Variance-Weighted Regression

What OLS with controls is actually estimating

Gov 2001 · Scott Cunningham · Spring 2026

Where we are: SE done; today, what the *coefficient* is

Monday: getting $\widehat{\mathrm{Var}}(\hat{\beta})$ right.

- Heteroskedasticity $\Rightarrow$ sandwich estimator (HC1)
- Within-cluster correlation $\Rightarrow$ CRVE (cluster-robust)
- Same OLS estimator $\hat{\beta}$; what changed was the variance formula

Today: getting the *interpretation* of $\hat{\beta}_1$ right.

- Run $Y = \alpha + \beta_1 D + \beta_2 G + \varepsilon$ with D, G binary
- What does $\hat{\beta}_1$ *actually* return as the sample grows?
- Spoiler: not the average treatment effect. Something close, but *not the same*.

Today the SE is fixed. We're going after the point estimate itself.

The question: what is $\hat{\beta}_1$ an average *of*?

You ran the regression. You have a number. What does that number aggregate?

Hope

$\hat{\beta}_1 \approx$ average treatment effect

$= \sum_g P(G = g)\, \tau_g$

Reality

$\hat{\beta}_1 \approx$ *variance*-weighted average

$= \sum_g w_g\, \tau_g, \quad w_g \neq P(G = g)$

Today: derive the weights. See them. Run code that proves the formula.

The decomposition is purely algebraic. No causal assumptions until the very end.

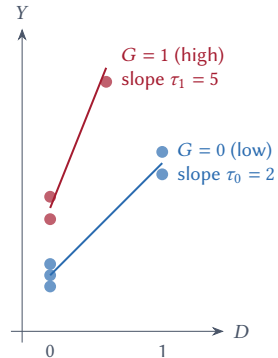
A puzzle: a job-training program with two strata

Setup.

- $D \in \{0, 1\}$: binary *treatment* (took the program / didn't)
- $G \in \{0, 1\}$: binary *baseline education* (0 = low, 1 = high)
- $\tau_g = \mathbb{E}[Y(1) - Y(0) \mid G = g]$: *CATE* within stratum g

The story. Two strata of people. Within each stratum, the program has its own causal effect τ_g. The fraction who actually took the program (*take-up*) may also differ across strata.

The question is purely descriptive: when you run a single OLS regression on the pooled data, what number does $\hat{\beta}_1$ return?



Two strata, two within-group slopes

A brief detour: what is a treatment effect?

Today is *not* a causal inference lecture. But the language of treatment effects is what we use to talk about τ_g — so we owe it a moment.

The potential outcomes framework. Rubin (1974, 1977), back to Neyman (1923):

- $Y_i(1)$: the outcome unit i would have if treated
- $Y_i(0)$: the outcome unit i would have if untreated
- $\tau_i = Y_i(1) - Y_i(0)$: the *individual* treatment effect

This is a definition, not yet an estimator. Causality is defined by contrasts between the same unit's outcomes under two interventions.

The switching equation: only one outcome is ever observed

For each unit i, what we actually observe is:

$$Y_i \;=\; D_i \cdot Y_i(1) \;+\; (1 - D_i) \cdot Y_i(0)$$

- If $D_i = 1$ (you went to college): you observe $Y_i(1)$. The counterfactual $Y_i(0)$ is gone — forever.
- If $D_i = 0$: you observe $Y_i(0)$. $Y_i(1)$ is gone.
- Either way: τ_i is the difference of two numbers, only one of which exists.

This is the fundamental problem of causal inference. The notation directs us toward *identification* of population estimands — not to measurement of individual effects.

So causal estimands are summaries of the unobservable $\{\tau_i\}$ — typically (weighted) averages.

When does OLS coincidentally recover the ATE?

Average treatment effect.

$$\text{ATE} \;=\; \mathbb{E}[\tau_i] \;=\; \mathbb{E}[Y(1)] - \mathbb{E}[Y(0)]$$

Under randomization (or more generally, $D \perp\!\!\!\perp \{Y(0), Y(1)\}$):

- $\mathbb{E}[Y(1) \mid D = 1] = \mathbb{E}[Y(1) \mid D = 0] = \mathbb{E}[Y(1)]$
- Same for $Y(0)$
- The conditional means *become* the marginal means

The BLP coefficient on D then *coincides* with the ATE. OLS reads as causal — by accident of the design.

Without randomization: a different story — today's story. Further reading: Imbens & Rubin (2015); Mostly Harmless Econometrics; *Causal Inference: The Mixtape*; Kosuke Imai's Gov classes.

But what does the *multivariate* BLP equal?

You ran:

$$Y = \alpha + \tau D + \beta X + \varepsilon$$

Does τ equal the ATE? Often no.

