Part B: Covariate Adjustment. Selection on Observables

Summary of Main Ideas

Kirill Borusyak

ARE 213 Applied Econometrics

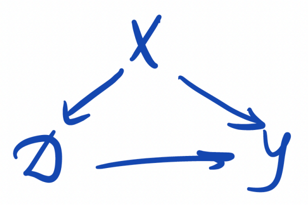
UC Berkeley, Fall 2025

Outline

What if treatment is not randomly assigned?

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

Unconfoundedness:



Violation of unconfoundedness:

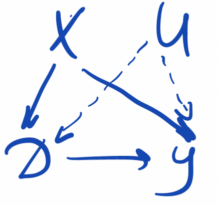


Illustration: Returns to College Degree

i	X_i (urban)	D_i (college)	Y (earnings)	Y_0	Y_1
1	1	**1**	**51**	?	51
2	1	**1**	**50**	?	50
3	1	**1**	**49**	?	49
4	1	0	40	40	?
5	0	**1**	**15**	?	15
6	0	0	11	11	?
7	0	0	9	9	?
Treated *minus control (not causal!)*			21.25		

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 *defining* that X_i is a control variable
- **Overlap**: $0 < P(D_i = 1 \mid X_i) < 1$ on the support of X_i
  - $P(D_i = 1 \mid X_i)$ is called the **propensity score**

Identification under CIA + overlap

- **Conditional average treatment effect**: $CATE(x) \equiv \mathbb{E}[Y_i(1) - Y_i(0) \mid X_i = x]$
  - Let $h_1(x) = \mathbb{E}[Y \mid D = 1, X = x]$ and $h_0(x) = \mathbb{E}[Y \mid D = 0, X = x]$
  - CATE can be identified as $CATE(x) = h_1(x) - h_0(x)$ *(prove this!)*
- $ATE = \mathbb{E}[CATE(X_i)] = \mathbb{E}[h_1(X_i) - h_0(X_i)]$, with expectation taken across X_i (in both treatment and control groups)
- And $ATT = \mathbb{E}[CATE(X_i) \mid D_i = 1]$, with expectation taken across X_i in the treated group

Imputation representation

- We also have
$$ATT = \mathbb{E}\left[Y_i - h_0(X_i) \mid D_i = 1\right]$$
  - For treated i, we impute $\widehat{Y_i(0)} = h_0(X_i)$and
- And
$$ATE = \mathbb{E}\left[(Y_i - h_0(X_i))\, D_i + (h_1(X) - Y_i)(1 - D_i)\right]$$
  - For untreated i, also impute $\widehat{Y_i(1)} = h_1(X_i)$

Matching / Imputation

i	X_i (urban)	D_i (college)	Y (earnings)	$\hat{Y}_0$	$\hat{Y}_1$	$\hat{Y}_1 - \hat{Y}_0$
1	1	1	**51**	40	51	11
2	1	1	**50**	40	50	10
3	1	1	**49**	40	49	9
4	1	0	40	40	$\frac{51+50+49}{3}$	10
			Urban averages & CATE	*40*	*50*	*10*
5	0	1	**15**	$\frac{11+9}{2}$	15	5
6	0	0	11	11	15	4
7	0	0	9	9	15	6
			Rural averages and CATE	*10*	*15*	*5*
			Overall averages and ATE:	*27.1*	*35*	*7.9*

Regression adjustment

Regress Y_i on X_i in the subsample $D_i = 1$; take fitted values as $\widehat{h}_1(X_i)$; repeat for $D_i = 0$

i	X_i (urban)	D_i (college)	Y (earnings)	$\hat{Y}_0$	$\hat{Y}_1$	$\hat{Y}_1 - \hat{Y}_0$
1	1	**1**	**51**	40	50	10
2	1	**1**	**50**	40	50	10
3	1	**1**	**49**	40	50	10
4	1	0	40	40	50	10
5	0	**1**	**15**	10	15	5
6	0	0	11	10	15	5
7	0	0	9	10	15	5
			Overall averages and ATE:	27.1	*35*	*7.9*
			Treatment group averages and ATT:	32.5	*41.3*	*8.8*

