Pooled OLS and Cluster-Robust Standard Errors

When 240 Observations Are Really Just 8

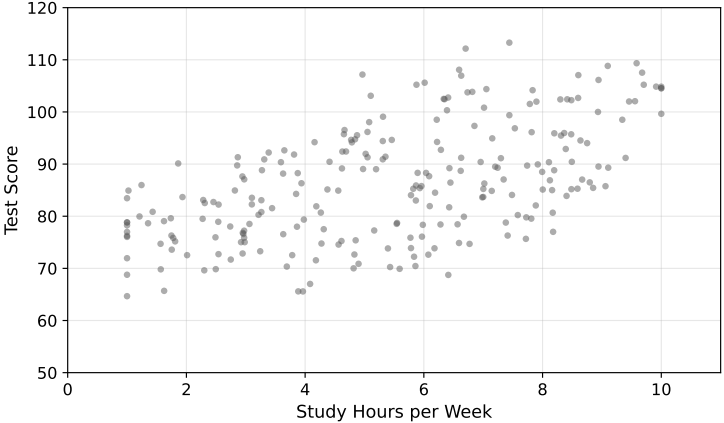
Jake Anderson

May 16, 2026

Outline

The Data

A school district tracks **study hours** vs. **test scores** for 240 students.



There appears to be a positive relationship. The true slope is $\beta_1 = 2.5$ points per hour. Can OLS recover it?

Setup: The Pooled OLS Model

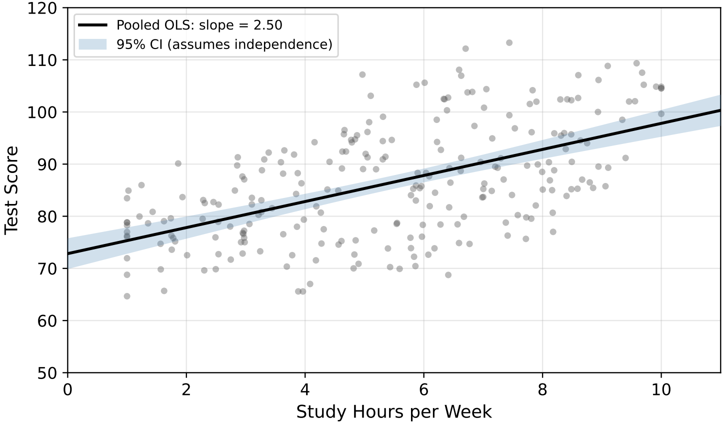
We ignore any group structure and run a single regression:

$$\text{Score}_i = \beta_0 + \beta_1 \,\text{Hours}_i + \varepsilon_i$$

This treats all 240 students as **independent observations**.

- One intercept, one slope, one error term
- No distinction between "within classroom" and "between classroom" variation
- Standard OLS assumptions: ε_i independent, $\text{Var}(\varepsilon_i) = \sigma^2$

Pooled OLS Result



Result: $\hat{\beta}_1 = 2.50$, SE $= 0.239$. The 95% CI is [2.03, 2.97]. Tight, precise, and contains the true slope of 2.5. Looks great!

But Something Is Wrong

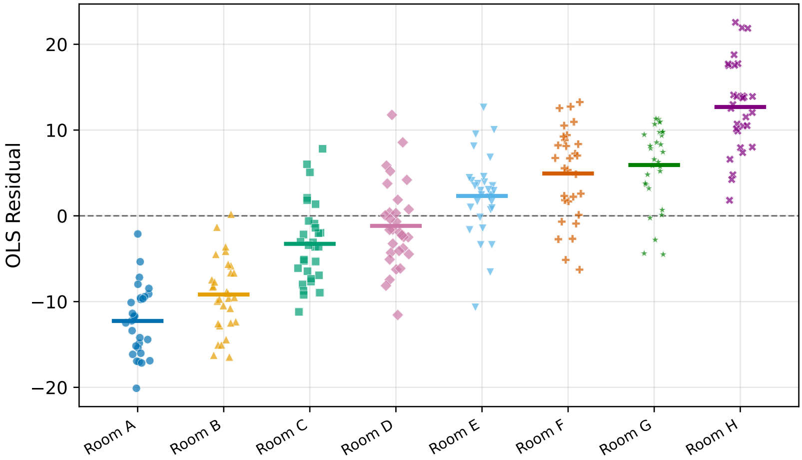
The slope estimate is right on target. So what's the problem?

The problem is not the slope. It's the standard error.

- OLS standard errors require that errors be **independent** across observations
- If students share unobserved factors (teacher quality, classroom culture, grading norms), their errors are **correlated**, not independent

$\Longrightarrow$ Let's look at the OLS residuals to see if the independence assumption holds.

