

# **OLS: Estimation, Properties, and Inference**

The sample analog of the BLP — and what we can learn from it

Gov 2001 · Scott Cunningham · Spring 2026

# End-of-semester announcements

## Logistics

- **Kaixiao:** unavailable for remainder of semester
- **Section:** canceled for rest of semester
- **Problem Set 4:** canceled — not required
- **April 28 (last class):** variance weights (Angrist), Goodman-Bacon, Callaway & Sant'Anna
- **Review guide:** distributed before May 7
- **Grading:** exams returned as quickly as possible

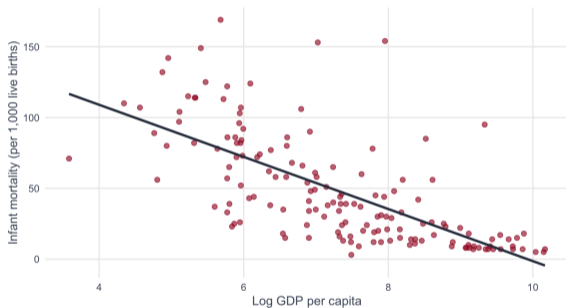
**Final exam:** Wednesday May 7 · 2:00 pm

## Coverage (100 pts):

- **Probs 1–2** (35 pts): inference & estimation — distributions, delta method, CIs, hypothesis testing
- **Prob 3** (20 pts): T/F spanning full course
- **Probs 4–5** (45 pts): regression — OLS, FWL, Gauss–Markov, variance weights

Regression-heavy, but inference is load-bearing throughout

# The question: why do some nations have ten times the infant mortality of others?



## UN data, 154 countries (late 1990s)

- Range: 3–169 deaths per 1,000 live births
- Some countries 50× higher than others
- Why?

## Model:

$$IM_i = \beta_0 + \beta_1 \ln(\text{GDP}_i) + \beta_2 TFR_i + \beta_3 Illit_i + \varepsilon_i$$

$IM$  = infant mortality,  $TFR$  = fertility,  $Illit$  = female illiteracy

## The complete OLS output – coefficients are fixed; only standard errors change

UN98: infant mortality on GDP, fertility, illiteracy ( $n = 154$ )

	$\hat{\beta}$	Naive SE	Naive $t$	Robust SE	Robust $t$
Intercept	53.34	11.29	4.72	13.49	3.95
ln(GDP)	-7.17	1.21	-5.95	1.39	-5.16
Fertility	8.83	1.39	6.37	1.56	5.66
Illiteracy	0.52	0.09	5.78	0.10	5.20

$R^2 = 0.803$     $F = 203.2$    BP:  $\chi^2(3) = 38.3$ ,  
 $p < 0.001$

$\hat{\beta}$  is **identical** in both columns.  
Only SE – and therefore  $t$  – changes.

Heteroskedasticity does not move the point estimates. It corrupts our uncertainty about them.

## Last week: derivation and optimality – today: inference

### Wednesday Apr 15 (done):

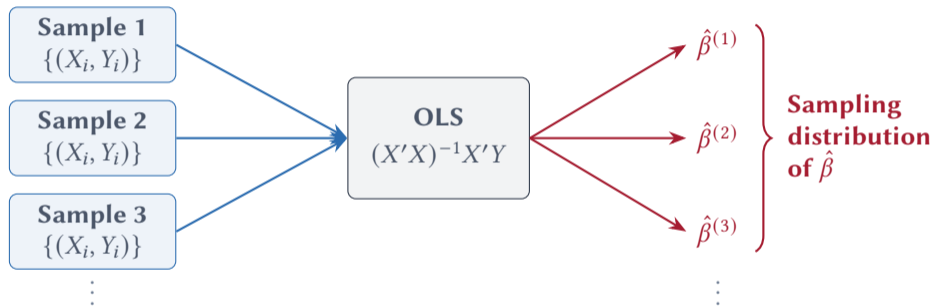
- $\hat{\beta} = (X'X)^{-1}X'Y$  from FOC
- $R^2$ , TSS, RSS, adjusted  $R^2$
- Gauss-Markov: BLUE under homoskedasticity

### Today:

- Sampling distribution of  $\hat{\beta}$
- Unbiasedness, consistency, asymptotic normality
- SE formulas: classical, robust, clustered
- $t$ -tests,  $F$ -tests, size distortion, coverage

The question today:  $\hat{\beta}$  is a number from our data – how uncertain should we be about it?

## OLS is a random variable – it has a sampling distribution



Each dataset gives a different  $\hat{\beta}$  – the distribution of  $\hat{\beta}$  across datasets is the sampling distribution.

# Becker, Mincer, and a century-old question: does schooling *cause* earnings?

## The question:

- **Human capital (Becker 1964):** skills are capital – invest in schooling now (pay tuition, forgo wages)  $\Rightarrow$  higher productivity  $\Rightarrow$  higher wages later
- **Mincer (1974):**  
 $\ln Y = a + b_1S + b_2E + \dots$  – measures the return
- **Cause or correlate?**  $D$  raises  $Y$  (human capital) vs.  $D$  signals ability (Spence 1973)
- **Policy:** subsidize college only if  $\beta_{causal} > 0$

## Long (causal) model:

$$Y_i = \beta_0 + \beta_{causal} D_i + \gamma A_i + u_i$$

$D_i$  = college,  $A_i$  = ability (unobserved)

## Short (OLS) model – ability omitted:

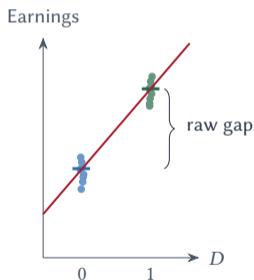
$$Y_i = \beta_0 + \beta_{BLP} D_i + \varepsilon_i, \quad \varepsilon_i = \gamma A_i + u_i$$

$$\beta_{BLP} = \beta_{causal} + \frac{\text{Cov}(D, \varepsilon)}{\text{Var}(D)}$$

