Plan for the rest of the semester

- Selected Advanced topics
  - Machine learning, including deep learning methods for causal effect estimation
  - Causal models of relational data, temporal data
  - Advanced methods for learning causal models
  - Causal effects generalizability (e.g., transportability)
- Selected Applications
  - Healthcare
  - Algorithmic fairness
  - Explainable ML
- Student project presentations
Machine learning methods for causal effect estimation

• Causal trees and forests
  • First introduce the ideas assuming randomized experiments (no confounding)
  • Then modify the approach to account for confounding

Causal inference: A missing data problem

<table>
<thead>
<tr>
<th>ID</th>
<th>Education</th>
<th>Treated</th>
<th>No job training</th>
<th>Job training</th>
<th>Treatment effect $\tau_i = Y_i(1) - Y_i(0)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>High school</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>High school</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>College</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>College</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>
Causal inference: A missing data problem

<table>
<thead>
<tr>
<th>ID</th>
<th>Education</th>
<th>Treated</th>
<th>No job training</th>
<th>Job training</th>
<th>Treatment effect</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>High school</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>High school</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>College</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>College</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

\[
\bar{\tau} = \bar{Y}_{i:W_i=1}(1) - \bar{Y}_{i:W_i=0}(0) \\
= 1 - 0.5 \\
= 0.5
\]

Causal inference: A missing data problem

<table>
<thead>
<tr>
<th>ID</th>
<th>Education</th>
<th>Treated</th>
<th>No job training</th>
<th>Job training</th>
<th>Treatment effect</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>High school</td>
<td>0</td>
<td>0</td>
<td>?</td>
<td>?</td>
</tr>
<tr>
<td>2</td>
<td>High school</td>
<td>1</td>
<td>?</td>
<td>1</td>
<td>?</td>
</tr>
<tr>
<td>3</td>
<td>College</td>
<td>0</td>
<td>1</td>
<td>?</td>
<td>?</td>
</tr>
<tr>
<td>4</td>
<td>College</td>
<td>1</td>
<td>?</td>
<td>1</td>
<td>?</td>
</tr>
</tbody>
</table>

If \( W_i \perp \{Y_i(0), Y_i(1)\}, \) then

\[
\hat{\tau} = \bar{Y}_{i:W_i=1} - \bar{Y}_{i:W_i=0} \\
= 1 - 0.5 \\
= 0.5
\]
Causal inference: A missing data problem

Potential employment

<table>
<thead>
<tr>
<th>ID</th>
<th>Education</th>
<th>Treated</th>
<th>No job training $Y_i(0)$</th>
<th>Job training $Y_i(1)$</th>
<th>Treatment effect $\tau_i = Y_i(1) - Y_i(0)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>High school</td>
<td>0</td>
<td>0</td>
<td>?</td>
<td>?</td>
</tr>
<tr>
<td>2</td>
<td>High school</td>
<td>1</td>
<td>?</td>
<td>1</td>
<td>?</td>
</tr>
<tr>
<td>3</td>
<td>College</td>
<td>0</td>
<td>1</td>
<td>?</td>
<td>?</td>
</tr>
<tr>
<td>4</td>
<td>College</td>
<td>1</td>
<td>?</td>
<td>1</td>
<td>?</td>
</tr>
</tbody>
</table>

What if we want to study $\tau_i = f(X_i)$?

<table>
<thead>
<tr>
<th>Education</th>
<th>Treated</th>
<th>No job training $Y_i(0)$</th>
<th>Job training $Y_i(1)$</th>
<th>Treatment effect $\tau_i = Y_i(1) - Y_i(0)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>High school</td>
<td>1</td>
<td>1 - 0.5</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>College</td>
<td>0</td>
<td>0.5</td>
<td>?</td>
<td>?</td>
</tr>
</tbody>
</table>

What if there are dozens of $X$ variables?
What if $X$ is continuous?
What if the effects of $X$ are heterogeneous?
• Start with a simpler prediction question.
• Which subgroups of $X$ have very different average outcomes?

