Econ 272/Mgtecon 607 - Section 1

Amar Venugopal

Stanford University

Spring 2025

Outline

1. General Info & Logistics
2. Discussion Questions
3. Intro to Causality
4. Completely Randomized Experiments
5. Stratified/Paired Experiments
6. Power Analyses
7. Practice Problems

General Info

- Email: amar.venugopal@stanford.edu
- Section: Thursdays 4:30-6:20pm in STLC 114
- Office Hours: Wednesdays 2-4pm in Econ 149
- Section Questions: Question Form
- Section Feedback: Feedback Form
- Note: please use office hours to ask questions about the material, rather than over email.

For the entirety of the course, section materials are drawn heavily from those prepared by Lea Bottmer in past years.

Announcements

- Remember to email me about whether you wish to take the final exam, and if you require any special accommodations for it
  - Unless you are an Econ PhD, in which case you have no choice :)
  - If you are a predoc and wish to waive the class next year, you must take the exam as well

Section Norms

In order to foster a healthy, inclusive classroom environment, we will all abide by the following **community norms**:

- One microphone: we will allow everyone to finish their thought without interruption
- Everyone got accepted to Stanford and is here for a reason. There is no need to prove anything to anyone.
- Asking questions is a great way to deepen your understanding and resolve any confusion you may have. It is not a signaling device.
- There is no such thing as a stupid or dumb question
- Questions asked during section will be answered by the TA

Homework

- First problem set due **Sunday April 6, 11pm**. No late submissions are accepted.
- All submissions should consist of 2 files:
  - A **typed** writeup, which should be self-contained (the grader should not have to view any code)
  - A .zip file with all your code and user-created programs
- Please make sure your writeup is easy to read, and code should be well-commented
- For empirical questions: try to make economic sense of your results, or explain why they may not make sense
- For theoretical questions: make sure to explain steps clearly (cite theorems, references, etc) for full credit
- For problem set-related questions, come to my office hours or reach out to me (cc Guido on all problem set-related emails)
  - I **will not** respond to problem set emails sent on the due date

Hints for Problem Set 1

- The lecture notes (on Canvas) may be helpful if you get stuck at estimating homoskedastic variance

Lecture Feedback

Guido and I are both committed to making lecture and section as helpful/instructive as possible.

Feel free to discuss or email either of us with any feedback you might have.

Can also you section feedback google form if you prefer anonymous submission of feedback (just note in your response that the feedback pertains to lecture).

1. General Info & Logistics
2. Discussion Questions
3. Intro to Causality
4. Completely Randomized Experiments
5. Stratified/Paired Experiments
6. Power Analyses
7. Practice Problems

1. General Info & Logistics
2. Discussion Questions
3. Intro to Causality
4. Completely Randomized Experiments
5. Stratified/Paired Experiments
6. Power Analyses
7. Practice Problems

Key question: what is the effect of changing someone's treatment status on some outcome?

3 key areas of interest:

1. Point estimation ($\hat{\tau}$)
2. Inference ($\hat{V}$)
3. Hypothesis testing

Fundamental problem of causal inference:

We never observe counterfactuals

Notation

i denotes unit

- Y_i: Outcome
- W_i: Treatment → $W_i \in \{0, 1\}$
- $Y_i(w)$: Potential outcome at $W_i = w$
- X_i: Covariates

Note: Via Neymans's repeated sampling approach, PO are assumed to be *fixed* (unlike observed outcomes)

$$Var(Y_i) = Var(W_i Y_i + (1 - W_i) Y_i) = Var(W_i Y_i) + Var((1-W_i)$$

$$= Var(W_i Y_i(1)) \quad (Cov$$

$$= Y_i(1)^2 \, Var(W_i)$$

Assumptions

- **Known treatment assignment** $(Y_i(0), Y_i(1)) \perp W_i$
- **Stable Unit Treatment Value Assumption** $Y_i(\vec{w}) = Y_i(w_i)$

1. General Info & Logistics
2. Discussion Questions
3. Intro to Causality
4. Completely Randomized Experiments
5. Stratified/Paired Experiments
6. Power Analyses
7. Practice Problems

Point Estimation

- Estimand: $\tau^{ATE} = \mathbb{E}\left[Y_i(1) - Y_i(0)\right]$ $\quad \tau^{ATT} = \mathbb{E}\left[Y(1) - Y(0) \mid W=1\right]$
- Estimator: $\hat{\tau}^{DiM} = \bar{Y}_1^{obs} - \bar{Y}_0^{obs}$
- Properties: Unbiased $\quad \mathbb{E}\left[\hat{\tau}^{DiM}\right] = \tau^{ATE}$

Inference

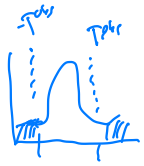
- True Variance:
- Estimator:
- Properties:
- Confidence Intervals:

Hypothesis Testing: Fisher Exact p-Values

H_0: $Y_i(1) = a\,Y_i(0) + b$

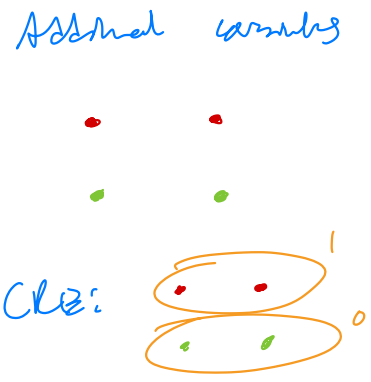
4-step procedure:

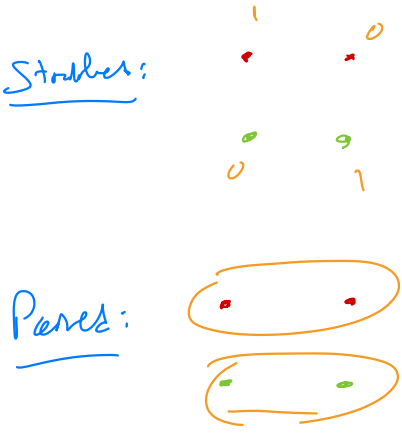
1. Define a sharp null H_0
   is H_0: $Y_i(0), Y_i(1)$ under H_0
2. Define test statistic, T $\quad (T = \bar{Y}_1 - \bar{Y}_0)$
3. Calculate distribution of T (exact numerical)
4. Calculate p-value $\quad p = \dfrac{\sum_{a \in A} 1\{|T^a| \geq |T^{obs}|\}}{|A|}$



1. General Info & Logistics
2. Discussion Questions
3. Intro to Causality
4. Completely Randomized Experiments
5. Stratified/Paired Experiments
6. Power Analyses
7. Practice Problems

What is the problem?





- Estimand: $\tau^{Strat} = E\left[Y_i(1) - Y_i(0)\right] = E\left[\tau(X_i)\right]$
- Estimator: $\hat{\tau}^{Strata} = \sum_{g=1}^{G} \frac{N_g}{N} \hat{\tau}_g$

$$Y_i^{obs} = \alpha + \tau W_i + \beta_1 S_{6i2g} + \varepsilon_i$$

$$\hat{\tau}^{RM} = \bar{Y}_1^{obs} - \bar{Y}_0^{obs}$$

Inference

- True Variance: $Var(\hat{\tau}^{strata}) = \sum_g \left(\frac{N_g}{N}\right)^2 Var(\hat{\tau}_g)$
- Estimator: $\widehat{Var}(\hat{\tau}^{strata}) = \sum_g \left(\frac{N_g}{N}\right)^2 \widehat{Var}(\hat{\tau}_g)$

Completely Randomized vs. Stratified vs. Paired

Design

	CRE	Strat	Pair
Pro	Simple, easy	Lower MSE vs CRE Heterogeneity	Lowest MSE in finite sample Higher power in large samples
Cons	Doesn't use potentially useful covariates		Cannot estimate variance of ATE (within strata)

$$V_C - V_S = A\left(\underbrace{n(X) \cdots}_{>0} - \underbrace{\text{Var}(\tilde{\tau})^2}_{\to >0}\right) \geq 0$$

Practical Recommendations

If sample large + predictive covs. → shrinkage or pool

Generally, small shrink > pool

1. General Info & Logistics
2. Discussion Questions
3. Intro to Causality
4. Completely Randomized Experiments
5. Stratified/Paired Experiments
6. Power Analyses
7. Practice Problems

Power Calculations

Power of a test (β): $P(\text{reject} \mid H_0 \text{ false})$

Size of a test (α): $P(\text{reject} \mid H_0 \text{ true})$

General procedure to get required sample size for given α, β:

1. Specify rejection criterion (e.g. $|T| \geq \Phi^{-1}(1-\alpha/2)$)
2. Specify probability $= \beta$
3. Solve for N

$\downarrow$

$P(\quad \cdot \quad) = \beta$

$\uparrow$

N

Useful Formulae

To determine minimum sample size when...