# College raises earnings – but OLS gives the raw association, not the causal effect

## Observational data:

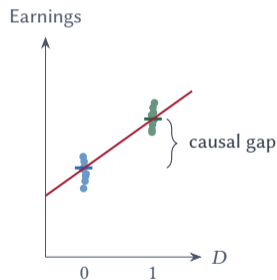
$\mathbb{E}[\varepsilon | D=1] > \mathbb{E}[\varepsilon | D=0] \Rightarrow$  slope too steep



college-goers have higher ability – raw gap  $\neq$  causal effect

## If college were randomized:

Randomization:  $\mathbb{E}[\varepsilon | D=1] = \mathbb{E}[\varepsilon | D=0] \Rightarrow$  no bias



ability balanced across  $D$  – raw gap = causal effect

## Under iid sampling alone, OLS always converges to the population BLP

**Population BLP** (always defined):

$$\beta_{BLP} = (\mathbb{E}[X_i X_i'])^{-1} \mathbb{E}[X_i Y_i] = \frac{\text{Cov}(D, \text{Earnings})}{\text{Var}(D)}$$

**OLS converges to it by LLN:**

$$\hat{\beta} = \left( \frac{1}{n} \sum_i X_i X_i' \right)^{-1} \frac{1}{n} \sum_i X_i Y_i \xrightarrow{p} \beta_{BLP}$$

$\hat{\beta} \xrightarrow{p} \beta_{BLP}$  under iid sampling alone — OLS always hits the population regression you would run with infinite data

In the college example, the raw earnings gap IS  $\beta_{BLP}$  — OLS hits it exactly; the question is whether that's the estimand you wanted.

## For binary $D$ , the OVB reduces to a difference in conditional means

**Any  $D$ :**  $\text{Cov}(D, \varepsilon)/\text{Var}(D)$

Projection of  $\varepsilon$  onto  $D$

= 0 if  $D \perp \varepsilon$     $\neq 0$  if correlated

**Binary  $D$ :**  $\text{Var}(D) = p(1-p)$

$$\frac{\text{Cov}(D, \varepsilon)}{\text{Var}(D)} = \mathbb{E}[\varepsilon \mid D=1] - \mathbb{E}[\varepsilon \mid D=0]$$

$p(1-p)$  cancels from numerator and denominator

Selection bias =  $\mathbb{E}[\varepsilon \mid D=1] - \mathbb{E}[\varepsilon \mid D=0]$  is OVB in binary- $D$  form

# Plug $D = 1$ and $D = 0$ into the regression equation – the selection bias falls out

**Model:**  $Y_i = \beta_0 + \beta_1 D_i + \varepsilon_i$

$\varepsilon_i$  = ability, family background, motivation

**Plug in  $D_i = 1$ :**

$$\mathbb{E}[Y_i | D_i = 1] = \beta_0 + \beta_1 + \mathbb{E}[\varepsilon_i | D_i = 1]$$

**Plug in  $D_i = 0$ :**

$$\mathbb{E}[Y_i | D_i = 0] = \beta_0 + \mathbb{E}[\varepsilon_i | D_i = 0]$$

**Subtract:**

$$\underbrace{\mathbb{E}[Y | D=1] - \mathbb{E}[Y | D=0]}_{\hat{\beta}_{OLS} \xrightarrow{p} \text{this}} = \underbrace{\beta_1}_{\text{causal}} + \underbrace{\mathbb{E}[\varepsilon | D=1] - \mathbb{E}[\varepsilon | D=0]}_{\text{selection bias}}$$

$$\beta_{BLP} = \beta_1 + (\mathbb{E}[\varepsilon | D = 1] - \mathbb{E}[\varepsilon | D = 0]) \quad \text{– OLS recovers the full gap, causal effect and selection combined}$$

**Why policymakers need  $\beta_1$ :**

- Can set tuition, grants, access
- **Cannot** change ability, family background, motivation

$\varepsilon_i$  is beyond policy's reach.

$\beta_1$  is not.

# Mean independence makes the BLP a causal parameter – without it, OLS is right about the wrong thing

Decompose  $\beta_{BLP}$  in the college example:

$$\beta_{BLP} = \underbrace{\beta_{causal}}_{\text{effect of college}} + \underbrace{(\mathbb{E}[\varepsilon | D = 1] - \mathbb{E}[\varepsilon | D = 0])}_{\text{selection bias: college-goers have higher ability}}$$

$\mathbb{E}[\varepsilon | D] \neq 0$ :

$\beta_{BLP} \neq \beta_{causal}$

- OLS correctly hits  $\beta_{BLP}$
- But  $\beta_{BLP}$  includes selection
- Right answer to wrong question

$\mathbb{E}[\varepsilon | D] = 0$ :

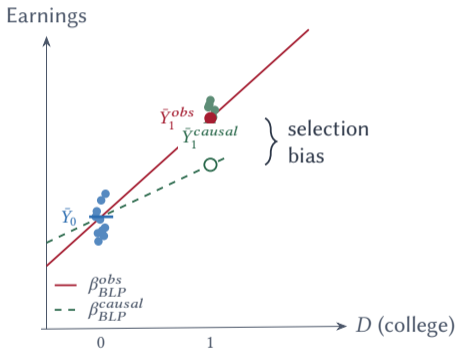
$\beta_{BLP} = \beta_{causal}$

- Selection term vanishes
- OLS hits  $\beta_{BLP}$
- Which IS the causal effect

OLS is never biased for  $\beta_{BLP}$ . Mean independence makes  $\beta_{BLP}$  worth having.

Both  $\beta_{causal}$  and the selection bias are BLP quantities – the “wrong” answer is the BLP of confounded data.

## Both lines are BLPs – mean independence aligns the observed BLP with the causal one



$$\beta_{BLP}^{obs} = \beta_{BLP}^{causal} + \text{selection bias} \quad \text{— both sides are BLP quantities}$$

## Unbiasedness vs. consistency – two different properties, two different standards

**Unbiased:**  $\mathbb{E}[\hat{\beta}] = \beta_{causal}$

**Requires:**  $\mathbb{E}[\varepsilon | X] = 0$

Finite-sample. Holds for any  $n$ .

**Consistent:**  $\hat{\beta} \xrightarrow{p} \beta_{BLP}$

**Requires:** iid sampling only

Asymptotic. Sample averages converge to population moments by LLN.

- An estimator can be **biased but consistent** – e.g., MLE of  $\sigma^2$  with  $n$  in denominator
- An estimator can be **unbiased but inconsistent** – e.g.,  $Y_1$  alone always targets  $\mu$  but never converges
- OLS is **consistent always; unbiased** additionally under  $\mathbb{E}[\varepsilon | X] = 0$

Consistency for  $\beta_{BLP}$ : guaranteed under iid alone. Unbiasedness for  $\beta_{causal}$ : requires the stronger  $\mathbb{E}[\varepsilon | X] = 0$ .