Prediction: One Tree

\[
\text{MSE}_0 = \frac{1}{n} \sum (Y_i - \bar{Y})^2
\]

\[
\text{MSE}_1 = \frac{1}{n} \sum (Y_i - \tilde{Y}^*_{j: x_j \in c(x_i|a_j)})^2
\]

All observations

\[ X_i < k \quad X_i \geq k \]

Choose $k$ to minimize $\text{MSE}_1$
Prediction: One Tree

Mathematical expressions:
\[ \text{MSE}_0 = \frac{1}{n} \sum (Y_i - \hat{Y})^2 \]
\[ \text{MSE}_1 = \frac{1}{n} \sum (Y_i - \hat{Y}_{j: X_j \in \{x_i|n\}})^2 \]
\[ \text{MSE}_2 = \frac{1}{n} \sum (Y_i - \hat{Y}_{j: X_j \in \{x_i|n\}})^2 \]

Choose \( k_1 \) or \( k_2 \) to minimize \( \text{MSE}_2 \)

- Could continue until all leaves had only one observation.
- Unbiased but uselessly high variance!
- So, regularize: keep only splits that improve MSE by more than \( c \).

Partition \( \Pi \in \mathbb{P} \) into:
\[ \{ \ell_1 = \{x_i : x_i < 16\} , \ell_2 = \{x_i : x_i \geq 16\} \} \]

Prediction rule for new \( x \):
\[ \hat{\mu}(x) = \bar{Y}_{j: X_j \in \ell(x_i|n)} \]

- Could we use this method to find causal effects that are homogeneous within and heterogeneous across the leaves of the tree?
Causal tree: What’s different?

- We do not observe the ground truth
- Honest estimation:
  - One sample to choose partition
  - One sample to estimate leaf effects

Why?

- Fitting both on the training sample risks **overfitting**
- Estimating many “heterogeneous effects” that are really just noise idiosyncratic to the sample.
- We want to search for heterogeneity, not noise

\[
\text{MSE}_\mu(S^{\text{te}}, S^{\text{est}}, \pi) \equiv \frac{1}{|S^{\text{te}}|} \sum_{i \in S^{\text{te}}} \left\{ (Y_i - \hat{Y}_i(S^{\text{te}}, S^{\text{est}}, \pi))^2 - \hat{Y}_i \right\}
\]

**Honest criterion:** Maximize

\[
Q^H(\pi) \equiv -E_{S^{\text{te}}, S^{\text{est}}, S^{\text{tr}}} \left[ \text{MSE}_\mu(S^{\text{te}}, S^{\text{est}}, \pi(S^{\text{tr}})) \right]
\]

where \( \pi : \mathbb{R}^{p+1} \to \mathbb{P} \) is a function that takes a training sample \( S^{\text{tr}} \in \mathbb{R}^{p+1} \) and outputs a partition \( \pi \in \mathbb{P} \).

Simplifies the math, shifts the estimator by a constant
• Estimate expected MSE using only the training sample.
• Use this to place splits when training a tree.

Analytic estimator for $\text{EMSE}_\mu(\Pi)$

Expected mean squared error of partition $\Pi$

$$-\text{EMSE}_\mu(\Pi) = -\mathbb{E}_{X \sim \mathbb{S}_N} \left[ (Y_i - \mu(X_i | S^m, \Pi))^2 - Y_i^2 \right]$$

$$= -\mathbb{E}_{X \sim \mathbb{S}_N} \left[ (Y_i - \mu(X_i | \Pi) + \mu(X_i | \Pi) - \mu(X_i | S^m, \Pi))^2 - Y_i^2 \right]$$

$$= -\mathbb{E}_{X \sim \mathbb{S}_N} \left[ (Y_i - \mu(X_i | \Pi))^2 - Y_i^2 \right]$$

$$- \mathbb{E}_{X \sim \mathbb{S}_N} \left[ (\mu(X_i | \Pi) - \mu(X_i | S^m, \Pi))^2 \right]$$

$$- \mathbb{E}_{X \sim \mathbb{S}_N} \left[ 2(Y_i - \mu(X_i | \Pi))(\mu(X_i | \Pi) - \mu(X_i | S^m, \Pi)) \right]$$
\[
E(Y_i, X_i), \text{ST} \left[ (Y_i - \mu(X_i | \Pi))^2 - Y_i^2 \right] \\
- E_{X_i, \text{ST}} \left[ \mu(X_i | \text{ST}, \Pi) - \mu(X_i | \Pi))^2 \right]
\]

\[
= - E(Y_i, X_i), \text{ST} \left[ Y_i^2 + \mu^2(X_i | \Pi) - 2Y_i \mu(X_i | \Pi) \right] \\
- E_{X_i, \text{ST}} \left[ \mu(X_i | \text{ST}, \Pi) - \mu(X_i | \Pi))^2 \right]
\]

\[
= - E(Y_i, X_i), \text{ST} \left[ \mu^2(X_i | \Pi) - 2\mu(X_i | \Pi)\mu(X_i | \Pi) \right]
\]