- Testing a mean against 0 (assuming random sample from normal distribution):

$$N = \left(\frac{\Phi^{-1}(\beta) + \Phi^{-1}(1 - \alpha/2)}{\mu_0/\sigma} \right)^2$$

- Testing DiM with unequal sample sizes and equal variances:

$$N = \frac{(\Phi^{-1}(\beta) + \Phi^{-1}(1 - \alpha/2))^2}{(\tau_0^2/\sigma^2)\gamma(1 - \gamma)}$$

where γ = share of treated units

Econ 272/Mgtecon 607 - Section 2

Amar Venugopal

Stanford University

Spring 2025

Outline

1. Announcements

2. Discussion Questions

3. Clustered Experiments

4. Clustering in Sampling and in Assignment

5. Practice Problems

Announcements

- Remember to email me about whether you wish to take the final exam, and if you require any special accommodations for it
  - Unless you are an Econ PhD, in which case you have no choice :)
  - If you are a predoc and wish to waive the class next year, you must take the exam as well

Hints for Problem Set 2 (Due April 13, 11pm)

- Within-group correlations of a single variable try to answer the question: does *my* value going up mean that *your* value is more likely to go up/down?
  - $Corr(Y_i, Y_j)$ for $i \neq j$ can be calculated by constructing pairs (combinatorially) of observations from the data
  - Recall the definition/equation for correlation - how can it be simplified in this context?

$$Corr(Y_i, Y_j) = \frac{Cov(Y_i, Y_j)}{\sqrt{Var(Y_i)Var(Y_j)}} = \frac{Cov(\tilde{Y}_i, \tilde{Y}_j)}{Var(\tilde{Y}_i)} = E[\tilde{Y}_i \tilde{Y}_j] - E[\tilde{Y}_i]E[\tilde{Y}_j]$$

$$\tilde{Y}_i = \frac{Y_i - \bar{Y}}{\sigma_Y}$$

$$\frac{1}{\binom{N_g}{2}} \sum_{i \neq j} \tilde{Y}_i \tilde{Y}_j$$

- Simplified equations for LZ/EHW variance estimators on slides (Lecture 4, Slide 4) are missing a factor of $1/N$

A note on terminology

- In the course, we will often use the term "population" to refer to the experimental sample since we are mainly thinking about design-based inference
- We will therefore refer to the broader population from which the experimental sample is drawn as the "super-population"
- We may occasionally refer to our experimental sample as the "sample", for example when we want to distinguish between sample and population ATEs
- It is confusing, I will try to be as explicit as possible

From last time: nontrivial sharp nulls

Test statistic for sharp null: $Y_i(1) = a \cdot Y_i(0) + b$

$\tilde{Y}_i(1)$ $\tilde{Y}_i(0)$

$$T = \bar{\tilde{Y}}_i(1) - \bar{\tilde{Y}}_i(0)$$

$$= \bar{Y}_i(1) - a\,\bar{Y}_i(0) - b$$

1. Announcements
2. Discussion Questions
3. Clustered Experiments
4. Clustering in Sampling and in Assignment
5. Practice Problems

1. Announcements

2. Discussion Questions

3. Clustered Experiments → Clustered Assignment

4. Clustering in Sampling and in Assignment

5. Practice Problems

Why cluster?

Spillovers, network effects $\Rightarrow$ SUTVA violations

My outcome depends on your treatment

Idea: If spillovers occur in specific clusters

treating clusters as units.

Clustered Sampling vs. Assignment

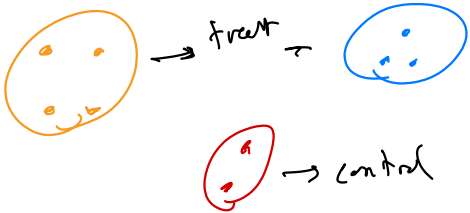
- **Clustered Sampling:**

C_1, C_2, C_3

Pick 2

- **Clustered Assignment:**

helps spillovers



Notation

- G: Clusters
- $G_i \in \{1, ..., G\}$: cluster assignment for unit i
- N_g: # units in cluster g ($N = \sum_g N_g$)

= if $N_g = n \ \forall g$ if TE cluster

- **Estimands**:

$\tau_{pop} = \frac{1}{N} \sum_i Y_i(1) - Y_i(0)$ — overall pop. ATE

$\tau_{cluster} = \frac{1}{G} \sum_g \left(\frac{1}{N_g} \sum_i Y_i(1) - Y_i(0) \right)$ — overall cluster-level ATE

- **Estimators**:

$\hat{\tau}_{pop} = \bar{Y}_1 - \bar{Y}_0$ — need to estimate previously

$\hat{\tau}_{cluster} = \frac{1}{G_1} \sum_g \left(\frac{1}{N_g} \sum_i W_i Y_i \right) - \frac{1}{G_0} \sum_g \left(\frac{1}{N_g} \sum_i (1 - W_i) Y_i \right)$

Both estimators are unbiased

Variance: $\hat{\tau}_{cluster}$

$$\Rightarrow V_{regression} = \frac{S_0^2}{n_0} + \frac{S_1^2}{n_1} - \frac{S_{01}^2}{n}$$

- **True Variance**:

$$Var(\hat{\tau}_{cluster}) = \frac{S_0^2}{G_0} + \frac{S_1^2}{G_1} - \frac{S_{01}^2}{G}$$

- **Estimator**:

$$\widehat{Var}(\hat{\tau}_{cluster}) = \frac{\hat{S}_0^2}{G_0} + \frac{\hat{S}_1^2}{G_1}$$

$$S_v^2 = \frac{1}{G-1} \sum_g \left(\bar{Y}_g(v) - \overline{\bar{Y}_g(v)}\right)^2$$

Variance: $\hat{\tau}_{pop}$

Challenge: treatment assignment not indep.

- **True Variance**: Asymptotics: Sequence of progs. $k = 1, 2, \ldots$
  In prog. k, sample all units, assign cluster g to treatment w.p. p
  Keep N_g fixed, let $G_k, N_k \to \infty$:
  $$\sqrt{N_k}\left(\hat{\tau}_{prog} - \tau_{prog}\right) \xrightarrow{d} N(0, V_k)$$
  V_k = var on strata 24

- **Estimator**: Conservative estimate with: $Y_i = \alpha + \beta W_i + \varepsilon_i$

$$\hat{V}_{(cluster} = \left(\frac{1}{N}\sum \tilde{W}_i \tilde{W}_i^T\right)^{-1} \left(\frac{1}{N}\sum_g \sum_i (\varepsilon_i \tilde{W}_i)^2\right) \left(\quad\right)^{-1}$$

$$\tilde{W} = W_i - \frac{1}{N}\sum_{j=1}^{N} W_j$$

$\hat{V}_{LZ}$ $\left(\geq V_{EHW}\right)$

Hypothesis Testing

Inference based on rand. dsgn. induced by clustered assignment, rather than model

Typically rely asymptotics.

But, Recent econ & STAT work, but most of analyses are clusters

1. Announcements

2. Discussion Questions

3. Clustered Experiments

4. Clustering in Sampling and in Assignment

5. Practice Problems

2 "Flavors" of Error / Variability

1. Sampling – how sample constructed
2. Assignment – how treatment assigned

DoS needed for inference in causal settings
→ Abadie et. al. (2023) (general framework)

(Even More) Notation

Sequences of populations indexed by $k = 1, 2, 3, ...$

- g_k: # clusters in k
- $n_{g,k}$: # units in cluster g in pop k
- $q_k \in (0, 1]$: cluster sampling probability in pop k { $q_k = 1$: random sampling; $q_k < 1$: cluster sampling } (sampling)
- p_k: prob. a unit is sampled from cluster
- $A_{k,g}$: prob. unit in cluster g in pop. k is treated: $A_{k,g} \sim (\mu_k, \sigma^2_k)$

Not all clusters need be sampled

Not all units need same treatment within cluster } Need C.S. + C.A.

Clustering Regimes

q_k, σ_k^2, μ_k

+

sampling assignment

		Assignment, $A_{k,g}$		
		Random	**Partially Clustered**	**Clustered**
Sampling q_k	**Random**	$q_k = 1$, $\sigma_k^2 \geq 0$	$q_k \geq 1$, $0 < \sigma_k^2 < \mu_k(1-\mu_k)$	$q_k = 1$, $\sigma_k^2 = \mu_k(1-\mu_k)$ ↓
	Clustered	$q_k < 1$, $\sigma_k^2 \geq 0$	$q_k < 1$,	$q_k < 1$,

Special Cases

- Random sampling of clusters from a large population of clusters → ρ_k small
- Random sampling from a large population → $\rho_k = 1$, P_k small
  - Completely random assignment $A_{k,g} = A_k$ $(\sigma^2_k = 0)$
  - Clustered random assignment $A_{k,g} \in \{0, 1\}$
- What can we test?

Cannot test $\rho_k = 1$ vs. < 1

Can test $\sigma^2_k > 0$ vs. $= 0$

Causal Cluster Variance (CCV) - Big Idea

Authors DiM converges: $\sqrt{N_k}\,(\hat{\tau}_k - \tau_k) \xrightarrow{d} N(0, V_k^{CCV})$

$V_k^{CCV} = V_k^{CCV}(q_k, \sigma_k^2, \eta_k, \mu_k)$

Special cases:

$V_k^{EHW} = V_k^{CCV}(q_k = 1, \sigma_k^2 = 0)$

$V_k^{CL} - V_k^{CCV} \geq 0 \rightarrow$ CL conservative

$\forall x \; \hat{\tau}^{DiM}$ to est. τ_{FT}

True Variance by Regime

		Assignment, $A_{k,g}$		
		Random	**Partially Clustered**	**Clustered**
Sampling q_k	**Random**			
	Clustered			

$\sigma_2^2 \to 0$, $\hat{V}^{ccv} \to \hat{V}^{EHW}$ $\quad\quad$ $\sigma_2^2 \to \mu_k(1-\mu_k)$, $\hat{V}^{ccv} \to \hat{V}^{L2}$

$\xi_k \to 0$

(1) Analytic : $\hat{V}_k^{ccv} = \hat{\xi}_k \cdot \hat{V}_k^{CCV}(1) + (1-\hat{\xi}_k)\hat{V}_k^{L2}$

$\hat{\xi}_k$: knowledge on the lack of clusters in ξ_{gm}

(2) Bootstrap : JSCB

Variance Estimation by Regime

		Assignment, $A_{k,g}$		
		Random	**Partially Clustered**	**Clustered**
Sampling q_k	**Random**	$\hat{V}^{EHW}$	$\hat{V}^{CCV} \to \hat{e}_{k,2}=1 \to \hat{V}^{CCV}(1)$	$\hat{V}^{LZ}$
	Clustered	$\hat{V}^{CCV}$	$\hat{V}^{CCV}$	$\hat{V}^{CCV}$

Note: the choice of estimator really matters here!! Standard errors can vary widely.

