Causal-trees, -forests and heterogenous treatment effects

Goal

- Systematically identify subpopulations with heterogenous treatment effects and estimate treatment effects within these under valid inference.

Use: Optimal policies

- Customized to subpopulations or even individuals.

Readings:

- Athey, Susan, and Guido Imbens. "Recursive partitioning for heterogeneous causal effects." Proceedings of the National Academy of Sciences 113.27 (2016): 7353-7360.
- Wager, Stefan, and Susan Athey. "Estimation and inference of heterogeneous treatment effects using random forests." Journal of the American Statistical Association 113.523 (2018): 1228-1242.
- Athey, Susan, and Stefan Wager. "Policy learning with observational data." Econometrica 89.1 (2021): 133-161.
- https://d2cml-ai.github.io/mgtecon634_r/intro.html

Causal trees

Differences to standard prediction trees

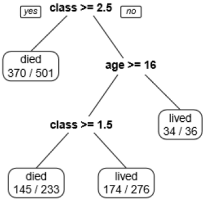
1. Splitting criterion penalizes variance (small leaf size).
   - Can be applied to standard prediction trees.
2. Estimates treatment effect rather than mean.
   - Diff in average outcomes for treated and control group within leaf.

Inference

- Holding tree from training sample fixed, we can use standard methods to conduct inference within each leaf of the tree on test sample.
- Requires no DGP sparsity assumptions, does not deteriorate as number of covariates increase

Prediction tree

Who survived the Titanic?



- Prediction: mean y in leaf
- Split criterion: *in-sample* min MSE
- Pruning to avoid over-fit.

Split criterion issues

The standard, *in-sample*, splitting criterion,

$$MSE\left(S^{tr}\right) = \frac{1}{N}\sum_{i=1}^{N}\left(\overline{Y}_{l_i} - Y_i\right)^2,$$

may lead to splits with poor *out-of-sample* prediction performance, e.g. mall number of outlier observations on one side of the split.

- Alternative: split on t-stats for test of equal means?

Honest split criterion

- Let S^{tr}, S^{est} and S^{te}be three independent samples.
- The goal is to asses algorithms $\pi()$ that minimizes the expected out-of-sample MSE.
    - S^{tr} is used to construct the tree.
    - S^{est} to compute the mean within each leaf and
    - S^{te} to evaluate the MSE given the tree and leaf means.
- Expected MSE removes S^{est} by providing analytical formula (as e.g. AIC).

Honest split criterion estimate

Minimize expected MSE in an independent sample, S^{est}, (standard tree algorithm min MSE in sample, S^{tr}).

$$-\widehat{EMSE}_{\mu}\left(S^{tr}, N^{est}, \Pi\right) = \frac{1}{N}\sum_{i=1}^{N}\overline{Y}_{l_i}^{2} - \left(\frac{1}{N^{tr}} + \frac{1}{N^{est}}\right)\sum_{l\in\Pi} S_{S^{tr}}^{2}\left(l\right),$$

where $S_{S^{tr}}^{2}\left(l\right)$ is the within leaf variance of the mean.

- Smaller leafs have larger variance in in-sample leaf means
- Honest criterion penalizes expected variance of the mean estimates (small leaf size), as would splitting on t-stat for difference test.

2. Conditional treatment in trees

Consider some tree Π, that assigns each individual i to a leaf l_i.The mean outcome in each leaf is

$$\mu\left(w,l\right) = E\left[Y_i \mid W_i = w, l_i = l\right]$$

and the mean treatment effect is

$$\tau(l) = \mu\left(1,l\right) - \mu\left(0,l\right).$$

Split criterion

Let an empirical estimate of $\tau(l)$ be $\overline{\tau}_l$. The standard tree split criterion minimizes the MSE

$$\frac{1}{N}\sum_{i=1}^{N}(\overline{\tau}_l - \tau_i)^2$$

But this is infeasible since τ_i is not observed!
Solutions:

1. Transform the outcome to something that captures τ_i on average.
2. Compute expected MSE (EMSE), similar to CP, AIC, etc.