## OLS converges to $\beta_{BLP}$ under iid alone — $\mathbb{E}[X_i\varepsilon_i] = 0$ is automatic by the projection

Why OLS is consistent:

$$\hat{\beta} - \beta_{BLP} = \left( \frac{1}{n} \sum_i X_i X_i' \right)^{-1} \frac{1}{n} \sum_i X_i \varepsilon_i$$

By LLN:

$$\frac{1}{n} \sum_i X_i X_i' \xrightarrow{p} \mathbb{E}[X_i X_i'] \equiv Q$$

$$\frac{1}{n} \sum_i X_i \varepsilon_i \xrightarrow{p} \mathbb{E}[X_i \varepsilon_i] = 0$$

By Slutsky:

$$\hat{\beta} - \beta_{BLP} \xrightarrow{p} Q^{-1} \cdot 0 = 0$$

$$\therefore \hat{\beta} \xrightarrow{p} \beta_{BLP}$$

$\mathbb{E}[X_i \varepsilon_i] = 0$  holds by construction for the BLP residual — OLS always converges to  $\beta_{BLP}$  under iid

No extra assumption needed for consistency —  $\mathbb{E}[\varepsilon | X] = 0$  is needed only if you want  $\beta_{BLP} = \beta_{causal}$ .

## CLT gives normality to sample means – OLS is not a sample mean

CLT applies directly to simple averages:

$$\sqrt{n} (\bar{X} - \mu) \xrightarrow{d} N(0, \text{Var}(X))$$

OLS is not a simple average:

$$\hat{\beta} = (X'X)^{-1}X'Y$$

- A ratio of matrix sums – nonlinear in the data
- CLT does not apply *directly* to  $\hat{\beta}$
- Saying OLS is asymptotically normal is asserting something nontrivial

“Isn’t everything normal under CLT?” – only sample means are. OLS earns normality by showing  $\hat{\beta} - \beta_{BLP}$  reduces to one.

# OLS is approximately a scaled sample mean – that's why CLT applies

Decompose the estimation error:

$$\hat{\beta} - \beta_{BLP} = \underbrace{\left(\frac{X'X}{n}\right)^{-1}}_{\xrightarrow{p} Q^{-1} \text{ (LLN)}} \cdot \underbrace{\frac{X'\varepsilon}{n}}_{\text{sample mean of } X_i\varepsilon_i}$$

CLT on the numerator:

$$\sqrt{n} \cdot \frac{1}{n} \sum_i X_i\varepsilon_i \xrightarrow{d} N(0, \Sigma) \quad \Sigma = \mathbb{E}[X_i X_i' \varepsilon_i^2]$$

Slutsky handles the denominator: multiply by  $Q^{-1} \xrightarrow{p} Q^{-1}$

$$\sqrt{n} (\hat{\beta} - \beta_{BLP}) \xrightarrow{d} N(0, Q^{-1}\Sigma Q^{-1})$$

CLT applies to  $X_i\varepsilon_i$  (a sample mean), not to  $Y_i$  – LLN and Slutsky handle the rest.

## Asymptotic normality is the foundation for every piece of inference we do

What  $\sqrt{n}(\hat{\beta} - \beta_{BLP}) \xrightarrow{d} N(0, V)$  gives us:

**For large  $n$ :**

$$\hat{\beta} \approx N\left(\beta_{BLP}, \frac{V}{n}\right)$$

We know the shape of the sampling distribution — even though  $\beta_{BLP}$  is unknown.

**What we still need:**

An estimate  $\hat{V}$  of  $V$

$\hat{V}$  determines our standard errors — and getting it wrong breaks all inference.

The question “what is  $V$ ?” is exactly the homoskedasticity vs. heteroskedasticity question — coming up next.

## Gauss-Markov: OLS is BLUE under four assumptions

### The four classical assumptions:

1. **Linearity:**  $Y_i = X_i' \beta + \varepsilon_i$
2. **Full rank:**  $(X'X)$  invertible — no perfect multicollinearity
3. **Strict exogeneity:**  $\mathbb{E}[\varepsilon_i | X_1, \dots, X_n] = 0$
4. **Homoskedasticity:**  $\text{Var}(\varepsilon_i | X) = \sigma^2$  for all  $i$

**Gauss-Markov Theorem:** Under (1)–(4), OLS is **BLUE** — Best Linear Unbiased Estimator

Among all estimators of the form  $\tilde{\beta} = AY$  that satisfy  $\mathbb{E}[\tilde{\beta}] = \beta$ , OLS achieves the smallest variance for every coefficient.

## BLUE: what each word means

**B**

**Best**

Minimum  
variance

**L**

**Linear**

$\tilde{\beta} = AY$   
a linear function  
of the data

**U**

**Unbiased**

$\mathbb{E}[\tilde{\beta}] = \beta$   
centered on  
 $\beta_{BLP}$

**E**

**Estimator**

$\hat{\beta}_k$ , one  
component  
at a time

- “Best among linear unbiased” — not best among *all* estimators
- Finite-sample result: no asymptotic argument needed
- **Requires homoskedasticity (assumption 4)** — this is the critical one
- Under heteroskedasticity: OLS is still unbiased, but **no longer best**

WLS is BLUE under heteroskedasticity — but in practice we fix the SEs rather than reweight.

## What you get – and what you need – under each assumption

Assumptions held	OLS property	What breaks without it
$\mathbb{E}[\varepsilon   X] = 0$	Unbiased for $\beta_{causal}$	$\hat{\beta} \rightarrow \beta_{BLP} \neq \beta_{causal}$
+ Homoskedasticity	BLUE (Gauss-Markov)	OLS is inefficient; WLS beats it
+ Large $n$ (LLN, CLT)	Consistent + asymp. normal	No sampling distribution
Heterosk. only	Unbiased, consistent, normal	<b>Standard errors are wrong</b>

Heteroskedasticity does not bias  $\hat{\beta}$  – it corrupts our estimate of  $\text{Var}(\hat{\beta})$

This is why heteroskedasticity is fundamentally an inference problem, not an estimation problem.

