

Standard CI: [0.04, 0.23]

AR CI: [0.04, 0.26]

AR is wider because it is honest about instrument uncertainty.

**Rule:** When  $F < 20$ , always report AR CI alongside standard.

## Anderson-Rubin CI in R: grid inversion

```
library(wooldridge); data(card)

# For each candidate beta0: test Cov(Y - beta0*D, Z) = 0
beta_grid <- seq(-0.1, 0.5, by = 0.001)
pvals <- sapply(beta_grid, function(b0) {
  e_tilde <- card$lwage - b0 * card$educ
  m <- lm(e_tilde ~ nearc4 + exper + expersq + black + south + smsa,
          data = card)
  coef(summary(m))["nearc4", 4] # p-value on Z
})
ar_ci <- range(beta_grid[pvals > 0.05]) # [0.039, 0.261]
```

AR CI is valid even when  $F$  is small. Standard CI requires  $F \gg 10$ .

## Section exercises

### Warm-up (5 min each)

1. **First stage.** Run the first stage. What is the coefficient on `nearc4`? Is  $F > 10$ ?
2. **Wald by hand.** Run the reduced form. Divide RF by FS. Does it match `iv_robust()`?


### Core (8 min each)

3. **Exclusion.** Name one concrete way the exclusion restriction could fail. How does Card address it?
4. **LATE.** Write one sentence interpreting the 2SLS coefficient: (i) what effect, (ii) for whom, (iii) what drives the variation.

### Challenge (if time)

5. Add `nearc2` (2-year college) as a second instrument. Does  $F$  increase? Does the 2SLS estimate change? Why might many weak instruments be worse than one strong one?

Run the `ggplot` code from the previous slide to visualize the first stage before attempting exercise 1.



IV finds causal effects where OLS can't.  
But it estimates a LATE  
— for compliers only.  
The instrument defines the question.