Estimate with \[\hat{\mu}(X | S_{\text{est}}, \Pi) = \frac{S_{\text{est}}^2 \left( x \mid \Pi \right)}{\sum_{\ell} S_{\text{est}}^2 (\ell)} \]

\[\hat{E}_{X_i} \left[ \hat{\mu}(X_i | S_{\text{est}}, \Pi) \right] = \sum_{\ell} p_{\ell} S_{\text{est}}^2 (\ell) \]

(assuming \# equal leaf sizes)

\[
- \text{EMSE}_\mu(\Pi) = E_{X_i} \left[ \mu^2(X_i | \Pi) \right] - E_{X_i, \text{ST}} \left[ V(\hat{\mu}(X_i | S_{\text{est}}, \Pi) \right]
\]
\[ \mathbb{V}(\hat{\mu} \mid x, \Pi) = \mathbb{E}(\hat{\mu}^2 \mid x, \Pi) - \left[ \mathbb{E}(\hat{\mu} \mid x, \Pi) \right]^2 \]

\[ \frac{S_{\text{tr}}^2(\ell(x \mid \Pi))}{N_{\text{tr}}(\ell(x \mid \Pi))} \approx \hat{\mu}^2(x \mid S_{\text{tr}}, \Pi) - \mu^2(x \mid \Pi) \]

\[ \mu^2(x \mid \Pi) \approx \hat{\mu}^2(x \mid S_{\text{tr}}, \Pi) - \frac{S_{\text{tr}}^2(\ell(x \mid \Pi))}{N_{\text{tr}}(\ell(x \mid \Pi))} \]

\[ \hat{E}_x(\mu^2(X_i \mid \Pi)) \approx \frac{1}{N_{\text{tr}}} \sum_{i \in S_{\text{tr}}} \hat{\mu}^2(x_i \mid S_{\text{tr}}, \Pi) - \frac{1}{\#\ell} \sum_{\ell} \frac{S_{\text{tr}}^2(\ell)}{N_{\text{tr}}(\ell)} \]

\[ = \frac{1}{N_{\text{tr}}} \sum_{i \in S_{\text{tr}}} \hat{\mu}^2(x_i \mid S_{\text{tr}}, \Pi) - \frac{1}{N_{\text{tr}}} \sum_{\ell} S_{\text{tr}}^2(\ell) \]

\[ -\text{EMSE}_{\mu}(\Pi) = \mathbb{E}_{X_i} \left[ \mu^2(X_i \mid \Pi) \right] - \mathbb{E}_{S_{\text{test}}, X_i} \left[ \mathbb{V}(\hat{\mu}(X_i \mid S_{\text{test}}, \Pi)) \right] \]

\[ -\text{EMSE}_{\mu}(S_{\text{tr}}, N_{\text{test}}, \Pi) = \frac{1}{N_{\text{tr}}} \sum_{i \in S_{\text{tr}}} \hat{\mu}^2(x_i \mid S_{\text{tr}}, \Pi) - \frac{1}{N_{\text{tr}}} \sum_{\ell \in \Pi} S_{\text{tr}}^2(\ell) \]

\[ - \frac{1}{N_{\text{test}}} \sum_{\ell \in \Pi} S_{\text{tr}}^2(\ell) \]

\[ = \frac{1}{N_{\text{tr}}} \sum_{i \in S_{\text{tr}}} \hat{\mu}^2(x_i \mid S_{\text{tr}}, \Pi) - \left( \frac{1}{N_{\text{tr}}} + \frac{1}{N_{\text{test}}} \right) \sum_{\ell \in \Pi} S_{\text{tr}}^2(\ell) \]

Conventional CART criterion

Uncertainty about leaf means

\[ -\text{EMSE}_{\mu}(\Pi) = \mathbb{E}_{X_i} \left[ \mu^2(X_i \mid \Pi) \right] - \mathbb{E}_{S_{\text{test}}, X_i} \left[ \mathbb{V}(\hat{\mu}(X_i \mid S_{\text{test}}, \Pi)) \right] \]
Honest inference for treatment effects

Population-average potential outcomes within leaves:

\[
\mu(w, x \mid \Pi) \equiv \mathbb{E} \left[ Y_i(w) \mid X_i \in \ell(x \mid \Pi) \right]
\]

Potential outcome for treatment \( w \)
(heterogeneous by \( X \))

Averaged over controls \( X_i \) in the leaf

Average causal effect:

\[
\tau(x \mid \Pi) \equiv \mathbb{E} \left[ \frac{Y_i(1)}{\ell(x \mid \Pi)} - \frac{Y_i(0)}{\ell(x \mid \Pi)} \right] - \mu(0, x \mid \Pi)
\]