## The variance of $\hat{\beta}$ is the pivot of all inference

From asymptotic normality:  $\hat{\beta}_k \approx N(\beta_{BLP,k}, \text{Var}(\hat{\beta}_k))$

Every inferential object depends on  $\text{Var}(\hat{\beta}_k)$ :

**Standard error**

$$\text{SE}(\hat{\beta}_k) = \sqrt{\widehat{\text{Var}}(\hat{\beta}_k)}$$

**t-statistic**

$$t = \frac{\hat{\beta}_k - \beta_k^0}{\text{SE}(\hat{\beta}_k)}$$

**Confidence interval**

$$\hat{\beta}_k \pm 1.96 \cdot \text{SE}(\hat{\beta}_k)$$

Wrong  $\widehat{\text{Var}}(\hat{\beta}_k) \Rightarrow$  wrong SE  $\Rightarrow$  wrong  $t \Rightarrow$  wrong  $p$ -values and CIs

Getting  $\widehat{\text{Var}}(\hat{\beta}_k)$  right IS the whole inference problem.

## Homoskedasticity: one number summarizes the error variance everywhere

**Assumption:**  $\text{Var}(\varepsilon_i | X_i) = \sigma^2$  (constant, does not depend on  $X_i$ )

**Population variance of  $\hat{\beta}$ :**

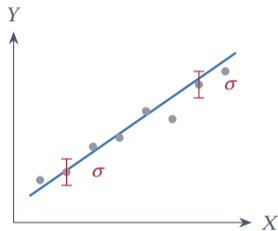
$$\text{Var}(\hat{\beta}) = \sigma^2 (X'X)^{-1}$$

**Estimated:**

$$\widehat{\text{Var}}(\hat{\beta}) = \hat{\sigma}^2 (X'X)^{-1}$$

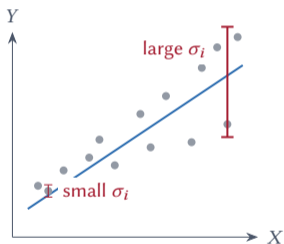
$$\hat{\sigma}^2 = \frac{1}{n-k} \sum_{i=1}^n \hat{\varepsilon}_i^2$$

One number  $\hat{\sigma}^2$  scales the whole variance matrix. Clean — but only valid under homoskedasticity.



## Heteroskedasticity: the error variance changes with $X$ – the homoskedastic formula breaks

Fan-shaped residuals:



$\text{Var}(\varepsilon_i | X_i) = \sigma_i^2$  varies with  $X_i$

The homoskedastic formula uses  $\hat{\sigma}^2$  – a single average:

$$\widehat{\text{Var}}_{\text{hom}}(\hat{\beta}) = \hat{\sigma}^2 (X'X)^{-1}$$

- Over-smooths the true variance structure
- Wrong in every part of the  $X$  range
- SEs too small where variance is large
- SEs too large where variance is small

**Result:  $t$ -statistics,  $p$ -values, and CIs are all wrong.**

## Three statisticians solved the problem – White brought it to econometrics

### Eicker (1963)

German statistician

Proved the sandwich estimator is *consistent* for the asymptotic variance under heteroskedasticity

Ann. Math. Stat. 1963

### Huber (1967)

Swiss statistician

Extended the theory to M-estimators broadly – robust statistics as a general framework

Berkeley Symp. 1967

### White (1980)

UC San Diego economist

Synthesized Eicker and Huber for econometrics; added the White test for heteroskedasticity

Econometrica 1980

The estimator is called **Eicker-Huber-White (EHW)** or **HC** (heteroskedasticity-consistent) standard errors

White's 1980 paper made the sandwich practical – before it, nearly everyone reported naive SEs by default.

## The sandwich estimator lets each residual speak for itself

### Eicker-Huber-White (robust) variance estimator:

$$\widehat{\text{Var}}_{robust}(\hat{\beta}) = \underbrace{(X'X)^{-1}}_{\text{bread}} \underbrace{\left( \sum_{i=1}^n \hat{e}_i^2 X_i X_i' \right)}_{\text{meat}} \underbrace{(X'X)^{-1}}_{\text{bread}}$$

#### Why it works:

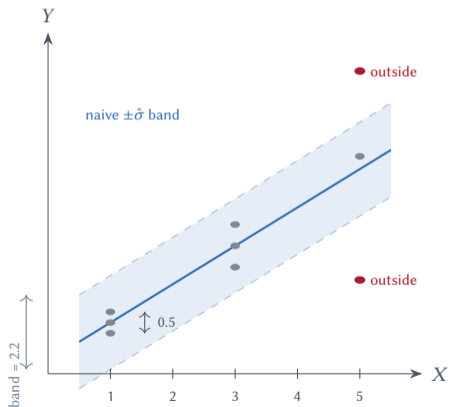
- $\hat{e}_i^2$  estimates  $\sigma_i^2$  at each observation
- Weights  $X_i X_i'$  by the local variance
- No assumption on the shape of  $\sigma_i^2$

#### In practice:

- `lm_robust()` in R (estimatr)
- `vce(robust)` in Stata
- Use by default — the cost of not using it can be severe

Angrist & Pischke: “always report robust SEs” — homoskedasticity is an assumption; robust SEs are not.

# Naive SE: one band forced everywhere – too wide at low $X$ , too narrow at high $X$



## What naive SE computes:

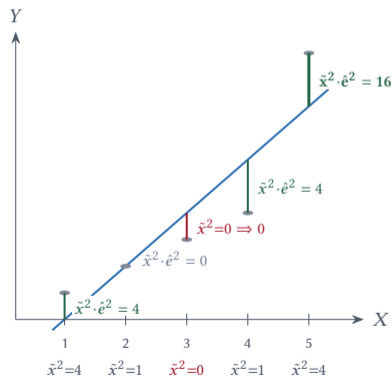
- Pool all  $\hat{e}_i^2$  into one  $\hat{\sigma}^2$
- Same band width *everywhere* on the  $X$  axis

## What actually happens:

- Low  $X$ : band is too wide – excess empty space
- High  $X$ : band is too narrow – **points escape**

Result:  $t$ -statistics and  $p$ -values are wrong in *both* directions simultaneously

# Robust SE: each residual weighted by how much that observation pulls on $\hat{\beta}_1$



$$\text{Meat} = \sum \tilde{x}_i^2 \hat{e}_i^2:$$

- Each  $\hat{e}_i^2$  is *weighted* by  $\tilde{x}_i^2 = (X_i - \bar{X})^2$
- $\tilde{x}_i^2$  = how much obs  $i$  pulls on  $\hat{\beta}_1$