- Even with $D \perp\!\!\!\perp \{Y(0), Y(1)\} \mid X$ (selection on observables)
- τ is a *weighted average* of within-stratum treatment effects τ_g
- The weights are not the population weights $\mathbb{P}(G = g)$
- They are *variance weights* — invisible without a tool to see them

FWL is that tool. The rest of today's lecture is the decomposition of τ into the weights and within-stratum effects that produced it.

Back to the puzzle: the joint distribution of (G, D)

Same job-training program, two strata. Here is the full population in a cross-tab:

	$D = 0$	$D = 1$	total: $\mathbb{P}(G)$
$G = 0$ (low)	54%	6%	60%
$G = 1$ (high)	20%	20%	40%
total: $\mathbb{P}(D)$	74%	26%	100%

Reading the cross-tab:

- Cells: joint probability $\mathbb{P}(G = g, D = d)$
- Right margin: stratum sizes $\mathbb{P}(G)$ — 60% low education, 40% high education
- Bottom margin: treatment shares $\mathbb{P}(D)$ — 74% untreated, 26% treated
- The four cells sum to 100%; both margins separately sum to 100%

Conditional take-up rates and CATEs from the cross-tab

Within-stratum take-up $p_g = \mathbb{P}(D = 1 \mid G = g)$, computed from the cells:

$$p_0 = \frac{\mathbb{P}(G = 0,\, D = 1)}{\mathbb{P}(G = 0)} = \frac{6\%}{60\%} = 0.10$$

$$p_1 = \frac{\mathbb{P}(G = 1,\, D = 1)}{\mathbb{P}(G = 1)} = \frac{20\%}{40\%} = 0.50$$

Within-stratum treatment effects: $\tau_0 = 2, \quad \tau_1 = 5.$

Two distinct probabilities. p_0 partitions *within* stratum 0 (10% treated, 90% not). p_1 partitions *within* stratum 1 (50/50). They do *not* sum to 1 across strata.

These five numbers ($\mathbb{P}(G)$, p_0, p_1, τ_0, τ_1) are everything we need to compute both averages on the next slide.

Two natural averages of τ_0 and τ_1 — which does OLS return?

Population-weighted (ATE)

$$\text{ATE} = \sum_g \mathbb{P}(G{=}g)\,\tau_g$$

weights = stratum sizes

$= 0.6 \cdot 2 + 0.4 \cdot 5 = \mathbf{3.20}$

Variance-weighted (OLS)

$$\beta_1 = \sum_g w_g\,\tau_g$$

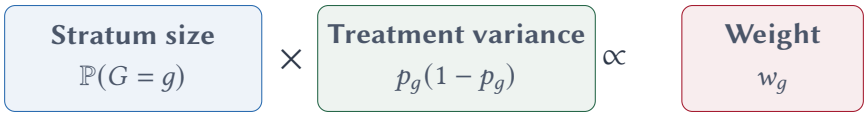
weights $\propto \mathbb{P}(G{=}g)\,p_g(1-p_g)$

$= 0.35 \cdot 2 + 0.65 \cdot 5 \approx \mathbf{3.95}$

Same data. Two different numbers. Gap ≈ 0.75.

The variance weights overweight the smaller stratum because its take-up is closer to 50/50.

Preview the answer: $w_g \propto \mathbb{P}(G{=}g) \cdot p_g(1 - p_g)$



- **Stratum size** matters — but it does *not* go in alone
- **Treatment variance** $p_g(1 - p_g)$ peaks at $p_g = 0.5$
- Strata where D is closer to a coin flip get *more* weight

The same OLS coefficient that worked perfectly for one regressor reweights observations the moment you add a control.

Cochran (1968): the matching tradition — compare within strata

William Cochran, "The Effectiveness of Adjustment by Subclassification in Removing Bias in Observational Studies" (1968).

The matching logic:

- Stratify on covariates
- Compare treated and untreated *within* each stratum
- Aggregate by stratum size

The aggregation step matters. Cochran used *population weights*:

$$\widehat{\text{ATE}} = \sum_g \widehat{\mathbb{P}}(G = g)\, \widehat{\tau}_g$$

Cochran's framework was the foundation for matching estimators in observational studies. It is the *benchmark* you compare regression's weighting scheme against.

Key idea

Within-stratum comparison + population aggregation

Frisch–Waugh (1933) and Lovell (1963): the partialling-out lemma

The result. The coefficient on D in $Y \sim D + G$ equals the coefficient from regressing the *residualized* Y on the *residualized* D — where both are residualized on G.

Algebraically:

$$\beta_1 = \frac{\mathrm{Cov}(\tilde{D}, Y)}{\mathrm{Var}(\tilde{D})}$$

where $\tilde{D} = D - \mathbb{E}[D \mid G]$.