Pscore reweighting (for ATE)

i	X_i (urban)	D_i (college)	Y (earnings)	$\hat{Y}_0$	$\hat{Y}_1$	$\hat{Y}_1 - \hat{Y}_0$
1	1	**1**	**51**	0	$51 \cdot \frac{1}{3/4}$	$51 \cdot \frac{1}{3/4}$
2	1	**1**	**50**	0	$50 \cdot \frac{1}{3/4}$	$50 \cdot \frac{1}{3/4}$
3	1	**1**	**49**	0	$49 \cdot \frac{1}{3/4}$	$49 \cdot \frac{1}{3/4}$
4	1	0	40	$40 \cdot \frac{1}{1/4}$	0	$-40 \cdot \frac{1}{1/4}$
5	0	**1**	**15**	0	$15 \cdot \frac{1}{1/3}$	$15 \cdot \frac{1}{1/3}$
6	0	0	11	$11 \cdot \frac{1}{2/3}$	0	$-11 \cdot \frac{1}{2/3}$
7	0	0	9	$9 \cdot \frac{1}{2/3}$	0	$-9 \cdot \frac{1}{2/3}$
			Overall averages and ATE:	*27.1*	*35*	*7.9*

Continuous covariates

- With continuous or high-dimensional covariates, can't do any of this
  - Approximate matching *(skipped)*
  - Regression adjustment
  - Propensity score methods
  - Doubly-robust methods; double machine learning
  - *Note:* Different methods are no longer equivalent
- Also, do simpler strategies, like regressing Y_i on D_i and X_i, recover anything useful?

Outline

Regression adjustment

- Recall that, 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 $h_1(\cdot)$, we get

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

- or via imputation:

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

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

How to estimate $h_0(\cdot), h_1(\cdot)$?

- $h_d(\cdot)$ is the CEF of Y_i given X_i for $D_i = d$
- 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 $\mathbb{E}\left[Y(d) \mid X\right] = \gamma_d' X$ 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)

$$Y_i = \gamma_0 + \gamma_1' X_i + \beta D_i + \tau' X_i D_i + \text{error}_i, \qquad \widehat{ATE} = \hat{\beta} + \hat{\tau}' \bar{X}$$

  - Or its convenient reformulation

$$Y_i = \gamma_0 + \gamma_1' X_i + \beta D_i + \tau' \left(X_i - \bar{X}\right) D_i + \text{error}_i, \qquad \widehat{ATE} = \hat{\beta}$$

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

Uninteracted regression

What about regression of Y_i on D_i and X_i without interactions?

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

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

  where $h_0(X) = \gamma' X$ and $h_1(X) = \beta + \gamma' X$

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

- But Mostly Harmless Econometrics advocates for uninteracted regressions even when the effects are heterogeneous. Why?

Uninteracted regression with heterogeneous effects

- Assume the propensity score $p(X_i) \equiv \mathbb{E}[D_i \mid X_i] = P(D_i = 1 \mid X_i)$ is linear in X_i
  - Angrist (1998) focused on saturated controls $X_i \implies$ 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 pscore, partialling out X_i from D_i yields residuals $\tilde{D}_i = D_i - \mathbb{E}\left[D_i \mid X_i\right]$
- 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, $\mathbb{E}\left[\tilde{D}_i D_i\right] = \mathbb{E}\left[\mathrm{Var}\left[D_i \mid X_i\right]\right]$

Outline

Propensity score theorems (Rosenbaum and Rubin, 1983)

- Consider binary D. Recall $p(X) \equiv P(D = 1 \mid X)$
- **Proposition 1**: $D \perp\!\!\!\perp X \mid p(X)$
  - i.e. propensity score balances X between treated and control groups
  - Proof: $P(D = 1 \mid X, p(X)) = P(D = 1 \mid X) = p(X)$
- **Proposition 2:** $D \perp\!\!\!\perp (Y_0, Y_1) \mid X \implies D \perp\!\!\!\perp (Y_0, Y_1) \mid p(X)$
  - i.e., under CIA, controlling for scalar $p(X)$ is enough
    - A stronger version of the OVB idea
  - Proof: $P\left(D = 1 \mid p(X), Y_0, Y_1\right) = \mathbb{E}\left[\mathbb{E}\left[D \mid p(X), X, Y_0, Y_1\right] \mid p(X), Y_0, Y_1\right] = p(X)$, doesn't depend on (Y_0, Y_1)

