

Mostly Harmless Econometrics

Session 3: Regression and Causality (MHE §3.2)

CIA → Linear causal model → Omitted variables bias → Bad control

Kaiwen Luo

Washington University in St.Louis

October 16, 2026

Today's Question: When Is a Regression Causal?

Section 3.1 showed that regression is the best linear approximation to the CEF. That is a statistical fact. Today's question is a causal one.

Angrist–Pischke's answer, in one line

A regression is causal when the CEF it approximates is causal.

- 1 **CIA:** when does the CEF $\mathbb{E}[Y_i | X_i, s_i]$ have a causal meaning?
- 2 **Credibility:** in which settings is the CIA believable?
- 3 **Linearity:** what extra assumptions turn the causal CEF into $\alpha + \rho s_i + X_i' \gamma$?
- 4 **OVB:** how do short and long regression coefficients relate?
- 5 **Choosing controls:** the job of X_i is to make the CIA credible.
- 6 **Bad control:** two ways that adding a control makes things worse.

Start with a Binary Treatment: College

Let $c_i \in \{0, 1\}$ indicate college attendance, with potential earnings

$$\text{Potential outcome} = \begin{cases} Y_{1i} & \text{if } c_i = 1 \\ Y_{0i} & \text{if } c_i = 0 \end{cases} \implies Y_i = Y_{0i} + (Y_{1i} - Y_{0i}) c_i.$$

We see only one of Y_{1i}, Y_{0i} . A naive comparison of means gives

$$\underbrace{\mathbb{E}[Y_i \mid c_i = 1] - \mathbb{E}[Y_i \mid c_i = 0]}_{\text{observed difference}} = \underbrace{\mathbb{E}[Y_{1i} - Y_{0i} \mid c_i = 1]}_{\text{ATT}} + \underbrace{\mathbb{E}[Y_{0i} \mid c_i = 1] - \mathbb{E}[Y_{0i} \mid c_i = 0]}_{\text{selection bias}}. \quad (3.2.1)$$

Why the naive comparison fails

College-goers would probably have earned more *even without college*: $\mathbb{E}[Y_{0i} \mid c_i = 1] > \mathbb{E}[Y_{0i} \mid c_i = 0]$. Selection bias is positive, so the raw gap **overstates** the return to college.

The Conditional Independence Assumption (CIA)

CIA (selection on observables)

$$\{Y_{0i}, Y_{1i}\} \perp\!\!\!\perp c_i \mid X_i. \quad (3.2.2)$$

Among people with the same X_i , college attendance is unrelated to potential earnings.

Compare people *with the same* X_i :

$$\begin{aligned} \mathbb{E}[Y_i \mid X_i, c_i = 1] - \mathbb{E}[Y_i \mid X_i, c_i = 0] &= \mathbb{E}[Y_{1i} \mid X_i, c_i = 1] - \mathbb{E}[Y_{0i} \mid X_i, c_i = 0] \\ &= \mathbb{E}[Y_{1i} \mid X_i] - \mathbb{E}[Y_{0i} \mid X_i] \quad (\text{CIA}) \\ &= \boxed{\mathbb{E}[Y_{1i} - Y_{0i} \mid X_i].} \end{aligned}$$

Within an X_i -cell, selection bias is zero: $\mathbb{E}[Y_{0i} \mid X_i, c_i = 1] - \mathbb{E}[Y_{0i} \mid X_i, c_i = 0] = 0$.

Many Treatment Levels: Years of Schooling s_i

Each person has a whole **potential earnings function**:

$$Y_{si} \equiv f_i(s) \quad = \text{what } i \text{ would earn with } s \text{ years of schooling.}$$

College vs. no college is the special case $Y_{0i} = f_i(12)$, $Y_{1i} = f_i(16)$. The data reveal only $Y_i = f_i(s_i)$.

