ECON 293/MGTECON 634:
Machine Learning and Causal Inference

Susan Athey and Stefan Wager
Stanford University

Lecture 3: Average Treatment Effect Estimation
in Observational Studies

20 April 2018

The potential outcomes model

For a set of i.i.d. subjects $i = 1, \ldots, n$, 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** $W_i \in \{0, 1\}$.

We assume that the treatment is **unconfounded** (aka selection on observables) (Rosenbaum & Rubin, 1983):

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

We estimate the average treatment effect $\tau = \mathbb{E}\left[Y_i(1) - Y_i(0)\right]$.

Two representations of the average treatment effect

Given **unconfoundedness** $\{Y_i(0),\ Y_i(1)\} \perp\!\!\!\perp W_i \mid X_i$, we have

$$\begin{aligned}
\tau &= \mathbb{E}\left[Y_i(1) - Y_i(0)\right] = \mathbb{E}\left[\mathbb{E}\left[Y_i(1) \,\middle|\, X_i\right] - \mathbb{E}\left[Y_i(0) \,\middle|\, X_i\right]\right] \\
&= \mathbb{E}\left[\frac{\mathbb{E}\left[W_i \,\middle|\, X_i\right]\mathbb{E}\left[Y_i(1) \,\middle|\, X_i\right]}{e(X_i)} - \frac{\mathbb{E}\left[1 - W_i \,\middle|\, X_i\right]\mathbb{E}\left[Y_i(0) \,\middle|\, X_i\right]}{1 - e(X_i)}\right] \\
&= \mathbb{E}\left[\frac{\mathbb{E}\left[W_i\, Y_i(1) \,\middle|\, X_i\right]}{e(X_i)} - \frac{\mathbb{E}\left[(1 - W_i)\, Y_i(0) \,\middle|\, X_i\right]}{1 - e(X_i)}\right] \\
&= \mathbb{E}\left[\frac{W_i Y_i}{e(X_i)} - \frac{(1 - W_i) Y_i}{1 - e(X_i)}\right],
\end{aligned}$$

where $e(x) = \mathbb{P}\left[W_i = 1 \,\middle|\, X_i = x\right]$ is the **propensity score**. This suggests estimating $e(x)$, and then **inverse-propensity weighting**

$$\hat{\tau} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{W_i Y_i}{\hat{e}(X_i)} - \frac{(1 - W_i) Y_i}{1 - \hat{e}(X_i)}\right).$$

Two representations of the average treatment effect

Given **unconfoundedness** $\{Y_i(0),\ Y_i(1)\} \perp\!\!\!\perp W_i \mid X_i$, we have

$$\begin{aligned}
\tau &= \mathbb{E}\left[Y_i(1) - Y_i(0)\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[Y_i(1) \,\middle|\, X_i\right] - \mathbb{E}\left[Y_i(0) \,\middle|\, X_i\right]\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[Y_i \,\middle|\, X_i,\ W_i = 1\right] - \mathbb{E}\left[Y_i \,\middle|\, X_i,\ W_i = 0\right]\right] \\
&= \mathbb{E}\left[\mu_{(1)}(X_i) - \mu_{(0)}(X_i)\right],
\end{aligned}$$

where $\mu_{(w)}(x) = \mathbb{E}\left[Y_i \,\middle|\, X_i = x,\ W_i = w\right]$. This suggests an **estimator** based on a **regression adjustment**:

$$\hat{\tau} = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{\mu}_{(1)}(X_i) - \hat{\mu}_{(0)}(X_i)\right),$$

where $\hat{\mu}_{(w)}(X_i)$ is obtained by regression Y_i on X_i on those observations with $W_i = w$.

ATE estimation via OLS

A simple instance of this idea involves estimating $\mu_{(0)}(x)$ and $\mu_{(1)}(x)$ via **ordinary least-squares regression** (OLS). Specifically, in `R` notation, we first run two separate regressions (recall that `lm` is the `R` command for running linear regression):

$$\hat{\beta}_{(0)} \leftarrow \texttt{lm}\left(Y_i \sim X_i, \texttt{ subset } W_i = 0\right),$$
$$\hat{\beta}_{(1)} \leftarrow \texttt{lm}\left(Y_i \sim X_i, \texttt{ subset } W_i = 1\right).$$

We then make predictions $\hat{\mu}_{(w)}(x) = \hat{\beta}_{(w)} x$, and obtain a **treatment effect** estimate as

$$\hat{\tau} = \frac{1}{n} \sum_{i=1}^{n} \left(\hat{\mu}_{(1)}(X_i) - \hat{\mu}_{(0)}(X_i)\right) = \left(\hat{\beta}_{(1)} - \hat{\beta}_{(0)}\right) \overline{X},$$

where $\overline{X} = \sum_{i=1}^{n} X_i$. Note that, X implicitly includes an **intercept**.

Example: ATE estimation via OLS

```
library(sandwich) # for robust standard errors
treat_data = read.table("...../nswre74_treated.txt")
control_data = read.table("...../psid3_controls.txt")
combined = rbind(treat_data, control_data)
X = combined[,2:9]; Y = combined[,10]; W = combined[,1]

# First center the X, then run OLS with full W:X
# interactions. With this construction, the
# W-coefficient can be interpreted as ATE.
X.centered = scale(X, center = TRUE, scale = FALSE)
ols.fit = lm(Y ~ W * X.centered)

# Use robust standard errors
tau.hat = coef(ols.fit)["W"]
tau.se = sqrt(sandwich::vcovHC(ols.fit)["W", "W"])
print(paste0("95% CI: ", round(tau.hat),
             " +/- ", round(1.96 * tau.se)))
"95% CI: 2107 +/- 2379"
```

ATE estimation via OLS

Note that we did **not** do the following: Fit a single linear regression on all the data

$$\left(\hat{\tau},\, \hat{\beta}\right) \leftarrow \texttt{lm}\left(Y_i \sim W_i + X_i\right),$$

and then report the W_i-coefficient of that regression.

This is very common, but **not justified by the potential outcomes** framework, and should only be interpreted as an ATE estimate if you **actually believe** the model

$$Y_i = W_i\tau + X_i\beta + \varepsilon.$$

It is often said that all models are **wrong**, but some are **useful**. Our first OLS-based method used linear modeling as a useful tool for estimation in the Neyman-Rubin causal model. We will discuss how to interpret the second in upcoming weeks.

ATE estimation via the lasso?

OLS is optimal for learning linear models in **low dimensions**, i.e., with p predictors using n samples when $p \ll n$. In many modern applications, however, p may be of comparable size (or larger than) the sample size n. In this case, we often use the **lasso**:

$$\hat{\beta}_{lasso} = \text{argmin} \left\{ \frac{1}{n} \sum_{i=1}^{n} (Y_i - X_i \beta)^2 + \lambda \|\beta\|_1 \right\}.$$

By analogy to the low-dimensional case, may be tempted to use

$$\begin{aligned}
\hat{\beta}_{(0)} &\leftarrow \texttt{lasso}\left(Y_i \sim X_i,\ \texttt{subset}\ \ W_i = 0\right), \\
\hat{\beta}_{(1)} &\leftarrow \texttt{lasso}\left(Y_i \sim X_i,\ \texttt{subset}\ \ W_i = 1\right), \\
\hat{\tau} &= \left(\hat{\beta}_{(1)} - \hat{\beta}_{(0)}\right) \overline{X}.
\end{aligned}$$

Is this any good? The fundamental difference between the lasso and OLS is that the lasso has **bias** (and, in fact, any method in high dimensions must have bias).

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 $X \in \mathbb{R}^{n \times p}$ satisfies a restricted eigenvalue condition: no small group of variables is nearly collinear. Then we can show:

$$\left\|\hat{\beta} - \beta\right\|_2 = \mathcal{O}_P\left(\sqrt{\frac{k \log(p)}{n}}\right), \quad \left\|\hat{\beta} - \beta\right\|_1 = \mathcal{O}_P\left(k\sqrt{\frac{\log(p)}{n}}\right).$$

At a high level, this error arises because the lasso **shrinks** each coefficient on the order of $\sqrt{\log(p)/n}$.