TSCB: not available

LZ: if treated untreated clusters in typ to estimate

EHW: if drawing randomly local to sample

CCV: if switchers, or if you use clustered sampling

Econ 272/Mgtecon 607 - Section 3

Amar Venugopal

Stanford University

Spring 2025

Outline

1. Discussion Questions

2. Multi-Armed Bandits

3. Identification

4. Practice Problems

1. Discussion Questions

2. Multi-Armed Bandits

3. Identification

4. Practice Problems

What's different from traditional experiments?

- **Setting:** 1 experimental sample → sequential samples ("online")

Inference is hard

- **Treatment Arms:** 2 → K (> 2)

- **Goal:** Estimation → right decision

"Dual Mandate" for Bandits

- Exploration : get a good sense of each arm's efficacy

- Exploitation : give units best possible treatment

→ balance these to minimize regret (suboptimal assignments)

give as many units as possible best treatment possible

- **Action/Arm:** $k_t \in \{1, \dots, K\}$ (a_t)
- **Outcome/Reward:** $Y_t \in \mathbb{R}$ (r_t)
- **Policy:** π : history $\rightarrow$ action

  $\pi(k_1, Y_1, \dots, k_{t-1}, Y_{t-1}) = k_t$

Naive Approach

Given: N units, K arms

1. Assign N/K units per arm, observe avg. outcome
2. Choose highest avg. arm: $k^* = \arg\max_k \bar{Y}_k$

Drawbacks:

- risk averse
- no priors
- equal confidence in each arm → could be wrong about best

no exploitation

Big idea: "learn as we go"

Dynamically assign units to arms to balance explore vs. exploit

2 common methods/algorithms:

1. Thompson Sampling (TS)
2. Upper Confidence Bound (UCB)

Quick Detour: Bayesian Updating

1. Form prior : $p(\theta)$
2. Observe data : $p(x \mid \theta)$
3. Update via Bayes' Rule : $p(\theta \mid x) = \frac{p(\theta, x)}{p(x)} = \frac{p(x \mid \theta)\, p(\theta)}{p(x)}$ (posterior)

$$\propto \underbrace{p(x \mid \theta)}_{\text{likelihood}} \underbrace{p(\theta)}_{\text{prior}}$$

Incalculable analytically → favor conjugate priors

distributions posterior, prior same class
ex. Normal, Beta (uniform)

Thompson Sampling: Idea

1. **Select:** according to probability arm is optimal
2. **Update:** observe Y_t, update posterior for selected arm

- **Exploration:** Each arm has nonzero probability of being selected
- **Exploitation:** Choosing arms we think are better more often

Modification: Exploration Sampling – choose according to variance
↑ Exploration, ↓ Exploitation

Thompson Sampling: Example

$\mu_2 = P(Y_2 = 1)$

$K = 3$ arms, binary rewards, $\mu_1 = 0.5, \mu_2 = 0.2, \mu_3 = 0.9$, uniform prior: $\underbrace{\mathcal{B}(1,1)}_{U(0,1)}$

1. Sample from prior: $\hat{\mu}_1 = 0.43$, $\hat{\mu}_2 = 0.75$, $\hat{\mu}_3 = 0.11$ $\rightarrow$ $B(1,1)$
2. Assign to highest: $k_1 = 2$, observe failure $(Y = 0)$
   Update: $\beta_2 = \beta_2 + 1$ $\rightarrow$ $B(1,2)$ $\rightarrow$ supp $= [0,1]$
3. Sample posterior: $\hat{\mu}_1 = 0.52$ (prior), $\hat{\mu}_2 = 0.27$ (posterior), $\hat{\mu}_3 = 0.86$ (prior)
4. $k_2 = 3$, observe, update, ...

$\hat{\mu}$ $\rightarrow$ $P(\text{wing}) \geq a$

E

E

am

Upper Confidence Bound: Idea

1. **Select:** Calculate CI for μ_k, choose arm with highest UB
2. **Update:** $UB_t^{(k)} = \hat{\mu}_{k_t} + \sqrt{\frac{2\log(1/\epsilon)}{T(k,t)}}$ → "confidence level" $(= 1/\epsilon)$
   → wider CI over time
   ↑ Exploration
   ↓ Exploitation

- **Exploration:** More uni's → narrower CI
- **Exploitation:** As we get better estimates of means, CIs shift accordingly

Upper Confidence Bound: Example

$K = 3$ arms, binary rewards, $\mu_1 = 0.5, \mu_2 = 0.2, \mu_3 = 0.9$

$t=1$ 1. Pull each once to get initial estimates: $\hat{\mu}_1 = 0$, $\hat{\mu}_2 = 0$, $\hat{\mu}_3 = 1$

$t=2$ 2. Calculate UCB: $UB_1 = 0 + \sqrt{\frac{2\log(2)}{1}} = ?$ $UB_3 = 1 + \sqrt{\frac{2\log(2)}{1}} = ?$ (larger)

$= UB_2$

$(\rho = 1/t)$

3. Pull arm 3, observe, update UB, ...

⋮

- **Regret:** $R(\pi) = \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T} Y_t \mid \pi^*\right] - \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T} Y_t \mid \pi\right]$

- **Lower Bound:** $\mathbb{E}[R(\pi)] \geq \frac{\ln(T)}{T} \sum_{k \neq k^*} \frac{\mu^* - \mu_k}{KL(\mu_k, \mu^*)} + o(1)$

  (1) Decreasing in T
  (2) Small if μ_k v. diff. μ^*
  (3) Hard to learn if $\mu_k \sim \mu^*$ ← close

- **TS Regret:** $\mathbb{E}[R(\pi^{TS})] = O\left(\frac{\ln(T)}{T} \sum_{k \neq k^*} \frac{1}{(\mu^* - \mu_k)^2}\right)$

Extensions

- **What if we observe covariates?**

  ↪ LinUCB

- **Inference?**

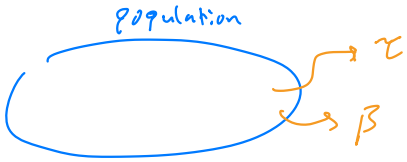
1. Discussion Questions

2. Multi-Armed Bandits

3. Identification

4. Practice Problems

Identification



- **Question:** Given "enough data", can we learn it?
- **Challenge:** What is "enough data"

3 Types of Identification

1. Point
2. Partial / Set
3. None

Setting

Focus on cross-sectional data. Consider sequence of populations $k = 1, ..., \infty$

OLS, $\theta = \beta$.

Assume: $Y_i | X_i \sim \mathcal{N}(\alpha + \underbrace{\beta X_i}_{\theta}, \sigma^2)$

$$\beta = \frac{Cov(X_i, Y_i)}{Var(X_i)}$$ → estimable from data

$$\rightarrow \mathcal{N}(\alpha + \beta X_i + \gamma\left(\frac{X_i}{2}\right), \sigma^2)$$

↳ not ID

$$\left(\beta + \frac{\gamma}{2}\right)$$

Partial Identification - Example

$W_i, Y_i \in \{0, 1\}$, observe outcome only for treated units. $\theta = E[Y_i] = P(Y_i = 1)$.

Never observe: $P(Y_i | W_i = 0) \in [0, 1]$

$LB \leq \theta \leq UB$

Econ 272/Mgtecon 607 - Section 4

Amar Venugopal

Stanford University

Spring 2025

Outline

1. Announcements

2. Directed Acyclic Graphs (DAGs)

3. Practice Problems

4. eCONometrics? Evolution of Views

Announcements

- Final Exam: Wednesday June 11, 8:30-11:30am (room TBD)
- Remember to email me about whether you wish to take the final exam, and if you require any special accommodations for it
  - Unless you are an Econ PhD, in which case you have no choice :)
  - If you are a predoc and wish to waive the class next year, you must take the exam as well
- Assignment 5 - Due May 11 (skipping a week, good luck with midterms!)

1. Announcements

2. Directed Acyclic Graphs (DAGs)

3. Practice Problems

4. eCONometrics? Evolution of Views

Structural Equation Models $\iff$ DAGs

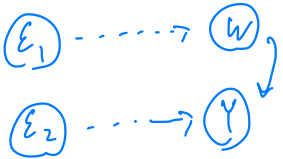
Consider a simple randomized experiment.

- **Structural Equation:**

$W = f_1(\varepsilon_1)$ $\quad$ $Y = f_2(W, \varepsilon_2)$

ε_1: random

- **DAG:**



Why DAGs? Help (visually) argue for identification

Terminology

- **Nodes:** Quantities/variables (A, B, C)
- **Parent/Child:** Immediate predecessor/successor
- **Ancestor:** (Non) immediate predecessor
  Descendant
- **(Dashed) Arrows:** (Unobserved) links
- **Directed Path:** Path with all arrows in same direction
  C → A → B dir. path
  C ← A → B path

Types of Nodes (on a given path)

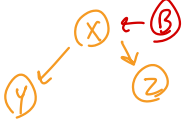
- **Collider:** Arrows coming in, but no out

Y collider

- **Mediator:** 1 arrow in, 1 arrow out

W mediator

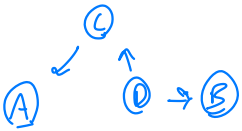
- **Confounder:** Arrows going out, but not in



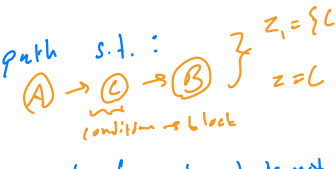
X confounder

Types of Paths (from A to B...)

