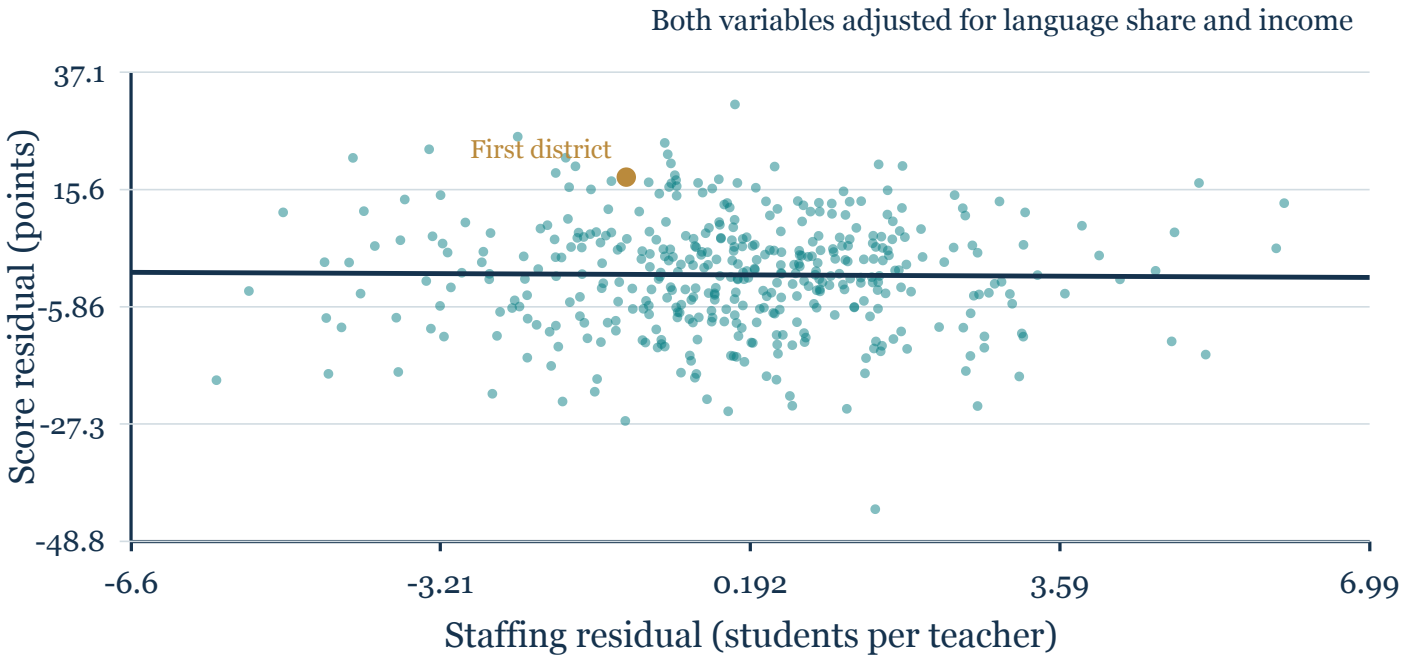
# What does holding other things fixed mean?

Show the comparison left after observed controls are removed.




Actual residuals after removing English-learner share and income from both variables. The slope matches the full regression.

# The full model sets the target coefficient

Define every component before removing controls.

> $Y_i = \beta_0 + \beta_1 X_i + \beta_2 Z_i + \beta_3 W_i + u_i$

> $E(u_i \mid X_i, Z_i, W_i) = 0$ for the stated mean model

# OLS solves a sample optimisation

Its residual identities follow from the fitted model.

> $\sum_i e_i = 0$

> $\sum_i X_i e_i = 0$ for each included regressor.

