Part B: Selection on Observables

B3: Doubly-Robust Methods

Kirill Borusyak

ARE 213 Applied Econometrics

UC Berkeley, Fall 2024

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 $h_d(X) = \mathbb{E}[Y_d \mid X]$, $d = 0, 1$
  - No assumptions on $p(X) = \mathbb{E}[D \mid X]$
- Propensity score methods are the opposite
- Doubly robust methods take a model of $h_d(X)$ and a model of $p(X)$ as inputs
  - But validity requires only one of those models to be correct
  - Gain some robustness. (How valuable? Any efficiency cost?)

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

Doubly-robust estimation with homogeneous effects

- But doubly-robust estimators can be produced for other models of $h_d(\cdot)$ and $p(\cdot)$
- Start with constant effects. We have moment condition:

$$\mathbb{E}\left[\left(Y - \beta D - \tilde{h}_0(X)\right)(D - \tilde{p}(X))\right] = 0 \qquad (*)$$

  which holds if $\tilde{h}_0(\cdot) = h_0(\cdot)$ **or** $\tilde{p}(\cdot) = p(\cdot)$

  - Get a preliminary estimate of $p(X)$, e.g. by logit of D on X
  - Get a preliminary estimate of $h_0(X)$, e.g. by linear regression of Y on D and X
  - Get $\hat{\beta}$ that sets the sample analog of (*) to zero

Heterogeneous effects: Mixed methods

With heterogeneous effects, we can mix pscore and regression adjustments:

- 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 $\hat{\beta}_b$ with appropriate weights

- 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 (AIPW)

- An important combination of reweighting + regression is **AIPW**. Idea:

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

  - 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

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

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

  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

- Efficiency:

  - When models of **both** $(h_0(\cdot), h_1(\cdot))$ and $p(\cdot)$ are correct, AIPW achieves the semi-parametric efficiency bound
  - But if $p(\cdot)$ is incorrect, the IPW correction 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*					
				Stratifying on the score			*Matching on the score*	
	(1) Unadjusted	*(2) Adjusted*[a]	*Quadratic in score*[b] *(3)*	*(4) Unadjusted*	*(5) Adjusted*	*(6) Observations*[c]	*(7) Unadjusted*	*(8) Adjusted*[d]
NSW	1,794	1,672						
	(633)	(638)						
PSID-1[e]	−15,205	731	294	1,608	1,494	1,255	1,691	1,473
	(1,154)	(886)	(1,389)	(1,571)	(1,581)		(2,209)	(809)
PSID-2[f]	−3,647	683	496	2,220	2,235	389	1,455	1,480
	(959)	(1,028)	(1,193)	(1,768)	(1,793)		(2,303)	(808)
PSID-3[f]	1,069	825	647	2,321	1,870	247	2,120	1,549
	(899)	(1,104)	(1,383)	(1,994)	(2,002)		(2,335)	(826)
CPS-1[g]	−8,498	972	1,117	1,713	1,774	4,117	1,582	1,616
	(712)	(550)	(747)	(1,115)	(1,152)		(1,069)	(751)
CPS-2[g]	−3,822	790	505	1,543	1,622	1,493	1,788	1,563
	(670)	(658)	(847)	(1,461)	(1,346)		(1,205)	(753)
CPS-3[g]	−635	1,326	556	1,252	2,219	514	587	662
	(657)	(798)	(951)	(1,617)	(2,082)		(1,496)	(776)

(From Dehejia-Wahba 1999)

Performance of estimators

There doesn't seem to be a clear winner across all estimators (regression, pscore, doubly-robust)

- See Monte Carlo simulations in e.g. Huber, Lechner, Wunsch (2013), Busso, DiNardo, McCrary (2014)

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

$$Y_i = \beta D_i + h_0(X_i) + \varepsilon_i, \qquad \mathbb{E}\left[\varepsilon_i \mid D_i, X_i\right] = 0.$$

- With enough nonlinear terms and interactions,

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

- But with many controls and/or high-order terms, $\dim(\gamma)$ is large, possibly $> N$ $\Longrightarrow$ 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

$$y_i = \beta' x_i + u_i, \qquad \mathbb{E}\left[u_i \mid x_i\right] = 0$$

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

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

  where λ 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 $\Longrightarrow$ covariate selection
- A good prediction of $\mathbb{E}\left[y_i \mid x_i\right]$ is guaranteed

More on LASSO

- How to choose the penalty?
  - Optimal penalty depends on the unknown $\mathrm{Var}[u]$
  - Cross-validation often used in practice: choose λ that minimizes prediction error on a validation set
  - *Note:* Root LASSO of Belloni, Chernozhukov, Wang (2011) bypasses the problem

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

- *Note:* LASSO is **not** invariant to transforming variables
  - E.g. including all dummies of a categorial variable does not have redundancy

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 highly correlated with D_i will be dropped, even if they have a non-zero γ-coefficient
   - LASSO regularizes $\gamma \implies$ "regularization bias" of β
2. Run LASSO of D_i on X_i; keep the selected subset of covariates $\widetilde{X}_i$; regress Y_i on D_i and $\widetilde{X}_i$
   - Covariates with a moderate effect on D may get dropped, even if they have a large effect on Y
   - 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 D **or** Y

Post-double-selection LASSO

Belloni, Chernozhukov, Hansen (2013) propose:

1. LASSO Y on $X \Longrightarrow$ subset $\widetilde{X_A}$ gets selected
2. LASSO D on $X \Longrightarrow$ subset $\widetilde{X_B}$ gets selected
3. Regress Y on D and $\widetilde{X_A} \cup \widetilde{X_B}$

Theorem: if both $\mathbb{E}[Y \mid X]$ and $\mathbb{E}[D \mid X]$ 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

More formally: Without double selection

- Recall $Y_i = \beta D_i + h_0(X_i) + \varepsilon_i$; let $D_i = p(X_i) + u_i$; $\mathrm{Cov}[\varepsilon_i, u_i] = 0$ — *why?*
- What if we naively regress Y_i on $D_i - \hat{p}(X_i)$? Letting $\hat{g}(X_i) = \beta\hat{p}(X_i) + h_0(X_i)$,

$$\hat{\beta} = \beta + \frac{\frac{1}{N}\sum_i \left(\hat{g}(X_i) + \varepsilon_i\right)\left(p(X_i) - \hat{p}(X_i) + u_i\right)}{\frac{1}{N}\sum_i \left(D - \hat{p}(X_i)\right)^2}$$

- Because of model selection errors, we have

$$\textit{Bias} \propto \frac{1}{N}\sum_i \hat{g}(X_i)\left(p(X_i) - \hat{p}(X_i)\right)$$

which is often $> O_p(1/\sqrt{N})$

More formally: With double selection

- Let $m(X_i) = \mathbb{E}[Y_i \mid X_i]$. Regressing $Y_i - \hat{m}(X_i)$ on $D_i - \hat{p}(X_i)$ yields

$$\hat{\beta} = \beta + \frac{\frac{1}{N}\sum_i (m(X_i) - \hat{m}(X_i) + \varepsilon_i)(p(X_i) - \hat{p}(X_i) + u_i)}{\frac{1}{N}\sum_i (D - \hat{p}(X_i))^2}$$

- Now

$$Bias \propto \frac{1}{N}\sum_i (m(X_i) - \hat{m}(X_i))(p(X_i) - \hat{p}(X_i)) = o_p(1/\sqrt{N})$$

  e.g. if prediction errors are $o_p(N^{-1/4})$, which is much easier

  - With double selection, prediction errors multiply!

Double/Debiased Machine Learning (DML)

Chernozhukov et al. (Econometrics J 2018) generalize beyond LASSO and improve performance:

1. Use estimators based on **Neyman-orthogonal** moments
   - Moments insensitive to mistakes in estimating nuisance parameters/functions around the true values
   - With constant effects: $\mathbb{E}\left[\{(Y_i - \mathbb{E}[Y_i \mid X_i]) - \beta(D_i - p(X_i))\} \cdot (D_i - p(X_i))\right] = 0$
   - 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 $I_1, \ldots, I_K$

2. For each k, use data *excluding* fold k to obtain predictors $\hat{m}_{-k}(x)$ for $\mathbb{E}[Y \mid X = x]$ and $\hat{p}_{-k}(x)$ for $\mathbb{E}[D \mid X = x]$ 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 and for ATE

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

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

- Split the data into K random folds $I_1, \ldots, I_K$
- For each fold k, obtain $\hat{\eta}_{-k}$ by some prediction tool on the data *excluding* fold k
- Obtain fitted values ψ_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) pension 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, no need for doubly-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	0.0143	0.0139
	(0.0085)	(0.0079)	(0.0077)
Medicaid Take-Up	0.172	0.159	0.150
	(0.0063)	(0.0062)	(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: $\mathrm{Cov}\left[X_i, W_i\right] = 0$
  - 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 β_{OLS} from a regression of Y_i on D_i controlling for X_i and for $\widetilde{D}_i = D_i - \pi_1 \cdot \gamma' X_i$,

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

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 $\gamma' X$,

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

  where β_{short} and R^2_{short} 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^2_{max} is from the infeasible regression on D_i, X_i, W_i
  - Standard regression output allows you to debias β given R^2_{max} and δ
  - R^2_{max} has to be chosen — no clear guidance
  - δ is either chosen or you estimate the $\max \delta$ such that your conclusion does not change, as measure of robustness to selection on unobservables
  - Note: δ is not how *correlated* W vs. $\gamma' X$ should be with the treatment
- General case: Oster provides a complicated formula to be used in practice

Alternative: Reusing the OVB formula

- Cinelli and Hazlett (2020) rewrite the OVB formula as

$$|\text{bias}| = \sqrt{R^2_{Y\sim W|D,X} \cdot \frac{R^2_{D\sim W|X}}{1 - R^2_{D\sim W|X}}} \cdot \frac{SD(Y^{\perp X,D})}{SD(D^{\perp X})}$$

  where $R^2_{A\sim B|C}$ is the R^2 from regressing A on B after partialling out C from both, and $SD(A^{\perp B})$ is SD of residuals from regressing A on B

- Define the Robustness Value statistic RV_q: if $R^2_{Y\sim W|D,X} = R^2_{D\sim W|X}$, which value should it take to reduce the coefficient by $100 \cdot q\%$

- Instead of arguing about $R^2_{Y\sim W|D,X}, R^2_{D\sim W|X}$, can bound their relative values for W vs. X_j