Introduction to Fixed Effects

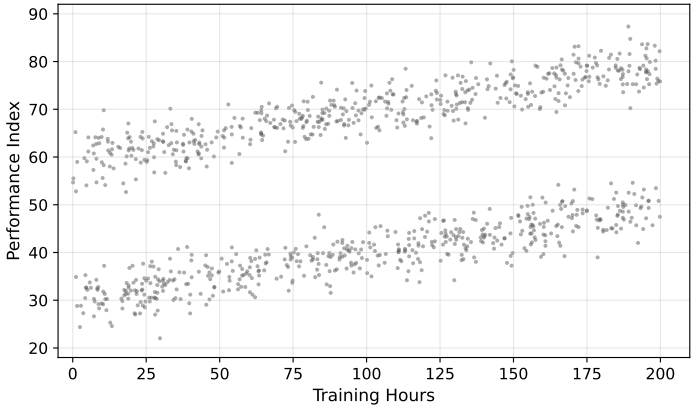
Why One Regression Line Isn't Enough

Jake Anderson

May 16, 2026

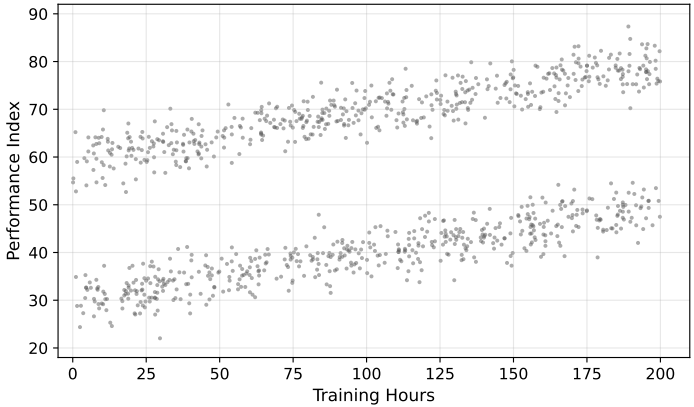
The Data

A coach tracks **training hours** vs. **performance** for their trainees.



The Data

A coach tracks **training hours** vs. **performance** for their trainees.



How could this data be generated?

The Setup

It turns out there are **two teams**: Varsity and Junior Varsity (JV). Same training program, different baseline ability.
Let's assume the following:

- JV players have a baseline of 30 "skill points"; Varsity have 60
- Each additional 10 hours of training $\rightarrow +1$ performance point (slope $= 0.1$)

The Setup

It turns out there are **two teams**: Varsity and Junior Varsity (JV). Same training program, different baseline ability.
Let's assume the following:

- JV players have a baseline of 30 "skill points"; Varsity have 60
- Each additional 10 hours of training $\rightarrow +1$ performance point (slope $= 0.1$)

If we ignore team membership and run a single regression:

$$\text{Performance}_i = \underbrace{\beta_0}_{\substack{\text{Assumes the same intercept} \\ \text{for both teams!}}} + \underbrace{\beta_1}_{\substack{\text{Same slope} \\ \text{for both — OK}}} \text{Hours}_i + \varepsilon_i$$

The Setup

It turns out there are **two teams**: Varsity and Junior Varsity (JV). Same training program, different baseline ability.
Let's assume the following:

- JV players have a baseline of 30 "skill points"; Varsity have 60
- Each additional 10 hours of training $\rightarrow +1$ performance point (slope $= 0.1$)

If we ignore team membership and run a single regression:

$$\text{Performance}_i = \underbrace{\beta_0}_{\substack{\text{Assumes the same intercept} \\ \text{for both teams!}}} + \underbrace{\beta_1}_{\substack{\text{Same slope} \\ \text{for both — OK}}} \text{Hours}_i + \varepsilon_i$$

But JV starts at 30 and Varsity starts at 60 — a single β_0 **cannot be right for both groups**.

The Setup

It turns out there are **two teams**: Varsity and Junior Varsity (JV). Same training program, different baseline ability.
Let's assume the following:

- JV players have a baseline of 30 "skill points"; Varsity have 60
- Each additional 10 hours of training $\rightarrow +1$ performance point (slope $= 0.1$)

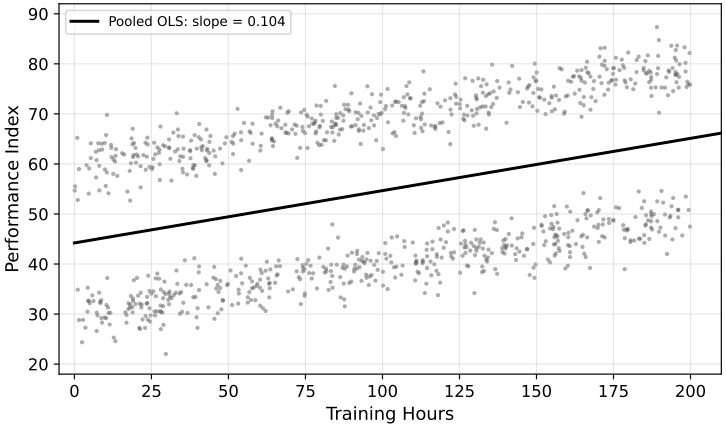
If we ignore team membership and run a single regression:

$$\text{Performance}_i = \underbrace{\beta_0}_{\substack{\text{Assumes the same intercept} \\ \text{for both teams!}}} + \underbrace{\beta_1}_{\substack{\text{Same slope} \\ \text{for both — OK}}} \text{Hours}_i + \varepsilon_i$$

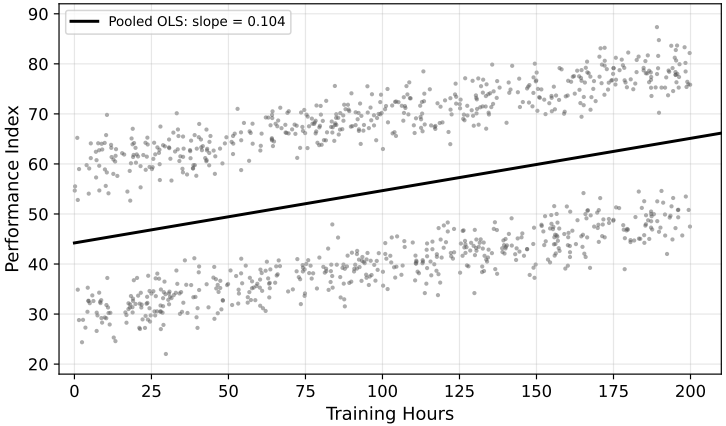
But JV starts at 30 and Varsity starts at 60 — a single β_0 **cannot be right for both groups**.

Question: What goes wrong when we force one intercept on data with two?

What Happens If We Run Naive OLS?



What Happens If We Run Naive OLS?



Pooled OLS: slope ≈ 0.10 — looks correct! But this only works because the sample is balanced (50/50) and x is identically distributed across groups. What if that changes?

First Check: Are Groups Sampled Differently?

In practice, the distribution of x often differs across groups.

- Varsity players may train **more** (selection into training)
- Or JV players may train more (catching up)

First Check: Are Groups Sampled Differently?

In practice, the distribution of x often differs across groups.

- Varsity players may train **more** (selection into training)
- Or JV players may train more (catching up)

Key question: Does $\text{Cov}(\text{Group}, x) \neq 0$?

First Check: Are Groups Sampled Differently?

In practice, the distribution of x often differs across groups.

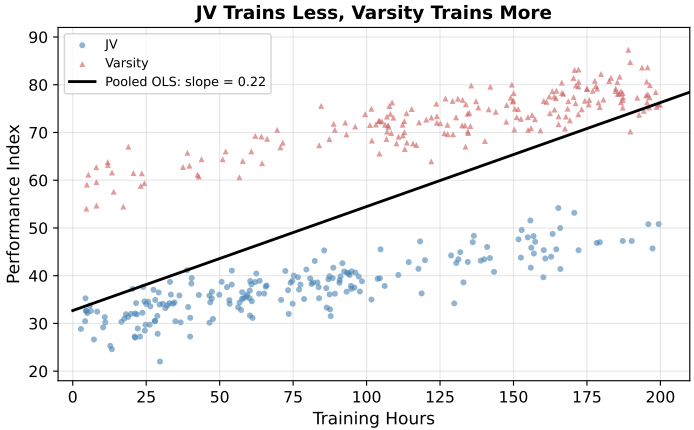
- Varsity players may train **more** (selection into training)
- Or JV players may train more (catching up)

Key question: Does $\text{Cov}(\text{Group}, x) \neq 0$?

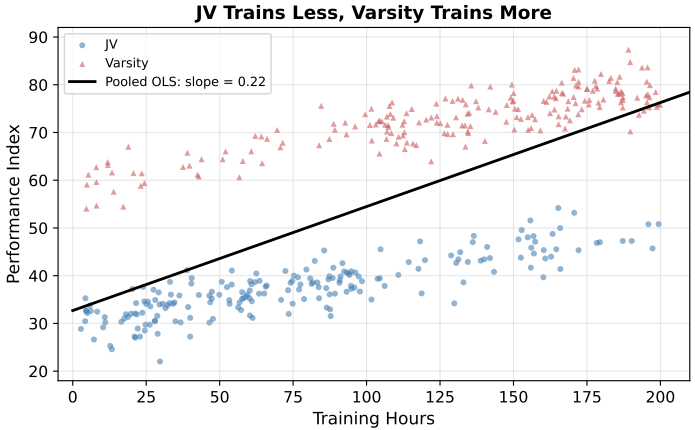
If so, pooled OLS conflates the group effect (α_j) with the treatment effect (β).

This is **omitted variable bias**.

Scenario: Varsity Trains More

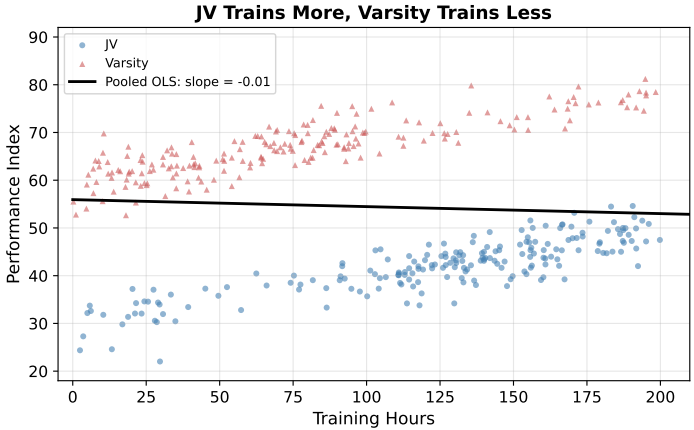


Scenario: Varsity Trains More

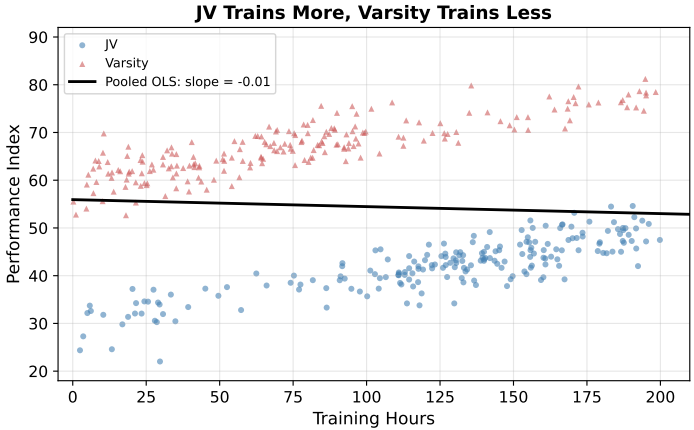


OLS slope = **0.22** (true = 0.10). Bias is **positive**: OLS attributes Varsity's higher baseline to their greater training hours.

Scenario: JV Trains More (Catching Up)



Scenario: JV Trains More (Catching Up)



OLS slope $= -\mathbf{0.02}$ (true $= 0.10$). Bias is **negative**: OLS thinks training has nearly *zero effect* because the high-training group has lower baseline ability.

The Omitted Variable Bias

Let's map this to the OVB framework you already know. Let X_2 = Group / team membership (OV)

The Omitted Variable Bias

Let's map this to the OVB framework you already know. Let X_2 = Group / team membership (OV)

Short regression (what we run → omits X_2):

$$Y_i = \beta_0 + \beta_1 X_{1i} + u_i$$

The Omitted Variable Bias

Let's map this to the OVB framework you already know. Let X_2 = Group / team membership (OV)

Short regression (what we run $\rightarrow$ omits X_2):

$$Y_i = \beta_0 + \beta_1 X_{1i} + u_i$$

Long regression (what we should run $\rightarrow$ includes X_2):

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i$$

The Omitted Variable Bias

Let's map this to the OVB framework you already know. Let X_2 = Group / team membership (OV)

Short regression (what we run $\rightarrow$ omits X_2):

$$Y_i = \beta_0 + \beta_1 X_{1i} + u_i$$

Long regression (what we should run $\rightarrow$ includes X_2):

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i$$

Auxiliary regression (relationship between included and omitted):

$$X_{2i} = \delta_0 + \delta_1 X_{1i} + v_i$$

The Omitted Variable Bias

Let's map this to the OVB framework you already know. Let X_2 = Group / team membership (OV)

Short regression (what we run $\rightarrow$ omits X_2):

$$Y_i = \beta_0 + \beta_1 X_{1i} + u_i$$

Long regression (what we should run $\rightarrow$ includes X_2):

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i$$

Auxiliary regression (relationship between included and omitted):

$$X_{2i} = \delta_0 + \delta_1 X_{1i} + v_i$$

OVB formula: (effect of X_2 on Y) $\times$ (relationship of X_2 to X_1)

$$\hat{\beta}_1^{\text{short}} = \hat{\beta}_1^{\text{long}} + \hat{\beta}_2 \times \hat{\delta}_1$$

OVB Applied to Our Example

$$\hat{\beta}_1^{\text{short}} = \underbrace{\hat{\beta}_1^{\text{long}}}_{=0.10} + \underbrace{\hat{\beta}_2}_{\substack{\text{effect of group} \\ \text{on Performance}}} \times \underbrace{\hat{\delta}_1}_{\substack{\text{relationship of} \\ \text{group to Hours}}}$$

OVB Applied to Our Example

$$\hat{\beta}_1^{\text{short}} = \underbrace{\hat{\beta}_1^{\text{long}}}_{=0.10} + \underbrace{\hat{\beta}_2}_{\substack{\text{effect of group} \\ \text{on Performance}}} \times \underbrace{\hat{\delta}_1}_{\substack{\text{relationship of} \\ \text{group to Hours}}}$$

Varsity has higher baseline performance, so $\hat{\beta}_2 > 0$.

OVB Applied to Our Example

$$\hat{\beta}_1^{\text{short}} = \underbrace{\hat{\beta}_1^{\text{long}}}_{=0.10} + \underbrace{\hat{\beta}_2}_{\substack{\text{effect of group} \\ \text{on Performance}}} \times \underbrace{\hat{\delta}_1}_{\substack{\text{relationship of} \\ \text{group to Hours}}}$$

Varsity has higher baseline performance, so $\hat{\beta}_2 > 0$.

Scenario	$\hat{\delta}_1$	**Bias** $(\hat{\beta}_2 \times \hat{\delta}_1)$	**OLS slope**
Varsity trains more	> 0	$+$	0.22
JV trains more	< 0	$-$	-0.02
Equal training	≈ 0	≈ 0	0.10

OVB Applied to Our Example

$$\hat{\beta}_1^{\text{short}} = \underbrace{\hat{\beta}_1^{\text{long}}}_{=0.10} + \underbrace{\hat{\beta}_2}_{\substack{\text{effect of group} \\ \text{on Performance}}} \times \underbrace{\hat{\delta}_1}_{\substack{\text{relationship of} \\ \text{group to Hours}}}$$

Varsity has higher baseline performance, so $\hat{\beta}_2 > 0$.

Scenario	$\hat{\delta}_1$	**Bias** ($\hat{\beta}_2 \times \hat{\delta}_1$)	**OLS slope**
Varsity trains more	> 0	$+$	0.22
JV trains more	< 0	$-$	-0.02
Equal training	≈ 0	≈ 0	0.10

Same data, same true effect. The OLS estimate swings from -0.02 to $+0.22$ just by changing which group trains more.

What About Class Imbalance?

Suppose x is distributed the same across groups ($\delta \approx 0$), but the **sample composition** is unbalanced.

- Does the slope change?
- What about the intercept?

What About Class Imbalance?

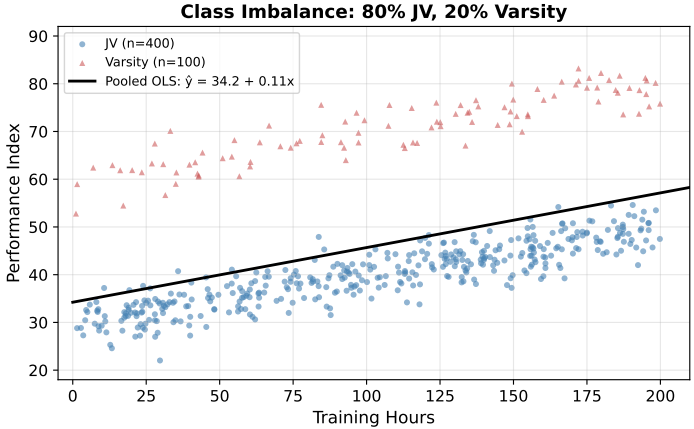
Suppose x is distributed the same across groups ($\delta \approx 0$), but the **sample composition** is unbalanced.

- Does the slope change?
- What about the intercept?

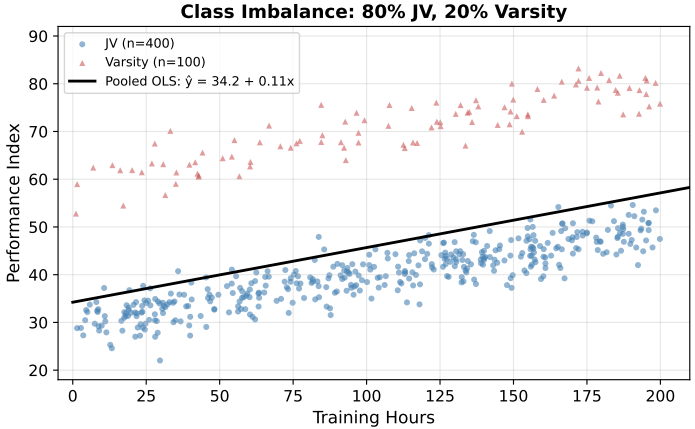
The OLS intercept is a **weighted average** of the group intercepts:

$$\begin{aligned}\hat{\beta}_0 &= \frac{\sum_{i=1}^{n} \mathbb{1}\{i \in \text{JV}\}}{n} \cdot \beta_{0,\text{JV}} + \frac{\sum_{i=1}^{n} \mathbb{1}\{i \in \text{Var}\}}{n} \cdot \beta_{0,\text{Var}} \\ &= \frac{n_{\text{JV}}}{n} \cdot \beta_{0,\text{JV}} + \frac{n_{\text{Var}}}{n} \cdot \beta_{0,\text{Var}} \\ &= (\text{Share JV}) \cdot \beta_{0,\text{JV}} + (\text{Share Var}) \cdot \beta_{0,\text{Var}}\end{aligned}$$

80% JV, 20% Varsity

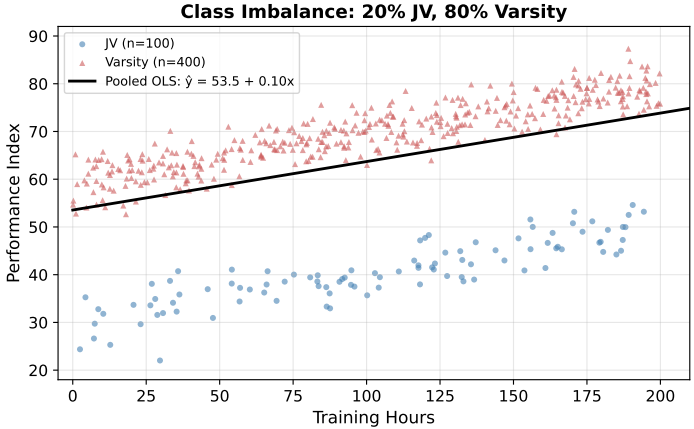


80% JV, 20% Varsity

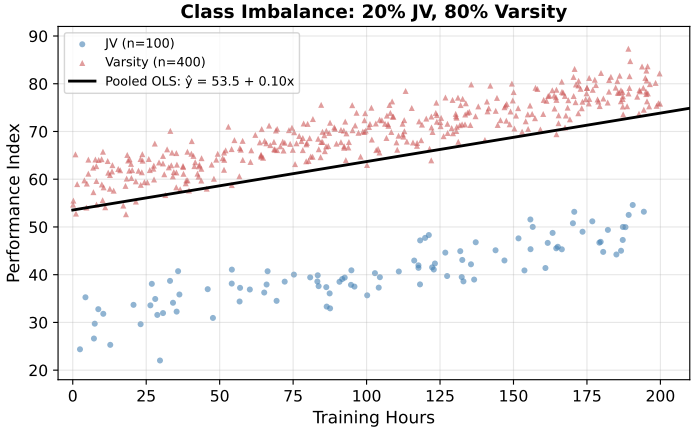


Slope $\approx$ correct. Intercept = **34** — pulled toward JV's true intercept (30). Predictions are wrong for most Varsity players.

20% JV, 80% Varsity



20% JV, 80% Varsity



Slope $\approx$ correct. Intercept = **54** — pulled toward Varsity's true intercept (60). Now predictions are wrong for most JV players.

Class Imbalance: The Intercept Shifts

Sample	**OLS intercept**	**True JV** ($\alpha = 30$)	**True Var** ($\alpha = 60$)
80% JV	34	close	off by 26
50/50	44	off by 14	off by 16
80% Varsity	54	off by 24	close

Class Imbalance: The Intercept Shifts

Sample	**OLS intercept**	**True JV** ($\alpha = 30$)	**True Var** ($\alpha = 60$)
80% JV	34	close	off by 26
50/50	44	off by 14	off by 16
80% Varsity	54	off by 24	close

$\Longrightarrow$ Even when the slope is approximately correct, pooled OLS uses a **single intercept** that is wrong for every subgroup. The error depends on sample composition, which the researcher may not control.

A Separate Intercept per Unit

Every "naive" example we just saw used the same model:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

A Separate Intercept per Unit

Every "naive" example we just saw used the same model:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

A single β_0 forces **one baseline** for all groups. That's where the bias comes from.

A Separate Intercept per Unit

Every "naive" example we just saw used the same model:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

A single β_0 forces **one baseline** for all groups. That's where the bias comes from.

The fix: let each group j have its own intercept α_j, where $j = \text{JV}$ or $j = \text{Varsity}$:

$$y_i = \alpha_j + \beta_1 x_i + \varepsilon_i$$

A Separate Intercept per Unit

Every "naive" example we just saw used the same model:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

A single β_0 forces **one baseline** for all groups. That's where the bias comes from.

The fix: let each group j have its own intercept α_j, where j = JV or j = Varsity:

$$y_i = \alpha_j + \beta_1 x_i + \varepsilon_i$$

The subscript j on the intercept is doing all the work:

- $\beta_0 \rightarrow$ one number, shared by everyone
- $\alpha_j \rightarrow$ a **different number for each group**

A Separate Intercept per Unit

Every "naive" example we just saw used the same model:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

A single β_0 forces **one baseline** for all groups. That's where the bias comes from.

The fix: let each group j have its own intercept α_j, where $j = \text{JV}$ or $j = \text{Varsity}$:

$$y_i = \alpha_j + \beta_1 x_i + \varepsilon_i$$

The subscript j on the intercept is doing all the work:

- $\beta_0 \rightarrow$ one number, shared by everyone
- $\alpha_j \rightarrow$ a **different number for each group**

This is the core idea behind **fixed effects**.

The Fixed Effects Model

Allow each group its own intercept:

$$\text{Performance}_{ij} = \alpha_j + \beta\,\text{Hours}_{ij} + \varepsilon_{ij}$$

The Fixed Effects Model

Allow each group its own intercept:

$$\text{Performance}_{ij} = \alpha_j + \beta\,\text{Hours}_{ij} + \varepsilon_{ij}$$

Equivalently, add a group dummy:

$$y_i = \beta\,x_i + \alpha_{\text{JV}} \cdot \mathbb{1}\{\text{JV}\} + \alpha_{\text{Var}} \cdot \mathbb{1}\{\text{Var}\} + \varepsilon_i$$

The Fixed Effects Model

Allow each group its own intercept:

$$\text{Performance}_{ij} = \alpha_j + \beta\,\text{Hours}_{ij} + \varepsilon_{ij}$$

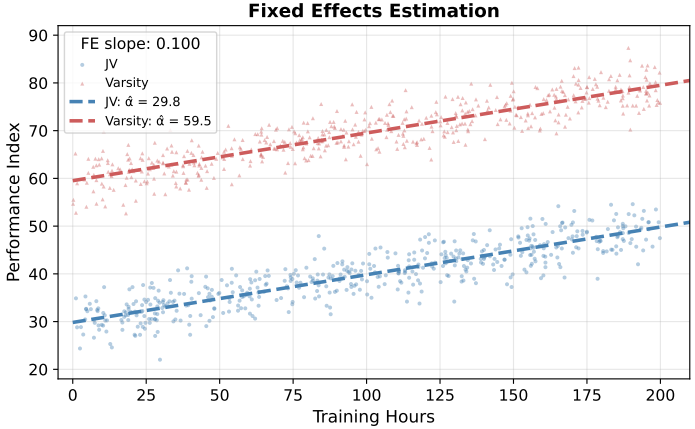
Equivalently, add a group dummy:

$$y_i = \beta\, x_i + \alpha_{\text{JV}} \cdot \mathbb{1}\{\text{JV}\} + \alpha_{\text{Var}} \cdot \mathbb{1}\{\text{Var}\} + \varepsilon_i$$

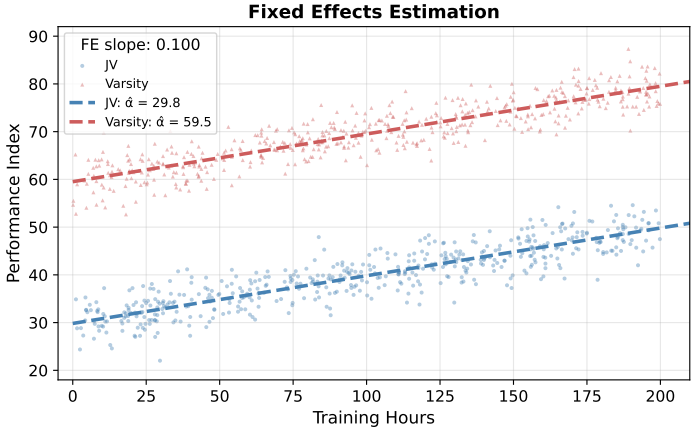
Key idea: FE estimates β using only *within-group* variation in x.

It asks: "Among JV players, do those who train more perform better?"