# Sample orthogonality is mechanical

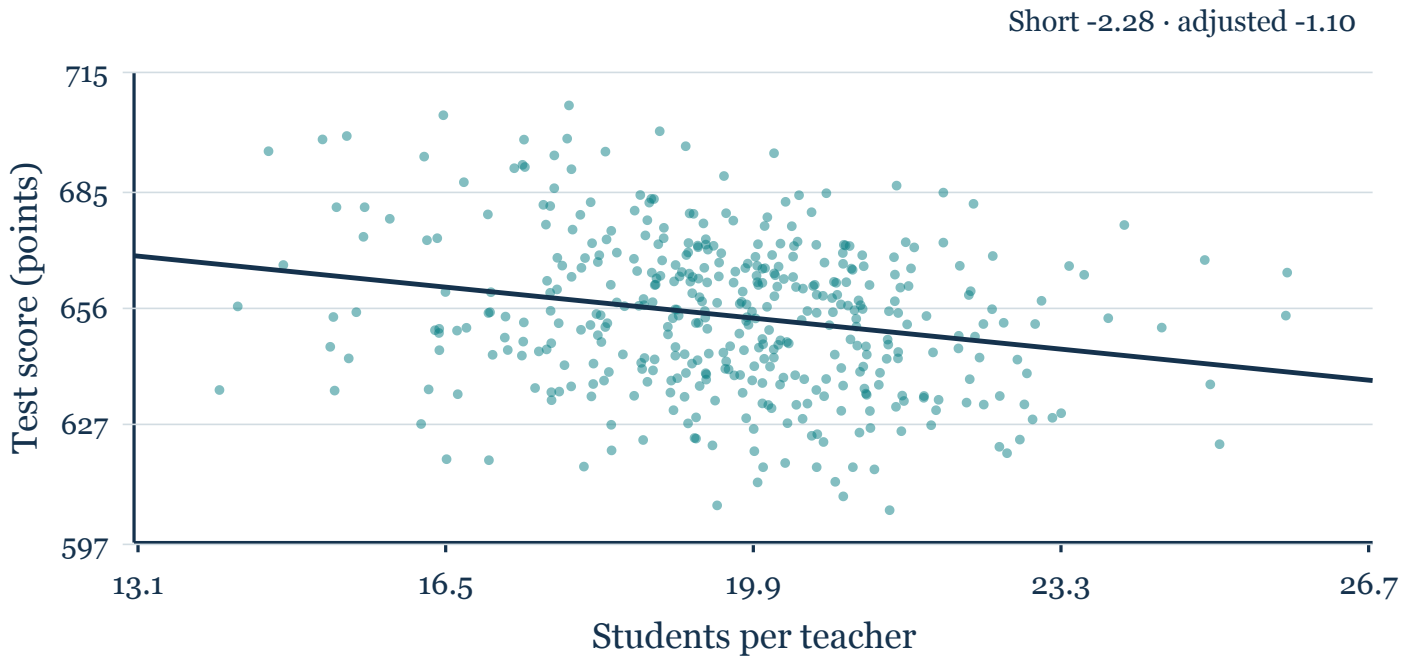
Population exogeneity is a substantive assumption.

> $\sum_i e_i = 0, \sum_i X_i e_i = 0$ with an OLS intercept

> Orthogonality here means a zero sum of paired products.

# What is hidden in the staffing comparison?

Use a second variable to understand a changed slope.




Observed school data. The control moves an illustrative slope between the two fitted endpoints; intermediate slopes are not new regressions.

# Write the longer relationship first

Define what the short model leaves out.

> $Y_i = \beta_0 + \beta_1 X_i + \beta_2 Z_i + u_i$

> Exogeneity for this mean model: $E(u_i \mid X_i, Z_i) = 0$

# Two relationships determine omitted-variable bias

Multiply the omitted effect by the omitted-variable association.