Smaller k (i.e., a sparser truth), is better for the lasso:

- If $k \ll n/\log(p)$, we say the problem is **sparse**, and the lasso will make accurate predictions. For example, if

$$X_{ij} \overset{\text{iid}}{\sim} \mathcal{N}(0, 1), \text{ then } \mathbb{E}\left[(X\hat{\beta} - X\beta))^2\right] = \left\|\hat{\beta} - \beta\right\|_2^2.$$

- If $k \ll \sqrt{n}/\log(p)$, we say the problem is **sparse**, we can build confidence intervals for β_j via the **debiased lasso**.

Why Lasso Regression Adjustments Don't Work

We have $\mathbb{E}\left[X \,\middle|\, W=0\right] = \mathbf{0}$, $\mathbb{E}\left[X \,\middle|\, W=1\right] = \mathbf{1}$, such that $Y(w) = X\beta_{(w)} + \text{noise}$, $\text{Var}\left[Y(w) \,\middle|\, X\right] = \sigma^2$ and

$$\left(\beta_{(0)}\right)_j = \left(\beta_{(1)}\right)_j = 1\left(\{i \leq k\}\right)\sigma\sqrt{\log(p)/n}, \text{ and so}$$
$$\tau = \mathbb{E}\left[X\right] \cdot \left(\beta_{(1)} - \beta_{(0)}\right) = 0.$$

The lasso fits, $\hat{\mu}_{(w)}(x) = \hat{a}_{(w)} + x\hat{\beta}_{(w)}$, with **intercept** $\hat{a}_{(w)}$.

In order to **zero-out noise terms**, the lasso must eliminate all signals smaller than $\sigma\sqrt{2\log(p)/n}$. Thus, with high probability,

$$\hat{a}_{(0)}^{lasso} \approx 0, \quad \hat{\beta}_{(0)}^{lasso} = 0, \quad \text{and}$$
$$\hat{a}_{(1)}^{lasso} \approx k\sqrt{\log(p)/n}, \quad \hat{\beta}_{(1)}^{lasso} = 0.$$

Combining these into an ATE estimate, we get

$$\hat{\tau}^{lasso} = \hat{a}_{(1)}^{lasso} - \hat{a}_{(0)}^{lasso} + \overline{X} \cdot \left(\hat{\beta}_{(1)}^{lasso} - \hat{\beta}_{(0)}^{lasso}\right) \approx k\sigma\sqrt{\log(p)/n}.$$

Thus, the lasso has an error on the order of $k\sigma\sqrt{\log(p)/n}$.

Why Lasso Regression Adjustments Don't Work

- The lasso only looks for strong relationships between X and the outcome Y.
- But, for estimating ATE, it's also important to capture variables with a strong relationship between X and W.
- Strong variables in the propensity model can leak confounding effects, even if the corresponding X-Y effect is so small the lasso ignores it.
- This was not a problem with OLS, because OLS is unbiased (so it tries to fit every coefficient accurately, even if it's close to 0).

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

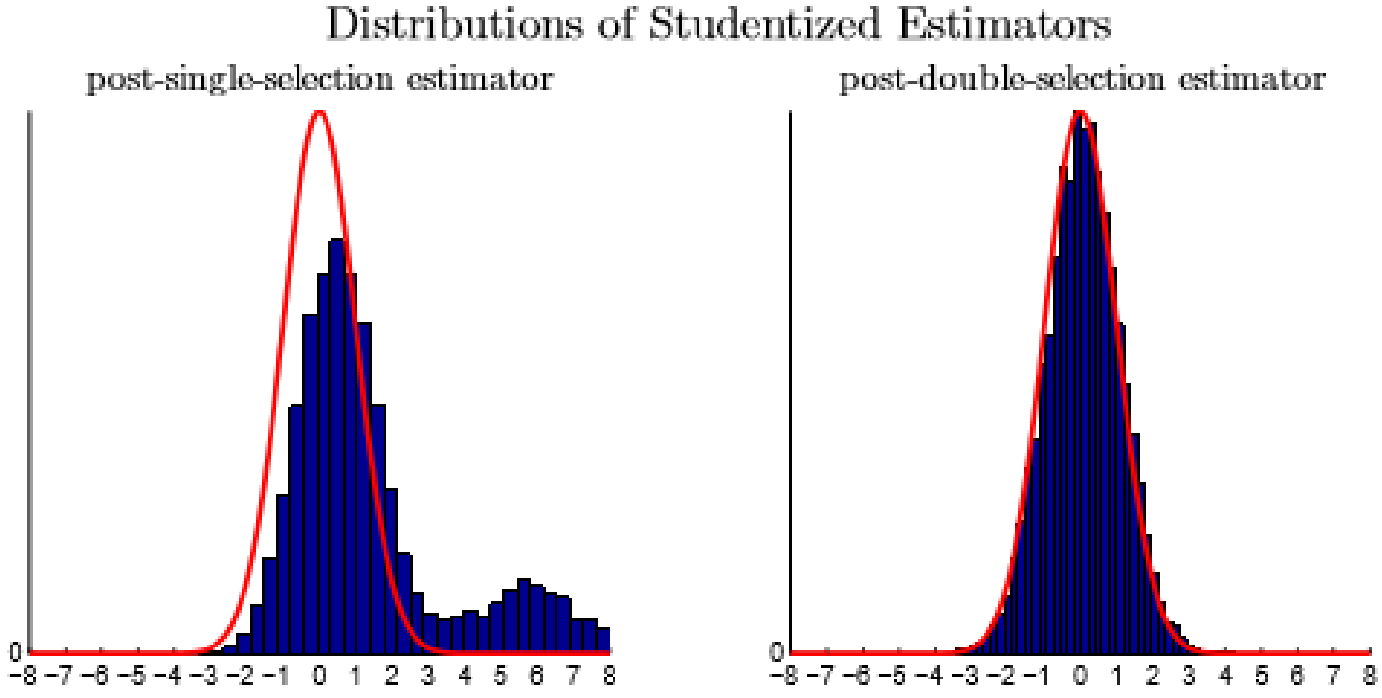
Belloni, Chernozukov, and Hansen (2014) propose a simple fix to this problem.

- Run a 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 on both the treated on control samples. Select variables with non-zero coefficients at a selected λ (may be different than first λ).
- Run OLS of Y on W interacted with the union of selected variables. Conclude as in the regular OLS case.

The third step above is not as good at purely **predicting** Y as using only second set. But it is more accurate for the ATE.

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

Single v. Double Selection in BCH Algorithm



Recap: ATE estimation via the lasso

The **lasso** can be use to estimate conditional response functions $\mu_{(w)}(x) = \mathbb{E}\left[Y_i(w) \,\middle|\, X_i = x\right]$ as

$$\hat{\mu}^{lasso}_{(w)}(x) = \hat{a}^{lasso}_{(w)} + x\hat{\beta}^{lasso}_{(w)}.$$

Because we're in high dimensions, we need to **regularize**, and this leads to **bias**.

- The lasso is calibrated to make good **predictions**, but not necessarily to make good **ATE estimates**.
- The simple lasso regression adjustment $\hat{\tau} = \frac{1}{n}\sum_{i=1}^{n}(\hat{\mu}^{lasso}_{(1)}(X_i) - \hat{\mu}^{lasso}_{(0)}(X_i))$ may fail badly.

The BCH method provides a fix to this problem, by "un-regularizing" features that are important in the propensity model.

- Does this idea generalize beyond **linear models**?
- What if the propensity model isn't **sparse**?

