Part B: Selection on Observables

B1: Matching and Regression Adjustment

Kirill Borusyak

ARE 213 Applied Econometrics

UC Berkeley, Fall 2024

General plan of attack

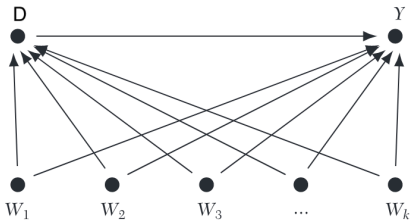
1. Selection on observables: Idea
2. Matching
3. Regression adjustment
4. Propensity score-based methods
5. Doubly-robust methods
6. ML methods for high-dimensional covariates
7. Violations of CIA and coefficient stability

Outline of B1

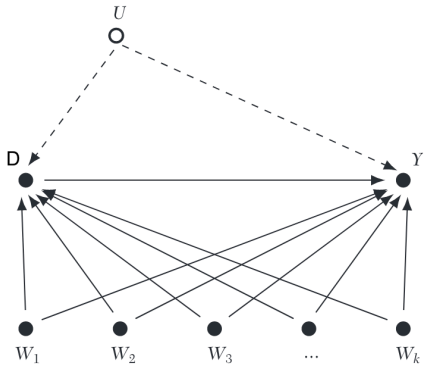
What if treatment is not randomly assigned?

One basic approach for causal inference with observational data is to control for observables

A: Unconfoundedness



B: Violation of Unconfoundedness



Identification assumptions

- Unconfoundedness = Ignorability = Conditional independence assumption (**CIA**) = Selection on observables: $(Y_i(0), Y_i(1)) \perp\!\!\!\perp D_i \mid X_i$
  - Sometimes viewed as a *definition* of a control variable
- **Overlap**: $0 < Pr(D_i = 1 \mid X_i) < 1$ on the support of X_i (testable)
  - $Pr(D_i = 1 \mid X_i)$ is called the **propensity score**

Identification under CIA + overlap

- Conditional average treatment effect $CATE(x) \equiv \mathbb{E}\left[Y_i(1) - Y_i(0) \mid X_i = x\right]$ is identified by
$$\mathbb{E}\left[Y_i \mid D_i = 1, X_i = x\right] - \mathbb{E}\left[Y_i \mid D_i = 0, X_i = x\right]$$
- $ATE = \mathbb{E}\left[CATE(X_i)\right]$ is identified by
$$\mathbb{E}\left[\mathbb{E}\left[Y_i \mid D_i = 1, X_i\right] - \mathbb{E}\left[Y_i \mid D_i = 0, X_i\right]\right]$$
where the outer expectation is taken w.r.t. (observed) distribution of X_i
- $ATT = \mathbb{E}\left[CATE(X_i) \mid D_i = 1\right]$ is identified by
$$\mathbb{E}\left[\mathbb{E}\left[Y_i \mid D_i = 1, X_i\right] - \mathbb{E}\left[Y_i \mid D_i = 0, X_i\right] \mid D_i = 1\right]$$
where the outer expectation is taken w.r.t. (observed) distribution of $X_i \mid D_i = 1$
- Estimation is non-trivial — stay tuned

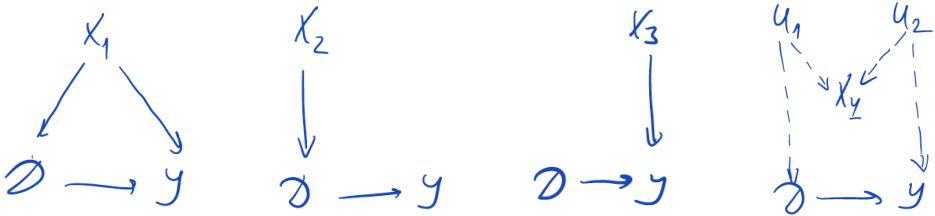
Is the CIA likely to hold?