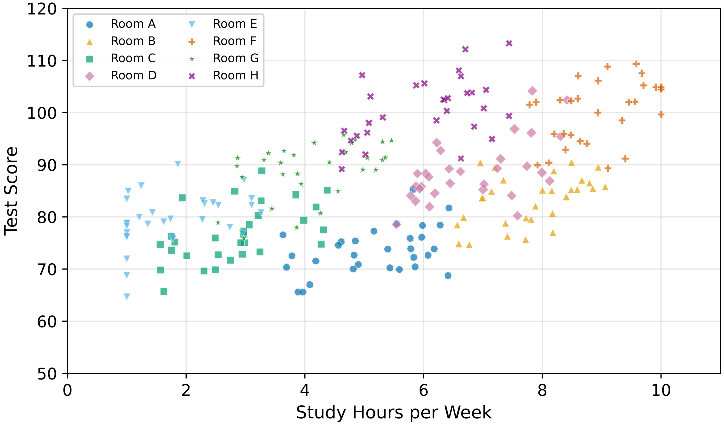
Residuals by Classroom



Room A residuals are **all negative** (mean $= -12.3$). Room H residuals are **all positive** (mean $= +12.7$). Within each classroom, residuals move together.

Reveal: Eight Classrooms

The 240 students come from **8 different classrooms**.



Students in the same classroom share a teacher, curriculum, and grading standard.

Why Residuals Are Correlated: The Composite Error

We now use double subscripts: student i within classroom j.

The true model has a **classroom-level component** that pooled OLS ignores:

$$\text{Score}_{ij} = \beta_0 + \beta_1 \,\text{Hours}_{ij} + \underbrace{u_j + e_{ij}}_{v_{ij}}$$

What the pooled OLS model calls ε_i is really the composite error $v_{ij} = u_j + e_{ij}$.

- u_j = classroom effect (shared by all students in classroom j)
- e_{ij} = idiosyncratic student noise (independent across students)
- $v_{ij} = u_j + e_{ij}$ = composite error that pooled OLS lumps together

Two students i and k in the **same classroom** j share the same u_j.

$\Longrightarrow$ Their composite errors v_{ij} and v_{kj} are correlated, even if e_{ij} and e_{kj} are independent.

Consequence: Standard Errors Are Too Small

Recall the OLS variance formula for the slope:

$$\text{Var}(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

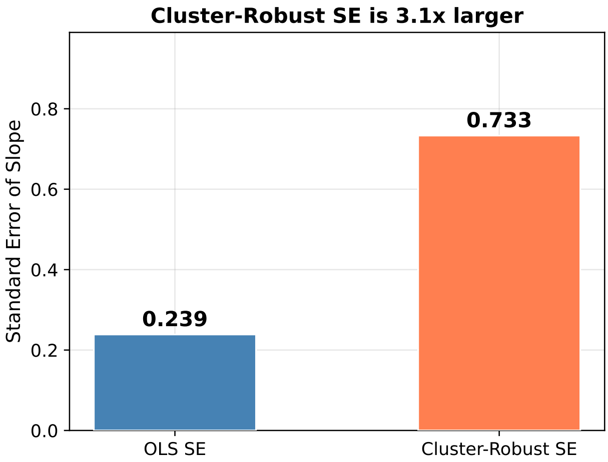
This formula assumes all errors are independent. It drops all $\text{Cov}(v_i, v_k)$ cross-terms. When within-cluster errors are positively correlated, those cross-terms are positive, making the true variance larger.

Intuitively:

- The denominator counts all $n = 240$ observations
- But many of those observations carry **overlapping information**
- The formula "over-counts" the effective information in the data

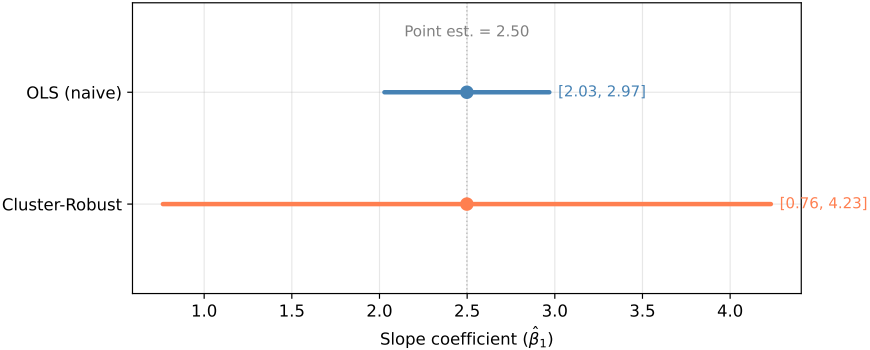
$\implies$ OLS standard errors are **too small**, confidence intervals are **too narrow**, and p-values are **too small**.

How Wrong: SE Comparison



In this dataset, the cluster-robust SE is **3.1x larger** than the naive OLS SE.

