Experiments: Part II

PS200C Quant Methods in Politics, Spring 2026

Chad Hazlett

UCLA

chazlett@ucla.edu

Where we are

- Part I: random assignment makes the simple difference in means an unbiased estimator of the ATE.
- Part II: Be better than **unbiased**, be closer to correct in your sample, using design and analysis levers.
- Roadmap:
  1. Why unbiasedness isn't the whole story
  2. Covariate adjustment (precision, not bias)
  3. Blocking (better-than-random balance by design)
  4. Non-independence: dependent residuals and clustered assignment
  5. Wrap and takeaways

Don't stop at unbiased

- Under random assignment, $\hat{\tau}_{DIM} = \bar{Y}_T - \bar{Y}_C$ is unbiased for the ATE. We proved this last time.
- But "unbiased" is a statement about *averages over hypothetical repetitions*. In any *single* experiment, $\hat{\tau}$ can land far from τ.
- What we actually care about is how far you are from the truth in your experiment.
- As a property of a procedure, we look at **root mean squared error**:

$$\text{RMSE}(\hat{\tau}) = \sqrt{\mathbb{E}[(\hat{\tau} - \tau)^2]} = \sqrt{\text{Var}(\hat{\tau}) + \text{Bias}(\hat{\tau})^2}$$

- For an unbiased estimator, RMSE = $\sqrt{\text{Var}\hat{\tau}} = SD$, so the goal is to "minmize variance".

Two different "variances"

It is easy to conflate two distinct things, both called "variance."

- **Actual sampling variance.** $\mathrm{Var}(\hat{\tau})$ across hypothetical repetitions of the experiment. This is real uncertainty and we want it to be small.

- **Estimated variance / standard error.** A number we compute from our *one* dataset, hoping it accurately reflects the actual sampling variance. We need this to be roughly correct so that confidence intervals cover at the right rate.

We care about both but for different reasons.

Running example: a GOTV experiment

- **Unit:** a registered voter.
- **Treatment** D**:** a get-out-the-vote contact (mailer, canvass, etc.) vs. no contact.
- **Outcome** Y**:** did they vote? (0/1)
- **Covariates** X**:**
  - `past_turnout`: an index of voting in the last several elections. Strong predictor of Y. (This is a pre-treatment proxy for $Y(0)$.)
  - `age`, `partisan strength`: weaker predictors.
- **Structure:** voters are nested in households (4 per household) and households in precincts (10 per precinct). 100 precincts, $N = 4000$.
- **True effect:** $\tau = +0.05$ on the probability of voting.
- **We will estimate** $\hat{\tau}$ **many times under different designs and analyses**, and compare the resulting distributions.

What does a single experiment look like? Check balance.

We randomly assign half the sample to treatment. Then look at the covariate means by treatment status (a "balance table"):

Variable	Treated	Control	Difference
past_turnout	-0.028	-0.001	-0.027
age	-0.015	-0.016	+0.002
partisan	-0.008	-0.006	-0.002

Randomization ensures:

- Balance *in expectation*, not in your sample.
- There is a gap, $\tau - \hat{\tau}$, in your sample, even though $\mathbb{E}[\tau - \hat{\tau}] = 0$.