Augmented Inverse-Propensity Weighting

There is a more flexible approach to using machine learning methods for ATE estimation that relies on the **propensity score**

$$e(x) = \mathbb{P}\left[W_i = 1 \,\middle|\, X_i = x\right].$$

Suppose that we have estimates $\hat{\mu}_{(w)}(x)$ from any machine learning method, and also have propensity estimate $\hat{e}(x)$. AIPW then uses:

$$\hat{\tau} = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{\mu}_{(1)}(X_i) - \hat{\mu}_{(0)}(X_i) + \frac{W_i}{\hat{e}(X_i)}\left(Y_i - \hat{\mu}_{(1)}(X_i)\right) - \frac{1 - W_i}{1 - \hat{e}(X_i)}\left(Y_i - \hat{\mu}_{(0)}(X_i)\right)\right).$$

In considerable generality, this is a good estimator of the ATE.

Augmented Inverse-Propensity Weighting

To interpret AIPW, it is helpful to write it as

$$\hat{\tau}_{AIPW} = D + R$$

$$D = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{\mu}_{(1)}(X_i) - \hat{\mu}_{(0)}(X_i)\right)$$

$$R = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{W_i}{\hat{e}(X_i)}\left(Y_i - \hat{\mu}_{(1)}(X_i)\right) - \frac{1-W_i}{1-\hat{e}(X_i)}\left(Y_i - \hat{\mu}_{(0)}(X_i)\right)\right).$$

D is the direct **regression adjustment** estimator using $\hat{\mu}_{(w)}(x)$, and R is an IPW estimator applied to the **residuals** $Y_i - \hat{\mu}_{(W_i)}(X_i)$.

Qualitatively, AIPW uses propensity weighting on the residuals to **debias** the direct estimate.

Augmented Inverse-Propensity Weighting

To understand why AIPW works, we can compare it to an **oracle** that gets to use the true values of $\mu_{(w)}(x)$ and $e(x)$:

$$\tilde{\tau}_{AIPW} = \frac{1}{n}\sum_{i=1}^{n}\left(\mu_{(1)}(X_i) - \mu_{(0)}(X_i) + \frac{W_i}{e(X_i)}\left(Y_i - \mu_{(1)}(X_i)\right) - \frac{1-W_i}{1-e(X_i)}\left(Y_i - \mu_{(0)}(X_i)\right)\right).$$

"Theorem." If the first-stage function estimates satisfy

$$\mathbb{E}\left[\left(\hat{\mu}_{(w)}(X) - \mu_{(w)}(X)\right)^2\right]^{\frac{1}{2}}\mathbb{E}\left[\left(\hat{e}(X) - e(X)\right)^2\right]^{\frac{1}{2}} = o_P\left(\frac{1}{\sqrt{n}}\right),$$

and we also have **overlap**, then $\hat{\tau}_{AIPW}$ and $\tilde{\tau}_{AIPW}$ satisfy

$$\sqrt{n}\left(\hat{\tau}_{AIPW} - \tilde{\tau}_{AIPW}\right) \to_p 0.$$

In other words, $\hat{\tau}_{AIPW}$ and $\tilde{\tau}_{AIPW}$ are first-order equivalent.

Augmented Inverse-Propensity Weighting

The upshot of this result is that we can study $\tilde{\tau}_{AIPW}$ instead of $\hat{\tau}_{AIPW}$. Because $\tilde{\tau}_{AIPW}$ is just an average of independent terms, a direct application of the **central limit theorem** implies that

$$\sqrt{n}\left(\tilde{\tau}_{AIPW} - \tau\right) \Rightarrow \mathcal{N}\left(0,\, V^*\right),$$

$$V^* = \text{Var}\left[\mu_{(1)}(X) - \mu_{(0)}(X)\right] + \mathbb{E}\left[\frac{\text{Var}\left[Y_i(1)\right] \mid X_i}{e(X_i)}\right] + \mathbb{E}\left[\frac{\text{Var}\left[Y_i(0)\right] \mid X_i}{1 - e(X_i)}\right].$$

Because $\hat{\tau}_{AIPW}$ and $\tilde{\tau}_{AIPW}$ are equivalent on the $\sqrt{n}$-scale, we then immediately get, whenever the result from the previous slide holds,

$$\sqrt{n}\left(\hat{\tau}_{AIPW} - \tau\right) \Rightarrow \mathcal{N}\left(0,\, V^*\right).$$

Moreover, it can be shown that this behavior is **optimal** for any ATE estimator, assuming a generic non-parametric setup.

Augmented Inverse-Propensity Weighting

To recap, we have considered $\hat{\tau}_{AIPW} = n^{-1} \sum_{i=1}^{n} \widehat{\Gamma}_i$, with

$$\widehat{\Gamma}_i = \hat{\mu}_{(1)}(X_i) - \hat{\mu}_{(0)}(X_i) + \frac{W_i}{\hat{e}(X_i)} \left(Y_i - \hat{\mu}_{(1)}(X_i) \right) - \dots$$

If the first-stage function estimates are **reasonably accurate**,

$$\mathbb{E}\left[\left(\hat{\mu}_{(w)}(X) - \mu_{(w)}(X) \right)^2 \right]^{\frac{1}{2}} \mathbb{E}\left[\left(\hat{e}(X) - e(X) \right)^2 \right]^{\frac{1}{2}} = o_P\left(\frac{1}{\sqrt{n}} \right),$$

this estimator is **first-order optimal**. Moreover, for **inference**, we can act as thought the $\widehat{\Gamma}_i$ were independent, and use

$$\widehat{\text{Var}}\left[\hat{\tau}_{AIPW} \right] = \widehat{V}_n := \frac{1}{n(n-1)} \sum_{i=1}^{n} \left(\widehat{\Gamma}_i - \hat{\tau}_{AIPW} \right)^2.$$

We can use this to build Gaussian **confidence intervals**:

$$\mathbb{P}\left[\tau \in \hat{\tau}_{AIPW} \pm z_{1-\alpha/2} \widehat{V}_n^{1/2} \right] \to 1 - \alpha.$$

Details #1: The assumptions

Our result relies on the assumption that

$$\mathbb{E}\left[\left(\hat{\mu}_{(w)}(X) - \mu_{(w)}(X)\right)^2\right]^{\frac{1}{2}} \mathbb{E}\left[\left(\hat{e}(X) - e(X)\right)^2\right]^{\frac{1}{2}} = o_P\left(\frac{1}{\sqrt{n}}\right).$$

One simple way to achieve this condition is if both $\hat{\mu}$ and $\hat{e}$ are $o(1/n^{1/4})$-consistent in root-mean squared error.

- In other words, our final estimate $\hat{\tau}_{AIPW}$ can be an **order of magnitude** more accurate than either nuisance component ($1/n^{1/2}$ vs $1/n^{1/4}$).
- The reason for this phenomenon is that, to first order, the errors in the $\hat{\mu}$ and $\hat{e}$ regressions **cancel out**.
- This is known as the **orthogonal moments** construction, and plays a key role in semiparametric statistics.

Of course, $o(1/n^{1/4})$-consistency is still a strong assumption, and may not always hold. The topic of when we can get $o(1/n^{1/4})$-consistency, and also when we can improve on the assumptions stated above, is the topic of a large literature.

Details #2: Cross-fitting

To get good behavior out of AIPW, we recommend **cross-fitting**

$$\hat{\tau} = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{\mu}_{(1)}^{(-i)}(X_i) - \hat{\mu}_{(0)}^{(-i)}(X_i) + \frac{W_i}{\hat{e}^{(-i)}(X_i)}\left(Y_i - \hat{\mu}_{(1)}^{(-i)}(X_i)\right) - \frac{1-W_i}{1-\hat{e}^{(-i)}(X_i)}\left(Y_i - \hat{\mu}_{(0)}^{(-i)}(X_i)\right)\right).$$

In other words, when estimating $e(X_i)$, use a model that **did not have access** to the i-th training example during training.