## Key insight:

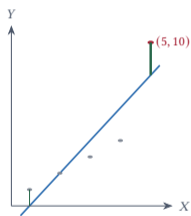
- Obs. 3 ( $\tilde{x}_3=0$ ): zero weight even though  $\hat{e}_3 \neq 0$
- Obs. 5 ( $\tilde{x}_5=2$ ): weight = 4 amplifies large residual

$$\text{Meat} = 4 + 0 + 0 + 4 + 16 = 24$$

# Fan-shaped data: naive SE underestimates uncertainty where it matters most

**Data:**  $Y = (1, 2, 3, 4, 10)$

Obs. 5 is a fan-shape outlier



**OLS:**  $\hat{\beta}_0 = -2, \hat{\beta}_1 = 2; \hat{Y}_i = -2 + 2X_i$

$i$	$X_i$	$Y_i$	$\hat{Y}_i$	$\hat{e}_i$	$\hat{e}_i^2$
1	1	1	0	+1	1
2	2	2	2	0	0
3	3	3	4	-1	1
4	4	4	6	-2	4
5	5	10	8	+2	4
					$\sum \hat{e}_i^2 = 10$

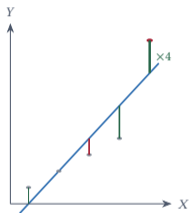
**Step 1:**  $\hat{\sigma}^2 = \frac{10}{3} = 3.33$       **Step 2:**  $\sum \tilde{x}_i^2 = 10$

**Step 3:**  $\widehat{\text{Var}}_{\text{naive}} = \frac{3.33}{10} = 0.333$

$$\text{SE}_{\text{naive}} = \sqrt{0.333} \approx 0.577$$

# Fan-shaped data: robust SE correctly flags more uncertainty at high X

Same data; now weight each  $\hat{e}_i^2$  by  $\tilde{x}_i^2$ :



$\text{Meat} = \sum \tilde{x}_i^2 \hat{e}_i^2$ , then  $\text{HC1} = \frac{n}{n-k} \times \text{meat} / S_{xx}^2$ :

$i$	$\tilde{x}_i$	$\tilde{x}_i^2$	$\hat{e}_i^2$	$\tilde{x}_i^2 \cdot \hat{e}_i^2$
1	-2	4	1	4
2	-1	1	0	0
3	0	0	1	0
4	+1	1	4	4
5	+2	4	4	16
meat =				24

$$\text{HC1: } \widehat{\text{Var}}_{\text{HC1}} = \frac{n}{n-k} \cdot \frac{\text{meat}}{S_{xx}^2} = \frac{5}{3} \cdot \frac{24}{100} = 0.400$$

$$\text{SE}_{\text{HC1}} = \sqrt{0.400} \approx 0.632$$

**Naive:** SE = 0.577

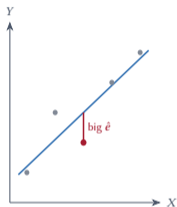
**Robust (HC1):** SE = 0.632    ↑ larger — correct

Obs. 5 contributes 16 of the 24 meat — the fan's wide end drives the robust SE above naive.

## Naive SE: one $\hat{\sigma}^2$ pools every $\hat{e}_i^2$ with equal weight

Data:  $Y = (1, 3, 2, 4, 5)$

Big residual at center ( $X = 3$ )



OLS:  $\hat{\beta}_0 = 0.3, \hat{\beta}_1 = 0.9; \hat{Y}_i = 0.3 + 0.9X_i$

$i$	$X_i$	$Y_i$	$\hat{Y}_i$	$\hat{e}_i$	$\hat{e}_i^2$
1	1	1	1.2	-0.20	0.04
2	2	3	2.1	+0.90	0.81
3	3	2	3.0	-1.00	1.00
4	4	4	3.9	+0.10	0.01
5	5	5	4.8	+0.20	0.04
					$\sum \hat{e}_i^2 = 1.90$

Step 1:  $\hat{\sigma}^2 = \frac{1.90}{3} = 0.633$       Step 2:  $\sum \tilde{x}_i^2 = 10$

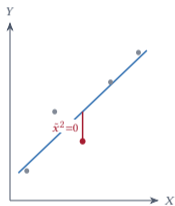
Step 3:  $\widehat{\text{Var}}_{\text{naive}} = \frac{0.633}{10} = 0.0633$

$$\text{SE}_{\text{naive}} = \sqrt{0.0633} \approx 0.252$$

The large residual at  $X = 3$  inflates  $\hat{\sigma}^2$  – but that observation has zero leverage on  $\hat{\beta}_1$ .

# Non-fan data: robust SE is *smaller* — the big residual is at a zero-leverage point

Same data; weight each  $\hat{e}_i^2$  by  $\tilde{x}_i^2$ :



Meat =  $\sum \tilde{x}_i^2 \hat{e}_i^2$ , then HC1 =  $\frac{n}{n-k} \times \text{meat} / S_{xx}^2$ :

$i$	$\tilde{x}_i$	$\tilde{x}_i^2$	$\hat{e}_i^2$	$\tilde{x}_i^2 \cdot \hat{e}_i^2$
1	-2	4	0.04	0.16
2	-1	1	0.81	0.81
3	0	0	1.00	0.00
4	+1	1	0.01	0.01
5	+2	4	0.04	0.16
meat =				1.14

$$\text{HC1: } \widehat{\text{Var}}_{\text{HC1}} = \frac{5}{3} \cdot \frac{1.14}{100} = 0.019$$

$$\text{SE}_{\text{HC1}} = \sqrt{0.019} \approx 0.138$$

Naive: SE = 0.252

Robust (HC1): SE = 0.138 ↓ smaller — correct

Robust SE goes *down* when the big residuals are at low-leverage points — it is honest in both directions.

