pdfbase

Causal Inference for Average Treatment Effects

Professor Susan Athey
Stanford University
Machine Learning and Causal Inference

Spring 2017

The potential outcomes framework

For a set of i.i.d. subjects $i = 1, ..., n$, we observe a tuple (X_i, Y_i, W_i), comprised of

- A **feature vector** $X_i \in \mathbb{R}^p$,
- A **response** $Y_i \in \mathbb{R}$, and
- A **treatment assignment** $W_i \in \{0, 1\}$.

Following the **potential outcomes** framework (Holland, 1986, Imbens and Rubin, 2015, Rosenbaum and Rubin, 1983, Rubin, 1974), we posit the existence of quantities $Y_i^{(0)}$ and $Y_i^{(1)}$.

- These correspond to the response we **would have measured** given that the i-th subject received treatment ($W_i = 1$) or no treatment ($W_i = 0$).
- **NB:** We only get to see $Y_i = Y_i^{(W_i)}$

The potential outcomes framework

For a set of i.i.d. subjects $i = 1, ..., n$, we observe a tuple (X_i, Y_i, W_i), comprised of

- A **feature vector** $X_i \in \mathbb{R}^p$,
- A **response** $Y_i \in \mathbb{R}$, and
- A **treatment assignment** $W_i \in \{0, 1\}$.

- Define the **average treatment effect (ATE)**, the **average treatment effect on the treated (ATT)**

$$\tau = \tau^{\text{ATE}} = \mathbb{E}\left[Y^{(1)} - Y^{(0)}\right]; \tau^{\text{ATT}} = \mathbb{E}\left[Y^{(1)} - Y^{(0)} \mid W_i = 1\right];$$

- and, the **conditional average treatment effect (CATE)**

$$\tau(x) = \mathbb{E}\left[Y^{(1)} - Y^{(0)} \mid X = x\right].$$

The potential outcomes framework

The potential outcomes framework

If we make no further assumptions, it is not possible to estimate ATE, ATT, CATE, and related quantities.

- This is a failure of identification (infinite sample size), not a small sample issue. Unobserved confounders correlated with both the treatment and the outcome make it impossible to separate correlation from causality.
- One way out is to assume that we have measured enough features to achieve **unconfoundedness** (Rosenbaum and Rubin, 1983)

$$\left\{ Y_i^{(0)},\ Y_i^{(1)} \right\} \perp\!\!\!\perp W_i \ \mid\ X_i.$$

- When this assumption + OVERLAP ($e(x) \in (0, 1)$) holds, causal effects are identified and can be estimated.

Identification

$$\mathbb{E}_{Y_i(1)}\left[Y_i^{(1)}\right] = \mathbb{E}_{X_i}\left[\mathbb{E}_{Y_i^{(1)}|X_i}\left[Y_i^{(1)} \mid X_i\right]\right]$$

$$= \mathbb{E}_{X_i}\left[\mathbb{E}_{Y_i(1)|X_i}\left[\frac{Y_i^{(1)} \cdot W_i}{Pr(W_i = 1|X_i)} \mid X_i\right]\right]$$

$$= \mathbb{E}_{X_i}\left[\mathbb{E}_{Y_i|X_i}\left[\frac{Y_i \cdot W_i}{Pr(W_i = 1|X_i)} \mid X_i\right]\right]$$

$$= \mathbb{E}_{Y_i}\left[\frac{Y_i \cdot W_i}{Pr(W_i = 1|X_i)}\right]$$

- Argument is analogous for $\mathbb{E}\left[Y^0\right]$, which leads to ATE; and similar arguments allow you to identify CATE as well as the counterfactual effect of any policy assigning units to treatments on the basis of covariates.
- This result suggests a natural estimator: propensity score weighting using the sample analog of the last equation.

The role of overlap

Note that we need $e(x) \in (0, 1)$ to be able to calculate treatment effects for all x.

- Intuitively, how could you possibly infer $[Y(0)|X_i = x]$ if $e(x) = 1$?
- Note that for discrete x, the variance of ATE is infinite when $e(x) = 0$.
- "Moving the goalposts": Crump, Hotz, Imbens, Miller (2009) analyze trimming, which entails dropping observations where $e(x)$ is too extreme. Typical approaches entail dropping bottom and top 5% or 10%.
- Approaches that don't directly require propensity score weighting may seem to avoid the need for this, but important to understand role of extrapolation.

Propensity Score Plots: Assessing Overlap

The causal inference literature has developed a variety of conventions, broadly referred to as "supplementary analysis," for assessing credibility of empirical studies. One of the most prevalent conventions is to plot the propensity scores of treated and control groups to assess overlap.

- Idea: for each $q \in (0, 1)$, plot the fraction of observations in the treatment group with $e(x) = q$, and likewise for the control group.
- Even if there is overlap, when there are large imbalances, this is a sign that it may be difficult to get an accurate estimate of the treatment effect.