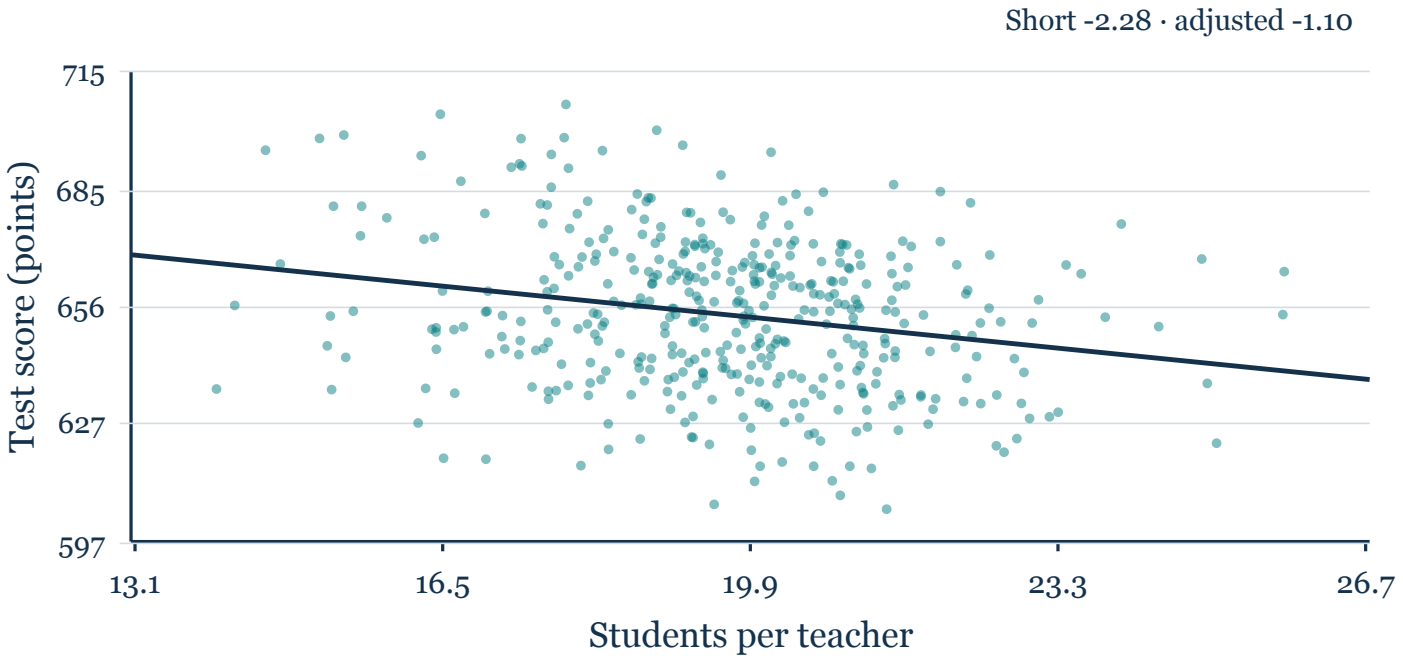
> $Z = \delta_0 + \delta_1 X + v$

> $Y = (\beta_0 + \beta_2 \delta_0) + (\beta_1 + \beta_2 \delta_1) X + (\beta_2 v + u)$

> Short slope minus longer slope: $\beta_2 \delta_1$

# Predict the direction before seeing the estimate

The sign depends on both parts of the product.




Observed school data. The control moves an illustrative slope between the two fitted endpoints; intermediate slopes are not new regressions.

# Predict omitted-variable bias

Try the comparison before opening the explanation.

**Practice**

In a wage model, the causal return to schooling is $\beta$. Ability raises wages by $\gamma > 0$, and the auxiliary regression of ability on schooling has slope $\delta > 0$. Under the scalar model, what is the omitted-variable bias?

# Predict omitted-variable bias: explanation

Follow the calculation and interpret its units.

> The short-regression slope is $\beta + \gamma\delta$, so its bias relative to $\beta$ is positive. Schooling picks up some of the association between ability and wages. This requires the stated linear model and exogeneity of the remaining error.