## Two formulas, two philosophies – naive pools, robust weights

**Naive (Homoskedastic)**

**Scalar (bivariate slope):**

$$\widehat{\text{Var}}_{\text{naive}}(\hat{\beta}_1) = \frac{\hat{\sigma}^2}{\sum \tilde{x}_i^2}$$

$$\text{where } \hat{\sigma}^2 = \frac{\sum \hat{e}_i^2}{n-k}$$

**Matrix:**

$$\widehat{\text{Var}}_{\text{naive}}(\hat{\beta}) = \hat{\sigma}^2 (X'X)^{-1}$$

One number  $\hat{\sigma}^2$  scales the whole matrix

**Robust HC1 (EHW)**

**Scalar (bivariate slope):**

$$\widehat{\text{Var}}_{\text{HC1}}(\hat{\beta}_1) = \frac{n}{n-k} \cdot \frac{\sum \tilde{x}_i^2 \hat{e}_i^2}{(\sum \tilde{x}_i^2)^2}$$

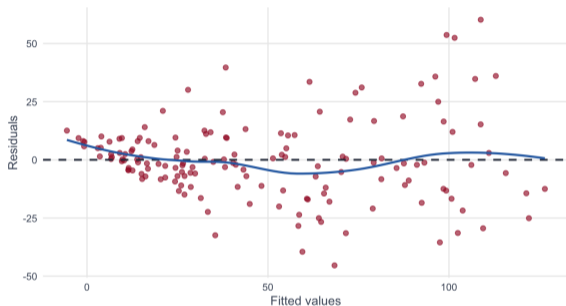
**Matrix:**

$$\widehat{\text{Var}}_{\text{HC1}} = \frac{n}{n-k} (X'X)^{-1} \left( \sum_i \hat{e}_i^2 X_i X_i' \right) (X'X)^{-1}$$

Each  $\hat{e}_i^2$  enters with its own leverage weight

HC0 drops the  $n/(n-k)$  correction; HC2 and HC3 apply further leverage corrections. Stata's `vce(robust)` uses HC1 by default.

## Back to UN98: the residuals reveal a clear fan shape



### What we see:

- Low fitted values: residuals spread widely
- High fitted values: residuals compress
- Classic fan shape — variance falls as predicted mortality rises

### Breusch-Pagan test:

$$\chi^2(3) = 38.3, \quad p < 0.001$$

One  $\hat{\sigma}^2$  for all 154 countries —  
over-weights rich, under-weights poor.

## Back to UN98: naive SEs are too small – robust SEs are 13–19% larger

	Naive SE	Robust SE	Change
Intercept	11.29	13.49	+19%
ln(GDP)	1.21	1.39	+15%
Fertility	1.39	1.56	+13%
Illiteracy	0.09	0.10	+18%

$\hat{\sigma}^2$  is too small because it averages over low-variance (high-income) and high-variance (low-income) countries together.

95% CI for  $\hat{\beta}_{\ln(\text{GDP})}$ :

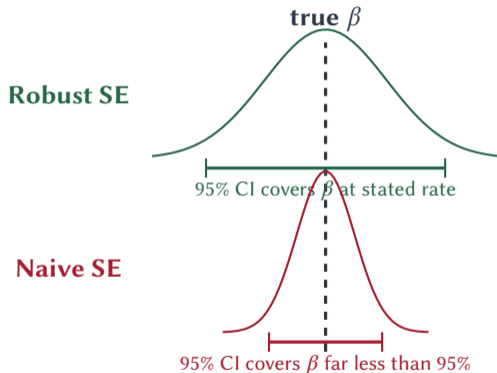
**Naive:** [-9.53, -4.81]

**Robust:** [-9.88, -4.45]

Robust CI is **15% wider**  
(5.43 vs. 4.72 in width)

The naive formula tells us we know  $\beta_{\ln(\text{GDP})}$  more precisely than we actually do

## What heteroskedasticity does to inference – a visual



$\hat{\beta}$  is in the same place — unbiasedness intact; the spread we *report* is wrong, and that IS the entire problem.

## Wrong SE: size distortion in tests, under-coverage in intervals

### Hypothesis tests: size distortion

Size =  $\mathbb{P}(\text{reject } H_0 \mid H_0 \text{ true})$

Correct: size =  $\alpha$

Naive SE: size  $> \alpha$

- $|t|$  inflated (SE too small)
- Over-reject the null
- False positives above stated rate

### Confidence intervals: under-coverage

Coverage =  $\mathbb{P}(\beta \in \text{CI})$

Correct: coverage =  $1 - \alpha$

Naive SE: coverage  $< 1 - \alpha$

- CI is too narrow
- Misses true  $\beta$  too often
- “95% CI” covering only 88%

Wrong SE  $\Rightarrow$  same problem, two faces: over-reject the null; under-cover the parameter

“You think you’re running a 5% test. You’re actually running a 10–15% test.” *That is size distortion.*

## HC robust SEs fix heteroskedasticity – clustering fixes a different problem

### HC (EHW) robust SEs

**Problem:**  $\text{Var}(\varepsilon_i | X_i) = \sigma_i^2$  varies

**Assumes:**  $\varepsilon_i$  independent across  $i$

**Fix:** weight each by  $\hat{\varepsilon}_i^2$  per obs

### Clustered SEs

**Problem:**  $\text{Cov}(\varepsilon_i, \varepsilon_j) \neq 0$  within groups

**Example:** students in same school, states, countries

**Fix:** sum the error over clusters, not obs

Two different violations. Two different fixes. Both called “robust” – they are **not** the same thing.

HC assumes independence; clustering allows within-group dependence – if you have both, cluster.

## The cluster-robust sandwich: group the meat by cluster, not by observation

EHW (HC) – sum over  $i$ :

$$\widehat{V}_{HC} = (X'X)^{-1} \left( \sum_{i=1}^n \hat{e}_i^2 X_i X_i' \right) (X'X)^{-1}$$

Cluster-robust – sum over  $g$ :

$$\widehat{V}_{cl} = (X'X)^{-1} \left( \sum_{g=1}^G X_g' \hat{e}_g \hat{e}_g' X_g \right) (X'X)^{-1}$$

$\hat{e}_g$  is the vector of residuals in cluster  $g$

When to cluster:

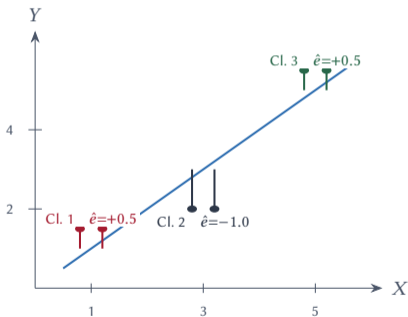
- Treatment assigned at group level (state, school, village)
- Panel data (same unit over time)
- Observations share a common shock