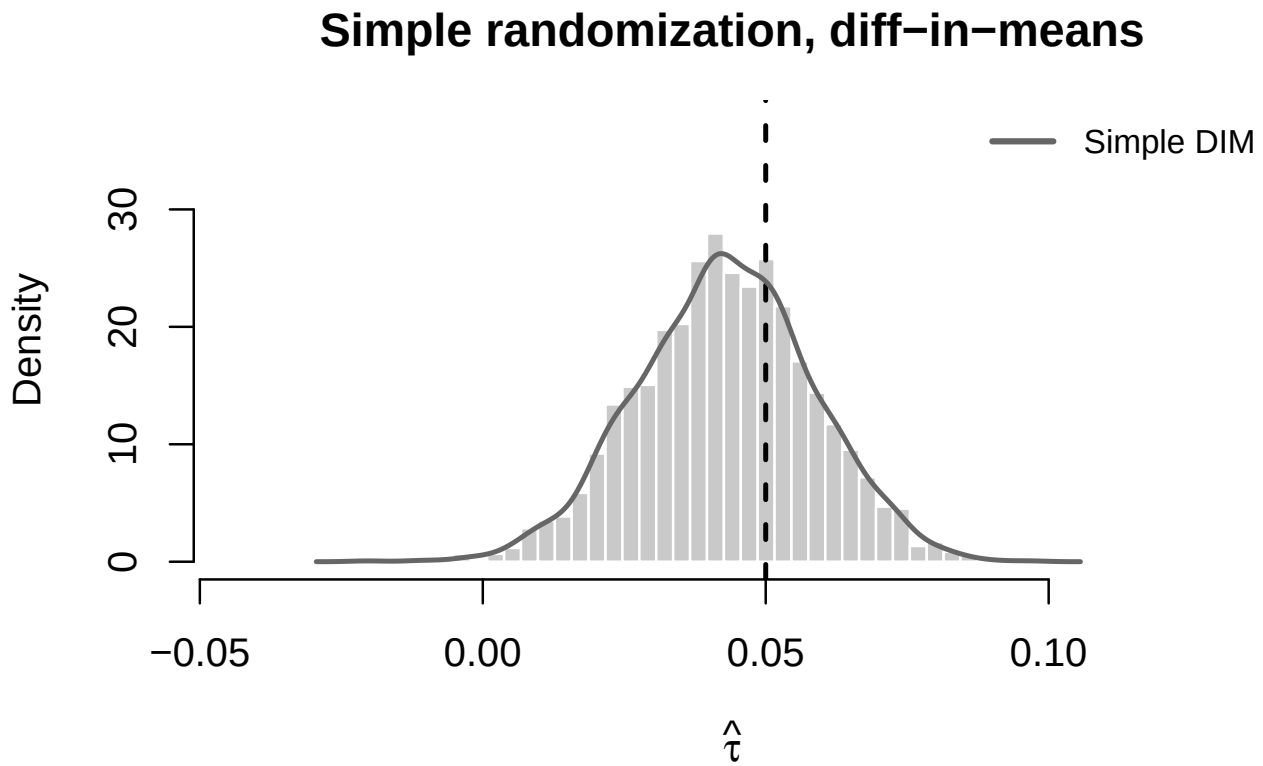
Imbalance on (prognostic) covariates surface reasons to worry about (and opportunities to reduce) this gap.

Aside: Y is binary. Can we use diff-in-means / OLS?

- Yes. ATE is a difference of two probabilities, DIM estimates it.
- `lm(Y ~ D)` returns the same point estimate as the DIM.
- `lm(Y ~ D + X)` (the linear probability model) is still unbiased for the ATE under random assignment.
- Logit/probit are fine too — but their coefficients are *not* the ATE; require computing marginal effects to recover the ATE.
- So: Don't let the binary outcome scare you away from OLS.

Simulating the simple design

2000 replications. Each one: assign D, draw Y, run `lm(Y ~ D)`.



- **Actual SD** of $\hat{\tau}$ across reps: 0.016.
- **Estimated SEs** get HC2 SE fro `sandwich`; on average 0.016.

We have real uncertainty, but the estimated SE knows it.

Covariate adjustment: motivation

- A pre-treatment X that predicts Y ("prognostic"), if imbalanced, influences $\tau - \hat{\tau}$.
- "Adjusting" for this X removes its contribution to this gap.
- Thus we are adjusting to reduce variance (gaps) *not* bias.
- Strong predictors (`past_turnout`) help a lot. Weak ones (e.g. `partisan strength`) barely help, but if you have enough sample, okay.

Adjustment "perfects" the balance the regression sees

Why does OLS with X work?

- One way to think of it: OLS splits D into the part predicted by X and the part orthogonal to X, and uses only the orthogonal part to estimate τ.
- Then the "effective" treatment indicator the regression uses is *exactly balanced* on X; the chance imbalance from the previous slide is gone.
- So whatever variance was driven by that imbalance also goes away.
- (Formal version — the Frisch-Waugh-Lovell theorem — in the appendix.)

