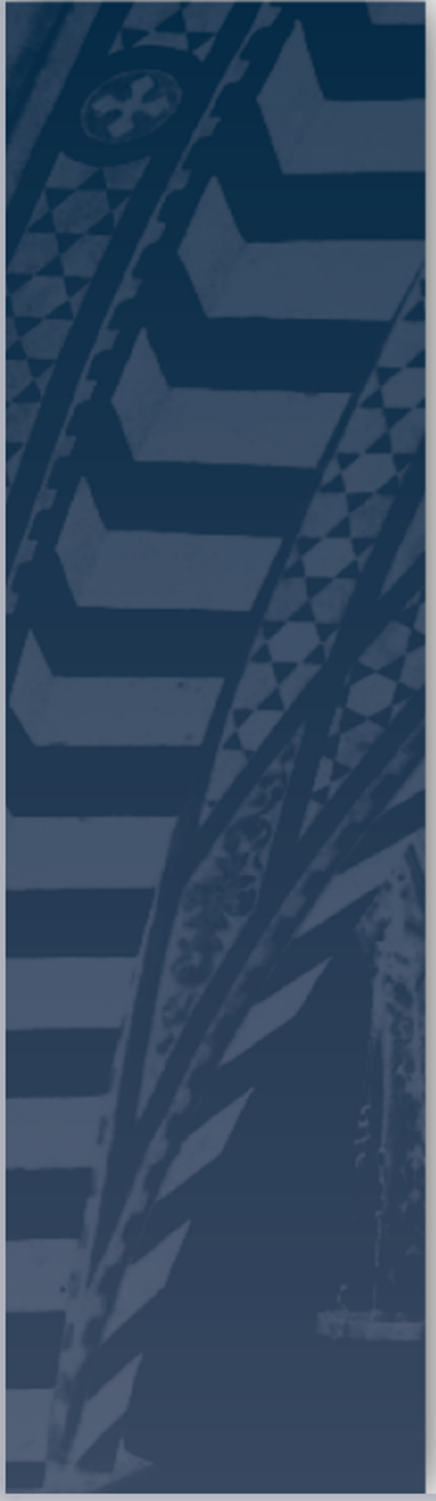


IMT SCHOOL FOR ADVANCED STUDIES LUCCA

ML Group Presentation

S.Mullainathan, J.Speiss (2019); S.Wager, S.Athey(2018)

Federico Nutarelli

22 April, 2020

IMT SCHOOL FOR ADVANCED STUDIES LUCCA

Contents

Problem: The many degrees of freedom available to researchers raise fears of post-hoc analyses (“ p -hacking”). Without properly sized tests we do not know whether to believe the findings.

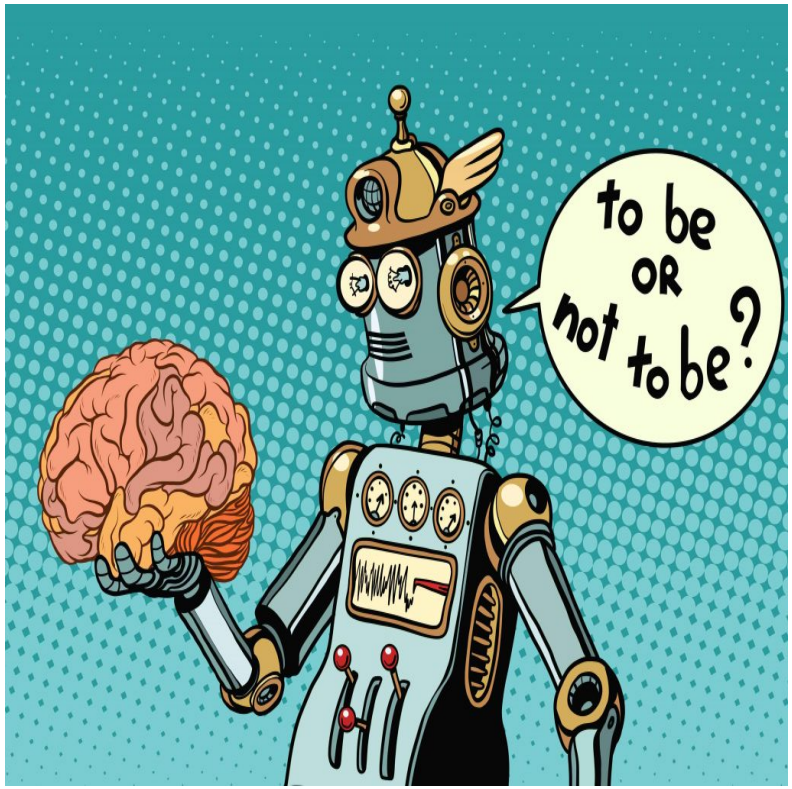
One solution is to restrict these freedoms and have researchers specify their analyses in detail ahead of time. And indeed such pre-analysis plans (PAPs) became popular.