- A simple approach is to cut the data into K **folds**. Then, for each $k = 1, ..., K$, train a model on all but the k-th fold, and evaluate its predictions on the k-th fold.
- With forests, **leave-one-out** estimation is natural, i.e., $\hat{e}^{(-i)}(X_i)$ is trained on all but the i-th sample.

Chernozhukov et al. (2017) emphasize the role of cross-fitting in proving flexible efficiency results for AIPW.

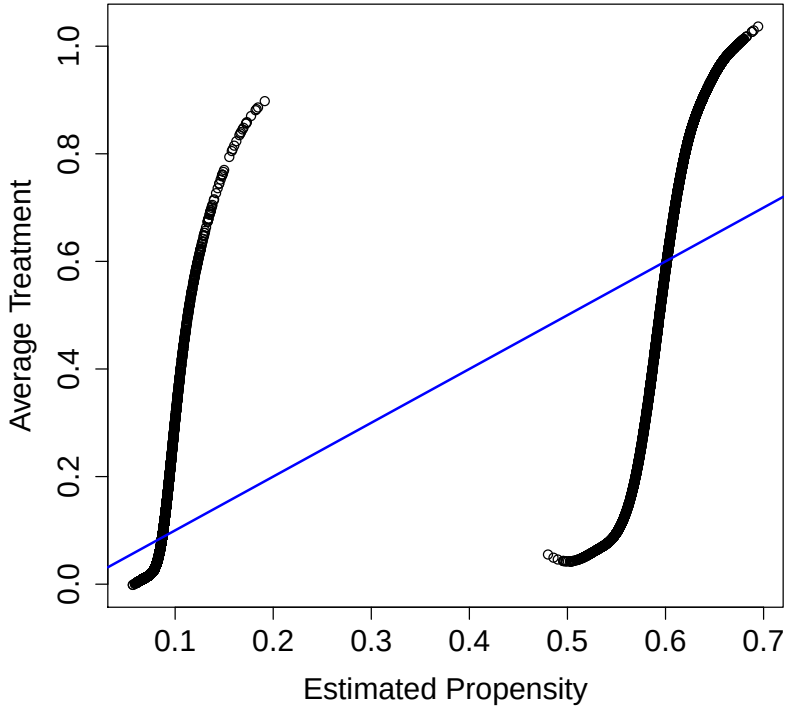
Details #2: Cross-fitting

Example from tutorial, with $n = 9750$ and $p = 21$.

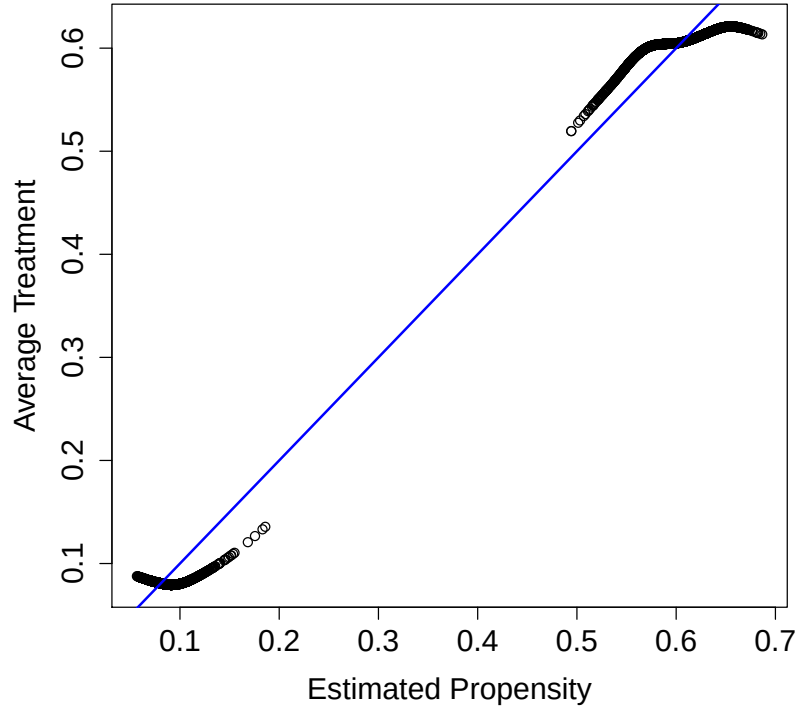
```
library(grf)
propensity_fit = regression_forest(Xmod, Wmod)

# If you ask a forest to predict without giving it a test
# set, it automatically does OOB on the training set.

ehat_oob = predict(propensity_fit)$predictions
ehat_naive = predict(propensity_fit,
                     newdata = Xmod)$predictions

c(OOB=mean(Wmod / ehat_oob), NAIVE=mean(Wmod / ehat_naive))

    OOB          NAIVE
    1.0173325    0.9093602
```

Details #2: Cross-fitting

Calibration plots run a single non-parametric regression of W_i against $\hat{e}(X_i)$, and are a good way to assess quality of a propensity fit. Ideally, the calibration curve should be close to the diagonal.



without crossfitting



with crossfitting

Details #3: Overlap

Overlap means that propensity scores are bounded away from 0 and 1:

$$\eta \leq \mathbb{P}\left[W_i = 1 \,\middle|\, X_i = x\right] \leq 1 - \eta, \quad \eta > 0,$$

for all possible value of x. The proof assumes overlap, and even the limiting **variance** gets bad as overlap gets bad:

$$V^* = \text{Var}\left[\mu_{(1)}(X) - \mu_{(0)}(X)\right] + \mathbb{E}\left[\frac{\text{Var}\left[Y_i(1)\right] \mid X_i}{e(X_i)}\right] + \mathbb{E}\left[\frac{\text{Var}\left[Y_i(0)\right] \mid X_i}{1 - e(X_i)}\right].$$

In applications, it is important to check overlap.

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.
- If we subset the data, need to be mindful of what the estimand is.

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.

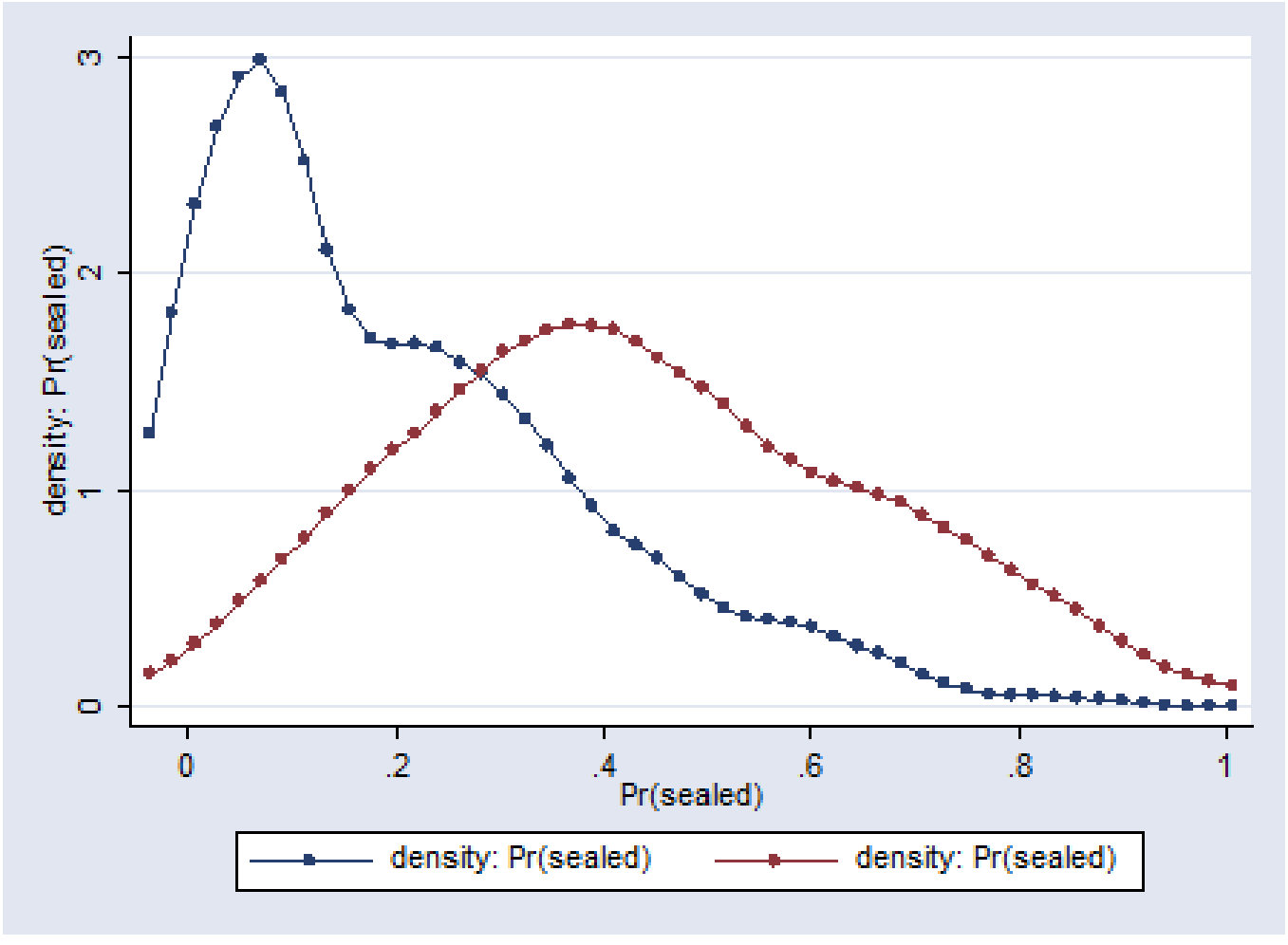
Propensity Score Plots: Assessing Overlap