Better still: The Lin Estimator

The simplest covariate-adjusted estimator is just:

$$Y_i = \alpha + \tau D_i + \beta' X_i + \varepsilon_i \qquad \text{(plain OLS, fine in large samples)}$$

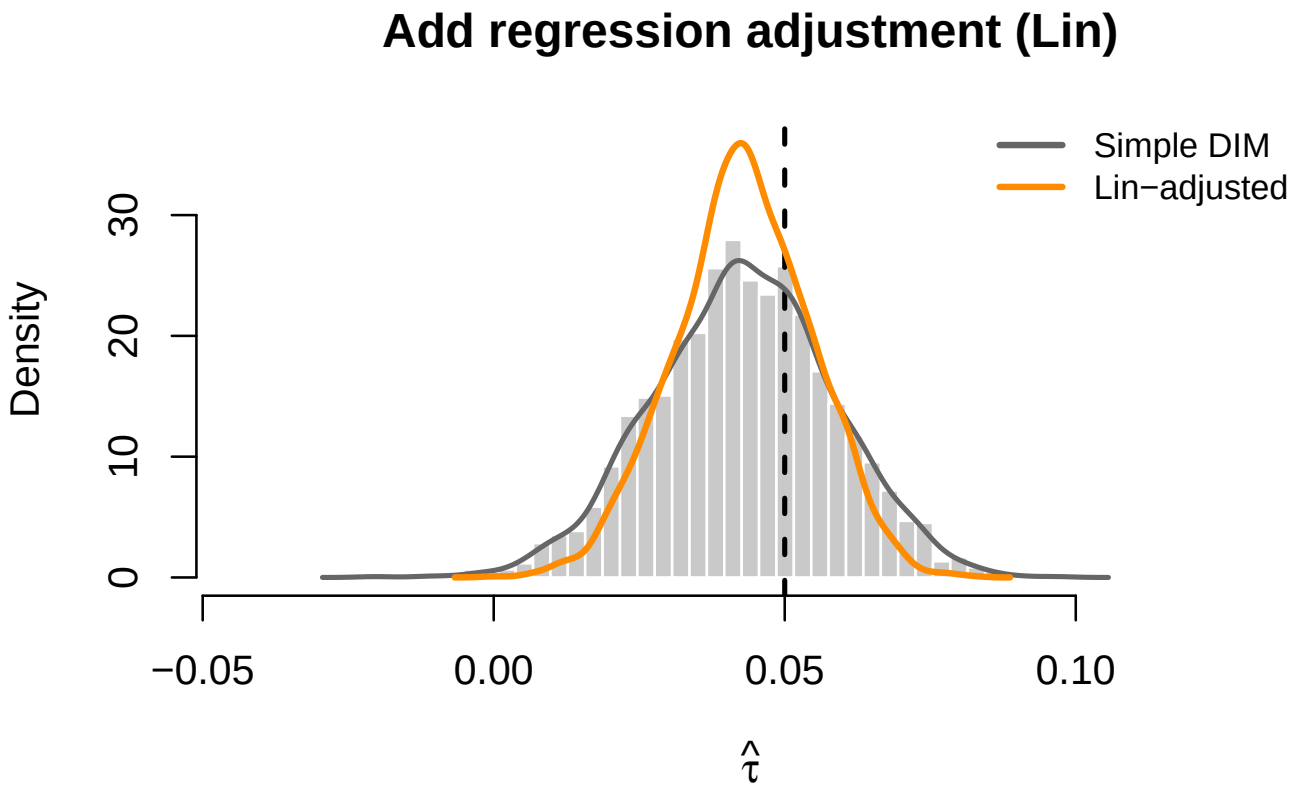
Lin (2013) showed a small modification is better:

$$Y_i = \alpha + \tau_{Lin} D_i + \beta_1'(X_i - \bar{X}) + \beta_2' D_i \cdot (X_i - \bar{X}) + \varepsilon_i$$

- Demean X, then interact with D.
- τ_{Lin} is interpretable directly as an ATE.
- Removes a small ($O(1/N)$) bias from plain regression adjustment that Freedman (2008) flagged; deals with "weird weights under heterogenous effects" we will get to later.
- Use as your default, pre-specify the covariates.

(See appendix for ample details)

Simulation: + Lin-adjusted



- The orange curve is the Lin-adjusted estimator. Same center, ∼25% tighter.
- SD(actual) = 0.012 (was 0.016 without adjustment)
- Mean HC2 SE = 0.012 (was 0.016 without adjustment)

Blocking: build the balance in, don't hope for it

- Adjustment fixes chance imbalances *after* they happen, in the analysis.
- Blocking prevents them *before* they happen, in the design:
  - Pre-stratify on a strong predictor (e.g., quartiles of `past_turnout`).
  - Randomize treatment *within* each stratum (block)
  - Obtains perfect balance on that feature, in your sample.
- The randomization is still doing its job balancing unobserved influences, in expectation.
- "Block on what you can, adjust for what you can't."