- Economists are suspicious of the CIA: if units self-select into treatment based on observables, why not on unobservables, too?
- But there are some settings where it makes sense:
  1. RCT with unequal probabilities of treatment, e.g. varying by age
  2. If you have a model of the factors driving D in which none of them but X affect Y (or are correlated with factors affecting Y); or the other way round
     - ⋆ E.g. (IW Lecture 1): firms optimally choose whether to adopt a new technology (D) based on the costs of adoption
     - ⋆ Given technology, the output (Y) is decided by random factors like weather

     Understanding sources of variation in D is key for estimating the effects of D
  3. Pragmatic approach: causal inference is always about comparing *some* treated and control outcomes $\Longrightarrow$ might as well compare units with similar X
     - ⋆ Caveat: this does not always reduce bias!

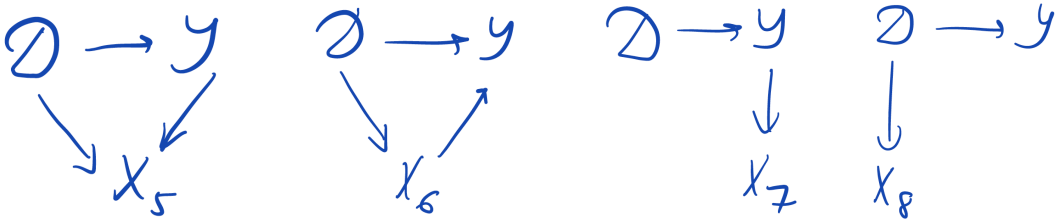
Good and bad controls: Pre-treatment controls

Which of these variables would you control for? Think identification and efficiency



Good and bad controls: Post-treatment controls

Which of these variables would you control for? Think identification and efficiency



Good and bad controls: Answer key

- X_1 Necessary for identification
- X_2 Bad for efficiency (but may be good for robustness)
- X_3 Good for efficiency
- X_4 "Collider," generates bias (but may serve as proxy for U_2 if U_2 affects D)
- X_5 Also collider, generates bias (e.g., $D =$ education, $Y =$ social skills, $X_5 =$ being employed)
- X_6 A mechanism, generate bias
- X_7 Another outcome, generate bias
- X_8 *Exercise for you*

How many controls?

Controlling for more pre-treatment variables:

- Typically relaxes CIA
  - But a clear narrative for a small set of controls is considered cleaner
- Weakens overlap but efficiency implications ambiguous
- Makes estimation trickier $\Longrightarrow$ machine learning solutions

Note: reasoning is different for covariate adjustments after an RCT

Leveraging CIA

- How can we estimate $ATE = \mathbb{E}\left[\mathbb{E}\left[Y_i \mid D_i = 1, X_i\right] - \mathbb{E}\left[Y_i \mid D_i = 0, X_i\right]\right]$ in practice?
  - If X_i is discrete and doesn't take too many values, can do exact matching
  - With continuous or high-dimensional X_i: approximate matching, regression adjustment, propensity score methods, doubly-robust methods, double machine learning
- Do simpler strategies, like regressing Y_i on D_i and X_i, recover anything useful?

Outline

Exact matching

- When X is discrete and there is overlap, we can estimate CATE directly by matching treated and untreated units with the same X_i:

$$\widehat{CATE}(x) = \frac{\sum_{i:X_i=x} D_i Y_i}{\sum_{i:X_i=x} D_i} - \frac{\sum_{i:X_i=x} (1 - D_i) Y_i}{\sum_{i:X_i=x} (1 - D_i)}.$$

- Then average them into ATE (*exercise*: rewrite for ATT)

$$\widehat{ATE} = \sum_x \hat{P}(X = x) \cdot \widehat{CATE}(x) \equiv \frac{1}{N} \sum_i \widehat{CATE}(X_i)$$

- Example: Angrist (1998) study of how voluntary military service affects employment and earnings
  - X = all combinations of race, application year, years of schooling, Armed Forces Qualification Test score group, and year of birth