Multiple regression = a sequence of simple regressions on residuals. The slope on D only "sees" the part of D orthogonal to G.

Why this matters today

FWL is the *tool* that lets us decompose $\hat{\beta}_1$ into a sum over strata.

Once we know $\beta_1 = \mathrm{Cov}(\tilde{D}, Y)/\mathrm{Var}(\tilde{D})$, the rest is iterated expectations.

Angrist (1998): FWL + binary D + discrete $G \Rightarrow$ variance weights

Joshua Angrist, "Estimating the Labor Market Impact of Voluntary Military Service Using Social Security Data on Military Applicants" (*Econometrica*, 1998).

- Substantive question: military service and earnings, instrumenting with the draft lottery
- Methodological contribution: a clean derivation of *what regression with discrete controls actually estimates*
- Combines: FWL (mechanics) + Bernoulli variance for binary D (algebra) + iterated expectations

$$\beta_1 = \sum_g w_g \, \tau_g, \qquad w_g = \frac{\mathbb{P}(G = g)\, p_g(1 - p_g)}{\sum_{g'} \mathbb{P}(G = g')\, p_{g'}(1 - p_{g'})}$$

Variance weights ≠ population weights. The deviation is the heart of today's lecture.

Słoczyński (2022): when the variance weights misbehave

Tymon Słoczyński, "Interpreting OLS Estimands When Treatment Effects Are Heterogeneous" (*Review of Economics and Statistics*, 2022).

- The variance weights can put *very* little mass on the larger group — and most of the weight on the smaller one
- In real applications, $\hat{\beta}_1$ can be far from *any* of τ_0, τ_1, ATE
- Diagnostic: report stratum-specific effects, weights, and the ATE alongside the OLS coefficient

> Modern critique: the OLS coefficient is a perfectly precise estimate *of the wrong quantity* when heterogeneity and uneven take-up coexist.

We will see this concretely at the end of the lecture. First: build the formula.

Chattopadhyay & Zubizarreta (2022): the modern generalization

Ambarish Chattopadhyay & José R. Zubizarreta, "On the Implied Weights of Linear Regression for Causal Inference" (*Biometrika*, 2022).

- Today's binary D and discrete G generalizes to *any* covariates
- For an arbitrary regression $Y = \alpha + \beta_1 D + \beta_2' X + \varepsilon$, the OLS coefficient $\hat{\beta}_1$ is a weighted average of unit-level contrasts — with weights you can read off the design
- The framework asks: *how well does linear regression emulate a randomized experiment?*

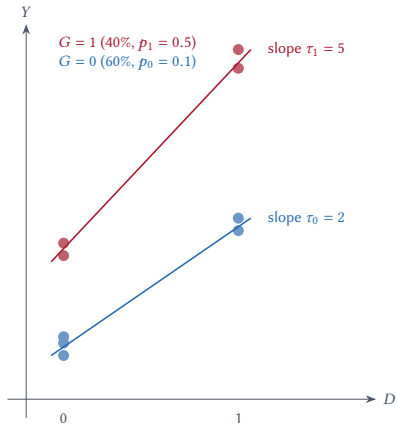
Practical tool: the `lmw` R package (Chattopadhyay, Greifer, Zubizarreta 2024)

`install.packages("lmw")` — computes the implied weights for any linear regression you run.

Use it as a diagnostic: which units are driving your coefficient?

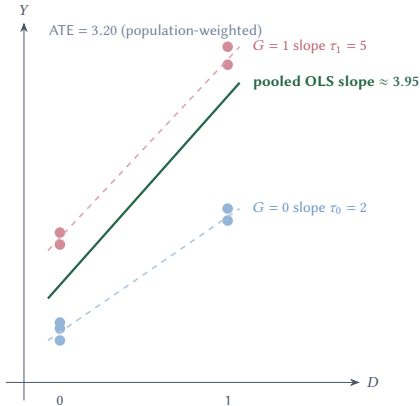
Today shows the simplest case (binary D + binary G). The Biometrika paper shows the same logic governs every linear regression with controls.

Two strata, two within-group slopes — before pooling



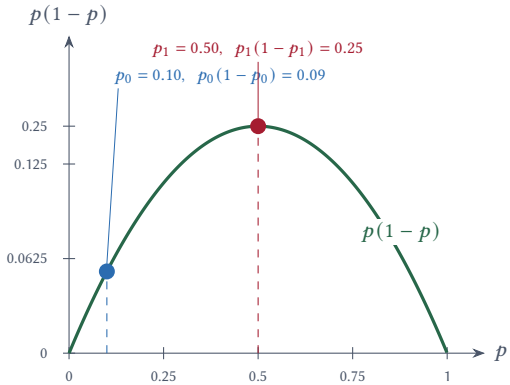
Each stratum has its own treatment effect. They differ. The question is what OLS returns when we run a single regression on the pooled data.