Transform the outcome

Suppose treatment with probability p.Let

$$Y_i^* = \frac{W_i - p}{p\left(1-p\right)} Y_i = \begin{array}{ll} \frac{1}{p} Y_i & \text{if } W_i = 1 \\ -\frac{1}{1-p} Y_i & \text{if } W_i = 0. \end{array}$$

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

Y_i^* is a very noisy measure of τ_i, but it is right on average.
Estimate $\widehat{p}\left(x\right)$ using standard methods. Criterion for evaluating split, in-sample MSE,

$$\frac{1}{N} \sum_{i=1}^{N} \left(\overline{Y}_l^* - Y_i^*\right)^2 + \lambda\left(\#leaves\right).$$

Transformed Outcome Trees (TOT) loss function

- Within leaf, $\overline{Y}_l^*$ is not most efficient estimator of treatment effect.
- Better estimate, sample leave-one-out average treatment effect within leaf,

$$\widehat{\tau}^{CT}_{(-i)} = \overline{Y}_l^{W=1} - \overline{Y}_l^{W=0}.$$

- Criterion for evaluating split, in-sample MSE,

$$\frac{1}{N}\sum_{i=1}^{N}\left(\widehat{\tau}^{CT}_{(-i)} - Y_i^*\right)^2 + \lambda\left(\#leaves\right).$$

- Select λ via cross-validation, $\widehat{\tau}^{CT}_{(-i)}$ computed on fold not containing i.

Honest splitting criterion for causal effects

Honest criteria

$$-\widehat{EMSE}_{\tau}\left(S^{tr}, N^{est}, \Pi\right) = \frac{1}{N^{tr}} \sum_{i \in S^{tr}} \widehat{\tau}^2\left(X_i; S^{tr}, \Pi\right)$$

$$-\left(\frac{1}{N^{tr}} + \frac{1}{N^{est}}\right) \sum_{l \in \Pi} \left(\frac{S^2_{S^{tr}_{treat}}}{p} - \frac{S^2_{S^{tr}_{control}}}{1-p}\right).$$

- $\widehat{\tau}$ is within-leaf differences in Y between treated and control.
- Penalizes variance in the leaf estimates, for example, leaves that are unbalanced in treatment and control.

Cross validation

$$-\widehat{EMSE}_{\tau}\left(S^{tr,cv}, N^{est}, \Pi\right) + \lambda\left(\#leaves\right)$$

is the same objective function as above, but evaluated in the cross-validation sample, $S^{tr,cv}$.

Inference

Independent samples for

- tree construction, S^{tr},
- estimation of within leaf means, S^{est},
- inference (testing) S^{te}

Holding tree from training sample fixed, use standard methods to conduct inference within each leaf of the tree on test sample.

- Use any valid method for treatment effect estimation, not just method used in training. Asymptotic theory as usual within a leaf.
- Once you have the partition, just run a regression on second sample interacting leaf dummies with treatment indicator. Everything is as usual.

No assumptions about sparsity of true data-generating process.

Causal forests

- For observational data, we need to "orthogonalize" the data prior to building trees:

$$\begin{aligned} \text{residual outcome:} \qquad & Y^* = Y - m(x) \\ \text{residual treatment:} \qquad & W^* = W - e(x). \end{aligned}$$

1. Use any method to estimate propensity score of treatment and mean of Y

$$\begin{aligned} e\left(x\right) &= P\left[W = 1 \mid X = x\right] \\ m\left(x\right) &= E\left[Y \mid X = x\right]. \end{aligned}$$

2. Solve

$$\widehat{\tau}\left(\cdot\right) = \arg\min_{\tau} \sum_{i=1}^{n} \left(\left(Y_i - \widehat{m}^{(-i)}\right) - \tau\left(X_i\right)\left(W_i - \widehat{e}^{(-i)}\right)\right)^2 + \Lambda_n\left(\tau\left(\cdot\right)\right),$$

where Λ_n is some regularizer and $\widehat{m}^{(-i)}$ and $\widehat{e}^{(-i)}$ are based on trees built on samples not including observation i.

grf

1. Fits two separate regression forests to estimate $\widehat{e}\left(x\right)$ and $\widehat{m}\left(x\right)$, and makes out of bag predictions to obtain $\widehat{e}^{(-i)}$ and $\widehat{m}^{(-i)}$ as average outputs of trees that did not include observation i.
2. Grows a causal forest via

$$\widehat{\tau}\left(x\right) = \frac{\sum_{i=1}^{n} \alpha_i\left(x\right)\left(Y_i - \widehat{m}^{(-i)}(X_i)\right)\left(W_i - \widehat{e}^{(-i)}(X_i)\right)}{\sum_{i=1}^{n} \alpha_i\left(x\right)\left(W_i - \widehat{e}^{(-i)}(X_i)\right)^2} + \Lambda_n\left(\tau\left(\cdot\right)\right),$$

where Λ_n is some regularizer and $\widehat{m}^{(-i)}$ and $\widehat{e}^{(-i)}$ are based on samples not including observation i, where $\alpha_i\left(x\right)$ is the share of trees in the forest where x falls in the same leaf as X_i.