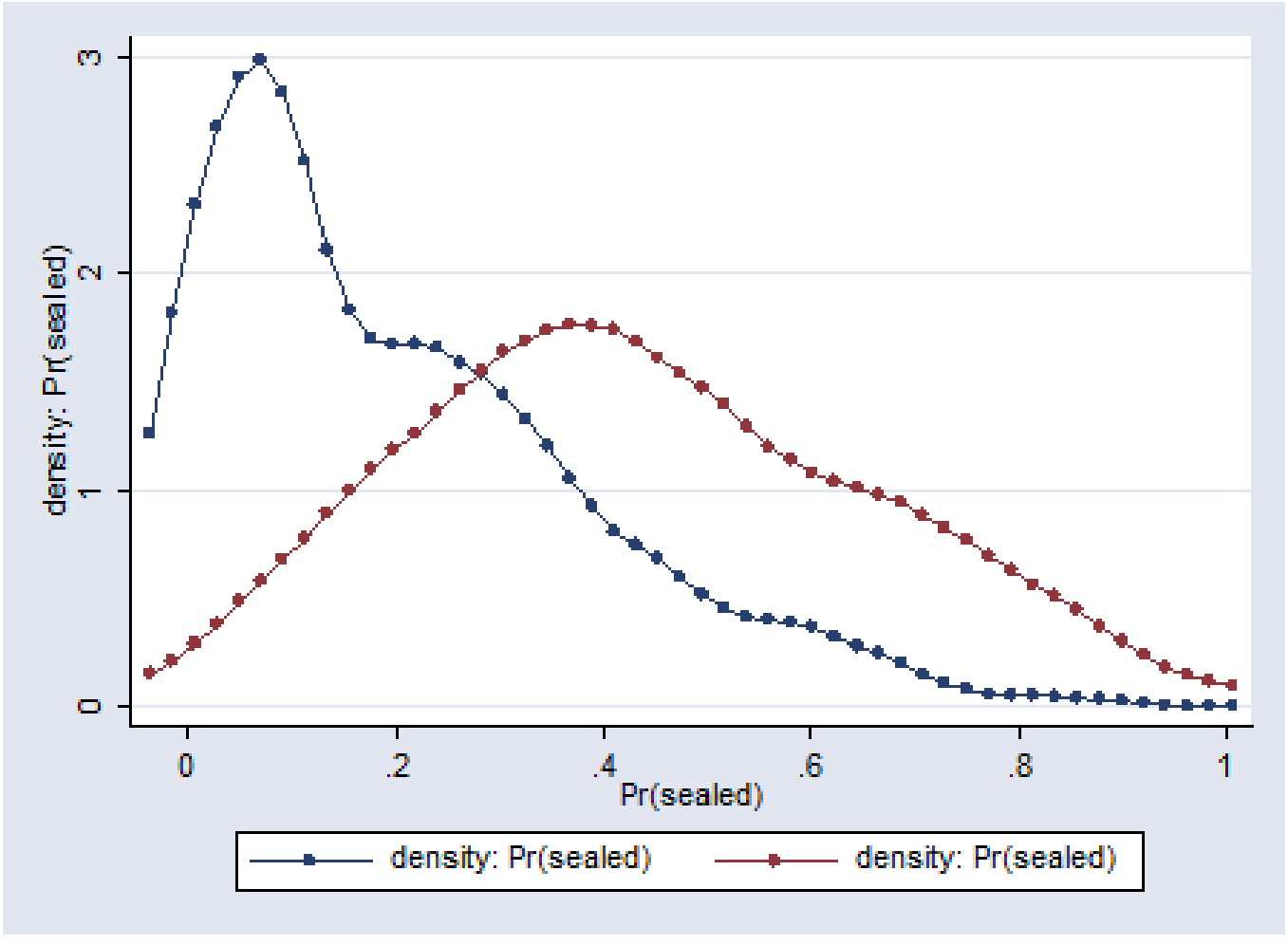
Propensity Score Plots: Assessing Overlap

Example: Athey, Levin and Seira analysis of timber.

- Assignment to first price or open ascending:
  - in ID, randomized for subset of tracts with different probabilities in different geographies;
  - in CA, small v. large sales (with cutoffs varying by geography).
- So $W = 1$ if auction is sealed, and X represents geography, size and year.

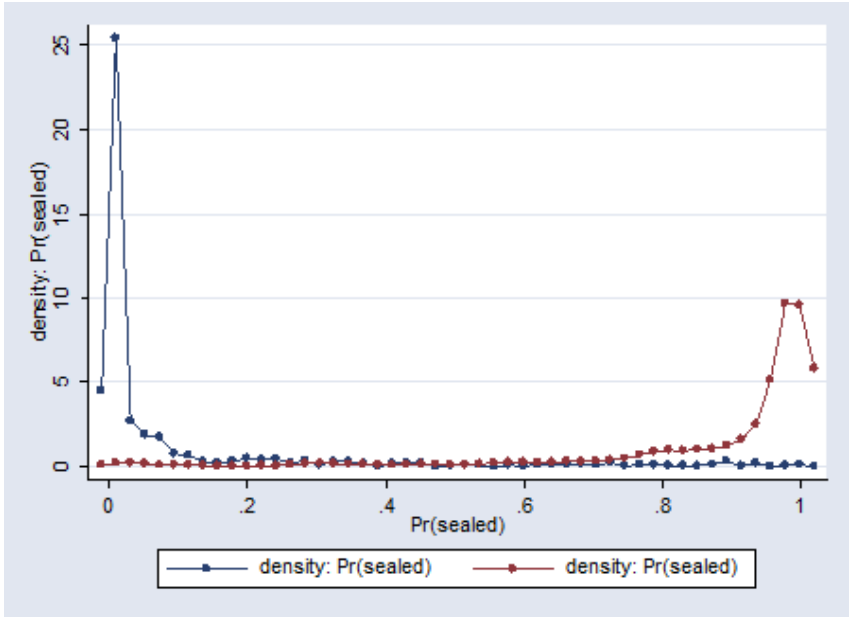
Propensity Score Plots: Assessing Overlap in ID