CIA for a multivalued treatment

$$Y_{si} \perp\!\!\!\perp s_i \mid X_i \quad \text{for all } s.$$

Then, for every value s ,

$$\mathbb{E}[Y_i \mid X_i, s_i = s] = \mathbb{E}[f_i(s) \mid X_i, s_i = s] = \boxed{\mathbb{E}[f_i(s) \mid X_i]}.$$

So the CEF is the average causal response function. Differences along it are causal:

$$\mathbb{E}[Y_i \mid X_i, s_i = 12] - \mathbb{E}[Y_i \mid X_i, s_i = 11] = \mathbb{E}[f_i(12) - f_i(11) \mid X_i].$$

From Cell-by-Cell Effects to One Number

The CIA gives one causal effect for *each value* of X_i . Summarize them by iterating expectations.

Average effect (ATE)

$$\begin{aligned}\mathbb{E}\{\mathbb{E}[Y_i \mid X_i, s_i=12] - \mathbb{E}[Y_i \mid X_i, s_i=11]\} \\ = \mathbb{E}[f_i(12) - f_i(11)]\end{aligned}$$

Weights: distribution of X_i in the population.

Effect on the treated (ATT)

$$\begin{aligned}\mathbb{E}\{\mathbb{E}[Y_i \mid X_i, s_i=12] - \mathbb{E}[Y_i \mid X_i, s_i=11] \mid s_i=12\} \\ = \mathbb{E}[f_i(12) - f_i(11) \mid s_i = 12]\end{aligned}$$

Weights: distribution of X_i among graduates.

The empirical counterpart is a **matching estimator**: compare within X -cells, then average.

- It takes two steps (match, then average), and standard errors are not automatic.
- With many values of s_i there is a separate effect for each increment $s - 1 \rightarrow s$.

⇒ Regression: an automatic way to turn the CIA into one causal summary.

The CIA Is a Claim About How s_i Was Assigned

CIA = “conditional on X_i , s_i is *as good as randomly assigned*.” It is credible when we know **what determines** s_i and we observe it.

Scenario	Example	Why the CIA is plausible
1. Random assignment given X_i	Black et al. (2003): mandatory retraining for UI claimants; Tennessee STAR	Workers sorted into groups by observed traits and job history; oversubscribed groups assigned training <i>by lottery</i> . Given the group variables, training is random by design.
2. Institutional knowledge of selection	Angrist (1998): voluntary military service	The military screens applicants on age, schooling, AFQT score. Given these, remaining variation comes from qualified applicants who drop out at the last minute.

Even then, not guaranteed: why a qualified applicant fails to enlist could be related to earnings potential.

Schooling: The Hard Case

- There is usually **no lottery** for finishing high school or going to college.
- Closest analog: an *encouragement* experiment among people with similar ability and family background.
 - ▶ Education Maintenance Allowance (Dearden et al. 2003): pays UK students in some areas to stay in school.
- Otherwise we must argue that X_i captures everything that drives both schooling and earnings:

$$X_i = \{\text{ability, family background, motivation, } \dots\}.$$

The question to ask of any control set

Besides X_i , what else makes people with different s_i different in their potential earnings?

If the answer is “nothing important,” the CIA is credible.

Step 0: The CIA Identifies the CEF but Not Its Shape

Define the CEF $m(s, x) = \mathbb{E}[Y_i \mid s_i = s, X_i = x]$. Identically,

$$Y_i = m(s_i, X_i) + \varepsilon_i, \quad \mathbb{E}[\varepsilon_i \mid s_i, X_i] = 0.$$

This is a decomposition (Theorem 3.1.1), not an economic model.

Under the CIA (plus overlap: s is in the support of $s_i \mid X_i = x$)

$$m(s, x) = \mathbb{E}[f_i(s) \mid X_i = x].$$

The marginal causal effect is $\partial m(s, x) / \partial s$. It may vary with the schooling level s *and* with characteristics x .

The CIA tells us what $m(s, x)$ means; it does not tell us what it looks like.

The regression $Y_i = \alpha + \rho s_i + X_i' \gamma + v_i$ needs three more restrictions.

A1: Constant Marginal Response in s

Assumption A1: no curvature in schooling (given X)

$$\frac{\partial m(s, x)}{\partial s} = \rho(x).$$

Within a group x , the 11th year pays the same as the 16th.