Approximate matching

- With continuous (or high-dimensional discrete) variables, can't match exactly
- For each treated unit, choose R untreated ones closest to X_i
  - Mahalanobis distance: $d(X_i, X_j)^2 = (X_i - X_j)' \operatorname{Var}[X]^{-1} (X_i - X_j)$
  - Or diagonal version: $d(X_i, X_j)^2 = (X_i - X_j)' \operatorname{diag}\left(\operatorname{Var}[X]^{-1}\right) (X_i - X_j)$
  - Typically $R = 1$; also "radius" or "caliper" matching: choose all j with $d(X_i, X_j) < \bar{d}$
- Matching with replacement: multiple treated units can match to the same control
- Without replacement: only match to controls previously unmatched (if $N_0 > N_1$)
  - Warning: the ordering of treated units matter
- This yields ATT. To get ATE, also find treated matches to each control unit

Matching caveats

- **Bias** from imperfect matches is of order $N_0^{-1/\dim X}$ (Abadie and Imbens 2002) — curse of dimensionality
  - Bias is small with $N_0 \gg N_1$ and if $\dim X = 1$
  - But $\dim X = 2 \implies$ same order as SE; $\dim X > 2 \implies$ bias dominates
  - Use Abadie-Imbens bias correction
- **Inference** is complicated. Asymptotic variance derived by Abadie, Imbens (2006)
  - Bootstrap fails (Abadie and Imbens 2008)
- Matching is **inefficient** with fixed R
  - But efficient with $R \to \infty$ and with bias correction (Lin, Ding, Han 2023)
- Matching can be **computationally difficult**

Outline

Regression adjustment

- Under CIA, we have for $d = 0, 1$

$$\mathbb{E}\left[Y(d) \mid X\right] = \mathbb{E}\left[Y(d) \mid D = d, X\right] = \mathbb{E}\left[Y \mid D = d, X\right] \equiv h_d(X)$$

- If we estimate $h_0(\cdot)$ and $h_1(\cdot)$ by regression methods, we get

$$\widehat{ATE} = \frac{1}{N}\sum_i \left(\hat{h}_1(X_i) - \hat{h}_0(X_i)\right), \qquad \widehat{ATT} = \frac{1}{N_1}\sum_i \left(\hat{h}_1(X_i) - \hat{h}_0(X_i)\right) D_i$$

- Since average fitted values equal average outcomes, we also get the **imputation** representation:

$$\widehat{ATT} = \frac{1}{N_1}\sum_i \left(Y_i - \hat{h}_0(X_i)\right) D_i \qquad \text{and}$$

$$\widehat{ATE} = \frac{1}{N}\sum_i \left\{\left(Y_i - \hat{h}_0(X_i)\right) D_i + \left(\hat{h}_1(X_i) - Y_i\right)(1 - D_i)\right\}$$

Estimating $h_0(\cdot), h_1(\cdot)$

- Can use nonparametric regression, e.g. local linear regression
  - For each x, estimate $h_d(x)$ by an intercept from a regression of Y_i on $(X_i - x)$, keeping observations in the neighborhood of x (and with $D_i = d$) only
- If $\mathbb{E}[Y(d) \mid X] = \gamma_d' X$ is linear in X (e.g. X is saturated): Oaxaca-Blinder estimator
  - Run linear regressions of Y on X within treated/control groups separately
  - Or a single fully-interacted regression (for X_i including an intercept)

$$Y_i = \gamma_0' X_i + \tau' X_i D_i + \text{error}_i, \qquad \widehat{ATE} = \hat{\tau}' \bar{X}$$

  - Or its convenient reformulation

$$Y_i = \gamma_0' X_i + \beta D_i + \tau' \left(X_i - \bar{X}\right) D_i + \text{error}_i, \qquad \widehat{ATE} = \hat{\beta}$$

  - Note: interactions are helpful even if you are not interested in effect heterogeneity

Regression adjustment with homogeneous effects

- If causal effects are homogeneous, $Y_i = \beta D_i + Y_i(0)$
- If additionally $\mathbb{E}[Y(0) \mid X] = \gamma' X$ is linear,

$$Y_i = \beta D_i + \gamma' X_i + \varepsilon_i, \qquad \mathbb{E}[\varepsilon_i \mid D_i, X_i] = 0$$