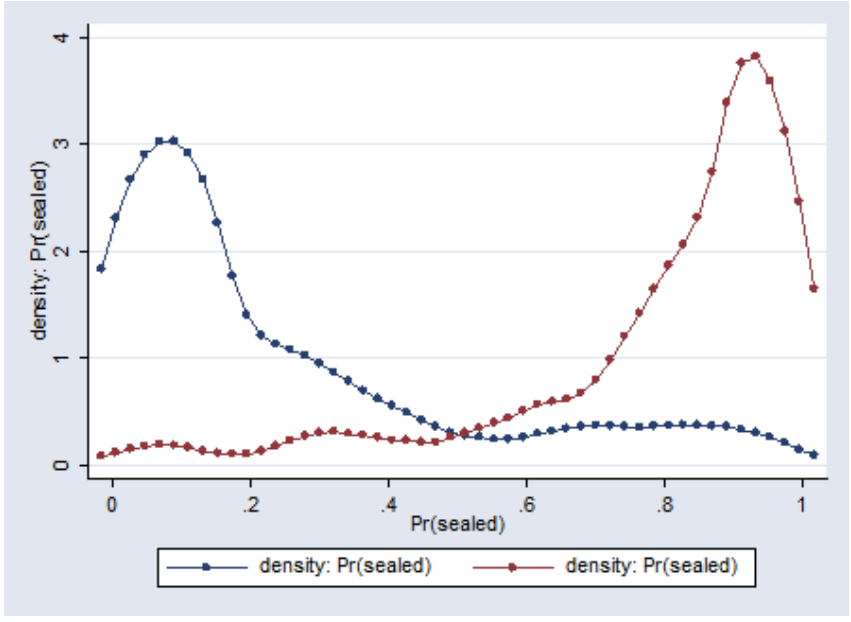
Very few observations with extreme propensity scores



Propensity Score Plots: Assessing Overlap in CA

Untrimmed v. trimmed so that $e(x) \in [.025, .975]$





Variance of Estimator: Discrete Case

- Suppose small number of realizations of X_i.
- Under unconfoundedness, can analyze these as separate experiments and average up the results.
- How does conditioning on X_i affect variance of estimator?

Variance of Estimator: Discrete Case

Let $\hat{\mathbb{E}}$ denote the sample average, $\mathbb{V}$ be the variance, $\pi(x)$ be the proportion of observations with $X_i = x$, and let $e(x)$ be the propensity score ($Pr(W_i = 1|X_i = x)$).

$$\mathbb{V}\left(\hat{\mathbb{E}}_{i:X_i=x,W_i=1}(Y_i)\right) = \frac{\sigma^2(x)}{n \cdot \pi(x) \cdot e(x)}$$

$$\mathbb{V}(\hat{\tau}(x)) = \frac{\sigma^2(x)}{n \cdot \pi(x) \cdot e(x)} + \frac{\sigma^2(x)}{n \cdot \pi(x) \cdot (1 - e(x))}.$$

$$\mathbb{V}(\hat{\text{ATE}}) = \sum_x \left[\frac{n(x)}{n} \frac{\sigma^2(x)}{n(x) \cdot e(x)} + \frac{\sigma^2(x)}{n(x) \cdot (1 - e(x))}\right].$$

$$= \sum_x \frac{\sigma^2(x)}{n} \left[\frac{1}{e(x)} + \frac{1}{(1 - e(x))}\right].$$

Estimation Methods

The following methods are efficient when the number of covariates is fixed:

- Propensity score weighting
- "Direct" model of the outcome (model of $\mathbb{E}\left[Y_i \,\middle|\, X_i, W_i\right]$), e.g. using regression
- Propensity-score weighted regression of Y on X, W (doubly robust)

The choice among these methods is widely studied:

- Other popular methods include matching, propensity score matching, propensity score blocking, which are not efficient but often do better in practice.
- Note: Hirano, Imbens, Ridder (2003) establish that more efficient to weight by estimated propensity score than actual.

Regression Case

Suppose that conditional mean function is given by

$$\mu(w, x) = \beta^{(w)} \cdot x.$$

If we estimate using OLS, then we can estimate the ATE as

$$\widehat{\text{ATE}} = \bar{X} \cdot (\hat{\beta}^{(1)} - \hat{\beta^{(0)}})$$

Note that OLS is unbiased and efficient, so the above quantity converges to the true values at rate $\sqrt{n}$:

$$\bar{X} \cdot (\hat{\beta}^{(1)} - \hat{\beta}^{(0)}) - \mu_x \cdot (\beta^{(1)} - \beta^{(0)}) = O_p\left(\frac{1}{\sqrt{n}}\right)$$

High-Dimensional Analogs??

