Part 4: Inference via double robustness

What we want

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

$$\left[\{Y_i(0),\ Y_i(1)\} \perp\!\!\!\perp W_i\right] \,\middle|\, X_i.$$

Our **desideratum** is an estimator $\hat{\tau}$ with familiar "parametric" behavior that satisfies a central limit theorem of the form

$$\sqrt{n}\left(\hat{\tau} - \tau\right) \Rightarrow \mathcal{N}\left(0,\ V\right),$$

thus enabling **confidence intervals**.

What we have

For this purpose, we assume access to a **predictive black box**: A machine learning method that can provide "pretty good" estimates of regression functions that satisfy

$$\sqrt{\mathbb{E}\left[\left(\hat{\mu}_{(w)}(X_i) - \mu_{(w)}(X_i)\right)^2\right]},\ \sqrt{\mathbb{E}\left[\left(\hat{e}(X_i) - e(X_i)\right)^2\right]} \ll \frac{1}{\sqrt[4]{n}}.$$

Why is this a natural thing to assume?

- We want to frame our assumption in terms of **RMSE**, because that's what machine learning methods claim to be good at (and what we tune them for).
- Assuming a parametric rate of convergence ($1/\sqrt{n}$) would be too optimistic. Assuming a 4-th root rate is a way of saying the **machine learning method is pretty accurate**, but doesn't achieve parametric rates.

How can we use this predictive black box to reach our goal?

Augmented Inverse-Propensity Weighting

The answer is use a method that combines estimates of both the **outcome regression** and the **propensity score**

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

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.

A Simple Example

Consider an example with $X_i \sim \mathcal{N}(0, I)$, $n = 1,000$ and $p = 20$:

$$
\begin{aligned}
e(x) &= 1/(1 + e^{-x_1}) \\
\mu_{(0)}(x) &= (x_1 + x_2)_+ \\
\mu_{(1)}(x) &= (x_1 + x_3)_+ - 0.05.
\end{aligned}
$$

First, model $\hat{\mu}_{(w)}(x)$ with a **random forest**, and use:

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

Across 100 trials, we get **0.09**; the true treatment effect is -**0.05**.

The random forest fit for $\hat{\mu}_{(w)}(x)$ is doing OK, but there's not enough data to nail the functional form.

A Simple Example

Now consider AIPW. We also fit $\hat{e}_{(w)}(x)$ with a **random forest**, and then take

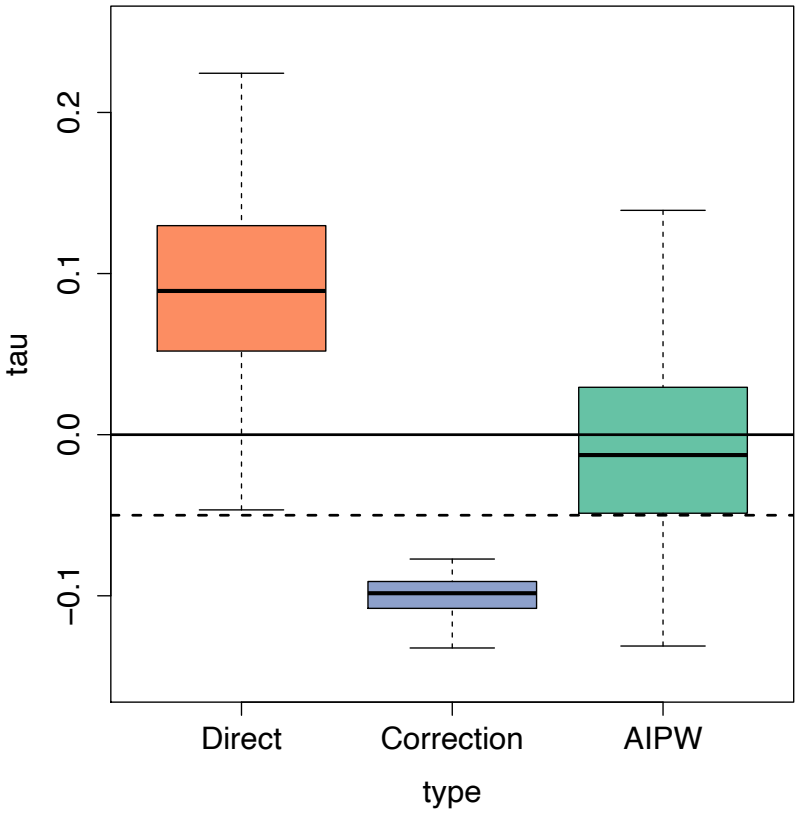
$$\hat{\tau}_{AIPW} = D + R$$

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

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

Here, the **direct** regression adjustment D is on average **0.09**, the **correction** R is on average -**0.10**, and **AIPW** gives us on average -**0.01**, which is closer to the **true** value -**0.05**.