https://d2cml-ai.github.io/mgtecon634_r/intro.html
https://github.com/grf-labs/grf/blob/master/REFERENCE.md

Caveat

Chernozhukov et al. (2023), the method by Wagner & Athey (2017) does not provide reliable estimates in high-dimensional settings where the number of covariates is much larger than the log of the number of observations.

Who (Mis)uses the Sickness Insurance System? Evidence from a Randomised Experiment

Yakymovych (2022)

Question

- How much predictable heterogeneity in duration response to monitoring, and who responds more?

Method

- Randomized experiment for sickness insurance recipients
- Causal forest

Results

- How much predictable heterogeneity?
  - Least affected decile $<$0.36 days, most affected decile $>$1.7 days
  - Targeted monitoring 40% more efficient than random.
- Who responds more?
  - High sick-leave uptake in pre-period
  - From disadvantaged neighborhoods
  - Low education and income, male, working at large plants with high sick leave-uptake.

Prediction guiding policy

- Heterogenous effects for policy targeting is a prediction problem (not causation).
- Program aimed at monitoring sickness insurance to reduce misuse.
    - We don't need to know whether education/gender causes the increased response to monitoring, just that it is a predictor of a strong response.

The experiment

Randomized experiment for sickness insurance recipients July-Dec 1988

- Sample: 70,000 in Jämtland 240,000 in Gothenburg.
    - After restrictions: 123,429 spells by 77,672 workers
- Treatment: days of absence spell before doctor's certificate required
    - 7 for odd birth dates (control)
    - 14 for even dates (treated)

Outcomes

- Duration of sickness spell in days
- Sickness spell 8-14 days

Worker Characteristics

- Health, demographic, family, education, neighborhood, career, workplace, etc

- Tests for heterogeneity across large number of covariates
- Strength
  - Avoids over-fitting when testing multiple hypotheses.
  - Selects model in data-driven way
  - Allows for interactions and non-linearities.

Method: Generalized Random Forest

How strong is the measured heterogeneity
Out-of-bag CATE

FIGURE 4. DISTRIBUTION OF PREDICTED TREATMENT EFFECTS ON SICKNESS ABSENCE SPELL DURATION.

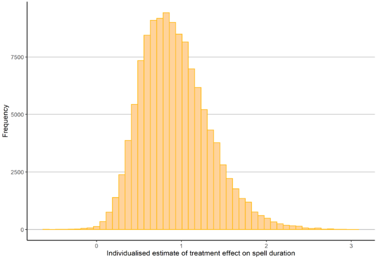
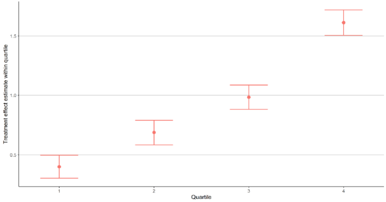


FIGURE 5. ESTIMATED TREATMENT EFFECTS FOR WORKERS WITHIN EACH QUARTILE OF PREDICTED CAUSAL FOREST $\hat{\tau}_x$ ESTIMATES

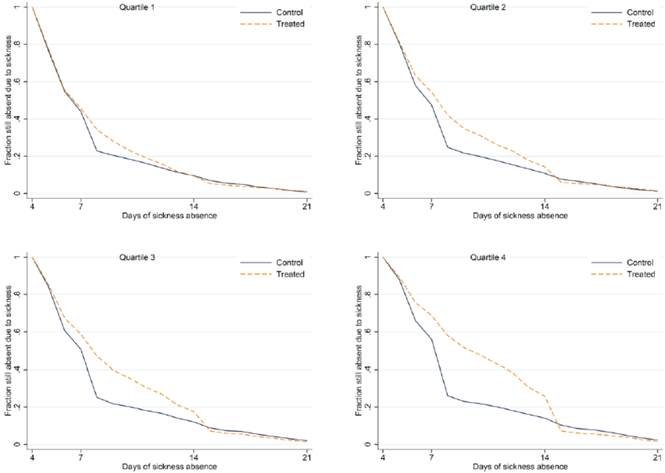


Note: Quartiles ranked according to causal forest estimated treatment effects, with Q1 containing those estimated to be least affected and Q4 those estimated to be most affected. Treatment effects within each of the quartiles estimated as $\hat{\tau} = \bar{y}_i|(W_w = 1) - \bar{y}_i|(W_w = 0)$. Confidence intervals at the 95 percent level shown.