Obvious possibility: substitute in the lasso (or ridge, or elastic net) for OLS. But bias is a big problem.
With lasso, for each component j:

$$\hat{\beta}_j^{(w)} - \beta_j^{(w)} = O_p\left(\sqrt{\frac{\log(p)}{n}}\right)$$

This adds up across all dimensions, so that we can only guarantee for the ATT:

$$\widehat{\text{ATT}} - \text{ATT} = O_p\left(\sqrt{\frac{\log(p)}{n}}\|\bar{X}_1 - \bar{X}_0\|_\infty \cdot \|\beta^{(0)}\|_0\right)$$

Imposing Sparsity: LASSO Crash Course

Assume linear model, and that there are at most a fixed number k of non-zero coefficients: $\|\beta\|_0 \leq k$.

Suppose X satisfies a "restricted eigenvalue" condition: no small group of variables is nearly collinear.

$$\|\hat{\beta} - \beta\|_2 = O_p\left(\sqrt{\frac{k \cdot \log(p)}{n}}\right)$$

$$\|\hat{\beta} - \beta\|_1 = O_p\left(k\sqrt{\frac{\log(p)}{n}}\right)$$

With the "de-biased lasso" (post-LASSO OLS) we can even build confidence intervals on $\hat{\beta}$ if $k << \frac{\sqrt{n}}{\log(p)}$.

Applying to the ATT, where we need to estimate $\bar{X}_1 \cdot \beta^{(0)}$ (the cf outcome for treated observations had they been control instead):

$$\hat{\text{ATT}} - \text{ATT} = O_p\left(k\|\bar{X}_1 - \bar{X}_0\|_\infty\sqrt{\frac{\log(p)}{n}}\right)$$

Improving the Properties of ATE Estimation in High Dimensions: A "Double-Selection" Method

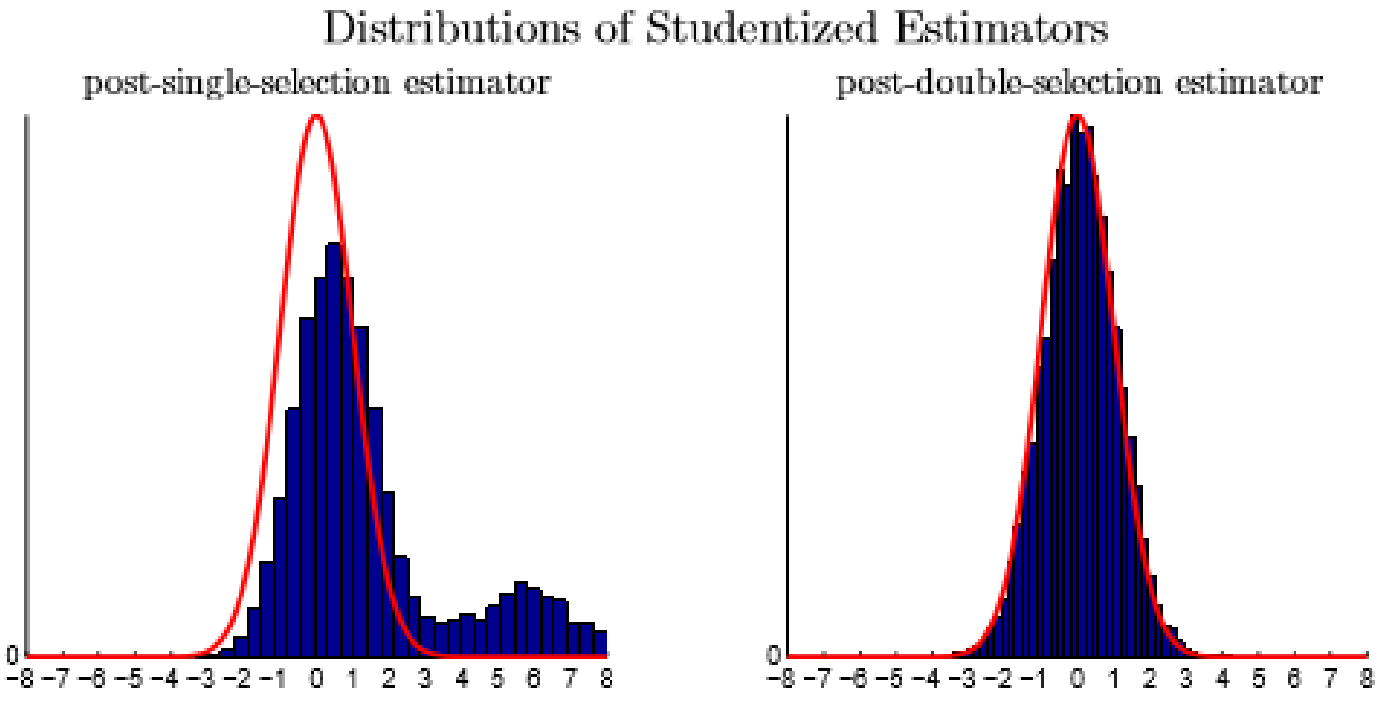
Belloni, Chernozukov, and Hansen (2013) observe that causal inference is not an off-the-shelf prediction problem: confounders might be important if they have a large effect on outcomes OR a large effect on treatment assignment. They propose:

- Run LASSO of W on X. Select variables with non-zero coefficients at a selected λ (e.g. cross-validation).
- Run a LASSO of Y on X. Select variables with non-zero coefficients at a selected λ (may be different than first λ).
- Run a OLS of Y on W and the union of selected variables. (Not as good at purely predicting Y as using only second set.)

