Non-spherical Standard Errors

Robust SEs, clustering, and the sampling distribution of $\hat{\beta}$

Gov 2001 · Scott Cunningham · Spring 2026

Where we are: OLS done; today we sharpen the standard errors

Last week — inference and SE calculation in OLS:

- Derived OLS as the sample plug-in for the BLP: $\hat{\beta} = (X'X)^{-1}X'Y$
- Proved unbiasedness, consistency, and asymptotic normality
- Built the homoskedastic variance formula: $\widehat{\text{Var}}(\hat{\beta}) = \hat{\sigma}^2(X'X)^{-1}$
- Saw that wrong $\widehat{\text{Var}}(\hat{\beta}) \Rightarrow$ wrong SE $\Rightarrow$ wrong t-stats, p-values, CIs

Today — when the homoskedastic formula breaks:

1. **Robust SEs** — revisited carefully (last week was too fast)
2. **Cluster-robust SEs** — a different violation, a different fix
3. **Practice exam worksheet** — distributed in class; also on Canvas & the course website

The point estimate $\hat{\beta}$ does not change. Only $\widehat{\text{Var}}(\hat{\beta})$ does.

What a standard error *is*: the sampling distribution of $\hat{\beta}$

The conceptual move. $\hat{\beta}$ is a random variable. Different samples from the same population would give different $\hat{\beta}$ values.

Sampling distribution of $\hat{\beta}_k$: the distribution of $\hat{\beta}_k$ across all hypothetical samples drawn from the same data-generating process.

Standard error: the standard deviation of this sampling distribution.

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

> The SE is not about your data. It is about how much $\hat{\beta}$ would jitter if you re-ran the experiment.

Every inferential statement — t-stats, p-values, confidence intervals — is a quantitative claim about that jitter.

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

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

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

Standard error

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

***t*-statistic**

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

Confidence interval

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

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

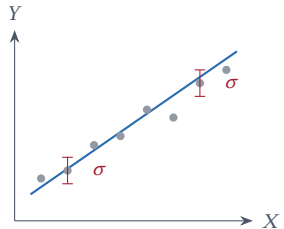
Getting $\widehat{\mathrm{Var}}(\hat{\beta}_k)$ right IS the whole inference problem.

Homoskedasticity: one number summarizes the error variance everywhere

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

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

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



Estimated:

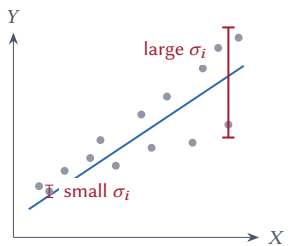
$$\widehat{\text{Var}}(\hat{\beta}) = \hat{\sigma}^2 (X'X)^{-1}$$

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

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:



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

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

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

- 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.

From homoskedasticity to heteroskedasticity: where the formula must change

The general variance of $\hat{\beta}$ is always:

$$\text{Var}(\hat{\beta}) \;=\; (X'X)^{-1}\Big(\textstyle\sum_i \sigma_i^2\, X_i X_i'\Big)(X'X)^{-1}$$

Homoskedastic case

$\sigma_i^2 = \sigma^2$ for every i

σ^2 *pulls out* of the sum:

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

one number summarizes all the noise

Heteroskedastic case

σ_i^2 varies with i

σ_i^2 *stays inside* the sum:

$\text{Var}(\hat{\beta}) = (X'X)^{-1}\big(\sum_i \sigma_i^2 X_i X_i'\big)(X'X)^{-1}$

no single σ^2 to pool

Replace each unobserved σ_i^2 with the observable $\hat{e}_i^2$. That is the sandwich.

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:

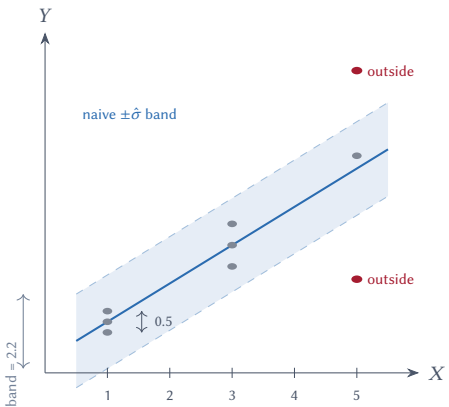
- $\hat{e}_i^2$ estimates σ_i^2 at each observation
- Weights $X_i X_i'$ by the local variance
- No assumption on the shape of σ_i^2

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

Two ingredients of the SE: residual noise and X-spread

Before we plug numbers into the formula, name the two pieces.

Numerator: residual noise

$$\hat{\sigma}^2 = \frac{1}{n-k}\sum_i \hat{e}_i^2$$

Y-residuals, squared, averaged

Denominator: X-spread

$$\sum_i \tilde{x}_i^2 \quad \text{where} \quad \tilde{x}_i = X_i - \bar{X}$$

X-deviations from the mean, squared, summed

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

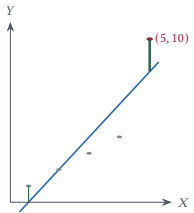
noise over information

For the example coming up: $X = (1, 2, 3, 4, 5)$, so $\bar{X} = 3$ and $\tilde{x} = (-2, -1, 0, +1, +2)$, giving $\sum \tilde{x}_i^2 = 10$.