- **Backdoor Path:** $A \leftarrow \dots \rightarrow B$



- **Blocked Path:** Z_1 blocks path if $\exists z \in Z_1$ on path s.t.:
  1. z is a non-collider that is conditioned on
  OR
  2. z is a collider that is not conditioned on, and all of its descendants not in Z_1



- **Open Path:** Any path that is not blocked

Blocking and Conditioning

Mediators, confounders don't block (by default)

Colliders do " " "

Conditioning reverses blocking rules, may be necessary for ID

↕

including in regression

Extension: d-Separation

- **Definition:** A set of variables "d-separates" 2 other variables if it blocks (conditionally) all paths between them

- **Implication:** If A is d-seperated from B by set Z_1, $A \perp B \mid Z_1$

Identification Criteria (for causal effect of A on B)

backdoor

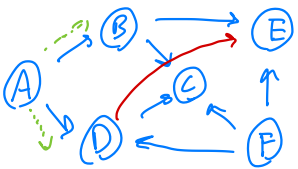
- Back door Criterion: Can we block all paths by conditioning on some set of nodes? If yes, ID causal effect of $A \to B$ (conditional)

- Front door Criterion: Set of nodes Z_1 that satisfies:
  1. Z_1 intercepts all dir. paths from $A \to B$
  2. No unblocked backdoor paths from $A \to Z_1$
  3. All backdoor paths from $Z_1 \to B$ blocked by A

  If all 3 hold, ID causal effect $A \to B$

Back Door Criterion - Example

TD: $D \to E$



What if A unobserved?

Yes, still have TD

$\{B, F\}$

Cannot condition on unobserved

Paths with unobserved nodes still need to be blocked!

Backdoor Paths:

$D \leftarrow A \to B \to E$

$D \leftarrow F \to E$

$D \leftarrow F \to C \leftarrow B \to E$ (collider)

$D \leftarrow A \to B \to C \leftarrow F \to E$

Block by:

$\{A\}$, $\{B\}$, $\{A, B\}$

$\{F\}$

$\emptyset$

$\emptyset$

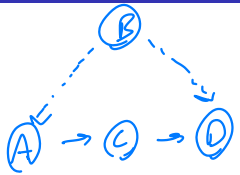
Options!

$\{A, F\}$, $\{B, F\}$, $\{A, B, F\}$ → condition on

prefer fewer

Front Door Criterion - Example

TD $A \to D$



(1) Intercept all directed paths

$A \to C \to D$ $\{C\}$

(2) No backdoor path $A \leftarrow \dots \to Z_1$

No backdoor $A \leftarrow \dots \to C$

(3) All backdoor $Z_1 \leftarrow \dots \to D$ blocked by A

$C \leftarrow \dots \to D$

$C \leftarrow A \leftarrow B \to D$ C blocked by A ✓

All 3 criteria satisfied by $Z_1 = \{C\}$

Collider Bias

Conditioning on a collider can unblock paths and introduce bias

"General" bias

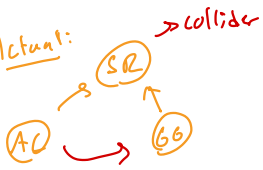
Goal: AC → GG

Method: Run survey

Ideal:

SR independent of both

Actual:



M-bias

Goal: education → diabetes

Think: confounded by mother's diabetes status

C is really a collider

Conditioning on it opens backdoor paths from X↔Y, introduce bias (no TD)

1. Announcements

2. Directed Acyclic Graphs (DAGs)

3. Practice Problems

4. eCONometrics? Evolution of Views

1. Announcements

2. Directed Acyclic Graphs (DAGs)

3. Practice Problems

4. eCONometrics? Evolution of Views

Econ 272/Mgtecon 607 - Section 5

Amar Venugopal

Stanford University

Spring 2025

Outline

1. Discussion Questions

2. Unconfoundedness

3. Semiparametric Efficiency Bounds

4. Practice Problems

1. Discussion Questions

2. Unconfoundedness

3. Semiparametric Efficiency Bounds

4. Practice Problems

How can we identify causal effects?

- **Experiments:** Randomization of treatment
- **Observational Data:** Treatment no longer random

Goal: "as-if" randomization

unconfoundedness + overlap

Unconfoundedness Assumption

Also known as "selection on observables" or "exogeneity".

Definition: $W_i \perp (Y_i(0), Y_i(1)) \mid X_i$

Observables may impact treatment assignment

But, conditional on them, we are "as-if" random.

Propensity Scores

- **Motivation:** With continuous covariates $P(\exists i,j : X_i = X_j) = 0$

- **Propensity Score:** $e(x) = P(W_i = 1 \mid X_i = x)$
  → "sufficient statistic" for selection on observables

- **(Alternative) Unconfoundedness:** $W_i \perp (Y_i(0), Y_i(1)) \mid e_i$
  → match on propensity scores, rather than raw covariates

Assessing Unconfoundedness

While unconfoundedness is not *testable*, its plausibility can be *assessed*.

2 assessment strategies:

1. Estimate effect on pseudo outcome
2. Estimate effects of pseudo treatments

Strategy #1: Estimating Effect on Pseudo Outcomes

Big Idea: Treat some covariates as outcomes, test independence.

$X_i = (X_i^p, X_i^r)$ (pseudo outcome)

If $X_i^p \approx Y_i(0)$, then under (U):

$X_i^p \perp W_i \mid X_i^r$ → testable!

Ex. Estimate avg. diff. in X_i^p by treatment status, after adjusting for differences in X_i^r as you choose.

One option: X_i^p = lagged outcome (if available)

Strategy #2: Estimating Effects of Pseudo Treatments

Big Idea: If we have multiple control groups, we can test if their observed outcomes differ.

$G_i \in \{t, c_1, c_2\}$

$W_i = \begin{cases} 0, & G_i \in \{c_1, c_2\} \\ 1, & G_i = t \end{cases}$

stronger version of (2):

$(Y_i(0), Y_i(1)) \perp G_i \mid X_i$

Testable implication:

$Y_i(0) \perp G_i \mid X_i, G_i \in \{c_1, c_2\} \iff Y_i^{obs} \perp G_i \mid X_i, G_i \in \{c_1, c_2\}$

Ex. As before, test whether Y_i differs on avg. b/w c_1, c_2 after adjusting for X_i

Conditional Average Treatment Effect (CATE): $\tau(X) = \mathbb{E}[Y_i(1) - Y_i(0)|X_i = x]$

Goal for ATE ID: $\tau = \mathbb{E}[\tau(X_i)]$

Let's see:

$$\mathbb{E}[\tau(X_i)] = \mathbb{E}\big[\mathbb{E}[Y_i(1) - Y_i(0) \mid X_i = x]\big] \quad \text{(definition of } \tau(x)\text{)}$$

$$= \mathbb{E}\big[\mathbb{E}[Y_i(1) \mid X_i = x] - \mathbb{E}[Y_i(0) \mid X_i = x]\big] \quad \text{(linearity of expectation)}$$

$$= \mathbb{E}\big[\mathbb{E}[Y_i(1) \mid W_i = 1, X_i = x] - \mathbb{E}[Y_i(0) \mid W_i = 0, X_i = x]\big] \quad \text{(unconfoundedness)}$$

$$= \mathbb{E}\big[\mathbb{E}[Y_i \mid W_i = 1, X_i = x] - \mathbb{E}[Y_i \mid W_i = 0, X_i = x]\big] \quad \text{(definition of PO)}$$

$\stackrel{?}{=} \tau$ → need something more

may not be separately estimable!

Overlap Assumption

Definition: $p(w_i = 1 \mid x_i) = e_i \in (0,1)$

For any x, we can find both treated and control

This assumption is *testable*. Some strategies:

1. Normalized differences
2. Plotting
3. Log odds ratio

Strategy #1: Normalized Differences

Big Idea: Check if covariate distributions vary across treatment states

$$nd = \frac{\bar{X}_t - \bar{X}_c}{\sqrt{\frac{s_t^2 + s_c^2}{2}}}$$

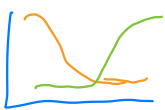
→ better than t statistic, which will increase in sample size

Rule of Thumb: If $nd > 0.25$, problem!

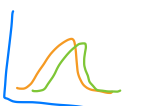
Strategy #2: Plotting

Big Idea: Check visually if distributions "look" overlapping

"Rule of Thumb":



Bad!



Good!

Strategy #3: Log Odds Ratio

Big Idea: Propensity scores should be similar between the 2 groups.

$$LOR = \log\left(\frac{e(x)}{(-e(x)}\right)$$ → Calculate means for each group and compare

Rule of Thumb: If means > 1 s.d. apart, problem!

std. dev.

What if we have only limited overlap?

Challenge: We cannot get good estimates of the seperate conditional outcomes.
Need a "better" sample.

2 strategies:

1. Matching
2. Subsampling

Strategy #1: (Propensity Score-) Matched Samples

Big Idea: Construct a new sample of similar units in each treatment group.

Better to match on propensity scores than raw covariates.

Algorithm:

1. Estimate $e(x_i)$ $\forall i$
2. Order treated units by decreasing propensity
3. For each, find control unit with closest propensity (without replacement)
4. Now have balanced sample of $2 \cdot N_T$ units
5. Proceed as usual, with this as your new sample

Strategy #2: Subsamples

Big Idea: Only consider units with covariate values where we have good overlap

$$\tau_c(A) = \frac{1}{|A|} \sum_{i: X_i \in A} \tau(X_i)$$ → subpopulation ATE

Choosing Subsample: Choose A to minimize variance of $\tau_c(A)$.