Result: under "approximate sparsity" of BOTH propensity and outcome models, and constant treatment effects, estimated ATE is asymptotically normal and estimation is efficient.

Intuition: with enough data, can find the variables relevant for bias. With approximate sparsity and constant treatment effect, there aren't too many, and OLS will be unbiased.

Single v. Double Selection in BCH Algorithm



More General Results

Belloni, Chernozukov, Fernandez-Val and Hansen (2016) (http://arxiv.org/abs/1311.2645, forthcoming *Econometrica*) have a variety of generalizations:

- Applies general approach to IV
- Allows for a continuum of outcome variables
- Observes that nuisance parameters can be estimated generally using ML methods without affecting the convergence rate, subject to orthoganality conditions
- Shows how to use a framework based on orthogonality in moment conditions

Doubly Robust Methods

With small data, a "doubly robust" estimator (though not the typical one, where typically people use inverse propensity score weighted regression) is (with $\hat{\gamma}_i = \frac{1}{\hat{e}(X_i)}$):

$$\hat{\mu}_1^0 = \bar{X}_1 \cdot \hat{\beta}^{(0)} + \hat{\mathbb{E}}_{i:W_i=0} \hat{\gamma}_i \left(Y_i - X_i \hat{\beta}^{(0)} \right)$$

To see why, note that the term in parentheses goes to 0 if we estimate $\beta^{(0)}$ well, while to show that we get the right answer if we estimate the propensity score well, we rearrange the expression to be

$$\hat{\mu}_1^0 = \left(\bar{X}_1 - \hat{\mathbb{E}}_{i:W_i=0} (\hat{\gamma}_i X_i) \right) \hat{\beta}^{(0)} + \hat{\mathbb{E}}_{i:W_i=0} \hat{\gamma}_i Y_i$$

The first term has expectation 0, and the second term gives the relevant counterfactual, if the propensity score is well-estimated.

Doubly Robust Methods: A High-Dimensional Analog?

$$\hat{\mu}_1^0 = \bar{X}_1 \cdot \hat{\beta}^{(0)} + \hat{\mathbb{E}}_{i:W_i=0} \hat{\gamma}_i \left(Y_i - X_i \hat{\beta}^{(0)} \right)$$

How does this relate to the truth?

$$\hat{\mu}_1^0 - \mu_1^0 = \bar{X}_1 \cdot (\hat{\beta}^{(0)} - \beta^{(0)}) + \hat{\mathbb{E}}_{i:W_i=0} \hat{\gamma}_i \left(\epsilon_i + X_i \beta^{(0)} - X_i \hat{\beta}^{(0)} \right)$$

$$= \left(\bar{X}_1 - \hat{\gamma}' \bar{X}_0 \right) \cdot (\hat{\beta}^{(0)} - \beta^{(0)}) + \hat{\mathbb{E}}_{i:W_i=0} \hat{\gamma}_i \epsilon_i$$

With high dimensions, we could try to estimate $\hat{\beta}$ and the propensity score with LASSO or post-LASSO rather than OLS. However, this may not be good enough. It is also not clear how to get good estimates of the inverse propensity score weights γ_i, in particular if we don't want to assume that the propensity model is sparse (e.g. if the treatment assignment is a complicated function of confounders).

Residuals on Residuals

- Small data approach (a la Robinson's 1988) analyzed a semi-parametric model
  - Model $Y_i = \tau W_i + g(X_i) + \epsilon_i$
  - Goal: estimate τ
  - Approach: residuals on residuals gives $\sqrt{n}$-consistent and asymptotically normal estimator
  - Regress $Y_i - \hat{g}(X_i)$ on $W_i - \widehat{\mathbb{E}[W_i|X_i]}$

Double Machine Learning

- Chernozhukov et al (2017):
  - Model $Y_i = \tau W_i + g(X_i) + \epsilon_i$, $\mathbb{E}[W_i|X_i] = h(X_i)$
  - Goal: estimate τ
  - Use a modern machine learning method like random forests to estimate the "nuisance parameters"
  - Regress $Y_i - \hat{g}(X_i)$ on $W_i - \widehat{\mathbb{E}[W_i|X_i]}$
  - If ML method converges at the rate $n^{\frac{1}{4}}$, residuals on residuals gives $\sqrt{n}$-consistent and asymptotically normal estimator

Comparing Straight Regression to Double ML

- Moments used in estimation:
  - Regression: $\mathbb{E}[(Y_i - W_i\tau - g(X_i)) \cdot W_i] = 0$
  - Double ML:
    $\mathbb{E}[((Y_i - \hat{g}(X_i) - (W_i - \hat{h}(X_i]))\tau) \cdot (W_i - \hat{h}(X_i))] = 0$
- Double robustness and orthogonality: Robinson's result implies that if $\hat{g}(X_i)$ is consistent, then $\hat{\tau}$ is the regression coefficient of the residual on residual regression, and even if $\hat{h}$ is wrong, the orthogonality of the residual of the outcome regression and the residual $W_i - \hat{h}$ still holds
- Neyman orthogonality: the Double ML moment condition has the property that when evaluated at $\hat{g} = g$ and $\hat{h} = h$, small changes in either of them do not change the moment condition. The moment condition is minimized at the truth.
- You are robust to small mistakes in estimation of nuisance parameters, unlike regression approach

Comparing Straight Regression to Double ML

Application to Ghana Data (Duflo et al, 2017) with 2000 controls

- Study effect of secondary education.
- Ground truth: experimental estimates of the effect of secondary education.
- Try to recover experimental estimates from observational/non-experimental data using **2,000** controls.