How strong is the heterogeneity

Hold-out (test) set

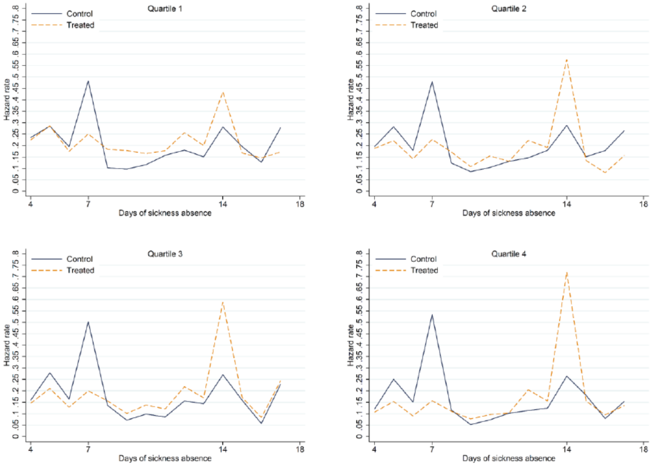
FIGURE 6. SURVIVAL GRAPHS FOR ABSENCE SPELLS AMONG WORKERS IN THE HELD-OUT TEST SET, RANKED BY QUARTILES OF PREDICTED TREATMENT EFFECTS



How strong is the heterogeneity

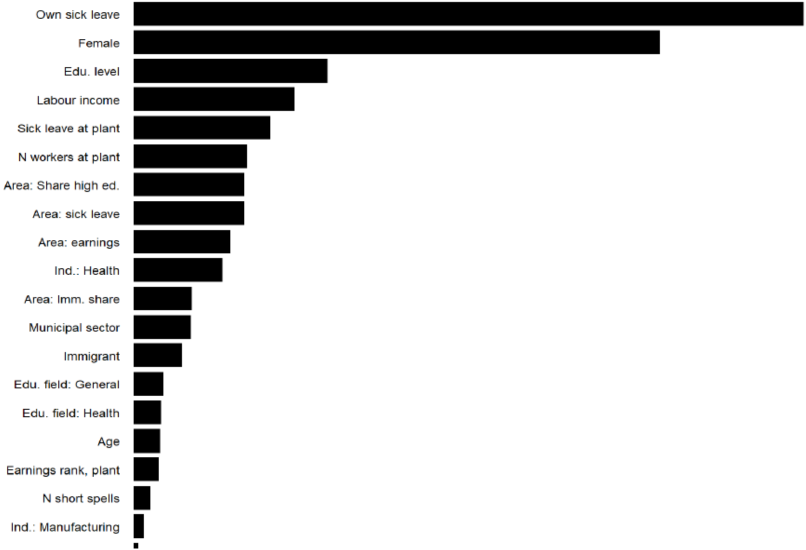
Hold-out (test) set

FIGURE 7. HAZARD GRAPHS FOR ABSENCE SPELLS AMONG WORKERS IN THE HELD-OUT TEST SET, RANKED BY QUARTILES OF PREDICTED TREATMENT EFFECTS.



What characteristics help predict heterogeneity?