Example: Imagine for example writing the PAP for the $300 million RAND Health Insurance Experiment before it happened. A key question is whether health insurance affects health..

Example: Imagine for example writing the PAP for the $300 million RAND Health Insurance Experiment before it happened. A key question is whether health insurance affects health..

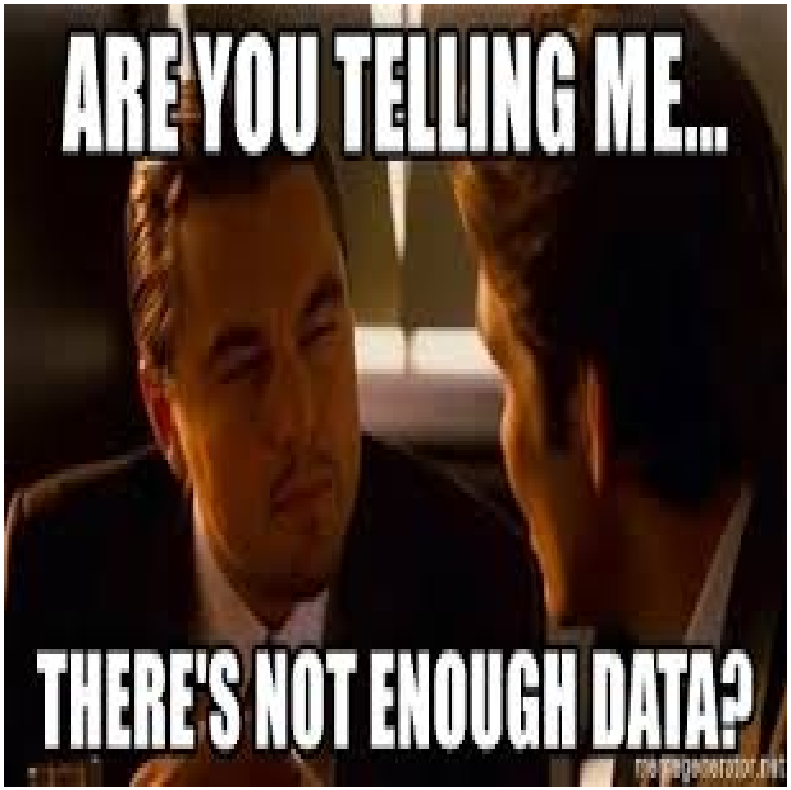
What is "health"? Is it self-reported health status? Or number of physically unhealthy days? Or specific conditions like heart disease, diabetes, or cancer (or others)? What cut-points should we use—just for example body mass index (BMI) over 30 (obesity), or also over 25 (overweight), or 40 (morbid obesity)?

Can ML be helpful?



Supervised ML algorithms focus on finding prediction functions that are as accurate as possible out of sample, using the data to select the right variables and functional forms rather than relying on the investigator to specify these choices. But...

... generic off-the-shelf ML algorithms are data-intensive. Because our social science RCTs often do not reach the scale of data typically used for ML, these ML methods are not a perfect substitute for the pre-specified analyses of current PAPs



What to do?!

Let's come back to the problem of whether there is any effect of health insurance on "health."