FE Estimation: The Result

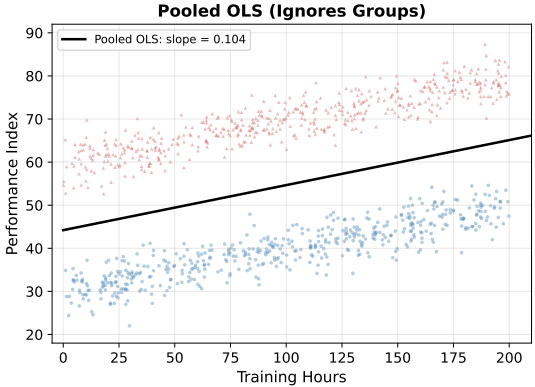


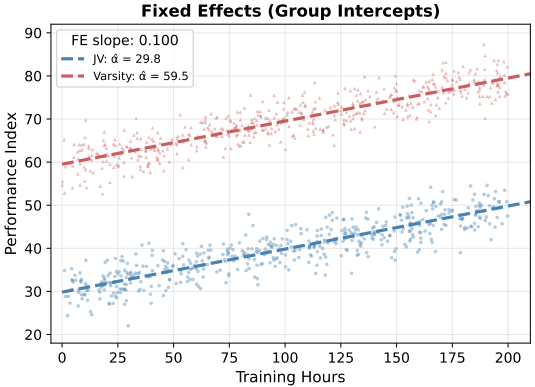
FE Estimation: The Result



FE slope = **0.100** (true = 0.10). Intercepts: JV = **29.8** (true 30), Varsity = **59.5** (true 60).

Pooled OLS vs. Fixed Effects





Why Does This Work? The Rank Condition

Basic idea: You need at least **two points to draw a line** (one intercept + one slope = 2 unknowns).

Why Does This Work? The Rank Condition

Basic idea: You need at least **two points to draw a line** (one intercept + one slope = 2 unknowns).

Pooled OLS: 2 unknowns (β_0, β_1) $\implies$ need ≥ 2 observations total. Easy.

Why Does This Work? The Rank Condition

Basic idea: You need at least **two points to draw a line** (one intercept + one slope = 2 unknowns).

Pooled OLS: 2 unknowns (β_0, β_1) $\implies$ need ≥ 2 observations total. Easy.

Fixed effects with 2 groups: 3 unknowns (α_{JV}, α_{Var}, β_1).

- You need ≥ 2 observations *per group* (to pin down each group's line)
- So at minimum: **4 observations** (2 JV + 2 Varsity)

Why Does This Work? The Rank Condition

Basic idea: You need at least **two points to draw a line** (one intercept + one slope = 2 unknowns).

Pooled OLS: 2 unknowns (β_0, β_1) $\implies$ need ≥ 2 observations total. Easy.

Fixed effects with 2 groups: 3 unknowns (α_{JV}, α_{Var}, β_1).

- You need ≥ 2 observations *per group* (to pin down each group's line)
- So at minimum: **4 observations** (2 JV + 2 Varsity)

In general: with J groups, you have $J + 1$ unknowns (J intercepts + 1 slope), so you need at least 2 observations per group to be identified.

Why Does This Work? The Rank Condition

Basic idea: You need at least **two points to draw a line** (one intercept + one slope = 2 unknowns).

Pooled OLS: 2 unknowns (β_0, β_1) $\implies$ need ≥ 2 observations total. Easy.

Fixed effects with 2 groups: 3 unknowns (α_{JV}, α_{Var}, β_1).

- You need ≥ 2 observations *per group* (to pin down each group's line)
- So at minimum: **4 observations** (2 JV + 2 Varsity)

In general: with J groups, you have $J + 1$ unknowns (J intercepts + 1 slope), so you need at least 2 observations per group to be identified.

$\implies$ More groups = more unknowns = more data required.

Connection to Panel Data

Our training example maps directly to the panel data framework:

$$y_{it} = \alpha_i + \beta\, x_{it} + \varepsilon_{it}$$

- i = individual (trainee $\rightarrow$ firm, person, country)
- t = time period
- α_i = **individual fixed effect** (unobserved, time-invariant)

Connection to Panel Data

Our training example maps directly to the panel data framework:

$$y_{it} = \alpha_i + \beta\, x_{it} + \varepsilon_{it}$$

- i = individual (trainee $\rightarrow$ firm, person, country)
- t = time period
- α_i = **individual fixed effect** (unobserved, time-invariant)

Two equivalent estimation approaches:

1. **Least Squares Dummy Variable** (when we want all of the individual fixed effects):

$$y_{it} = \beta\, x_{it} + \sum_i \alpha_i D_i + \varepsilon_{it}$$

2. **Within / Demeaning Estimator:** Subtract individual means

$$(y_{it} - \bar{y}_i) = \beta(x_{it} - \bar{x}_i) + (\varepsilon_{it} - \bar{\varepsilon}_i)$$

Connection to Panel Data

Our training example maps directly to the panel data framework:

$$y_{it} = \alpha_i + \beta\, x_{it} + \varepsilon_{it}$$

- i = individual (trainee $\rightarrow$ firm, person, country)
- t = time period
- α_i = **individual fixed effect** (unobserved, time-invariant)

Two equivalent estimation approaches:

1. **Least Squares Dummy Variable** (when we want all of the individual fixed effects):

$$y_{it} = \beta\, x_{it} + \sum_i \alpha_i D_i + \varepsilon_{it}$$

2. **Within / Demeaning Estimator:** Subtract individual means

$$(y_{it} - \bar{y}_i) = \beta(x_{it} - \bar{x}_i) + (\varepsilon_{it} - \bar{\varepsilon}_i)$$

Both give the **same** $\hat{\beta}$. Let's walk through the within estimator step by step.

The Within Estimator: Step 1 $\rightarrow$ Compute Group Means

Start with the model:

$$y_{it} = \alpha_i + \beta\, x_{it} + \varepsilon_{it}$$

The Within Estimator: Step 1 $\rightarrow$ Compute Group Means

Start with the model:

$$y_{it} = \alpha_i + \beta\, x_{it} + \varepsilon_{it}$$

For each individual i, take the **time average** of both sides:

$$\bar{y}_i = \alpha_i + \beta\, \bar{x}_i + \bar{\varepsilon}_i$$

The Within Estimator: Step 1 $\rightarrow$ Compute Group Means

Start with the model:

$$y_{it} = \alpha_i + \beta\, x_{it} + \varepsilon_{it}$$

For each individual i, take the **time average** of both sides:

$$\bar{y}_i = \alpha_i + \beta\, \bar{x}_i + \bar{\varepsilon}_i$$

where:

- $\bar{y}_i = \frac{1}{T}\sum_{t=1}^{T} y_{it}$ (mean outcome for individual i)
- $\bar{x}_i = \frac{1}{T}\sum_{t=1}^{T} x_{it}$ (mean regressor for individual i)

The Within Estimator: Step 1 → Compute Group Means

Start with the model:

$$y_{it} = \alpha_i + \beta\, x_{it} + \varepsilon_{it}$$

For each individual i, take the **time average** of both sides:

$$\bar{y}_i = \alpha_i + \beta\, \bar{x}_i + \bar{\varepsilon}_i$$

where:

- $\bar{y}_i = \frac{1}{T}\sum_{t=1}^{T} y_{it}$ (mean outcome for individual i)
- $\bar{x}_i = \frac{1}{T}\sum_{t=1}^{T} x_{it}$ (mean regressor for individual i)

Notice that α_i survives averaging because it **doesn't vary over time**.

The Within Estimator: Step 2 $\rightarrow$ Subtract

Subtract the individual mean equation from the original:

$$y_{it} - \bar{y}_i = (\alpha_i - \alpha_i) + \beta(x_{it} - \bar{x}_i) + (\varepsilon_{it} - \bar{\varepsilon}_i)$$

The Within Estimator: Step 2 $\rightarrow$ Subtract

Subtract the individual mean equation from the original:

$$y_{it} - \bar{y}_i = (\alpha_i - \alpha_i) + \beta(x_{it} - \bar{x}_i) + (\varepsilon_{it} - \bar{\varepsilon}_i)$$

The fixed effect cancels:

$$\ddot{y}_{it} = \beta\, \ddot{x}_{it} + \ddot{\varepsilon}_{it}$$

where $\ddot{y}_{it} \equiv y_{it} - \bar{y}_i$ is the **demeaned** variable.

The Within Estimator: Step 2 $\rightarrow$ Subtract

Subtract the individual mean equation from the original:

$$y_{it} - \bar{y}_i = (\alpha_i - \alpha_i) + \beta(x_{it} - \bar{x}_i) + (\varepsilon_{it} - \bar{\varepsilon}_i)$$

The fixed effect cancels:

$$\ddot{y}_{it} = \beta\, \ddot{x}_{it} + \ddot{\varepsilon}_{it}$$

where $\ddot{y}_{it} \equiv y_{it} - \bar{y}_i$ is the **demeaned** variable.

$\Longrightarrow$ We have a simple regression with **no intercept** and no α_i.

The Within Estimator: Step 2 $\rightarrow$ Subtract

Subtract the individual mean equation from the original:

$$y_{it} - \bar{y}_i = (\alpha_i - \alpha_i) + \beta(x_{it} - \bar{x}_i) + (\varepsilon_{it} - \bar{\varepsilon}_i)$$

The fixed effect cancels:

$$\ddot{y}_{it} = \beta\, \ddot{x}_{it} + \ddot{\varepsilon}_{it}$$

where $\ddot{y}_{it} \equiv y_{it} - \bar{y}_i$ is the **demeaned** variable.

$\Longrightarrow$ We have a simple regression with **no intercept** and no α_i.
Just run OLS on the demeaned data.

The Within Estimator: Step 3 $\rightarrow$ Estimate β

OLS on the demeaned regression gives:

$$\hat{\beta} = \frac{\text{Cov}(\ddot{x}_{it},\, \ddot{y}_{it})}{\text{Var}(\ddot{x}_{it})} = \frac{\sum_i \sum_t (x_{it} - \bar{x}_i)(y_{it} - \bar{y}_i)}{\sum_i \sum_t (x_{it} - \bar{x}_i)^2}$$

The Within Estimator: Step 3 $\rightarrow$ Estimate β

OLS on the demeaned regression gives:

$$\hat{\beta} = \frac{\text{Cov}(\ddot{x}_{it},\, \ddot{y}_{it})}{\text{Var}(\ddot{x}_{it})} = \frac{\sum_i \sum_t (x_{it} - \bar{x}_i)(y_{it} - \bar{y}_i)}{\sum_i \sum_t (x_{it} - \bar{x}_i)^2}$$

$\implies$ The covariance of the demeaned variables equals the covariance of the original variables:

$$\text{Cov}(\ddot{x}_{it},\, \ddot{y}_{it}) = \text{Cov}(x_{it} - \bar{x}_i,\, y_{it} - \bar{y}_i) = \text{Cov}(x_{it},\, y_{it})$$

The Within Estimator: Step 3 $\rightarrow$ Estimate β

OLS on the demeaned regression gives:

$$\hat{\beta} = \frac{\text{Cov}(\ddot{x}_{it},\, \ddot{y}_{it})}{\text{Var}(\ddot{x}_{it})} = \frac{\sum_i \sum_t (x_{it} - \bar{x}_i)(y_{it} - \bar{y}_i)}{\sum_i \sum_t (x_{it} - \bar{x}_i)^2}$$

$\implies$ The covariance of the demeaned variables equals the covariance of the original variables:

$$\text{Cov}(\ddot{x}_{it},\, \ddot{y}_{it}) = \text{Cov}(x_{it} - \bar{x}_i,\, y_{it} - \bar{y}_i) = \text{Cov}(x_{it},\, y_{it})$$

Why? Let's show it briefly.

Proof: Demeaning Doesn't Change Covariance

Let $c_i = \bar{x}_i$ and $d_i = \bar{y}_i$ (constants within group i). Then:

$$\begin{aligned}\text{Cov}(x_{it} - c_i,\ y_{it} - d_i) = \text{Cov}(x_{it},\ y_{it}) - \text{Cov}(x_{it},\ d_i) \\ - \text{Cov}(c_i,\ y_{it}) + \text{Cov}(c_i,\ d_i)\end{aligned}$$

Proof: Demeaning Doesn't Change Covariance

Let $c_i = \bar{x}_i$ and $d_i = \bar{y}_i$ (constants within group i). Then:

$$\begin{aligned}\text{Cov}(x_{it} - c_i,\ y_{it} - d_i) = \text{Cov}(x_{it},\ y_{it}) - \text{Cov}(x_{it},\ d_i) \\ - \text{Cov}(c_i,\ y_{it}) + \text{Cov}(c_i,\ d_i)\end{aligned}$$

Within each group i, the means c_i and d_i are **constants**, so:

$$\text{Cov}(x_{it},\ d_i) = 0, \quad \text{Cov}(c_i,\ y_{it}) = 0, \quad \text{Cov}(c_i,\ d_i) = 0$$

Proof: Demeaning Doesn't Change Covariance

Let $c_i = \bar{x}_i$ and $d_i = \bar{y}_i$ (constants within group i). Then:

$$\begin{aligned}\text{Cov}(x_{it} - c_i,\ y_{it} - d_i) = \text{Cov}(x_{it},\ y_{it}) - \text{Cov}(x_{it},\ d_i) \\ - \text{Cov}(c_i,\ y_{it}) + \text{Cov}(c_i,\ d_i)\end{aligned}$$

Within each group i, the means c_i and d_i are **constants**, so:

$$\text{Cov}(x_{it},\ d_i) = 0, \quad \text{Cov}(c_i,\ y_{it}) = 0, \quad \text{Cov}(c_i,\ d_i) = 0$$

Therefore:

$$\text{Cov}(x_{it} - \bar{x}_i,\ y_{it} - \bar{y}_i) = \text{Cov}(x_{it},\ y_{it}) \qquad \blacksquare$$

Proof: Demeaning Doesn't Change Covariance

Let $c_i = \bar{x}_i$ and $d_i = \bar{y}_i$ (constants within group i). Then:

$$\begin{aligned}\mathrm{Cov}(x_{it} - c_i,\ y_{it} - d_i) = \mathrm{Cov}(x_{it},\ y_{it}) - \mathrm{Cov}(x_{it},\ d_i) \\ - \mathrm{Cov}(c_i,\ y_{it}) + \mathrm{Cov}(c_i,\ d_i)\end{aligned}$$

Within each group i, the means c_i and d_i are **constants**, so:

$$\mathrm{Cov}(x_{it},\ d_i) = 0, \quad \mathrm{Cov}(c_i,\ y_{it}) = 0, \quad \mathrm{Cov}(c_i,\ d_i) = 0$$

Therefore:

$$\mathrm{Cov}(x_{it} - \bar{x}_i,\ y_{it} - \bar{y}_i) = \mathrm{Cov}(x_{it},\ y_{it}) \qquad \blacksquare$$

$\Longrightarrow$ The within estimator uses **only within-group variation**. All between-group differences are "absorbed" by the fixed effects.

Thank you!
jakeanderson@g.ucla.edu

Introduction to Random Effects

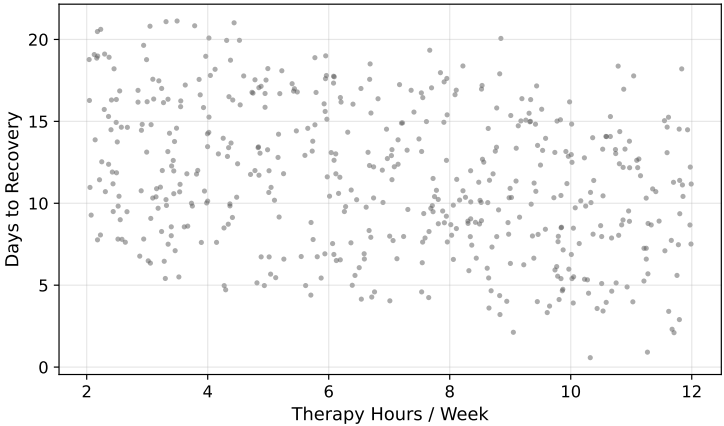
From Fixed Parameters to Random Draws

Jake Anderson

May 16, 2026

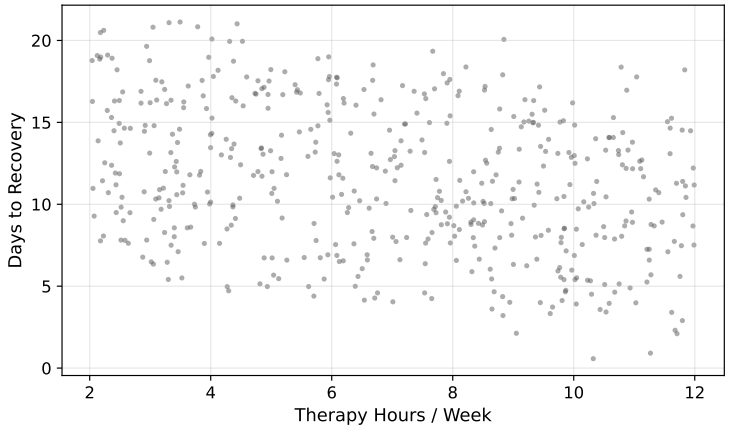
The Data

A researcher tracks **days to recovery** vs. **hours of physical therapy per week** across patients at several hospitals.



The Data

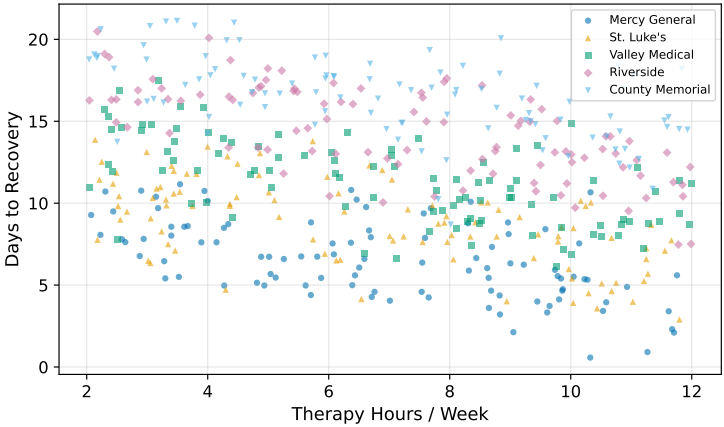
A researcher tracks **days to recovery** vs. **hours of physical therapy per week** across patients at several hospitals.



How could this data be generated?

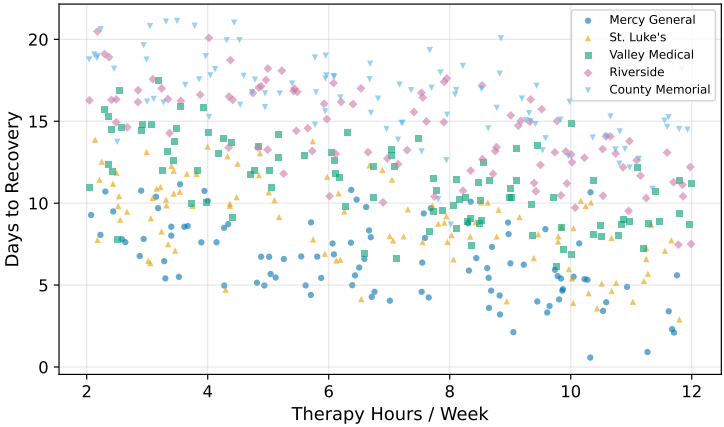
Reveal: Five Hospitals

Patients come from **5 hospitals**, each with a different baseline recovery time.



Reveal: Five Hospitals

Patients come from **5 hospitals**, each with a different baseline recovery time.



Recall: ignoring groups biases OLS. FE solved this by giving each group its own intercept.

FE Recap: Group-Specific Intercepts

The **fixed effects** model treats each hospital's baseline as a fixed unknown:

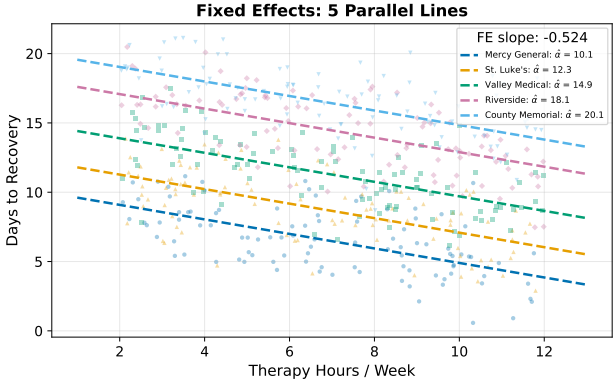
$$y_{ij} = \alpha_j + \beta\, x_{ij} + \varepsilon_{ij}$$

FE Recap: Group-Specific Intercepts

The **fixed effects** model treats each hospital's baseline as a fixed unknown:

$$y_{ij} = \alpha_j + \beta\, x_{ij} + \varepsilon_{ij}$$

FE estimates a **separate intercept** for each hospital:



The Cost of FE

With 5 hospitals, FE estimates **5 separate intercepts**. That's manageable.

The Cost of FE

With 5 hospitals, FE estimates **5 separate intercepts**. That's manageable.

But what if you have:

- 50 hospitals? $\rightarrow$ 50 intercepts
- 500 hospitals? $\rightarrow$ 500 intercepts
- 5,000 hospitals? $\rightarrow$ 5,000 intercepts

The Cost of FE

With 5 hospitals, FE estimates **5 separate intercepts**. That's manageable.

But what if you have:

- 50 hospitals? $\rightarrow$ 50 intercepts
- 500 hospitals? $\rightarrow$ 500 intercepts
- 5,000 hospitals? $\rightarrow$ 5,000 intercepts

Two problems with FE:

1. Uses up degrees of freedom (one parameter per group)
2. **Cannot** estimate the effect of time-invariant variables

The Cost of FE

With 5 hospitals, FE estimates **5 separate intercepts**. That's manageable.

But what if you have:

- 50 hospitals? $\rightarrow$ 50 intercepts
- 500 hospitals? $\rightarrow$ 500 intercepts
- 5,000 hospitals? $\rightarrow$ 5,000 intercepts

Two problems with FE:

1. Uses up degrees of freedom (one parameter per group)
2. **Cannot** estimate the effect of time-invariant variables

For example: "Do *teaching* hospitals have faster recovery?" Teaching status doesn't change, so FE absorbs it into α_j.

The Cost of FE

With 5 hospitals, FE estimates **5 separate intercepts**. That's manageable.

But what if you have:

- 50 hospitals? $\rightarrow$ 50 intercepts
- 500 hospitals? $\rightarrow$ 500 intercepts
- 5,000 hospitals? $\rightarrow$ 5,000 intercepts

Two problems with FE:

1. Uses up degrees of freedom (one parameter per group)
2. **Cannot** estimate the effect of time-invariant variables

For example: "Do *teaching* hospitals have faster recovery?" Teaching status doesn't change, so FE absorbs it into α_j.

$\Longrightarrow$ What if we could model the group effects with **fewer parameters**?

The RE Assumption

Fixed effects: Each α_j is a fixed, unknown parameter to estimate.

$$y_{ij} = \alpha_j + \beta\, x_{ij} + \varepsilon_{ij}$$

The RE Assumption

Fixed effects: Each α_j is a fixed, unknown parameter to estimate.

$$y_{ij} = \alpha_j + \beta x_{ij} + \varepsilon_{ij}$$

Random effects: Decompose α_j into a common mean plus a random deviation:

$$\alpha_j = \bar{\alpha} + u_j \qquad \text{where } u_j \sim (0, \sigma_u^2)$$

The RE Assumption

Fixed effects: Each α_j is a fixed, unknown parameter to estimate.

$$y_{ij} = \alpha_j + \beta\, x_{ij} + \varepsilon_{ij}$$

Random effects: Decompose α_j into a common mean plus a random deviation:

$$\alpha_j = \bar{\alpha} + u_j \qquad \text{where } u_j \sim (0, \sigma_u^2)$$

- $\bar{\alpha}$ = average baseline across *all* hospitals
- u_j = hospital j's **deviation** from that average (same role as α_j, but a random draw instead of a free parameter)
- σ_u^2 = how spread out hospital baselines are

The RE Assumption

Fixed effects: Each α_j is a fixed, unknown parameter to estimate.

$$y_{ij} = \alpha_j + \beta\, x_{ij} + \varepsilon_{ij}$$

Random effects: Decompose α_j into a common mean plus a random deviation:

$$\alpha_j = \bar{\alpha} + u_j \qquad \text{where } u_j \sim (0, \sigma_u^2)$$

- $\bar{\alpha}$ = average baseline across *all* hospitals
- u_j = hospital j's **deviation** from that average (same role as α_j, but a random draw instead of a free parameter)
- σ_u^2 = how spread out hospital baselines are

Instead of estimating 5 (or 500) separate α_j's, we estimate $\bar{\alpha}$ and **one variance** σ_u^2.

The RE Assumption

Fixed effects: Each α_j is a fixed, unknown parameter to estimate.

$$y_{ij} = \alpha_j + \beta\, x_{ij} + \varepsilon_{ij}$$

Random effects: Decompose α_j into a common mean plus a random deviation:

$$\alpha_j = \bar{\alpha} + u_j \qquad \text{where } u_j \sim (0, \sigma_u^2)$$

- $\bar{\alpha}$ = average baseline across *all* hospitals
- u_j = hospital j's **deviation** from that average (same role as α_j, but a random draw instead of a free parameter)
- σ_u^2 = how spread out hospital baselines are

Instead of estimating 5 (or 500) separate α_j's, we estimate $\bar{\alpha}$ and **one variance** σ_u^2.

The critical assumption:

$$\boxed{\text{Cov}(u_j, x_{ij}) = 0}$$

i.e., patients at a **better** (worse) hospital can't tend to receive **more** (less) therapy.

The Error Components Model

Substitute $\alpha_j = \bar{\alpha} + u_j$ into the FE model:

$$y_{ij} = \bar{\alpha} + \beta x_{ij} + \underbrace{u_j + e_{ij}}_{v_{ij}}$$

The Error Components Model

Substitute $\alpha_j = \bar{\alpha} + u_j$ into the FE model:

$$y_{ij} = \bar{\alpha} + \beta\, x_{ij} + \underbrace{u_j + e_{ij}}_{v_{ij}}$$

where:

- u_j = hospital-specific random component (same for all patients at hospital j)
- e_{ij} = idiosyncratic error (patient-specific noise)
- $v_{ij} = u_j + e_{ij}$ = **composite error**

The Error Components Model

Substitute $\alpha_j = \bar{\alpha} + u_j$ into the FE model:

$$y_{ij} = \bar{\alpha} + \beta\, x_{ij} + \underbrace{u_j + e_{ij}}_{v_{ij}}$$

where:

- u_j = hospital-specific random component (same for all patients at hospital j)
- e_{ij} = idiosyncratic error (patient-specific noise)
- $v_{ij} = u_j + e_{ij}$ = **composite error**

Problem: v_{ij} is **not** iid.

The Error Components Model

Substitute $\alpha_j = \bar{\alpha} + u_j$ into the FE model:

$$y_{ij} = \bar{\alpha} + \beta\, x_{ij} + \underbrace{u_j + e_{ij}}_{v_{ij}}$$

where:

- u_j = hospital-specific random component (same for all patients at hospital j)
- e_{ij} = idiosyncratic error (patient-specific noise)
- $v_{ij} = u_j + e_{ij}$ = **composite error**

Problem: v_{ij} is **not** iid.

Two patients at the same hospital share u_j, so their errors are **correlated**. OLS ignores this.