Example: Athey, Levin and Seira analysis of timber auctions.

- The paper studies consequences of awarding contracts to harvest timber via first price sealed auction or open ascending auction.
- Assignment to first price sealed auction or open ascending auction:
  - In Idaho, auction mechanism is randomized for subset of tracts with different probabilities in different geographies;
  - In California, auction mechanism is determined by small v. large sales (with cutoffs varying by geography).
- So $W = 1$ if auction is sealed, and X represents geography, size and year.

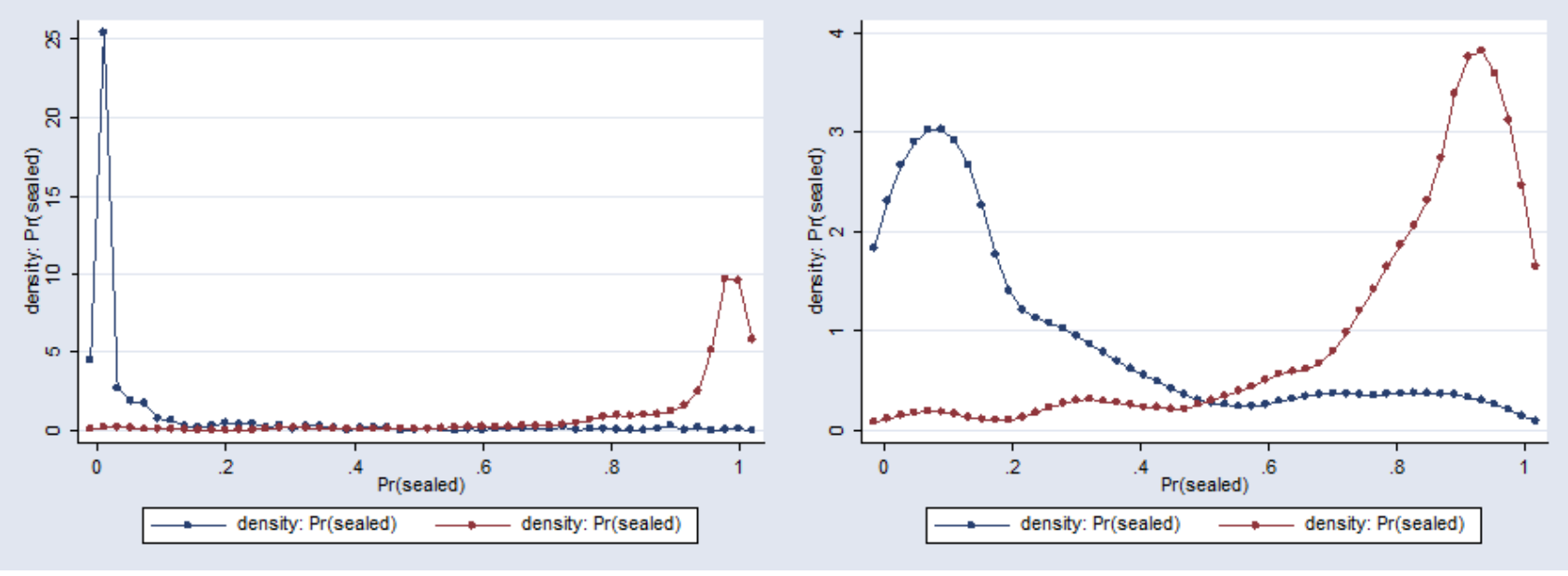
Propensity Score Plots: Assessing Overlap in ID