- a standard PAP would fully specify how different health variables are aggregated into a single test;
- a pure ML approach could start with a set of variables and aggregate them into a single index by solving a prediction problem;
- rather than picking one of those two extremes, the work combines both approaches into a single test

ML part $\left\{\begin{array}{l} \text{Fit an ML prediction function} \\ \hat{f} \text{ of } T \text{ from the group} \\ \text{of outcomes } Y = (Y_1, Y_2, \dots) \\ \text{(where subscripts denote} \\ \text{different variables) minimizing MSE.} \\ \text{By K-fold cross validation} \\ \text{obtain an “outcome index”} \\ \hat{f}(Y) \text{ for each unit in our sample.} \end{array}\right\}$

ML part $\left\{\begin{array}{l} \text{Fit an ML prediction function} \\ \hat{f} \text{ of } T \text{ from the group} \\ \text{of outcomes } Y = (Y_1, Y_2, \dots) \\ \text{(where subscripts denote} \\ \text{different variables) minimizing MSE.} \\ \text{By K-fold cross validation} \\ \text{obtain an "outcome index"} \\ \hat{f}(Y) \text{ for each unit in our sample.} \end{array}\right\}$

Standard part $\left\{\begin{array}{l} \text{Choose a subgroup of outcomes} \\ Y^* = (Y_1^*, ..., Y_n^*) \text{ which} \\ \text{are a subgroup of Y} \end{array}\right\}$

Put together

Joint (Wald) test on whether there is an average effect on any of these variables (γ_0 and γ_1) in our "health" group.

$$Y = \alpha + \beta_0 * T + \overbrace{\gamma_0 * \hat{f}(Y)}^{MLpart} + \underbrace{\gamma_1 * Y^*}_{Standardpart}$$

General results

Ex. a researcher fits an ML predictor $\hat{f}(Z)$ with K -fold cross-validation; then runs a regression of Y on regressors Z_0, Z^* and $\hat{f}(Z)$ to obtain OLS estimates $(\hat{\alpha}, \hat{\beta}, \hat{\gamma})$

General results

Ex. a researcher fits an ML predictor $\hat{f}(Z)$ with K -fold cross-validation; then runs a regression of Y on regressors Z_0, Z^* and $\hat{f}(Z)$ to obtain OLS estimates $(\hat{\alpha}, \hat{\beta}, \hat{\gamma})$

The population regression is

$$Y = \alpha' Z_0 + \beta' Z^* + \gamma f(Z) + \epsilon$$

where f is the limit of $\hat{f}$. In the what follows authors compare the output of this procedure to a standard linear regression of Y on Z_0 and Z^* to make inference on β .

PROPOSITION 1 (Cheap Lunch): *Assume regularity conditions (see the online Appendix) and*

$$E\left[\left(\hat{f}(Z)-f(Z)\right)^2\right] \to 0$$

as $n \to \infty$. *Then, asymptotically for* $n \to \infty$:

(*i*) $\left(\hat{\beta},\hat{\gamma}\right) \xrightarrow{P} (\beta,\gamma)$;

(*ii*) *If* $\gamma = 0$ *and* $E\left[\varepsilon \hat{f}(Z)\right] = 0$, *then* $\sqrt{n}\left(\left(\hat{\beta},\hat{\gamma}\right)-(\beta,\gamma)\right)$ *has the same weak Normal limit as in an OLS regression on* $Z^0, Z^*, f(Z)$; *in particular*:
- (*a*) *A Wald test of a null hypothesis with* $\gamma = 0, E\left[\varepsilon \hat{f}(Z)\right] = 0$ *is still valid*;
- (*b*) *For alternatives with* $\beta \neq \mathbf{0}$, *the power loss is at most that of adding one irrelevant regressor in the OLS regression on* Z^0, Z^*;
- (*c*) *If* $f(Z)$ *predicts the residual of the linear regression on* Z^0, Z^* *better than trivial (with respect to MSE), then power goes to* 100 *percent as* $n \to \infty$.

(*iii*) *If* $E\left[Z^* \hat{f}(Z)\right] = \mathbf{0}$ *and* $E\left[Z^*\left(Z^0\right)'\right] = \mathbf{0}\mathbf{0}'$ *then* $\sqrt{n}\left(\hat{\beta}-\beta\right)$ *has the same weak Normal limit as in an OLS regression on* $Z^0, Z^*, f(Z)$; *in particular:*
- (*a*) *OLS inference on* β *remains valid;*
- (*b*) *If* $\gamma = 0$, *then the variance of* $\hat{\beta}$ *is that in an OLS regression on* Z^0, Z^*;
- (*c*) *If* $f(Z)$ *predicts the residual of the linear regression on* Z^0, Z^* *with* $E[Z^*\varepsilon^2(Z^*)'] \prec E[Z^*(\gamma f(Z)+\varepsilon)^2(Z^*)']$, *then the variance of* $\hat{\beta}$ *goes down.*

Practical application

Practical case: *Heterogeneous effects*

Which control variables to include in estimating an average treatment effect?