Blocking: the estimator

Within each block j, compute the block-specific diff-in-means $\hat{\tau}_j$. Then weight by block size:

$$\hat{\tau}_B = \sum_{j=1}^{J} \frac{N_j}{N} \hat{\tau}_j$$

Equivalently, when treatment probability is the same in every block, we can run OLS with block fixed effects:

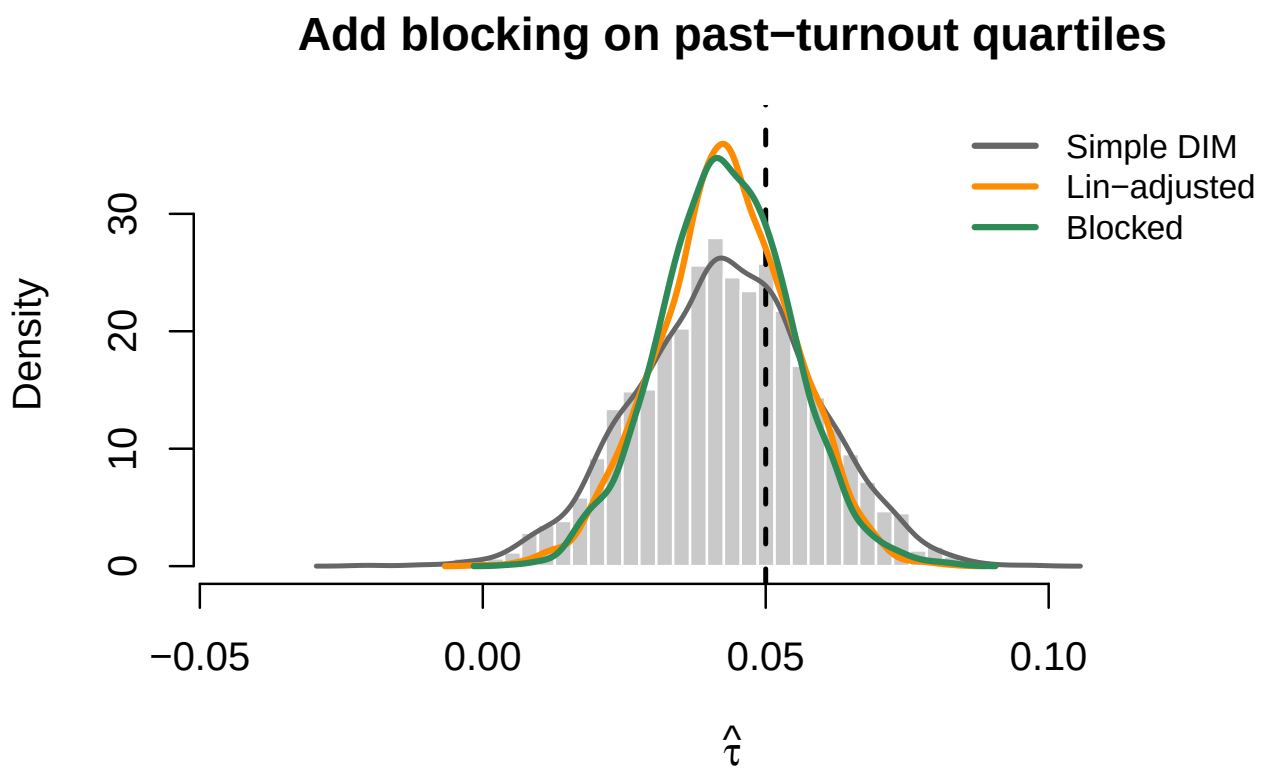
$$Y_i = \alpha + \tau D_i + \sum_{j=2}^{J} \beta_j B_{j[i]} + \varepsilon_i$$

If treatment probability is not equal in all blocks, the Lin-style interacted regression, IPW, and the block-DIM average all recover the ATE and are numerically identical (Shinkre & Hazlett 2025).

Simulation: + blocking

A great choice for blocking: the pre-treatment outcome!

- Here, block on quartiles of `past_turnout`.



- Green = dist. of $\hat{\tau}$ from blocking
- Matches Lin here, not always the case.
- SD(actual) = 0.012; Mean HC2 SE = 0.012

When does blocking help most? Intuition

- Blocking helps a lot when the blocking variable *strongly predicts* Y. Then within-block variation in Y is small, so the diff-in-means within each block is sharp.
- Blocking on something unrelated to Y is free except DoF cost, but useless.
- In our sim, strong gain because cor(post_turnout,turnout) ≈ 0.8.
- Formal variance comparison in the appendix.

