OLS Foundations and Omitted-Variable Bias

28 Sep 2026 | Chapters 5-6

ECO 351 – Introduction to Econometrics

Shi Qiu – Fall 2026

Quiz 2 debrief: calculate the fitted comparison

Question 1

	Coefficient	95% CI
STR	-2.2798	[-3.223,-1.337]

$$\widehat{Score}(18) - \widehat{Score}(20) = (-2.2798)(18 - 20) = 4.56$$

The model predicts about 4.56 score points higher at STR 18. State the direction and the units.

Quiz 2 debrief: an interval is not a causal design

Question 2

$$CI_{95\%}(\beta_{STR}) = [-3.223, -1.337]$$

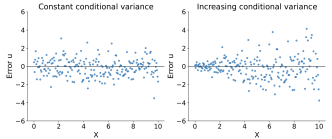
What the interval supports

Under the maintained inference assumptions, the population regression slope is negative and zero is outside the interval.

What it does not establish

The interval cannot show that $E(u \mid STR) = 0$. Without that assumption, the slope remains an association.

Quiz 2 debrief: two different problems need two different repairs



Changing spread

The fan-shaped panel violates constant $Var(u \mid STR)$. The errors can still be centered at zero.

Changing mean

If high-need districts receive extra teachers, need enters u and is related to STR, so $E(u \mid STR)$ can fail.

Robust SEs address the first problem. A better comparison or design is needed for the second.

From the quiz concern to omitted-variable bias

English-learner share is one possible omitted score determinant. For omitting it to bias the STR slope, two links must both be present.

Outcome link

At fixed STR, English-learner share must be related to scores: $\beta_{EL} \neq 0$.

Regressor link

English-learner share must be related to STR: $\delta_1 \neq 0$.

The sign of the bias depends on the signs of both links.

Put the omitted variable into the population equation

Let X be STR and Z English-learner share. Suppose the fuller population model is

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 Z_i + v_i, \qquad E(v_i \mid X_i, Z_i) = 0.$$

Here β_1 is the conditional STR slope. If Z is omitted, its contribution enters the short-regression error.

See the omitted term enter the short regression

$$Z = \delta_0 + \delta_1 X + r, \qquad \text{Cov}(X, r) = 0$$

Substitute this projection into the fuller outcome equation:

$$Y = (\beta_0 + \beta_2\delta_0) + (\beta_1 + \beta_2\delta_1)X + (\beta_2 r + v).$$

The short-regression slope contains the extra term $\beta_2\delta_1$.

Use the signs before using the numbers

$$\widetilde{\beta}_1 \xrightarrow{p} \beta_1 + \beta_2\delta_1$$

The arrow means that, as the sample grows, the short-regression estimate becomes increasingly concentrated near $\beta_1 + \beta_2\delta_1$. For the California story, higher English-learner share predicts lower scores ($\beta_2 < 0$) and is positively associated with STR ($\delta_1 > 0$).

$$\beta_2\delta_1 < 0$$

The simple STR slope is therefore more negative than the conditional STR slope in this model.

Exercise: reverse the assignment link

Suppose districts with more English learners receive extra teachers. English-learner share is then negatively associated with STR, while its score relationship remains negative.

What sign does the omitted-variable bias have? Is the simple STR slope shifted upward or downward?

Solution: reverse the assignment link

$$\beta_2 < 0, \qquad \delta_1 < 0 \qquad \Longrightarrow \qquad \beta_2 \delta_1 > 0$$

The simple STR slope is shifted upward relative to the conditional slope. It may look less negative, or even positive.

Stata: compare the simple and controlled equations

Predict the direction of the STR coefficient movement before running the models.

```
do "setup.do" schools
quietly regress score str, vce(robust)
estimates store simple
regress score str el, vce(robust)
estimates store controlled
estimates table simple controlled, b(%9.3f) se(%9.3f)
   stats(N)
```

What to compare

Read the STR row across the two columns. Both regressions use the same 420 districts.

Stata output excerpt (1/2)

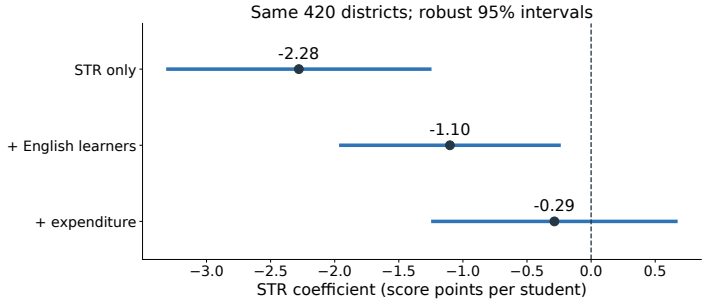
```
. regress score str el, vce(robust)
Number of obs     =         420

------------------------------------------------------------------------------
             |               Robust
       score | Coefficient  std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
         str |  -1.101296   .4328472    -2.54   0.011      -1.95213   -.2504616
          el |  -.6497768   .0310318   -20.94   0.000      -.710775   -.5887786
       _cons |   686.0322   8.728224    78.60   0.000     668.8754     703.189
------------------------------------------------------------------------------
```

Stata output excerpt (2/2)

```
. estimates table simple controlled, b(%9.3f) se(%9.3f) stats(N)
-------------------------------------
    Variable |   simple     control~d
-------------+-----------------------
         str |     -2.280       -1.101
             |      0.519        0.433
          el |                  -0.650
             |                   0.031
       _cons |    698.933      686.032
             |     10.364        8.728
-------------+-----------------------
           N |        420          420
-------------------------------------
                          Legend: b/se
```

The estimate moves as the comparison changes



The first two rows use the same 420 districts. The third row previews adding expenditure per student.

Coefficient movement shows that the models make different comparisons; it does not prove that the final model is causal.

Exercise: calculate a short-regression slope

$$\beta_1 = -1.5, \qquad \beta_2 = -0.8, \qquad \delta_1 = 0.5$$

Find the large-sample short-regression slope and the omitted-variable bias.

Solution: calculate a short-regression slope

$$\beta_2\delta_1 = -0.4, \qquad \widetilde{\beta}_1 \xrightarrow{p} -1.5 - 0.4 = -1.9$$

The short regression is 0.4 units too negative for the stated conditional target.

What coefficient movement can and cannot tell us

No bias through this pathway

If $\beta_2 = 0$ or $\delta_1 = 0$, the product $\beta_2\delta_1$ is zero.

No proof of causality

A stable coefficient can coexist with a missing confounder or offsetting omitted pathways. Movement is evidence to explain, not a test of identification.

Two problems require two different repairs

Unequal conditional variance

Use an appropriate variance estimator to improve inference about the maintained regression target.

Omitted-variable bias

Improve the comparison using credible controls, a suitable design, or other identifying variation.

What omitted-variable bias teaches us

Conclusion

An omitted determinant distorts a simple slope when it affects the outcome and is related to the included regressor.

Explain the sign of a bias by naming both links. Next we define exactly what a controlled coefficient measures.