The pooled OLS line: between the two, but not at the average



The pooled OLS line lies between the two within-stratum slopes — but *closer* to the smaller stratum's slope. Why?

Why $p_g(1-p_g)$? Bernoulli variance peaks where treatment is at $50/50$



Reading the curve:

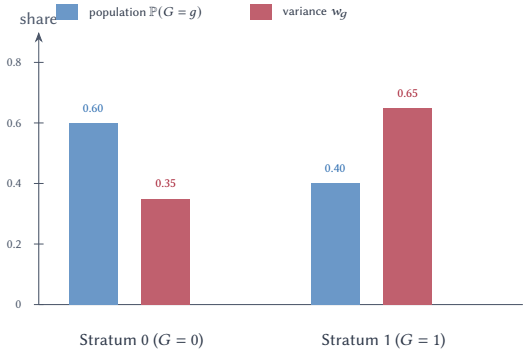
- Peaks at $p = 0.5$, where $p(1-p) = 0.25$
- Falls to zero at $p = 0$ and $p = 1$
- Quadratic, symmetric

Why it matters:

Treatment variance *within* a stratum measures how much D varies there. More variation = more information about τ_g.

Stratum 1 ($p = 0.5$) is $\sim 2.8\times$ as informative as stratum 0 ($p = 0.1$).

Variance weights vs. population weights: the bar chart



The smaller stratum (40% of the population) gets 65% of the OLS weight.

Because its take-up is at 0.5, where Bernoulli variance is maximal.

A toy population: compute ATE and the variance-weighted sum

	$\mathbb{P}(G=g)$	p_g	$p_g(1-p_g)$	τ_g
$G=0$	0.60	0.10	0.09	2
$G=1$	0.40	0.50	0.25	5

ATE (population-weighted)

$$0.60 \cdot 2 + 0.40 \cdot 5$$
$$= 1.20 + 2.00 = \mathbf{3.20}$$

Variance-weighted (OLS)

$\beta_1 = \sum_g w_g \tau_g$ — compute next

Computing the variance weights: numerator, denominator, ratio

Step 1. Numerator of each weight: $\mathbb{P}(G = g)\, p_g(1 - p_g)$

$$G = 0: \quad 0.60 \cdot 0.09 = 0.054 \qquad\qquad G = 1: \quad 0.40 \cdot 0.25 = 0.100$$

Step 2. Denominator: $\sum_g \mathbb{P}(G = g)\, p_g(1 - p_g) = 0.054 + 0.100 = 0.154$

Step 3. Weights: divide each numerator by the denominator

$$w_0 = \frac{0.054}{0.154} \approx 0.351 \qquad w_1 = \frac{0.100}{0.154} \approx 0.649$$

Stratum 1 is 40% of the population but gets $\approx$ 65% of the OLS weight.

Putting it together: the OLS coefficient and the ATE differ

	ATE	OLS (β_1)
Weight on $\tau_0 = 2$	0.60	0.351
Weight on $\tau_1 = 5$	0.40	0.649
Weighted sum	**3.20**	**3.95**

Same stratum-specific effects τ_0, τ_1. Different weights. Different aggregate. Gap = 0.75.

Next: derive these weights from FWL. Then run R code that produces all three numbers — OLS, FWL by hand, and $\sum w_g \tau_g$.

FWL recap: $\beta_1 = \text{Cov}(\tilde{D}, Y)/\text{Var}(\tilde{D})$

Frisch–Waugh–Lovell. For the regression $Y = \alpha + \beta_1 D + \beta_2' G + \varepsilon$:

$$\beta_1 = \frac{\text{Cov}(\tilde{D},\, Y)}{\text{Var}(\tilde{D})}$$

where $\tilde{D} = D - \mathbb{E}[D \mid G]$ is D with the part predicted by G removed.

- $\tilde{D}$ has mean zero (by construction)
- $\tilde{D}$ is uncorrelated with G
- Multiple regression on D, G = simple regression of Y on $\tilde{D}$

The whole derivation today rides on this one equation.

FWL step 1: regress D on G to get the residualized treatment $\tilde{D}$

The auxiliary regression:

$$D_i = \alpha + \gamma G_i + u_i$$

Population OLS gives:

- $\alpha = \mathbb{E}[D \mid G = 0] = p_0$ (intercept = mean of D in stratum 0)
- $\alpha + \gamma = \mathbb{E}[D \mid G = 1] = p_1$ (intercept + slope = mean of D in stratum 1)

Therefore the fitted value at observation i is the within-stratum mean of D:

$$\widehat{D}_i = \mathbb{E}[D \mid G_i] = p_{G_i} \qquad \text{and} \qquad \tilde{D}_i = D_i - \widehat{D}_i = D_i - p_{G_i}$$

Residualizing D on G = subtracting the within-stratum mean of D from each observation.

