

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

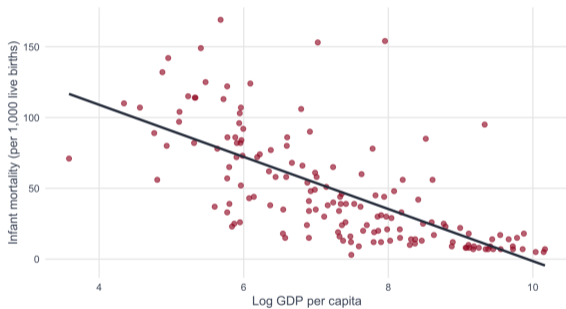
Final exam: Wednesday May 7 · 2:00 pm

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):

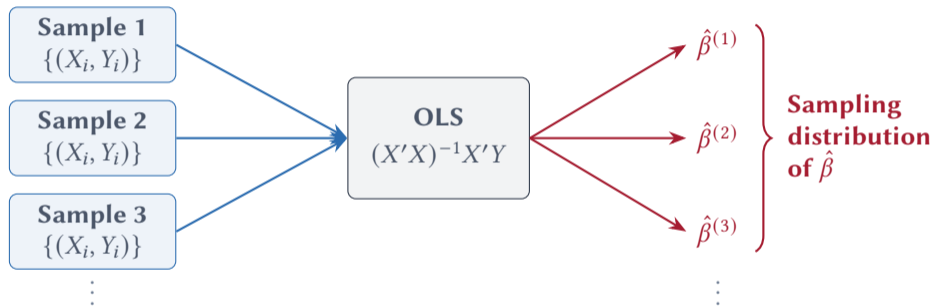
- $\hat{\beta} = (X'X)^{-1}X'Y$ from FOC
- R^2 , TSS, RSS, adjusted R^2
- Gauss-Markov: BLUE under homoskedasticity

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:

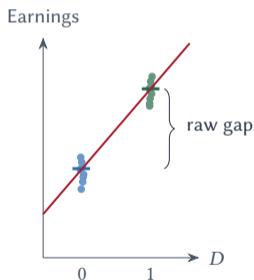
$$Y_i = \beta_0 + \beta_{BLP} D_i + \varepsilon_i, \quad \varepsilon_i = \gamma A_i + u_i$$

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

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

Observational data:

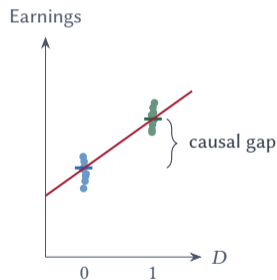
$\mathbb{E}[\varepsilon | D=1] > \mathbb{E}[\varepsilon | D=0] \Rightarrow$ slope too steep



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 β_{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 ε 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: $Y_i = \beta_0 + \beta_1 D_i + \varepsilon_i$

ε_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 β_1 :

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

ε_i is beyond policy's reach.

β_1 is not.

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

Decompose β_{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 β_{BLP}
- But β_{BLP} includes selection
- Right answer to wrong question

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

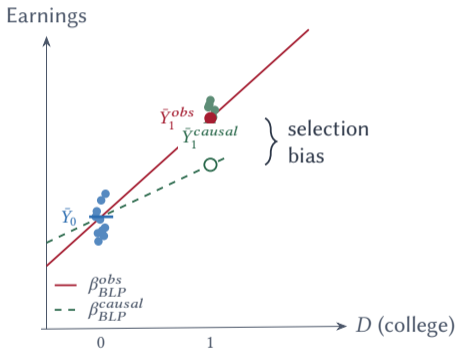
$\beta_{BLP} = \beta_{causal}$

- Selection term vanishes
- OLS hits β_{BLP}
- Which IS the causal effect

OLS is never biased for β_{BLP} . Mean independence makes β_{BLP} worth having.

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

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

OLS converges to β_{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 β_{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 β_{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
 β_{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 β_{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	Standard errors are wrong

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 $\hat{\beta}$:

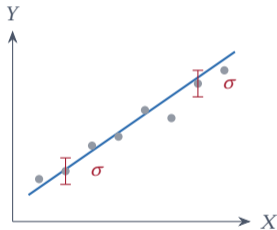
$$\text{Var}(\hat{\beta}) = \sigma^2 (X'X)^{-1}$$

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.



Today's arc – and where we go on Monday

Today.

- **OLS** as the sample plug-in for the BLP: $\hat{\beta} = (X'X)^{-1}X'Y$
- **Unbiased** under strict exogeneity – finite-sample
- **Consistent** under the weaker $\mathbb{E}[X_i\varepsilon_i] = 0$
- **Asymptotically normal** because $X'\varepsilon/n$ is a sample mean – CLT + Slutsky
- **BLUE** (Gauss-Markov) under homoskedasticity – loses “B” the moment errors are heteroskedastic
- **Homoskedastic variance:** $\widehat{\text{Var}}(\hat{\beta}) = \hat{\sigma}^2(X'X)^{-1}$ – one number scales everything

Wrong $\widehat{\text{Var}}(\hat{\beta}) \Rightarrow$ wrong SE \Rightarrow wrong t -stats and CIs. Getting the variance right *is* the inference problem.

Monday: when the homoskedastic formula breaks. Robust SEs (revisited carefully), clustering, $t/F/R^2$, and the practice exam worksheet.