Generally, this means:

$$A = \{x \in X : \alpha < e(x) < 1-\alpha\}$$ → throw out extrema of propensity score

$$\alpha = 2 \cdot E\left[\frac{1}{e(x)(1-e(x))} \;\middle|\; \frac{1}{e(x)(1-e(x))} < \alpha \right]$$ → $\alpha \approx 0.1$

Identification!

$$
\begin{aligned}
\mathbb{E}[\tau(X_i)] &= \mathbb{E}[\mathbb{E}[Y_i(1) - Y_i(0)|X_i = x]] && \text{by definition} \\
&= \mathbb{E}[\mathbb{E}[Y_i(1)|X_i = x] - \mathbb{E}[Y_i(0)|X_i = x]] && \text{by linearity} \\
&= \mathbb{E}[\mathbb{E}[Y_i(1)|W_i = 1, X_i = x] - \mathbb{E}[Y_i(0)|W_i = 0, X_i = x]] && \text{by unconfoundedness} \\
&= \mathbb{E}[\mathbb{E}[Y_i|W_i = 1, X_i = x] - \mathbb{E}[Y_i|W_i = 0, X_i = x]] && \text{by definition} \\
&= \tau && \text{by overlap}
\end{aligned}
$$

1. Discussion Questions

2. Unconfoundedness

3. Semiparametric Efficiency Bounds

4. Practice Problems

Types of Models

- **Parametric:** Well-defined models, finite dimensional space
  e.g. OLS, $y \sim N(\alpha + \beta X, \sigma^2)$

- **Semiparametric:** Parameter of interest is finite-dimensional, with another that is infinite-dimensional
  e.g. Cox Proportional Hazards, $y \sim f(y \mid \beta, \lambda_0)$ (β: finite-dim; λ_0: function (∞-dim); "nuisance parameter")

- **Nonparametric:** All parameters can span an infinite space
  e.g. decision trees

Given: $f_X(x|\theta)$ pdf

$\hat{\theta}(x)$ unbiased for θ

Fisher Information: $I(\theta) = -E\left[\frac{\partial^2 \log(f)}{\partial\theta\,\partial\theta}(X|\theta)\right]$

Cramer-Rao Lower Bound: $V(\hat{\theta}(X)) \geq I(\theta_0)^{-1}$

MLE generally unbiased + efficient

$\hat{\theta}$ is asymptotically efficient if:

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} N(0, I(\theta_0)^{-1})$$

Semiparametric model: $f(x|\theta, h(\gamma)) = f(x|\theta,\gamma)$ → parametric sub-model

d-dim $h(\gamma)$ known up to finite parameter γ

CRLB tells us: $\underbrace{ASV(\hat{\theta})}_{\text{asymptotic variance}} \geq \underbrace{I^{\theta\theta'}}_{\text{inverse Fisher info. for } \theta} = \left(I_{\theta\theta'} - I_{\theta\gamma'} I_{\gamma\gamma'}^{-1} I_{\gamma\theta'}\right)^{-1}$

Goal: rewrite $I^{\theta\theta'}$ in simpler terms

Score Function = 1st deriv of log. max. likelihood

Let $\beta = (\theta, \gamma)$

$$S_\beta(x) = \begin{pmatrix} S_\theta(x) \\ S_\gamma(x) \end{pmatrix} = \begin{pmatrix} \frac{\partial \log f}{\partial \theta}(x; \theta, \gamma) \\ \frac{\partial \log f}{\partial \gamma}(x; \theta, \gamma) \end{pmatrix}$$

Can then write Fisher info:

$$I(\theta, \gamma) = \begin{pmatrix} I_{\theta\theta'} & I_{\theta\gamma'} \\ I_{\gamma\theta'} & I_{\gamma\gamma'} \end{pmatrix} = E\left[\begin{pmatrix} S_\theta(x) S_\theta(x)' & S_\theta(x) S_\gamma(x)' \\ S_\gamma(x) S_\theta(x)' & S_\gamma(x) S_\gamma(x)' \end{pmatrix} \right]$$

Efficient Score

CRLB then gives:

$$(I^{\theta\theta'})^{-1} = I_{\theta\theta'} - I_{\theta\gamma} I_{\gamma\gamma'}^{-1} I_{\gamma\theta'}$$

$$= E\left[\psi(x)\, \psi(x)' \right]$$

$$\psi(x) = S_\theta(x) - E\left[S_\gamma(x)\, S_\theta'(x) \right]' E\left[S_\gamma(x)\, S_\gamma'(x) \right]^{-1} S_\gamma(x)$$

↳ residual from projection of S_θ onto S_γ = "efficient score"

In practice, we do not know S_γ, so we look over all possible values, called the tangent set → generally tricky to find

(Efficient) Influence Function

Efficient Influence Function = influence function with smallest variance

EIF = orthogonal to nuisance tangent space, so can be obtained by rescaling efficient score:

$$\phi(x) = \mathbb{E}\left[\psi(x)\psi(x)'\right]^{-1}\psi(x)$$

Intuition: tells you how much a single data point "influences" your point estimate

If $\hat{\theta}$ is efficient, we say it is asymptotically linear:

$$\sqrt{n}(\hat{\theta} - \theta) = \frac{1}{n}\sum_{i=1}^{n}\phi(x_i) + o_p(n^{-1/2}) \xrightarrow{d} N(0, I^{\theta\theta'})$$

ϕ gives "impact" of x_i on $\hat{\theta}$

Semiparametric Variance Bound

Bound given by variance of influence function:

$$E[\phi(X_i)\phi(X_i)'] = \left(E[\psi(X_i)\psi(X_i)']\right)^{-1} \quad \text{(SVB)}$$

$\forall\ \hat{\theta}$ semiparametric, $\underbrace{Var(\hat{\theta}) \geq SVB}$

if $=$, $\hat{\theta}$ "semiparametrically efficient"

(many such estimators, coming soon)

Simpler way to calculate variance bound.

Differentiable Parameter: θ differentiable if $\exists a(\cdot)$ s.t.

$$\frac{\partial \theta}{\partial \gamma}(\gamma) = E\left[a(z_i)' S(z_i;\gamma)\right]$$ Riesz Representer

If differentiability holds, we can characterize the Cramer-Rao bound via the Riesz representer:

$$E\left[\phi(z_i;\gamma)\phi(z_i;\gamma)'\right] = \underbrace{E\left[a(z_i)' S(z_i;\gamma)\right]\left(E\left[S(z_i;\gamma)S(z_i;\gamma)'\right]\right)^{-1} E\left[S(z_i;\gamma)a(z_i)\right]}_{SVB}$$

Where IF given by projection of RR on Full score function:

$$\phi(z;\gamma) = S(z;\gamma)'\left(E\left[S(z;\gamma)S(z;\gamma)'\right]\right)^{-1} E\left[S(z;\gamma)'a(z)\right]$$

Can estimate SVB by just estimating each of the unknown components

Econ 272/Mgtecon 607 - Section 6

Amar Venugopal

Stanford University

Spring 2025

Outline

1. Announcements

2. Discussion Questions

3. Doubly Robust Methods

4. Linear IV

5. Practice Problems

Announcements

- Makeup Lectures
  - May 14, 21 $\rightarrow$ 8-9:50am GSB C105
  - June 2, 4 no lecture (maybe section?) $\rightarrow$ May 22, 4:30-6:20pm STLC 114
- EOY Pizza Party: Date/Time TBD (June 5-12?)

Hints for Problem Set 5 (Due May 11, 11pm)

- Imbens (2004)[1] may be helpful in finding ATT estimators
- When bootstrapping, you must re-estimate relevant propensity scores for each bootstrap sample!
- If you have any covariates that are constant/perfectly colinear within a given stratum, you can just drop them
- For theoretical calculations: think about what we can add/subtract to the propensity score to express it as the difference of the covariate densities. Would doing so change the variance estimate?

[1] NONPARAMETRIC ESTIMATION OF AVERAGE TREATMENT EFFECTS UNDER EXOGENEITY: A REVIEW

1. Announcements
2. Discussion Questions
3. Doubly Robust Methods
4. Linear IV
5. Practice Problems

1. Announcements

2. Discussion Questions

3. Doubly Robust Methods

4. Linear IV

5. Practice Problems

Quick Recap: Semiparametric Efficiency

- **Goal:** Given a model with finite PoI θ and ∞-dim nuisance param $h(\cdot)$, find the best variance possible for PoI
- **Big Idea:** Considering all values of $h(\cdot)$, find best var. for θ
- **Asymptotic Linearity:** Can express asymptotic dist. of $\hat{\tau}$ as a linear combo of influence function $\hat{\tau} = \tau + \frac{1}{N}\sum_{i=1}^{N} \phi(Y_i, w_i, X_i) + o_p(N^{-1/2})$
- **ATE Bound:** $V^{SPVB} = E\left[\frac{\sigma_1^2(X_i)}{e(X_i)} + \frac{\sigma_0^2(X_i)}{1-e(X_i)} + \left(\tau(X_i) - \tau\right)^2\right]$

Under unconfoundedness

Given: outcome models $\mu_1(x), \mu_0(x)$, propensity score $\hat{e}(x)$

1. **Influence Function:**

$$\tilde{\tau}^{IF} = \frac{1}{N}\sum_{i=1}^{N}\left[\mu_1(x_i) - \mu_0(x_i) + w_i\frac{y_i - \mu_1(x_i)}{e(x_i)} - (1-w_i)\frac{y_i - \mu_0(x_i)}{1-e(x_i)}\right]$$

↓ AIPW

2. **Regression:**

$$\hat{\tau}^{reg} = \frac{1}{N}\sum_{i=1}^{N}\mu_1(x_i) - \mu_0(x_i)$$

Infeasible because all rely on μ_1, μ_0, e

3. **IPW:**

$$\tilde{\tau}^{IPW} = \frac{1}{N}\sum_{i=1}^{N}\left[w_i\frac{y_i}{e(x_i)} - (1-w_i)\frac{y_i}{1-e(x_i)}\right]$$

Infeasible Estimators - Variance

1. **Influence Function:** Asymptotic linearity $\Rightarrow$ efficient

$$\sqrt{N}\left(\tilde{\tau}^{IF} - \tau\right) \xrightarrow{d} \mathcal{N}(0, V^{SPVB})$$

2. **Regression:** Lower than SPVB

(assume knowledge of $\mu \rightarrow$ more info $\rightarrow$ lower variance)

3. **IPW:** Higher than SPVB

$$\tilde{\tau}^{IPW} = \frac{1}{N}\sum_{i=1}^{N}\left(W_i \frac{Y_i}{e(X_i)} - (1-W_i)\frac{Y_i}{1-e(X_i)}\right)$$

Comparison: $\mathbb{V}^{Reg} \leq \mathbb{V}^{IF} = \mathbb{V}^{SPVB} \leq \mathbb{V}^{IPW}$

Feasible Estimators

Big Idea: Replace μ_1, μ_0, e with estimates $\hat{\mu}_1, \hat{\mu}_0, \hat{e}$

Variance Comparison: $\mathbb{V}^{Reg} = \mathbb{V}^{IF} = \mathbb{V}^{SPVB} = \mathbb{V}^{IPW}$