Different SEs, Different Conclusions



Same point estimate ($\hat{\beta}_1 = 2.50$), but the cluster-robust CI $[0.76, 4.23]$ is about **3.7x wider** than the OLS CI $[2.03, 2.97]$.

Why wider than 3.1x? The t critical value also changes: $t_{0.025,238} = 1.97$ vs. $t_{0.025,7} = 2.36$, so fewer clusters mean a higher bar for significance.

The narrow OLS interval gives a false sense of precision.

How Correlated: The Intraclass Correlation

How correlated are errors within the same classroom?

Define $\sigma_u^2 =$ variance of the classroom effect u_j, and $\sigma_e^2 =$ variance of the idiosyncratic error e_{ij}.

The **intraclass correlation** is the share of total error variance that comes from the classroom level:

$$\rho = \frac{\sigma_u^2}{\sigma_u^2 + \sigma_e^2}$$

In this dataset, the estimated intraclass correlation is $\hat{\rho} \approx 0.75$: about 75% of the residual variance is **between classrooms**, not between individual students.

This is why Room A's residuals were all negative and Room H's were all positive: the classroom effect u_j dominates.

OLS Assumption Violated

Standard OLS assumes:

$$\text{Corr}(\varepsilon_i, \varepsilon_k) = 0 \quad \text{for all } i \neq k$$

With clustered data:

$$\text{Corr}(v_{ij}, v_{kj}) = \rho \approx 0.75 \neq 0$$

What does OLS "think" it has?

- 240 independent observations $\implies$ 240 independent pieces of information

What does it actually have?

- 240 observations clustered in 8 groups
- About 75% of the residual variance is shared (between-cluster), so 30 students per classroom contribute far less than 30 independent data points
- Effective sample size: $n_{\text{eff}} = \frac{n}{1+(m-1)\hat{\rho}} = \frac{240}{1+29\times 0.75} \approx 11$

Outline

The Cluster-Robust Variance Estimator

Instead of assuming independence, **cluster-robust SEs** allow arbitrary correlation within each cluster.

For a single regressor with G clusters, the idea is:

1. Run pooled OLS as usual $\implies$ get $\hat{\beta}_1$ and residuals $\hat{e}_i$
2. Group the residuals by cluster
3. Compute a variance estimate that accounts for within-cluster correlation

Software handles the calculation. In Stata:

```
reg score hours, vce(cluster classroom)
```

In Python (`statsmodels`):

```
OLS(y, X).fit(cov_type='cluster', cov_kwds={'groups': classroom})
```

Intuition: Why Does Clustering Fix the SE?

Standard OLS: assumes each residual is an independent draw.

Cluster-robust: groups residuals by cluster and asks "how much do they move together?"

Analogy: surveying 240 people about a policy.

- **Scenario 1:** 240 people from 240 different households

  $\Longrightarrow$ Each response is independent. SE is small.

- **Scenario 2:** 240 people from 8 large families (30 per family)

  $\Longrightarrow$ Family members think alike. You really only have $\sim$8 independent opinions.

$\Longrightarrow$ Cluster-robust SEs recognize that 240 people from 8 families carry less information than 240 independent people.

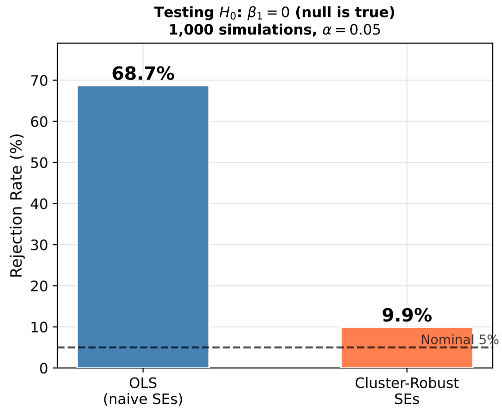
When Does Clustering Increase SEs?

The SE inflation depends on three factors:

1. **Positive within-cluster error correlation** ($\rho > 0$)
   - If $\rho = 0$, cluster-robust SEs $\approx$ OLS SEs
   - The larger ρ, the more inflation
2. **The regressor varies across clusters** (not just within)
   - If all classrooms have the same mean study hours, little inflation
   - If some classrooms study much more than others, large inflation
3. **Large cluster sizes**
   - 8 clusters of 30 students: more inflation
   - 80 clusters of 3 students: less inflation (closer to independence)

In this dataset, all three factors are present: $\hat{\rho} \approx 0.75$, hours vary by classroom, and each classroom has 30 students.

Simulation: How Often Does OLS Wrongly Reject?



We simulated 1,000 datasets where $\beta_1 = 0$ and tested $H_0: \beta_1 = 0$ at $\alpha = 0.05$.