Practical application

Practical case: *Heterogeneous effects*

Which control variables to include in estimating an average treatment effect?

Framework:

$$Y = \alpha + \tau T + \beta' X^* + (\tau^*)' X^* T + \gamma \hat{\tau}(X)(T - E[T]) + \epsilon \quad (1)$$

which incorporates both interaction effects with X^* as well as ML prediction $\hat{\tau}(X)$ of heterogeneous treatment effects

IMT SCHOOL FOR ADVANCED STUDIES LUCCA

Predicting $\hat{\tau}(X)$ with Random Forests

Aim:

- ⋆ Estimating true ATE: $\tau(X) = \mathbb{E}(Y_i^{(1)} - Y_i^{(0)} | X_i = x)$ (2)
- ⋆ Assuming: $\{Y_i^{(1)}, Y_i^{(0)}\} \perp\!\!\!\perp W_i | X_i$ (unconfoundedness) (3)
- ⋆ More indirect approach w.r.t. literature (where propensities $e(X)$ are directly estimated)

Main idea

Start by recursively splitting the feature space into L leaves. Then, given a test point x, evaluate the prediction. How?

Classification

$$\hat{\mu}(x) = \frac{1}{|\{i : X_i \in L(x)\}|} \sum_{\{i:X_i \in L(x)\}} Y_i. \tag{4}$$

Regression

$$\hat{\tau}(x) = \frac{1}{|\{i : W_i = 1,\, X_i \in L\}|} \sum_{\{i:W_i=1,\, X_i \in L\}} Y_i - \frac{1}{|\{i : W_i = 0,\, X_i \in L\}|} \sum_{\{i:W_i=0,\, X_i \in L\}} Y_i. \tag{5}$$

where W_i is the treatment indicator.

IMT SCHOOL FOR ADVANCED STUDIES LUCCA

Question time!

- Are the above predictions efficient/consistent?

Question time!

- ? Are the above predictions efficient/consistent?
- ? Do they guarantee asymptotic normality?

Question time!

- Are the above predictions efficient/consistent?
- Do they guarantee asymptotic normality?
- Under which conditions do any of their properties hold?

A little bit of context...

Under which conditions:

(a) training examples $Z_i = (X_i, Y_i) \quad i = 1, \ldots, n$;
We want to estimate true conditional mean function $\mathbb{E}(Y|X = x)$.

(b) regression tree T which can be used to get estimates of the conditional mean function at x: $T(x; \xi; Z_1, \ldots, Z_n)$, where $\xi \sim \Xi$ is a source of auxiliary randomness.

(c) Our goal is to use this tree-growing scheme to build a random forest: a random forest is an average of trees trained over all possible size-s subsamples of the training data, marginalizing over the auxiliary noise ξ:

$$RF(x; Z_1, \ldots, Z_n) = \binom{n}{s}^{-1} \sum_{1 \leq_1, \ldots, \leq i_s \leq n} \mathbb{E}_{\xi \sim \Xi}[T(x; \xi; Z_{i_1}, \ldots, Z_{i_s})]$$

...a few definitions...

Definition 1

The random forest with base learner T and subsample size s is practically computed as

$$RF(x; Z_1, \ldots, Z_n) \approx \frac{1}{B} \sum_{b=1}^{B} T(x; \xi_b^*; Z_{b1}^*, \ldots, Z_{bn}^*) \qquad (2)$$

where $\{Z_{b1}^*, \ldots, Z_{bn}^*\}$ is drawn without replacement from $\{Z_1, \ldots, Z_n\}$, ξ_b^* is a random draw from Ξ, and B is the number of Monte Carlo replicates we can afford to perform.