# Check the actual coefficient movement

Use the same districts in both regressions.

```
. quietly regress score str el

. matrix list e(b), format(%10.6f)

e(b)[1,3]
           str          el       _cons
y1   -1.101296   -0.649777  686.032250
```

Actual Stata coefficient-only display; reproduce with coefficient_workshop.do.

# Verify the omitted-variable calculation

An exact sample identity can be checked independently.

```
. quietly regress el str

. matrix list e(b), format(%10.6f)

e(b)[1,2]
           str       _cons
y1    1.813719  -19.854055
```

Actual Stata coefficient-only display; reproduce with coefficient_workshop.do.

$$b_{1,\text{short}} - b_{1,\text{long}} = b_2\hat{\delta}_1$$

# Derive the omitted-variable contribution from population moments

The omitted effect and its association with the included regressor both matter.

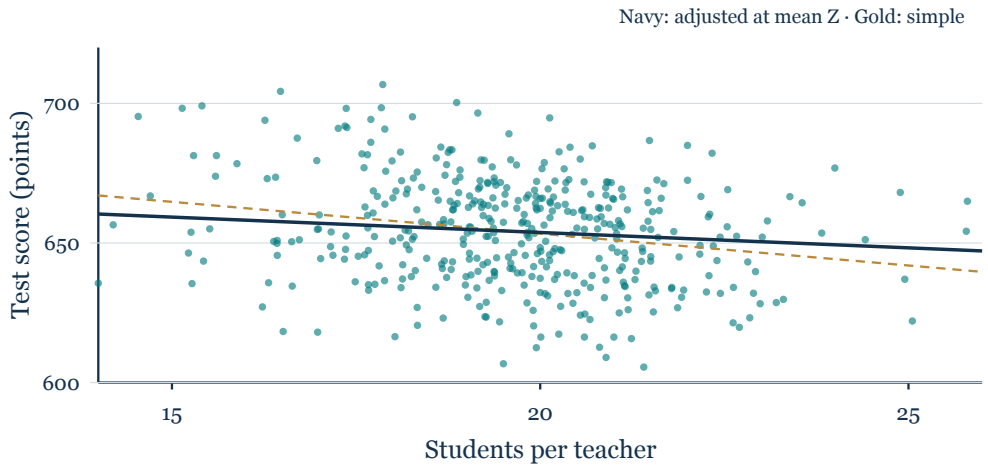
> $Y = \beta_0 + \beta_1 X + \beta_2 W + u,\ E[u \mid X, W] = 0.$

> $\operatorname{Cov}(X, Y) = \beta_1 \operatorname{Var}(X) + \beta_2 \operatorname{Cov}(X, W).$

> $\alpha_1 = \beta_1 + \beta_2 \operatorname{Cov}(X, W) / \operatorname{Var}(X),\ \operatorname{Var}(X) > 0.$

# Explain why the estimated slope changed

Adding Z changes the fitted comparison from about +4.56 to +2.20 score points.




Adjusted line fixes English-learner share at its sample mean. Both fits use the same 420 districts.

## Practice

Report the two fitted score differences for 22 → 20 students per teacher. Explain why they need not be the same.

**Explanation:** The simple model gives about +4.56 points. The adjusted model gives about +2.20 points while holding English-learner share fixed. The first comparison also reflects observed differences in language composition across districts; the second accounts for those differences within its stated model.

# Read the adjusted regression

Read both coefficients with their units and the stated comparison.

```
. quietly regress score str el

. matrix list e(b), format(%10.6f)

e(b)[1,3]
           str          el       _cons
y1   -1.101296   -0.649777  686.032250
```

Actual Stata coefficient-only display; reproduce with coefficient_workshop.do.