  $\Rightarrow$ Regression that linearly controls for X_i identifies the causal effect

- Overlap now matters for robustness to violations of linearity:
  $\hat{\beta} = \left(\overline{Y_1} - \overline{Y_0}\right) - \hat{\gamma}' \left(\overline{X_1} - \overline{X_0}\right)$
  (because $\hat{\beta}$ can be obtained by regressing $Y_i - \hat{\gamma}' X_i$ on D_i — *why?*)
  - If the distribution of X is the same among treated and controls, $\hat{\gamma}$ does not matter $\Longrightarrow$ does not matter which nonlinear terms and interactions are included in X
  - If the distributions are very different, regression is doing a lot of extrapolation, and functional form becomes important

Uninteracted regression with heterogeneous effects

Mostly Harmless Econometrics advocates for $Y_i = \beta D_i + \gamma' X_i + \varepsilon_i$ regressions even when the effects are heterogeneous. Why?

- Assume the propensity score $p(X_i) \equiv \mathbb{E}[D_i \mid X_i] = Pr(D_i = 1 \mid X_i)$ is linear in X_i
  - Angrist (1998) focused on saturated controls $X_i \Longrightarrow$ trivially satisfied
- Then

$$\beta_{OLS} = \frac{\mathbb{E}[CATE(X_i) \cdot \omega(X_i)]}{\mathbb{E}[\omega(X_i)]}, \qquad \omega(X_i) = \mathrm{Var}[D_i \mid X_i] = p(X_i)(1 - p(X_i))$$

  - Groups with $p(X_i) \approx 1/2$ get the most weight (relative to their size)
  - Groups where overlap is limited ($p(X_i) \approx 0$ or $p(X_i) \approx 1$) get little weight
  - $\beta_{OLS} = ATE$ if $CATE(X_i)$ is constant, $\omega(X_i)$ is constant, or they are uncorrelated with each other
- *Note*: linearity of $\mathbb{E}[Y_i(0) \mid X_i]$ is not needed

Variance weighting: Proof

- By linearity of the p-score, partialling out X_i from D_i yields residuals $\tilde{D}_i = D_i - \mathbb{E}[D_i \mid X_i]$
- By Frisch-Waugh-Lovell, $\beta_{OLS} = \mathbb{E}\left[\tilde{D}_i Y_i\right] / \mathbb{E}\left[\tilde{D}_i D_i\right]$ (where $\mathbb{E}\left[\tilde{D}_i D_i\right] = \mathrm{Var}\left[\tilde{D}_i\right]$)
- Using CIA and $\mathbb{E}\left[\tilde{D}_i \mid X_i\right] = 0$,

$$\begin{aligned}\mathbb{E}\left[\tilde{D}_i Y_i\right] &= \mathbb{E}\left[\mathbb{E}\left[\tilde{D}_i \left(Y_i(0) + \left(Y_i(1) - Y_i(0)\right) D_i\right) \mid X_i\right]\right] \\ &= \mathbb{E}\left[CATE(X_i) \cdot \mathbb{E}\left[\tilde{D}_i D_i \mid X_i\right]\right] = \mathbb{E}\left[CATE(X_i) \cdot \mathrm{Var}\left[D_i \mid X_i\right]\right]\end{aligned}$$

- Analogously, $\mathbb{E}\left[\tilde{D}_i D_i\right] = \mathbb{E}\left[\mathrm{Var}\left[D_i \mid X_i\right]\right]$