Integrating along a connected treatment interval containing zero:

$$m(s, x) - m(0, x) = \int_0^s \rho(x) dt = s\rho(x).$$

Hence, writing $g(x) = m(0, x)$,

$$m(s, x) = g(x) + s\rho(x).$$

Economic content: linear dose response *within* a group; returns may still differ *across* groups.

A2: Same Average Return Across Groups

Assumption A2: common conditional average marginal effect

$$\rho(x) = \rho \quad \text{for all relevant } x.$$

Combining A1 and A2,

$$m(s, x) = \rho s + g(x) \implies Y_i = \rho s_i + g(X_i) + \varepsilon_i, \quad \mathbb{E}[\varepsilon_i \mid s_i, X_i] = 0.$$

Partially linear regression (PLR)

The causal effect of s is constant, but the baseline relationship $g(x)$ is left unrestricted.

What “homogeneous” means here: for any supported s_1, s_0 ,

$$\mathbb{E}[f_i(s_1) - f_i(s_0) \mid X_i = x] = \rho(s_1 - s_0).$$

This restricts *average* effects across observed groups, not every individual's $f_i(\cdot)$.

A3: A Linear Baseline in X

Assumption A3: additive and linear baseline

$$g(x) = \alpha + x'\gamma.$$

No nonlinear X terms, no interactions in the baseline.

Substituting into the PLR model:

$$\mathbb{E}[Y_i \mid s_i, X_i] = \alpha + \rho s_i + X_i' \gamma \iff Y_i = \alpha + \rho s_i + X_i' \gamma + v_i, \quad \mathbb{E}[v_i \mid s_i, X_i] = 0.$$

ρ

Causal: needs the CIA + A1 + A2.

γ

Not causal: the CIA is about s_i , not about X_i . γ just describes how baseline earnings vary with X_i .

Angrist–Pischke's Version: A Linear Constant-Effects Model

MHE imposes the same structure in one step, at the *individual* level:

$$f_i(s) = \alpha + \rho s + \eta_i, \quad \mathbb{E}[\eta_i] = 0. \quad (3.2.7)$$

Everyone has the same slope ρ ($A1 + A2$, and more); η_i collects unobserved determinants of potential earnings. Then decompose η_i by its regression on X_i (this is $A3$):

$$\eta_i = X_i' \gamma + v_i, \quad \mathbb{E}[\eta_i | X_i] = X_i' \gamma.$$

By the CIA,

$$\mathbb{E}[f_i(s) | X_i, s_i] = \mathbb{E}[f_i(s) | X_i] = \alpha + \rho s + X_i' \gamma.$$

The causal regression

$$Y_i = \alpha + \rho s_i + X_i' \gamma + v_i. \quad (3.2.9)$$

v_i is uncorrelated with s_i and X_i , so the population regression coefficient on s_i is the causal effect ρ .

Key assumption: X_i is the *only* reason why η_i and s_i are correlated (Barnow, Cain, and Goldberger, 1981).

Summary: The Assumption Ladder

Step	Restriction	What we obtain
Identity	$m(s, x) = \mathbb{E}[Y_i \mid s_i = s, X_i = x]$	A prediction function, nothing causal
Identification	CIA + overlap	$m(s, x) = \mathbb{E}[f_i(s) \mid X_i = x]$
A1	Return does not vary with s given x	$g(x) + s\rho(x)$
A2	$\rho(x) = \rho$ across x	$g(x) + \rho s$ (PLR)
A3	$g(x) = \alpha + x'\gamma$	$\alpha + \rho s + x'\gamma$
MHE (3.2.7)	$f_i(s) = \alpha + \rho s + \eta_i, \mathbb{E}[\eta_i \mid X_i] = X_i'\gamma$	A1–A3 at once (stronger: same ρ for every i)

Central message

Causal identification (CIA) and **functional form** (A1–A3) are logically separate. If A1–A3 fail, regression still approximates the causal CEF (MHE §3.3).