> Adjusted staffing slope: $-1.101296$

> At fixed language share: $(-2)(-1.101296) = 2.202592$ points

# Attempt the adjusted policy contrast

Keep the arithmetic and the interpretation together.

```
. display %10.6f -2*_b[str]
  2.202592
```

Actual Stata coefficient-only display; reproduce with coefficient_workshop.do.

$$\Delta\widehat{Y} = (-2)(-1.101296) = 2.202592 \text{ points}$$

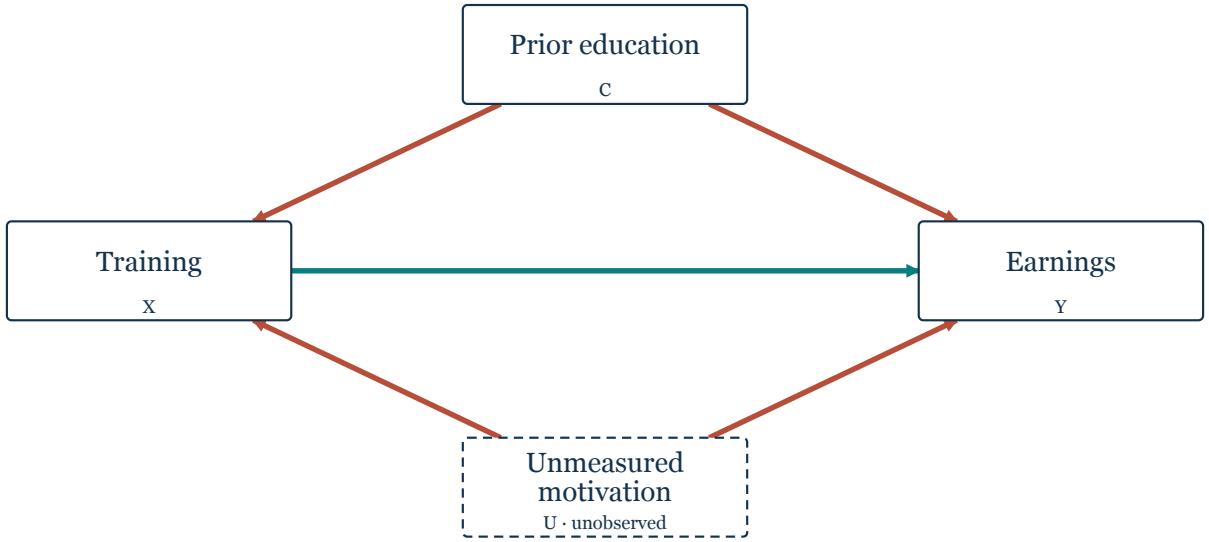
**Practice**

Calculate the adjusted fitted contrast for a ratio reduction from 22 to 20. Explain what is held fixed.

**Explanation:** About +2.20 score points, with English-learner share fixed in the fitted model.

# Choose controls for the research question

State what each variable is meant to address.




1 open directed route · 2 open alternatives · gold: conditioned on

Assumed causal structure · dashed boxes: unobserved · gold boxes: conditioned on

> New application: training → earnings.

> Prior education is a proposed common cause; motivation remains unmeasured.

# Adjustment changes a comparison; it does not guarantee identification

A richer model can still leave an important factor out.

> Identification: do the data and assumptions determine the target?

> Estimation: use the sample to calculate it.

# Add income and inspect what changes

An additional observed factor can substantially alter a slope.

```
. quietly regress score str el income

. matrix list e(b), format(%10.6f)

e(b)[1,4]
           str          el      income       _cons
y1   -0.068776   -0.488267    1.494517  640.315497
```

Actual Stata coefficient-only display; reproduce with coefficient_workshop.do.

# Keep the estimation sample fixed

Otherwise a coefficient change can have two explanations.

> Compare identical observations

> Then investigate why observations are missing

# Explain an observed omitted-variable contribution

Reproduce the evidence, then explain its economic meaning.

**Practice**

Repeat the omitted-variable decomposition for income, retaining language share in every supporting equation. Verify the staffing-coefficient movement on one common sample. Keep your commands, log, output and interpretation.