A Simple Example



Implementing AIPW with Random Forests

The AIPW may have a daunting functional form. But once you know you need to use doubly robust estimators when deploying machine learning methods for causal inference, you can look for packages that implement this.

The `grf` package has a forest-based **implementation** of AIPW:

```
cf = causal_forest(X, Y, W)
ate.hat = average_treatment_effect(cf)
```

NB: The causal forest function fits random forest estimates of $\hat{\mu}_{(w)}(x)$ and $\hat{e}(x)$ under the hood. The average treatment effect function extracts them to form doubly robust scores.

Understanding 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 machine learning predictions satisfy

$$\mathbb{E}\left[\left(\hat{\mu}_{(w)}(X) - \mu_{(w)}(X)\right)^2\right]^{\frac{1}{2}},\ \mathbb{E}\left[\left(\hat{e}(X) - e(X)\right)^2\right]^{\frac{1}{2}} \ll \frac{1}{\sqrt[4]{n}},$$

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.

Understanding 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

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

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

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.

Inference with Augmented Inverse-Propensity Weighting

We are considering $\hat{\tau}_{AIPW} = n^{-1} \sum_{i=1}^{n} \widehat{\Gamma}_i$, with

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

If the first-stage estimates are reasonably accurate, this estimator is to first order **not affected by errors** in $\hat{e}(\cdot)$ and $\hat{\mu}_{(w)}(\cdot)$.

As a consequence, for **inference**, we can act as though $\hat{\tau}$ were the average of independent terms $\widehat{\Gamma}_i$, with variance

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

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

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

Details #1: 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. (2018) emphasize the role of cross-fitting in proving flexible efficiency results for AIPW.

Details #2: 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:

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

In applications, it is important to check overlap.

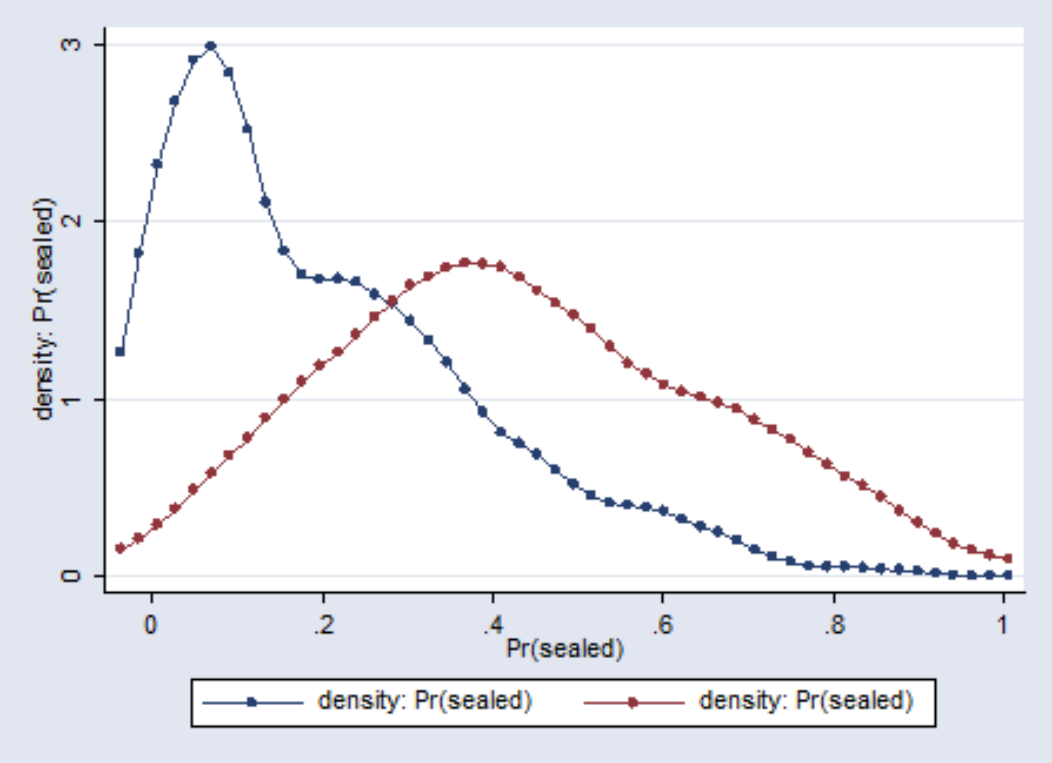
Example: Propensity Scores and Overlap

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.