$\tilde{D}$ takes exactly two values within each stratum

For each unit in stratum $G = g$, $\tilde{D}_i = D_i - p_g$ takes one of two values:

- If $D_i = 1$: $\tilde{D}_i = 1 - p_g$ (positive)
- If $D_i = 0$: $\tilde{D}_i = 0 - p_g = -p_g$ (negative)

In the running example ($p_0 = 0.1$, $p_1 = 0.5$):

Stratum	p_g	$\tilde{D}$ if $D = 0$	$\tilde{D}$ if $D = 1$	probability of each
$G = 0$ (low)	0.10	−0.10	+0.90	0.90 / 0.10
$G = 1$ (high)	0.50	−0.50	+0.50	0.50 / 0.50

Within each stratum, $\tilde{D}$ has mean zero by construction. So $\mathbb{E}[\tilde{D} \mid G] \equiv 0$, and by the tower property $\mathbb{E}[\tilde{D}] = 0$.

Conditional variance of $\tilde{D}$: the Bernoulli variance falls out

Because $\mathbb{E}[\tilde{D} \mid G = g] = 0$, the conditional variance equals the second moment:

$$\text{Var}(\tilde{D} \mid G = g) \;=\; \mathbb{E}[\tilde{D}^2 \mid G = g]$$

Two outcomes per stratum:

- With probability p_g: $\tilde{D} = 1 - p_g$, so $\tilde{D}^2 = (1 - p_g)^2$
- With probability $1 - p_g$: $\tilde{D} = -p_g$, so $\tilde{D}^2 = p_g^2$

Putting it together:

$$\begin{aligned}\mathbb{E}[\tilde{D}^2 \mid G = g] &= p_g\,(1 - p_g)^2 \;+\; (1 - p_g)\,p_g^2 \\ &= p_g(1 - p_g)\big[(1 - p_g) + p_g\big] \\ &= p_g(1 - p_g)\end{aligned}$$

$\text{Var}(\tilde{D} \mid G = g) = p_g(1 - p_g)$ — the Bernoulli variance, exactly the curve from slide 18.

Compute $\mathrm{Var}(\tilde{D})$: average the within-stratum variances over G

Law of total variance:

$$\mathrm{Var}(\tilde{D}) \;=\; \mathbb{E}\big[\mathrm{Var}(\tilde{D} \mid G)\big] \;+\; \mathrm{Var}\big(\mathbb{E}[\tilde{D} \mid G]\big)$$

But $\mathbb{E}[\tilde{D} \mid G] \equiv 0$ (we showed this on slide 24b), so the second term is zero:

$$\mathrm{Var}(\tilde{D}) \;=\; \mathbb{E}\big[\mathrm{Var}(\tilde{D} \mid G)\big] \;=\; \sum_{g} \mathbb{P}(G = g)\, p_g(1 - p_g)$$

In the running example:

$$\begin{aligned}
\mathrm{Var}(\tilde{D}) &= \mathbb{P}(G = 0)\, p_0(1 - p_0) \;+\; \mathbb{P}(G = 1)\, p_1(1 - p_1) \\
&= 0.60 \cdot 0.09 \;+\; 0.40 \cdot 0.25 \\
&= 0.054 \;+\; 0.100 \;=\; \mathbf{0.154}
\end{aligned}$$

$\mathrm{Var}(\tilde{D})$ is the population-weighted average of within-stratum Bernoulli variances.

Compute $\mathrm{Cov}(\tilde{D}, Y)$: condition on G first

Since $\mathbb{E}[\tilde{D}] = 0$:

$$\mathrm{Cov}(\tilde{D}, Y) \;=\; \mathbb{E}[\tilde{D}\,Y] \;-\; \mathbb{E}[\tilde{D}]\;\mathbb{E}[Y] \;=\; \mathbb{E}[\tilde{D}\,Y]$$

By iterated expectations, condition on G first:

$$\mathbb{E}[\tilde{D}\,Y] \;=\; \mathbb{E}\big[\mathbb{E}[\tilde{D}\,Y \mid G]\big]$$

The trick is to look *within* a stratum first — where $\tilde{D}$ takes only two values — then average over G.

Same iterated-expectations move you saw in 04b_cef. Here it pays off because $\tilde{D}$ has a closed-form distribution within each stratum.