Returns To Secondary School Completion for Males

Outcome	Experimental	Observ.: OLS (5 controls)	Observ.: DML
Standardized Score	0.502	0.595	**0.486**
	(0.205)	(0.069)	(0.066)
Wage Worker	0.057	0.091	**0.082**
	(0.109)	(0.036)	(0.037)
Log Earnings	-0.195	**-0.094**	-0.064
	(0.245)	(0.087)	(0.088)
Partner pregnant	-0.089	-0.167	**-0.120**
	(0.093)	(0.032)	(0.030)

Doubly Robust with Nuisance Parameters

Another approach:

- Estimate $\hat{\tau}(x)$ (the CATE, $E[Y_i(1) - Y_i(0)|X = x]$, $\hat{\mu}(w, x)$, and $\hat{e}(x)$ using your favorite ML estimator, e.g. use causal forest from generalized random forest package for $\hat{\tau}(x)$.
- Average up efficient scores (doubly robust approach):

$$\hat{\tau} = \sum_i \hat{\tau}(X_i) + W_i \cdot \frac{Y_i - \hat{\mu}(1, X_i)}{\hat{e}(x)}$$

$$-(1 - W_i) \cdot \frac{Y_i - \hat{\mu}(0, X_i)}{1 - \hat{e}(x)}$$

- Standard error is just the standard deviation of each component of the sum, across units, divided by the square root of the number of observations

An Efficient Approach with Non-Sparse Propensity

The solution proposed in Athey, Imbens and Wager (2016) for attacking the gap

$$\hat{\mu}_1^0 - \mu_1^0 = \left(\bar{X}_1 - \hat{\gamma}'\bar{X}_0\right) \cdot (\hat{\beta}^{(0)} - \beta^{(0)}) + \hat{\mathbb{E}}_{i:W_i=0}\hat{\gamma}_i\epsilon_i$$

is to bound 1st term by selecting γ_i's using brute force. In particular:

$$\hat{\gamma} = \text{argmin}_\gamma \zeta \cdot \|\bar{X}_1 - \gamma'\bar{X}_0\|_\infty + (1-\zeta)\|\gamma\|_2^2$$

The parameter ζ is a tuning parameter; the paper shows that ζ exists such that the γ's exist to tightly bound the first term above. With overlap, we can make $\|\bar{X}_1 - \gamma'\bar{X}_0\|_\infty$ be $O(\sqrt{\frac{log(p)}{n}})$.
Result: If the outcome model is sparse, estimate β using LASSO yielding bias of second term $O_p\left(k\sqrt{\frac{\log(p)}{n}}\right)$, so the bias term is $O(k\frac{log(p)}{n})$, so for k small enough, the last term involving $\hat{\gamma}_i\epsilon_i$ dominates, and ATE estimator is $O(\frac{1}{\sqrt{n}})$.

Why Approximately Balancing Beats Propensity Weighting

One question is why the balancing weights perform better than the propensity score weights. To gain intution, suppose the propensity score has the following logistic form,

$$e(x) = \frac{\exp(x \cdot \theta)}{1 + \exp(x \cdot \theta)}.$$

After normalization, the inverse propensity score weights satisfy

$$\gamma_i \propto \exp(x \cdot \theta).$$

The efficient estimator for θ is the maximum likelihood estimator,

$$\hat{\theta}_{\mathrm{ml}} = \arg\max_{\theta} \sum_{i=1}^{n} \left\{ W_i X_i \cdot \theta - \ln(1 + \exp(X_i \cdot \theta)) \right\}.$$

An alternativeis the method of moments estimator $\hat{\theta}_{\mathrm{mm}}$ that balances the covariates exactly:

$$\bar{X}_0 = \sum_{\{i:W_i=0\}} X_i \frac{\exp(X_i \cdot \theta)}{\sum_{\{j:W_j=0\}} \exp(X_j \cdot \theta)}.$$

Why Approximately Balancing Beats Propensity Weighting

An alternative is the method of moments estimator $\hat{\theta}_{\text{mm}}$ that balances the covariates exactly:

$$\bar{X}_0 = \sum_{\{i:W_i=0\}} X_i \frac{\exp(X_i \cdot \theta)}{\sum_{\{j:W_j=0\}} \exp(X_j \cdot \theta)}.$$

with implied weights $\gamma_i \propto \exp(X_i \cdot \hat{\theta}_{\text{mm}})$.