**Rule of thumb:** need  $G \geq 50$  clusters

Few clusters ( $G < 20$ ): clustered SEs themselves unreliable – use wild cluster bootstrap

Ignoring clustering can make SEs wrong by a factor of 2–3, not just 15%

## Naive SE on clustered data: squares each residual independently – misses that same-cluster obs move together



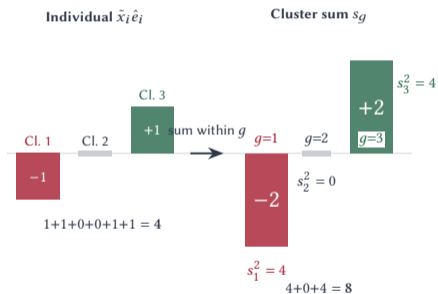
### Setting:

- $n=6$ ,  $G=3$  clusters of 2
- $X = (1, 1, 3, 3, 5, 5)$
- Same  $\hat{e}$  within each cluster

### HC1 (obs treated as independent):

- Scores  $\tilde{x}_i \hat{e}_i$ :  $(-1, -1, 0, 0, 1, 1)$
- Squares each separately:  
 $1+1+0+0+1+1 = 4$
- Ignores that Cl. 1 obs both pull  $\hat{\beta}_1$  the same direction

# Cluster-robust SE: sum scores within each cluster first, then square — cross-terms capture the correlation



Why CRVE meat > naive:

$$(a + b)^2 = a^2 + 2ab + b^2$$

Cluster 1 ( $a=b=-1$ ):

- Naive:  $(-1)^2 + (-1)^2 = 2$
- CRVE:  $(-1-1)^2 = 4$
- Extra:  $2(-1)(-1) = +2$

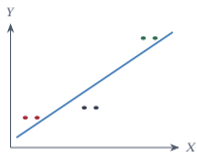
CRVE meat = 8

Naive meat = 4

# Clustered data: HC1 treats each of 6 obs as independent — $SE_{HC1} = 0.153$

## Clustered data:

$\hat{\beta}_0 = 0, \hat{\beta}_1 = 1$ ; line  $y = x$



Meat =  $\sum_i \tilde{x}_i^2 \hat{e}_i^2$  (obs treated independently):

$i$	$g$	$\tilde{x}_i$	$\hat{e}_i$	$\hat{e}_i^2$	$\tilde{x}_i^2 \cdot \hat{e}_i^2$
1	1	-2	+0.5	0.25	1.00
2	1	-2	+0.5	0.25	1.00
3	2	0	-1.0	1.00	0.00
4	2	0	-1.0	1.00	0.00
5	3	+2	+0.5	0.25	1.00
6	3	+2	+0.5	0.25	1.00
meat =				4.00	

$$\text{HC1: } \widehat{\text{Var}}_{\text{HC1}} = \frac{n}{n-k} \cdot \frac{\text{meat}}{S_{xx}^2} = \frac{6}{4} \cdot \frac{4.00}{256} = 0.02344$$

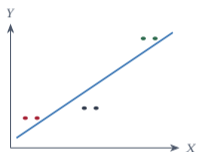
$$SE_{\text{HC1}} = \sqrt{0.02344} \approx 0.153$$

HC1 treats obs 1 and obs 2 (both in Cluster 1) as independent. They are not: same  $\tilde{x}$ , same  $\hat{e}$ , same pull on  $\hat{\beta}_1$ .

## Cluster-robust SE: group scores by cluster before squaring —

$$SE_{CRVE} = 0.242 > 0.153$$

Same data; now  
sum within cluster:



Step 1: obs-level scores; Step 2: cluster sums  $s_g$ :

$i$	$g$	$\tilde{x}_i$	$\hat{e}_i$	$\tilde{x}_i \hat{e}_i$			
1	1	-2	+0.5	-1	$g$	$s_g$	$s_g^2$
2	1	-2	+0.5	-1	1	-2	4
3	2	0	-1.0	0	2	0	0
4	2	0	-1.0	0	3	+2	4
5	3	+2	+0.5	+1	meat =		8
6	3	+2	+0.5	+1			

$$CRVE: \widehat{Var}_{CRVE} = \frac{G}{G-1} \cdot \frac{n-1}{n-k} \cdot \frac{\text{meat}}{S_{xx}^2} = \frac{3}{2} \cdot \frac{5}{4} \cdot \frac{8}{256} = 0.05859$$

$$SE_{CRVE} = \sqrt{0.05859} \approx 0.242$$

HC1 (naive): SE = 0.153

CRVE: SE = 0.242    ↑ 58% larger — correct

## Two SE formulas for correlated data – HC1 sums over obs, CRVE sums over clusters

### HC1 (obs as independent)

Scalar:

$$\widehat{\text{Var}}_{\text{HC1}}(\hat{\beta}_1) = \frac{n}{n-k} \cdot \frac{\sum_i \tilde{x}_i^2 \hat{e}_i^2}{\left(\sum_i \tilde{x}_i^2\right)^2}$$

Matrix:

$$\widehat{\text{Var}}_{\text{HC1}} = \frac{n}{n-k} (X'X)^{-1} \left( \sum_i \hat{e}_i^2 X_i X_i' \right) (X'X)^{-1}$$

Each  $\hat{e}_i^2$  enters separately

### CRVE (cluster-robust)

Scalar:  $s_g = \sum_{i \in g} \tilde{x}_i \hat{e}_i$

$$\widehat{\text{Var}}_{\text{CRVE}}(\hat{\beta}_1) = \frac{G}{G-1} \cdot \frac{n-1}{n-k} \cdot \frac{\sum_g s_g^2}{\left(\sum_i \tilde{x}_i^2\right)^2}$$

Matrix:  $(s_g = \sum_{i \in g} X_i \hat{e}_i)$

$$\widehat{\text{Var}}_{\text{CRVE}} = \frac{G}{G-1} \cdot \frac{n-1}{n-k} \times (X'X)^{-1} \left( \sum_g s_g s_g' \right) (X'X)^{-1}$$

Scores summed within  $g$ , then squared

The correction  $\frac{G}{G-1} \cdot \frac{n-1}{n-k}$  inflates CRVE; need  $G \geq 50$  clusters for reliability. Use wild cluster bootstrap for  $G < 20$ .