The Correlation Structure

Within hospital j (patients $i \neq k$):

$$\text{Corr}(v_{ij},\, v_{kj}) = \frac{\sigma_u^2}{\sigma_u^2 + \sigma_e^2} > 0$$

The Correlation Structure

Within hospital j (patients $i \neq k$):

$$\text{Corr}(v_{ij},\, v_{kj}) = \frac{\sigma_u^2}{\sigma_u^2 + \sigma_e^2} > 0$$

Across hospitals ($j \neq l$):

$$\text{Corr}(v_{ij},\, v_{kl}) = 0$$

The Correlation Structure

Within hospital j (patients $i \neq k$):

$$\text{Corr}(v_{ij},\, v_{kj}) = \frac{\sigma_u^2}{\sigma_u^2 + \sigma_e^2} > 0$$

Across hospitals ($j \neq l$):

$$\text{Corr}(v_{ij},\, v_{kl}) = 0$$

What this means:

- Patients at the same hospital have positively correlated errors
- The correlation equals the share of total variance due to hospital effects
- OLS standard errors are **too small** because they count within-hospital observations as independent

The Correlation Structure

Within hospital j (patients $i \neq k$):

$$\text{Corr}(v_{ij},\, v_{kj}) = \frac{\sigma_u^2}{\sigma_u^2 + \sigma_e^2} > 0$$

Across hospitals ($j \neq l$):

$$\text{Corr}(v_{ij},\, v_{kl}) = 0$$

What this means:

- Patients at the same hospital have positively correlated errors
- The correlation equals the share of total variance due to hospital effects
- OLS standard errors are **too small** because they count within-hospital observations as independent

$\Longrightarrow$ Even if the OLS slope is OK, the inference is wrong.

Why This Correlation Is Useful

Define the **intraclass correlation**:

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

Why This Correlation Is Useful

Define the **intraclass correlation**:

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

ρ = the share of total variance explained by **which hospital** a patient is in.

Why This Correlation Is Useful

Define the **intraclass correlation**:

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

ρ = the share of total variance explained by **which hospital** a patient is in.

- $\rho \approx 0$: hospitals are basically the same $\rightarrow$ grouping doesn't help much
- $\rho \approx 1$: almost all variation is between hospitals $\rightarrow$ knowing the hospital tells you almost everything

Why This Correlation Is Useful

Define the **intraclass correlation**:

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

ρ = the share of total variance explained by **which hospital** a patient is in.

- $\rho \approx 0$: hospitals are basically the same $\rightarrow$ grouping doesn't help much
- $\rho \approx 1$: almost all variation is between hospitals $\rightarrow$ knowing the hospital tells you almost everything

This is where RE gets its power. Instead of estimating each hospital in isolation, RE **borrows strength from the ensemble** (Tukey, 1970; Efron & Morris, 1973): it pulls each hospital's estimate toward the overall mean, especially when a hospital has few patients.

Why This Correlation Is Useful

Define the **intraclass correlation**:

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

ρ = the share of total variance explained by **which hospital** a patient is in.

- $\rho \approx 0$: hospitals are basically the same $\rightarrow$ grouping doesn't help much
- $\rho \approx 1$: almost all variation is between hospitals $\rightarrow$ knowing the hospital tells you almost everything

This is where RE gets its power. Instead of estimating each hospital in isolation, RE **borrows strength from the ensemble** (Tukey, 1970; Efron & Morris, 1973): it pulls each hospital's estimate toward the overall mean, especially when a hospital has few patients.

$\implies$ A hospital with only 5 patients gets a better estimate by "learning" from the other 495. The more hospitals look alike (small ρ), the more we can borrow.

GLS Intuition: Transforming the Data

OLS ignores the correlation in v_{ij}. GLS accounts for it by **transforming** the data.

GLS Intuition: Transforming the Data

OLS ignores the correlation in v_{ij}. GLS accounts for it by **transforming** the data.

Introduce the parameter $\hat{\alpha}_j$:

$$\hat{\alpha}_j = 1 - \frac{\sigma_e}{\sqrt{N_j \sigma_u^2 + \sigma_e^2}}$$

GLS Intuition: Transforming the Data

OLS ignores the correlation in v_{ij}. GLS accounts for it by **transforming** the data.

Introduce the parameter $\hat{\alpha}_j$:

$$\hat{\alpha}_j = 1 - \frac{\sigma_e}{\sqrt{N_j \sigma_u^2 + \sigma_e^2}}$$

where N_j = observations in group j (patients per hospital in our example).

GLS Intuition: Transforming the Data

OLS ignores the correlation in v_{ij}. GLS accounts for it by **transforming** the data.

Introduce the parameter $\hat{\alpha}_j$:

$$\hat{\alpha}_j = 1 - \frac{\sigma_e}{\sqrt{N_j \sigma_u^2 + \sigma_e^2}}$$

where N_j = observations in group j (patients per hospital in our example).

The RE transformation is a partial demeaning:

$$y_{ij} - \hat{\alpha}_j \bar{y}_j \qquad \text{and} \qquad x_{ij} - \hat{\alpha}_j \bar{x}_j$$

GLS Intuition: Transforming the Data

OLS ignores the correlation in v_{ij}. GLS accounts for it by **transforming** the data.

Introduce the parameter $\hat{\alpha}_j$:

$$\hat{\alpha}_j = 1 - \frac{\sigma_e}{\sqrt{N_j \sigma_u^2 + \sigma_e^2}}$$

where N_j = observations in group j (patients per hospital in our example).

The RE transformation is a partial demeaning:

$$y_{ij} - \hat{\alpha}_j \bar{y}_j \qquad \text{and} \qquad x_{ij} - \hat{\alpha}_j \bar{x}_j$$

Then run OLS on the transformed data. This is **feasible GLS** (FGLS).

What $\hat{\alpha}$ Does: A Spectrum

The RE estimator lives on a spectrum between pooled OLS and FE:

What $\hat{\alpha}$ Does: A Spectrum

The RE estimator lives on a spectrum between pooled OLS and FE:

$\hat{\alpha}$	**Transformation**	**Equivalent to**
0	$y_{ij} - 0 \cdot \bar{y}_j = y_{ij}$	Pooled OLS (no group effect)
$0 < \hat{\alpha} < 1$	$y_{ij} - \hat{\alpha}\,\bar{y}_j$	**RE: weighted average**
1	$y_{ij} - \bar{y}_j$	FE (full demeaning)

What $\hat{\alpha}$ Does: A Spectrum

The RE estimator lives on a spectrum between pooled OLS and FE:

$\hat{\alpha}$	**Transformation**	**Equivalent to**
0	$y_{ij} - 0 \cdot \bar{y}_j = y_{ij}$	Pooled OLS (no group effect)
$0 < \hat{\alpha} < 1$	$y_{ij} - \hat{\alpha}\,\bar{y}_j$	**RE: weighted average**
1	$y_{ij} - \bar{y}_j$	FE (full demeaning)

When does $\hat{\alpha} \to 1$?

- σ_u^2 is large relative to σ_e^2 (strong group effects)
- N_j is large (many obs per group)

What $\hat{\alpha}$ Does: A Spectrum

The RE estimator lives on a spectrum between pooled OLS and FE:

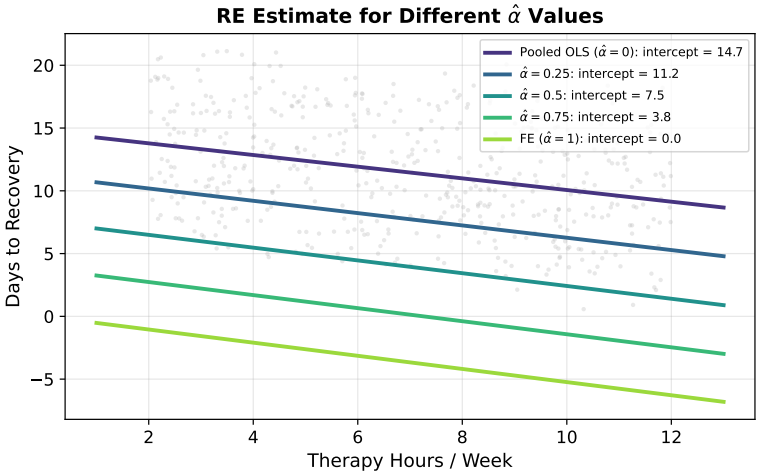
$\hat{\alpha}$	**Transformation**	**Equivalent to**
0	$y_{ij} - 0 \cdot \bar{y}_j = y_{ij}$	Pooled OLS (no group effect)
$0 < \hat{\alpha} < 1$	$y_{ij} - \hat{\alpha}\,\bar{y}_j$	**RE: weighted average**
1	$y_{ij} - \bar{y}_j$	FE (full demeaning)

When does $\hat{\alpha} \to 1$?

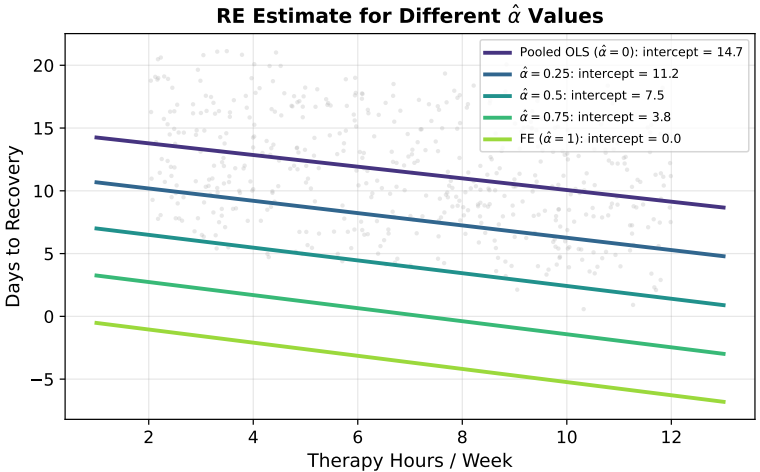
- σ_u^2 is large relative to σ_e^2 (strong group effects)
- N_j is large (many obs per group)

$\Longrightarrow$ With strong group effects or large panels, RE $\approx$ FE.

The $\hat{\alpha}$ Spectrum: Visualized



The $\hat{\alpha}$ Spectrum: Visualized



As $\hat{\alpha}$ increases from 0 to 1, we demean more aggressively, trusting each hospital's own data rather than pulling it toward the overall mean.

The RE Assumption Revisited

RE requires:

$$\text{Cov}(u_j, x_{ij}) = 0$$

The RE Assumption Revisited

RE requires:

$$\text{Cov}(u_j, x_{ij}) = 0$$

In our hospital example: hospital quality must be **uncorrelated** with therapy hours.

The RE Assumption Revisited

RE requires:

$$\text{Cov}(u_j, x_{ij}) = 0$$

In our hospital example: hospital quality must be **uncorrelated** with therapy hours.

Is that realistic?

- Better hospitals might prescribe *more* therapy (better protocols)
- Or *less* therapy (patients recover faster anyway)
- Patient selection: sicker patients go to better hospitals

The RE Assumption Revisited

RE requires:

$$\text{Cov}(u_j, x_{ij}) = 0$$

In our hospital example: hospital quality must be **uncorrelated** with therapy hours.

Is that realistic?

- Better hospitals might prescribe *more* therapy (better protocols)
- Or *less* therapy (patients recover faster anyway)
- Patient selection: sicker patients go to better hospitals

If $\text{Cov}(u_j, x_{ij}) \neq 0$:

- FE is still consistent (it eliminates u_j entirely)
- RE is **inconsistent** (the partial demeaning doesn't fully remove u_j)

When RE Is Appropriate

Use RE when:

When RE Is Appropriate

Use RE when:

1. Groups are **random draws** from a larger population
   → Sampled hospitals from all hospitals in the country

When RE Is Appropriate

Use RE when:

1. Groups are **random draws** from a larger population
   → Sampled hospitals from all hospitals in the country
2. No reason to think group effects correlate with regressors
   → Random assignment, natural experiment

When RE Is Appropriate

Use RE when:

1. Groups are **random draws** from a larger population
   → Sampled hospitals from all hospitals in the country
2. No reason to think group effects correlate with regressors
   → Random assignment, natural experiment
3. You want to estimate effects of **time-invariant variables**
   → Teaching hospital status, rural vs. urban

When RE Is Appropriate

Use RE when:

1. Groups are **random draws** from a larger population
   → Sampled hospitals from all hospitals in the country
2. No reason to think group effects correlate with regressors
   → Random assignment, natural experiment
3. You want to estimate effects of **time-invariant variables**
   → Teaching hospital status, rural vs. urban
4. **Efficiency**: RE uses both within and between variation
   → Smaller standard errors than FE

When FE Is Safer

Use FE when:

When FE Is Safer

Use FE when:

1. Groups are **specific entities** you care about
   → These particular 5 hospitals, not a random sample

When FE Is Safer

Use FE when:

1. Groups are **specific entities** you care about
   → These particular 5 hospitals, not a random sample
2. Plausible that $\text{Cov}(u_j, x_{ij}) \neq 0$
   → Better hospitals may assign more/less therapy

When FE Is Safer

Use FE when:

1. Groups are **specific entities** you care about
   → These particular 5 hospitals, not a random sample
2. Plausible that $\text{Cov}(u_j, x_{ij}) \neq 0$
   → Better hospitals may assign more/less therapy
3. Micro data (individuals, firms) $\implies$ FE is **almost always safer**
   → Unobserved ability, management quality, etc.

When FE Is Safer

Use FE when:

1. Groups are **specific entities** you care about
   → These particular 5 hospitals, not a random sample
2. Plausible that $\text{Cov}(u_j, x_{ij}) \neq 0$
   → Better hospitals may assign more/less therapy
3. Micro data (individuals, firms) $\implies$ FE is **almost always safer**
   → Unobserved ability, management quality, etc.
4. You only care about **within-group** effects
   → "Among patients at the same hospital, does more therapy help?"

When FE Is Safer

Use FE when:

1. Groups are **specific entities** you care about
   → These particular 5 hospitals, not a random sample
2. Plausible that $\text{Cov}(u_j, x_{ij}) \neq 0$
   → Better hospitals may assign more/less therapy
3. Micro data (individuals, firms) $\implies$ FE is **almost always safer**
   → Unobserved ability, management quality, etc.
4. You only care about **within-group** effects
   → "Among patients at the same hospital, does more therapy help?"

$\implies$ When in doubt, FE is the conservative choice. But how do we decide formally?

Motivation: What Difference Does It Make?

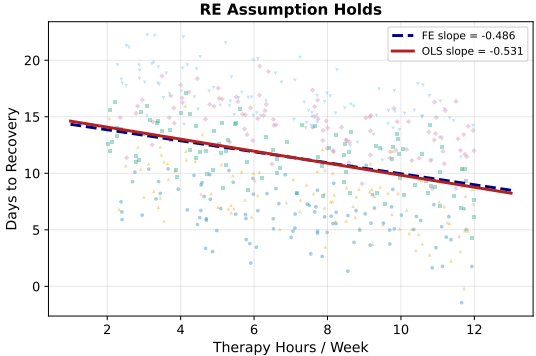
If RE assumption holds: Both FE and RE are consistent, but RE is more efficient.

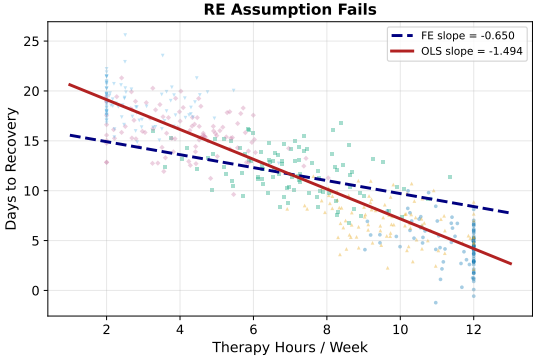
If RE assumption fails: Only FE is consistent. RE gives the wrong answer.

Motivation: What Difference Does It Make?

If RE assumption holds: Both FE and RE are consistent, but RE is more efficient.

If RE assumption fails: Only FE is consistent. RE gives the wrong answer.

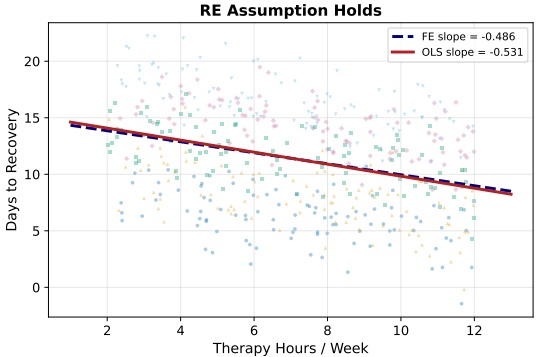


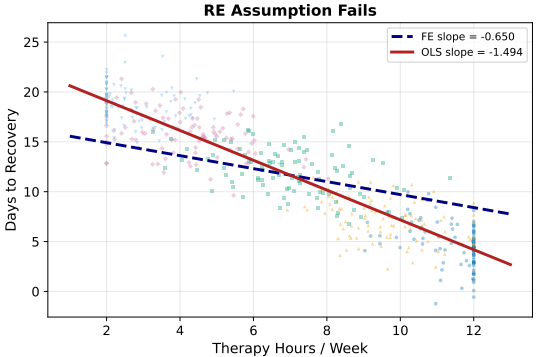


Motivation: What Difference Does It Make?

If RE assumption holds: Both FE and RE are consistent, but RE is more efficient.

If RE assumption fails: Only FE is consistent. RE gives the wrong answer.





When the assumption holds, slopes are similar. When it fails, they **diverge**.

The Hausman Test: Setup

Idea: If RE is valid, then $\hat{\beta}_{FE}$ and $\hat{\beta}_{RE}$ should be close. If they diverge, something is wrong with RE.

The Hausman Test: Setup

Idea: If RE is valid, then $\hat{\beta}_{FE}$ and $\hat{\beta}_{RE}$ should be close. If they diverge, something is wrong with RE.

Hypotheses:

- H_0: $\text{Cov}(u_j, x_{ij}) = 0 \iff$ RE is consistent (and more efficient)
- H_1: $\text{Cov}(u_j, x_{ij}) \neq 0 \iff$ only FE is consistent

The Hausman Test: Setup

Idea: If RE is valid, then $\hat{\beta}_{FE}$ and $\hat{\beta}_{RE}$ should be close. If they diverge, something is wrong with RE.

Hypotheses:

- H_0: $\text{Cov}(u_j, x_{ij}) = 0 \iff$ RE is consistent (and more efficient)
- H_1: $\text{Cov}(u_j, x_{ij}) \neq 0 \iff$ only FE is consistent

Test statistic (single regressor):

$$t = \frac{\hat{\beta}_{FE} - \hat{\beta}_{RE}}{\sqrt{\widehat{\text{Var}}(\hat{\beta}_{FE}) - \widehat{\text{Var}}(\hat{\beta}_{RE})}}$$

The Hausman Test: Setup

Idea: If RE is valid, then $\hat{\beta}_{FE}$ and $\hat{\beta}_{RE}$ should be close. If they diverge, something is wrong with RE.

Hypotheses:

- H_0: $\text{Cov}(u_j, x_{ij}) = 0 \iff$ RE is consistent (and more efficient)
- H_1: $\text{Cov}(u_j, x_{ij}) \neq 0 \iff$ only FE is consistent

Test statistic (single regressor):

$$t = \frac{\hat{\beta}_{FE} - \hat{\beta}_{RE}}{\sqrt{\widehat{\text{Var}}(\hat{\beta}_{FE}) - \widehat{\text{Var}}(\hat{\beta}_{RE})}}$$

With multiple regressors, the test generalizes to a χ^2_K statistic (where K = number of regressors). Software handles this automatically.

The Hausman Test: Why Does It Work?

Under H_0: Both estimators converge to the true β, so the difference $\hat{\beta}_{FE} - \hat{\beta}_{RE}$ is small.

The Hausman Test: Why Does It Work?

Under H_0: Both estimators converge to the true β, so the difference $\hat{\beta}_{FE} - \hat{\beta}_{RE}$ is small.

Under H_1: RE is biased, so the difference is systematically large.

The Hausman Test: Why Does It Work?

Under H_0: Both estimators converge to the true β, so the difference $\hat{\beta}_{FE} - \hat{\beta}_{RE}$ is small.

Under H_1: RE is biased, so the difference is systematically large.

Why is the denominator well-defined?

Under H_0, FE is less efficient than RE:

$$\text{Var}(\hat{\beta}_{FE}) > \text{Var}(\hat{\beta}_{RE})$$

so $\widehat{\text{Var}}(\hat{\beta}_{FE}) - \widehat{\text{Var}}(\hat{\beta}_{RE}) > 0$ and the square root exists.

Interpreting the Hausman Test

Decision rule:

Result	**Conclusion**	**Action**
$p < 0.05$	Reject H_0	Use **FE**
$p \geq 0.05$	Fail to reject H_0	Can use **RE**

Interpreting the Hausman Test

Decision rule:

Result	Conclusion	Action
$p < 0.05$	Reject H_0	Use **FE**
$p \geq 0.05$	Fail to reject H_0	Can use **RE**

Intuition:

- Reject → FE and RE give **different** answers → RE is biased → use FE
- Fail to reject → FE and RE give **similar** answers → RE is OK and more efficient

Interpreting the Hausman Test

Decision rule:

Result	Conclusion	Action
$p < 0.05$	Reject H_0	Use **FE**
$p \geq 0.05$	Fail to reject H_0	Can use **RE**

Intuition:

- Reject $\rightarrow$ FE and RE give **different** answers $\rightarrow$ RE is biased $\rightarrow$ use FE
- Fail to reject $\rightarrow$ FE and RE give **similar** answers $\rightarrow$ RE is OK and more efficient

$\Longrightarrow$ The Hausman test is a **specification test**: it checks whether $\text{Cov}(u_j, x_{ij}) = 0$ is reasonable.

Worked Example: Hospital Data

Using our hospital recovery data:

Worked Example: Hospital Data

Using our hospital recovery data:

	FE	**RE**
Slope ($\hat{\beta}$)	-0.500	-0.497
$\widehat{\text{Var}}(\hat{\beta})$	0.00180	0.00165

Worked Example: Hospital Data

Using our hospital recovery data:

	FE	**RE**
Slope ($\hat{\beta}$)	-0.500	-0.497
$\widehat{\text{Var}}(\hat{\beta})$	0.00180	0.00165

Hausman statistic:

$$t = \frac{-0.500 - (-0.497)}{\sqrt{0.00180 - 0.00165}} = \frac{-0.003}{\sqrt{0.00015}} = \frac{-0.003}{0.0122} = -0.25$$

Worked Example: Hospital Data

Using our hospital recovery data:

	FE	**RE**
Slope ($\hat{\beta}$)	-0.500	-0.497
$\widehat{\text{Var}}(\hat{\beta})$	0.00180	0.00165

Hausman statistic:

$$t = \frac{-0.500 - (-0.497)}{\sqrt{0.00180 - 0.00165}} = \frac{-0.003}{\sqrt{0.00015}} = \frac{-0.003}{0.0122} = -0.25$$

$|t| = 0.25 < 1.96 \implies$ **Fail to reject** H_0.

Worked Example: Hospital Data

Using our hospital recovery data:

	FE	**RE**
Slope ($\hat{\beta}$)	-0.500	-0.497
$\widehat{\text{Var}}(\hat{\beta})$	0.00180	0.00165

Hausman statistic:

$$t = \frac{-0.500 - (-0.497)}{\sqrt{0.00180 - 0.00165}} = \frac{-0.003}{\sqrt{0.00015}} = \frac{-0.003}{0.0122} = -0.25$$

$|t| = 0.25 < 1.96 \implies$ **Fail to reject** H_0.

$\implies$ RE and FE agree. RE is appropriate here, and more efficient.

FE vs. RE: Decision Flowchart

Step 1: Is there unobserved group heterogeneity?

- No $\rightarrow$ Pooled OLS is fine | Yes $\rightarrow$ Go to Step 2

FE vs. RE: Decision Flowchart

Step 1: Is there unobserved group heterogeneity?

- No $\rightarrow$ Pooled OLS is fine | Yes $\rightarrow$ Go to Step 2

Step 2: Do you need to estimate effects of time-invariant variables?

- Yes $\rightarrow$ Must use RE (FE absorbs them) | No $\rightarrow$ Go to Step 3

FE vs. RE: Decision Flowchart

Step 1: Is there unobserved group heterogeneity?

- No $\rightarrow$ Pooled OLS is fine | Yes $\rightarrow$ Go to Step 2

Step 2: Do you need to estimate effects of time-invariant variables?

- Yes $\rightarrow$ Must use RE (FE absorbs them) | No $\rightarrow$ Go to Step 3

Step 3: Is $\text{Cov}(u_j, x_{ij}) = 0$ plausible?

- Clearly no $\rightarrow$ Use FE | Maybe $\rightarrow$ Run the Hausman test

FE vs. RE: Decision Flowchart

Step 1: Is there unobserved group heterogeneity?

- No $\rightarrow$ Pooled OLS is fine | Yes $\rightarrow$ Go to Step 2

Step 2: Do you need to estimate effects of time-invariant variables?

- Yes $\rightarrow$ Must use RE (FE absorbs them) | No $\rightarrow$ Go to Step 3

Step 3: Is $\text{Cov}(u_j, x_{ij}) = 0$ plausible?

- Clearly no $\rightarrow$ Use FE | Maybe $\rightarrow$ Run the Hausman test

Step 4: Hausman test result?

- Reject H_0 $\rightarrow$ Use FE | Fail to reject $\rightarrow$ Use RE (more efficient)

FE vs. RE: Decision Flowchart

Step 1: Is there unobserved group heterogeneity?

- No $\rightarrow$ Pooled OLS is fine | Yes $\rightarrow$ Go to Step 2

Step 2: Do you need to estimate effects of time-invariant variables?

- Yes $\rightarrow$ Must use RE (FE absorbs them) | No $\rightarrow$ Go to Step 3

Step 3: Is $\text{Cov}(u_j, x_{ij}) = 0$ plausible?

- Clearly no $\rightarrow$ Use FE | Maybe $\rightarrow$ Run the Hausman test

Step 4: Hausman test result?

- Reject H_0 $\rightarrow$ Use FE | Fail to reject $\rightarrow$ Use RE (more efficient)

$\Longrightarrow$ When in doubt, FE is the sa**fe** default. RE is the **re**ward for being able to argue $\text{Cov}(u_j, x_{ij}) = 0$.

Thank you!
jakeanderson@g.ucla.edu

Pooled OLS and Cluster-Robust Standard Errors

When 240 Observations Are Really Just 8

Jake Anderson

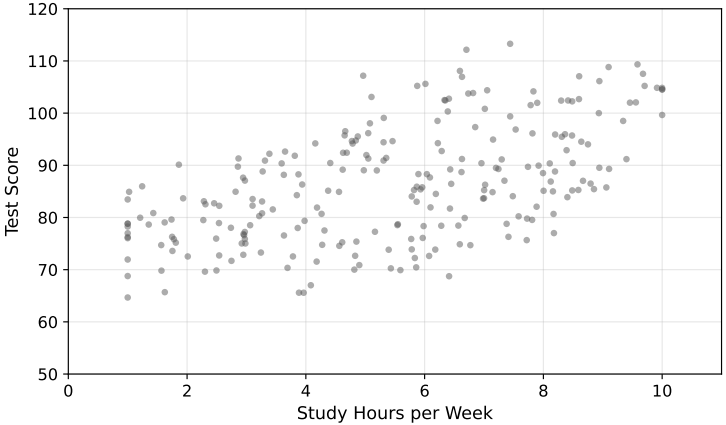
May 16, 2026

Outline

1. The Problem: Clustered Data
2. The Cluster-Robust Fix
3. When to Cluster
4. Summary

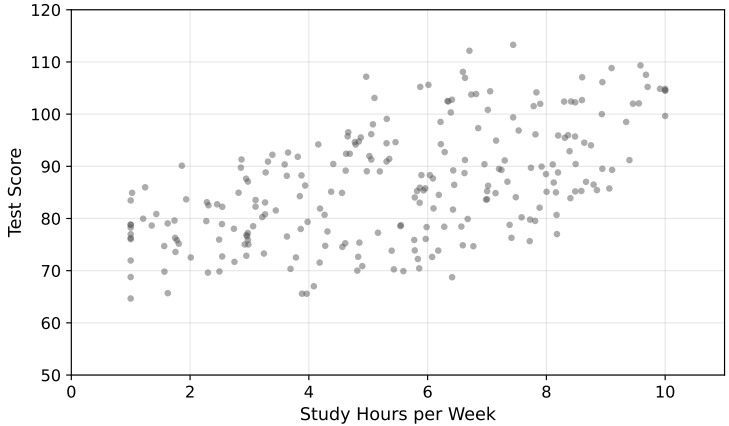
The Data

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



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:

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

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**.