Multi-valued and multiple treatments

This is easy to extend to multi-valued (e.g. continuous) treatments, still assuming linear $\mathbb{E}[D \mid X]$ (see MHE p.58 for details)

$$\beta_{OLS} = \mathbb{E}\left[\int \frac{\partial \mathbb{E}\left[Y(\tilde{d}) \mid X\right]}{\partial \tilde{d}} \omega(\tilde{d}, X) d\tilde{d}\right] \Big/ \mathbb{E}\left[\int \omega(\tilde{d}, X) d\tilde{d}\right]$$

where

$$\begin{aligned}\omega(\tilde{d}, x) &= \mathrm{Cov}\left[\mathbf{1}\left[D \geq \tilde{d}\right], D \mid X\right] \\ &= \left(\mathbb{E}\left[D \mid D \geq \tilde{d}, X\right] - \mathbb{E}\left[D \mid D < \tilde{d}, X\right]\right) P\left(D \geq \tilde{d} \mid X\right) P\left(D < \tilde{d} \mid X\right)\end{aligned}$$

But extra care is needed with multiple treatments, even dummies of multi-valued treatments (Goldsmith-Pinkham, Hull, and Kolesar, forthcoming)

Outline

Application: NSW

Lalonde (1986) studied National Supported Work (NSW) Demonstration

- A government program in 1970s for groups with weak labor-force attachment (e.g. ex-convicts)
- Guaranteed a job for 9–18 months and paid for it. Expensive: \$7–9k per person
- How did going through NSW in 1976–77 affect wage earnings in 1978?

NSW was designed as an RCT: random selection among qualified applicants

- So we know the ATE (=ATT) for the population of applicants
- But could one get the right answer by covariate adjustment?
- This setting has become the testing ground for CIA estimators

Application: NSW (2)

- Lalonde constructed several control groups:
  - Full CPS and PSID
  - Same but pre-screened: e.g. unemployed in 1976 and below poverty line in 1975
  - We focus on male workers
- And applied several estimators for ATT
  - Both cross-sectional (our focus) and diff-in-diffs (1978 minus 1975)
  - (Also Heckit-style selection corrections and more)

NSW: Covariate imbalance

Variable	NSW		Full Samples	
	Treated	Control	CPS-1	CPS-3
	(1)	(2)	(3)	(4)
Age	25.82	25.05	33.23	28.03
Years of schooling	10.35	10.09	12.03	10.24
Black	0.84	0.83	0.07	0.20
Hispanic	0.06	0.11	0.07	0.14
Dropout	0.71	0.83	0.30	0.60
Married	0.19	0.15	0.71	0.51
1974 earninigs	2,096	2,107	14,017	5,619
1975 earnings	1,532	1,267	13,651	2,466
Number of Obs.	185	260	15,992	429

(From MHE Table 3.3.2)

Reporting covariate imbalance

- It's good to report **normalized** (standardized) **differences** between treatment and control groups:

$$\text{Nor.Dif.} = \frac{\overline{X_1} - \overline{X_0}}{\sqrt{\left(\sigma_1^2 + \sigma_0^2\right)/2}}$$

  where σ_1, σ_0 are SD of X_i in treated & control groups

- This is *not* the t-stat for $\mathbb{E}[X \mid D = 1] = \mathbb{E}[X \mid D = 0]$:

$$t = \frac{\overline{X_1} - \overline{X_0}}{\sqrt{\sigma_1^2/N_1 + \sigma_0^2/N_0}}$$

  - t-stat is larger in larger samples — but larger samples are not an indication of a more difficult causal inference problem

NSW: Regression adjustment estimates

Specification	Full Samples		
	NSW	CPS-1	CPS-3
	(1)	(2)	(3)
Raw Difference	1,794	-8,498	-635
	(633)	(712)	(657)
Demographic controls	1,670	-3,437	771
	(639)	(710)	(837)
1975 Earnings	1,750	-78	-91
	(632)	(537)	(641)
Demographics, 1975 Earnings	1,636	623	1,010
	(638)	(558)	(822)
Demographics, 1974 and 1975 Earnings	1,676	794	1,369
	(639)	(548)	(809)

(From MHE Table 3.3.3. Demographics are age, years of schooling, dummies for Black, Hispanic, high school dropout, and married)

- Lalonde concludes it'd be hard for an analyst to pick a good estimate
- Note: Oaxaca-Blinder doesn't help either (IW Lecture 2, Table 2)