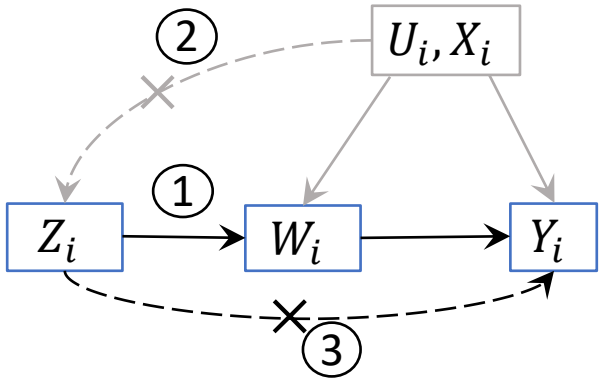
Lecture 10
Non-compliance in randomized experiments, instrumental variables
Part II

Outline

- IV in observational studies: the draft lottery
- Estimating ITT effects and two-stage least squares
- Conditional IV and population CACE
- Standard errors and bootstrap
- Weak instruments, exclusion, and interpretation
- Suggested reading: Imbens and Rubin Chapters 24.6, Peng Chapter 21.3-21.4, 23

Causal diagram for randomized encouragement

X: measured baseline causes. U: unmeasured causes.



- 1. Relevance: Z changes received treatment W.
- 2. Randomization: the instrument Z is independent of the background causes X and U.
- 3. Exclusion: Z affects Y only through W.
- Monotonicity adds the no-defiers condition for the binary-treatment CACE result.

Adjustment and IV use different assumptions

Measured X and unmeasured U may affect both treatment receipt W and outcome Y.

Direct adjustment for X requires unconfounded treatment receipt:

$$\{Y_i(0), Y_i(1)\} \perp W_i^{\text{obs}} \mid X_i$$

IV instead requires an unconfounded instrument:

$$\{Y_i(0), Y_i(1), W_i(0), W_i(1)\} \perp Z_i \mid X_i$$

Exclusion and monotonicity then connect instrument contrasts to CACE/LATE.

Suppose defendants are randomly assigned to judges who differ in sentencing severity. Why might judge assignment be unconfounded, while imprisonment is confounded?

The draft lottery as a natural experiment

- Question: how did military service during the Vietnam era affect later earnings?
- Veterans and nonveterans can differ in unobserved ways even after adjusting for observed covariates.
- Draft eligibility Z was randomized by the lottery. Military service W was not directly randomized (observational).
    - For each birth year of birth cohort 1950-1952, a random ordering of the 365 days was constructed, a cutoff number was pre-determined, young men of that birth year who had a birth date with order before the cutoff "won" the lottery
- Two-sided non-compliance:
  Eligible men could fail medical or education requirements. Ineligible men could volunteer.

The natural experiment supplies variation in eligibility.
Military service is the treatment whose effect we seek.

Angrist (1990). Document the actual lottery, cohort cutoffs, and assignment unit.

Draft lottery: assignment and the teaching model

- For each relevant birth cohort (1950–1952), lottery ranks and a cohort-specific cutoff determined draft eligibility.
- Birth dates received lottery ranks. Men with the same birth date in the same cohort shared eligibility.
- The actual assignment has clusters defined by birth date within cohort, with stratification by cohort.

Teaching simplification for the supplied numerical example:

- We treat each individual as a unit in a completely randomized experiment, ignoring birth-date clustering and cohort stratification.
- Ignoring shared assignment can understate uncertainty. An actual analysis must respect the lottery design.

Veterans and nonveterans in the teaching data

Summaries group people by actual military service W.

Variable	Nonveterans (N = 6,675)	Veterans (N = 2,030)
Draft eligible	0.24 (0.43)	0.40 (0.49)
Yearly earnings, $1,000s	11.8 (11.5)	11.7 (11.8)
Positive earnings	0.88 (0.32)	0.91 (0.29)
Birth year − 1900	51.1 (0.8)	50.9 (0.8)

Entries are means (SDs). Earnings ranges: 0–62.8 and 0–50.7, respectively.

Is 0.40 − 0.24 the IV ITT_W?
Which conditioning direction do we need?

The IV assumptions in the draft example

Assumption	Evidence or concern
Independence	Lottery mechanism, conditional on cohort and design features
Relevance	Eligibility increases the probability of service.
Monotonicity	No one volunteers only when ineligible but resists serving when eligible.
Exclusion restriction	Could eligibility change schooling or migration without service?

- Exclusion restriction could fail if eligibility changes schooling, career planning, or migration independently of actual service.

A man avoids service by entering school when eligible.
Which assumption is threatened if schooling changes his earnings? Why (think about a never taker)?

Draft-lottery estimates

Classroom calculation: treat individuals as completely randomized.