OLS rejects at **68.7%** (should be 5%). Cluster-robust rejects at **9.9%**: still above 5% because with only 8 clusters the CR variance estimate is imprecise; the wider $t(7)$ critical value partially compensates but does not fully correct.

What Clustering Does NOT Fix

Cluster-robust SEs fix the **standard errors**, not the **slope**.

Suppose better-teacher classrooms also assign more study hours. The slope is biased (omitted variable bias). Clustering gives a consistent estimate of the sampling variance of the (biased) slope estimator, but does not correct the bias in the point estimate.

	Clustering alone	**FE + clustering**
SE correct?	Yes	Yes
Slope unbiased?	No (if OVB)	Yes

$\Longrightarrow$ Clustering corrects *inference* (SEs, CIs, p-values). It does not correct *estimation* (the slope itself). For that, you need **fixed effects**.

Comparison: Pooled OLS vs. Pooled+CR vs. FE

	Pooled OLS	**Pooled + CR SE**	**Fixed Effects**
Slope consistent?	Yes*	Yes*	Yes
SEs correct?	No	Yes	Yes**
Handles OVB?	No	No	Yes
Removes u_j?	No	No	Yes
When to use	Baseline	Clustered data, no OVB concern	OVB concern

*Consistent only if $\text{Cov}(u_j, \text{Hours}_{ij}) = 0$ (no omitted variable bias).

**FE standard errors should also be clustered when $n_j > 1$.

$\Longrightarrow$ Cluster-robust SEs and fixed effects solve **different problems**. You often need both: FE to remove bias, plus clustering on the FE residuals to get correct inference.

Outline

When to Cluster

Cluster your standard errors whenever observations share unobserved common shocks:

- **Shared environment:** students in the same classroom, workers in the same firm, patients in the same hospital
- **Group-level treatment:** a policy that affects everyone in a state, a school-wide intervention
- **Repeated observations:** the same individual observed over multiple time periods (panel data)

Rule of thumb: if you can point to a grouping variable that might create shared unobservables, cluster on it. The cost of clustering when it's unnecessary is small (slight efficiency loss). The cost of *not* clustering when you should is large (invalid inference).

What to Cluster On

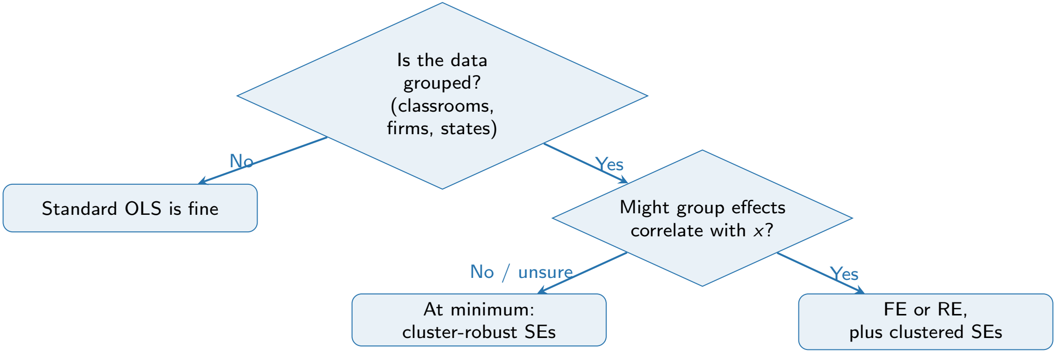
Principle: cluster at the level where treatment varies or common shocks arise.

Setting	Cluster on
Students in classrooms	Classroom
Workers in firms	Firm
State-level policy, individual data	State
Panel data (same person over time)	Individual

How many clusters are enough?

- With too few clusters (< 30), cluster-robust SEs can be unreliable
- In this dataset, we had only 8 clusters $\implies$ the CR test over-rejected slightly (9.9% instead of 5%)
- With few clusters, consider the wild cluster bootstrap (an advanced technique beyond our scope) or small-sample corrections

Decision Flowchart



Outline

Summary

We started with 240 students, a clean regression, and a tight CI. It looked perfect. But the 240 students came from only 8 classrooms, and those 240 observations carried about as much independent information as 11.

1. Pooled OLS on clustered data can give a **reasonable slope estimate**, but the standard errors are **too small** because they ignore within-cluster error correlation.
2. The intraclass correlation ρ measures how much error variance is between clusters vs. within clusters. Higher $\rho \implies$ worse SE distortion.
3. **Cluster-robust SEs** fix the standard errors by allowing arbitrary within-cluster correlation. The point estimate does not change.
4. **Always cluster** when data has a group structure (classrooms, firms, states, panel individuals). The cost of unnecessary clustering is small; the cost of missing it is large.
5. Clustering fixes *inference*, not *estimation*. If group effects are correlated with the regressor (OVB), you also need **fixed effects**.

Thank you!

jakeanderson@g.ucla.edu