Long and Short Regressions

Let A_i (“ability”) collect family background, intelligence, motivation. The **long regression**:

$$Y_i = \alpha + \rho s_i + A_i' \gamma + e_i, \quad (3.2.10)$$

where e_i is a regression residual, uncorrelated with s_i and A_i *by construction*. If the CIA holds given A_i , then ρ is the causal effect in (3.2.7). **Problem:** A_i is rarely observed (e.g., the CPS has no ability or family background), so we run the **short regression** of Y_i on s_i alone.

Omitted variables bias formula

$$\underbrace{\frac{\text{Cov}(Y_i, s_i)}{V(s_i)}}_{\text{short}} = \underbrace{\rho}_{\text{long}} + \underbrace{\gamma'}_{\text{effect of omitted}} \underbrace{\delta_{As}}_{\text{omitted on included}}, \quad (3.2.11)$$

where δ_{As} is the vector of coefficients from regressing each element of A_i on s_i .

Short equals long plus the effect of omitted times the regression of omitted on included.

Deriving OVB: Two Lines

Plug the long regression into the short-regression slope:

$$\begin{aligned}\text{Cov}(Y_i, s_i) &= \text{Cov}(\alpha + \rho s_i + A_i' \gamma + e_i, s_i) \\ &= \rho V(s_i) + \gamma' \text{Cov}(A_i, s_i) + \underbrace{\text{Cov}(e_i, s_i)}_{=0 \text{ (long residual)}}.\end{aligned}$$

Divide by $V(s_i)$ and note $\delta_{As} = \text{Cov}(A_i, s_i)/V(s_i)$:

$$\frac{\text{Cov}(Y_i, s_i)}{V(s_i)} = \rho + \gamma' \delta_{As}. \quad \blacksquare$$

Three things to notice

- 1 **Mechanical:** the formula holds whether or not the long regression is causal.
- 2 **Short = long** iff $\gamma' \delta_{As} = 0$, e.g. omitted variables are uncorrelated with s_i (cf. regression anatomy).
- 3 **Multivariate version:** $\beta_1^s = \beta_1^l + \mathbb{E}[X_{1i} X_{1i}']^{-1} \mathbb{E}[X_{1i} X_{2i}'] \beta_2^l$.

Signing the Bias

With a scalar omitted variable, the bias is $\gamma \delta_{As}$: two signs to guess.

	$\delta_{As} > 0$ (omitted $\uparrow$ with s_i)	$\delta_{As} < 0$ (omitted $\downarrow$ with s_i)
$\gamma > 0$ (omitted raises Y_i)	short too big	short too small
$\gamma < 0$ (omitted lowers Y_i)	short too small	short too big

Ability and schooling

- Ability raises wages: $\gamma > 0$. More able people likely get more schooling: $\delta_{As} > 0$.
- ⇒ The short regression is probably **too big**.
- But $\delta_{As} > 0$ is not a theorem: for very able people, the opportunity cost of schooling is high (Mick Jagger left the LSE, Bill Gates left Harvard).

OVB at Work: Returns to Schooling in the NLSY

Coefficient on years of schooling, log wage regressions, NLSY79 men (MHE Table 3.2.1):

	(1)	(2)	(3)	(4)	(5)
Controls	None	Age dummies	(2) + parents' educ., race, region	(3) + AFQT	(4) + occupation dummies
$\hat{\rho}$	.132 (.007)	.131 (.007)	.114 (.007)	.087 (.009)	.066 (.010)

- (1)→(3)→(4): $\hat{\rho}$ falls as family background and AFQT are added.
- OVB reading: these controls are *positively correlated with both wages and schooling* ($\gamma > 0$, $\delta_{As} > 0$).
- Column (5) also falls. Is occupation one more omitted variable? **No: hold that thought.**

Controls Exist to Make the CIA Credible

Saying “my estimate has no OVB” means “my regression is the long regression I want,” and the regression we want is the one with a causal interpretation:

$$\text{no OVB claim} \iff Y_{si} \perp\!\!\!\perp s_i \mid X_i \text{ for my chosen } X_i.$$

So the right question is not ...