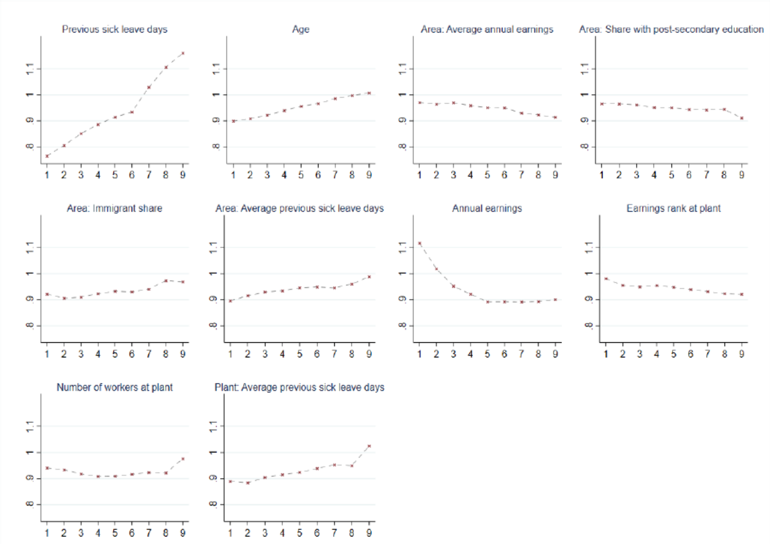
FIGURE 8. IMPORTANCE OF WORKER CHARACTERISTICS FOR HETEROGENEITY, BASED ON THE NUMBER OF TIMES THE CAUSAL FOREST'S TREES SPLIT ON THE CHARACTERISTIC



What characteristics help predict heterogeneity?

Partial response: One variable is at a fixed value, the rest are at their empirical values. What is the average treatment effect for this value?

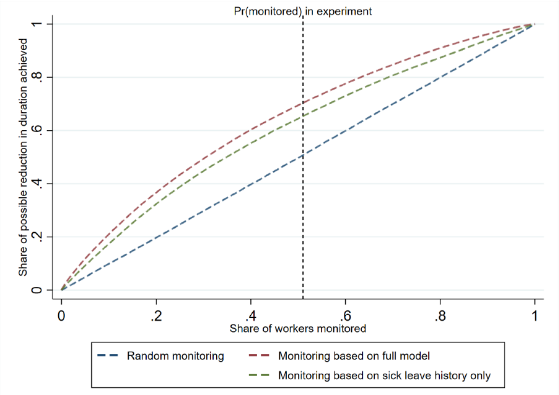
FIGURE A3. PARTIAL DEPENDENCE PLOTS FOR THE CONTINUOUS COVARIATES.



Gains from targeting

Target highest $\widehat{\tau}_i$ or longest sick-leave history.

Figure 11. Effect in terms of reducible sickness absence duration for given share of workers monitored according to different monitoring policies



Problem set: Voter preferences for female candidates

We now use data from an experiment among likely Democratic presidential primary voters fielded by YouGov. The experiment studied which attributes likely Democratic primary voters valued in prospective candidates for president in 2020. Respondents were shown profiles for two hypothetical candidates, Candidate A and Candidate B, who were each given a randomly generated race, gender, age, climate plan, health care plan, general election strategy, and background with respect to their relationship to the Washington "establishment."

The data we use are the first pair of candidates the respondent chose between in matchups that featured a man versus a woman and without missing covariates. The key dependent variable is picked_cand_a, indicating that the responded picked candidate A. The key independent variable is cand_a_female, indicating that candidate A was female.

The data and code book are uploaded on Athena. The data is analyzed in the paper "Strategic Discrimination in the 2020 Democratic Primary.", Green, Schaffner, and Luks, Public Opinion Quarterly (2022).

1. Run an OLS regression of picked_cand_a on cand_a_female. What is the average preference for or against female candidates? Run regressions of picked_cand_a on cand_a_female and one of each of the other included regressions, main effects and interactions. What is the most significant heterogeneity? Why can't we use the standard errors from these estimates for analyzing hetergeneity?

2. Estimate a causal forest model using the grf package. The tutorial at

https://d2cml-ai.github.io/mgtecon634_r/intro.html

shows all the steps for a deeper analysis. We will only do a simple analysis.

The grf::causal_forest command needs as inputs: Y (dependent variable), W (treatment indicator) and X (covariates). In our case, Y is picked_cand_a, W is cand_a_female, and X contains all the other variables command to convert data frame t matrix).

```
# train causal forest, syntax
cf <- grf::causal_forest( Y = train_Y,  X = train_X,  W = train_W,  num.trees = 5000,  seed = 10101 )
```

Before running the above command you need to: (1) create numeric vectors Y and W from picked_cand_a and cand_a_female; (2) create dummy variables for each factor level of each factor variable, see one_hot command.

```
X <- mltools::one_hot(
 data.table::data.table(
 df %>%
 dplyr::select(-cand_a_female, -picked_cand_a) ) )
```

Randomly select 75% of data to training and 25% to test. Train the causal forest on the training data, using 5000 trees and setting the random seed.