Setup: The Pooled OLS Model

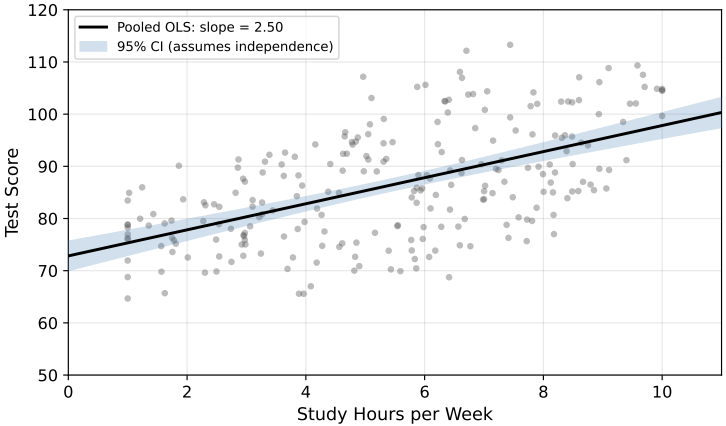
We ignore any group structure and run a single regression:

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

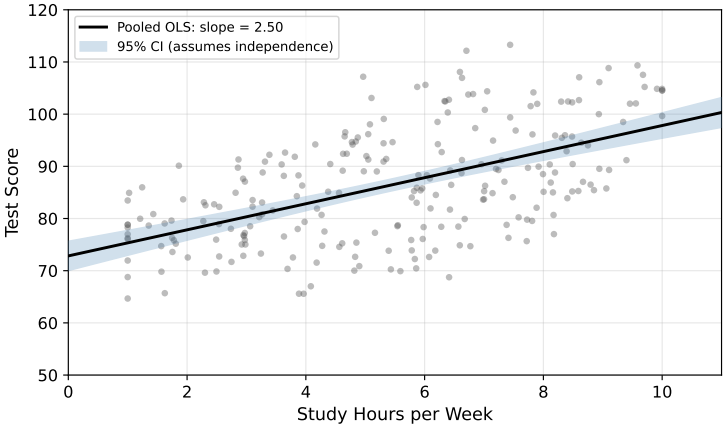
This treats all 240 students as **independent observations**.

- 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

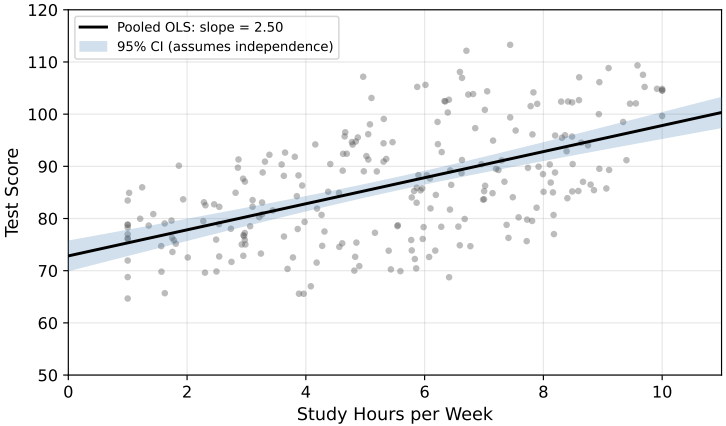


Pooled OLS Result



Result: $\hat{\beta}_1 = 2.50$, SE $= 0.239$.

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?

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.

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

But Something Is Wrong

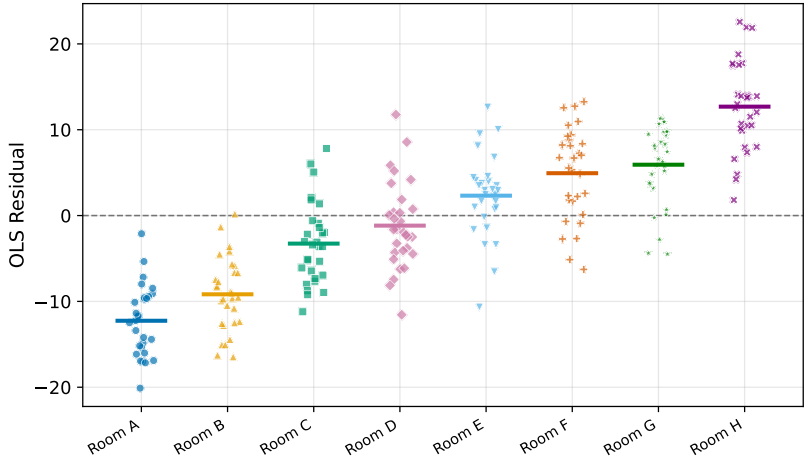
The slope estimate is right on target. So what's the problem?

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

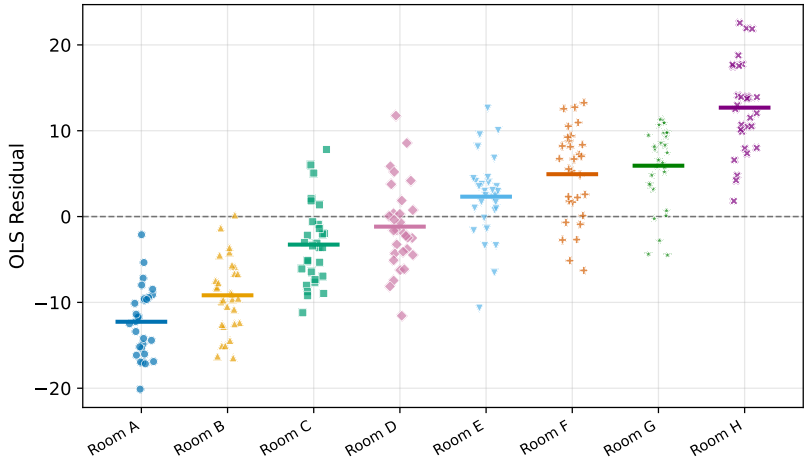
- OLS standard errors require that errors be **independent** across observations
- If students share unobserved factors (teacher quality, classroom culture, grading norms), their errors are **correlated**, not independent

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

Residuals by Classroom



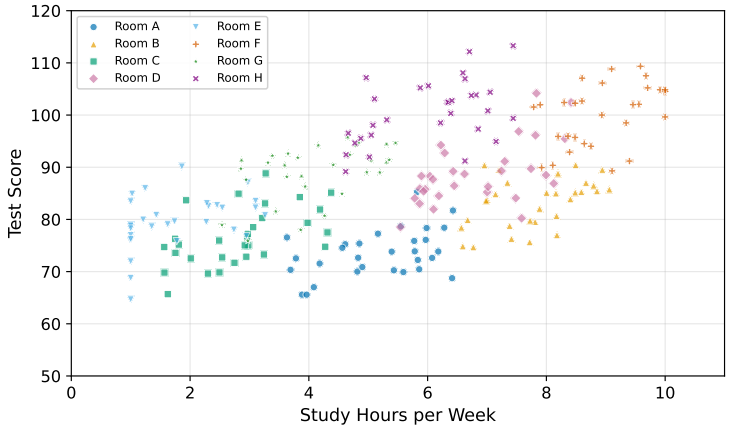
Residuals by Classroom



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

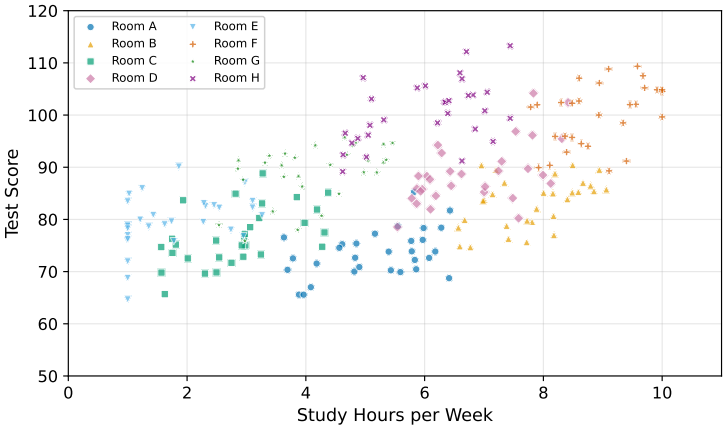
Reveal: Eight Classrooms

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



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:

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

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 $v_{ij} = u_j + e_{ij}$.

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

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.

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:

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

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.

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

Consequence: Standard Errors Are Too Small

Recall the OLS variance formula for the slope:

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

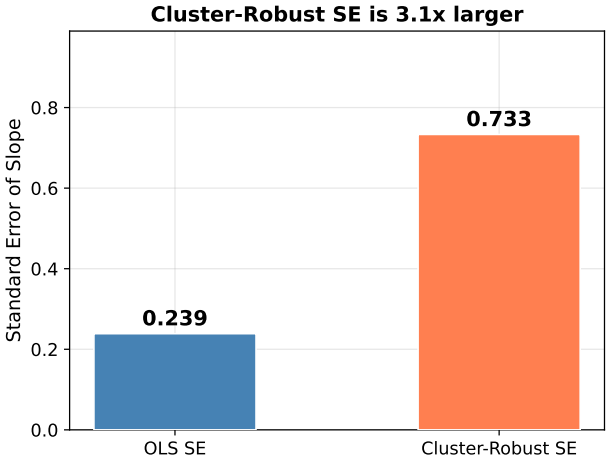
This formula assumes all errors are independent. It drops all $\text{Cov}(v_i, v_k)$ cross-terms. When within-cluster errors are positively correlated, those cross-terms are positive, making the true variance larger.

Intuitively:

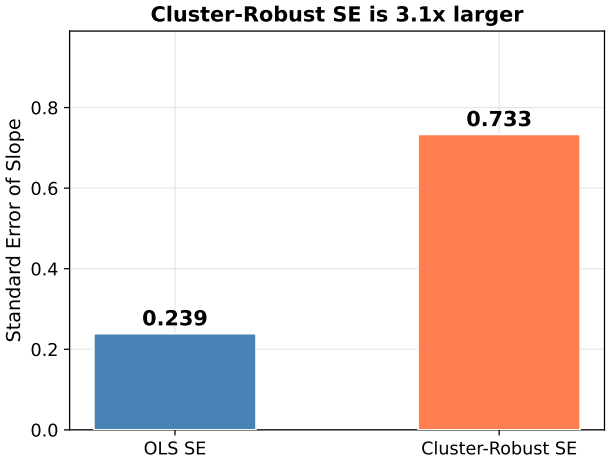
- The denominator counts all $n = 240$ observations
- But many of those observations carry **overlapping information**
- The formula "over-counts" the effective information in the data

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

How Wrong: SE Comparison

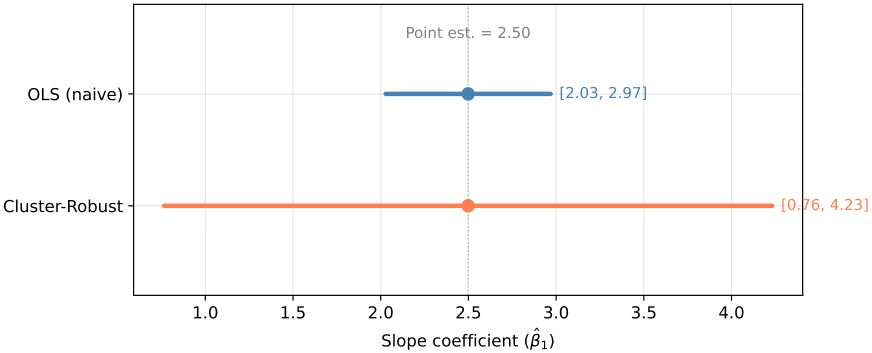


How Wrong: SE Comparison

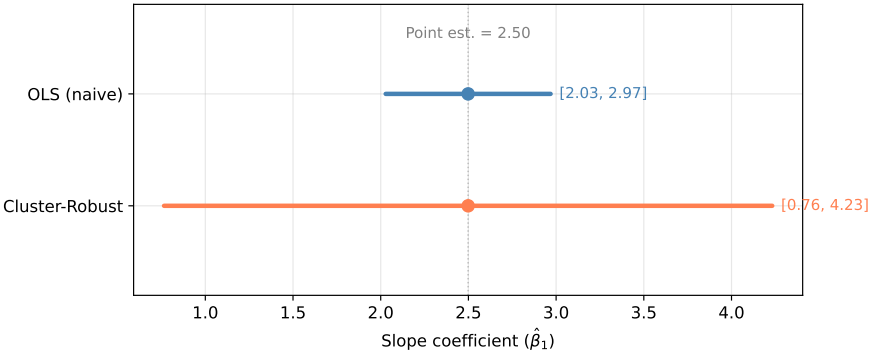


In this dataset, the cluster-robust SE is **3.1x larger** than the naive OLS SE.

Different SEs, Different Conclusions

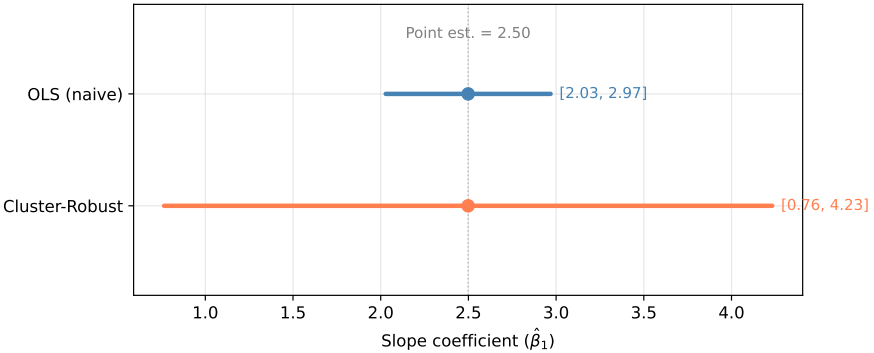


Different SEs, Different Conclusions



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 $[2.03, 2.97]$.

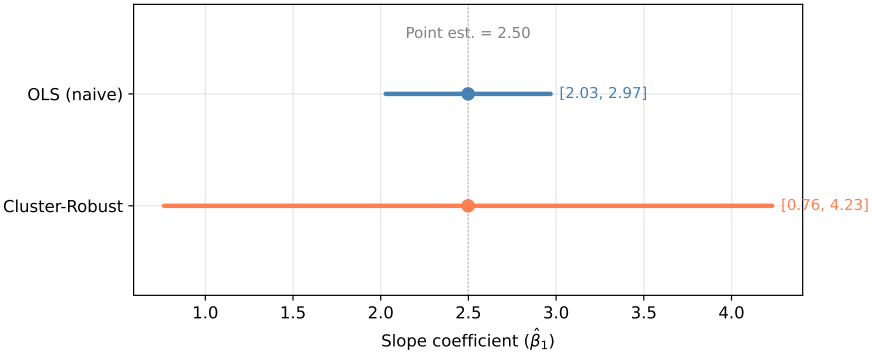
Different SEs, Different Conclusions



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.

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?

How Correlated: The Intraclass Correlation

How correlated are errors within the same classroom?

Define σ_u^2 = variance of the classroom effect u_j, and σ_e^2 = variance of the idiosyncratic error e_{ij}.

How Correlated: The Intraclass Correlation

How correlated are errors within the same classroom?

Define σ_u^2 = variance of the classroom effect u_j, and σ_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:

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

How Correlated: The Intraclass Correlation

How correlated are errors within the same classroom?

Define σ_u^2 = variance of the classroom effect u_j, and σ_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.

How Correlated: The Intraclass Correlation

How correlated are errors within the same classroom?

Define σ_u^2 = variance of the classroom effect u_j, and σ_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:

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

OLS Assumption Violated

Standard OLS assumes:

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

With clustered data:

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

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

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

1 The Problem: Clustered Data

2 The Cluster-Robust Fix

3 When to Cluster

4 Summary

The Cluster-Robust Variance Estimator

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

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

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)
```

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?"

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.

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.

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:

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

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

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)

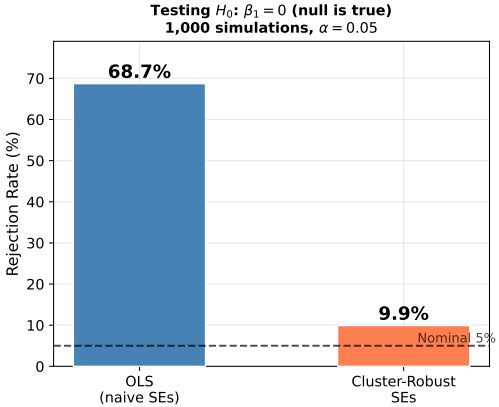
When Does Clustering Increase SEs?

The SE inflation depends on three factors:

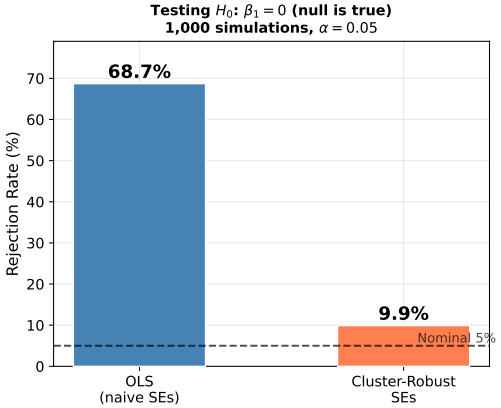
1. **Positive within-cluster error correlation** ($\rho > 0$)
   - If $\rho = 0$, cluster-robust SEs $\approx$ OLS SEs
   - The larger ρ, the more inflation
2. **The regressor varies across clusters** (not just within)
   - If all classrooms have the same mean study hours, little inflation
   - If some 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?

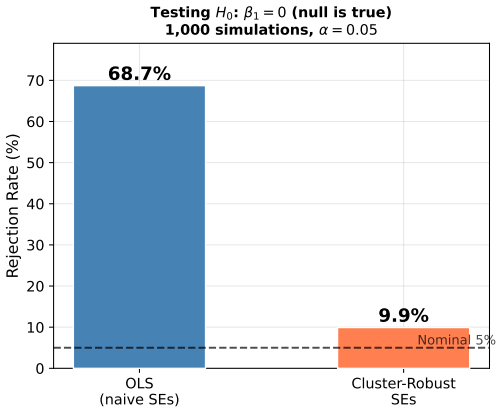


Simulation: How Often Does OLS Wrongly Reject?



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

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**.

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.

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

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

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 $n_j > 1$.

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 $n_j > 1$.

$\implies$ 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

1 The Problem: Clustered Data

2 The Cluster-Robust Fix

3 When to Cluster

4 Summary

When to Cluster

Cluster your standard errors whenever observations share unobserved common shocks:

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

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

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)

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.

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

What to Cluster On

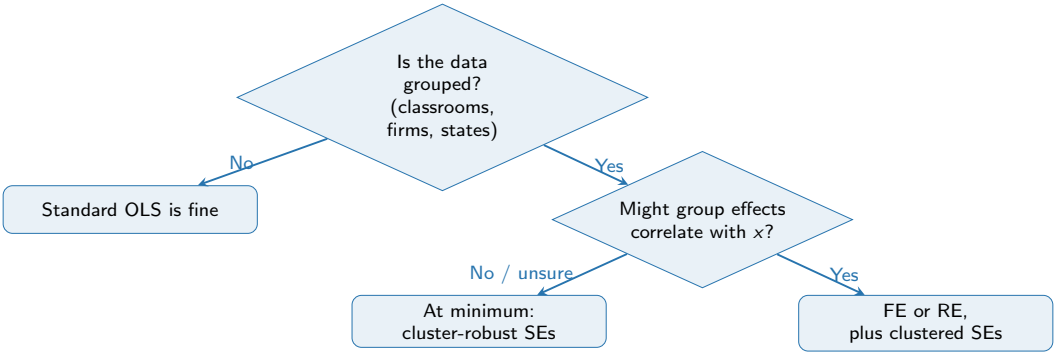
Principle: cluster at the level where treatment varies or common shocks arise.

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

1 The Problem: Clustered Data

2 The Cluster-Robust Fix

3 When to Cluster

4 Summary

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.

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.

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.

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.

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.

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

Dynamic Panel Data Models

Jake Anderson

May 16, 2026

Outline

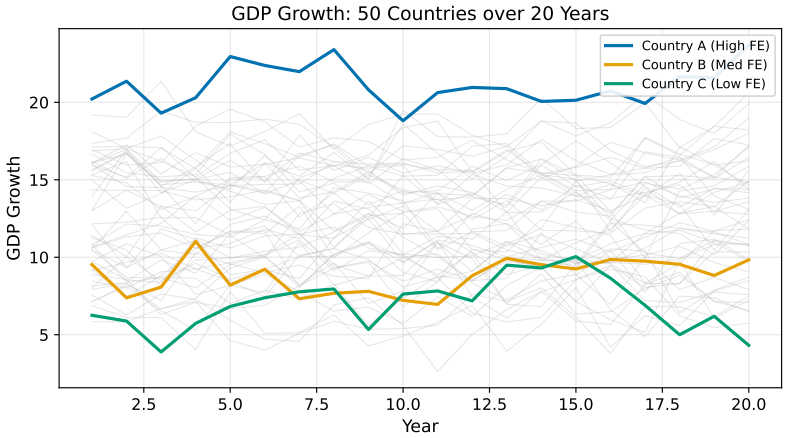
1. The Problem: Persistence in Panel Data
2. The FE Attempt and Nickell Bias
3. Arellano-Bond GMM
4. Decision Framework
5. Summary

GDP Growth Across 50 Countries

Does this year's GDP growth depend on last year's? If so, how do we estimate that persistence?

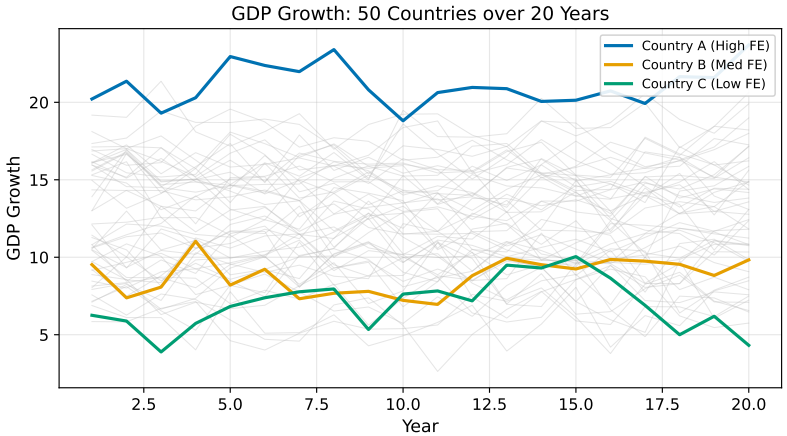
GDP Growth Across 50 Countries

Does this year's GDP growth depend on last year's? If so, how do we estimate that persistence?



GDP Growth Across 50 Countries

Does this year's GDP growth depend on last year's? If so, how do we estimate that persistence?



Some countries consistently grow faster than others. And growth seems **sticky**: a good year tends to follow a good year.

Persistence: The Pattern

Two features stand out in this panel:

Persistence: The Pattern

Two features stand out in this panel:

1. **Cross-country heterogeneity.** Country-specific factors (institutions, geography, culture) generate permanently different growth levels $\implies$ fixed effects α_i.

Persistence: The Pattern

Two features stand out in this panel:

1. **Cross-country heterogeneity.** Country-specific factors (institutions, geography, culture) generate permanently different growth levels $\implies$ fixed effects α_i.
2. **Within-country persistence.** Even after accounting for country differences, growth this year is correlated with growth last year $\implies$ the lagged dependent variable $y_{i,t-1}$ belongs in the model.

Persistence: The Pattern

Two features stand out in this panel:

1. **Cross-country heterogeneity.** Country-specific factors (institutions, geography, culture) generate permanently different growth levels $\implies$ fixed effects α_i.
2. **Within-country persistence.** Even after accounting for country differences, growth this year is correlated with growth last year $\implies$ the lagged dependent variable $y_{i,t-1}$ belongs in the model.

$\implies$ We need a model that includes **both** fixed effects and a lagged dependent variable.

The Dynamic Panel Model

$$y_{it} = \rho\, y_{i,t-1} + \beta\, \text{invest}_{it} + \alpha_i + \varepsilon_{it}$$

The Dynamic Panel Model

$$y_{it} = \rho\, y_{i,t-1} + \beta\, \text{invest}_{it} + \alpha_i + \varepsilon_{it}$$

- y_{it}: GDP growth for country i in year t
- $y_{i,t-1}$: last year's growth (the **lagged dependent variable**)
- invest_{it}: investment rate (exogenous regressor)
- α_i: country fixed effect (unobserved, time-invariant)
- ε_{it}: idiosyncratic error ($\text{E}[\varepsilon_{it}] = 0$, serially uncorrelated)

The Dynamic Panel Model

$$y_{it} = \rho\, y_{i,t-1} + \beta\, \text{invest}_{it} + \alpha_i + \varepsilon_{it}$$

- y_{it}: GDP growth for country i in year t
- $y_{i,t-1}$: last year's growth (the **lagged dependent variable**)
- invest_{it}: investment rate (exogenous regressor)
- α_i: country fixed effect (unobserved, time-invariant)
- ε_{it}: idiosyncratic error ($\text{E}[\varepsilon_{it}] = 0$, serially uncorrelated)

The parameter ρ measures **persistence**: how much of last year's growth carries forward. If $|\rho| < 1$, the process is stationary.

The Dynamic Panel Model

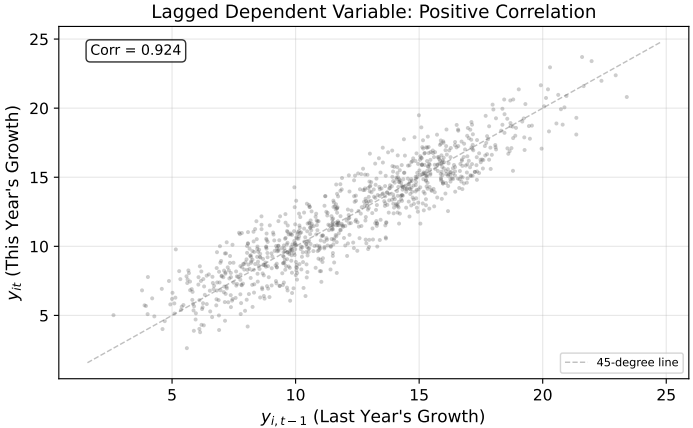
$$y_{it} = \rho\, y_{i,t-1} + \beta\, \text{invest}_{it} + \alpha_i + \varepsilon_{it}$$

- y_{it}: GDP growth for country i in year t
- $y_{i,t-1}$: last year's growth (the **lagged dependent variable**)
- invest_{it}: investment rate (exogenous regressor)
- α_i: country fixed effect (unobserved, time-invariant)
- ε_{it}: idiosyncratic error ($\text{E}[\varepsilon_{it}] = 0$, serially uncorrelated)

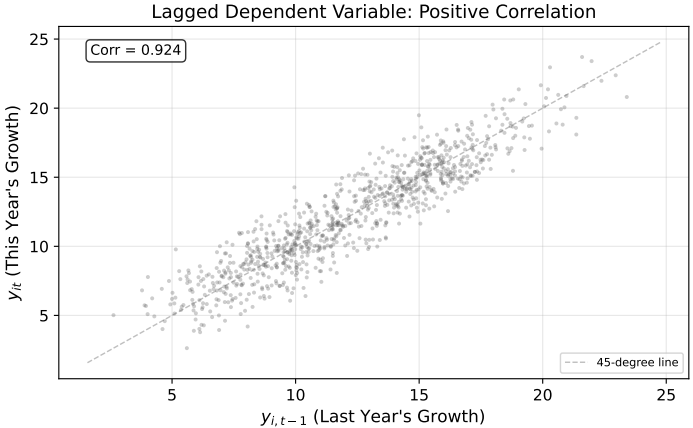
The parameter ρ measures **persistence**: how much of last year's growth carries forward. If $|\rho| < 1$, the process is stationary.

Goal: estimate ρ consistently. True value in our simulation: $\rho = 0.40$.

The Lagged Relationship

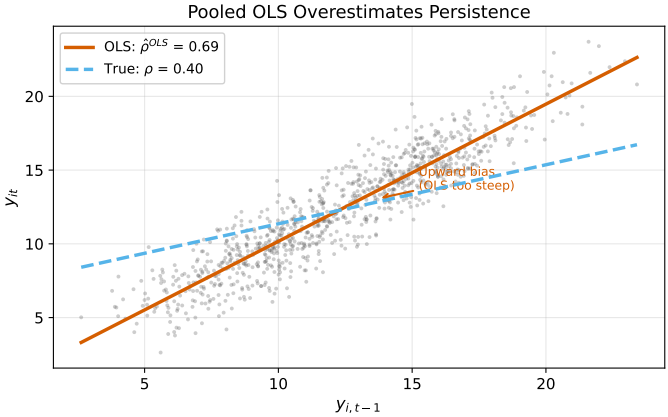


The Lagged Relationship

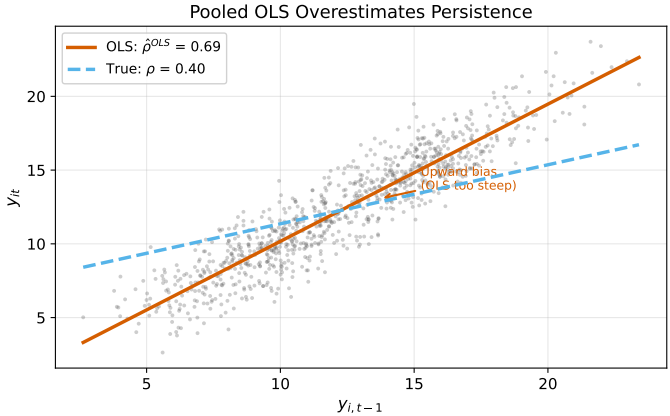


Strong positive correlation between y_{it} and $y_{i,t-1}$. But how much is true persistence, and how much is driven by the unobserved α_i?

Naive OLS: Just Regress y_{it} on $y_{i,t-1}$



Naive OLS: Just Regress y_{it} on $y_{i,t-1}$



OLS estimates $\hat{\rho}^{OLS} \approx 0.69$. The true ρ is 0.40. Why the overestimate?

Why OLS Fails: Omitted Variable Bias

Think of two countries:

Why OLS Fails: Omitted Variable Bias

Think of two countries:

- **Country A** has strong institutions (α_i high). Growth is *always* high. Both $y_{i,t-1}$ and y_{it} are large.

Why OLS Fails: Omitted Variable Bias

Think of two countries:

- **Country A** has strong institutions (α_i high). Growth is *always* high. Both $y_{i,t-1}$ and y_{it} are large.
- **Country C** has weak institutions (α_i low). Growth is *always* low. Both $y_{i,t-1}$ and y_{it} are small.

Why OLS Fails: Omitted Variable Bias

Think of two countries:

- **Country A** has strong institutions (α_i high). Growth is *always* high. Both $y_{i,t-1}$ and y_{it} are large.
- **Country C** has weak institutions (α_i low). Growth is *always* low. Both $y_{i,t-1}$ and y_{it} are small.

OLS sees: high $y_{i,t-1}$ followed by high y_{it}, low followed by low. It concludes the relationship is very strong. But the correlation is driven by **permanent country differences**, not persistence.

Why OLS Fails: Omitted Variable Bias

Think of two countries:

- **Country A** has strong institutions (α_i high). Growth is *always* high. Both $y_{i,t-1}$ and y_{it} are large.
- **Country C** has weak institutions (α_i low). Growth is *always* low. Both $y_{i,t-1}$ and y_{it} are small.

OLS sees: high $y_{i,t-1}$ followed by high y_{it}, low followed by low. It concludes the relationship is very strong. But the correlation is driven by **permanent country differences**, not persistence.

Formally, $y_{i,t-1}$ depends on α_i (since $y_{i,t-1} = \rho\, y_{i,t-2} + \beta\, \text{invest}_{i,t-1} + \alpha_i + \varepsilon_{i,t-1}$), so:

$$\text{bias} = \frac{\text{Cov}(y_{i,t-1}, \alpha_i)}{\text{Var}(y_{i,t-1})} > 0$$

Why OLS Fails: Omitted Variable Bias

Think of two countries:

- **Country A** has strong institutions (α_i high). Growth is *always* high. Both $y_{i,t-1}$ and y_{it} are large.
- **Country C** has weak institutions (α_i low). Growth is *always* low. Both $y_{i,t-1}$ and y_{it} are small.

OLS sees: high $y_{i,t-1}$ followed by high y_{it}, low followed by low. It concludes the relationship is very strong. But the correlation is driven by **permanent country differences**, not persistence.

Formally, $y_{i,t-1}$ depends on α_i (since $y_{i,t-1} = \rho\, y_{i,t-2} + \beta\, \text{invest}_{i,t-1} + \alpha_i + \varepsilon_{i,t-1}$), so:

$$\text{bias} = \frac{\text{Cov}(y_{i,t-1}, \alpha_i)}{\text{Var}(y_{i,t-1})} > 0$$

$\Longrightarrow$ OLS **overestimates** ρ because it confuses level differences with persistence.

Outline

1. The Problem: Persistence in Panel Data
2. The FE Attempt and Nickell Bias
3. Arellano-Bond GMM
4. Decision Framework
5. Summary

Fixed Effects: The Natural Fix?

FE removes α_i by demeaning. The within-transformed model:

$$\underbrace{y_{it} - \bar{y}_i}_{\ddot{y}_{it}} = \rho \underbrace{(y_{i,t-1} - \bar{y}_i)}_{\ddot{y}_{i,t-1}} + \beta \, \ddot{\text{invest}}_{it} + \underbrace{(\varepsilon_{it} - \bar{\varepsilon}_i)}_{\ddot{\varepsilon}_{it}}$$

Fixed Effects: The Natural Fix?

FE removes α_i by demeaning. The within-transformed model:

$$\underbrace{y_{it} - \bar{y}_i}_{\ddot{y}_{it}} = \rho \underbrace{(y_{i,t-1} - \bar{y}_i)}_{\ddot{y}_{i,t-1}} + \beta \, \ddot{\text{invest}}_{it} + \underbrace{(\varepsilon_{it} - \bar{\varepsilon}_i)}_{\ddot{\varepsilon}_{it}}$$

No more α_i. Problem solved?

Fixed Effects: The Natural Fix?

FE removes α_i by demeaning. The within-transformed model:

$$\underbrace{y_{it} - \bar{y}_i}_{\ddot{y}_{it}} = \rho \underbrace{(y_{i,t-1} - \bar{y}_i)}_{\ddot{y}_{i,t-1}} + \beta \ddot{\text{invest}}_{it} + \underbrace{(\varepsilon_{it} - \bar{\varepsilon}_i)}_{\ddot{\varepsilon}_{it}}$$

No more α_i. Problem solved?

Not quite. There is a subtle problem with demeaning when you have a lagged dependent variable.

Nickell Bias: Demeaning Creates New Correlation

The demeaned lag $\ddot{y}_{i,t-1} = y_{i,t-1} - \bar{y}_i$ is correlated with the demeaned error $\ddot{\varepsilon}_{it} = \varepsilon_{it} - \bar{\varepsilon}_i$.

Nickell Bias: Demeaning Creates New Correlation

The demeaned lag $\ddot{y}_{i,t-1} = y_{i,t-1} - \bar{y}_i$ is correlated with the demeaned error $\ddot{\varepsilon}_{it} = \varepsilon_{it} - \bar{\varepsilon}_i$.

Concrete example. Suppose $\varepsilon_{i,5}$ is a large positive shock.

- $\bar{y}_i$ includes $y_{i,5}$, which is inflated by $\varepsilon_{i,5} \implies \bar{y}_i$ goes **up**.
- Look at year 4: $\ddot{y}_{i,4} = y_{i,4} - \bar{y}_i$ goes **down** (the mean was inflated by the year-5 shock, but $y_{i,4}$ was not).
- Meanwhile $\ddot{\varepsilon}_{i,5} = \varepsilon_{i,5} - \bar{\varepsilon}_i$ goes **up** ($\varepsilon_{i,5}$ is positive).

Nickell Bias: Demeaning Creates New Correlation

The demeaned lag $\ddot{y}_{i,t-1} = y_{i,t-1} - \bar{y}_i$ is correlated with the demeaned error $\ddot{\varepsilon}_{it} = \varepsilon_{it} - \bar{\varepsilon}_i$.

Concrete example. Suppose $\varepsilon_{i,5}$ is a large positive shock.

- $\bar{y}_i$ includes $y_{i,5}$, which is inflated by $\varepsilon_{i,5} \implies \bar{y}_i$ goes **up**.
- Look at year 4: $\ddot{y}_{i,4} = y_{i,4} - \bar{y}_i$ goes **down** (the mean was inflated by the year-5 shock, but $y_{i,4}$ was not).
- Meanwhile $\ddot{\varepsilon}_{i,5} = \varepsilon_{i,5} - \bar{\varepsilon}_i$ goes **up** ($\varepsilon_{i,5}$ is positive).

$\implies$ Positive demeaned error at t=5 is associated with a negative demeaned lag at t=4. The correlation is **negative**, so FE is biased **downward**.

Nickell Bias: Demeaning Creates New Correlation

The demeaned lag $\ddot{y}_{i,t-1} = y_{i,t-1} - \bar{y}_i$ is correlated with the demeaned error $\ddot{\varepsilon}_{it} = \varepsilon_{it} - \bar{\varepsilon}_i$.

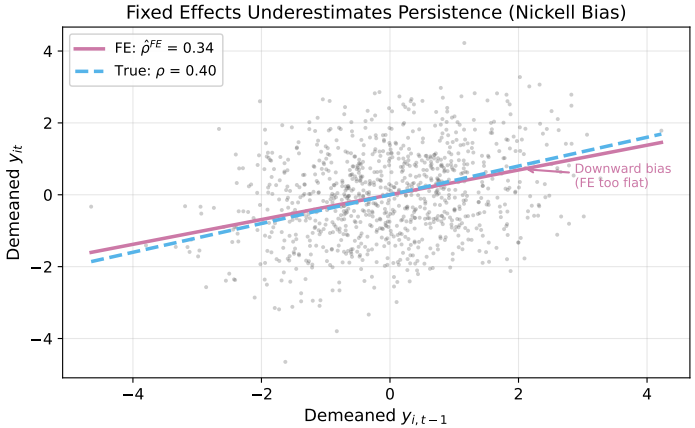
Concrete example. Suppose $\varepsilon_{i,5}$ is a large positive shock.

- $\bar{y}_i$ includes $y_{i,5}$, which is inflated by $\varepsilon_{i,5} \implies \bar{y}_i$ goes **up**.
- Look at year 4: $\ddot{y}_{i,4} = y_{i,4} - \bar{y}_i$ goes **down** (the mean was inflated by the year-5 shock, but $y_{i,4}$ was not).
- Meanwhile $\ddot{\varepsilon}_{i,5} = \varepsilon_{i,5} - \bar{\varepsilon}_i$ goes **up** ($\varepsilon_{i,5}$ is positive).

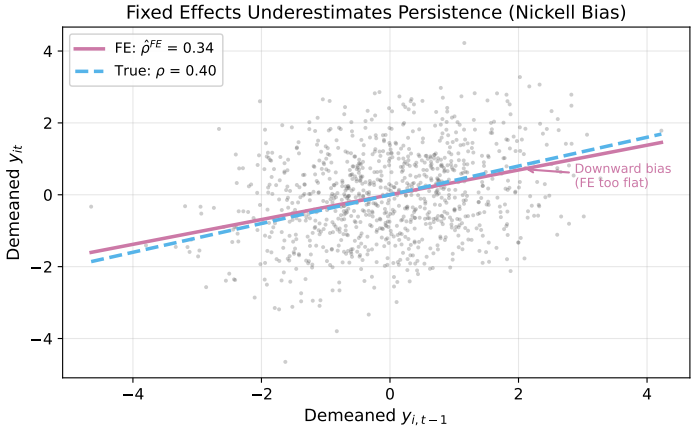
$\implies$ Positive demeaned error at t=5 is associated with a negative demeaned lag at t=4. The correlation is **negative**, so FE is biased **downward**.

This is the **Nickell bias** (Nickell, 1981). It arises purely from the mechanical relationship between demeaning and the lagged dependent variable.

FE Estimate: Biased Downward

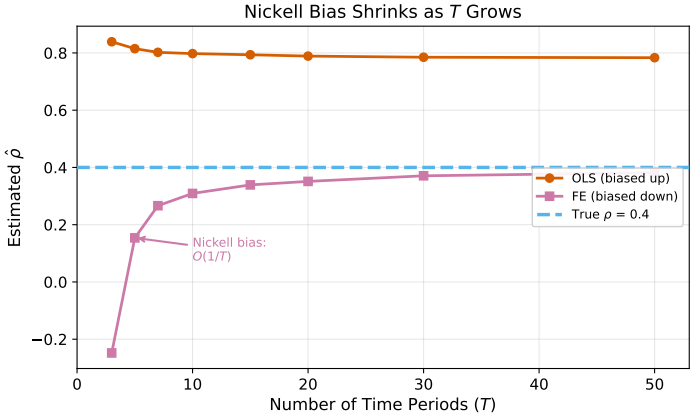


FE Estimate: Biased Downward

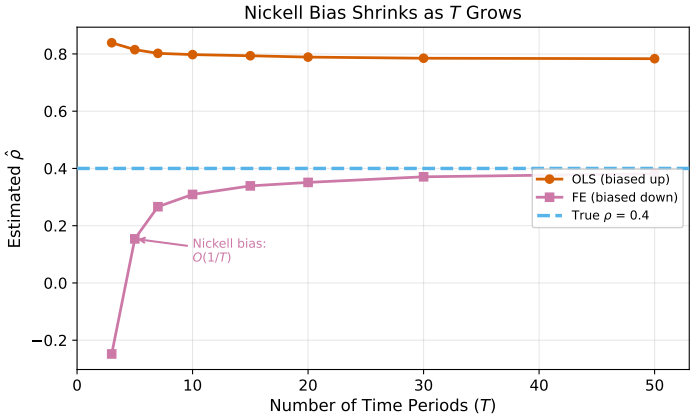


FE estimates $\hat{\rho}^{FE} \approx 0.34$, well below the true $\rho = 0.40$. FE overcorrects: it removes α_i but introduces a new bias in the opposite direction.

Nickell Bias Shrinks as T Grows



Nickell Bias Shrinks as T Grows



The FE bias is $O(1/T)$ (roughly proportional to $1/T$): it is severe for short panels ($T = 3, 5$) but shrinks as T grows. The OLS bias barely moves because it stems from cross-sectional heterogeneity, not from T.

The Bounds: OLS Up, FE Down

We now have two biased estimators that bracket the truth:

Estimator	$\hat{\rho}$	**Bias direction**
OLS	≈ 0.69	Upward (ignores α_i)
FE	≈ 0.34	Downward (Nickell bias)
True	**0.40**	

The Bounds: OLS Up, FE Down

We now have two biased estimators that bracket the truth:

Estimator	$\hat{\rho}$	**Bias direction**
OLS	≈ 0.69	Upward (ignores α_i)
FE	≈ 0.34	Downward (Nickell bias)
True	**0.40**	

In well-behaved dynamic panels, the true ρ lies **between** OLS and FE. This gives a useful sanity check for any estimator you apply.

The Bounds: OLS Up, FE Down

We now have two biased estimators that bracket the truth:

Estimator	$\hat{\rho}$	**Bias direction**
OLS	≈ 0.69	Upward (ignores α_i)
FE	≈ 0.34	Downward (Nickell bias)
True	**0.40**	

In well-behaved dynamic panels, the true ρ lies **between** OLS and FE. This gives a useful sanity check for any estimator you apply.

Both OLS and FE are biased. Is there a way out?

Two Problems, One Strategy

Where we stand:

- **Problem 1:** OLS ignores α_i $\implies$ biased upward.
- **Problem 2:** FE removes α_i by demeaning, but demeaning creates Nickell bias $\implies$ biased downward.

Two Problems, One Strategy

Where we stand:

- **Problem 1:** OLS ignores α_i $\implies$ biased upward.
- **Problem 2:** FE removes α_i by demeaning, but demeaning creates Nickell bias $\implies$ biased downward.

The root cause of Nickell bias is *demeaning*: it spreads every error across every time period. What if we eliminated α_i a different way?

Two Problems, One Strategy

Where we stand:

- **Problem 1:** OLS ignores α_i $\implies$ biased upward.
- **Problem 2:** FE removes α_i by demeaning, but demeaning creates Nickell bias $\implies$ biased downward.

The root cause of Nickell bias is *demeaning*: it spreads every error across every time period. What if we eliminated α_i a different way?

The plan:

1. Eliminate α_i by **first-differencing** instead of demeaning.
2. First-differencing creates its own endogeneity problem, so find **valid instruments** for the resulting equation.

Two Problems, One Strategy

Where we stand:

- **Problem 1:** OLS ignores α_i $\implies$ biased upward.
- **Problem 2:** FE removes α_i by demeaning, but demeaning creates Nickell bias $\implies$ biased downward.

The root cause of Nickell bias is *demeaning*: it spreads every error across every time period. What if we eliminated α_i a different way?

The plan:

1. Eliminate α_i by **first-differencing** instead of demeaning.
2. First-differencing creates its own endogeneity problem, so find **valid instruments** for the resulting equation.

$\implies$ This is the Arellano-Bond (1991) approach.

Outline

1 The Problem: Persistence in Panel Data

2 The FE Attempt and Nickell Bias

3 Arellano-Bond GMM

4 Decision Framework

5 Summary

Step 1: First-Difference to Eliminate α_i

AB = Arellano-Bond. GMM = Generalized Method of Moments.

Instead of demeaning (which creates Nickell bias), **first-difference** the model:

$$\Delta y_{it} = \rho \, \Delta y_{i,t-1} + \beta \, \Delta \text{invest}_{it} + \Delta \varepsilon_{it}$$

Step 1: First-Difference to Eliminate α_i

AB = Arellano-Bond. GMM = Generalized Method of Moments.

Instead of demeaning (which creates Nickell bias), **first-difference** the model:

$$\Delta y_{it} = \rho \, \Delta y_{i,t-1} + \beta \, \Delta \text{invest}_{it} + \Delta \varepsilon_{it}$$

where $\Delta y_{it} = y_{it} - y_{i,t-1}$.

Step 1: First-Difference to Eliminate α_i

AB = Arellano-Bond. GMM = Generalized Method of Moments.

Instead of demeaning (which creates Nickell bias), **first-difference** the model:

$$\Delta y_{it} = \rho \, \Delta y_{i,t-1} + \beta \, \Delta \text{invest}_{it} + \Delta \varepsilon_{it}$$

where $\Delta y_{it} = y_{it} - y_{i,t-1}$.

The fixed effect α_i is gone (differenced out). But we still cannot run OLS on this equation.

Step 1: First-Difference to Eliminate α_i

AB = Arellano-Bond. GMM = Generalized Method of Moments.

Instead of demeaning (which creates Nickell bias), **first-difference** the model:

$$\Delta y_{it} = \rho \, \Delta y_{i,t-1} + \beta \, \Delta \text{invest}_{it} + \Delta \varepsilon_{it}$$

where $\Delta y_{it} = y_{it} - y_{i,t-1}$.

The fixed effect α_i is gone (differenced out). But we still cannot run OLS on this equation.

What goes wrong if we try?

OLS on First-Differenced Data: Still Wrong

OLS on First-Differenced Data: Still Wrong

OLS on the differenced data gives a slope of ≈ -0.32, wildly wrong (true $\rho = 0.40$). First-differencing removed α_i, but something else went wrong. What?

The Problem: $\Delta y_{i,t-1}$ and $\Delta \varepsilon_{it}$ Share a Term

Expand the terms:

$$\Delta y_{i,t-1} = y_{i,t-1} - y_{i,t-2} \quad \text{contains } \varepsilon_{i,t-1}$$
$$\Delta \varepsilon_{it} = \varepsilon_{it} - \varepsilon_{i,t-1} \quad \text{also contains } \varepsilon_{i,t-1}$$

The Problem: $\Delta y_{i,t-1}$ and $\Delta \varepsilon_{it}$ Share a Term

Expand the terms:

$$\Delta y_{i,t-1} = y_{i,t-1} - y_{i,t-2} \quad \text{contains } \varepsilon_{i,t-1}$$
$$\Delta \varepsilon_{it} = \varepsilon_{it} - \varepsilon_{i,t-1} \quad \text{also contains } \varepsilon_{i,t-1}$$

$\implies$ $\Delta y_{i,t-1}$ and $\Delta \varepsilon_{it}$ share $\varepsilon_{i,t-1}$, so:

$$\text{Cov}(\Delta y_{i,t-1},\ \Delta \varepsilon_{it}) \neq 0$$

The Problem: $\Delta y_{i,t-1}$ and $\Delta \varepsilon_{it}$ Share a Term

Expand the terms:

$$\Delta y_{i,t-1} = y_{i,t-1} - y_{i,t-2} \quad \text{contains } \varepsilon_{i,t-1}$$
$$\Delta \varepsilon_{it} = \varepsilon_{it} - \varepsilon_{i,t-1} \quad \text{also contains } \varepsilon_{i,t-1}$$

$\implies$ $\Delta y_{i,t-1}$ and $\Delta \varepsilon_{it}$ share $\varepsilon_{i,t-1}$, so:

$$\text{Cov}(\Delta y_{i,t-1},\ \Delta \varepsilon_{it}) \neq 0$$

OLS on the first-differenced equation is **inconsistent**. We need an instrument for $\Delta y_{i,t-1}$.

The Problem: $\Delta y_{i,t-1}$ and $\Delta \varepsilon_{it}$ Share a Term

Expand the terms:

$$\Delta y_{i,t-1} = y_{i,t-1} - y_{i,t-2} \quad \text{contains } \varepsilon_{i,t-1}$$
$$\Delta \varepsilon_{it} = \varepsilon_{it} - \varepsilon_{i,t-1} \quad \text{also contains } \varepsilon_{i,t-1}$$

$\implies$ $\Delta y_{i,t-1}$ and $\Delta \varepsilon_{it}$ share $\varepsilon_{i,t-1}$, so:

$$\text{Cov}(\Delta y_{i,t-1},\ \Delta \varepsilon_{it}) \neq 0$$

OLS on the first-differenced equation is **inconsistent**. We need an instrument for $\Delta y_{i,t-1}$.
What variable is correlated with $\Delta y_{i,t-1}$ but uncorrelated with $\Delta \varepsilon_{it}$?

Step 2: Use $y_{i,t-2}$ as an Instrument

The **Arellano-Bond (1991)** insight: $y_{i,t-2}$ is a valid instrument for $\Delta y_{i,t-1}$.

Step 2: Use $y_{i,t-2}$ as an Instrument

The **Arellano-Bond (1991)** insight: $y_{i,t-2}$ is a valid instrument for $\Delta y_{i,t-1}$.

Relevance:

- $\Delta y_{i,t-1} = y_{i,t-1} - y_{i,t-2}$, and $y_{i,t-1}$ depends on $y_{i,t-2}$ through the dynamic model
- $\implies \text{Corr}(y_{i,t-2}, \Delta y_{i,t-1}) \neq 0$ ✓

Step 2: Use $y_{i,t-2}$ as an Instrument

The **Arellano-Bond (1991)** insight: $y_{i,t-2}$ is a valid instrument for $\Delta y_{i,t-1}$.

Relevance:

- $\Delta y_{i,t-1} = y_{i,t-1} - y_{i,t-2}$, and $y_{i,t-1}$ depends on $y_{i,t-2}$ through the dynamic model
- $\implies$ $\text{Corr}(y_{i,t-2}, \Delta y_{i,t-1}) \neq 0$ ✓

Validity (exclusion restriction):

- $\Delta \varepsilon_{it} = \varepsilon_{it} - \varepsilon_{i,t-1}$, which contains only period-t and period-$(t-1)$ errors
- $y_{i,t-2}$ depends on $\varepsilon_{i,t-2}$ and earlier, not on $\varepsilon_{i,t-1}$ or ε_{it}
- $\implies$ $\text{Cov}(y_{i,t-2}, \Delta \varepsilon_{it}) = 0$ ✓ (assuming no serial correlation in ε)

Step 2: Use $y_{i,t-2}$ as an Instrument

The **Arellano-Bond (1991)** insight: $y_{i,t-2}$ is a valid instrument for $\Delta y_{i,t-1}$.

Relevance:

- $\Delta y_{i,t-1} = y_{i,t-1} - y_{i,t-2}$, and $y_{i,t-1}$ depends on $y_{i,t-2}$ through the dynamic model
- $\implies$ $\text{Corr}(y_{i,t-2}, \Delta y_{i,t-1}) \neq 0$ ✓

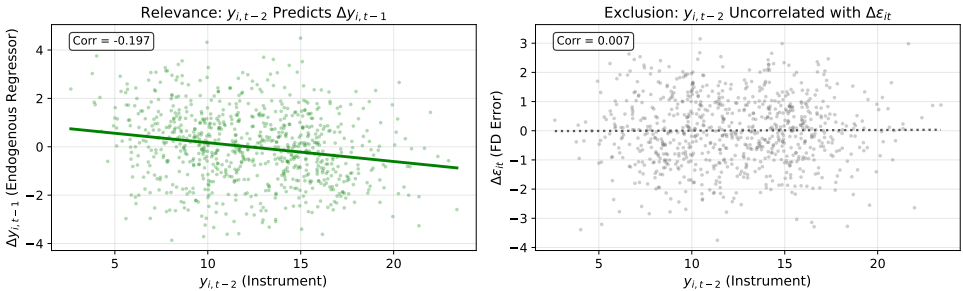
Validity (exclusion restriction):

- $\Delta \varepsilon_{it} = \varepsilon_{it} - \varepsilon_{i,t-1}$, which contains only period-t and period-$(t-1)$ errors
- $y_{i,t-2}$ depends on $\varepsilon_{i,t-2}$ and earlier, not on $\varepsilon_{i,t-1}$ or ε_{it}
- $\implies$ $\text{Cov}(y_{i,t-2}, \Delta \varepsilon_{it}) = 0$ ✓ (assuming no serial correlation in ε)

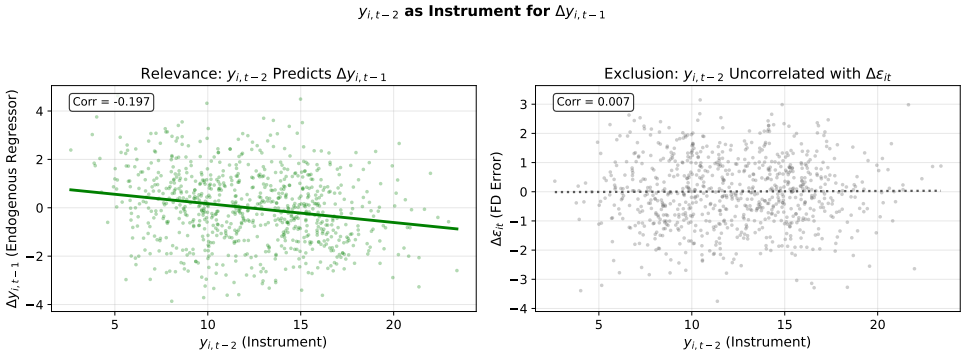
This exclusion restriction is an **assumption**. It fails if the original errors ε_{it} are serially correlated, since then $y_{i,t-2}$ would correlate with $\varepsilon_{i,t-1}$ inside $\Delta \varepsilon_{it}$.

The Instrument in Action

$y_{i,t-2}$ **as Instrument for** $\Delta y_{i,t-1}$

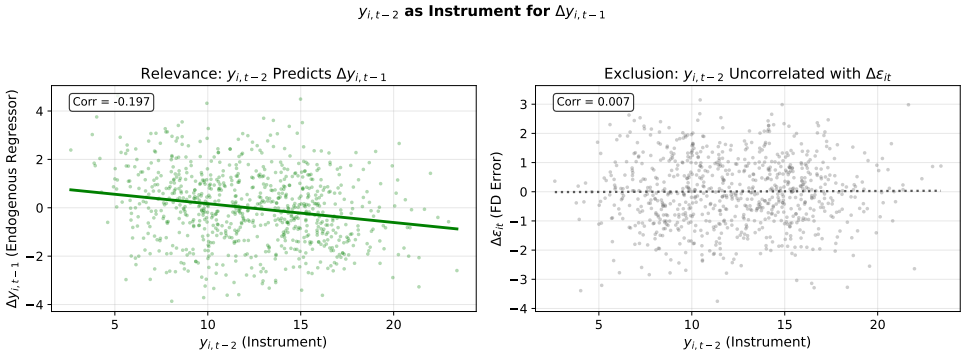


The Instrument in Action



Left: $y_{i,t-2}$ predicts $\Delta y_{i,t-1}$ (relevance). The correlation is negative because $y_{i,t-2}$ appears with a minus sign in $\Delta y_{i,t-1} = y_{i,t-1} - y_{i,t-2}$. Any nonzero correlation suffices for relevance.

The Instrument in Action



Left: $y_{i,t-2}$ predicts $\Delta y_{i,t-1}$ (relevance). The correlation is negative because $y_{i,t-2}$ appears with a minus sign in $\Delta y_{i,t-1} = y_{i,t-1} - y_{i,t-2}$. Any nonzero correlation suffices for relevance.

Right: $y_{i,t-2}$ is uncorrelated with $\Delta \varepsilon_{it}$ (exclusion). This single-period instrument looks weak, but AB uses many lags, which collectively provide strong identification.

Expanding the Instrument Set

The set of valid instruments **grows with** t:

Period	**Available instruments for** $\Delta y_{i,t-1}$
$t = 3$	$y_{i,1}$
$t = 4$	$y_{i,1},\ y_{i,2}$
$t = 5$	$y_{i,1},\ y_{i,2},\ y_{i,3}$
$\vdots$	$\vdots$
$t = T$	$y_{i,1},\ y_{i,2},\ \ldots,\ y_{i,T-2}$

Expanding the Instrument Set

The set of valid instruments **grows with** t:

Period	**Available instruments for** $\Delta y_{i,t-1}$
$t = 3$	$y_{i,1}$
$t = 4$	$y_{i,1},\ y_{i,2}$
$t = 5$	$y_{i,1},\ y_{i,2},\ y_{i,3}$
$\vdots$	$\vdots$
$t = T$	$y_{i,1},\ y_{i,2},\ \ldots,\ y_{i,T-2}$

This creates many **moment conditions**. A moment condition is a restriction on population averages. Here, it says each instrument is uncorrelated with the error, the same logic as an IV exclusion restriction:

$$\mathrm{E}[y_{i,s} \cdot \Delta\varepsilon_{it}] = 0 \quad \text{for all } s \leq t-2$$

Expanding the Instrument Set

The set of valid instruments **grows with** t:

Period	**Available instruments for** $\Delta y_{i,t-1}$
$t = 3$	$y_{i,1}$
$t = 4$	$y_{i,1},\ y_{i,2}$
$t = 5$	$y_{i,1},\ y_{i,2},\ y_{i,3}$
$\vdots$	$\vdots$
$t = T$	$y_{i,1},\ y_{i,2},\ \ldots,\ y_{i,T-2}$

This creates many **moment conditions**. A moment condition is a restriction on population averages. Here, it says each instrument is uncorrelated with the error, the same logic as an IV exclusion restriction:

$$\mathrm{E}[y_{i,s} \cdot \Delta\varepsilon_{it}] = 0 \quad \text{for all } s \leq t-2$$

We need a way to combine all these moment conditions into a single estimate. That is what GMM does.

GMM: Combining Many Instruments

We have more instruments than parameters (the model is **overidentified**). Standard 2SLS can handle multiple instruments, but GMM provides optimal weighting when the number of moment conditions grows.

GMM: Combining Many Instruments

We have more instruments than parameters (the model is **overidentified**). Standard 2SLS can handle multiple instruments, but GMM provides optimal weighting when the number of moment conditions grows.

Intuition: GMM finds the $\hat{\rho}$ that makes all the instrument-error correlations as close to zero as possible.

GMM: Combining Many Instruments

We have more instruments than parameters (the model is **overidentified**). Standard 2SLS can handle multiple instruments, but GMM provides optimal weighting when the number of moment conditions grows.

Intuition: GMM finds the $\hat{\rho}$ that makes all the instrument-error correlations as close to zero as possible.

In Stata:

```
xtabond growth invest, lags(1) twostep
```

You specify the dependent variable, exogenous regressors, and the number of lags. The software constructs the full instrument matrix and solves the GMM problem.

GMM: Combining Many Instruments

We have more instruments than parameters (the model is **overidentified**). Standard 2SLS can handle multiple instruments, but GMM provides optimal weighting when the number of moment conditions grows.

Intuition: GMM finds the $\hat{\rho}$ that makes all the instrument-error correlations as close to zero as possible.

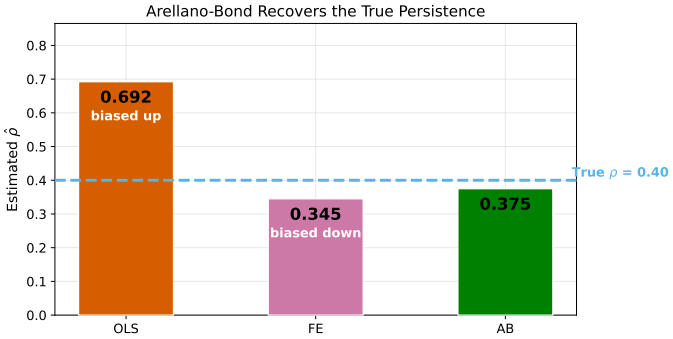
In Stata:

```
xtabond growth invest, lags(1) twostep
```

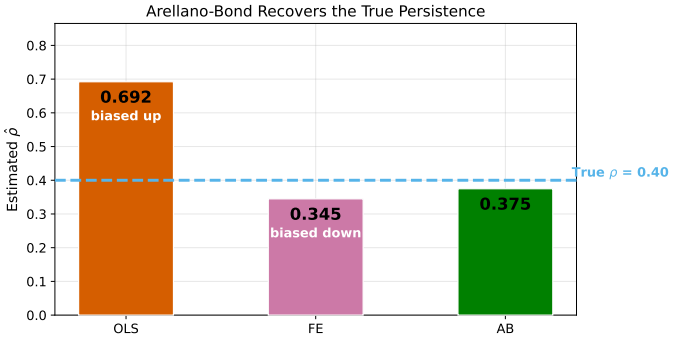
You specify the dependent variable, exogenous regressors, and the number of lags. The software constructs the full instrument matrix and solves the GMM problem.

Note: "One-step" GMM uses a preliminary weighting matrix; "two-step" re-estimates with an optimal weighting matrix from step-1 residuals. Two-step is more efficient.

The Result: AB Recovers the True ρ

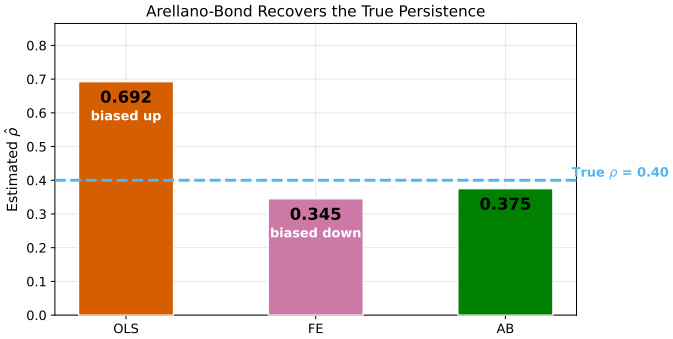


The Result: AB Recovers the True ρ



AB estimates $\hat{\rho}^{AB} \approx 0.37$, close to the true $\rho = 0.40$. It falls between the OLS upper bound (0.69) and the FE lower bound (0.34), as expected.

The Result: AB Recovers the True ρ



AB estimates $\hat{\rho}^{AB} \approx 0.37$, close to the true $\rho = 0.40$. It falls between the OLS upper bound (0.69) and the FE lower bound (0.34), as expected.

$\Longrightarrow$ By first-differencing (to remove α_i) and instrumenting (to handle the $\Delta\varepsilon_{it}$ correlation), AB produces a consistent estimator.

Checking the AB Assumptions

AB relies on **no serial correlation** in ε_{it}. Two standard tests:

Checking the AB Assumptions

AB relies on **no serial correlation** in ε_{it}. Two standard tests:

1. **AR(1) and AR(2) tests on $\Delta\varepsilon_{it}$:**
   - AR(1) in $\Delta\varepsilon$: we *expect* this (by construction, $\Delta\varepsilon_{it}$ and $\Delta\varepsilon_{i,t-1}$ share $\varepsilon_{i,t-1}$)
   - AR(2) in $\Delta\varepsilon$: we should *not* find this. Why? $\Delta\varepsilon_{it}$ and $\Delta\varepsilon_{i,t-2}$ share no terms unless ε_{it} itself is autocorrelated. So AR(2) in the differenced errors is really a test for AR(1) in the original errors.

Checking the AB Assumptions

AB relies on **no serial correlation** in ε_{it}. Two standard tests:

1. AR(1) and AR(2) tests on $\Delta\varepsilon_{it}$:

- AR(1) in $\Delta\varepsilon$: we *expect* this (by construction, $\Delta\varepsilon_{it}$ and $\Delta\varepsilon_{i,t-1}$ share $\varepsilon_{i,t-1}$)
- AR(2) in $\Delta\varepsilon$: we should *not* find this. Why? $\Delta\varepsilon_{it}$ and $\Delta\varepsilon_{i,t-2}$ share no terms unless ε_{it} itself is autocorrelated. So AR(2) in the differenced errors is really a test for AR(1) in the original errors.

2. Sargan/Hansen test of overidentifying restrictions:

- We have more instruments than strictly needed. The Sargan test checks whether they all agree. If some instruments give very different answers, at least one is probably invalid.
- H_0: all instruments are valid. Rejection $\implies$ at least one instrument is correlated with the error.

Checking the AB Assumptions

AB relies on **no serial correlation** in ε_{it}. Two standard tests:

1. AR(1) and AR(2) tests on $\Delta\varepsilon_{it}$:

- AR(1) in $\Delta\varepsilon$: we *expect* this (by construction, $\Delta\varepsilon_{it}$ and $\Delta\varepsilon_{i,t-1}$ share $\varepsilon_{i,t-1}$)
- AR(2) in $\Delta\varepsilon$: we should *not* find this. Why? $\Delta\varepsilon_{it}$ and $\Delta\varepsilon_{i,t-2}$ share no terms unless ε_{it} itself is autocorrelated. So AR(2) in the differenced errors is really a test for AR(1) in the original errors.

2. Sargan/Hansen test of overidentifying restrictions:

- We have more instruments than strictly needed. The Sargan test checks whether they all agree. If some instruments give very different answers, at least one is probably invalid.
- H_0: all instruments are valid. Rejection $\implies$ at least one instrument is correlated with the error.

$\implies$ Report both tests. If AR(2) is significant or Sargan rejects, the AB assumptions are in doubt.

Diagnostics: Our Simulation

In our simulated data (ε_{it} i.i.d. by construction), the tests should confirm AB is valid:

Diagnostics: Our Simulation

In our simulated data (ε_{it} i.i.d. by construction), the tests should confirm AB is valid:

Test	H_0	**Expected result**	**Conclusion**
AR(1) in $\Delta\varepsilon$	No AR(1)	Reject (by construction)	Normal
AR(2) in $\Delta\varepsilon$	No AR(2)	Fail to reject	Instruments valid
Sargan/Hansen	All IVs valid	Fail to reject	Instruments valid

Diagnostics: Our Simulation

In our simulated data (ε_{it} i.i.d. by construction), the tests should confirm AB is valid:

Test	H_0	**Expected result**	**Conclusion**
AR(1) in $\Delta\varepsilon$	No AR(1)	Reject (by construction)	Normal
AR(2) in $\Delta\varepsilon$	No AR(2)	Fail to reject	Instruments valid
Sargan/Hansen	All IVs valid	Fail to reject	Instruments valid

Since we generated the data with serially uncorrelated errors, we know the AB assumptions hold. In real applications, you must rely on these tests.

Diagnostics: Our Simulation

In our simulated data (ε_{it} i.i.d. by construction), the tests should confirm AB is valid:

Test	H_0	**Expected result**	**Conclusion**
AR(1) in $\Delta\varepsilon$	No AR(1)	Reject (by construction)	Normal
AR(2) in $\Delta\varepsilon$	No AR(2)	Fail to reject	Instruments valid
Sargan/Hansen	All IVs valid	Fail to reject	Instruments valid

Since we generated the data with serially uncorrelated errors, we know the AB assumptions hold. In real applications, you must rely on these tests.

$\Longrightarrow$ Always report AR(2) and Sargan/Hansen p-values alongside your AB estimates.

When Does AB Struggle?

AB uses only the **first-differenced equation**, instrumented with lagged levels. This works well when ρ is moderate.

When Does AB Struggle?

AB uses only the **first-differenced equation**, instrumented with lagged levels. This works well when ρ is moderate.

But when ρ is close to 1, AB runs into trouble. Why?

When Does AB Struggle?

AB uses only the **first-differenced equation**, instrumented with lagged levels. This works well when ρ is moderate.

But when ρ is close to 1, AB runs into trouble. Why?

- When $\rho \approx 1$, the series behaves like a random walk: levels barely change over time.
- Past levels carry almost no information about *changes* $\Longrightarrow$ the instruments (lagged levels) are **weak**.
- Weak instruments produce imprecise, unreliable estimates.

When Does AB Struggle?

AB uses only the **first-differenced equation**, instrumented with lagged levels. This works well when ρ is moderate.

But when ρ is close to 1, AB runs into trouble. Why?

- When $\rho \approx 1$, the series behaves like a random walk: levels barely change over time.
- Past levels carry almost no information about *changes* $\implies$ the instruments (lagged levels) are **weak**.
- Weak instruments produce imprecise, unreliable estimates.

$\implies$ We need additional instruments. Where do they come from?

System GMM: Blundell-Bond

Blundell and Bond (1998) proposed **System GMM**: stack two equations:

1. First-differenced equation (as in AB), instrumented with lagged *levels*
2. Levels equation, instrumented with lagged *differences*

System GMM: Blundell-Bond

Blundell and Bond (1998) proposed **System GMM**: stack two equations:

1. First-differenced equation (as in AB), instrumented with lagged *levels*
2. Levels equation, instrumented with lagged *differences*

The extra moment conditions improve efficiency, especially when:

- ρ is close to 1 (near unit root)
- T is small
- The variance of α_i is large relative to ε_{it}

System GMM: Blundell-Bond

Blundell and Bond (1998) proposed **System GMM**: stack two equations:

1. First-differenced equation (as in AB), instrumented with lagged *levels*
2. Levels equation, instrumented with lagged *differences*

The extra moment conditions improve efficiency, especially when:

- ρ is close to 1 (near unit root)
- T is small
- The variance of α_i is large relative to ε_{it}

In Stata: `xtabond2 growth L.growth invest, gmm(L.growth) iv(invest) twostep`

System GMM: Blundell-Bond

Blundell and Bond (1998) proposed **System GMM**: stack two equations:

1. First-differenced equation (as in AB), instrumented with lagged *levels*
2. Levels equation, instrumented with lagged *differences*

The extra moment conditions improve efficiency, especially when:

- ρ is close to 1 (near unit root)
- T is small
- The variance of α_i is large relative to ε_{it}

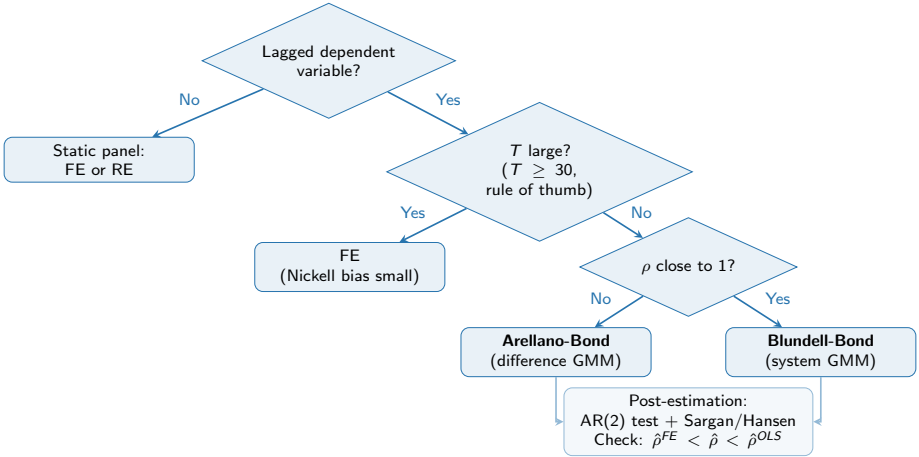
In Stata: `xtabond2 growth L.growth invest, gmm(L.growth) iv(invest) twostep`

$\Longrightarrow$ System GMM adds an assumption: $\mathrm{E}[\Delta y_{i,t} \cdot \alpha_i] = 0$ (each country's growth has settled to its long-run average by the start of the sample). This is testable via the difference-in-Sargan test.

Outline

1. The Problem: Persistence in Panel Data
2. The FE Attempt and Nickell Bias
3. Arellano-Bond GMM
4. Decision Framework
5. Summary

Decision Flowchart



Check whether $y_{i,t-1}$ is in the model; if T is small, choose AB or BB based on persistence; always run post-estimation diagnostics.

Outline

1. The Problem: Persistence in Panel Data
2. The FE Attempt and Nickell Bias
3. Arellano-Bond GMM
4. Decision Framework
5. Summary

Summary

We started with a panel of 50 countries and wanted to estimate the persistence of GDP growth. Both OLS and FE are biased: OLS overestimates, FE underestimates.

Summary

We started with a panel of 50 countries and wanted to estimate the persistence of GDP growth. Both OLS and FE are biased: OLS overestimates, FE underestimates.

1. **OLS** ignores α_i $\implies$ confuses level differences with persistence $\implies$ $\hat{\rho}^{OLS}$ biased **upward**.

Summary

We started with a panel of 50 countries and wanted to estimate the persistence of GDP growth. Both OLS and FE are biased: OLS overestimates, FE underestimates.

1. **OLS** ignores α_i $\implies$ confuses level differences with persistence $\implies$ $\hat{\rho}^{OLS}$ biased **upward**.
2. **FE** removes α_i by demeaning, but demeaning creates a mechanical correlation between $\ddot{y}_{i,t-1}$ and $\ddot{\varepsilon}_{it}$ $\implies$ $\hat{\rho}^{FE}$ biased **downward** (Nickell bias, $O(1/T)$).

Summary

We started with a panel of 50 countries and wanted to estimate the persistence of GDP growth. Both OLS and FE are biased: OLS overestimates, FE underestimates.

1. **OLS** ignores α_i $\implies$ confuses level differences with persistence $\implies$ $\hat{\rho}^{OLS}$ biased **upward**.
2. **FE** removes α_i by demeaning, but demeaning creates a mechanical correlation between $\ddot{y}_{i,t-1}$ and $\ddot{\varepsilon}_{it}$ $\implies$ $\hat{\rho}^{FE}$ biased **downward** (Nickell bias, $O(1/T)$).
3. **Arellano-Bond** first-differences to remove α_i, then uses $y_{i,t-2}$ (and deeper lags) as instruments for $\Delta y_{i,t-1}$. Consistent when ε_{it} is serially uncorrelated.

Summary

We started with a panel of 50 countries and wanted to estimate the persistence of GDP growth. Both OLS and FE are biased: OLS overestimates, FE underestimates.

1. **OLS** ignores α_i $\implies$ confuses level differences with persistence $\implies$ $\hat{\rho}^{OLS}$ biased **upward**.
2. **FE** removes α_i by demeaning, but demeaning creates a mechanical correlation between $\ddot{y}_{i,t-1}$ and $\ddot{\varepsilon}_{it}$ $\implies$ $\hat{\rho}^{FE}$ biased **downward** (Nickell bias, $O(1/T)$).
3. **Arellano-Bond** first-differences to remove α_i, then uses $y_{i,t-2}$ (and deeper lags) as instruments for $\Delta y_{i,t-1}$. Consistent when ε_{it} is serially uncorrelated.
4. **Blundell-Bond** (System GMM) adds moment conditions from the levels equation. Preferable when ρ is close to 1 or T is very small.

Summary

We started with a panel of 50 countries and wanted to estimate the persistence of GDP growth. Both OLS and FE are biased: OLS overestimates, FE underestimates.

1. **OLS** ignores α_i $\implies$ confuses level differences with persistence $\implies$ $\hat{\rho}^{OLS}$ biased **upward**.
2. **FE** removes α_i by demeaning, but demeaning creates a mechanical correlation between $\ddot{y}_{i,t-1}$ and $\ddot{\varepsilon}_{it}$ $\implies$ $\hat{\rho}^{FE}$ biased **downward** (Nickell bias, $O(1/T)$).
3. **Arellano-Bond** first-differences to remove α_i, then uses $y_{i,t-2}$ (and deeper lags) as instruments for $\Delta y_{i,t-1}$. Consistent when ε_{it} is serially uncorrelated.
4. **Blundell-Bond** (System GMM) adds moment conditions from the levels equation. Preferable when ρ is close to 1 or T is very small.
5. **Sanity check:** a credible estimate should satisfy $\hat{\rho}^{FE} < \hat{\rho} < \hat{\rho}^{OLS}$.

Summary

We started with a panel of 50 countries and wanted to estimate the persistence of GDP growth. Both OLS and FE are biased: OLS overestimates, FE underestimates.

1. **OLS** ignores α_i $\implies$ confuses level differences with persistence $\implies$ $\hat{\rho}^{OLS}$ biased **upward**.
2. **FE** removes α_i by demeaning, but demeaning creates a mechanical correlation between $\ddot{y}_{i,t-1}$ and $\ddot{\varepsilon}_{it}$ $\implies$ $\hat{\rho}^{FE}$ biased **downward** (Nickell bias, $O(1/T)$).
3. **Arellano-Bond** first-differences to remove α_i, then uses $y_{i,t-2}$ (and deeper lags) as instruments for $\Delta y_{i,t-1}$. Consistent when ε_{it} is serially uncorrelated.
4. **Blundell-Bond** (System GMM) adds moment conditions from the levels equation. Preferable when ρ is close to 1 or T is very small.
5. **Sanity check:** a credible estimate should satisfy $\hat{\rho}^{FE} < \hat{\rho} < \hat{\rho}^{OLS}$.

$\implies$ When your panel model has a lagged dependent variable and small T, use AB or System GMM, not OLS or FE. First-difference to remove the fixed effect, then instrument to restore consistency.

Thank you!

jakeanderson@g.ucla.edu

The Hausman-Taylor Estimator

Estimating Time-Invariant Effects with Endogeneity

Jake Anderson

May 16, 2026

Outline

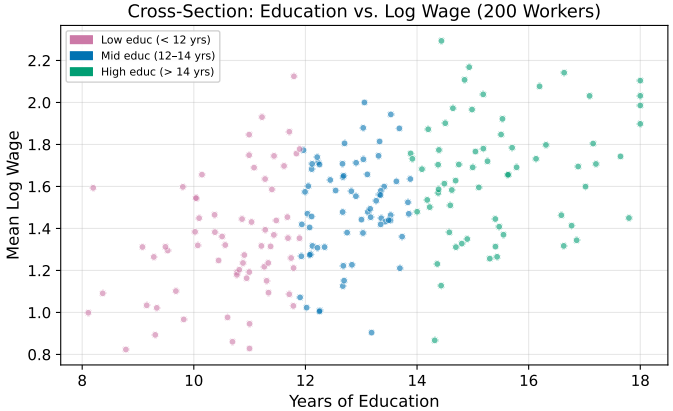
1. The Problem: Returns to Education
2. The Hausman-Taylor Estimator
3. When HT Works (and When It Doesn't)
4. Decision Framework
5. Summary

The Data

A policymaker wants to know the return to one more year of schooling. We have panel data on **200 workers** over **5 years**. How hard can it be?

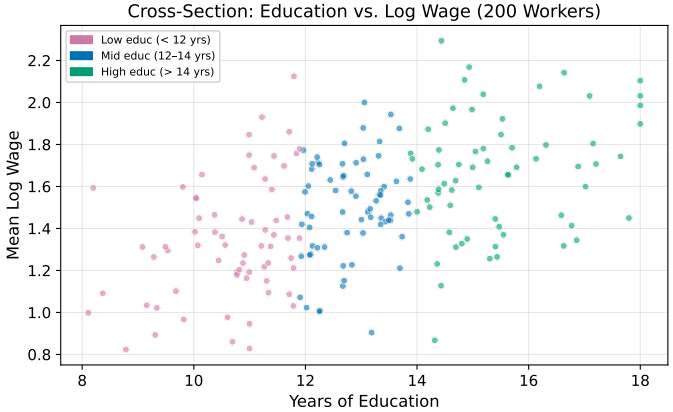
The Data

A policymaker wants to know the return to one more year of schooling. We have panel data on **200 workers** over **5 years**. How hard can it be?



The Data

A policymaker wants to know the return to one more year of schooling. We have panel data on **200 workers** over **5 years**. How hard can it be?



Higher education is associated with higher wages. But is the relationship **causal**?

Pooled OLS: A First Pass

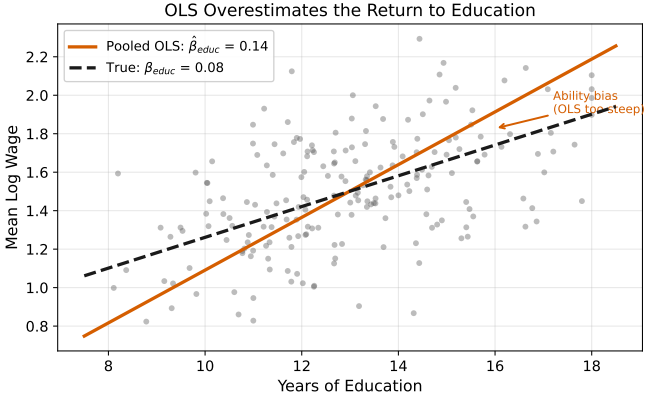
Run pooled OLS of log wages on education, experience, union, and race:

$$\log(\text{wage}_{it}) = \beta_0 + \beta_1 \text{educ}_i + \beta_2 \text{exper}_{it} + \beta_3 \text{union}_{it} + \beta_4 \text{black}_i + \varepsilon_{it}$$

Pooled OLS: A First Pass

Run pooled OLS of log wages on education, experience, union, and race:

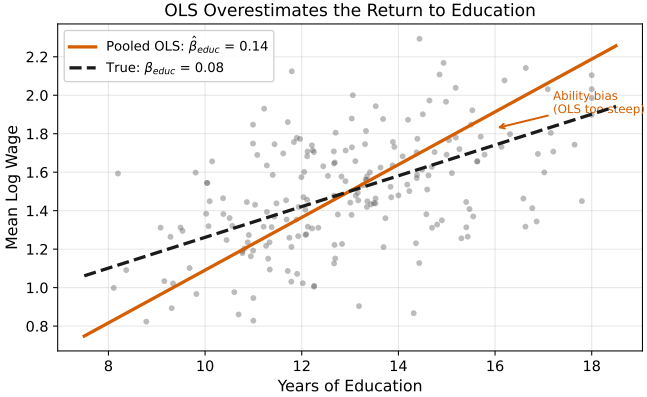
$$\log(\text{wage}_{it}) = \beta_0 + \beta_1 \text{educ}_i + \beta_2 \text{exper}_{it} + \beta_3 \text{union}_{it} + \beta_4 \text{black}_i + \varepsilon_{it}$$



Pooled OLS: A First Pass

Run pooled OLS of log wages on education, experience, union, and race:

$$\log(\text{wage}_{it}) = \beta_0 + \beta_1 \text{educ}_i + \beta_2 \text{exper}_{it} + \beta_3 \text{union}_{it} + \beta_4 \text{black}_i + \varepsilon_{it}$$



OLS estimates $\hat{\beta}_{\text{educ}} \approx 0.14$, but the true return is 0.08. Why the overestimate?

The Endogeneity Problem: Omitted Ability

Unobserved **ability** (α_i) affects both wages and education:

The Endogeneity Problem: Omitted Ability

Unobserved **ability** (α_i) affects both wages and education:

- High-ability workers earn more $\implies$ α_i belongs in the wage equation
- High-ability workers get more education $\implies$ $\text{Cov}(\alpha_i, \text{educ}_i) > 0$

The Endogeneity Problem: Omitted Ability

Unobserved **ability** (α_i) affects both wages and education:

- High-ability workers earn more $\implies$ α_i belongs in the wage equation
- High-ability workers get more education $\implies$ $\text{Cov}(\alpha_i, \text{educ}_i) > 0$

OVB formula (single-regressor version):

$$\hat{\beta}^{\text{OLS}}_{\text{educ}} = \underbrace{\beta_{\text{educ}}}_{\text{true}} + \underbrace{\beta_\alpha}_{\substack{\text{effect of ability} \\ \text{on wages}}} \times \underbrace{\delta_1}_{\substack{\text{relationship of} \\ \text{ability to educ}}}$$

The Endogeneity Problem: Omitted Ability

Unobserved **ability** (α_i) affects both wages and education:

- High-ability workers earn more $\implies$ α_i belongs in the wage equation
- High-ability workers get more education $\implies$ $\text{Cov}(\alpha_i, \text{educ}_i) > 0$

OVB formula (single-regressor version):

$$\hat{\beta}^{\text{OLS}}_{\text{educ}} = \underbrace{\beta_{\text{educ}}}_{\text{true}} + \underbrace{\beta_\alpha}_{\substack{\text{effect of ability} \\ \text{on wages}}} \times \underbrace{\delta_1}_{\substack{\text{relationship of} \\ \text{ability to educ}}}$$

Both $\beta_\alpha > 0$ and $\delta_1 > 0$ $\implies$ OLS is **biased upward**.

Fixed Effects to the Rescue?

We know FE eliminates time-invariant unobservables. Write the model as:

$$\log(\text{wage}_{it}) = \alpha_i + \beta_1 \text{educ}_i + \beta_2 \text{exper}_{it} + \beta_3 \text{union}_{it} + \beta_4 \text{black}_i + \varepsilon_{it}$$

Fixed Effects to the Rescue?

We know FE eliminates time-invariant unobservables. Write the model as:

$$\log(\text{wage}_{it}) = \alpha_i + \beta_1 \text{educ}_i + \beta_2 \text{exper}_{it} + \beta_3 \text{union}_{it} + \beta_4 \text{black}_i + \varepsilon_{it}$$

If we demean the model, which variables survive?

Fixed Effects to the Rescue?

We know FE eliminates time-invariant unobservables. Write the model as:

$$\log(\text{wage}_{it}) = \alpha_i + \beta_1 \text{educ}_i + \beta_2 \text{exper}_{it} + \beta_3 \text{union}_{it} + \beta_4 \text{black}_i + \varepsilon_{it}$$

If we demean the model, which variables survive?

FE subtracts the worker mean from each variable:

$$\log(\text{wage}_{it}) - \overline{\log(\text{wage})}_i = \beta_2(\text{exper}_{it} - \bar{\text{exper}}_i) + \beta_3(\text{union}_{it} - \bar{\text{union}}_i) + (\varepsilon_{it} - \bar{\varepsilon}_i)$$

Fixed Effects to the Rescue?

We know FE eliminates time-invariant unobservables. Write the model as:

$$\log(\text{wage}_{it}) = \alpha_i + \beta_1 \text{educ}_i + \beta_2 \text{exper}_{it} + \beta_3 \text{union}_{it} + \beta_4 \text{black}_i + \varepsilon_{it}$$

If we demean the model, which variables survive?

FE subtracts the worker mean from each variable:

$$\log(\text{wage}_{it}) - \overline{\log(\text{wage})}_i = \beta_2(\text{exper}_{it} - \bar{\text{exper}}_i) + \beta_3(\text{union}_{it} - \bar{\text{union}}_i) + (\varepsilon_{it} - \bar{\varepsilon}_i)$$

α_i is gone. But so are $\beta_1 \text{educ}_i$ and $\beta_4 \text{black}_i$ (since educ_i does not vary over time, $\text{educ}_i - \bar{\text{educ}}_i = 0$).

Fixed Effects to the Rescue?

We know FE eliminates time-invariant unobservables. Write the model as:

$$\log(\text{wage}_{it}) = \alpha_i + \beta_1 \text{educ}_i + \beta_2 \text{exper}_{it} + \beta_3 \text{union}_{it} + \beta_4 \text{black}_i + \varepsilon_{it}$$

If we demean the model, which variables survive?

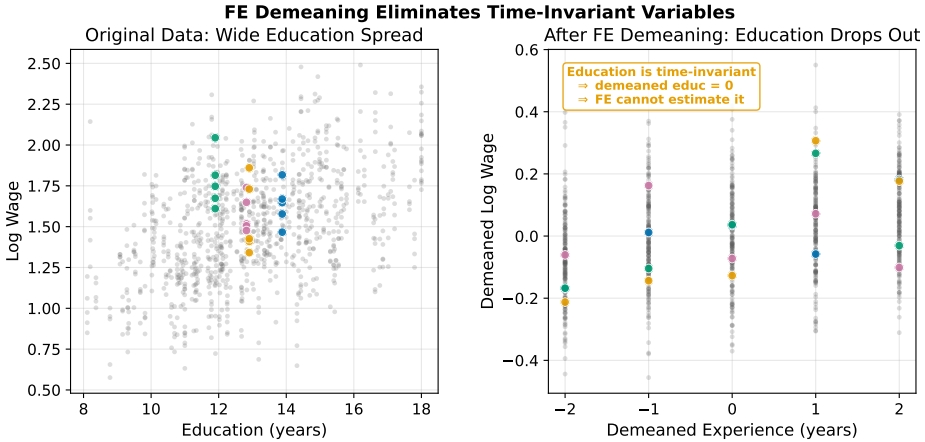
FE subtracts the worker mean from each variable:

$$\log(\text{wage}_{it}) - \overline{\log(\text{wage})}_i = \beta_2(\text{exper}_{it} - \bar{\text{exper}}_i) + \beta_3(\text{union}_{it} - \bar{\text{union}}_i) + (\varepsilon_{it} - \bar{\varepsilon}_i)$$

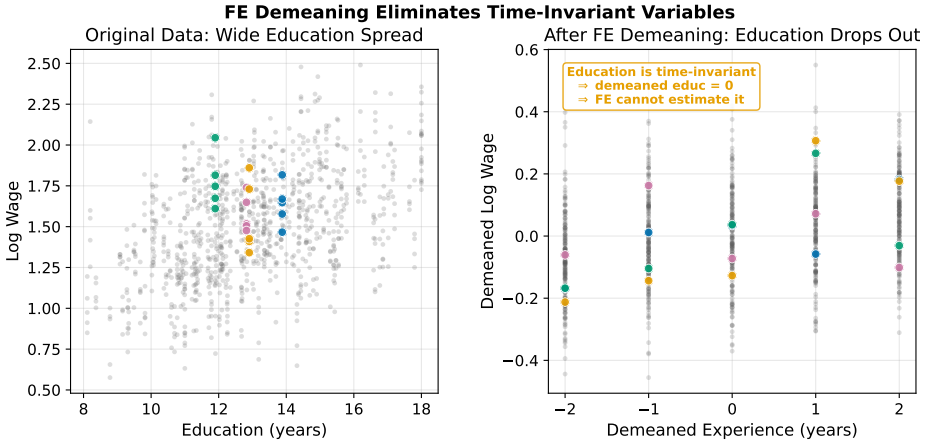
α_i is gone. But so are $\beta_1 \text{educ}_i$ and $\beta_4 \text{black}_i$ (since educ_i does not vary over time, $\text{educ}_i - \bar{\text{educ}}_i = 0$).

$\implies$ **Education is time-invariant.** Demeaning removes it entirely.

FE Demeaning: Education Disappears



FE Demeaning: Education Disappears



Each worker has **one education level** across all 5 years. After demeaning, the within-worker variation in education is exactly zero. FE cannot estimate β_{educ}.

Random Effects: Can It Estimate Education?

RE treats α_i as a random draw, uncorrelated with the regressors:

$$\log(\text{wage}_{it}) = \beta_0 + \beta_1 \text{educ}_i + \beta_2 \text{exper}_{it} + \beta_3 \text{union}_{it} + \beta_4 \text{black}_i + \alpha_i + \varepsilon_{it}$$

Random Effects: Can It Estimate Education?

RE treats α_i as a random draw, uncorrelated with the regressors:

$$\log(\text{wage}_{it}) = \beta_0 + \beta_1 \text{educ}_i + \beta_2 \text{exper}_{it} + \beta_3 \text{union}_{it} + \beta_4 \text{black}_i + \alpha_i + \varepsilon_{it}$$

RE uses both within and between variation $\implies$ it **can** estimate β_{educ}.

Random Effects: Can It Estimate Education?

RE treats α_i as a random draw, uncorrelated with the regressors:

$$\log(\text{wage}_{it}) = \beta_0 + \beta_1 \text{educ}_i + \beta_2 \text{exper}_{it} + \beta_3 \text{union}_{it} + \beta_4 \text{black}_i + \alpha_i + \varepsilon_{it}$$

RE uses both within and between variation $\implies$ it **can** estimate β_{educ}.

Will RE give us a better estimate than OLS?

Random Effects: Can It Estimate Education?

RE treats α_i as a random draw, uncorrelated with the regressors:

$$\log(\text{wage}_{it}) = \beta_0 + \beta_1 \text{educ}_i + \beta_2 \text{exper}_{it} + \beta_3 \text{union}_{it} + \beta_4 \text{black}_i + \alpha_i + \varepsilon_{it}$$

RE uses both within and between variation $\implies$ it **can** estimate β_{educ}.

Will RE give us a better estimate than OLS?

RE requires:

$$\text{Cov}(\alpha_i, \text{educ}_i) = 0$$

Random Effects: Can It Estimate Education?

RE treats α_i as a random draw, uncorrelated with the regressors:

$$\log(\text{wage}_{it}) = \beta_0 + \beta_1 \text{educ}_i + \beta_2 \text{exper}_{it} + \beta_3 \text{union}_{it} + \beta_4 \text{black}_i + \alpha_i + \varepsilon_{it}$$

RE uses both within and between variation $\implies$ it **can** estimate β_{educ}.

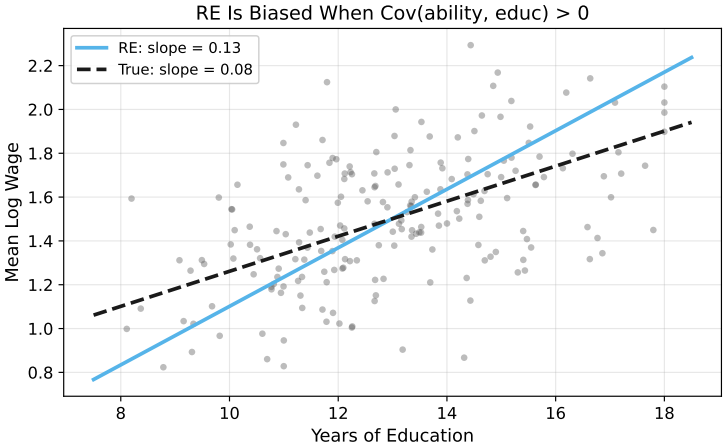
Will RE give us a better estimate than OLS?

RE requires:

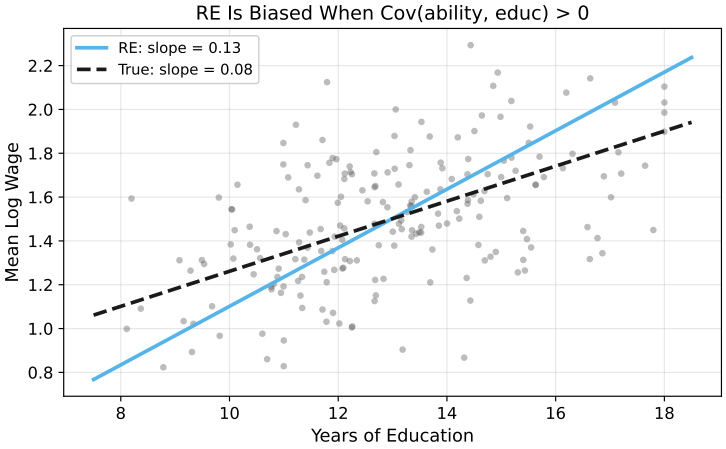
$$\text{Cov}(\alpha_i, \text{educ}_i) = 0$$

We just argued this is **violated** (ability correlates with education). So the RE estimate is **biased**.

RE Is Biased


RE Is Biased When Cov(ability, educ) > 0
RE: slope = 0.13
True: slope = 0.08
Mean Log Wage
Years of Education
0.8
1.0
1.2
1.4
1.6
1.8
2.0
2.2
8
10
12
14
16
18

RE Is Biased



RE estimates $\hat{\beta}_{\text{educ}} \approx 0.13$ (true $= 0.08$). The ability bias persists because RE does not fully eliminate α_i from the estimation.

The Dilemma

Abbreviations: TV = time-varying; TI = time-invariant.

Method	**Estimate $\hat{\beta}_{\text{educ}}$**	**Consistent?**	**Problem**
OLS	≈ 0.14	No	Omits ability
FE	N/A		Cannot estimate TI vars
RE	≈ 0.13	No	Requires $\text{Cov}(\alpha_i, x) = 0$
True	**0.08**		

The Dilemma

Abbreviations: TV = time-varying; TI = time-invariant.

Method	**Estimate $\hat{\beta}_{\text{educ}}$**	**Consistent?**	**Problem**
OLS	≈ 0.14	No	Omits ability
FE	N/A		Cannot estimate TI vars
RE	≈ 0.13	No	Requires $\text{Cov}(\alpha_i, x) = 0$
True	**0.08**		

- FE is consistent for time-varying coefficients, but **drops** education
- RE can estimate education, but is **biased** when $\text{Cov}(\alpha_i, \text{educ}_i) \neq 0$

The Dilemma

Abbreviations: TV = time-varying; TI = time-invariant.

Method	**Estimate $\hat{\beta}_{\text{educ}}$**	**Consistent?**	**Problem**
OLS	≈ 0.14	No	Omits ability
FE	N/A		Cannot estimate TI vars
RE	≈ 0.13	No	Requires $\text{Cov}(\alpha_i, x) = 0$
True	**0.08**		

- FE is consistent for time-varying coefficients, but **drops** education
- RE can estimate education, but is **biased** when $\text{Cov}(\alpha_i, \text{educ}_i) \neq 0$

The paradox: the only estimator that controls for ability (FE) is the one that cannot estimate the coefficient we care about.

The Dilemma

Abbreviations: TV = time-varying; TI = time-invariant.

Method	**Estimate $\hat{\beta}_{\text{educ}}$**	**Consistent?**	**Problem**
OLS	≈ 0.14	No	Omits ability
FE	N/A		Cannot estimate TI vars
RE	≈ 0.13	No	Requires $\text{Cov}(\alpha_i, x) = 0$
True	**0.08**		

- FE is consistent for time-varying coefficients, but **drops** education
- RE can estimate education, but is **biased** when $\text{Cov}(\alpha_i, \text{educ}_i) \neq 0$

The paradox: the only estimator that controls for ability (FE) is the one that cannot estimate the coefficient we care about.

$\Longrightarrow$ We need an estimator that handles **time-invariant endogenous** variables.

Can We Recycle What FE Gave Us?

FE gave us consistent estimates of the time-varying coefficients ($\hat{\beta}_{\text{exper}}$, $\hat{\beta}_{\text{union}}$). Those estimates produced residuals. And the variables themselves vary across individuals.

Can We Recycle What FE Gave Us?

FE gave us consistent estimates of the time-varying coefficients ($\hat{\beta}_{\text{exper}}$, $\hat{\beta}_{\text{union}}$). Those estimates produced residuals. And the variables themselves vary across individuals.

What if we could *reuse* those time-varying exogenous variables as instruments for the time-invariant endogenous variable?

Can We Recycle What FE Gave Us?

FE gave us consistent estimates of the time-varying coefficients ($\hat{\beta}_{\text{exper}}$, $\hat{\beta}_{\text{union}}$). Those estimates produced residuals. And the variables themselves vary across individuals.

What if we could *reuse* those time-varying exogenous variables as instruments for the time-invariant endogenous variable?

Think about it:

- Experience and union status are uncorrelated with ability (exogenous)
- But their **worker-level averages** are correlated with education (workers who studied longer entered the workforce later $\implies$ less experience)

Can We Recycle What FE Gave Us?

FE gave us consistent estimates of the time-varying coefficients ($\hat{\beta}_{\text{exper}}$, $\hat{\beta}_{\text{union}}$). Those estimates produced residuals. And the variables themselves vary across individuals.

What if we could *reuse* those time-varying exogenous variables as instruments for the time-invariant endogenous variable?

Think about it:

- Experience and union status are uncorrelated with ability (exogenous)
- But their **worker-level averages** are correlated with education (workers who studied longer entered the workforce later $\implies$ less experience)

$\implies$ We already have instruments sitting inside the panel. We just need a procedure that uses them.

Outline

1. The Problem: Returns to Education
2. The Hausman-Taylor Estimator
3. When HT Works (and When It Doesn't)
4. Decision Framework
5. Summary

The Hausman-Taylor Idea

Hausman and Taylor (1981) proposed using information *already in the panel* to construct instruments for the endogenous time-invariant variables.

The Hausman-Taylor Idea

Hausman and Taylor (1981) proposed using information *already in the panel* to construct instruments for the endogenous time-invariant variables.

The insight:

- Some time-varying variables (like experience) are **exogenous**: $\text{Cov}(\alpha_i, \text{exper}_{it}) = 0$
- Their **within-group means** ($\text{ex}\bar{\text{p}}\text{er}_i$) vary across individuals
- These means are correlated with education (workers who studied longer entered the workforce later $\implies$ less experience)
- But they are **uncorrelated with ability** (by assumption)

The Hausman-Taylor Idea

Hausman and Taylor (1981) proposed using information *already in the panel* to construct instruments for the endogenous time-invariant variables.

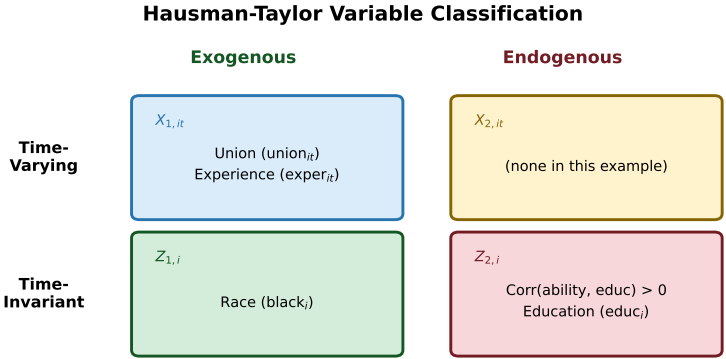
The insight:

- Some time-varying variables (like experience) are **exogenous**: $\text{Cov}(\alpha_i, \text{exper}_{it}) = 0$
- Their **within-group means** ($\text{ex}\bar{\text{p}}\text{er}_i$) vary across individuals
- These means are correlated with education (workers who studied longer entered the workforce later $\implies$ less experience)
- But they are **uncorrelated with ability** (by assumption)

$\implies$ The within-group means of time-varying exogenous variables are **valid instruments** for education.

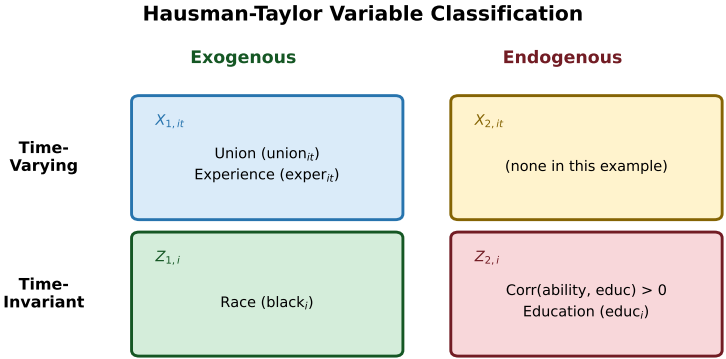
Variable Classification: Four Categories

HT requires classifying every variable into one of four groups:



Variable Classification: Four Categories

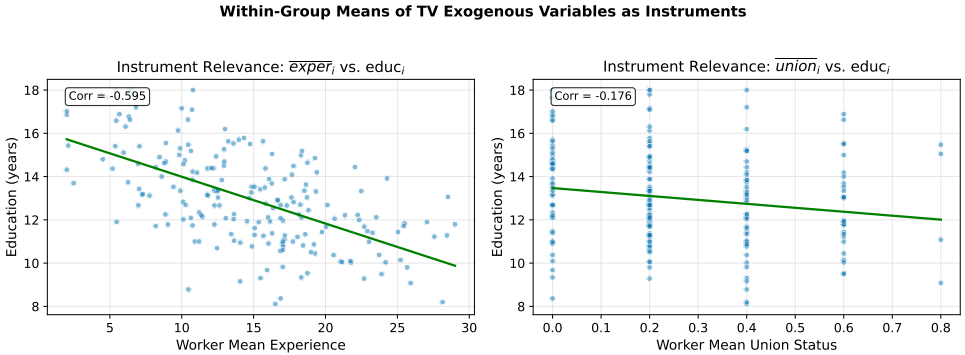
HT requires classifying every variable into one of four groups:



The classification determines which variables need instruments and which variables *provide* instruments.

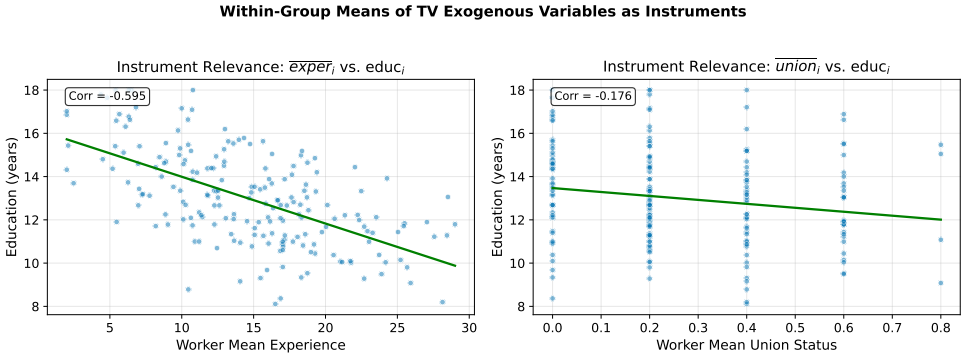
The Instruments: Within-Group Means

Why do $\overline{\text{exper}}_i$ and $\overline{\text{union}}_i$ work as instruments for educ_i?



The Instruments: Within-Group Means

Why do $\overline{\text{exper}}_i$ and $\overline{\text{union}}_i$ work as instruments for educ_i?



Relevance: Workers with more education have less experience (delayed entry) and lower unionization.
Validity: These means are uncorrelated with unobserved ability (exogenous by classification).

Instrument Requirements

Recall: X_1 = TV exogenous (exper, union); Z_2 = TI endogenous (educ).

For $\bar{X}_{1,i}$ (worker-level means of TV exogenous vars) to be valid instruments for $Z_{2,i}$:

Instrument Requirements

Recall: X_1 = TV exogenous (exper, union); Z_2 = TI endogenous (educ).

For $\bar{X}_{1,i}$ (worker-level means of TV exogenous vars) to be valid instruments for $Z_{2,i}$:

1. Relevance:

$$\text{Corr}(\bar{X}_{1,i},\ Z_{2,i}) \neq 0$$

In our data: $\text{Corr}(\bar{\text{exper}}_i, \text{educ}_i) \approx -0.60$. Workers who went to school longer started working later.

Instrument Requirements

Recall: X_1 = TV exogenous (exper, union); Z_2 = TI endogenous (educ).

For $\bar{X}_{1,i}$ (worker-level means of TV exogenous vars) to be valid instruments for $Z_{2,i}$:

1. Relevance:

$$\text{Corr}(\bar{X}_{1,i},\ Z_{2,i}) \neq 0$$

In our data: $\text{Corr}(\text{ex}\bar{\text{p}}\text{er}_i, \text{educ}_i) \approx -0.60$. Workers who went to school longer started working later.

2. Validity (exclusion restriction):

$$\text{Cov}(\bar{X}_{1,i},\ \alpha_i) = 0$$

This is an **assumption**, not something we can verify directly. It follows from classifying $X_{1,it}$ as exogenous: if $\text{Cov}(\alpha_i, \text{exper}_{it}) = 0$ for all t, then $\text{Cov}(\alpha_i, \text{ex}\bar{\text{p}}\text{er}_i) = 0$.

Instrument Requirements

Recall: X_1 = TV exogenous (exper, union); Z_2 = TI endogenous (educ).

For $\bar{X}_{1,i}$ (worker-level means of TV exogenous vars) to be valid instruments for $Z_{2,i}$:

1. Relevance:

$$\text{Corr}(\bar{X}_{1,i},\ Z_{2,i}) \neq 0$$

In our data: $\text{Corr}(\bar{\text{exper}}_i, \text{educ}_i) \approx -0.60$. Workers who went to school longer started working later.

2. Validity (exclusion restriction):

$$\text{Cov}(\bar{X}_{1,i},\ \alpha_i) = 0$$

This is an **assumption**, not something we can verify directly. It follows from classifying $X_{1,it}$ as exogenous: if $\text{Cov}(\alpha_i, \text{exper}_{it}) = 0$ for all t, then $\text{Cov}(\alpha_i, \bar{\text{exper}}_i) = 0$.

When might this fail? If high-ability workers accumulate more experience through promotions or job mobility, then $\text{Cov}(\alpha_i, \text{exper}_{it}) \neq 0$ and the instruments are invalid. The maintained assumption in this example rules that out: experience here reflects labor market entry timing, not ability-driven career advancement.

Instrument Requirements

Recall: X_1 = TV exogenous (exper, union); Z_2 = TI endogenous (educ).

For $\bar{X}_{1,i}$ (worker-level means of TV exogenous vars) to be valid instruments for $Z_{2,i}$:

1. Relevance:

$$\text{Corr}(\bar{X}_{1,i},\ Z_{2,i}) \neq 0$$

In our data: $\text{Corr}(\bar{\text{exper}}_i, \text{educ}_i) \approx -0.60$. Workers who went to school longer started working later.

2. Validity (exclusion restriction):

$$\text{Cov}(\bar{X}_{1,i},\ \alpha_i) = 0$$

This is an **assumption**, not something we can verify directly. It follows from classifying $X_{1,it}$ as exogenous: if $\text{Cov}(\alpha_i, \text{exper}_{it}) = 0$ for all t, then $\text{Cov}(\alpha_i, \bar{\text{exper}}_i) = 0$.

When might this fail? If high-ability workers accumulate more experience through promotions or job mobility, then $\text{Cov}(\alpha_i, \text{exper}_{it}) \neq 0$ and the instruments are invalid. The maintained assumption in this example rules that out: experience here reflects labor market entry timing, not ability-driven career advancement.

$\Longrightarrow$ Same logic as any IV estimation, but the instruments come from **within the panel**.

HT Procedure: Step 1 (Within Estimation)

We need consistent time-varying coefficients so we can strip them out and isolate the between-individual variation.

HT Procedure: Step 1 (Within Estimation)

We need consistent time-varying coefficients so we can strip them out and isolate the between-individual variation.

Run FE on the full model. FE wipes out both α_i (good) and the time-invariant regressors (the cost):

$$\ddot{y}_{it} = \hat{\beta}^{FE}_{\text{exper}} \ddot{\text{exper}}_{it} + \hat{\beta}^{FE}_{\text{union}} \ddot{\text{union}}_{it} + \ddot{\varepsilon}_{it}$$

where $\ddot{x}_{it} = x_{it} - \bar{x}_i$ denotes the demeaned variable.

HT Procedure: Step 1 (Within Estimation)

We need consistent time-varying coefficients so we can strip them out and isolate the between-individual variation.

Run FE on the full model. FE wipes out both α_i (good) and the time-invariant regressors (the cost):

$$\ddot{y}_{it} = \hat{\beta}^{FE}_{\text{exper}} \ddot{\text{exper}}_{it} + \hat{\beta}^{FE}_{\text{union}} \ddot{\text{union}}_{it} + \ddot{\varepsilon}_{it}$$

where $\ddot{x}_{it} = x_{it} - \bar{x}_i$ denotes the demeaned variable.

$\implies$ We now have consistent $\hat{\beta}^{FE}_{\text{exper}}$ and $\hat{\beta}^{FE}_{\text{union}}$. We still need $\hat{\beta}_{\text{educ}}$ and $\hat{\beta}_{\text{black}}$.

HT Procedure: Step 2a (Between Equation)

Now we have a cross-sectional equation where education is still present but ability is in the error.

HT Procedure: Step 2a (Between Equation)

Now we have a cross-sectional equation where education is still present but ability is in the error.

Take worker-level averages and subtract the FE-estimated time-varying effects:

$$\underbrace{\bar{y}_i - \hat{\beta}^{FE}_{\text{exper}} \bar{\text{exper}}_i - \hat{\beta}^{FE}_{\text{union}} \bar{\text{union}}_i}_{d_i} = \beta_0 + \beta_{\text{educ}}\, \text{educ}_i + \beta_{\text{black}}\, \text{black}_i + \alpha_i + \bar{\varepsilon}_i$$

HT Procedure: Step 2a (Between Equation)

Now we have a cross-sectional equation where education is still present but ability is in the error.

Take worker-level averages and subtract the FE-estimated time-varying effects:

$$\underbrace{\bar{y}_i - \hat{\beta}^{FE}_{\text{exper}}\bar{\text{exper}}_i - \hat{\beta}^{FE}_{\text{union}}\bar{\text{union}}_i}_{d_i} = \beta_0 + \beta_{\text{educ}}\,\text{educ}_i + \beta_{\text{black}}\,\text{black}_i + \alpha_i + \bar{\varepsilon}_i$$

This is a **cross-sectional equation** with one observation per worker. Education is back, but the problem remains: α_i is in the error and $\text{Cov}(\alpha_i, \text{educ}_i) \neq 0$.

HT Procedure: Step 2a (Between Equation)

Now we have a cross-sectional equation where education is still present but ability is in the error.

Take worker-level averages and subtract the FE-estimated time-varying effects:

$$\underbrace{\bar{y}_i - \hat{\beta}^{FE}_{\text{exper}} \bar{\text{exper}}_i - \hat{\beta}^{FE}_{\text{union}} \bar{\text{union}}_i}_{d_i} = \beta_0 + \beta_{\text{educ}}\, \text{educ}_i + \beta_{\text{black}}\, \text{black}_i + \alpha_i + \bar{\varepsilon}_i$$

This is a **cross-sectional equation** with one observation per worker. Education is back, but the problem remains: α_i is in the error and $\text{Cov}(\alpha_i, \text{educ}_i) \neq 0$.

$\implies$ OLS on this equation would still be biased. We need instruments.

HT Procedure: Step 2b (IV Estimation)

This is a standard IV problem: endogenous regressor (educ_i), error contains α_i.

HT Procedure: Step 2b (IV Estimation)

This is a standard IV problem: endogenous regressor (educ_i), error contains α_i.

Run 2SLS on the between equation, using $\overline{\text{exper}}_i$ and $\overline{\text{union}}_i$ as instruments for educ_i.

HT Procedure: Step 2b (IV Estimation)

This is a standard IV problem: endogenous regressor (educ_i), error contains α_i.

Run 2SLS on the between equation, using exper_i and union_i as instruments for educ_i.

Why does this work?

- **Relevance:** exper_i and union_i are correlated with educ_i (through labor market timing)
- **Validity:** They are uncorrelated with α_i (by the TV exogeneity assumption)

HT Procedure: Step 2b (IV Estimation)

This is a standard IV problem: endogenous regressor (educ_i), error contains α_i.

Run 2SLS on the between equation, using $\bar{\text{exper}}_i$ and $\bar{\text{union}}_i$ as instruments for educ_i.

Why does this work?

- **Relevance:** $\bar{\text{exper}}_i$ and $\bar{\text{union}}_i$ are correlated with educ_i (through labor market timing)
- **Validity:** They are uncorrelated with α_i (by the TV exogeneity assumption)

$\Longrightarrow$ 2SLS isolates the variation in education that is driven by experience and union patterns, not by ability.

HT Procedure: Step 3 (FGLS Combination)

Combining within and between estimates optimally gives efficiency, just as RE does when its assumptions hold.

HT Procedure: Step 3 (FGLS Combination)

Combining within and between estimates optimally gives efficiency, just as RE does when its assumptions hold.

The full HT estimator uses feasible GLS to combine:

1. **Within information** (from FE): identifies time-varying coefficients
2. **Between information** (from IV): identifies time-invariant coefficients

HT Procedure: Step 3 (FGLS Combination)

Combining within and between estimates optimally gives efficiency, just as RE does when its assumptions hold.

The full HT estimator uses feasible GLS to combine:

1. **Within information** (from FE): identifies time-varying coefficients
2. **Between information** (from IV): identifies time-invariant coefficients

FGLS uses the variance components ($\hat{\sigma}^2_\alpha$, $\hat{\sigma}^2_\varepsilon$) to weight observations: when σ^2_α is large relative to σ^2_ε, the weighting leans more heavily on the within variation (similar to FE); when it is small, the weighting uses more of the between variation.

HT Procedure: Step 3 (FGLS Combination)

Combining within and between estimates optimally gives efficiency, just as RE does when its assumptions hold.

The full HT estimator uses feasible GLS to combine:

1. **Within information** (from FE): identifies time-varying coefficients
2. **Between information** (from IV): identifies time-invariant coefficients

FGLS uses the variance components ($\hat{\sigma}_\alpha^2$, $\hat{\sigma}_\varepsilon^2$) to weight observations: when σ_α^2 is large relative to σ_ε^2, the weighting leans more heavily on the within variation (similar to FE); when it is small, the weighting uses more of the between variation.

In practice, software handles this. In Stata:

```
xthtaylor lwage exper union educ black, endog(educ) constant(educ black)
```

HT Procedure: Step 3 (FGLS Combination)

Combining within and between estimates optimally gives efficiency, just as RE does when its assumptions hold.

The full HT estimator uses feasible GLS to combine:

1. **Within information** (from FE): identifies time-varying coefficients
2. **Between information** (from IV): identifies time-invariant coefficients

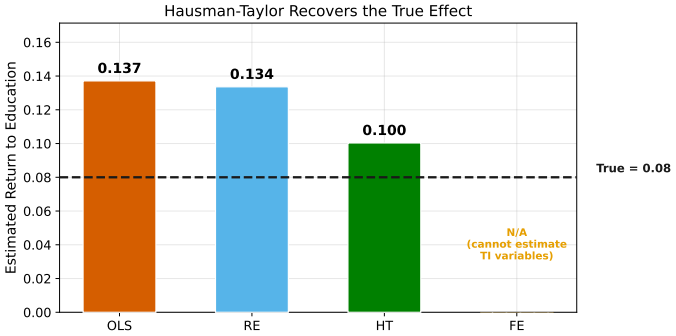
FGLS uses the variance components ($\hat{\sigma}_\alpha^2$, $\hat{\sigma}_\varepsilon^2$) to weight observations: when σ_α^2 is large relative to σ_ε^2, the weighting leans more heavily on the within variation (similar to FE); when it is small, the weighting uses more of the between variation.

In practice, software handles this. In Stata:

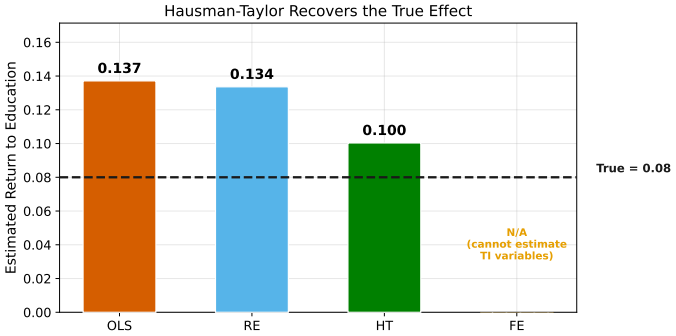
```
xthtaylor lwage exper union educ black, endog(educ) constant(educ black)
```

$\Longrightarrow$ You specify which variables are time-invariant and which are endogenous. The estimator does the rest.

The Result: HT Recovers the True Effect

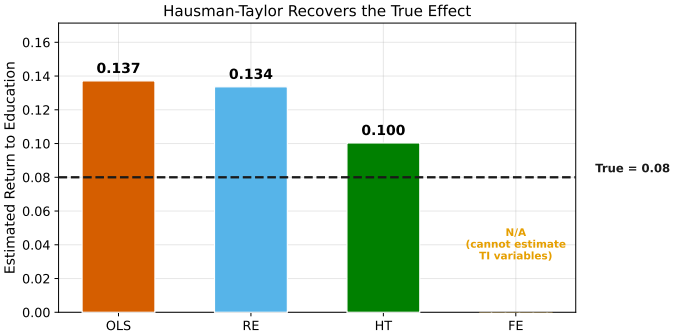


The Result: HT Recovers the True Effect



OLS (≈ 0.14) and RE (≈ 0.13) are both biased upward. HT estimates ≈ 0.10, much closer to the true $\beta_{\text{educ}} = 0.08$.

The Result: HT Recovers the True Effect



OLS (≈ 0.14) and RE (≈ 0.13) are both biased upward. HT estimates ≈ 0.10, much closer to the true $\beta_{\text{educ}} = 0.08$.

$\Longrightarrow$ By instrumenting education with within-group means of experience and union status, HT removes most of the ability bias.

Note: the HT estimate shown is from the IV-between step (Step 2b). The full FGLS estimator (Step 3) produces a slightly different number; Stata's `xthtaylor` implements the complete procedure.

How Each HT Component Solves a Problem

Problem	**HT Component**	**Result**
TV coefficients biased	FE (within estimation)	Consistent $\hat{\beta}_{\text{exper}}$, $\hat{\beta}_{\text{union}}$
TI endogenous variable	IV using $\bar{X}_{1,i}$ as instruments	Consistent $\hat{\beta}_{\text{educ}}$
Combine efficiently	GLS weighting	Efficient estimates

How Each HT Component Solves a Problem

Problem	**HT Component**	**Result**
TV coefficients biased	FE (within estimation)	Consistent $\hat{\beta}_{\text{exper}}$, $\hat{\beta}_{\text{union}}$
TI endogenous variable	IV using $\bar{X}_{1,i}$ as instruments	Consistent $\hat{\beta}_{\text{educ}}$
Combine efficiently	GLS weighting	Efficient estimates

Each step addresses a specific failure of the standard estimators:

- FE handles TV coefficients (consistent) ✓
- IV handles TI endogenous variables (via internal instruments) ✓
- GLS combines them efficiently ✓

How Each HT Component Solves a Problem

Problem	**HT Component**	**Result**
TV coefficients biased	FE (within estimation)	Consistent $\hat{\beta}_{\text{exper}}$, $\hat{\beta}_{\text{union}}$
TI endogenous variable	IV using $\bar{X}_{1,i}$ as instruments	Consistent $\hat{\beta}_{\text{educ}}$
Combine efficiently	GLS weighting	Efficient estimates

Each step addresses a specific failure of the standard estimators:

- FE handles TV coefficients (consistent) ✓
- IV handles TI endogenous variables (via internal instruments) ✓
- GLS combines them efficiently ✓

$\Longrightarrow$ HT fills the gap between FE (drops TI variables) and RE (requires full exogeneity).

Outline

1 The Problem: Returns to Education

2 The Hausman-Taylor Estimator

3 When HT Works (and When It Doesn't)

4 Decision Framework

5 Summary

When Does HT Work?

HT requires three conditions:

When Does HT Work?

HT requires three conditions:

1. **Correct classification.** You must correctly identify which variables are exogenous and which are endogenous. If a "time-varying exogenous" variable is actually endogenous, the instruments are invalid.

When Does HT Work?

HT requires three conditions:

1. **Correct classification.** You must correctly identify which variables are exogenous and which are endogenous. If a "time-varying exogenous" variable is actually endogenous, the instruments are invalid.

2. **Instrument relevance.** The within-group means of TV exogenous variables must be correlated with the TI endogenous variables. If $\text{Corr}(\bar{X}_{1,i}, Z_{2,i}) \approx 0$, the instruments are weak and HT will perform poorly.

When Does HT Work?

HT requires three conditions:

1. **Correct classification.** You must correctly identify which variables are exogenous and which are endogenous. If a "time-varying exogenous" variable is actually endogenous, the instruments are invalid.

2. **Instrument relevance.** The within-group means of TV exogenous variables must be correlated with the TI endogenous variables. If $\text{Corr}(\bar{X}_{1,i}, Z_{2,i}) \approx 0$, the instruments are weak and HT will perform poorly.

3. **Enough instruments.** You need at least as many TV exogenous variables as TI endogenous variables (order condition). More instruments enable overidentification testing.

When Does HT Work?

HT requires three conditions:

1. **Correct classification.** You must correctly identify which variables are exogenous and which are endogenous. If a "time-varying exogenous" variable is actually endogenous, the instruments are invalid.

2. **Instrument relevance.** The within-group means of TV exogenous variables must be correlated with the TI endogenous variables. If $\text{Corr}(\bar{X}_{1,i}, Z_{2,i}) \approx 0$, the instruments are weak and HT will perform poorly.

3. **Enough instruments.** You need at least as many TV exogenous variables as TI endogenous variables (order condition). More instruments enable overidentification testing.

In our example: 2 TV exogenous (exper, union) instrumenting 1 TI endogenous (educ) $\implies$ overidentified (good: we can test validity).

Formal Assumptions

Partition the variables:

- $X_{1,it}$: time-varying, exogenous (exper, union)
- $X_{2,it}$: time-varying, endogenous (none here)
- $Z_{1,i}$: time-invariant, exogenous (black)
- $Z_{2,i}$: time-invariant, endogenous (educ)

Formal Assumptions

Partition the variables:

- $X_{1,it}$: time-varying, exogenous (exper, union)
- $X_{2,it}$: time-varying, endogenous (none here)
- $Z_{1,i}$: time-invariant, exogenous (black)
- $Z_{2,i}$: time-invariant, endogenous (educ)

The Hausman-Taylor assumptions:

1. $E[\varepsilon_{it} \mid X_1, X_2, Z_1, Z_2, \alpha_i] = 0$ (strict exogeneity of idiosyncratic error)

   The idiosyncratic error is pure noise.

2. $E[\alpha_i \mid X_1, Z_1] = 0$ (α_i uncorrelated with exogenous variables)

   Ability is unrelated to experience, union status, and race.

3. $E[\alpha_i \mid X_2, Z_2] \neq 0$ (endogeneity of X_2, Z_2)

   Ability IS related to education. This is why we need HT.

Formal Assumptions

Partition the variables:

- $X_{1,it}$: time-varying, exogenous (exper, union)
- $X_{2,it}$: time-varying, endogenous (none here)
- $Z_{1,i}$: time-invariant, exogenous (black)
- $Z_{2,i}$: time-invariant, endogenous (educ)

The Hausman-Taylor assumptions:

1. $E[\varepsilon_{it} \mid X_1, X_2, Z_1, Z_2, \alpha_i] = 0$ (strict exogeneity of idiosyncratic error)

   The idiosyncratic error is pure noise.

2. $E[\alpha_i \mid X_1, Z_1] = 0$ (α_i uncorrelated with exogenous variables)

   Ability is unrelated to experience, union status, and race.

3. $E[\alpha_i \mid X_2, Z_2] \neq 0$ (endogeneity of X_2, Z_2)

   Ability IS related to education. This is why we need HT.

$\Longrightarrow$ Assumption 2 is what makes $\bar{X}_{1,i}$ a valid instrument: it inherits exogeneity from $X_{1,it}$.

Testing HT: Overidentification

When you have more TV exogenous variables than TI endogenous variables, the model is **overidentified**.

Testing HT: Overidentification

When you have more TV exogenous variables than TI endogenous variables, the model is **overidentified**.

This allows a **Sargan/Hansen test** of instrument validity:

- H_0: instruments are valid (uncorrelated with α_i)
- H_1: at least one instrument is invalid

Testing HT: Overidentification

When you have more TV exogenous variables than TI endogenous variables, the model is **overidentified**.

This allows a **Sargan/Hansen test** of instrument validity:

- H_0: instruments are valid (uncorrelated with α_i)
- H_1: at least one instrument is invalid

In our example:

- 2 instruments ($\text{ex}\bar{\text{p}}\text{er}_i$, $\text{un}\bar{\text{i}}\text{on}_i$) for 1 endogenous variable (educ_i)
- Degrees of overidentification $= 2 - 1 = 1$
- Test statistic $\sim \chi^2(1)$ under H_0

Testing HT: Overidentification

When you have more TV exogenous variables than TI endogenous variables, the model is **overidentified**.

This allows a **Sargan/Hansen test** of instrument validity:

- H_0: instruments are valid (uncorrelated with α_i)
- H_1: at least one instrument is invalid

In our example:

- 2 instruments ($\bar{\text{exper}}_i$, $\bar{\text{union}}_i$) for 1 endogenous variable (educ_i)
- Degrees of overidentification $= 2 - 1 = 1$
- Test statistic $\sim \chi^2(1)$ under H_0

$\implies$ Rejection suggests that at least one of experience or union is not truly exogenous to ability. This would invalidate the HT approach.

Hausman Test: HT vs. FE

For **time-varying** coefficients, both FE and HT are consistent (under their respective assumptions). We can compare them:

Hausman Test: HT vs. FE

For **time-varying** coefficients, both FE and HT are consistent (under their respective assumptions). We can compare them:

$$H_0 : \text{Cov}(\alpha_i, X_{1,it}) = 0 \quad \text{(necessary for HT consistency)}$$

Hausman Test: HT vs. FE

For **time-varying** coefficients, both FE and HT are consistent (under their respective assumptions). We can compare them:

$$H_0 : \text{Cov}(\alpha_i, X_{1,it}) = 0 \quad \text{(necessary for HT consistency)}$$

The Hausman test compares $\hat{\beta}^{FE}_{\text{TV}}$ and $\hat{\beta}^{HT}_{\text{TV}}$:

- If they are "close," the HT assumptions are not rejected
- If they diverge, the classification of exogenous variables may be wrong

Hausman Test: HT vs. FE

For **time-varying** coefficients, both FE and HT are consistent (under their respective assumptions). We can compare them:

$$H_0 : \text{Cov}(\alpha_i, X_{1,it}) = 0 \quad \text{(necessary for HT consistency)}$$

The Hausman test compares $\hat{\beta}^{FE}_{\text{TV}}$ and $\hat{\beta}^{HT}_{\text{TV}}$:

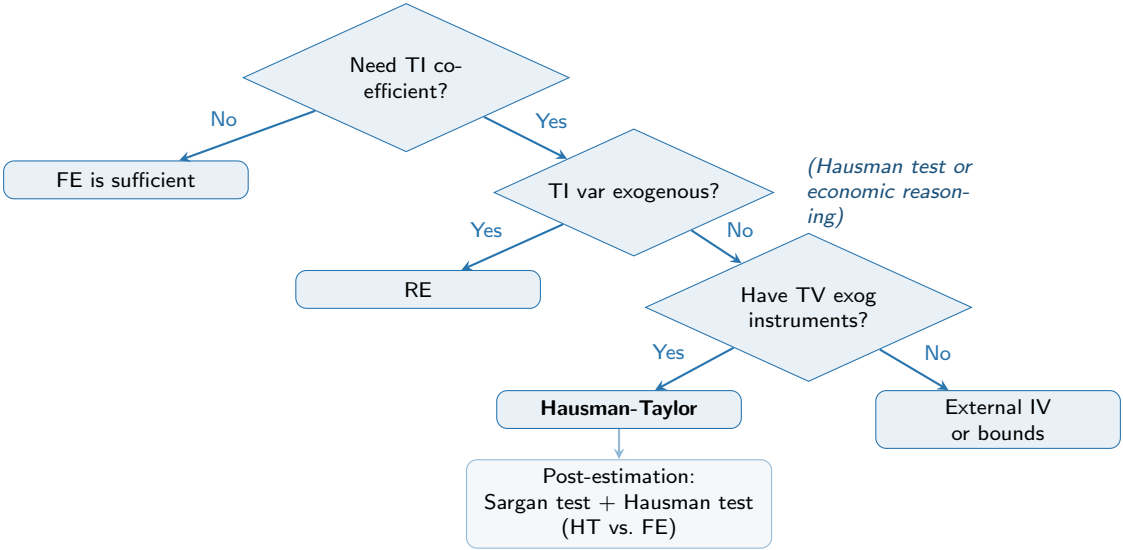
- If they are "close," the HT assumptions are not rejected
- If they diverge, the classification of exogenous variables may be wrong

$\Longrightarrow$ This tests whether the TV variables you classified as exogenous truly are. It does *not* test the validity of the TI classification.

Outline

1. The Problem: Returns to Education
2. The Hausman-Taylor Estimator
3. When HT Works (and When It Doesn't)
4. Decision Framework
5. Summary

Decision Flowchart



Outline

1. The Problem: Returns to Education
2. The Hausman-Taylor Estimator
3. When HT Works (and When It Doesn't)
4. Decision Framework
5. Summary

Summary

We started with a panel of 200 workers and wanted to estimate the return to education. OLS and RE both overestimated it because unobserved ability is correlated with education. FE eliminates ability but also eliminates education.

Summary

We started with a panel of 200 workers and wanted to estimate the return to education. OLS and RE both overestimated it because unobserved ability is correlated with education. FE eliminates ability but also eliminates education.

1. **FE** removes time-invariant unobservables but **cannot estimate** coefficients on time-invariant regressors.

Summary

We started with a panel of 200 workers and wanted to estimate the return to education. OLS and RE both overestimated it because unobserved ability is correlated with education. FE eliminates ability but also eliminates education.

1. **FE** removes time-invariant unobservables but **cannot estimate** coefficients on time-invariant regressors.
2. **RE** can estimate time-invariant coefficients but requires all regressors to be uncorrelated with α_i. When this fails, RE is biased.

Summary

We started with a panel of 200 workers and wanted to estimate the return to education. OLS and RE both overestimated it because unobserved ability is correlated with education. FE eliminates ability but also eliminates education.

1. **FE** removes time-invariant unobservables but **cannot estimate** coefficients on time-invariant regressors.
2. **RE** can estimate time-invariant coefficients but requires all regressors to be uncorrelated with α_i. When this fails, RE is biased.
3. **Hausman-Taylor** uses within-group means of TV exogenous variables as instruments for TI endogenous variables. It fills the gap between FE and RE.

Summary

We started with a panel of 200 workers and wanted to estimate the return to education. OLS and RE both overestimated it because unobserved ability is correlated with education. FE eliminates ability but also eliminates education.

1. **FE** removes time-invariant unobservables but **cannot estimate** coefficients on time-invariant regressors.
2. **RE** can estimate time-invariant coefficients but requires all regressors to be uncorrelated with α_i. When this fails, RE is biased.
3. **Hausman-Taylor** uses within-group means of TV exogenous variables as instruments for TI endogenous variables. It fills the gap between FE and RE.
4. HT requires **correct variable classification**, **relevant instruments**, and at least as many TV exogenous vars as TI endogenous vars.

Summary

We started with a panel of 200 workers and wanted to estimate the return to education. OLS and RE both overestimated it because unobserved ability is correlated with education. FE eliminates ability but also eliminates education.

1. **FE** removes time-invariant unobservables but **cannot estimate** coefficients on time-invariant regressors.
2. **RE** can estimate time-invariant coefficients but requires all regressors to be uncorrelated with α_i. When this fails, RE is biased.
3. **Hausman-Taylor** uses within-group means of TV exogenous variables as instruments for TI endogenous variables. It fills the gap between FE and RE.
4. HT requires **correct variable classification**, **relevant instruments**, and at least as many TV exogenous vars as TI endogenous vars.
5. Use the **Sargan/Hansen test** (overidentification) and the **Hausman test** (HT vs. FE) to check the assumptions.

Summary

We started with a panel of 200 workers and wanted to estimate the return to education. OLS and RE both overestimated it because unobserved ability is correlated with education. FE eliminates ability but also eliminates education.

1. **FE** removes time-invariant unobservables but **cannot estimate** coefficients on time-invariant regressors.
2. **RE** can estimate time-invariant coefficients but requires all regressors to be uncorrelated with α_i. When this fails, RE is biased.
3. **Hausman-Taylor** uses within-group means of TV exogenous variables as instruments for TI endogenous variables. It fills the gap between FE and RE.
4. HT requires **correct variable classification**, **relevant instruments**, and at least as many TV exogenous vars as TI endogenous vars.
5. Use the **Sargan/Hansen test** (overidentification) and the **Hausman test** (HT vs. FE) to check the assumptions.

$\Longrightarrow$ When you need a time-invariant coefficient but suspect endogeneity, check whether your TV exogenous variables can serve as internal instruments.

Thank you!

jakeanderson@g.ucla.edu