- “What raises my R^2 ?”
- “What is correlated with Y_i ?”
- “What changes $\hat{\rho}$ when I add it?”

...but

Conditional on which X_i is s_i as good as randomly assigned?

OVB says to include variables with $\gamma \neq 0$ and $\delta_{As} \neq 0$. That is necessary but **not sufficient**: occupation is correlated with both schooling and wages, and is still a bad control.

Good Controls and Bad Controls

More control is not always better.

Good control

A variable we can think of as **fixed at the time** s_i **was determined**.

e.g. parents' education, race, 8th-grade test score.

Bad control

A variable that is itself an **outcome of** s_i : it might just as well be a dependent variable.

e.g. occupation, a post-schooling test score.

Two versions (MHE §3.2.3)

- ① **Outcome as control:** conditioning on occupation in a schooling regression.
- ② **Proxy control:** using a later ability measure, itself raised by schooling, in place of early ability.

Both are versions of **selection bias**, more subtle than the one in (3.2.1).

Bad Control I: Setup

College c_i affects two outcomes: earnings Y_i and white-collar status w_i .

$$Y_i = c_i Y_{1i} + (1 - c_i) Y_{0i}, \quad w_i = c_i w_{1i} + (1 - c_i) w_{0i}.$$

Make the best case: c_i is **randomly assigned**, so $c_i \perp\!\!\!\perp \{Y_{1i}, Y_{0i}, w_{1i}, w_{0i}\}$. Then both causal effects are easy:

$$\begin{aligned} \mathbb{E}[Y_i \mid c_i = 1] - \mathbb{E}[Y_i \mid c_i = 0] &= \mathbb{E}[Y_{1i} - Y_{0i}], \\ \mathbb{E}[w_i \mid c_i = 1] - \mathbb{E}[w_i \mid c_i = 0] &= \mathbb{E}[w_{1i} - w_{0i}]. \end{aligned}$$

Just regress Y_i and w_i on c_i .

The tempting idea

Occupation is correlated with both college and pay. So compare college and non-college workers *within* white-collar jobs, i.e. add w_i as a control, or regress Y_i on c_i in the $w_i = 1$ sample.

Bad Control I: Conditioning on an Outcome

Among white-collar workers, $\mathbb{E}[Y_i \mid w_i = 1, c_i = 1] - \mathbb{E}[Y_i \mid w_i = 1, c_i = 0]$

$$\begin{aligned} &= \mathbb{E}[Y_{1i} \mid w_{1i} = 1, c_i = 1] - \mathbb{E}[Y_{0i} \mid w_{0i} = 1, c_i = 0] \\ &= \mathbb{E}[Y_{1i} \mid w_{1i} = 1] - \mathbb{E}[Y_{0i} \mid w_{0i} = 1] \quad (\text{random } c_i) \\ &= \underbrace{\mathbb{E}[Y_{1i} - Y_{0i} \mid w_{1i} = 1]}_{\text{causal effect}} + \underbrace{\mathbb{E}[Y_{0i} \mid w_{1i} = 1] - \mathbb{E}[Y_{0i} \mid w_{0i} = 1]}_{\text{selection bias}}. \end{aligned}$$

Why: college changes *who* is in the white-collar pool

- Treated group: people who are white collar *with* college ($w_{1i} = 1$).
- Control group: people who are white collar *without* college ($w_{0i} = 1$): probably special, with high Y_{0i} .

⇒ Apples vs. oranges, even though c_i was randomized. Selection bias here is probably **negative**.

Even if $Y_{1i} = Y_{0i}$ for everyone, this comparison is generally nonzero. Nor is it “the effect of college not explained by occupation.”