# Explain an observed omitted-variable contribution: explanation

A reproducible result needs both a calculation and an interpretation.

> Original language contribution: −1.178512.

> Income contribution, retaining language share: −1.032520.

> Check compatible models and the same 420 districts.

# From a control set to the information OLS uses

> From a control set to the information OLS uses

# Why can adding a regressor change a slope?

A raw comparison can mix several observed differences.

> A short and a longer regression describe different conditional comparisons.

> Derive the difference before interpreting observed coefficient movement.

# Recall the difference between an error and a residual

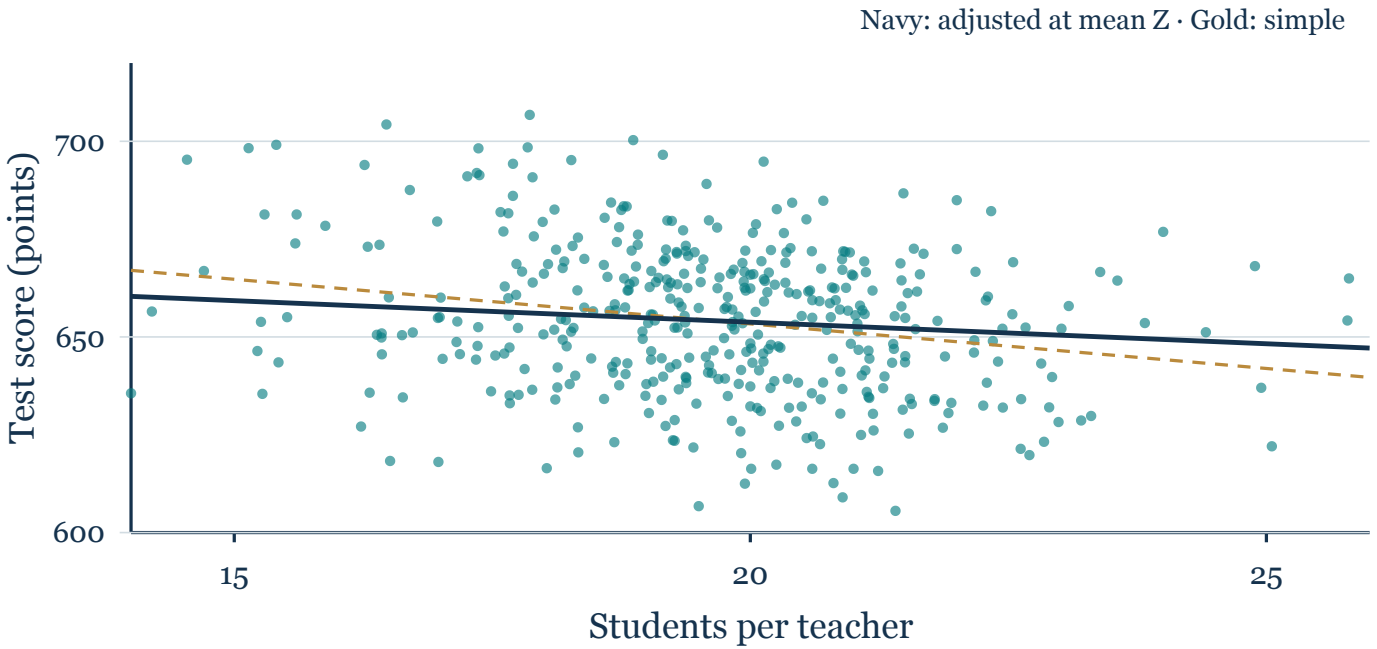
A property of a fitted line is not a test of the design.

> $e_i = Y_i - \widehat{Y}_i$

> $\sum_i e_i = 0$ with an OLS intercept

# Holding an observed factor fixed

Multiple regression compares ratios and scores at a fixed value of another variable.

