

# Part B: Selection on Observables

## B3: Doubly-Robust Methods

Kirill Borusyak

ARE 213 Applied Econometrics

UC Berkeley, Fall 2023

# Outline

- 1 Doubly-robust methods
- 2 ML methods for high-dimensional covariates
- 3 Violations of CIA. Coefficient stability

# Idea of double robustness

- Regression adjustment methods model and estimate  $h_d(X) = \mathbb{E}[Y_d | X]$ ,  $d = 0, 1$ 
  - ▶ No assumptions on  $p(X) = \mathbb{E}[D | X]$
- Propensity score methods are reverse
- Doubly robust methods require a model of  $h_d(X)$  and a model of  $p(X)$ 
  - ▶ But validity requires only one of those models 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  $X$

## Mixed methods

But doubly-robust estimators can be produced for other models of  $h_d(\cdot)$  and  $p(\cdot)$ ...

- Blocking on p-score + regression: in each block  $b$  regress

$$Y_i = \beta_b D_i + \gamma'_b X_i + \tau'_b D_i (X_i - \bar{X}) + \text{error}_i,$$

then average  $\hat{\beta}_b$

- IPW + regression (Hirano-Imbens 2003): regress

$$Y_i = \beta D_i + \gamma'_0 X_i + \tau' D_i (X_i - \bar{X}) + \text{error}_i$$

with weights  $\frac{D_i}{p(X_i)} + \frac{1-D_i}{1-p(X_i)}$ ; take  $\hat{\beta}$

- Both approaches are valid if  $h_0(X)$ ,  $h_1(X)$  are linear in  $X$  **or** the model of p-scores is correct

## Augmented IPW

Another popular combination of reweighting + regression is **AIPW**. Based on:

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

if  $h_0(X)$ ,  $h_1(X)$  **or**  $p(X)$  is correctly specified

- If models of  $h_0(\cdot)$ ,  $h_1(\cdot)$  are correct, IPW adjustment doesn't change the estimand
- If the model of  $p(\cdot)$  is correct, the adjustment fixes mistakes in  $h_0(\cdot)$ ,  $h_1(\cdot)$
- Thus, the sample analog is doubly-robust
- Efficiency:
  - ▶ When both  $\hat{h}_0(\cdot)$ ,  $\hat{h}_1(\cdot)$  and  $\hat{p}(\cdot)$  are correct, AIPW achieves the semi-parametric efficiency bound
  - ▶ But if  $\hat{p}(\cdot)$  is incorrect, the IPW part increases variance

# NSW application: Mixed methods

Table 3. Estimated Training Effects for the NSW Male Participants Using Comparison Groups From PSID and CPS

	NSW earnings less comparison group earnings		NSW treatment earnings less comparison group earnings, conditional on the estimated propensity score					
	(1) Unadjusted	(2) Adjusted <sup>a</sup>	Quadratic in score <sup>b</sup> (3)	Stratifying on the score			Matching on the score	
				(4) Unadjusted	(5) Adjusted	(6) Observations <sup>c</sup>	(7) Unadjusted	(8) Adjusted <sup>d</sup>
NSW	1,794 (633)	1,672 (638)						
PSID-1 <sup>e</sup>	−15,205 (1,154)	731 (886)	294 (1,389)	1,608 (1,571)	1,494 (1,581)	1,255	1,691 (2,209)	1,473 (809)
PSID-2 <sup>f</sup>	−3,647 (959)	683 (1,028)	496 (1,193)	2,220 (1,768)	2,235 (1,793)	389	1,455 (2,303)	1,480 (808)
PSID-3 <sup>f</sup>	1,069 (899)	825 (1,104)	647 (1,383)	2,321 (1,994)	1,870 (2,002)	247	2,120 (2,335)	1,549 (826)
CPS-1 <sup>g</sup>	−8,498 (712)	972 (550)	1,117 (747)	1,713 (1,115)	1,774 (1,152)	4,117	1,582 (1,069)	1,616 (751)
CPS-2 <sup>g</sup>	−3,822 (670)	790 (658)	505 (847)	1,543 (1,461)	1,622 (1,346)	1,493	1,788 (1,205)	1,563 (753)
CPS-3 <sup>g</sup>	−635 (657)	1,326 (798)	556 (951)	1,252 (1,617)	2,219 (2,082)	514	587 (1,496)	662 (776)