Very few observations with extreme propensity scores

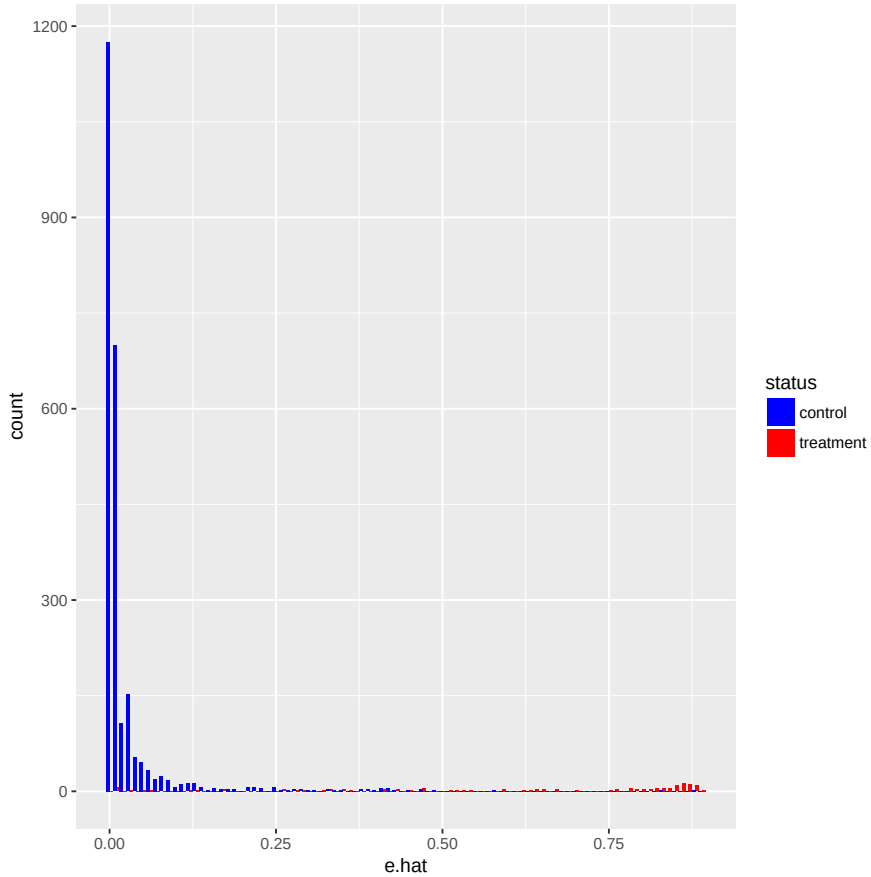


Propensity Score Plots: Assessing Overlap in CA

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

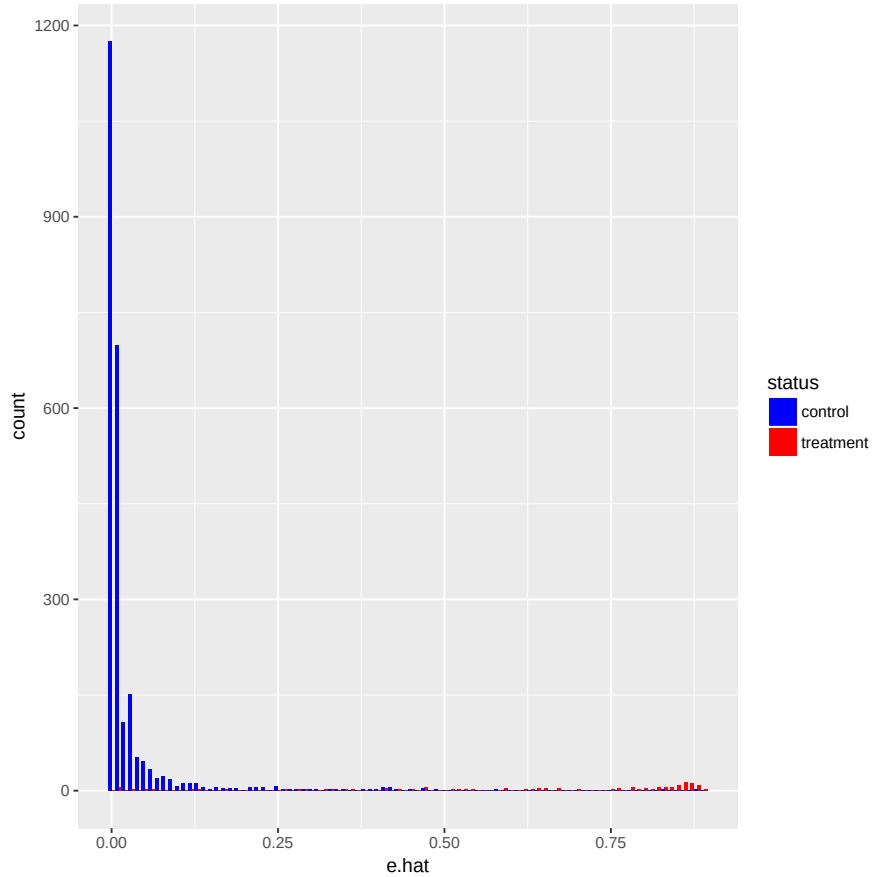


Overlap in the Lalonde data



Overlap on the Lalonde dataset, with full set of PSID controls. Many of the controls have essentially 0 propensity, but there is no overlap problem near 1. (The previous Lalonde example used a dataset filtered by unemployment, for better overlap.)

Overlap in the Lalonde data



We can find **propensity-matches** for the treated units, but not for all the controls. A simple way to trim away the overlap problem is to estimate an **average treatment effect on the treated**. Here, this may also be better conceptually justified.

Average treatment effect on the treated

Recall that the average treatment effect on the treated is

$$\tau_{ATT} = \mathbb{E}\left[Y_i(1) - Y_i(0) \,\middle|\, W_i = 1\right].$$

As usual, we can estimate it via several strategies. The direct **regression adjustment** estimator fits a model $\hat{\mu}_{(0)}(x)$ to the controls, and then uses it to impute what would have happened to the treated units (on average) in the control condition

$$\hat{\tau}_{ATT} = \frac{1}{n_1} \sum_{W_i=1} \left(Y_i - \hat{\mu}_{(0)}(X_i)\right).$$

The **propensity-weighted** estimator uses

$$\hat{\tau}_{ATT} = \frac{1}{n_1} \sum_{W_i=1} Y_i - \sum_{W_i=0} \frac{\hat{e}(X_i)}{1-\hat{e}(X_i)} Y_i \Bigg/ \sum_{W_i=0} \frac{\hat{e}(X_i)}{1-\hat{e}(X_i)}.$$

We never divide by $\hat{e}(x)$, so **propensities near 0** aren't a problem.

Average treatment effect on the treated

The **augmented propensity-weighted** estimator combines both

$$\hat{\tau}_{ATT} = \frac{1}{n_1} \sum_{W_i=1} \left(Y_i - \hat{\mu}_{(0)}(X_i)\right) - \sum_{W_i=0} \frac{\hat{e}(X_i)}{1-\hat{e}(X_i)} \left(Y_i - \hat{\mu}_{(0)}\right) \Bigg/ \sum_{W_i=0} \frac{\hat{e}(X_i)}{1-\hat{e}(X_i)}.$$

Again, this estimator is **asymptotically optimal** via an **orthogonal moments** argument. Also, we can write $\hat{\tau}_{ATT} = n^{-1} \sum_{i=1}^{n} \widehat{\Gamma}_i$,

$$\frac{\widehat{\Gamma}_i}{n} = \frac{W_i \left(Y_i - \hat{\mu}_{(0)}(X_i)\right)}{n_1} - \frac{(1-W_i)\frac{\hat{e}(X_i)}{1-\hat{e}(X_i)} \left(Y_i - \hat{\mu}_{(0)}\right)}{\sum_{W_i=0} \frac{\hat{e}(X_i)}{1-\hat{e}(X_i)}},$$

and, by the same argument as before, we can estimate variance via $V_n = \sum_{i=1}^{n} (\widehat{\Gamma}_i - \hat{\tau}_{ATT})^2 / (n(n-1))$ to build **confidence intervals**.

Overlap in the Lalonde data

```
library(grf) # for random forests
treat_data = read.table("...../nswre74_treated.txt")
control_data = read.table("...../psid_controls.txt")
combined = rbind(treat_data, control_data)
X = combined[,2:9]; Y = combined[,10]; W = combined[,1]
cf = causal_forest(X, Y, W)

ate.hat = average_treatment_effect(cf,
                                   target.sample = "all")
print(paste0("95% CI: ", round(ate.hat["estimate"]),
             " +/- ", round(1.96 * ate.hat["std.err"])))
Warning: Estimated treatment propensities go as low as 0.003...
[1] "95% CI: -3039 +/- 6388"

att.hat = average_treatment_effect(cf,
                                   target.sample = "treated")
print(paste0("95% CI: ", round(att.hat["estimate"]),
             " +/- ", round(1.96 * att.hat["std.err"])))
[1] "95% CI: 1142 +/- 1510"
```

Addressing failures in overlap

If there are some observations with propensities very near 0 and some very near 1, we need **more aggressive** methods:

- One idea is to fit a model for $\hat{e}(x)$, throw away all observations with $\hat{e}(X_i) \leq 0.1$ or $\hat{e}(X_i) \geq 0.9$, and estimate an ATE on the rest.
- Another idea is the weight observations by $\hat{e}(X_i)(1 - \hat{e}(X_i))$, so all observations with extreme weights are strongly enough discounted not to inflate variance.

In both cases, **interpretation** requires care.

Balancing vs Propensity-Weighting

The **augmented propensity-weighted** ATT estimator is

$$\hat{\tau}_{ATT} = \frac{1}{n_1} \sum_{W_i=1} \left(Y_i - \hat{\mu}_{(0)}(X_i)\right) - \sum_{W_i=0} \frac{\hat{e}(X_i)}{1-\hat{e}(X_i)} \left(Y_i - \hat{\mu}_{(0)}\right) \Bigg/ \sum_{W_i=0} \frac{\hat{e}(X_i)}{1-\hat{e}(X_i)}.$$

We first estimate $\hat{e}(x)$ in a way that is tuned for accurate **prediction** of the W_i, and then use it to **debias** the $\hat{\mu}_{(0)}(\cdot)$ fit.

- But does **optimal prediction** lead to **optimal debiasing**?

In the case of the **lasso**, it's also possible to use a more direct approach.