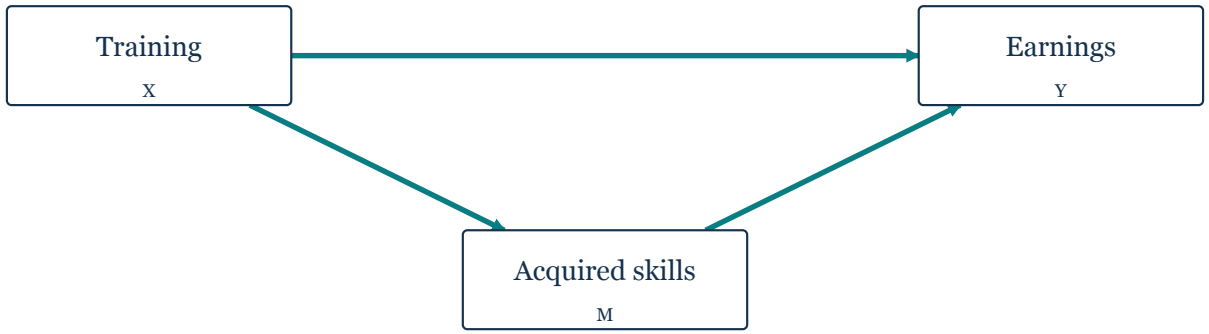



Adjusted line fixes English-learner share at its sample mean. Both fits use the same 420 districts.

- $\widehat{Y_i} = b_0 + b_1 X_i + b_2 Z_i$
- $b_1$: staffing slope at fixed $Z$
- The graph fixes language share at its sample average.

# A confounder and a mediator do different jobs

The timing of a control changes its meaning.



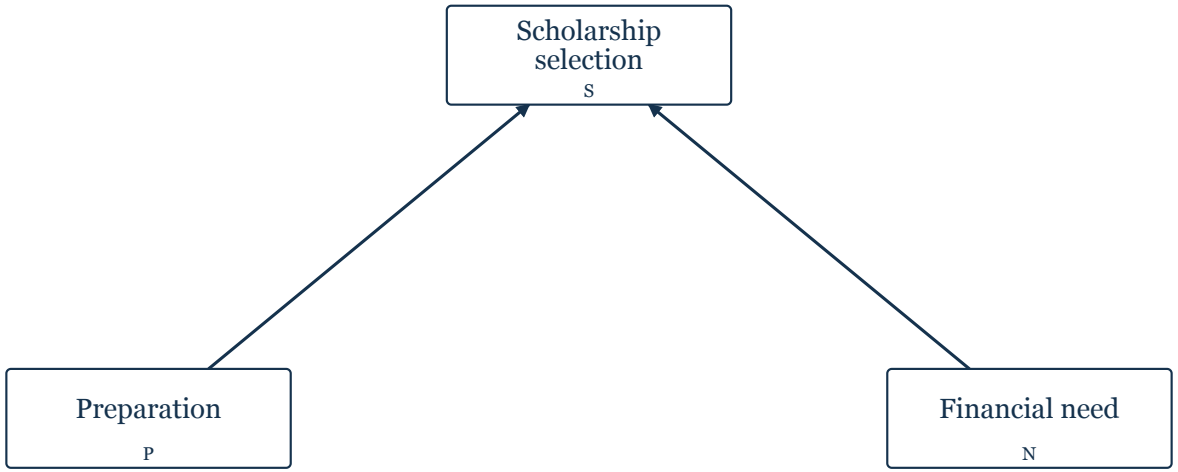

2 open directed routes · 0 open alternatives · gold: conditioned on

Assumed causal structure · dashed boxes: unobserved · gold boxes: conditioned on

- Confounder: a prior common influence
- Mediator: part of the effect pathway

# Why a collider can be a bad control

Selection into a shared consequence can create a relationship.




0 open directed routes · 0 open alternatives · gold: conditioned on

Assumed causal structure · dashed boxes: unobserved · gold boxes: conditioned on

> Preparation → Selection ← Need

# Confounder or mediator?

Try the comparison before opening the explanation.