Within stratum $G = g$: $\mathbb{E}[\tilde{D}\, Y \mid G = g] = p_g(1 - p_g)\, \tau_g$

Within stratum g, $\tilde{D}$ takes two values, each with known probability:

- With probability p_g: $\tilde{D} = 1 - p_g$, $\quad Y = \mathbb{E}[Y \mid D = 1, G = g]$
- With probability $1 - p_g$: $\tilde{D} = -p_g$, $\quad Y = \mathbb{E}[Y \mid D = 0, G = g]$

$$
\begin{aligned}
\mathbb{E}[\tilde{D}\, Y \mid G = g] &= p_g \cdot (1 - p_g)\ \mathbb{E}[Y \mid D = 1, G = g] \ + \ (1 - p_g) \cdot (-p_g)\ \mathbb{E}[Y \mid D = 0, G = g] \\
&= p_g(1 - p_g) \Big(\mathbb{E}[Y \mid D = 1, G = g] \ - \ \mathbb{E}[Y \mid D = 0, G = g] \Big) \\
&= p_g(1 - p_g)\, \tau_g
\end{aligned}
$$

The Bernoulli variance $p_g(1 - p_g)$ multiplies the within-stratum effect τ_g. That product becomes the building block of the formula.

Take the outer expectation: $\mathrm{Cov}(\tilde{D}, Y) = \sum_g \mathbb{P}(G = g)\, p_g(1 - p_g)\, \tau_g$

Now average the within-stratum quantity over G:

$$\mathrm{Cov}(\tilde{D}, Y) \;=\; \mathbb{E}\Big[\mathbb{E}[\tilde{D}\, Y \mid G]\Big] \;=\; \sum_g \mathbb{P}(G = g)\, p_g(1 - p_g)\, \tau_g$$

In the running example ($\tau_0 = 2$, $\tau_1 = 5$):

$$\begin{aligned}
\mathrm{Cov}(\tilde{D}, Y) &= 0.60 \cdot 0.09 \cdot 2 \;+\; 0.40 \cdot 0.25 \cdot 5 \\
&= 0.054 \cdot 2 \;+\; 0.100 \cdot 5 \\
&= 0.108 \;+\; 0.500 \;=\; \mathbf{0.608}
\end{aligned}$$

Each stratum contributes $\mathbb{P}(G = g) \cdot p_g(1 - p_g) \cdot \tau_g$. Sum over strata.

The formula: $\beta_1 = \sum_g w_g \, \tau_g$ with variance weights

Numerator (just derived):

$$\mathrm{Cov}(\tilde{D}, Y) = \sum_g \mathbb{P}(G = g)\, p_g(1 - p_g)\, \tau_g$$

Denominator:

$$\mathrm{Var}(\tilde{D}) = \mathbb{E}\big[\mathrm{Var}(D \mid G)\big] = \sum_g \mathbb{P}(G = g)\, p_g(1 - p_g)$$

Ratio:

$$\beta_1 = \frac{\sum_g \mathbb{P}(G = g)\, p_g(1 - p_g)\, \tau_g}{\sum_g \mathbb{P}(G = g)\, p_g(1 - p_g)} = \sum_g w_g \, \tau_g$$

$$w_g = \frac{\mathbb{P}(G = g)\, p_g(1 - p_g)}{\sum_{g'} \mathbb{P}(G = g')\, p_{g'}(1 - p_{g'})} \qquad \text{weights nonnegative; sum to one}$$

R: simulate a population matching the toy example

```r
# Population parameters (matching the toy example)
set.seed(2026)
n     <- 50000
mu_G  <- 0.4                # P(G = 1)
p0    <- 0.1                # P(D = 1 | G = 0)
p1    <- 0.5                # P(D = 1 | G = 1)
tau0  <- 2                  # within-stratum effect, G = 0
tau1  <- 5                  # within-stratum effect, G = 1

# Generate data
G  <- rbinom(n, 1, mu_G)
pG <- ifelse(G == 1, p1, p0)
D  <- rbinom(n, 1, pG)
tauG <- ifelse(G == 1, tau1, tau0)
Y  <- D * tauG + rnorm(n, 0, 1)
```

$n = 50{,}000$ is generous; we want sample analogs to be close to population values.

R: run OLS — this is the number we want to decompose

```
fit_ols <- lm(Y ~ D + G)
beta_OLS <- coef(fit_ols)["D"]
beta_OLS
#>         D
#> 3.926...
```

$\hat{\beta}_1^{\text{OLS}} \approx 3.93$. Now: derive the same number two other ways.

Path 1 just ran. Path 2 = FWL by hand. Path 3 = $\sum_g \hat{w}_g \hat{\tau}_g$. All three should agree.