(From Dehejia-Wahba 1999)

# Outline

- 1 Doubly-robust methods
- 2 ML methods for high-dimensional covariates
- 3 Violations of CIA. Coefficient stability



# High-dimensional regression adjustment: Setting

- CIA is weaker with more pre-determined controls
- Assume constant effects for now. Then we can write

$$Y_i = \beta D_i + h(X_i) + \varepsilon_i, \quad \mathbb{E}[\varepsilon_i \mid D_i, X_i] = 0.$$

- With enough nonlinear terms and interactions,

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

- But with many controls and/or high-order terms,  $\dim(\gamma)$  is large, possibly  $> N$   
 $\implies$  cannot run this regression
- Manual model selection is dangerous!
- We will use penalized regression: LASSO

# Least Absolute Shrinkage and Selection Operator (LASSO)

- In general, suppose we would like to estimate

$$y_i = \beta' x_i + u_i, \quad \mathbb{E}[u_i | x_i] = 0$$

- Assume  $\dim(\beta)$  is large but a small (unknown) set of covariates approximates  $\mathbb{E}[y | x]$  well — **approximate sparsity**
- Then this set can be approximately recovered by LASSO (Tibshirani 1996):

$$\hat{\beta} = \arg \min_b \sum_{i=1}^N (y_i - b' x_i)^2 + \lambda \sum_{j=1}^{\dim(\beta)} |b_j| \cdot s_j$$

where  $\lambda$  is penalty and  $s_j$  are penalty loadings (e.g. SD of  $x_j$ )

- Because  $|\cdot|$  has a kink at zero, many  $\hat{\beta}_j$  will be exactly zero  $\implies$  covariate selection
- A good prediction of  $\mathbb{E}[y_i | x_i]$  is guaranteed

# Naïve LASSO covariate adjustments fail

1. Estimate  $Y_i = \beta D_i + \gamma' X_i + \varepsilon_i$  by LASSO without penalty on  $D_i$  to make sure it's not dropped
  - ▶ Covariates that are highly correlated with  $D_i$  but don't help predicting  $Y_i$  will be dropped
2. Run LASSO of  $Y_i$  on  $X_i$ ; keep the selected subset of covariates  $\tilde{X}_i$ ; regress  $Y_i$  on  $D_i$  and  $\tilde{X}_i$ 
  - ▶ Covariates with a moderate effect on  $Y$  may get dropped, even if they have a large effect on  $D$
  - ▶ Yet they generate moderate OVB which is typically larger than  $O(1/\sqrt{N})$
  - ▶ We'd be okay if we only dropped variables which don't strongly predict  $Y$  **or**  $D$

# Post-double-selection LASSO

Belloni, Chernozhukov, Hansen (2013) propose:

1. LASSO  $Y$  on  $X \implies$  subset  $\widetilde{X}_A$  gets selected
2. LASSO  $D$  on  $X \implies$  subset  $\widetilde{X}_B$  gets selected
3. Regress  $Y$  on  $D$  and  $\widetilde{X}_A \cup \widetilde{X}_B$

Theorem: if both  $\mathbb{E}[Y | X]$  and  $\mathbb{E}[D | X]$  are approximately sparse,

- this **post-double-selection LASSO**  $\hat{\beta}$  is  $\sqrt{N}$ -consistent and asymptotically normal;
- conventional SE are valid, i.e. model selection does not affect the asy.variance

# Double/Debiased Machine Learning (DML)

Chernozhukov et al. (Econometrics Journal 2018) generalize and improve this strategy:

## 1. Estimators based on **Neyman-orthogonal** ( $\approx$ doubly-robust) moments

- ▶ Moments are insensitive to mistakes in estimating nuisance parameters/functions
- ▶ With constant effects:  $\mathbb{E} [\{ (Y_i - \mathbb{E} [Y_i | X_i]) - \beta (D_i - p(X_i)) \} \cdot (D_i - p(X_i))] = 0$
- ▶ With heterogeneous effects:

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

## 2. Better to use **cross-fitting** (or “sample splitting”): estimate nuisance parameters for each observation on the data excluding this observation

# DML algorithm for homogeneous effects

1. Split the data into  $K$  random folds  $I_1, \dots, I_K$
2. For each  $k$ , use data *excluding* fold  $k$  to obtain predictors  $\hat{m}_{-k}(x)$  for  $\mathbb{E}[Y | X = x]$  and  $\hat{p}_{-k}(x)$  for  $\mathbb{E}[D | X = x]$  by your favorite prediction tool (LASSO, neural network, random forest, etc.)
3. Obtain fitted values for  $i \in I_k$ :  $\hat{m}_{-k}(X_i)$  and  $\hat{p}_{-k}(X_i)$
4. Get  $\hat{\beta}$  by regressing  $Y_i - \hat{m}_{-k(i)}(X_i)$  on  $D_i - \hat{p}_{-k(i)}(X_i)$  in the full sample
5. Use conventional standard errors: noise from estimation does not matter

# DML algorithm in general

- Choose a Neyman-orthogonal moment  $\mathbb{E}[\psi(\text{data}, \theta, \eta)] = 0$  where  $\theta$  is parameter of interest and  $\eta$  is a nuisance parameter, e.g.  $\theta = ATE$ ,  $\eta = (h_0(\cdot), h_1(\cdot), p(\cdot))$ ,

$$\psi = h_1(X_i) + \frac{D_i}{p(X_i)} (Y_i - h_1(X_i)) - h_0(X_i) - \frac{1 - D_i}{1 - p(X_i)} (Y_i - h_0(X_i)) - \theta$$

- Split the data into  $K$  random folds  $I_1, \dots, I_K$
- For each fold  $k$ , obtain  $\hat{\eta}_{-k}$  by some prediction tool on the data *excluding* fold  $k$
- Obtain fitted values  $\psi_i$  for observations *in* fold  $k$
- Get  $\hat{\theta}$  by setting  $\sum_i \psi_i = 0$  in the full sample (e.g. for ATE, use AIPW)

## Application w/ observational data (Chernozhukov et al. 2018)

- Effect of eligibility for 401(k) plan on net financial assets (from Poterba et al. 1994)
- CIA: argue that workers choose jobs based on income not 401(k) availability, especially when 401(k) plans first became available
- Covariates: age, income, family size, education, marriage status, two-earner status, pension status, IRA participation, home ownership
- Chernozhukov et al. apply DML with various ML algorithms: lasso, random forest, boosting, neural network, ensemble methods
- Obtain similar magnitudes; sadly no parallel linear regression specifications to compare SE



# ML methods for RCT data (Wager, Du, Taylor, Tibshirani 2016)

- In a RCT, enough to do regression adjustment
- No need for double robust methods (although can use them too)
- Can use any ML algorithm for  $h_d(X) = \mathbb{E}[Y \mid D = d, X]$  with cross-fitting

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

(note the denominators!)