Canonical example: Gerber, Green, & Larimer (2008)

- Experiment on what can influence voter turnout
- Several mailings, including one revealing each voter's past turnout to their neighbors.
- **Design:**
  - Blocked on household composition and prior vote history.
- **Result:** the "neighbors" treatment moved turnout by ≈ 8 percentage points — enormous for a piece of mail.
- Relevance here:
  - The blocking is powerful – shows importance of good design.
  - Randomization is clustered at household level – we will talk about its implications next.

Clustering: Two problems from non-independence

So far we've quietly assumed observations are (conditionally) independent.

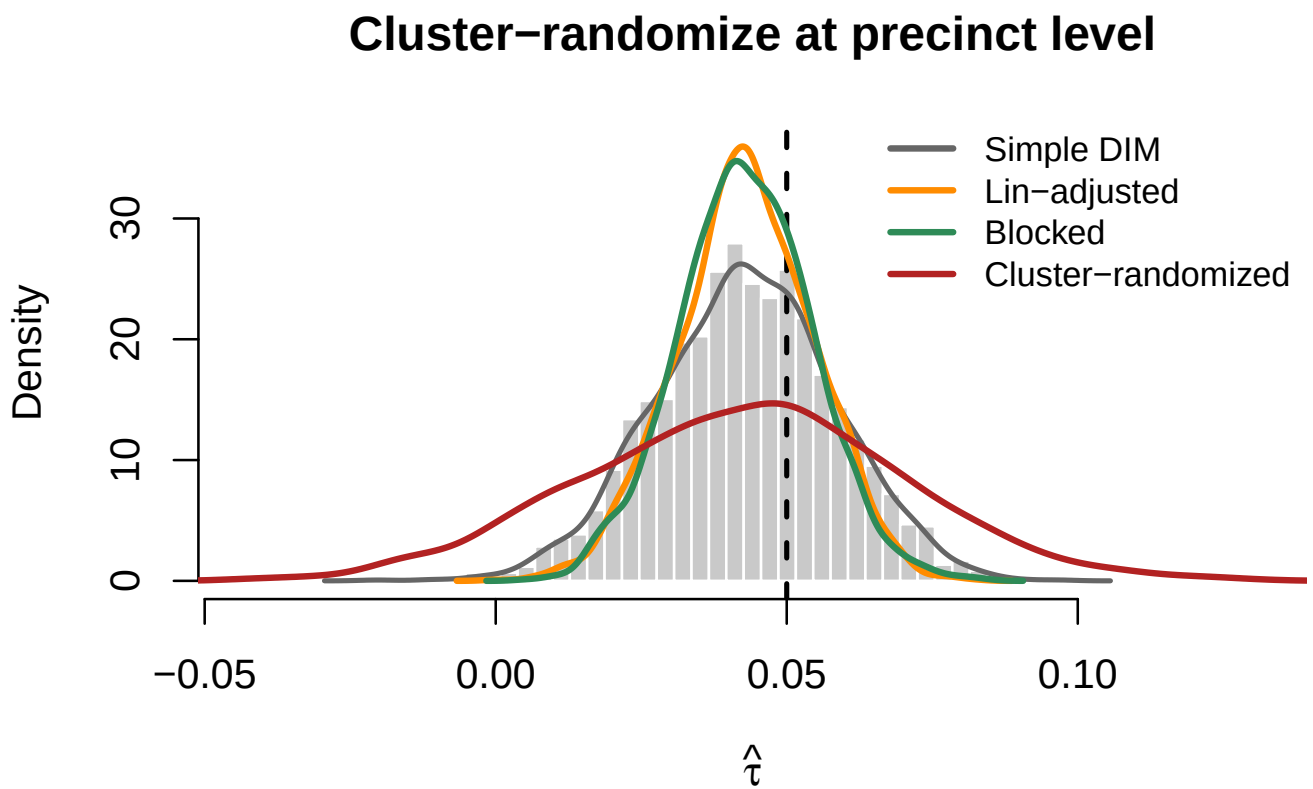
- Real settings often violate this — voters in the same precinct, students in the same classroom, repeated measures on the same person.