Quantity	Estimate	SE	Approximate 95% CI
First stage: service probability	0.1460	0.0108	[0.1248, 0.1672]
Reduced form: earnings ($1,000s)	−0.2129	0.1980	[−0.6010, 0.1752]
IV: complier earnings effect ($1,000s)	−1.46	1.36	[−4.13, 1.21]

Under the IV assumptions: about $1,460 lower yearly earnings for compliers.
The interval also permits no effect and positive effects.

Why can the IV estimate be imprecise even with a precise ITT_W?

The supplied standard errors ignore the actual birth-date clustering and cohort stratification.

Estimating CACE using familiar ITT methods

Under the IV assumptions, estimate the two population ITT effects and take their ratio.

$$\tau_c = \frac{\text{ITT}_Y}{\text{ITT}_W} \qquad \widehat{\tau_c} = \frac{\widehat{\text{ITT}_Y}}{\widehat{\text{ITT}_W}}$$

- When Z is randomized, reuse differences in means or regression methods from previous lectures to estimate each ITT.
- Neyman's approach gives variance estimates for differences in means. Uncertainty in the IV ratio depends on both ITTs and their covariance.
- Use methods appropriate to the assignment design and the same target population for both ITTs.
- If instrument independence requires conditioning on X, ITT estimation must account for X. We will discuss those estimation methods separately.

Would two valid ITT estimation methods have to give exactly the same IV estimate in a sample?

Two-stage least squares (2SLS)

Binary instrument, no covariates. Include an intercept in both stages.

Stage 1: regress received treatment W on the instrument Z.

$$\widehat{W_i} = \widehat{\alpha_W} + \widehat{\mathrm{ITT}_W} Z_i = \overline{W}_0 + (\overline{W}_1 - \overline{W}_0) Z_i$$

Fitted receipt equals the mean receipt in each instrument group.

Stage 2: regress outcome Y on fitted receipt.

$$Y_i = \alpha + \tau_{\mathrm{2SLS}} \widehat{W_i} + e_i$$

With only two fitted W values, the OLS line passes through both group means:

$$\overline{Y}_z = \hat{\alpha} + \widehat{\tau_{\mathrm{2SLS}}} \overline{W}_z, \qquad z = 0, 1$$

Subtract the two group equations and solve for the slope:

$$\widehat{\tau_{\mathrm{2SLS}}} = \frac{\overline{Y}_1 - \overline{Y}_0}{\overline{W}_1 - \overline{W}_0} = \frac{\widehat{\mathrm{ITT}_Y}}{\widehat{\mathrm{ITT}_W}} = \widehat{\tau_{\mathrm{IV}}}$$

The usual OLS standard errors from second-stage regression are generally incorrect. Why?

Why would replacing fitted receipt by a 0/1 classification change the IV estimator?

Conditional IV within baseline groups

- Our target remains population CACE. If Z is as-if random only within baseline groups X, we need comparisons within X. Earlier calculations ignored these differences.
- Draft example: eligibility probabilities and earnings can differ across birth cohorts.
  - Cohort 1950 has eligibility probability 53% (cutoff at 195), and cohort 1952 has 26% (cutoff 95)

Compare $Z = 1$ and $Z = 0$ within the same $X = x$ (for example, birth cohort):

$$\mathrm{ITT}_Y(x) = \mathrm{E}[Y \mid Z = 1, X = x] - \mathrm{E}[Y \mid Z = 0, X = x]$$

$$\mathrm{ITT}_W(x) = \mathrm{E}[W \mid Z = 1, X = x] - \mathrm{E}[W \mid Z = 0, X = x]$$

Within each baseline group ($X = x$), assume that the instrument is as-if randomized, exclusion restriction holds, and there are no defiers. Also assume both instrument values are possible.

$$\mathrm{ITT}_W(x) = \pi_c(x), \qquad \mathrm{ITT}_Y(x) = \pi_c(x)\tau_c(x)$$

$\pi_c(x)$: complier share. $\tau_c(x)$: average treatment effect among compliers in group x.

Suppose eligible and ineligible men have the same average earnings within every birth cohort. Could their average earnings differ when we combine all cohorts without adjusting for birth year? Why?

Population CACE and the ITT ratio

- Each group's weight is its share of all compliers.

$$\tau_c = \sum_x \mathrm{P}(X = x \mid C)\tau_c(x)$$

Using the conditional ITT identities from slide 12:

$$\tau_c = \frac{\sum_x p_x \mathrm{ITT}_Y(x)}{\sum_x p_x \mathrm{ITT}_W(x)} = \frac{\mathrm{ITT}_Y}{\mathrm{ITT}_W}$$

Hypothetical example: two cohorts of 1,000 people each

Cohort	Complier share	Number of compliers	Conditional CACE	Share of all compliers
A	0.10	100	2	0.20
B	0.40	400	1	0.80

$$\tau_c = 0.20 \times 2 + 0.80 \times 1 = 0.30 / 0.25 = 1.20$$

Why do the ITT averages use population shares p_x, while CACE uses shares of compliers $\mathrm{P}(X = x \mid C)$?