Average effect evaluated at (potentially moderating) covariate value \( x \)
Estimate:
\[ \hat{\mu}(w, x | S, \Pi) \equiv \frac{1}{\#(\{i \in S_w : X_i \in \tau(x|\Pi)\})} \sum_{i \in S_w : X_i \in \tau(x|\Pi)} Y_{i}^{obs} \]

MSE for treatment effects:
\[ \text{MSE}_T(S^{te}, S^{est}, \Pi) \equiv \frac{1}{\#(S^{te})} \sum_{i \in S^{te}} \left\{ \left( \tau_i - \hat{\tau}(X_i | S^{est}, \Pi) \right)^2 - \tau_i^2 \right\} \]

Challenge! \( \tau_i \) is never observed.

Adapt EMSE\(_\mu\) to estimate EMSE\(_T\)

\[-\text{EMSE}_\mu(S^{tr}, N^{est}, \Pi) = \frac{1}{N^{tr}} \sum_{i \in S^{tr}} \hat{\mu}^2(X_i | S^{tr}, \Pi) - \left( \frac{1}{N^{tr}} + \frac{1}{N^{est}} \right) \sum_{i \in \Pi} S^2_{S^i} \]

\[-\text{EMSE}_T(S^{tr}, N^{est}, \Pi) = \frac{1}{N^{tr}} \sum_{i \in S^{tr}} \hat{\tau}^2(X_i | S^{tr}, \Pi) \]

- Prefers leaves with heterogeneous effects
- Prefers leaves with good fit (leaf-specific effects estimated precisely)

- Conventional CART criterion
- Uncertainty about leaf means
- Valence of treatment effects across leaves
- Uncertainty about leaf treatment effects
Four partitioning estimators

1. Causal trees

Split by

\[-\text{EMSE}_r(S^{tr}, N^{est}, \pi) = \frac{1}{N^{tr}} \sum_{i \in S^{tr}} \frac{1}{\hat{\pi}(X_i | S^{tr}, \pi)} \]

- Prefer leaves with heterogeneous effects
- Prefer leaves with good fit (leaf-specific effects estimated precisely)

- **Benefit**: Prioritizes heterogeneity (\(\hat{\pi}\) varies a lot) and fit (within-leaf precision)
- **Drawback**: Cannot be done with off-the-shelf CART methods
2. Transformed outcome trees

Transform the outcome

\[ Y_i^* = \frac{Y_i W_i - p}{p(1-p)} = \mathbb{E}(Y_i^* \mid X_i = x) = \tau(x) \]

\[ \mathbb{E}(Y_i^*) = \mathbb{E}\left[ \frac{Y_i W_i - p}{p(1-p)} \right] \\
= \mathbb{E}\left[ \frac{Y_i W_i}{p(1-p)} \right] - \mathbb{E}\left[ \frac{p}{p(1-p)} \right] \\
= \mathbb{E}\left[ Y_i(1) \frac{W_i}{p(1-p)} \right] - \mathbb{E}\left[ \left( Y_i(1) W_i + Y_i(0)(1 - W_i) \right) \frac{p}{p(1-p)} \right] \\
= Y_i(1) \frac{1}{p(1-p)} \mathbb{E}[W_i] - Y_i(1) \frac{p}{p(1-p)} \mathbb{E}[W_i] - Y_i(0) \frac{p}{p(1-p)} \mathbb{E}[1 - W_i] \\
= Y_i(1) \frac{1}{p(1-p)} - Y_i(0) \frac{p}{p(1-p)} \mathbb{E}[1 - W_i] \\
= \tau_i
\]

2. Transformed outcome trees

- **Benefit:** Can use off-the-shelf CART methods for prediction
- **Drawbacks:** Inefficient. Treatment is ignored after transforming outcome.
- If within a leaf \( \mathcal{W} = p \) (by chance), then sample average within leaf is a poor estimator of \( \tau^* \).
3. Fit-based trees

Replace

$$\text{MSE}_\mu(S^{\text{te}}, S^{\text{est}}, \mathcal{P}) \equiv \frac{1}{\#(S^{\text{te}})} \sum_{i \in S^{\text{te}}} \left( Y_i - \hat{\mu}(X_i; S^{\text{est}}, \mathcal{P}) \right)^2$$

with the fit-based split rule

$$\text{MSE}_{\mu,W}(S^{\text{te}}, S^{\text{est}}, \mathcal{P}) \equiv \sum_{i \in S^{\text{te}}} \left( Y_i - \hat{\mu}_W(W_iX_i; S^{\text{est}}, \mathcal{P}) \right)^2$$

which loss by model fit within each leaf: the difference from the expected value for the treatment group of observation $i$.