3. Predict the heterogenous treatment effect in the test data. Tips, use the predict command as described in the tutorial:

```
test_pred <- predict(cf, newdata=as.matrix(test_X), ), estimate.variance=TRUE)

test_X$preds <- test_pred$predictions
```

Plot the predicted treatment effects. Does the predicted treatment effects predict treatment heterogeneity in the test data? Regress picked_cand_a on cand_a_female interacted with the predicted treatment effect (and including main effects). Is the interaction significant?

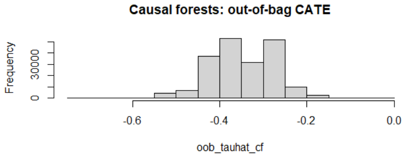
4. What are the ten most important variables for hetergeneity? You can find this using the variable_importance(cf) command in grf and using default settings.

Plot the predicted treatment effect as a function of two most important variables. If it is a count variable (like age), then plot the average treatment effect for each age. If it is a dichotomous variable (like "Favor #Metoo"), then plot two density plots (or histograms) one for "Favor #Metoo"=0 and another for "Favor #Metoo"=1.

Additional material: ML for policy

- Heterogenous effects, $\tau\left(X_i\right)$, for policy targeting is a prediction problem (not causation).
- Consider a program aimed at reducing unemployment.
    - Suppose people with low education respond more to program.
    - We don't need to know whether education causes the increased response, just that it is a predictor of a strong response.

Finding best policy



- Policy $\pi : X_i \rightarrow W_i$ with cost C.
- Value

$$V(\pi) = E\left[\pi(X_i)(\tau(X_i) - C)\right]$$

- Maximize $V(\pi)$ by treating if $\tau(X) > C$.

Random policy benchmark

$$A(\pi) = 2\left(V(\pi) - V(\pi_{random})\right)$$

Let

$$\begin{aligned} \tau^1 &= E[\tau \mid \pi_i = 1] & p_\pi^1 &= P(\pi_i = 1) \\ \tau^0 &= E[\tau \mid \pi_i = 0] & p_\pi^0 &= 1 - P(\pi_i = 1) \\ \tau^e &= E[\tau] = p_\pi^1 \tau^1 + p_\pi^0 \tau^0 & & \end{aligned}$$

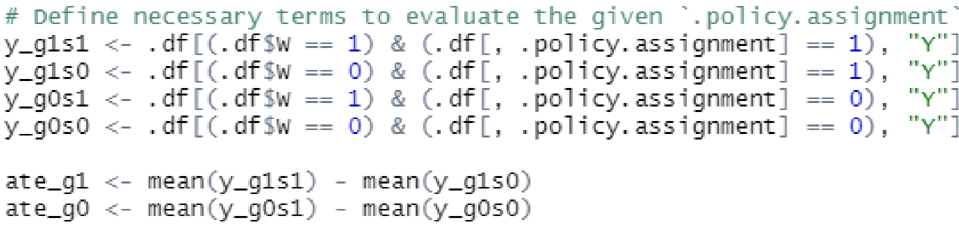
$$\begin{aligned} A(\pi) &= 2\left(V(\pi) - V(\pi_{random})\right) \\ &= 2(p_\pi^1(\tau^1 - C) - \frac{1}{2}(\tau^e - C)) \\ &= \left(2p_\pi^1\tau^1 - \tau^e\right) - \left(2p_\pi^1 - 1\right) C) \\ &= 2p_\pi^1\tau^1 - \left(p_\pi^1\tau^1 + p_\pi^0\tau^0\right) - \left(2p_\pi^1 - 1\right) C) \\ &= p_\pi^1\tau^1 - p_\pi^0\tau^0 - (2p_\pi^1 - 1)c \end{aligned}$$

Evaluating plugin policy

- Plugin policy: treat if $\widehat{\tau}^{(-i)} > c$.

$$\widehat{A}\left(\pi\right) = \widehat{p}^1_\pi \widehat{\tau}^1 - \widehat{p}^0_\pi \widehat{\tau}^0 - \left(2\widehat{p}^1_\pi - 1\right) C.$$