2SLS with baseline covariates

You may adjust for X in both stages.

Stage 1: regress received treatment W on Z and baseline covariates X.

$$\widehat{W_i} = \widehat{\alpha_W} + \hat{\pi} Z_i + \widehat{\gamma_W}^{\mathrm{T}} X_i$$

For example, X can contain indicators for birth cohort.

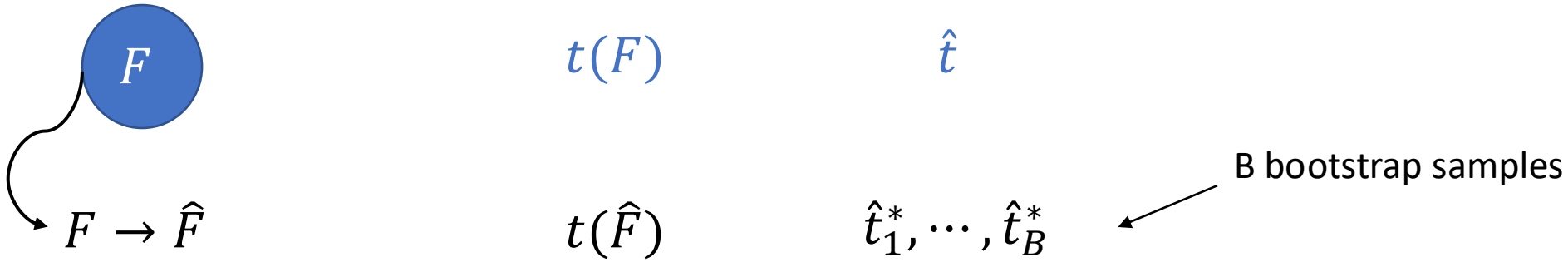
Stage 2: regress Y on fitted receipt and the same covariates X.

$$Y_i = \alpha_Y + \tau_{2\mathrm{SLS}} \widehat{W_i} + \gamma_Y{}^{\mathrm{T}} X_i + e_i$$

- You get the IV estimator as adjusting for X when estimating ITT effects and then take ratio
- Include an intercept in both stages. When X are confounding variables, choosing and modeling X requires care. We will discuss adjustment methods later.

Standard errors and the bootstrap

The general bootstrap idea



- For independent units, draw N complete records (Z, W, Y, X) with replacement.
- If holding assignment counts fixed, resample separately within the Z = 0 and Z = 1 groups.
- Repeat B times. Refit both stages and recompute $\widehat{\tau_{\mathrm{IV}}}$ in every replicate.
- The replicate histogram approximates the sampling distribution. Its SD estimates the standard error.

What would be lost if we refit the outcome model but kept the original ITT_W estimate in every replicate?

What makes an instrument weak?

An instrument is weak when the first-stage signal is small relative to its sampling uncertainty.

$$\mathrm{SE}(\hat{\tau}_{\mathrm{IV}}) \approx \frac{\mathrm{SE}(\widehat{\mathrm{ITT}_Y} - \tau_c \widehat{\mathrm{ITT}_W})}{|\mathrm{ITT}_W|}$$

- Small changes in the estimated denominator can cause large changes in the ratio.
- Normal intervals and the ordinary bootstrap may then be unreliable.
- A small receipt effect can still be estimated precisely in a large sample.

First-stage diagnostics

$$W_i = \alpha + \pi Z_i + \gamma^T X_i + u_i$$

Test whether Z predicts treatment receipt W after accounting for X.

$$\text{One excluded instrument: } F = \left[\frac{\hat{\pi}}{\text{SE}(\hat{\pi})}\right]^2$$

Rule of thumb for IV strong enough: $F > 10$
(a rough historical rule, not a universal guarantee).

Could an instrument have a very large first-stage F statistic and still violate exclusion?

Weak instruments magnify exclusion violations

$$\mathrm{ITT}_Y = \pi_c \tau_c + \delta \quad (\delta = \text{outcome effect outside the receipt pathway})$$

$$\frac{\mathrm{ITT}_Y}{\mathrm{ITT}_W} = \tau_c + \frac{\delta}{\pi_c}$$

Hypothetical direct contribution δ = 0.002:

Complier fraction	**Bias in the Wald estimand**
0.40	0.002 / 0.40 = 0.005
0.04	0.002 / 0.04 = 0.050

Assessing an observational IV analysis

- What changes treatment receipt, and why is the instrument independence assumption plausible in this study?
- Could the instrument change outcomes through another pathway?
- Is the no-defiers assumption plausible in this study?
- Who are the compliers in this study?
- After assessing the IV assumptions and identifying whose effect we estimate, how informative is the estimate?
    - How strong is the first stage ITT_W, and how precise is the IV estimate?

Two credible instruments give different effect estimates. Does that necessarily mean one is invalid?