Definition 3

A tree predictor is **regular** if (standard case) each split leaves at least a fraction $\alpha > 0$ of the available training examples on each side of the split and, moreover, the trees are fully grown to depth $k \quad k \in N$

Definition 4

A tree predictor is **symmetric** if the output of the predictor does not depend on the order $(i = 1, 2, \ldots)$ in which the training examples are indexed

IMT SCHOOL FOR ADVANCED STUDIES LUCCA

...a few definitions...

Definition 1

The random forest with base learner T and subsample size s is practically computed as

$$RF(x; Z_1, \ldots, Z_n) \approx$$
$$\frac{1}{B} \sum_{b=1}^{B} T(x; \xi_b^*; Z_{b1}^*, \ldots, Z_{bn}^*) \qquad (2)$$

where $\{Z_{b1}^*, \ldots, Z_{bn}^*\}$ is drawn without replacement from $\{Z_1, \ldots, Z_n\}$, ξ_b^* is a random draw from Ξ, and B is the number of Monte Carlo replicates we can afford to perform.

Definition 2

A tree grown with training samples $(Z_1 = (X_1, Y_1), \ldots, Z_s = (X_s, Y_s))$ is **honest** if the tree dos not use the responses $Y_1, \ldots, Y_s$ in choosing where to place splits

IMT SCHOOL FOR ADVANCED STUDIES LUCCA

...a few definitions...

Definition 1

The random forest with base learner T and subsample size s is practically computed as

$$RF(x; Z_1, \ldots, Z_n) \approx \frac{1}{B} \sum_{b=1}^{B} T(x; \xi_b^*; Z_{b1}^*, \ldots, Z_{bn}^*) \qquad (2)$$

where $\{Z_{b1}^*, \ldots, Z_{bn}^*\}$ is drawn without replacement from $\{Z_1, \ldots, Z_n\}$, ξ_b^* is a random draw from Ξ, and B is the number of Monte Carlo replicates we can afford to perform.

Definition 2

A tree grown with training samples $(Z_1 = (X_1, Y_1), \ldots, Z_s = (X_s, Y_s))$ is **honest** if the tree dos not use the responses $Y_1, \ldots, Y_s$ in choosing where to place splits

Definition 3

A tree predictor is **regular** if (standard case) each split leaves at least a fraction $\alpha > 0$ of the available training examples on each side of the split and, moreover, the trees are fully grown to depth $k \quad k \in N$

...a few definitions...

Definition 1

The random forest with base learner T and subsample size s is practically computed as

$$RF(x; Z_1, \ldots, Z_n) \approx \frac{1}{B} \sum_{b=1}^{B} T(x; \xi_b^*; Z_{b1}^*, \ldots, Z_{bn}^*) \quad (2)$$

where $\{Z_{b1}^*, \ldots, Z_{bn}^*\}$ is drawn without replacement from $\{Z_1, \ldots, Z_n\}$, ξ_b^* is a random draw from Ξ, and B is the number of Monte Carlo replicates we can afford to perform.

Definition 2

A tree grown with training samples $(Z_1 = (X_1, Y_1), \ldots, Z_s = (X_s, Y_s))$ is **honest** if the tree dos not use the responses $Y_1, \ldots, Y_s$ in choosing where to place splits

Definition 3

A tree predictor is **regular** if (standard case) each split leaves at least a fraction $\alpha > 0$ of the available training examples on each side of the split and, moreover, the trees are fully grown to depth $k \quad k \in N$

Definition 4

A tree predictor is **symmetric** if the output of the predictor does not depend on the order $(i = 1, 2, \ldots)$ in which the training examples are indexed

...and theorems are served!

Bias of tree predictions:

Lemma 2. *Let T be a regular, random-split tree and let $L(x)$ denote its leaf containing x. Suppose that $X_1, ..., X_s \sim U\left([0,1]^d\right)$ independently. Then, for any $0 < \eta < 1$, and for large enough s,*

$$\mathbb{P}\left[\text{diam}_j\left(L(x)\right) \geq \left(\frac{s}{2k-1}\right)^{-\frac{0.99\,(1-\eta)\log\left((1-\alpha)^{-1}\right)}{\log\left(\alpha^{-1}\right)}\frac{\pi}{d}}\right] \leq \left(\frac{s}{2k-1}\right)^{-\frac{\eta^2}{2}\frac{1}{\log\left(\alpha^{-1}\right)}\frac{\pi}{d}}.$$

This lemma then directly translates into a bound on the bias of a single regression tree. Since a forest is an average of independently-generated trees, the bias of the forest is the same as the bias of a single tree.