↓

generally prefer

→ Doubly Robust

$$\hat{\tau}^{AIPW} = \frac{1}{N}\sum_{i=1}^{N}(\hat{\mu}_1(X_i) - \hat{\mu}_0(X_i)) + W_i\frac{Y_i - \hat{\mu}_1(X_i)}{\hat{e}(X_i)} - (1 - W_i)\frac{Y_i - \hat{\mu}_0(X_i)}{1 - \hat{e}(X_i)}$$

Double Robustness: $\hat{\tau}^{AIPW} \xrightarrow{p} \tau$ if either μ or e correctly specified

consistent unbiased estimator

Solves WLS: $\min_{\alpha,\beta,\tau} \sum_{i=1}^{N} (Y_i - \alpha - \tau W_i - X_i\beta)^2 w(W_i, X_i)$

$= \frac{W_i}{e(X_i)} + \frac{1-W_i}{1-e(X_i)}$

$$\hat{\tau}^{AIPW} = \left[\frac{1}{N}\sum_{i=1}^{N}(\hat{\mu}_1(X_i) - \hat{\mu}_0(X_i))\right] + \left[\frac{1}{N}\sum_{i=1}^{N} W_i \frac{Y_i - \hat{\mu}_1(X_i)}{\hat{e}(X_i)} - (1 - W_i)\frac{Y_i - \hat{\mu}_0(X_i)}{1 - \hat{e}(X_i)}\right]$$

$\to 0$ $\to 0$

$E[\mu_1(x) - \mu_0(x)]$

$= E[E[Y_i(1) - Y_i(0) \mid X = x]]$

$= E[\tau(x)]$

$= \tau$

$E[\cdot] = 0$

Alternatively, WLS:

$\mu_w(x) = \alpha + \tau w + x\beta$

$E[Y - \mu_w(x) \mid X, w] = 0$

WLS obj $= 0$

$\mu_1(x) = E[Y \mid W = 1, X = x]$

$\hat{\mu}_1$: $Y \sim X, w$

Influence Function/AIPW - Correctly Specified Propensity

Key: in weighted sample $W_i \perp X_i$

$$\hat{\tau}^{AIPW} = \left[\frac{1}{N}\sum_{i=1}^{N} W_i \frac{Y_i}{\hat{e}(X_i)} - (1-W_i)\frac{Y_i}{1-\hat{e}(X_i)}\right] + \left[\frac{1}{N}\sum_{i=1}^{N}(\hat{e}(X_i) - W_i)\left(\frac{\hat{\mu}_1(X_i)}{\hat{e}(X_i)} + \frac{\hat{\mu}_0(X_i)}{1-\hat{e}(X_i)}\right)\right]$$

$\mathbb{E}[\cdot] = \mu_1$ $\quad$ $\mathbb{E}[\cdot] = \mu_0$

$\mathbb{E}[Y|W=1] - \mathbb{E}[Y|W=0]$

$= \tau$

$\mathbb{E}[V_i] - \mathbb{E}[V_i]$

$= 0$

Correct specification:

$\hat{e} \to e$

$\hat{\mu} \to \mu$

Cross-Fitting/Sample Splitting

Idea: Deriving formal properties if estimate μ, e on separate sample from τ (enforcing independence)

Procedure:

1. Split sample into K folds, $I_i = \{1, \dots, K\}$
2. Estimate $\hat{\mu}_w^j(x)$, $\hat{e}^j(x)$ on sample $I_i \neq j$
3. Estimate $\hat{\tau}^j(\hat{\mu}_w^j, \hat{e}^j)$
4. Average $\frac{1}{K} \sum_{j=1}^{K} \hat{\tau}^j$ to get efficient estimator

Goal: Find optimal assignment of units to treatment

- **Policy Rule:** $\pi : X \to \{0, 1\}$
- **Desirable Properties:**
  - $\hat{\pi}(x) = 1\{\mu_1(x) - \mu_0(x) \geq 0\}$
  - Interpretable, well-behaved assignment rules
    e.g. $\pi(x) = 1\{x'\beta \geq 0\}$

Estimating Optimal Policies

Goal: Maximize gain over treating nobody : $V(\pi) = \mathbb{E}\left[\pi(X_i)\left(Y_i(1) - Y_i(0)\right)\right]$
$= \mathbb{E}\left[\pi(X_i)\,\tau(X_i)\right]$

ATE Estimator: $\hat{\tau}^{IPW} = \frac{1}{n}\sum_{i=1}^{n} \hat{\Gamma}_i$ → exactly summation term from before

Effect Estimator: $\hat{V}(\pi) = \frac{1}{n}\sum_{i=1}^{n} \pi(X_i)\,\hat{\Gamma}_i$

Policy Estimator: $\hat{\pi} = \underset{\pi \in \Pi}{\arg\max} \underbrace{\frac{1}{n}\sum_{i=1}^{n} \pi(X)\,\hat{\Gamma}_i}_{= \hat{V}(\pi)}$ ↳ typically estimated via cross-fitting

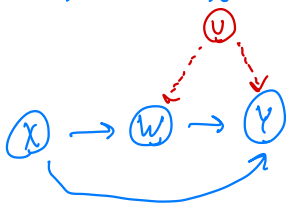
1. Announcements
2. Discussion Questions
3. Doubly Robust Methods
4. Linear IV
5. Practice Problems

What if unconfoundedness breaks down?

In observational settings, **unconfoundedness** let's us "pretend" that we are dealing with a randomized experiment.

When might unconfoundedness be violated?

We have some unobserved variable that impacts both outcome and treatment



Ex. $\text{Wage}_i = \alpha + \beta \text{Educ}_i + \varepsilon_i$

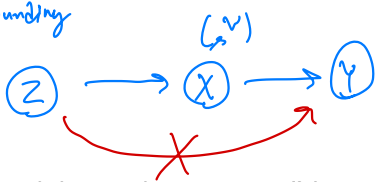
U = ability

Terminology

- **Endogeneity:** Variable correlated with error term (contain effects of all unobserved components)
- **Omitted Variable Bias:** When we cannot control for confounder, estimated coeffic. will be biased
- **Reverse Causality:** Values regressor takes on may be determined by function for outcome
  e.x. regressor = price, outcome = demand } price set by demand function, so price endogenous
  $d = D(p, U)$

Instrumental Variables

Goal: Find exogenous "instrument" that allows us to remove effect of confounding



To obtain valid causal estimates, IV must satisfy:

- **Exclusion:** Z affects Y only through X → not testable (argued)
- **Relevance:** Z has causal effect on X → testable ("strong 1st stage")

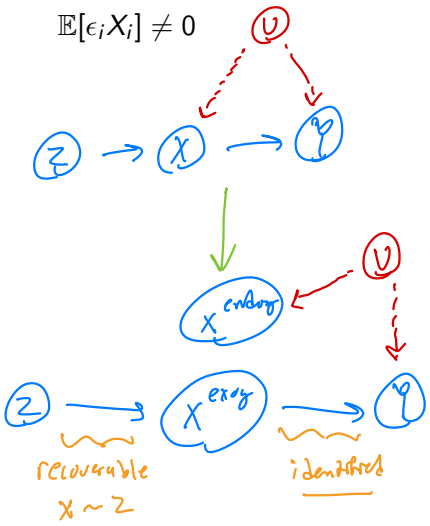
Estimation - Setting

$$Y_i = \alpha + \beta X_i + \epsilon_i \qquad \mathbb{E}[\epsilon_i X_i] \neq 0$$

Goal: Recover causal effect of X on Y removing bias in $\hat{\beta}^{naive}$

Estimand:

$$\beta = \frac{Cov(Y_i, Z_i)}{Cov(X_i, Z_i)} = \frac{Cov(Y_i, Z_i)/Var(Z_i)}{Cov(X_i, Z_i)/Var(Z_i)}$$



Estimation Strategy #1 - Two Stage Least Squares (TSLS)

Big Idea:

① Regress X on Z, get $\hat{X}$

② Regress Y on $\hat{X}$

Estimator:

① $X_i = \alpha_0 + \alpha_1 Z_i + u_i$ → contains confounding

② $Y_i = \beta_0 + \beta^{TSLS} \hat{X}_i + \varepsilon_i$ → $\hat{X} = \hat{\alpha}_0 + \hat{\alpha}_1 Z \perp \varepsilon_i$

$$\hat{\beta}^{TSLS} = \frac{Cov(Y_i, Z_i)}{Cov(X_i, Z_i)} = (Z'X)^{-1}(Z'Y)$$

Inference:

Under homoskedasticity: $\hat{\beta}^{TSLS} - \beta \approx N(0, \sigma^2(\hat{X}'\hat{X})^{-1})$

is only from 2nd stage

larger than OLS

Hausman Test

Goal: If X_i exogenous, TSLS and OLS consistent.
But, only OLS efficient → only want TSLS if really worth it.