## $t$ -statistic and confidence interval for a single coefficient $\hat{\beta}_k$

From asymptotic normality:

$$\frac{\hat{\beta}_k - \beta_k}{\text{SE}(\hat{\beta}_k)} \xrightarrow{d} N(0, 1)$$

Hypothesis test ( $H_0 : \beta_k = 0$ ):

$$t = \frac{\hat{\beta}_k}{\text{SE}(\hat{\beta}_k)}$$

Reject  $H_0$  at 5% if  $|t| > 1.96$

$p$ -value =  $2\mathbb{P}(Z > |t|)$ ,  $Z \sim N(0, 1)$

95% confidence interval:

$$\hat{\beta}_k \pm 1.96 \cdot \text{SE}(\hat{\beta}_k)$$

Covers  $\beta_{BLP,k}$  in 95%  
of repeated samples

Use  $t_{n-k}$  critical values for small  $n$

$$\text{SE}(\hat{\beta}_k) = \sqrt{\widehat{\text{Var}}(\hat{\beta}_k)} \text{ where } \widehat{\text{Var}} \text{ is the sandwich — robust or classical}$$

The formula for  $t$  and the CI is the same regardless. What changes is how we estimate  $\text{Var}(\hat{\beta}_k)$ .

## F-test: testing joint hypotheses on multiple coefficients

### Why not just run multiple $t$ -tests?

- Testing  $\beta_1 = 0$  and  $\beta_2 = 0$  separately inflates the Type I error rate
- A joint test holds the false rejection rate at  $\alpha$  for the *joint* null

### Restricted vs. unrestricted models:

$$F = \frac{(SSR_R - SSR_U)/q}{SSR_U/(n - k)} \sim F_{q, n-k} \text{ under } H_0$$

- $SSR_R$ : sum of squared residuals imposing  $H_0$
- $SSR_U$ : sum of squared residuals of the full model
- $q$ : number of restrictions
- Intuition: how much does imposing  $H_0$  hurt the fit?

Special case  $q = 1$ :  $F = t^2$  — the  $F$ -test is the  $t$ -test generalized to  $q$  simultaneous restrictions.

## **F-test example: are all slope coefficients jointly zero?**

**Null:**  $H_0 : \beta_1 = \beta_2 = \dots = \beta_{k-1} = 0$

**Restricted model** (intercept only):

$$Y_i = \beta_0 + \varepsilon_i$$

$$SSR_R = \sum (Y_i - \bar{Y})^2 = TSS$$

**Unrestricted model:**

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

$$SSR_U = \sum \hat{\varepsilon}_i^2$$

$$F = \frac{(TSS - SSR_U)/(k - 1)}{SSR_U/(n - k)} = \frac{R^2/(k - 1)}{(1 - R^2)/(n - k)}$$

When  $F$  is large, the regressors together explain enough variance to reject  $H_0$

This is the “overall  $F$ -statistic” in every regression output – does the model explain anything at all?

$R^2$  is the fraction of  $Y$ -variance explained by the regression

Variance decomposition:

$$\underbrace{\sum_i (Y_i - \bar{Y})^2}_{TSS} = \underbrace{\sum_i (\hat{Y}_i - \bar{Y})^2}_{ESS} + \underbrace{\sum_i \hat{e}_i^2}_{SSR}$$

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{SSR}{TSS} \in [0, 1]$$

- $R^2 = 0$ : regression explains nothing (flat line through  $\bar{Y}$ )
- $R^2 = 1$ : perfect fit, all points on the line
- Bivariate OLS:  $R^2 = \widehat{\text{Corr}}(X, Y)^2$

Adding *any* regressor — even noise — can only increase  $R^2$  or leave it flat. Never decreases.

## Adjusted $R^2$ penalizes complexity – $\bar{R}^2$ can fall when you add irrelevant variables

Adjusted  $R^2$ :

$$\bar{R}^2 = 1 - \frac{SSR/(n-k)}{TSS/(n-1)} = 1 - (1 - R^2) \frac{n-1}{n-k}$$

What it does:

- Divides by degrees of freedom, not raw sums
- Penalizes adding regressors that don't reduce  $SSR$  enough
- $\bar{R}^2$  can decrease – signals the new variable wasn't worth it

$\bar{R}^2$  is a better model-comparison tool than  $R^2$  – but neither is a causal object.

When  $\bar{R}^2$  falls:

$$\Delta \bar{R}^2 < 0 \iff F_{\text{new var}} < 1$$

The new variable's  $F$ -statistic is less than 1 – it actively hurts the adjusted fit.

## What $R^2$ tells you – and what it doesn't

### $R^2$ does tell you:

- How well the line fits this sample
- How much variance is explained by the included regressors
- Whether two models fit differently (via  $\bar{R}^2$ )

### $R^2$ does NOT tell you:

- Whether  $\hat{\beta}$  is unbiased or consistent
- Whether the causal interpretation is valid
- Whether the standard errors are correct
- Whether you should add more variables

Low  $R^2$  with correct specification beats high  $R^2$  with the wrong model

Causal inference cares about specification, not fit –  $R^2 = 0.04$  with a valid instrument beats  $R^2 = 0.90$  with endogenous regressors.

## OLS inference: everything rests on getting $\text{Var}(\hat{\beta})$ right

### The arc of today:

- **OLS** is the sample plug-in for the BLP:  $\hat{\beta} = (X'X)^{-1}X'Y$
- **Unbiased** under strict exogeneity — finite-sample result
- **Consistent** under the weaker moment condition  $\mathbb{E}[X_i\varepsilon_i] = 0$
- **Asymptotically normal** because  $\frac{X'\varepsilon}{n}$  is a sample mean — CLT + Slutsky
- **BLUE** (Gauss-Markov) under homoskedasticity — loses “B” under heteroskedasticity
- **Heteroskedasticity** is an inference problem: use the sandwich (EHW) estimator
- **Clustering**: within-group correlation requires a different fix — group the meat by cluster
- **Inference**:  $t$ -stats and CIs for single  $\hat{\beta}_k$ ;  $F$ -tests for joint hypotheses
- $R^2$  measures fit, not correctness —  $\bar{R}^2$  penalizes complexity

Monday: diagnostics — outliers, leverage, Cook's  $D$  — and the Angrist (1998) variance-weights decomposition.