Two different things called "clustering" can be trouble:

1. **Clustered assignment/ cluster-randomization**. Sometimes necessary to assigne treatment at group level (precinct, village, school). This is bad news for actual variance/noise and may mess up estimated SE relative to that.
2. **Dependent residuals within clusters**. Even with individual assignment, residuals can be correlated within a group (e.g. panel data); estimated SEs that ignore this can be wrong, usually too small.

Simulation: cluster-randomize at the precinct level

Same DGP but now everyone in a precinct gets the same D.



Eek! Red curve shows you get much bigger gaps (high variance)

And the (HC2) SE is wrong...but CR-SE fixes it.

Estimator	SD(actual)	HC2 SE	CR-SE (precinct)
Simple DIM (`lm(Y ~ D)`)	0.016	0.016	0.016
Lin-adjusted	0.012	0.012	—
Blocked (block FEs)	0.012	0.012	—
Cluster-randomized	**0.028**	**0.016**	**0.028**

- HC2 assumes independent residuals but cluster-influences + cluster-assignment badly violates this.
- Cluster-robust SE (clustered at the precinct level) recovers the actual sampling variance almost exactly. In R:

```
sandwich::vcovCL(fit, cluster = precinct_id).
```

When do you actually need CR-SEs? The principle

You need cluster-robust SEs when there is (i) a source of common influence on outcomes within a cluster, *and* (ii) that influence is not canceled out by treatment variation within the cluster.

1. A cluster-shared component in Y, call it U_c – a precinct effect, a school's teacher quality, a household's news diet, a person's baseline level in panel data).
2. A failure of within-cluster T/C variation to cancel that component out of the residuals.

Either ingredient *alone* is harmless (for the SE *estimation*)

- No cluster-shared component $\Rightarrow$ residuals are iid $\Rightarrow$ HC2 is fine, even if cluster-randomized.
- Cluster-shared component, but treatment varies within cluster $\Rightarrow$ the diff-in-means cancels u_c $\Rightarrow$ HC2 is fine.

Both at once is the dangerous case. Cluster-level influences on Y and cluster-assignment – our case above.

Where “both at once” shows up in practice

- **Cluster-randomized treatment** (precincts, villages, schools, clinics). Same D for the whole cluster; u_c rides with D.
- **Repeated measures / panel data with unit-level treatment.** Same person observed many times, “treated” or “control” the whole way through. The person plays the role of the cluster; their idiosyncratic level rides with D. Mechanically identical to cluster-randomized.
- **Treatment probabilities that differ by cluster.** Even with individual assignment, asymmetric T/C proportions correlated with u_c prevent it from cancelling cleanly.
- **Heterogeneous treatment effects by cluster.** τ_c varies across clusters; the residual carries a $\tau_c D_i$ piece that is cluster-correlated on the treated side and doesn’t cancel.

Catches with CR-SEs

- **They are not free.** CR-SEs can be conservative when— intervals widen, power drops.
- **But it is not** as bad as "having only $N_{cluster}$ observations": CR-SE estimates within-cluster correlation.
- **Few clusters is a real problem.** The standard CR-SE estimator (`vcovCL`) is downward biased with fewer then (roughly!) 50 clusters.
- **Fixes for few clusters:**
  - "Block boostrap": Resample whole clusters with replacement.
  - Wild cluster bootstrap `fwildclusterboot` in R.
  - Randomization inference at the cluster level.
  - Aggregate to the cluster level and run the analysis there (extremely conservative)

Takeaways

- **Unbiasedness is not enough.** The same experiment can be run more or less precisely; that's what we're paying attention to today.
- **Adjustment is good.** Covariate adjustment recovers real precision when X predicts Y. Use Lin's interacted form, with HC2 SEs; pre-specify the covariates.
- **Design is paramount.** Block on what you can; adjust for what you can't. The two are complements, not substitutes.
- **Cluster on the source of dependence in your residuals**, not on every group you can think of. Watch out for too-few-clusters and use an appropriate alternative when needed.
- Next: what if we can't randomize at all? **Selection on observables.**

Appendix

Frisch-Waugh-Lovell (FWL) theorem

Two facts:

1. Any bivariate regression of Y on X has $\hat{\beta} = \widehat{\text{Cov}}(Y, X)/\widehat{\text{Var}}(X)$.
2. The coefficient on the kth regressor in a multivariate regression can be obtained by (i) regressing X_k on the other X's and saving the residual $\tilde{X}_k$, then (ii) regressing Y (or its residual $\tilde{Y}$) on $\tilde{X}_k$.