Note: a good, valid IV required

H_0: $\varepsilon_i \perp X_i \rightarrow \beta^{TSLS} = \beta^{OLS} \rightarrow \hat{\beta}^{TSLS} - \hat{\beta}^{OLS} \approx \mathcal{N}(0, \hat{V}^{TSLS} - \hat{V}^{OLS})$

Test Statistic: $t = \dfrac{\hat{\beta}^{TSLS} - \hat{\beta}^{OLS}}{\sqrt{\hat{V}^{TSLS} - \hat{V}^{OLS}}}$ → if large, reject null and conclude X truly endogenous

Econ 272/Mgtecon 607 - Section 7

Amar Venugopal

Stanford University

Spring 2025

Outline

1. Announcements & Discussion Questions
2. Local Average Treatment Effects (LATE)
3. Weak Instruments & Many Instruments
4. Practice Problems

1. Announcements & Discussion Questions

2. Local Average Treatment Effects (LATE)

3. Weak Instruments & Many Instruments

4. Practice Problems

Announcements

- Next Week Lectures
  - May 21 $\rightarrow$ 8-9:50am GSB C105
  - June 2, 4 no lecture (maybe section?) $\rightarrow$ May 22, 4:30-6:20pm STLC 114

1. Announcements & Discussion Questions

2. Local Average Treatment Effects (LATE)

3. Weak Instruments & Many Instruments

4. Practice Problems

Motvation: Effect Heterogeneity

Goal: Allow for: heterogeneity in TE across units
more flexible functional forms
→ endogeneity ≠ correlation with residual

Challenge: Cannot ID population ATE
Can only ID ATE for sub-pop'n → compliers

- **(Binary)** Z: Instrument
- **(Binary, endogenous)** W: $W_i = W_i(Z_i) = \begin{cases} W_i(1) & \text{if } Z_i = 1 \\ W_i(0) & \text{if } Z_i = 0 \end{cases}$
- Y: $Y_i = Y_i(W_i) = Y_i(W_i(Z_i)) = Y_i(Z_i, W_i)$

Key Assumption: Independence of Instrument

$$Z_i \perp\!\!\!\perp (Y_i(0), Y_i(1), W_i(0), W_i(1))$$

→ IV does not directly affect outcome (nonparametric assumption)

Conditions:

- **Randomization of Z:**

$Z_i \perp\!\!\!\perp (W_i(0), W_i(1), Y_i(0,0), Y_i(1,0), Y_i(0,1), Y_i(1,1))$

- **Exclusion:** → not testable, must be argued

$\forall z', z, w : Y_i(z, w) = Y_i(z', w)$

True Compliance Types

$Z_i = 0$

		$W_i(0)$ = 0	$W_i(0)$ = 1
$W_i(1)$	0	never-taker	defiers
$Z_i = 1$	1	complier	always-takers

Observed Compliance Types

		Z_i = 0	Z_i = 1
W_i	0	never-taker / complier	never-taker / defier
	1	always-taker / defier	always-taker / complier

Monotonicity

Goal: Disambiguate type of unit from (Z_i, W_i)

Assumption: Monotonicity: $W_i(1) \geq W_i(0)$ $\Leftrightarrow$ no defiers

$=$ $\Leftrightarrow$ always-/never-takers

$>$ $\Leftrightarrow$ compliers

↓ sometimes testable

more plausible when IV incentive

		Z_i = 0	Z_i = 1
W_i	0	never-taker / complier	never-taker
	1	always-taker	always-taker / complier

Conditional Means

Let proportions of compliers, never-takers, and always-takers be given by π_c, π_n, π_a.

- $\mathbb{E}[W_i|Z_i = 0] = \pi_a$
- $\mathbb{E}[W_i|Z_i = 1] = \pi_a + \pi_c$

ID $\pi_a, \pi_c, \pi_n = 1 - \pi_a - \pi_c$

- $\mathbb{E}[Y_i|W_i = 0, Z_i = 0] = \frac{\pi_c}{\pi_c + \pi_n} \mathbb{E}[Y_i(0)|c] + \frac{\pi_n}{\pi_c + \pi_n} \mathbb{E}[Y_i(0)|n]$
- $\mathbb{E}[Y_i|W_i = 1, Z_i = 0] = \mathbb{E}[Y_i(1)|a]$
- $\mathbb{E}[Y_i|W_i = 0, Z_i = 1] = \mathbb{E}[Y_i(0)|n]$
- $\mathbb{E}[Y_i|W_i = 1, Z_i = 1] = \frac{\pi_c}{\pi_a + \pi_c} \mathbb{E}[Y_i(1)|c] + \frac{\pi_a}{\pi_a + \pi_c} \mathbb{E}[Y_i(1)|a]$

Estimating Complier ATE

Why only compliers? Only group with members in both treatment groups

Estimand (LATE): $LATE = E[Y_i(1) - Y_i(0) \mid \text{complier}]$

Estimation Strategies:

- Use previous conditional means
- ILS

- $Y \sim Z$ **Slope Coefficient:** $E[Y_i|Z_i=1] - E[Y_i|Z_i=0]$
  $(E[Y_i(1)|c]\pi_c + E[Y_i(0)|n]\pi_n + E[Y_i(1)|a]\pi_a) = E[Y_i(1)-Y_i(0)|c]\pi_c$
  $= -(E[Y_i(0)|c]\pi_c + E[Y_i(0)|n]\pi_n + E[Y_i(1)|a]\pi_a)$
- $W \sim Z$ **Slope Coefficient:** $E[W_i|Z_i=1] - E[W_i|Z_i=0] = \pi_c$
- **Ratio:** $\tau^{IV} = \frac{E[Y_i|Z_i=1] - E[Y_i|Z_i=0]}{E[W_i|Z_i=1] - E[W_i|Z_i=0]} = E[Y_i(1)-Y_i(0)|c]$

ATE Bounds (Binary Outcomes)

Goal: Partial ID of full pop'n ATE

$$ATE = E[Y_i(1) - Y_i(0)] = E[Y_i(1) - Y_i(0) \mid a]\,\pi_a + E[Y_i(1) - Y_i(0) \mid n]\,\pi_n + E[Y_i(1) - Y_i(0) \mid c]\,\pi_c$$

→ cannot estimate $E[Y_i(0) \mid a]$, $E[Y_i(1) \mid n]$

LB: $E[Y_i(0) \mid a] = 1$, $E[Y_i(1) \mid n] = 0$

UB: $E[Y_i(0) \mid a] = 0$, $E[Y_i(1) \mid n] = 1$

What if we (mistakenly) included Z as a control, rather than an instrument?

DiM estimator, after adjusting for Z_i, becomes:

$$E[Y_i \mid W_i = 1, Z_i = 1] - E[Y_i \mid W_i = 0, Z_i = 1] = E[Y_i(1) \mid W_i = 1, Z_i = 1] - E[Y_i(0) \mid W_i = 0, Z_i = 1]$$

$$= \frac{\pi_a}{\pi_a + \pi_c} E[Y_i(1) \mid a] + \frac{\pi_c}{\pi_a + \pi_c} E[Y_i(1) \mid c] - E[Y_i(0) \mid n]$$

not causally interpretable

1. Announcements & Discussion Questions

2. Local Average Treatment Effects (LATE)

3. Weak Instruments & Many Instruments

4. Practice Problems

"Weak" Instruments (F-stat < 10)

Challenge: Standard asymptotics generally rely on nonzero correlation between IV and endog

strong 1st stage

If corr ≈ 0, approx inaccurate even in large samples

Goal: Adjust s.e. to allow for this violation

Conventional Asymptotics (Fixed # Instruments, Fixed Parameters)

- **LIML vs. TSLS:** $\sqrt{N}\left(\hat{\beta}_{LIML} - \hat{\beta}_{TSLS}\right) \to 0$ s.e. should be similar
- **Just-Identified:** # IV = # endog
  

  (just become wider)
  CI not uniformly valid, but still "decent" coverage in "reasonable" regions
- **Over-Identified:** # IV > # endog
  CI potentially very misleading (especially TSLS)
- **Standard Estimator:** $X_i = \pi_0 + \pi_1 Z_i + \eta_i$, $Y_i = \alpha_0 + \alpha_1 Z_i + \nu_i$, $\beta_1 = \frac{\alpha_1}{\pi_1} \to \hat{\beta}^{IV} = \frac{\frac{1}{N}\sum (Y_i - \bar{Y})(Z_i - \bar{Z})}{\frac{1}{N}\sum (X_i - \bar{X})(Z_i - \bar{Z})}$
- **Concentration:** $\lambda = \pi_1^2 \sum \frac{(Z_i - \bar{Z})^2}{\sigma_\eta^2}$ → strength of instrument
  weak: $\lambda \to 0$

Big Idea: Let 1st stage param. depend on sample size: $\pi_{1N} = \frac{C}{\sqrt{N}}$

Then, $\lambda \to C^2 \cdot V(Z_i)$

Demonstrates that for any sample size, coverage can vary a lot from desired level.

Suggested methods for uniformly valid CIs.

Alternative CIs

Goal: Provide uniformly valid CIs

- **Anderson-Rubin CI:**
- **Kleibergen Test:**
- **Moreira's Similar Test:**

Many Weak Instruments

2 problems:

1. Bias in standard estimators
2. Misleadingly small s.e.

} not fixable via bootstrap

Main take-away: LIML works well, but s.e. must be adjusted

2 strategies to adjust standard errors:

1. Better
2. REQML

Bekker-Adjusted Standard Errors → generally recommended

Big Idea: TSLS is not consistent in large sample, but 2SLS is

Considers new asymptotics to derive uniformly valid CIs

$$V^{Bekker} = (V^{2SLS}) \left(\text{increasing } \# \text{ of instruments}\right)$$

Random Effects Quasi Maximum Likelihood (REQML)

Big Idea: Model 1st stage coefs. π_{ic} as iid draws from $N(\mu_\pi, \sigma^2_\pi)$

Assume joint normality, normalized Z_i, likelihood function:

$$L(\beta_0, \beta_1, \pi_0, \mu_\pi, \sigma^2_\pi, \Omega)$$

6-parameter problem

TSLS: posterior mode given flat priors ($\sigma^2_\pi = \infty$)

LIML: $\mu_\pi = 0$, Ω known

Correction: REQML corrections generally larger than Bekker, because ML estimator for σ^2_π will take into account noise in $\hat{\pi}$

Econ 272/Mgtecon 607 - Section 8

Amar Venugopal

Stanford University

Spring 2025

Outline

1. Announcements & Discussion Questions

2. Difference-in-Differences

3. Synthetic Control

4. Alternative/New Panel Methods

5. Practice Problems

Announcements

- Last section: Wednesday June 4, 11:30am-1:20pm, GSB Z301 (in place of regular lecture)
  - Focus on exam prep, review + practice problems (?)
- Last Problem Set: HW7 due June 4, 11pm
- Final Exam: June 11, 8:30-11:30am, 200-002
- Pizza Party! June 9 (exact time TBD) - RSVP by June 1