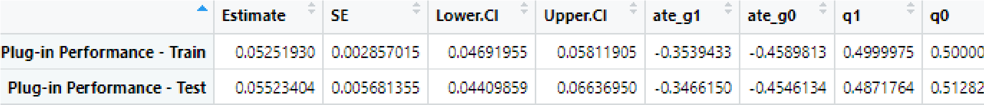
- $\widehat{p}^1_\pi$: share assigned to treatment
- $\widehat{\tau}^1$ average estimated treatment for those assigned to treatment
  - Difference in outcome means among treated and non-treatment in this group in experiment.



```
# Define necessary terms to evaluate the given `.policy.assignment`
y_g1s1 <- .df[(.df$W == 1) & (.df[, .policy.assignment] == 1), "Y"]
y_g1s0 <- .df[(.df$W == 0) & (.df[, .policy.assignment] == 1), "Y"]
y_g0s1 <- .df[(.df$W == 1) & (.df[, .policy.assignment] == 0), "Y"]
y_g0s0 <- .df[(.df$W == 0) & (.df[, .policy.assignment] == 0), "Y"]

ate_g1 <- mean(y_g1s1) - mean(y_g1s0)
ate_g0 <- mean(y_g0s1) - mean(y_g0s0)
```

Evaluating plugin policy



	Estimate	SE	Lower.CI	Upper.CI	ate_g1	ate_g0	q1	q0
Plug-in Performance - Train	0.05251930	0.002857015	0.04691955	0.05811905	-0.3539433	-0.4589813	0.4999975	0.50000
Plug-in Performance - Test	0.05523404	0.005681355	0.04409859	0.06636950	-0.3466150	-0.4546134	0.4871764	0.51282

The difference in average effect between those assigned to treatment and control is .1 (.45-.35) and the share treated is around .5 Hence the value is around .05.

Evaluating non-randomized policy

$$
\begin{aligned}
A(\pi) &= 2\left(V(\pi) - V(\pi_{random})\right) \\
&= 2E\left[\pi_i \tau_i\right] - E\left[\tau_i\right] - \left(2p^1 - 1\right)C.
\end{aligned}
$$

$$
\widehat{A}(\pi) = \frac{1}{n}\sum_i \left(2\pi_i - 1\right)\widehat{\Gamma}_i - \left(2\widehat{p}^1 - 1\right)C,
$$

In this case, the unconditional difference in means among treated and untreated is not a consistent measure of the treatment effect. Use

$$
\begin{aligned}
\widehat{\Gamma}_i &= \widehat{\tau}^{(-i)} + \widehat{Y}_i^*, \\
\widehat{Y}_i^* &= \frac{W_i - \widehat{e}^{(-i)}}{\widehat{e}^{(-i)}\left(1 - \widehat{e}^{(-i)}\right)}\left(Y_i - \widehat{m}^{(-i)}\right).
\end{aligned}
$$

Estimating the conditional average treatment effect function $\tau(X)$ and learning a good policy $\pi(X_i)$ are different problems.

- The correct loss function for policy learning is not mean-squared error on $\tau(X)$.
- The $\tau(X)$ function may change with variables we cannot use for targeting (e.g., variables only measured after the fact).
- We may wish to impose other constraints on policy functions.

Policy learning

- Maximize expected value $E\left[\tau\left(X\right)\pi\left(X_i\right)\right]$ for π in some class of allowed policies Π.
- Let $\widehat{\Gamma}_i$ be an estimate of the individual-level treatment effect (similar to the transformed outcome we discussed above)

$$\widehat{\Gamma}_i = \widehat{\tau}^{(-i)} : \text{ for experiments}$$

$$\widehat{\Gamma}_i = \widehat{\tau}^{(-i)} + \frac{W_i - \widehat{e}^{(-i)}}{\widehat{e}^{(-i)}\left(1 - \widehat{e}^{(-i)}\right)}\left(Y_i - \widehat{\mu}_{W_i}^{(-i)}\right) : \text{ observational data.}$$

- Find $\pi \in \Pi$ that maximizes

$$\frac{1}{n}\sum_{i=1}^{n}\left(2\pi\left(x\right)-1\right)\left(\widehat{\Gamma}_i - C\right)$$

- Rephrase as classification problem

$$\widehat{\Gamma}_i = \left|\widehat{\Gamma}_i - C\right| sign\left(\widehat{\Gamma}_i - C\right),$$

to classify positive effects, weighted by absolute magnitude of effect. Can be solved by standard methods.