Benefit: Prefers splits that lead to better fit.

Drawback: Does not prefer splits that lead to variation in treatment effects.

4. Squared T-statistic trees

Split based on:

$$T^2 \equiv \frac{N_L \left( \bar{Y}_L - \bar{Y}_R \right)^2}{S^2 / N_L + S^2 / N_R}$$

Benefit: Prefers splits that lead to variation in treatment effects.

Drawback: Missed opportunity to improve fit: ignores useful splits between leaves with similar treatment effects but very different average values.
From trees to forests: Double-sample trees

An individual tree can be noisy. Instead, we might fit a forest.

1. Draw a sample of size $s$
2. Split into an $I$ and $J$ sample.
3. Grow a tree on the $J$ sample
4. Estimate leaf-specific $\hat{\tau}_l$ using the $I$ sample

Repeat many times.

Advantages of forests:
- Consistent for true $\tau(x)$
- Asymptotic normality
- Asymptotic variance is estimable

Why double-sample forests:
- Advantage: Trees search for heterogeneous effects
- Disadvantage: Requires sample splitting

Wager and Athey, 2017

From trees to forests: Propensity trees

An individual tree can be noisy. Instead, we might fit a forest.

1. Draw a sample of size $s$
2. Grow a tree on the $J$ sample to predict $W$
   - Each leaf must have at least $k$ observations of each treatment class
3. Estimate $\hat{\tau}_l$ on each leaf

Repeat many times.

Advantages of forests:
- Consistent for true $\tau(x)$
- Asymptotic normality
- Asymptotic variance is estimable

Why propensity forests:
- Advantage: Can use full sample
- Disadvantage: Does not search for heterogeneous effects

Wager and Athey, 2017
Causal Trees

There is no ground truth: We never observe $\tau$.
Causal trees search for leaves with heterogeneous effects across leaves precisely-estimated leaf effects. 
Requirements include extra sample splitting.

Work well with randomized treatments. With selection on observables, the general recommendation is propensity forests:

- Maximizes the goal of addressing confounding by ignoring heterogeneous effects when choosing splits.
- Generalized random forests also perform well (Athey, Tibshirani, & Wager 2017).
- But “the challenge in using adaptive methods... is that selection bias can be difficult to quantify” (Wager & Athey, 2017).

If treatment is not randomized

- Causal trees find heterogeneous effects but cannot guarantee that confounding is addressed.
- How can we deal with high-dimensional confounding?
Addressing confounding

- What does address confounding? **Standardization**
- Why is tree-based standardization biased? **Regularization**
- Is there anything we can do? **Chernozhukov et al.**

---

**What works: Nonparametric standardization**

What if \( \{Y_i(0), Y_i(1)\} \perp W_i \) but \( \{Y_i(0), Y_i(1)\} \perp W_i | X_i \)?

We need to estimate \( \tau \) within each level of \( X_i \).

<table>
<thead>
<tr>
<th>ID</th>
<th>Education</th>
<th>Treated</th>
<th>No job training</th>
<th>Job training</th>
<th>Potential employment</th>
<th>Treatment effect</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>High school</td>
<td>0</td>
<td>0</td>
<td>?</td>
<td>?</td>
<td>?</td>
</tr>
<tr>
<td>2</td>
<td>High school</td>
<td>0</td>
<td>0</td>
<td>?</td>
<td>?</td>
<td>?</td>
</tr>
<tr>
<td>3</td>
<td>High school</td>
<td>1</td>
<td>?</td>
<td>1</td>
<td>?</td>
<td>?</td>
</tr>
<tr>
<td>4</td>
<td>College</td>
<td>0</td>
<td>1</td>
<td>?</td>
<td>?</td>
<td>?</td>
</tr>
<tr>
<td>5</td>
<td>College</td>
<td>1</td>
<td>?</td>
<td>1</td>
<td>?</td>
<td>?</td>
</tr>
<tr>
<td>6</td>
<td>College</td>
<td>1</td>
<td>?</td>
<td>1</td>
<td>?</td>
<td>?</td>
</tr>
</tbody>
</table>
What works: Nonparametric standardization

\[
\hat{P}(X = x) = \sum_{x \in \text{Support of } X} P(X = x) \left( \bar{Y}_{i; W_i = 1, X_i = x} - \bar{Y}_{i; W_i = 0, X_i = x} \right)
\]

\[
= P(X_i = \text{High school}) \left( \bar{Y}_{i; W_i = 1, X_i = \text{High school}} - \bar{Y}_{i; W_i = 0, X_i = \text{High school}} \right)
\]

\[
+ P(X_i = \text{College}) \left( \bar{Y}_{i; W_i = 1, X_i = \text{College}} - \bar{Y}_{i; W_i = 0, X_i = \text{College}} \right)
\]

\[
= \frac{1}{2}(1 - 0) + \frac{1}{2}(1 - 1) = 0.5 + 0 =
\]