R: FWL by hand — residualize, then regress

```r
# Step 1: residualize D on G  (compute D - E[D|G])
fit_D <- lm(D ~ G)
D_tilde <- residuals(fit_D)

# Step 2: residualize Y on G  (compute Y - E[Y|G])
fit_Y <- lm(Y ~ G)
Y_tilde <- residuals(fit_Y)

# Step 3: regress Y_tilde on D_tilde (no intercept needed)
fit_FWL <- lm(Y_tilde ~ D_tilde - 1)
beta_FWL <- coef(fit_FWL)["D_tilde"]
beta_FWL
#> D_tilde
#>   3.926...
```

Same as $\hat{\beta}_1^{\text{OLS}}$. FWL is not just a theorem; it works on the data.

R: compute the variance weights $\hat{w}_g$ from the sample

```
# Sample analogs of population quantities
mu_hat <- mean(G)                  # P(G = 1) ~ 0.40
p_hat  <- tapply(D, G, mean)       # p_0, p_1 (~ 0.10, 0.50)

# Numerators of each weight: P(G=g) * p_g * (1 - p_g)
prG <- c(`0` = 1 - mu_hat, `1` = mu_hat)
num <- prG * p_hat * (1 - p_hat)

# Weights = numerators / sum of numerators
w_hat <- num / sum(num)
round(w_hat, 3)
#>   0     1
#> 0.352 0.648
```

Stratum 0 is 60% of the sample but gets $\approx$ 35% of the OLS weight; stratum 1 (40%) gets $\approx$ 65%.

R: compute the within-stratum effects $\hat{\tau}_g$

```r
# Within stratum: mean of Y for D=1 minus mean of Y for D=0
means <- tapply(Y, list(D, G), mean)
means
#>       0           1
#> 0    0.000...    0.011...
#> 1    1.981...    4.994...

tau_hat <- means["1", ] - means["0", ]
round(tau_hat, 3)
#>     0     1
#> 1.981 4.983
```

The within-stratum effects are essentially the population $\tau_0 = 2, \tau_1 = 5$ — as they should be at $n = 50{,}000$.

R: $\sum_g \hat{w}_g \hat{\tau}_g$ — and the three numbers agree

```
beta_VW <- sum(w_hat * tau_hat)

# Three paths to the same number:
c(OLS    = beta_OLS,
  FWL    = beta_FWL,
  VarWtd = beta_VW)
#>     OLS      FWL   VarWtd
#> 3.92576  3.92576  3.92576

# Compare to ATE = sum(P(G=g) * tau_g)
ATE <- sum(prG * tau_hat)
ATE
#> 3.191...
```

Three different computations. One number ≈ 3.93. The ATE is a different number ≈ 3.19.

When OLS ≠ ATE: the gap requires *both* kinds of variation

The gap $\beta_1^{\text{OLS}} - \text{ATE}$ is zero in two cases:

Constant treatment effect

$\tau_g = \tau$ for all g

$\beta_1 = \sum_g w_g \tau = \tau = \text{ATE}$

Constant take-up

$p_g = p$ for all g

$w_g = \mathbb{P}(G{=}g)$, so $\beta_1 = \text{ATE}$

The gap is non-zero when τ_g varies *and* p_g varies. Both are needed.

Randomized experiments with $p_g = 0.5$ everywhere: gap is zero. Observational data with selection: gap is generally non-zero.

Słoczyński's diagnostic: when one stratum dominates the weights

The pathological case. A small stratum where treatment is at 0.5 can capture most of the OLS weight. If its τ differs from the rest, $\hat{\beta}_1$ can be far from any sensible average.

- Example: 90% of population has $p = 0.95$; 10% has $p = 0.5$
- $p_{0.95}(1-p_{0.95}) = 0.0475$ vs $p_{0.5}(1-p_{0.5}) = 0.25$
- Numerators: $0.90 \cdot 0.0475 = 0.0428$ vs $0.10 \cdot 0.25 = 0.0250$
- Weights: **63%** on the 90% group, **37%** on the 10% group
- OLS still tilted toward the small group — but less extreme than our toy example

The OLS coefficient is a very precise estimate. Of *what* is the question.

What to report alongside $\hat{\beta}_1$ when controls are discrete

The minimum honest output:

- $\hat{\beta}_1$ from OLS, with its (robust or clustered) SE
- Stratum-specific effects $\hat{\tau}_g$ for each level of the discrete control
- Stratum-specific weights $\hat{w}_g$ — so the reader can see which strata are driving the average
- ATE as a benchmark: $\sum_g \widehat{\mathbb{P}}(G = g)\, \hat{\tau}_g$

> If $\hat{\beta}_1$ and ATE agree closely, you can report just $\hat{\beta}_1$. If they disagree materially, report both — and be explicit about which one answers your research question.