1. Announcements & Discussion Questions
2. Difference-in-Differences
3. Synthetic Control
4. Alternative/New Panel Methods
5. Practice Problems

Key feature: Observe both units $(i=1,\dots,N)$ and time periods $(t=1,\dots,T)$

Y_{it}

3 main types of panel data:

1. **Proper Panel:** Observe same units in each period

   Y_{it}, $i=1,\dots,N$, $t=1,\dots,T$

2. **Repeated Cross-Sections:** Observe different individuals in each time period

   Y_i for $i=1,\dots,N$ in period T_i

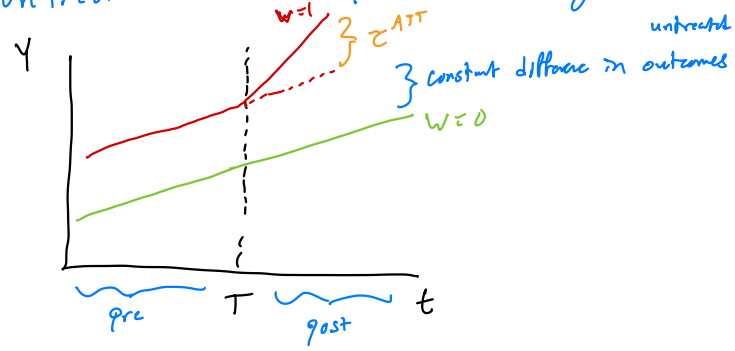
   G_i = group assignment (treatment/control)

3. **Repeated Cross-Sections with Multiple Groups:**

   Same, $G_i = \{1,\dots,G\}$

Key Assumption: Parallel Trends (2 × 2)

Big Idea: Untreated outcomes parallel across groups



Testable? Not directly

But, can assess by checking if parallel pre-treatment

Repeated Cross-Section, 2 Groups × 2 Periods - Setup

Observed data: (Y_i, G_i, T_i)

$G \in \{0,1\}$ $\quad T_i \in \{0,1\}$

Treatment assignment: Group 1 is treated in period 1

$W_i = G_i T_i$, $\quad Y_i = (1 - W_i) Y_i(0) + W_i Y_i(1)$

Modeling Assumption: $Y_i(0) = \alpha + \beta T_i + \gamma G_i + \varepsilon_i$

(β: time effect; γ: group effect)

→ can add covariates $+ \delta^T X_i$

Typically assume constant τ: $Y_i = \alpha + \beta T_i + \gamma G_i + \tau W_i + \varepsilon_i$

If τ heterogeneous, identifies average effect

Repeated Cross-Section, 2 Groups × 2 Periods - Estimation

Estimand: $\tau^{DiD} = \left(E[Y_i \mid G_i = 1, T_i = 1] - E[Y_i \mid G_i = 1, T_i = 0] \right)$ — treated group evolution

$- \left(E[Y_i \mid G_i = 0, T_i = 1] - E[Y_i \mid G_i = 0, T_i = 0] \right)$ → control group evolution

ATT

Estimator: $\hat{\tau}^{DiD} = \bar{Y}_{11} - \bar{Y}_{10} - \left(\bar{Y}_{01} - \bar{Y}_{00} \right)$

$\min_{\alpha, \beta, \gamma, \tau} \sum_{i=1}^{N} \left(Y_i - \alpha - \beta T_i - \gamma G_i - \tau W_i \right)^2$ } both equivalent

Variance:

$$V(\tau)_{robust} = \frac{\sigma_{11}^2}{N_{11}} + \frac{\sigma_{10}^2}{N_{10}} + \frac{\sigma_{01}^2}{N_{01}} + \frac{\sigma_{00}^2}{N_{00}}$$

Under homoskedasticity: $\sigma_{11}^2 = \sigma_{10}^2 = \sigma_{01}^2 = \sigma_{00}^2 \equiv \sigma^2$

(2 periods)

Observed data: $(G_i, Y_{i0}, Y_{i1}, W_{i0}, W_{i1})$

Modeling Assumption: $Y_{it} = \alpha + \beta t + \gamma G_i + \tau W_{it} + \varepsilon_{it}$

Estimator: Could do double difference

or

$$Y_{it} = \mu + \alpha_i + \rho_t + \tau W_{it} + \varepsilon_{it}$$

→ TWPE

α_i, ρ_t: fixed effects

Parallel Trends vs. Unconfoundedness

Unconfoundedness Assumption: Independence conditional on lagged outcomes

$$Y_{i1} - Y_{i0} = \alpha + \beta Y_{i0} + \tau G_i + \eta_i$$

DiD Assumption: $Y_{i1} - Y_{i0} = \alpha + \tau G_i + \varepsilon_i$

How to choose? Unconfoundedness allows for flexible estimation of β

But DiD assumption Y_{i0} may be correlated with ε_i

Nice result: If (U) correct, but estimate DiD, overestimate τ

If DiD right, but estimate (U), underestimate τ

} yield bounds

Changes-in-Changes: Continuous Data (2×2)

Motivation: DiD sensitive to transformations of data, functional form

Modeling Assumption: $Y_i(0) = h(U_i, T_i)$, $h(\cdot)$ strictly increasing in U (unobserved)

$U_i \perp T_i \mid G_i$ → distribution of U_i vary across groups, but not time

$U_i = G_i T_i$

Estimand: $\tau^{ATT} = \underbrace{E[Y_i(1) \mid G_i = 1, T_i = 1]}_{\text{observed}} - \underbrace{E[Y_i(0) \mid G_i = 1, T_i = 1]}_{\text{estimate distribution}}$

Estimation: $F_{Y(0),11}(y) = F_{Y(0),10}\left(F^{-1}_{Y(0),00}\left(F_{Y(0),01}(y)\right)\right)$

observed

Modern Approaches: DiD With Heterogeneous Effects, Multiple Groups/Multiple Periods

Big Idea: In these settings, TWFE ≠ DiD if effects heterogeneous

3 proposed estimators:

- **Chaisemartin & d'Haultfoeuille (2020):** Weighted avg. of 1-period ahead DiD estimators
  - → never-treated may be fundamentally different
  - → constant additive effects more plausible in short run
- **Callaway & Sant'Anna (2021):** Compare evolution of outcomes to never-treated group
  - 1-period ahead may miss long-run effects
- **Sun & Abraham (2021):** Similar to CS, allowing for comparison to last-treated group if never-treated unavailable

All require addition/modification to parallel trends

1. Announcements & Discussion Questions

2. Difference-in-Differences

3. Synthetic Control

4. Alternative/New Panel Methods

5. Practice Problems

What kind of panel must we have? Proper panel

$$W = \begin{pmatrix} 0 & 0 & \cdots & 0 \\ 0 & 0 & & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix} \Big\} N$$

$\underbrace{\qquad\qquad}_{T}$

Estimand:

ATT

$$\tau^{ATT} = \frac{\sum_{i,t} W_{it}\left(Y_{it}(1) - Y_{it}(0)\right)}{\sum_{i,t} W_{it}}$$

In case with 1 unit i treated in 1 period T, this becomes:

$$\tau^{i} = \underbrace{Y_{iT}(1)}_{\text{observed}} - \underbrace{Y_{iT}(0)}_{\text{unobserved}}$$

Vertical vs. Horizontal Regression

Goal: $Y_{iT}(0)$

$$Y = \begin{pmatrix} Y_{11} & Y_{12} & \cdots & Y_{1T} \\ Y_{21} & Y_{22} & \cdots & Y_{2T} \\ \vdots & \vdots & & \vdots \\ Y_{N1} & Y_{N2} & \cdots & Y_{NT} \end{pmatrix}$$

Vertical Regression: $T-1$ obs., $N-1$ regressors

e.g. SC

Main idea: similar units behave similarly (stable relationship across time over units)

Horizontal Regression: $N-1$ obs., $T-1$ regressors

e.g. unconfoundedness

Main idea: history explains all (stable relationship across units over time)

Takeaway: If T large, N small (fat), go vertical

If skinny, go horizontal

Synthetic Control - Assumption

Factor Model: Assume movement in outcome is summarized by shared latent factors

$$Y_{it}(0) = \mu_i^T \lambda_t + \varepsilon_{it}$$

μ_i: loadings; λ_t: factor values

Synthetic Control - Estimation (No Covariates)

Ideal: Match units based on factor loadings

$$\sum_{j=1}^{N-1} \hat{w}_j \mu_j = \mu_N$$ → unobserved μ

In Practice: Instead match on pre-treatment outcomes

$$\hat{Y}_{NT}(0) = \sum_{j=1}^{N-1} \hat{w}_j Y_{jT}$$

$$\hat{w} = \underset{w \in \Delta^{N-1}}{\operatorname{argmin}} \sum_{t=1}^{T-1} \left(Y_{Nt} - \sum_{j=1}^{N-1} w_j Y_{jt} \right)^2$$

Synthetic Control - Estimation (Covariates)

Big Idea: Use covariates to "impute" missing lagged outcomes

If all lagged outcomes, yields case with no covariates.

2-step optimization:

1. **Outer:**

2. **Inner:**

1. Announcements & Discussion Questions

2. Difference-in-Differences

3. Synthetic Control

4. Alternative/New Panel Methods

5. Practice Problems

Big Idea: Merge insights from DiD and SC

Estimation: $\hat{\tau}^{SDiD} = \underset{\tau, \gamma, \alpha}{\operatorname{argmin}} \sum_i \sum_t \left(Y_{it} - \gamma_t - \alpha_i - \tau W_{it} \right)^2 \cdot \omega_i^{SDiD} \cdot \lambda_t$

$\left(Y_{it} - \gamma_t - \alpha_i - \tau W_{it} \right)$: DiD

ω_i^{SDiD}: SC ↓ weight units by similarity

λ_t: ↓ weight time periods by recency

Challenge: If I treated i, t, no averaging so no CLT

Placebo analyses generally used, but still have a choice to make:

1. **Placebo on Units:**

2. **Placebo on Time Periods:**

Both are valid on average

Use version for dimension with less heteroskedasticity.