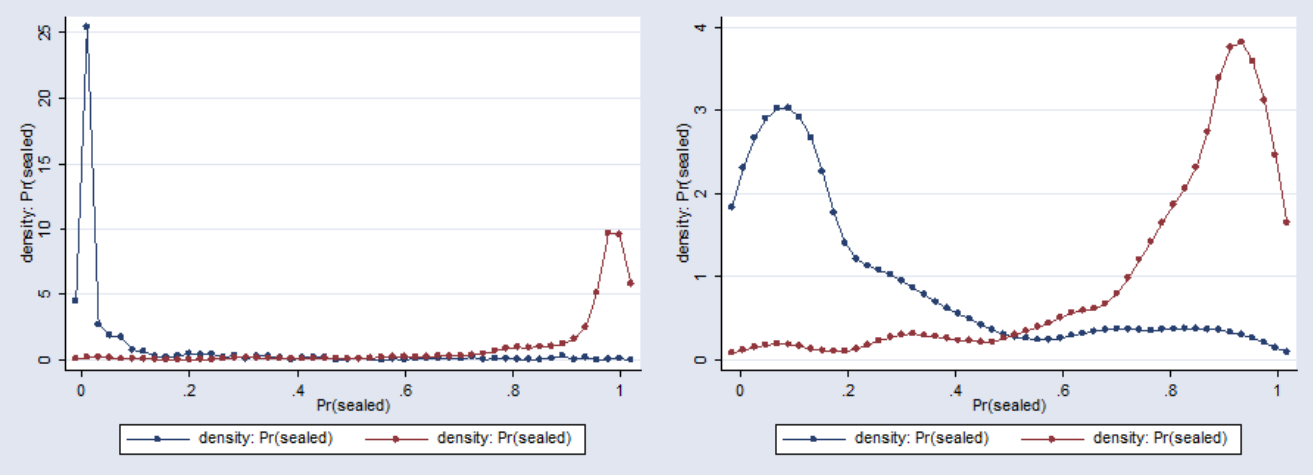
Propensity Score Plots: Assessing Overlap in ID

Very few observations with extreme propensity scores



Propensity Score Plots: Assessing Overlap in CA

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



Recap

Treatment effects are important in many scientific analyses.

Once we have **identified** treatment effects via unconfoundedness, we can **estimate** them by combining flexible **machine learning** methods with **augmented IPW**.

- Formally, AIPW yields **semiparametrically efficient** estimates of the treatment effect, provided the inputs from machine learning methods are accurate enough.
- In practice, AIPW makes our procedure robust to **regularization bias**.
- AIPW allows from simple **confidence intervals** that do not depend on which specific machine learning method we used.

AIPW lets machine learning focus on what it's good at (i.e., accurate predictions), and then uses its outputs for efficient treatment effect estimation.

	linear specification	general setting
linear regression	parametric behavior	inconsistent
machine learning	consistent, but with finite sample bias	consistent, but with finite sample bias

When deploying machine learning in causal inference, we need to use strategies that are **robust** to moderate finite sample errors.

In the context of **average treatment effect estimation**, AIPW is one solution with this property.

References

AIPW + machine learning, and generalizations.

- Chernozhukov, Victor, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. **"Double/debiased machine learning for treatment and structural parameters."** *The Econometrics Journal*, 21(1), 2018.
- Farrell, Max H., Tengyuan Liang, and Sanjog Misra. **"Deep neural networks for estimation and inference."** *Econometrica*, 89(1), 2021.
- van der Laan, Mark J., and Sherri Rose. *Targeted learning: causal inference for observational and experimental data.* Springer Science & Business Media, 2011.

Treatment effect estimation in high-dimensional problems.

- Athey, Susan, Guido Imbens, and Stefan Wager. **"Approximate residual balancing: Debiased inference of average treatment effects in high dimensions."** *Journal of the Royal Statistical Society Series B*, 80(4), 2018.
- Belloni, Alexandre, Victor Chernozhukov, and Christian Hansen. **"Inference on treatment effects after selection among high-dimensional controls."** *The Review of Economic Studies*, 81(2), 2014.