Theorem 3. *Under the conditions of Lemma 2, suppose moreover that $\mu(x)$ is Lipschitz continuous, that $|Y| \leq M$, and that the trees T comprising the random forest are honest. Then, provided that $\alpha \leq 0.2$, the bias of the random forest at x is bounded by*

$$\left|\mathbb{E}\left[\hat{\mu}(x)\right] - \mu(x)\right| = \mathcal{O}\left(s^{-\frac{1}{2}\frac{\log\left((1-\alpha)^{-1}\right)}{\log\left(\alpha^{-1}\right)}\frac{\pi}{d}}\right),$$

where the constant in the $\mathcal{O}$-bound depends on M, d, α, and k; the exact dependence is given in the proof.

What about asymptotic normality?!

Asymptotic Normality of Random Forests:

(i) From standard asymptotic theory: define $\mathring{T}$ (Hajek projection of predictor T) as: $\mathring{T} = \mathbb{E}[T] + \sum_{i=1}^{n}(\mathbb{E}[T|Z_i] - \mathbb{E}[T])$;

(ii) Since the Hajek projection is a sum of independent random variables, we should expect it to be asymptotically normal under all but pathological conditions. Thus whenever the ratio of the variance of $\mathring{T}$ to that of T tends to 1, the theory of Hajek projections almost automatically guarantees that T will be asymptotically normal

$$\lim_{n\to\infty} \mathrm{Var}\left[\mathring{T}\right]/\mathrm{Var}\left[T\right] = 1, \text{ then } \lim_{n\to\infty} \mathbb{E}\left[\left\|\mathring{T} - T\right\|_2^2\right]/\mathrm{Var}\left[T\right] = 0.$$

Of course...

If T is a regression tree, condition (ii) does not apply$\longrightarrow$ classical theory of Hajek projections fails...but...it does not if we are less strigent on requirements ($\nu(s)$-incrementality):

$$Var[\mathring{T}(x;Z)]/Var[T(x;Z)] \gtrsim \nu(s)$$

Step by step

The followings apply only to regular, PNN-k trees (e.g. a tree that makes axis-aligned split and having leaves bounded to be of a given size).

First step

Establish lower bounds for the incremenatality of regression trees (not all steps are reported here):

Theorem 5. *Suppose that the conditions of Lemma 4 hold and that T is an honest k-regular symmetric tree in the sense of Definitions 2, 4, and 5. Suppose moreover that the conditional moments $\mu(x)$ and $\mu_2(x)$ are both Lipschitz continuous at x, and that $\mu_2(x)$ is uniformly bounded for all $x \in [0, 1]^d$. Finally, suppose that $\mathrm{Var}\left[Y \,\middle|\, X = x\right] > 0$. Then T is $\nu(s)$-incremental at x with*

$$\nu(s) = C_{f,d}/\log(s)^d,$$

Step by step

Second step How can we turn weakly incremental predictors T into 1-incremental ensembles by subsampling (Lemma 7), thus bringing us back into the realm of classical theory?

Lemma 7. *Let $\hat{\mu}(x)$ be the estimate for $\mu(x)$ generated by a random forest with base learner T as defined in (12), and let $\mathring{\hat{\mu}}$ be the Hájek projection of $\hat{\mu}$ (18). Then*

$$\mathbb{E}\left[\left(\hat{\mu}(x) - \mathring{\hat{\mu}}(x)\right)^2\right] \leq \left(\frac{s(n)}{n}\right)^2 \mathrm{Var}\left[T\left(x; \xi, Z_1, ..., Z_s\right)\right]$$

whenever the variance $\mathrm{Var}[T]$ of the base learner is finite.

This technical result paired with Theorem 5 or Corollary 6 leads to an asymptotic Gaussianity result; from a technical point of view, it suffices to check Lindeberg-style conditions for the central limit theorem.