Bad Control I: A Numerical Illustration

Suppose college has **zero** effect on earnings ($Y_{1i} = Y_{0i}$) but does open white-collar doors. Three types:

Type	Share	(w_{0i}, w_{1i})	$Y_{0i} = Y_{1i}$	White collar if ...
Always white collar	0.2	(1, 1)	60	always
White collar only with college	0.3	(0, 1)	40	$c_i = 1$
Never white collar	0.5	(0, 0)	30	never

With c_i randomized, among white-collar workers:

$$\mathbb{E}[Y_i \mid w_i=1, c_i=1] = \frac{0.2 \cdot 60 + 0.3 \cdot 40}{0.5} = 48, \quad \mathbb{E}[Y_i \mid w_i=1, c_i=0] = 60.$$

Result

The “within-occupation return to college” is $48 - 60 = -12$, while the true effect is 0. The gap comes entirely from the change in who is white collar.

Back to the NLSY: Column (5)

	(4) + AFQT	(5) + occupation dummies
$\hat{\rho}$	.087	.066

- Adding two-digit occupation dummies lowers the schooling coefficient.
- OVB would read this as “occupation was an omitted variable.” But occupation is an **outcome of schooling**.
- The drop may be nothing more than the selection bias from the previous slides.

Rule

Control only for variables that are **not themselves caused by** the regressor of interest.

Bad Control II: Proxy Control

The regression we want uses **early ability** a_i (e.g. an 8th-grade IQ score, measured before schooling choices):

$$Y_i = \alpha + \rho s_i + \gamma a_i + e_i, \quad \mathbb{E}[s_i e_i] = \mathbb{E}[a_i e_i] = 0. \quad (3.2.13)$$

a_i is a good control but unavailable. We only have **late ability** a_{li} , measured after schooling (e.g. a job-screening test), which schooling also raises:

$$a_{li} = \pi_0 + \pi_1 s_i + \pi_2 a_i. \quad (3.2.14)$$

Solve for $a_i = (a_{li} - \pi_0 - \pi_1 s_i)/\pi_2$ and substitute into (3.2.13):

$$Y_i = \left(\alpha - \gamma \frac{\pi_0}{\pi_2} \right) + \boxed{\left(\rho - \gamma \frac{\pi_1}{\pi_2} \right)} s_i + \frac{\gamma}{\pi_2} a_{li} + e_i. \quad (3.2.15)$$

Result

With $\gamma, \pi_1, \pi_2 > 0$, the coefficient on s_i is **too small**, unless $\pi_1 = 0$ (schooling does not affect the late test).

Proxy Control Can Still Help: Bounds

Unlike the occupation case, the intention here is good: we *want* to control for ability.

No control (OVB formula)

$$\rho + \gamma \delta_{as} \geq \rho$$

if $\gamma > 0$, $\delta_{as} > 0$.

Proxy control (3.2.15)

$$\rho - \gamma \frac{\pi_1}{\pi_2} \leq \rho$$

if $\gamma, \pi_1, \pi_2 > 0$.

$$\rho - \gamma \frac{\pi_1}{\pi_2} \leq \rho \leq \rho + \gamma \delta_{as}$$

- The causal effect lies **between** the short estimate and the proxy-controlled estimate.
- The proxy estimate may be closer to ρ than no control at all.
- π_1 can be partly checked: if a_{li} does not move with s_i , assuming $\pi_1 = 0$ is more comfortable.

The Moral: Timing Matters

	Example	Determined	Effect on $\hat{\rho}$
Good control	Parents' education, early IQ a_i	Before s_i	Removes OVB if the CIA holds given it
Outcome control	Occupation w_i	After, <i>by</i> s_i	Selection bias, any sign; not interpretable
Proxy control	Late ability a_{li}	After, partly by s_i	Biased toward zero; useful as a bound

Practical rule

- Variables measured **before** s_i was determined are generally good controls: they cannot be outcomes of s_i .
- If timing is unclear, state explicitly what happened first, or argue that no control is caused by s_i .

Takeaways

- 1 **CIA makes the CEF causal:** $\mathbb{E}[Y_i | X_i, s_i = s] = \mathbb{E}[f_i(s) | X_i]$. Within X -cells, selection bias vanishes.
- 2 **The CIA is credible** when we know how s_i was assigned: lotteries within groups, or well-understood institutional screening.
- 3 **Linearity is extra:** constant returns in s (A1), common across groups (A2), linear baseline (A3), or MHE's $f_i(s) = \alpha + \rho s + \eta_i$.
- 4 **OVB:** short = long + $\gamma' \delta_{As}$. Mechanical, and a quick way to sign bias.
- 5 **Controls serve the CIA:** include what makes s_i as good as random; **exclude** what s_i causes.