For continuous controls the same logic generalizes; the weights become harder to compute but the principle is identical.

The decomposition is algebraic. Causal interpretation needs more.

What we proved today. An algebraic identity: $\beta_1 = \sum_g w_g \, \tau_g$ where $\tau_g = \mathbb{E}[Y \mid D = 1, G = g] - \mathbb{E}[Y \mid D = 0, G = g]$.

What we did *not* prove. That τ_g is a causal effect.

- τ_g is a difference in conditional means — a population quantity
- It equals the within-stratum causal effect only under *conditional unconfoundedness*: $D \perp (Y_0, Y_1) \mid G$
- With unconfoundedness, $\hat{\beta}_1 \to \sum_g w_g \, \text{ATE}_g$ — a variance-weighted causal effect
- Without unconfoundedness, $\hat{\beta}_1$ is still well-defined but is not a causal quantity

> Algebra is unconditional. Causal interpretation is an extra layer of assumption.

Closing the loop: the asymptotic distribution of $\hat{\beta}_1$

Under Gauss-Markov with homoskedastic errors $\text{Var}(\varepsilon \mid D, G) = \sigma^2$:

$$\sqrt{n}\,(\hat{\beta}_1 - \beta_1) \xrightarrow{d} \mathcal{N}\left(0,\ \frac{\sigma^2}{\text{Var}(\tilde{D})}\right)$$

Substituting today's expression for $\text{Var}(\tilde{D})$**:**

$$\sqrt{n}\,(\hat{\beta}_1 - \beta_1) \xrightarrow{d} \mathcal{N}\left(0,\ \frac{\sigma^2}{(1-\mu)\,p_0(1-p_0) \ +\ \mu\,p_1(1-p_1)}\right)$$

Numerator from 14a (Gauss-Markov + homoskedastic OLS variance).
Denominator from today (variance-weighted decomposition).

Reuse the $\text{Var}(\tilde{D})$ you derived earlier — do not re-derive it from scratch.

Today's arc: three numbers, one decomposition

1. **The puzzle.** OLS with binary D and discrete G returns a number. Which number?
2. **Cochran → FWL → Angrist.** Within-stratum comparison aggregated by population is the matching benchmark. FWL converts multiple regression into a sequence of simple regressions on residuals. Angrist combined the two for binary D.
3. **The formula.** $\beta_1 = \sum_g w_g \tau_g$ with $w_g \propto \mathbb{P}(G = g)\, p_g(1 - p_g)$.
4. **Variance weights.** Strata where treatment is closer to 50/50 carry more weight. Smaller strata can dominate.
5. **R verified.** OLS = FWL by hand = $\sum_g \hat{w}_g \hat{\tau}_g$. Three paths, one number.
6. **ATE ≠ OLS** when τ_g varies and p_g varies. Report both.

This is exactly Practice Exam Problem 5

From the practice exam:

- *Compute* $\mathbb{E}[\tilde{D}]$ *and* $\mathrm{Var}(\tilde{D})$ *in terms of* μ, p_0, p_1 $\Leftrightarrow$ today's denominator
- *Show* $\mathrm{Cov}(\tilde{D}, Y) = (1-\mu)\, p_0(1-p_0)\, \tau_0 + \mu\, p_1(1-p_1)\, \tau_1$ $\Leftrightarrow$ today's numerator
- *Use FWL to write* $\beta_1 = w_0\tau_0 + w_1\tau_1$ $\Leftrightarrow$ today's formula
- *Asymptotic distribution of* $\sqrt{n}(\hat{\beta}_1 - \beta_1)$ $\Leftrightarrow$ Gauss–Markov from 14a + $\mathrm{Var}(\tilde{D})$ from today

Today's lecture is the *derivation* for the technique you have already practiced. The exam asks you to execute the steps; today shows you why those steps work.

Final exam: what to expect and how to prepare

The exam.

- Five problems, 100 points
- Asymptotics + plug-in + delta (P1)
- CLT vs. Chebyshev CIs (P2)
- Conceptual T/F (P3)
- Matrix OLS + FWL (P4)
- Variance weights (P5) — today

Best preparation.

- Practice exam: 15 problems in 5 scaffolded parts
- Do the c-problems first, closed-book and timed
- Companion Shiny app at `scunning.com/.../ols_weights`
- Solutions to the practice exam will be posted later this week

Pattern recognition over memorization. Recognize the family; execute the steps.

Closing thought: ask what your number is an average *of*

Run the regression. Then ask:

What is that number an average *of*?

Today gave you the answer for the simplest case — binary treatment, discrete control.

The same logic generalizes to continuous controls and multi-valued treatments.

It is the foundation of the modern critique of OLS as a causal-effects machine.

Thanks for a good semester. See you in section and at the review.