<table>
<thead>
<tr>
<th>Education</th>
<th>Treated</th>
<th>No job training</th>
<th>Job training</th>
<th>Treatment effect</th>
<th>(\tau_i = Y_i(1) - Y_i(0))</th>
</tr>
</thead>
<tbody>
<tr>
<td>ID</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>High school</td>
<td>0</td>
<td>0</td>
<td>?</td>
<td>?</td>
</tr>
<tr>
<td>2</td>
<td>High school</td>
<td>0</td>
<td>0</td>
<td>?</td>
<td>?</td>
</tr>
<tr>
<td>3</td>
<td>High school</td>
<td>1</td>
<td>?</td>
<td>1</td>
<td>?</td>
</tr>
<tr>
<td>4</td>
<td>College</td>
<td>0</td>
<td>1</td>
<td>?</td>
<td>?</td>
</tr>
<tr>
<td>5</td>
<td>College</td>
<td>1</td>
<td>?</td>
<td>1</td>
<td>?</td>
</tr>
<tr>
<td>6</td>
<td>College</td>
<td>1</td>
<td>?</td>
<td>1</td>
<td>?</td>
</tr>
</tbody>
</table>

What works: Nonparametric standardization

- But when there are many cells of the covariates X, nonparametric standardization is impossible!
Why is tree-based standardization biased? Regularization

With no regularization, a tree would grow until each leaf was completely homogenous in $X_i$.

But this tree would be very noisy! We prune our trees so that leaves contain more observations.

- Treatment effects are more **precisely estimated**
- But treatment effects are **biased** if there is confounding within leaves

Is there anything we can do? Chernozhukov et al.

Outcome equation

$$Y = D\theta_0 + g_0(X) + U$$

Treatment assignment

$$D = m_0(X) + V$$

One might be tempted to estimate $\hat{g}_0(X)$ by machine learning and then state:

$$\hat{\theta}_0 = \frac{1}{n} \sum_{i \in I} D_i (Y_i - \hat{g}_0(X_i))$$

$$\frac{1}{n} \sum_{i \in I} D_i^2$$
Is there anything we can do? Chernozhukov et al.

Outcome equation
\[ Y = D\theta_0 + g_0(X) + U \]

Treatment assignment
\[ D = m_0(X) + V \]

One might be tempted to estimate \( \hat{g}_0(X) \) by machine learning and then state:
\[
\hat{\theta}_0 = \frac{1}{n} \sum_{i \in I} D_i (Y_i - \hat{g}_0(X_i)) \\
\frac{1}{n} \sum_{i \in I} D_i^2
\]

This will be biased because the estimator \( \hat{g}_0 \) is regularized.

Does not have mean 0
\[
b = \frac{1}{E(D_i^2)} \frac{1}{\sqrt{n}} \sum_{i \in I} \left( m_0(X_i)(g_0(X_i) - \hat{g}_0(X_i)) \right) + o_p(1)
\]

Key: \( D_i \) is centered at \( m_0(X) \neq 0 \). We should recenter \( D_i \).
Is there anything we can do? Chernozhukov et al.

\[ Y = D\theta_0 + \tilde{g}_0(X) + U \]
\[ D = m_0(X) + V \]

1. Split the sample into \( I \) and \( J \)
2. Estimate \( \tilde{g}_0(X) \) using sample \( J \)
3. Estimate \( \tilde{m}_0(X) \) using sample \( J \)
4. Orthogonalize \( D \) on \( X \) (approximately)
   \[ \tilde{V} = D - \tilde{m}_0(X) \]
5. Estimate the treatment effect
   - Biased
     \[ \hat{\theta}_0 = \frac{1}{n} \sum_{i \in I} D_i (Y_i - \tilde{g}_0(X_i)) \]
   - De-biased
     \[ \hat{\theta}_0 = \frac{1}{n} \sum_{i \in I} \tilde{V}_i (Y_i - \tilde{g}_0(X_i)) \]

Bias remaining in de-biased estimator (Chernozhukov et al.)

\[ \sqrt{n}(\hat{\theta}_0 - \theta_0) = a^* + b^* + c^* \]

Regularization bias:

\[ b^* = \frac{1}{E(V^2)} \frac{1}{\sqrt{n}} \sum_{i \in I} \left( \tilde{m}_0(X_i) - m_0(X_i) \right) \left( \tilde{g}_0(X_i) - g_0(X_i) \right) \]

Vanishes “under a broad range of data-generating processes.”