Good control: fixed before s_i

Bad control: an outcome of s_i

References

- Angrist and Pischke (2009), *Mostly Harmless Econometrics*, §3.2.
- Barnow, Cain, and Goldberger (1981), “Issues in the Analysis of Selectivity Bias,” *Evaluation Studies Review Annual*.
- Black, Smith, Berger, and Noel (2003), “Is the Threat of Reemployment Services More Effective than the Services Themselves?” *AER*.
- Angrist (1998), “Estimating the Labor Market Impact of Voluntary Military Service Using Social Security Data on Military Applicants,” *Econometrica*.
- Griliches and Mason (1972), “Education, Income, and Ability,” *JPE*.
- Robinson (1988), “Root- N -Consistent Semiparametric Regression,” *Econometrica*.

Next: MHE §3.3

When A1–A3 fail, what does the regression coefficient on s_i average? (Backup slides follow.)

Backup: OLS Exists Even When the CEF Is Nonlinear

The population linear projection is defined by

$$(\alpha^{LP}, \rho^{LP}, \gamma^{LP}) = \arg \min_{a, b, c} \mathbb{E}[(Y_i - a - bs_i - X_i'c)^2].$$

Write $R_i = s_i - \text{Proj}_{\text{lin}}(s_i \mid 1, X_i)$. By regression anatomy (FWL),

$$\rho^{LP} = \frac{\mathbb{E}[R_i Y_i]}{\mathbb{E}[R_i^2]}.$$

Under the CIA,

$$\rho^{LP} = \frac{\mathbb{E}[R_i m(s_i, X_i)]}{\mathbb{E}[R_i^2]}.$$

Warning

When m is nonlinear, the OLS coefficient still exists, but it **need not equal an ATE or an average causal derivative**.

Backup: Partialling Out with Heterogeneous Effects

Allow $m(s, x) = g(x) + s\rho(x)$: keep A1, drop A2 and A3. Let $V_i = s_i - \mathbb{E}[s_i | X_i]$ and define

$$\rho^{PO} = \frac{\mathbb{E}[V_i \{Y_i - \mathbb{E}[Y_i | X_i]\}]}{\mathbb{E}[V_i^2]}.$$

Iterating expectations,

$$\mathbb{E}[V_i Y_i | X_i] = \rho(X_i) \mathbb{E}[V_i s_i | X_i] = \rho(X_i) V(s_i | X_i).$$

Thus

$$\rho^{PO} = \frac{\mathbb{E}[V(s_i | X_i) \rho(X_i)]}{\mathbb{E}[V(s_i | X_i)]}.$$

Interpretation

Groups with more residual variation in schooling get more weight: a *variance-weighted* causal slope, not $\mathbb{E}[\rho(X_i)]$. This is the continuous analog of MHE (3.3.7).

Backup: Linear OLS vs. Nonparametric Partialling Out

	Linear OLS controls	Full CEF controls
Residual treatment	$R_i = s_i - \text{Proj}_{lin}(s_i \mid 1, X_i)$	$V_i = s_i - \mathbb{E}[s_i \mid X_i]$
Conditioning on X	Linear projection	Potentially nonlinear
General estimand	$\mathbb{E}[R_i Y_i] / \mathbb{E}[R_i^2]$	$\mathbb{E}[V_i Y_i] / \mathbb{E}[V_i^2]$
With $m(s, x) = g(x) + s\rho(x)$	May mix baseline misspecification with the treatment effect	Variance-weighted mean of $\rho(X_i)$

When do they coincide?

If $\mathbb{E}[s_i \mid X_i]$ is linear in $(1, X_i)$, e.g. a saturated model in X_i , then $R_i = V_i$ and the two coefficients coincide.