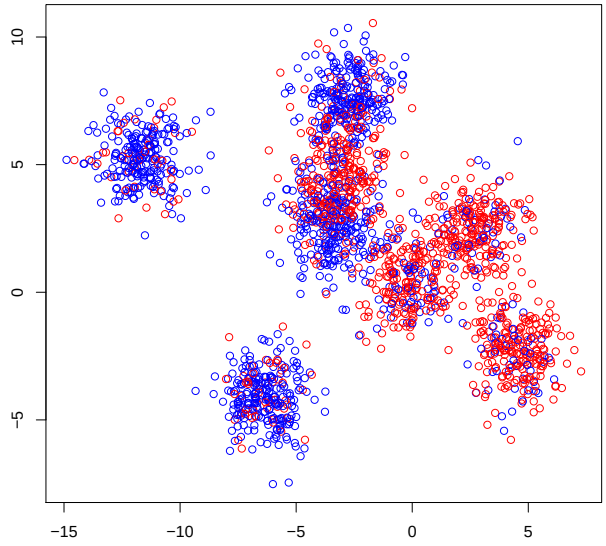
- The only difference between the two sets of weights is that the parameter estimates $\hat{\theta}$ differ.
- The estimator $\hat{\theta}_{\text{mm}}$ leads to weights that achieve exact balance on the covariates, in contrast to either the true value θ, or the maximum likelihood estimator $\hat{\theta}_{\text{ml}}$.
- The goal of balancing (leading to $\hat{\theta}_{\text{mm}}$) is different from the goal of estimating the propensity score (for which $\hat{\theta}_{\text{ml}}$ is optimal).

Summarizing the Approximate Residual Balancing Method of Athey, Imbens, Wager (2016)

- Estimate lasso (or elastic net) of Y on X in control group.
- Find "approximately balancing" weights that make the control group look like the treatment group in terms of covariates, while attending to the sum of squares of the weights. With many covariates, balance is not exact.
- Adjust the lasso prediction of the counterfactual outcome for the treatment group (if it had been control) using approximately balancing weights to take a weighted average of the residuals from the lasso model.

Main result: if the model relating outcomes to covariates is sparse, and there is overlap, then this procedure achieves the semi-parametric efficiency bound. No other method is known to do this for non-sparse propensity models.
Simulations show that it performs much better than alternatives when propensity is not sparse.

Simulation Experiment



The design X is "clustered." We study the following settings for β:

Dense: $\beta \propto (1, 1/\sqrt{2}, ..., 1/\sqrt{p})$,

Harmonic: $\beta \propto (1/10, 1/11, ..., 1/(p+9))$,

Moderately sparse: $\beta \propto (\underbrace{10, ..., 10}_{10}, \underbrace{1, ..., 1}_{90}, \underbrace{0, ..., 0}_{p-100})$,

Very sparse: $\beta \propto (\underbrace{1, ..., 1}_{10}, \underbrace{0, ..., 0}_{p-10})$.

Simulation Experiment

Beta Model	dense		harmonic		moderately sparse		very sparse	
Overlap (η)	0.1	0.25	0.1	0.25	0.1	0.25	0.1	0.25
Naive	0.672	0.498	0.688	0.484	0.686	0.484	0.714	0.485
Elastic Net	0.451	0.302	0.423	0.260	0.181	0.114	0.031	**0.021**
Approximate Balance	0.470	0.317	0.498	0.292	0.489	0.302	0.500	0.302
Approx. Resid. Balance	**0.412**	**0.273**	**0.399**	**0.243**	**0.172**	**0.111**	**0.030**	**0.021**
Inverse Prop. Weight	0.491	0.396	0.513	0.376	0.513	0.388	0.533	0.380
Inv. Prop. Resid. Weight	0.463	0.352	0.479	0.326	0.389	0.273	0.363	0.248
Double-Select + OLS	0.679	0.368	0.595	0.329	0.239	0.145	0.047	0.023

Simulation results, with $n = 300$ and $p = 800$. Approximate residual balancing estimates $\hat{\beta}$ using the elastic net. Inverse propensity residual weighting is like our method, except with $\gamma_i = 1/\hat{e}(X_i)$. We report root-mean-squared error for τ.

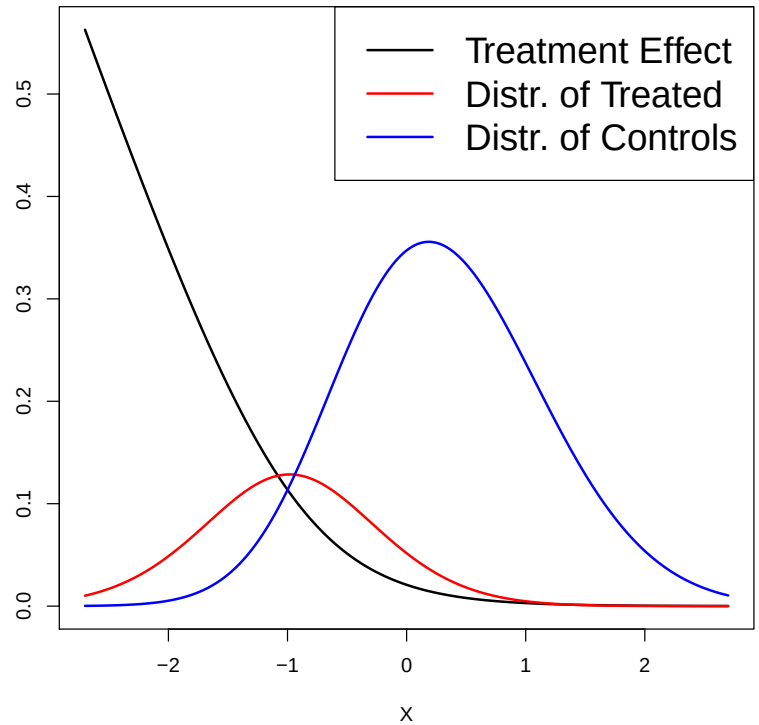
Observation: Weighting regression residuals works better than weighting the original data; balanced weighting works better inverse-propensity weighting.