Theorem 8. *Let $\hat{\mu}(x)$ be a random forest estimator trained according the conditions of Theorem 5 or Corollary 6, with Y restricted to a bounded interval $Y \in [-M, M]$. Suppose, moreover, that the subsample size $s(n)$ satisfies*

$$\lim_{n\to\infty} s(n) = \infty \text{ and } \lim_{n\to\infty} s(n)\log(n)^d/n = 0.$$

Then, there exists a sequence $\sigma_n(x) \to 0$ such that

$$\frac{\hat{\mu}(x) - \mathbb{E}[\hat{\mu}(x)]}{\sigma_n(x)} \Rightarrow \mathcal{N}(0, 1),$$

where $\mathcal{N}(0, 1)$ is the standard normal distribution.

Heterogeneous Treatment Effects

Modifications to honesty and regularity:

- **Honesty:** an honest causal tree is not allowed to look that the responses Y_i when making splits but can look at the treatment assignments W_i
- **Regularity:** a regular causal tree must have at least k examples from both treatment classes in each leaf

Notice: honesty is important to preserve independence among Y_i and W_i

Heterogeneous Treatment Effects

Given the above conditions, denoting $\mathcal{I}^{(1)}(x)$ and $\mathcal{I}^{(0)}(x)$ the indices of the treatment and control units in the leaf around x, we then find that after the splitting stage:

$$\begin{aligned}
&\mathbb{E}\left[\Gamma\left(x\right) \,\middle|\, X, W\right] \\
&\quad = \frac{\sum_{\{i \in \mathcal{I}^{(1)}(x)\}} \mathbb{E}\left[Y^{(1)} \,\middle|\, X = X_i,\, W = 1\right]}{\left|\mathcal{I}^{(1)}(x)\right|} - \frac{\sum_{\{i \in \mathcal{I}^{(0)}(x)\}} \mathbb{E}\left[Y^{(0)} \,\middle|\, X = X_i,\, W = 0\right]}{\left|\mathcal{I}^{(0)}(x)\right|} \\
&\quad = \frac{\sum_{\{i \in \mathcal{I}^{(1)}(x)\}} \mathbb{E}\left[Y^{(1)} \,\middle|\, X = X_i\right]}{\left|\mathcal{I}^{(1)}(x)\right|} - \frac{\sum_{\{i \in \mathcal{I}^{(0)}(x)\}} \mathbb{E}\left[Y^{(0)} \,\middle|\, X = X_i\right]}{\left|\mathcal{I}^{(0)}(x)\right|},
\end{aligned}$$

The second equivalence due to unconfoundedness. Do the two terms consistently estimate: $\mathbb{E}[Y_i^{(0)}]$ and $\mathbb{E}[Y_i^{(1)}]$?
YES! But remember $\longrightarrow$ modifications to honest and regularity + unconfoundedness and overlap needed now. Then theorems above apply.

Quick recap

Slide 4 presented a practical application of the PAP procedure using ML in a context with heterogeneous treatment effects

Slides 12-23 helped us understanding how ML can estimate $\hat{\tau}(X)$ through random forests

Now that we have $\hat{\tau}(X)$, we can put it in the main specification
$Y = \alpha + \tau T + \beta' X^* +$
$(\tau^*)' X^* T + \gamma \hat{\tau}(X)(T - E[T]) + \epsilon$

What's next? (see next slide)

By Proposiition I (slide 10) we know that if $\gamma = 0$ and $\mathbb{E}[\epsilon \hat{f}(Z)] = 0$, then $\sqrt{n}((\hat{\beta}, \hat{\gamma}) - (\beta, \gamma))$ has the same normal distribution as in an OLS regression on $Z, Z^*, f(Z)$.

Hence the above specification allows tests for whether there are treatment effects; whether treatment effects are heterogeneous; and whether all heterogeneity is captured by the specific covariates X^*

The worst-case cost in power of adding ML is limited by that of the inclusion of an irrelevant interaction term, and if the ML estimate indeed picks up additional heterogeneous treatment effects, these tests will detect this in the limit.

By including ML on the right-hand side of OLS, we do not lose any sample size for the full linear regression, while allowing the data to decide how much weight to put on the ML component (via $\hat{\gamma}$).