P-score methods: Steps

1. Obtain/estimate $p(X)$
   - Known in complex RCTs
   - Parametric estimation, e.g. logit of D on X
   - Non-parametric regression of D on X
2. Verify balance
   - Within bins of $\hat{p}(X)$ compare X among treated and controls
   - If balance fails (with sufficiently many bins), make the p-score model richer
3. Assess overlap
   - Compare p-score distributions in treated & control groups
4. **Adjust for pscore differences** between treated and control groups
   - Regression, matching, blocking, reweighting

P-score adjustment methods: Regression

- With constant effects, enough to control linearly

$$Y_i = \beta D_i + \gamma p(X_i) + \text{error}$$

  - *Exercise*: Why?
  - *Exercise*: if $p(X)$ is estimated from a linear probability model, both ways are numerically the same as linearly controlling for X

- With heterogeneous effects, this yields the variance-weighted average of effects (*Exercise*: Why?)

P-score adjustment methods: Blocking/Matching

- **Blocking** (stratifying) = exact matching on binned pscore
  - Split data into bins of $p(X_i)$
  - Estimate difference-in-means within bins
  - Average across bins weighting by # obs. (ATE) or # treated obs. (ATT)
- Approximate **matching:** For each treated obs., compare the outcome with the untreated one with the closest $p(X_i)$; and the other way round
  - Discard observations with the pscore outside the range for the other group, such that the nearest match is very far

P-score adjustment methods: Reweighting (IPW)

- In the bin with $p(X_i) = \pi$ we have fraction π of observed $Y_i(1)$ (for treated) and fraction $1 - \pi$ of comparable but missing $Y_i(1)$ (for controls)
- Horvitz-Thompson (1952) "**inverse probability weighting**" (IPW): reweighting by $1/\pi$ makes the sample of $Y_i(1)$ representative

$$\mathbb{E}\left[\frac{YD}{p(X)}\right] = \mathbb{E}\left[\frac{Y_1 D}{p(X)}\right] = \mathbb{E}\left[\mathbb{E}\left[\frac{Y_1 D}{p(X)} \mid X\right]\right] = \mathbb{E}\left[Y_1\right]$$

- Similarly for Y_0: $\mathbb{E}\left[\frac{Y(1-D)}{1-p(X)}\right] = \mathbb{E}\left[Y_0\right]$. Thus, under CIA+overlap:

$$ATE = \mathbb{E}\left[\left(\frac{D}{p(X)} - \frac{1-D}{1-p(X)}\right) Y\right] = \mathbb{E}\left[\frac{D - p(X)}{p(X)(1-p(X))} \cdot Y\right]$$

- *Exercise:* derive the reweighting expression for ATT

P-score adjustment methods: Reweighting (2)

- Plug-in (Horvitz-Thompson) estimator: $\widehat{ATE}_{HT} = \frac{1}{N}\sum_i \left(\frac{D_i}{\hat{p}(X_i)} - \frac{1-D_i}{1-\hat{p}(X_i)} \right) Y_i$

- Issue: weights on treated $\left(\frac{1}{N}\frac{1}{\hat{p}(X_i)}\right)$ and control $\left(\frac{1}{N}\frac{1}{1-\hat{p}(X_i)}\right)$ do not exactly sum to 1
    - So adding a constant to Y_i changes the estimator

- Normalizing the weights improves performance: the Hajek estimator

$$\widehat{ATE}_{Hajek} = \frac{\sum_i \frac{D_i Y_i}{\hat{p}(X_i)}}{\sum_i \frac{D_i}{\hat{p}(X_i)}} - \frac{\sum_i \frac{(1-D_i)Y_i}{1-\hat{p}(X_i)}}{\sum_i \frac{1-D_i}{1-\hat{p}(X_i)}}$$