# Application with a RCT (List, Muir, Sun 2022)

Oregon Health Insurance Experiment (from Finkelstein et al. 2016):

- Effects of lottery-assigned Medicaid eligibility on ER visits and Medicaid take-up
- Covariates: gender, age, prior health, education, prior ER visits
- $SM$  = diff-in-means;  $LRA$  = OLS + x-fitting;  $FRA$  = random forest + x-fitting

Table 7: Variance Reduction for OHIE

	SM	LRA	FRA
ER Visits	0.0132 (0.0085)	0.0143 (0.0079)	0.0139 (0.0077)
Medicaid Take-Up	0.172 (0.0063)	0.159 (0.0062)	0.150 (0.0062)

- Also: efficiency gain of FRA over LRA is bigger when controls are poorly scaled: levels instead of logs

# Outline

- 1 Doubly-robust methods
- 2 ML methods for high-dimensional covariates
- 3 Violations of CIA. Coefficient stability

# Learning from coefficient stability

- Suppose you are not sure if CIA holds but you find that the coefficient on  $D$  doesn't change too much with more covariates added
  - ▶ Can you conclude your estimates are close to causal?
  - ▶ Can you use these patterns to put bounds on the true causal effects?
- Altonji, Elder, Taber (2005) and Oster (2019): yes, but carefully
  - ▶ Requires extra assumptions on how “selection on unobservables” relates to “selection on observables”
  - ▶ Coef stability is only informative if the regression  $R^2$  increases significantly at the same time

# Setting

- Constant effects; impose model  $Y_i = \beta D_i + \gamma' X_i + W_i + u_i$  where:
  - ▶  $X_i$  are observed covariates
  - ▶  $W_i$  is the effect of unobserved covariates
  - ▶  $u_i$  is innocuous, e.g. measurement error ( $u_i = 0$  in Altonji et al.)
- Exogenous covariates:  $\text{Cov}[X_i, W_i] = 0$ 
  - ▶ Diegert, Masten, Poirier (2023) relax this assumption, at a cost
- Selection on both observables and unobservables:

$$D_i = \pi_1 \cdot \gamma' X_i + \pi_2 W_i + \text{error}_i$$

- ▶ **Equal selection:**  $\pi_2 = \pi_1$  (contrast with CIA:  $\pi_2 = 0$ )
  - ★ Justification: observe a random subset of many covariates
- ▶ Or  $\pi_2 = \delta \pi_1$  for unknown  $\delta \in [0, 1]$

## Altonji et al. (2005): Identification of bias

- For  $\beta_{OLS}$  from a regression of  $Y_i$  on  $D_i$  controlling for  $X_i$  and for  $\tilde{D}_i = D_i - \pi_1 \cdot \gamma' X_i$ ,

$$\beta_{OLS} - \beta = \frac{\text{Cov} [\tilde{D}_i, W_i]}{\text{Var} [\tilde{D}_i]} = \frac{\text{Cov} [D_i, W_i]}{\text{Var} [\tilde{D}_i]} = \frac{\pi_2 \text{Var} [W_i]}{\text{Var} [\tilde{D}_i]} = \delta \frac{\pi_1 \text{Var} [W_i]}{\text{Var} [\tilde{D}_i]}$$

using FWL, exogenous covariates, and equal selection

## Oster (2019): Link to coefficient stability

Oster expresses bias in terms of how much the coef moves:

- Approximation: if  $\delta \approx 1$  and if selection based on  $X$  is via  $\gamma'X$ ,

$$\beta \approx \beta_{OLS} + \delta (\beta_{OLS} - \beta_{short}) \frac{R_{max}^2 - R^2}{R^2 - R_{short}^2}$$

where  $\beta_{short}$  and  $R_{short}^2$  are from a regression of  $Y_i$  on  $D_i$  only,  $R^2$  is from the main regression on  $D_i, X_i$ , and  $R_{max}^2$  is from the infeasible regression on  $D_i, X_i, W_i$

- ▶ Standard regression output allows you to debias  $\beta$  given  $R_{max}^2$  and  $\delta$
- ▶  $R_{max}^2$  has to be chosen — no clear guidance
- ▶  $\delta$  is either chosen or you estimate the  $\max \delta$  such that your conclusion does not change, as measure of robustness to selection on unobservables
- ▶ *Exercise\*\**: prove Oster's result for  $\delta = 1$  without tedious math
- General case: Oster provides a complicated formula to be used in practice