Balancing in High Dimensions

We consider the **broader class** of lasso-based estimators generalizing AIPW, using $\hat{\tau} = \overline{Y}_1 - \hat{m}^{(0)}$ with

$$\hat{m}^{(0)} = \overline{X}_1 \cdot \hat{\beta}^{(0)} + \sum_{\{i:W_i=0\}} \hat{\gamma}_i \left(Y_i - X_i \cdot \hat{\beta}^{(0)} \right).$$

Proposition. (Athey, Imbens, Wager; 2016) Suppose that $Y_i = X_i \cdot \beta^{(W_i)} + \varepsilon_i$. Writing $\text{err} = \hat{m}^{(0)} - \overline{X}_1 \cdot \beta^{(0)}$, we have

$$|\text{err}| \leq \left\| \overline{X}_1 - X(0)^\top \hat{\gamma} \right\|_\infty \left\| \hat{\beta}^{(0)} - \beta^{(0)} \right\|_1 + \left| \sum_{\{i:W_i=1\}} \hat{\gamma}\, \varepsilon_i \right|.$$

Here $X(0)$ is a subset of X corresponding to the control examples.

- Weighting and regression adjustments play helpful, **complementary roles**; neither is as good as both together.
- Does not depend on **estimability of the propensity scores**, or a relation of the form $\hat{\gamma}_i \propto e(X_i)/(1 - e(X_i))$.
- Does not rely on **screening**, a.k.a., accurate model selection.

Approximate Residual Balancing

Motivated by this proposition, we estimate τ as follows.

1. Estimate $\hat{\beta}^{(0)}$ using a lasso or elastic net (Hastie et al., 2015) on the control cases.
2. Estimate weights $\hat{\gamma}$ by quadratic programming:

$$\hat{\gamma} = \text{argmin}_{\gamma} \left\{ \|\gamma\|_2^2 + \left\| \overline{X}_1 - X(0)^\top \gamma \right\|_\infty^2 \right\},$$

   subject to constraints $\gamma_i \geq 0$ and $\sum_i \gamma_i = 1$.
3. Finally, our treatment effect estimate is $\hat{\tau} = \overline{Y}_1 - \hat{m}^{(0)}$,

$$\hat{m}^{(0)} = \overline{X}_1 \cdot \hat{\beta}^{(0)} + \sum_{\{i: W_i = 0\}} \hat{\gamma}_i \left(Y_i - X_i \cdot \hat{\beta}^{(0)} \right).$$

Software for `R` is available in the package `balanceHD`.

Motivating Approximate Residual Balancing

To motivate this method, suppose the propensity score has the following logistic form,

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

After normalization, the ATT propensity weights satisfy

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

The usual estimator for θ is the **maximum likelihood** estimator,

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

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

$$\overline{X}_1 = \sum_{\{i:W_i=0\}} X_i \frac{\exp(X_i \cdot \theta)}{\sum_{\{j:W_j=0\}} \exp(X_j \cdot \theta)}.$$

Motivating Approximate Residual Balancing

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

$$\overline{X}_1 = \sum_{\{i:W_i=0\}} X_i \frac{\exp(X_i \cdot \theta)}{\sum_{\{j:W_j=0\}} \exp(X_j \cdot \theta)},$$

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 **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 **propensity estimation** (for which $\hat{\theta}_{\text{ml}}$ is optimal).

At a high level, approximate residual balancing is an extension of this idea to the **high-dimensional** context.

ARB vs AIPW

Augmented IPW and approximate residual balancing are both **orthogonalizing** methods, and seek to **debias** a pilot lasso estimate, and get **efficient** treatment effect estimates.

At a very high level (see papers in course outline for more details), to get efficiency, AIPW needs

$$\mathbb{E}\left[\left(\hat{\mu}_{(w)}(X) - \mu_{(w)}(X)\right)^2\right]^{\frac{1}{2}} \mathbb{E}\left[\left(\hat{e}(X) - e(X)\right)^2\right]^{\frac{1}{2}} = o_P\left(\frac{1}{\sqrt{n}}\right),$$

whereas approximate residual balancing needs

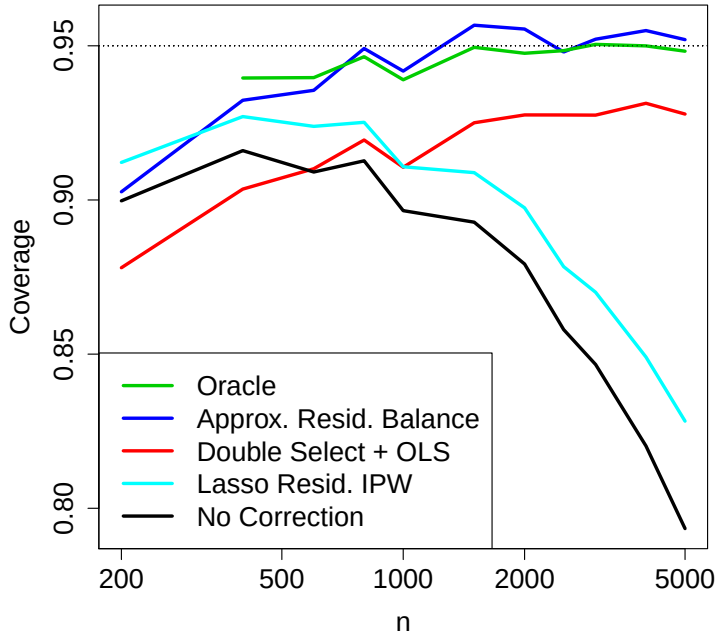
$$\mathbb{E}\left[\left(\hat{\mu}_{(w)}(X) - \mu_{(w)}(X)\right)^2\right]^{\frac{1}{2}} = o_P\left(\frac{1}{n^{1/4}}\right).$$

Which one is better depends on **how fast you can estimate the propensity score**. In high dimensions, 4-th root rates in L_2 error correspond to sparsity as $k \ll \sqrt{n}/\log(p)$.

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 $n = 19170$.



Extension: Average effect of a continuous treatment

For a set of i.i.d. subjects $i = 1, ..., n$, we observe

- A **feature vector** $X_i \in \mathcal{X}$,
- A **response** $Y_i \in \mathbb{R}$, and
- A **treatment assignment** $W_i \in \mathbb{R}$.

We posit a **conditionally linear model**, such that

$$Y_i = \mu(X_i) + W_i \, \tau(X_i) + \varepsilon_i, \quad \mathbb{E}\left[\varepsilon_i \,\middle|\, X_i, \, W_i = 0\right],$$

and seek to estimate $\tau_{ATT} = \mathbb{E}\left[\tau(X)\right]$.

In the case of a **binary treatment**, this setup can be motivated as average treatment effect estimation under **unconfoundedness** (Rosenbaum and Rubin, 1983),

$$\{Y_i(0), \, Y_i(1)\} \perp\!\!\!\perp W_i \;\; \Big| \;\; X_i,$$

where $Y_i(0), \, Y_i(1)$ denote Neyman-Rubin potential outcomes.

Extension: Average effect of a continuous treatment

Let $\mathcal{M}$ be a set of functions that can be defined via sparse linear combinations of a pre-determined dictionary. Then, the analogous **balancing** estimator is

$$\hat{\tau} = n^{-1} \sum_{i=1}^{n} \left(\hat{\tau}(X_i) + \hat{\gamma}_i \left(Y_i - \hat{\mu}(X_i) - W_i \hat{\tau}(X_i) \right) \right),$$

$$\hat{\gamma} = \text{argmin}_{\gamma} \left\{ \|\gamma\|_2^2 + \sup_{\mu \in \mathcal{M}} \left\{ \sum_{i=1}^{n} \gamma_i \mu(X_i) \right\}^2 + \sup_{\tau \in \mathcal{M}} \left\{ \sum_{i=1}^{n} \left(W_i \gamma_i - 1 \right) \tau(X_i) \right\}^2 \right\}.$$

Procedurally, the bias term enforces **approximate balance**, but now in a continuous sense. See Hirshberg and Wager (2018) for details + extensions.