Outline

Idea of double robustness

- Regression adjustment methods require estimates of $h_0(X), h_1(X)$
  - No need for $p(X) = \mathbb{E}[D \mid X]$
- Propensity score methods are the opposite
- Doubly robust methods take estimates of both $h_d(X)$ and $p(X)$ as inputs
  - But validity requires only one of those estimates to be correct

Automatic double robustness

Some methods already possess some double robustness

- Under constant effects, regression $Y_i = \beta D_i + \gamma' X_i + \text{error}_i$ is causal if:
    - $h_0(X)$ is linear in X **or** $p(X)$ is linear in X
- Kline (2011): the Oaxaca-Blinder estimator for ATT is consistent if:
    - $h_0(X), h_1(X)$ are linear in X **or** $\frac{p(X)}{1-p(X)}$ is linear in X

Augmented IPW (AIPW)

- But double-robustness can be achieved for arbitrary estimators of $h_d(\cdot)$ and $p(\cdot)$
- Augmented IPW **(AIPW)** idea:

$$
\begin{aligned}
\mathbb{E}\left[Y_{1i}\right] = \mathbb{E}\left[h_1(X_i)\right] = \mathbb{E}\left[\frac{D_i}{p(X_i)}Y_i\right] &= \mathbb{E}\left[h_1(X_i) + \frac{D_i}{p(X_i)}\left(Y_i - h_1(X_i)\right)\right] \\
&= \mathbb{E}\left[\tilde{h}_1(X_i) + \frac{D_i}{\tilde{p}(X_i)}\left(Y_i - \tilde{h}_1(X_i)\right)\right] \quad \text{if } p(\cdot) = \tilde{p}(\cdot) \textbf{ or } h_1(\cdot) = \tilde{h}_1(\cdot)
\end{aligned}
$$

  - If the model of $h_1(\cdot)$ is correct, IPW adjustment doesn't change the estimand
  - If the model of $p(\cdot)$ is correct, the adjustment fixes mistakes in $h_1(\cdot)$

Augmented IPW (2)

- Combining with a similar expression for Y_{0i},

$$ATE = \mathbb{E}\left[h_1(X_i) + \frac{D_i}{p(X_i)}(Y_i - h_1(X_i)) - h_0(X_i) - \frac{1-D_i}{1-p(X_i)}(Y_i - h_0(X_i))\right]$$

  if $(h_0(\cdot), h_1(\cdot))$ **or** $p(\cdot)$ are correctly specified

  - The sample analog based on preliminary estimates of h_0, h_1, p yields the estimator

Double robustness and Neyman orthogonality

- The usefulness of double robustness is not obvious
- But a closely related concept is very helpful when X_i are high-dimensional:
  - Estimation of **nuisance functions** $h_0(\cdot), h_1(\cdot), p(\cdot)$, e.g. via machine learning methods, is inevitably imprecise
  - So we want to use moments that have low sensitivity to such mistakes around the true values — **Neyman-orthogonal** moments
    - ⋆ *Optional exercise*: verify that the "derivative" of the IAPW moment w.r.t. nuisance functions is zero at the true values
  - Covariate adjustment by Double Machine Learning (Chernozhukov et al., 2018) uses this idea

References I

Angrist, J. (1998): "Estimating the labor market impact of voluntary military service using social security data on military applicants," *Econometrica*, 66, 249–288.

Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins (2018): "Double/debiased machine learning for treatment and structural parameters," *Econometrics Journal*, 21, C1–C68.

Horvitz, D. G. and D. J. Thompson (1952): "A generalization of sampling without replacement from a finite universe," *Journal of the American statistical Association*, 47, 663–685.

Kline, P. (2011): "Oaxaca-Blinder as a reweighting estimator," *American Economic Review*, 101, 532–537.

Rosenbaum, P. R. and D. B. Rubin (1983): "The Central Role of the Propensity Score in Observational Studies for Causal Effects," *Biometrika*, 70, 41–55.