Bounded above by

\[ \sqrt{n n^{-\psi_m} n^{-\psi_E}} \]

Rate of convergence of \( \tilde{m}_0 \rightarrow m \)
Rate of convergence of \( \tilde{g}_0 \rightarrow g \)
Bias remaining in de-biased estimator (Chernozhukov et al.)

$$\sqrt{n}(\hat{\theta}_0 - \theta_0) = a^* + b^* + c^*$$

An example of the third term in the partially linear model:

$$c^* = \frac{1}{\sqrt{n}} \sum_{i \in I} V_i \left( \hat{g}_0(X_i) - g_0(X_i) \right)$$

If $$\hat{g}_0$$ is estimated on an auxiliary sample $$J$$, then $$V_i$$ and $$\hat{g}_0(X_i)$$ will be uncorrelated and $$\mathbb{E}(c^*) = 0$$.

BART: Bayesian Additive Regression Trees

Differs from random forests:
- Fixed number of trees
- Backfits repeatedly over the fixed number of trees
- Strong prior encourages shallow trees
- Uncertainty comes automatically from posterior samples
Tree-based Methods

- Bayesian Additive Regression Trees (BART)
  - A Bayesian “sum-of-trees” model
  - Nonparametric Bayesian regression model

Tree-based Methods

- **Classification And Regression Trees (CART)**
  - Recursively partition the data space
  - Fit a simple prediction model for each partition
  - Represent every partitioning as a decision tree

- Leaf specific effect:
  
  \[
  \mu(w, x | \Pi) \equiv E \left[ Y_i(w) \mid X_i \in \ell(x | \Pi) \right] 
  \]

  \[
  \tau(x | \Pi) \equiv \mu(1, x | \Pi) - \mu(0, x | \Pi) 
  \]

---

**BART model**

\[
Y = \sum_{j=1}^{m} g_j(x | T_j, M_j) + \epsilon 
\]

\[
 \epsilon \sim N(0, \sigma^2) 
\]

\[
T_j \text{ prior} 
\]

\[
P( T_j \neq d ) = \alpha (1 + d)^{-\beta} 
\]

Tree-Split

Split variable \( \sim \text{Uniform(Available variables)} \)

Split value \( \sim \text{Uniform(Available split values)} \)

\[
\mu_{ij} | T_j \sim N \left( \frac{\mu_i \sigma^2_j}{\alpha_i} \right) 
\]

\[
\sigma \sim \frac{\nu_{ij}}{\nu_{ij}^2} \text{ (Inverse chi-square)} 
\]

They recommend \( \{ \alpha = .95, \beta = 2 \} \rightarrow 97\% \) of prior probability is on 4 or fewer terminal nodes.
BART for causal inference

Goal: Model the response surface as a function of treatment and pre-treatment covariates

1. Fit a flexible model for $Y = f(X, W)$
2. Set $W = 0$ to predict $\hat{Y}_i(0)$ for all $i$
3. Set $W = 1$ to predict $\hat{Y}_i(1)$ for all $i$
4. Difference to estimate $\hat{r}_i$
5. Plot effects

Tree-based Methods

• Advantages of BART:
  • Easy to implement
  • Posterior can provide uncertainty of the estimation
  • Can deal with a large number of covariates
  • Handle continuous treatment variables and missing data
BART: Benefits and drawbacks

Benefits
- Less researcher discretion for tuning parameters
- Automatic posterior uncertainty estimates

Drawbacks
- Not guaranteed to address confounding due to regularization
- No theoretical guarantees of centering over truth
- Splitting is based on prediction and is not explicitly optimized for causal inference within leaves

Summary
- Causal trees can detect high-dimensional covariate-based treatment effect heterogeneity
- Work well with high-order interactions
- Causal forests give theoretically valid confidence intervals
- Bayesian approaches (BART) are less theoretically verified but give easy uncertainty
- With high-dimensional confounding, all methods are biased but can be designed to be consistent.
Subspace Learning for Causal Effect Estimation