Balancing in practice

So far, we have chosen our weights $\hat{\gamma}$ to minimize a weighted combination of variance and a bound on the conditional bias,

$$\sup_{\mu \in \mathcal{M}} \left\{ \sum_{i=1}^{n} \gamma_i \mu(X_i) \right\}^2 + \sup_{\tau \in \mathcal{M}} \left\{ \sum_{i=1}^{n} (W_i \gamma_i - 1) \tau(X_i) \right\}^2.$$

We chose the function class $\mathcal{M}$ because we believed that it asymptotically captures the behavior of $\hat{\mu}(x) - \mu(x)$ and $\hat{\tau}(x) - \tau(x)$.

In practice, we can make our procedure **more robust** by balancing on more, i.e., balancing $f \in \mathcal{M} \cup \mathcal{G}$ where $\mathcal{G}$ contains:

- Strata of any **fitted propensity model** $\hat{e}(x)$.
- Leaves from a pruned **causal tree**.
- ...

What is the baseline?

We posit a **conditionally linear model**, such that

$$Y_i = \mu(X_i) + W_i\,\tau(X_i) + \varepsilon_i, \quad \mathbb{E}\left[\varepsilon_i \,\middle|\, X_i,\, W_i = 0\right],$$

and seek to estimate $\tau_{ATE} = \mathbb{E}\left[\tau(X)\right]$. The **Riesz representer** is

$$g(X_i,\, W_i) = \frac{W_i - e(X_i)}{V(X_i)},\; e(X_i) = \mathbb{E}_{X_i}\left[W_i\right],\; V(X_i) = \mathrm{Var}_{X_i}\left[W_i\right].$$

A natural baseline is to use the **doubly robust** form:

$$\hat{\tau}_{DR} = n^{-1}\sum_{i=1}^{n}\left(\hat{\tau}(X_i) + \frac{W_i - \hat{e}(X_i)}{\widehat{V}(X_i)}\left(Y_i - \hat{\mu}(X_i) - W_i\hat{\tau}(X_i)\right)\right).$$

In the **binary case**, using $\widehat{V}(x) = \hat{e}(x)(1 - \hat{e}(x))$ recovers **AIPW**.
We also consider **oracle DR**, which uses the true $g(\cdot)$.

	method			double rob. plugin			augm. balance			augm. balance+			double rob. oracle		
	n	p	κ	rmse	bias	covg	rmse	bias	covg	rmse	bias	covg	rmse	bias	covg
setup A	600	6	3	**0.13**	0.03	0.98	0.14	0.03	0.98	**0.13**	0.00	0.98	0.18	-0.01	0.96
	600	6	4	0.16	0.06	0.92	0.16	0.04	0.94	**0.15**	0.03	0.93	0.21	0.00	0.92
	600	12	3	0.22	0.09	0.78	0.18	-0.00	0.87	**0.17**	0.05	0.90	0.27	-0.04	0.90
	600	12	4	0.21	0.14	0.78	**0.15**	0.01	0.94	0.17	0.09	0.90	0.23	-0.03	0.93
	1200	6	3	**0.10**	0.03	0.94	0.11	0.06	0.92	**0.10**	0.02	0.96	0.12	0.00	0.98
	1200	6	4	0.11	0.03	0.94	0.11	0.05	0.92	**0.10**	0.02	0.96	0.13	0.00	0.94
	1200	12	3	0.11	0.02	0.90	**0.10**	0.01	0.95	**0.10**	0.02	0.94	0.14	0.00	0.94
	1200	12	4	0.15	0.06	0.86	**0.11**	0.00	0.92	0.12	0.04	0.90	0.16	-0.00	0.94
setup B	600	6	3	0.23	0.23	0.04	0.14	0.13	0.44	**0.11**	0.09	0.72	0.08	-0.00	0.96
	600	6	4	0.20	0.20	0.12	0.13	0.11	0.54	**0.10**	0.09	0.72	0.07	-0.00	0.96
	600	12	3	0.25	0.24	0.03	0.21	0.20	0.10	**0.12**	0.10	0.70	0.08	-0.01	0.95
	600	12	4	0.21	0.20	0.09	0.18	0.17	0.16	**0.11**	0.10	0.72	0.08	-0.01	0.94
	1200	6	3	0.20	0.19	0.01	0.10	0.09	0.55	**0.07**	0.05	0.78	0.05	-0.01	0.97
	1200	6	4	0.18	0.18	0.01	0.08	0.07	0.68	**0.06**	0.05	0.85	0.05	-0.01	0.96
	1200	12	3	0.23	0.22	0.00	0.16	0.15	0.02	**0.08**	0.07	0.76	0.05	-0.00	0.96
	1200	12	4	0.19	0.19	0.00	0.14	0.14	0.13	**0.08**	0.07	0.70	0.05	0.00	0.94
setup C	600	6	4	0.22	0.16	0.84	0.16	-0.03	0.94	**0.11**	-0.02	1.00	0.16	0.03	0.94
	600	6	5	0.20	0.14	0.88	0.15	-0.05	0.93	**0.11**	-0.02	1.00	0.15	0.00	0.93
	600	12	4	0.23	0.15	0.86	0.18	-0.09	0.88	**0.14**	-0.04	0.96	0.17	-0.01	0.91
	600	12	5	0.24	0.17	0.82	0.19	-0.09	0.89	**0.13**	-0.05	0.97	0.17	-0.01	0.94
	1200	6	4	0.13	0.09	0.90	0.10	-0.03	0.94	**0.07**	-0.01	1.00	0.10	0.00	0.96
	1200	6	5	0.14	0.08	0.91	0.11	-0.05	0.94	**0.08**	-0.01	1.00	0.11	0.00	0.94
	1200	12	4	0.14	0.08	0.88	0.13	-0.07	0.88	**0.08**	-0.02	0.98	0.11	-0.00	0.94
	1200	12	5	0.14	0.09	0.87	0.13	-0.07	0.90	**0.08**	-0.02	1.00	0.11	-0.00	0.96

Simulations based on balancing an L_1-ball of polynomial basis functions (and interactions). Target coverage is 95%.
Non-augmented methods did worse (not shown).

Implementation + replication: `amlinear` for `R` on `github`.

The Effect of Winning the Lottery on Earnings

Data from Imbens, Rubin and Sacerdote (2001) on "big" lottery winners, who won a "jackpot" of \$1k–\$100k per year over 20 years.

Question: Does lottery income reduce earnings? I/R/S model a **constant treatment effect**.

We examine robustness of their findings to potential **treatment heterogeneity**.

There is **confounding** due to survey non-response (42% of contacted lottery winners responded).

estimator	estimate	std. err
$*$ OLS without controls	-0.176	0.039
$*$ OLS with controls	-0.106	0.032
$*$ residual-on-residual OLS	-0.110	0.032
plugin Riesz weighting	-0.175	—
doubly robust plugin	-0.108	0.042
minimax linear weighting	-0.074	—
augm. minimax linear	-0.091	0.044
minimax linear+ weighting	-0.083	—
augm. minimax linear+	-0.097	0.045

Outcome is mean earnings in 6 years following win. We have $p = 12$ and $n = 194$. Estimators marked by $*$ assume a constant effect.

Closing thoughts

Augmented inverse-propensity weighting and approximate residual balancing are both special cases of a very broad class of methods studied in **semiparametric statistics**.

- The goal is to estimate a single **"parameter"** (e.g., an ATE) while controlling for non-parametric confounding (aka nuisance components).
- Semiparametric problems arise frequently in causal inference.

A big idea at the intersection of machine learning and economics is that we can put **machine learning** and **optimization** tools to good use in semiparametric problems.

1. Carefully discuss **identification** of causal parameters .
2. Estimate non-causal aspects of the problem via **any method** that does the job.
3. Use **orthogonal moments** for accurate inference.