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

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

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

Three ingredients of the robust SE: per-observation noise and leverage

Naive pools the noise. Robust keeps each observation's noise tied to that observation's leverage.

Per-obs noise

$\hat{e}_i^2$

each squared residual stays separate — no pooling

Per-obs leverage

$\tilde{x}_i^2$

how much obs i pulls on $\hat{\beta}_1$

Combined: the meat

$\sum_i \tilde{x}_i^2 \, \hat{e}_i^2$

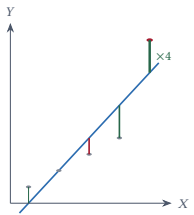
noise weighted by leverage

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

Same denominator structure as naive (regressor information). Numerator is the new piece: each $\hat{e}_i^2$ keeps its own weight $\tilde{x}_i^2$ instead of being pooled into one $\hat{\sigma}^2$.

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

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



Meat = $\sum \tilde{x}_i^2 \hat{e}_i^2$, then HC1 = $\frac{n}{n-k}\times$ 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

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

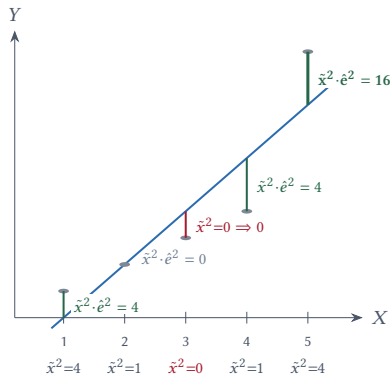
$$\mathrm{SE}_{\mathrm{HC1}} = \sqrt{0.400} \approx 0.632$$

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.

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



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

Meat = $4 + 0 + 0 + 4 + 16 = 24$

The opposite case: when the big residual sits at zero leverage

So far: fan-shape data, $Y = (1, 2, 3, 4, 10)$. Big residual at the high-X end. Robust SE *larger* than naive.

Now: switch the dataset. Same formulas. Different shape.

- New data: $Y = (1, 3, 2, 4, 5)$
- Big residual at $X = 3 = \bar{X}$ — exactly at the center, where $\tilde{x}_i = 0$
- That observation's leverage on $\hat{\beta}_1$ is *zero*

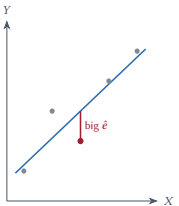
A residual at zero leverage gets weight $\tilde{x}^2 = 0$ in the meat. Robust ignores it. Naive averages it in. Robust will be *smaller* than naive.

Same machinery, opposite result. The lesson: robust isn't always bigger; it's correctly responsive to where the noise lives.

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

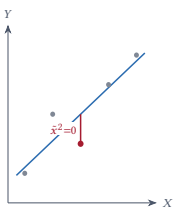
Step 3: $\widehat{\text{Var}}_{\text{naive}} = \dfrac{0.633}{10} = 0.0633$

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

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** $= \frac{n}{n-k} \times$ **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

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):

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

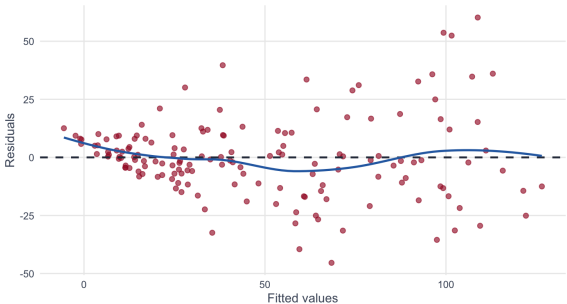
Matrix:

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

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.

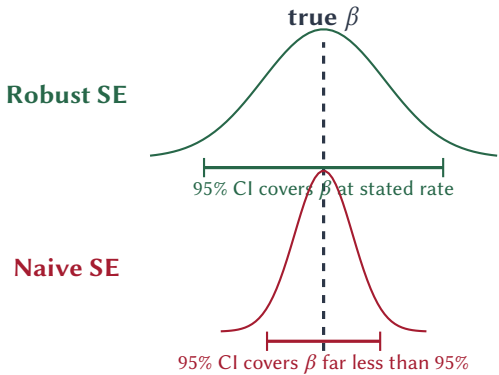
95% CI for $\hat{\beta}_{\ln(\text{GDP})}$:

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 $\beta_{\ln(GDP)}$ 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 = α

Naive SE: size > α

- $|t|$ 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 β 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: $\text{Var}(\varepsilon_i \mid X_i) = \sigma_i^2$ varies