**Practice**

Training changes skills, and skills change earnings. The target is training's total effect. Should post-training skill automatically be included as a control?

# Confounder or mediator?: explanation

Follow the calculation and interpret its units.

> No. Skill is a mediator on a causal path from training to earnings. Holding it fixed changes the estimand, meaning the population quantity we want and may introduce other bias. A direct-effect interpretation needs additional assumptions, not merely an added control.

# Keep the sample fixed

Try the comparison before opening the explanation.

**Practice**

A coefficient changes after adding income, but income is missing for 20% of observations. Why can you not attribute the entire change to controlling for income?

# Keep the sample fixed: explanation

Follow the calculation and interpret its units.

> The comparison changes both the regressors and the estimation sample. Refit the short and long models on the same complete-case sample to isolate the specification change. Then consider whether missingness itself creates selection bias.

# Interpret a control coefficient cautiously

The variable may be useful without having a causal interpretation.

> The coefficient on English-learner share is an adjusted district association.

> The staffing study may include this variable to address a possible source of confounding, without trying to identify the causal effect of language composition itself.

# Fit and causal validity are different properties

Explaining more variation need not improve the comparison.

$$R^2 = 1 - \frac{\sum_i e_i^2}{\sum_i (Y_i - \overline{Y})^2}$$

# A useful control plan fits in a short table

Make the assumptions inspectable.

> For each proposed control, write a sentence about timing and a sentence about purpose.

> This exercise reveals disagreements about the economic model before they become disagreements about software.

# Use the design to decide what data to collect next

More variables are useful when they address a named problem.

> If earlier learning difficulties drive teacher recruitment, prior outcome measurements may be useful.

> These suggestions follow from different explanations of the error term.

# Write the result as an empirical argument

A coefficient is one part of the explanation.

> A useful report starts with the staffing question and identifies a district as the observation.

> Avoid saying that the model 'controls for everything'.

# Check the two requirements for omitted-variable bias

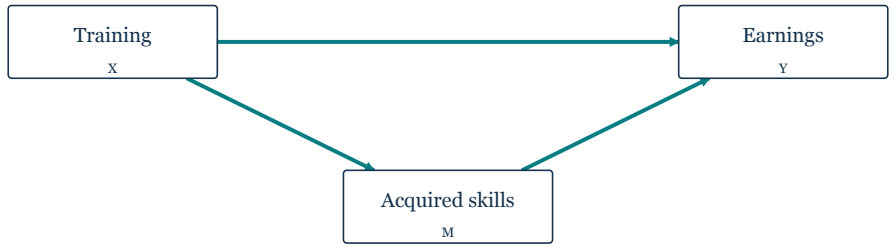
A left-out factor must matter in both relationships.

> If Z does not affect Y in the longer model, its omitted-effect coefficient is zero and the product vanishes.

> Explain these two cases using an economic example.

# DAG practice: preserve the intended effect

Trace the routes before choosing a control.




2 open directed routes · 0 open alternatives · gold: conditioned on

Assumed causal structure · dashed boxes: unobserved · gold boxes: conditioned on

**Practice**

Training can affect earnings directly and through acquired skills. The target is the total effect of training. Use the diagram to explain what conditioning on acquired skills removes. Does the graph alone guarantee that an adjusted regression identifies a direct effect?

# DAG practice: preserve the intended effect: explanation

Explain what conditioning changes in this stated graph.

> Conditioning on skills closes the route $X \to M \to Y$, removing a mechanism included in the total-effect question. The direct route $X \to Y$ remains in the diagram. A skills-held-fixed comparison therefore concerns a different target. A direct-effect estimate also needs suitable assumptions about treatment assignment, mediator–outcome confounding, intervention meaning, support and the fitted model. Adding a post-training mediator is not, by itself, evidence that these conditions hold. Start with the intended effect, then justify the control choice.