Together:

$$\hat{\beta}_k = \frac{\widehat{\text{Cov}}(\tilde{Y}_i, \tilde{X}_{ki})}{\widehat{\text{Var}}(\tilde{X}_{ki})}$$

Why experimental findings are robust to specification

Apply FWL to D:

$$\hat{\beta}_D = \frac{\widehat{\text{Cov}}(\tilde{Y}_i, \tilde{D}_i)}{\widehat{\text{Var}}(\tilde{D}_i)}$$

where $\tilde{D}_i$ is the residual from regressing D_i on X_i.

- Under randomization, $D \perp\!\!\!\perp X$, so on average $\tilde{D} \approx D$.
- Hence adjusting for X doesn't move $\hat{\beta}_D$ much in expectation — it just removes the variance contribution from chance imbalances.
- This is why experimental results are typically robust to "what did you control for?"

Lin (2013): why the demeaning?

$$Y_i = \alpha + \tau_{Lin} D_i + \beta_1'(X_i - \bar{X}) + \beta_2' D_i(X_i - \bar{X}) + \varepsilon_i$$

- $\mathbb{E}[Y_i \mid D_i = 0, X_i = \bar{X}] = \alpha$
- $\mathbb{E}[Y_i \mid D_i = 1, X_i = \bar{X}] = \alpha + \tau_{Lin}$
- So τ_{Lin} is the comparison of treated vs. control *at the mean of X*.
- Because $\mathbb{E}_X[\text{ATE}(X)] = \tau_{Lin}$ for this model, it recovers the ATE without assuming constant treatment effects.
- Plain OLS without the interaction works in large samples but is biased $O(1/N)$ (Freedman 2008). Lin's form removes this and is now standard practice.

When does blocking help? (variance comparison)

Models for complete and block-randomized designs:

$$
\begin{aligned}
Y_i &= \alpha + \tau_{CR} D_i + \varepsilon_i \\
Y_i &= \alpha + \tau_{BR} D_i + \sum_{j=2}^{J} \beta_j B_{j[i]} + \varepsilon_i^*
\end{aligned}
$$

Then

$$
\begin{aligned}
\text{Var}[\hat{\tau}_{CR}] &= \frac{\sigma_{\varepsilon}^2}{\sum_i (D_i - \bar{D})^2} \\
\text{Var}[\hat{\tau}_{BR}] &= \frac{\sigma_{\varepsilon^*}^2}{\sum_i (D_i - \bar{D})^2 (1 - R_D^2)}
\end{aligned}
$$

where R_D^2 is from regressing D on the block dummies.

When does blocking help? (intuition)

- If treatment proportions are the same in every block, $R^2_D = 0$ and the denominator is unchanged — blocking doesn't *cost* you anything.
- If the blocks strongly explain Y, then $\sigma^2_{\varepsilon*} \ll \sigma^2_{\varepsilon}$ — blocking *wins* a lot.
- Combine: blocking on a strong predictor of Y, with equal-share randomization, weakly improves precision and usually does so a lot.

Cluster-robust variance: a sketch

For OLS $\hat{\beta} = (X'X)^{-1}X'Y$, the cluster-robust variance is

$$\widehat{\text{Var}}_{CR}(\hat{\beta}) = (X'X)^{-1}\left(\sum_{c=1}^{C} X_c'\hat{e}_c\hat{e}_c'X_c\right)(X'X)^{-1}$$

- X_c and $\hat{e}_c$ are the design matrix and residuals for cluster c.
- This collapses cluster-level scores, so it allows arbitrary intra-cluster correlation in residuals while assuming clusters are independent.
- The "effective $N = C$" approximation is the limit case $\rho \to 1$. The actual penalty is data-driven via $\hat{e}_c$.
- With $C \lesssim 30$, this estimator is downward biased; use a wild cluster bootstrap (Cameron, Gelbach, & Miller 2008) instead.

LPM caveats (binary outcomes)

Using OLS / LPM with a binary Y:

- **Pro:** unbiased for the ATE under random assignment; coefficients are interpretable as probability differences; works in every standard software environment.
- **Cons:**
  - Predicted probabilities can fall outside $[0, 1]$.
  - Heteroskedasticity is built in — use robust SEs.
  - In observational settings with extreme covariates, linearity can be a poor approximation.
- For the experimental settings in this lecture, none of these are reasons to abandon LPM.