• Matching in the original feature space is unreliable when the feature space is high dimensional
• What if we perform the matching in lower dimensional subspaces?
• How do we identify subspaces?
  • Linear subspaces e.g., PCA
  • Random Subspaces
  • Informative Subspace
  • Balanced and Nonlinear Subspace

Linear Subspace Learning for Causal Estimation

• Examples:
  • Principal component analysis (PCA)
  • Locality preserving projections (LPP)
  • Canonical correlation analysis (CCA)
  and their variants
Nearest Neighbor Matching (NNM)

- For a treated unit $i$, nearest neighbor matching (NNM) finds its nearest neighbor in control group in terms of covariates.
- NNM usually uses standard distance measures such as Euclidean distance, Mahalanobis distance, etc.
- NNM is unreliable when the number of covariates is large.
- The bias of NNM based estimator increases with the number of covariates $d$ for a given number of samples $N$ as $O(N^{-1/d})$


Nearest neighbor matching with Random Subspaces

- Map data to a lower dimensional subspace via linear projection while preserving the distances between data samples in the original feature space
- Johnson-Lindenstrauss (JL) Lemma

**Johnson-Lindenstrauss (JL) Lemma.** For any $0 < \epsilon < 1/2$ and $x_1, \ldots, x_N \in \mathbb{R}^d$, there exists a mapping $f : \mathbb{R}^d \rightarrow \mathbb{R}^k$, with $k = O(\epsilon^{-2} \log N)$, such that

$$\forall i, j \ (1-\epsilon) \|x_i - x_j\|^2 \leq \|f(x_i) - f(x_j)\|^2 \leq (1+\epsilon) \|x_i - x_j\|^2.$$  

- Unclear whether preserving distances in the original feature space is a good idea
- Why?

**NNM with Random Subspaces**

Randomized Nearest Neighbor Matching (RNNM)

- Choose $K$ via JL Lemma
- Random Projection

**Informative Subspace Learning**

- Hilbert-Schmidt Independence Criterion (HSIC) based NNM
- HSIC-NNM learns two linear projections for predicting the outcomes under treatment and control
- Maximizes nonlinear dependency between the projected subspace and the outcome by

$$M_w = \max_{M_w} \text{HSIC}(X_m^w, M_w, Y^F) - R(M_w)$$

where $X_m^w$ is the transformed subspace, $Y^F$ is the observed control/treated outcome, and $R$ denotes the regularization term

Nonlinear and Balanced Subspace Learning

• Challenges
  • Estimation bias increases with the number of covariates
  • Distribution of covariates can be complex

• Approach
  • Convert counterfactual prediction to a multi-class classification problem with pseudo labels
  • Ordinal scatter discrepancy criterion to extract nonlinear representations
  • Maximum mean discrepancy criterion to learn balanced representations


Nonlinear and Balanced Subspace Learning

• Objective Function

\[ \arg \max_P F(P, \Phi(X), Y) - \beta \text{Dist}(\Psi(X_C), \Psi(X_T)) \]
\[ = \text{tr}(P^T(K_I - \alpha K_W)P) - \beta \text{tr}(P^T K L K P) \]

s.t. \[ P^T P = I, \]

• P is the learned nonlinear projection
• Closed form solution
Matching on Balanced Representation

- Evaluate distance between samples in a kernel induced feature space
  - Allows nonlinear mappings to feature space
- Ensure balanced representation
  - Treated and control group have similar distributions thus ensuring exchangeability

\[
\arg \max_P \left\{ \frac{\text{Dist}(\Psi(X_0), \Psi(X_1))}{\text{trace}(K) - \alpha \text{trace}(K_0)} \right\}
\]

Idea: similar units should have similar outcomes
  - \( K_i \): inter-class scatter
  - \( K_0 \): within-class scatter

"class": based on outcome

Idea: achieve balanced distribution

\[
\text{Is}\text{ }\text{dist}(\Phi(X_0), \Phi(X_1)) = \frac{1}{n_0} \sum_{i=1}^{n_0} \| x_i - \bar{x}_0 \|_2 + \frac{1}{n_1} \sum_{i=1}^{n_1} \| x_i - \bar{x}_1 \|_2
\]

- Perform matching by computing distance in the feature space.


---

Summary: Subspace methods for causal estimation

- **Subspace methods for Causal Estimation**
  - (+) Most methods have closed-form solutions and are computationally efficient
  - (-) Subspace learning methods often make strong distributional assumptions
  - (-) Rely on Matching estimators, and are only as good as the matching measures used