Assumes: ε_i independent across i

Fix: weight meat by $\hat{e}_i^2$ 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 meat 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:

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

Cluster-robust — sum over g:

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

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

When to cluster:

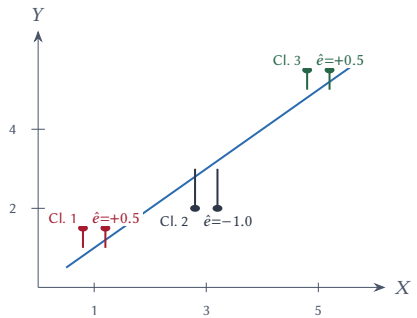
- Treatment assigned at group level (state, school, village)
- Panel data (same unit over time)
- Observations share a common shock

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:

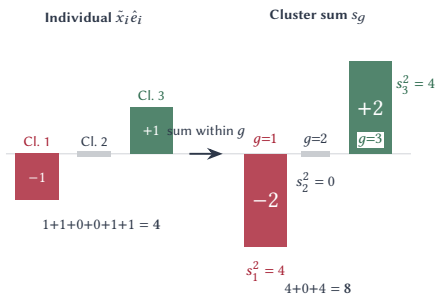
- n=6, G=3 clusters of 2
- $X = (1, 1, 3, 3, 5, 5)$
- Same $\hat{e}$ within each cluster

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

Within-cluster correlation ⇒ effective $n < 6$.

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: $2(-1)(-1) = +2$

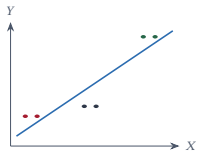
CRVE meat = 8

Naive meat = 4

Clustered data: HC1 treats each of 6 obs as independent — $\text{SE}_{\text{HC1}} = 0.153$

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

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

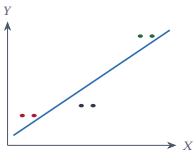
$$\text{SE}_{\text{HC1}} = \sqrt{0.02344} \approx 0.153$$

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 —

$\text{SE}_{\text{CRVE}} = 0.242 > 0.153$

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
2	1	−2	+0.5	−1
3	2	0	−1.0	0
4	2	0	−1.0	0
5	3	+2	+0.5	+1
6	3	+2	+0.5	+1

g	s_g	s_g^2
1	−2	4
2	0	0
3	+2	4
meat =		8

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

$$\text{SE}_{\text{CRVE}} = \sqrt{0.05859} \approx 0.242$$

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:

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

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

CRVE (cluster-robust)

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

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

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

$$\widehat{\text{Var}}_{\text{CRVE}} = \tfrac{G}{G-1} \cdot \tfrac{n-1}{n-k} \times (X'X)^{-1}\left(\textstyle\sum_g \mathbf{s}_g \mathbf{s}_g'\right)(X'X)^{-1}$$

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 $G < 20$.

We have the right SE. Now: what do we *do* with it?

So far today — getting $\widehat{\mathrm{Var}}(\hat{\beta})$ right:

- Heteroskedasticity $\Rightarrow$ the sandwich (HC1) keeps each $\hat{e}_i^2$ separate
- Within-cluster correlation $\Rightarrow$ CRVE sums the meat by cluster

Next — what we *do* with the corrected SE:

1. t-statistic and confidence interval for one coefficient
2. F-test for joint hypotheses on multiple coefficients
3. R^2 and adjusted R^2 as fit summaries

The inferential machinery is unchanged. Plug in the corrected SE; everything else works as before.

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

$\text{SE}(\hat{\beta}_k) = \sqrt{\widehat{\text{Var}}(\hat{\beta}_k)}$ where $\widehat{\text{Var}}$ 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 α 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: $H_0 : \beta_1 = \beta_2 = \cdots = \beta_{k-1} = 0$

Restricted model (intercept only):

$$Y_i = \beta_0 + \varepsilon_i$$

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

Unrestricted model:

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

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

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

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

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.

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

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 $\mathrm{Var}(\hat{\beta})$ 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

Wednesday: variance-weighted regression (Angrist 1998 & Słoczyński 2022).

Practice exam worksheet — pick one up on the way out

What's in it.

- **15 practice problems** in five scaffolded parts
- Each part builds toward one of the five families on the real final
- Worked solutions packet — separate document
- "Mock final path": 5 problems to do timed, closed-book

Where to find it.

- Paper copies passed around now
- Canvas → Files → `final_practice.pdf`
- Course website slides folder

Recommended order if you're tight on time:

1. **Practice 5c** (highest leverage — 25 pts of the exam)
2. **Practice 4c** (matrix OLS + FWL on a substantive control)
3. **Practice 1c** (variance-stabilizing transform)
4. **Practice 2c** (one-sided Bernoulli; CLT *vs.* Chebyshev)
5. **Practice 3a or 3b** (T/F traps)

The c-problems are the full "in-family" rehearsals. The a's and b's are scaffolds — only do them if you're stuck.

Solutions to the practice exam will be posted later this week.