Simulation Experiment

		$\beta_j \propto 1\left(\{j \leq 10\}\right)$		$\beta_j \propto 1/j^2$		$\beta_j \propto 1/j$	
n	p	$\eta = 0.25$	$\eta = 0.1$	$\eta = 0.25$	$\eta = 0.1$	$\eta = 0.25$	$\eta = 0.1$
200	400	0.90	0.84	0.94	0.88	0.84	0.71
200	800	0.86	0.76	0.92	0.85	0.82	0.71
200	1600	0.84	0.74	0.93	0.85	0.85	0.73
400	400	0.94	0.90	0.97	0.93	0.90	0.78
400	800	0.93	0.91	0.95	0.90	0.88	0.76
400	1600	0.93	0.88	0.94	0.90	0.86	0.76
800	400	0.96	0.95	0.98	0.96	0.96	0.90
800	800	0.96	0.94	0.97	0.96	0.94	0.90
800	1600	0.95	0.92	0.97	0.95	0.93	0.86

We report coverage of τ for 95% confidence intervals constructed by approximate residual balancing.

Simulation Experiment



We are in a misspecified linear model; the "main effects" model is 10-sparse and linear.

Simulation Experiment

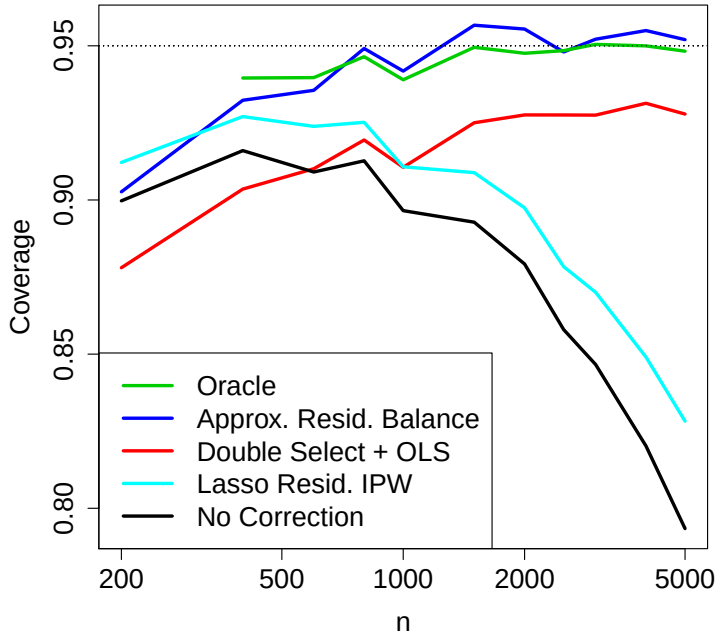
n	400					1000				
p	100	200	400	800	1600	100	200	400	800	1600
Naive	1.72	1.73	1.73	1.72	1.74	1.71	1.70	1.72	1.70	1.72
Elastic Net	0.44	0.46	0.50	0.51	0.54	0.37	0.39	0.39	0.40	0.42
Approximate Balance	0.48	0.55	0.61	0.63	0.70	0.24	0.30	0.38	0.40	0.45
Approx. Resid. Balance	**0.24**	**0.26**	**0.28**	**0.29**	**0.32**	**0.16**	**0.17**	**0.18**	**0.19**	**0.20**
Inverse Prop. Weight	1.04	1.07	1.11	1.13	1.18	0.82	0.84	0.88	0.89	0.94
Inv. Prop. Resid. Weight	1.29	1.30	1.31	1.31	1.33	1.25	1.25	1.26	1.25	1.28
Double-Select + OLS	0.28	0.29	0.31	0.31	0.34	0.24	0.25	0.25	0.25	0.26

Approximate residual balancing estimates $\hat{\beta}$ using the elastic net. Inverse propensity residual weighting is like our method, except with $\gamma_i = 1/\hat{e}(X_i)$. We report root-mean-squared error for τ_1.

Estimating the Effect of a Welfare-to-Work Program

Data from the California GAIN Program, as in Hotz et al. (2006).

- Program separately randomized in: Riverside, Alameda, Los Angeles, San Diego.
- Outcome: mean earnings over next 3 years.
- We hide county information. Seek to compensate with $p = 93$ controls.
- Full dataset has $n = 19170$.



Closing Thoughts

What are the pros and cons of approximate residual balancing vs. inverse-propensity residual weighting?

Pros of balancing:

- Works under weaker assumptions (only overlap).
- Algorithmic transparency.
- Hirshberg and Wager (2017) extend to nonlinearities in outcome function

Pros of propensity methods:

- Potential for double robustness in traditional sense.
- Potential for efficiency under heteroskedasticity.
- Generalizations beyond linearity.
- ...