Recursive Partitioning for Heterogeneous Causal Effects

Motivation I: Experiments and Data-Mining

- Concerns about ex-post "data-mining"
  - In medicine, scholars required to pre-specify analysis plan
  - In economic field experiments, calls for similar protocols
- But how is researcher to predict all forms of heterogeneity in an environment with many covariates?
- Goal:
  - Allow researcher to specify set of potential covariates
  - Data-driven search for heterogeneity in causal effects with valid standard errors

Motivation II: Treatment Effect Heterogeneity for Policy

- Estimate of treatment effect heterogeneity needed for optimal decision-making
- This paper focuses on estimating treatment effect as function of attributes directly, not optimized for choosing optimal policy in a given setting
- This "structural" function can be used in future decision-making by policy-makers without the need for customized analysis

Preview

- Distinguish between causal effects and attributes
- Estimate treatment effect heterogeneity:
  - Introduce estimation approaches that combine ML prediction & causal inference tools
- Introduce and analyze new cross-validation approaches for causal inference
- Inference on estimated treatment effects in subpopulations
  - Enabling post-experiment data-mining

- NOTE: estimation versus prediction objective

"Moving the Goalpost": What is Question?

- Estimate $\tau(x) = E[\tau_i|X_i = x]$ as well as possible
  - Why? Want to hold some covariates fixed and look at the effect of others.
- Estimate $\mathrm{BLP}[\tau_i|X_i = x]$
  - Why? "Interpretable"? The best linear predictor is a bit hard to interpret without the whole variance-covariance matrix of nonlinear functions and interactions; you have omitted variable bias on the coefficients you are explaining, relative to $\tau(x)$. My view is that simple models can be more "mis-interpretable" than interpretable.
- Causal Tree: Find partition of covariate space and estimate $E[\tau_i|X_i \in S]$ for each element of partition
  - Why? Easier to interpret than BLP, but still important to report mean, median, percentiles of all covariates for each leaf to understand how leaves are different, when covariates are correlated.
- Which units have highest or lowest treatment effects?
  - Why? Helps understand who could be treated. Can be estimated directly or can draw inferences based on output of causal tree or non-parametric estimates of $\tau(x)$
  - Common practice to display differences between covariates; see Chernozhukov and Duflo (2018)
- What is the best policy mapping from X to treatments W?
  - Why? Sometimes this is the direct object of interest.
  - Fully nonparametric? See e.g. Hirano and Porter (2009)
  - With limited complexity or other constraints? See e.g. Kitagawa and Tetenov (2015), Athey and Wager (2017).
- What is the full set of covariates for which there is statistically significant heterogeneity?
  - List, Shaikh, and Xu (2016) (multiple testing)
- Tradeoffs: More personalization, reliable confidence intervals, role of assumptions, interpretability

Regression Trees for Prediction

Data

- Outcomes Y_i, attributes X_i.
- Support of X_i is $\mathcal{X}$.
- Have training sample with independent obs.
- Want to predict on new sample
- Ex: Predict how many clicks a link will receive if placed in the first position on a particular search query

Build a "tree":

- Partition of $\mathcal{X}$ into "leaves" $\mathcal{X}_j$
- Predict Y conditional on realization of X in each region $\mathcal{X}_j$ using the sample mean in that region
- Go through variables and leaves and decide whether and where to split leaves (creating a finer partition) using in-sample goodness of fit criterion
- Select tree complexity using cross-validation based on prediction quality

Regression Trees for Prediction: Components

1. Model and Estimation
   A. Model type: Tree structure
   B. Estimator $\hat{Y}_i$: sample mean of Y_i within leaf
   C. Set of candidate estimators C: correspond to different specifications of how tree is split
2. Criterion function (for fixed tuning parameter λ)
   A. In-sample Goodness-of-fit function:

   $$Q^{is} = \text{-MSE (Mean Squared Error)} = -\frac{1}{N}\sum_{i=1}^{N}\left(\hat{Y}_i - Y_i\right)^2$$

   A. Structure and use of criterion
      i. Criterion: $Q^{crit} = Q^{is} - \lambda \times \#$ leaves
      ii. Select member of set of candidate estimators that maximizes Q^{crit}, given λ
3. Cross-validation approach
   A. Approach: Cross-validation on grid of tuning parameters. Select tuning parameter λ with highest Out-of-sample Goodness-of-Fit Q^{os}.
   B. Out-of-sample Goodness-of-fit function: Q^{os} = -MSE

Using Trees to Estimate Causal Effects

Model:

$$Y_i = Y_i(W_i) = \begin{cases} Y_i(1) & if \;\; W_i = 1, \\ Y_i(0) & otherwise. \end{cases}$$

- Suppose random assignment of W_i
- Want to predict individual *i*'s treatment effect
  - $\tau_i = Y_i(1) - Y_i(0)$
  - This is not observed for any individual
  - Not clear how to apply standard machine learning tools
- Let

$$\mu(w, x) = \mathbb{E}[Y_i | W_i = w, X_i = x]$$
$$\tau(x) = \mu(1, x) - \mu(0, x)$$

Using Trees to Estimate Causal Effects

$$\mu(w, x) = \mathbb{E}[Y_i | W_i = w, X_i = x]$$
$$\tau(x) = \mu(1, x) - \mu(0, x)$$

- Approach 1: Analyze two groups separately
  - Estimate $\hat{\mu}(1, x)$ using dataset where $W_i = 1$
  - Estimate $\hat{\mu}(0, x)$ using dataset where $W_i = 0$
  - Use propensity score weighting (PSW) if needed
  - Do within-group cross-validation to choose tuning parameters
  - Construct prediction using
    $$\hat{\mu}(1, x) - \hat{\mu}(0, x)$$
- Approach 2: Estimate $\mu(w, x)$ using tree including both covariates
  - Include PS as attribute if needed
  - Choose tuning parameters as usual
  - Construct prediction using
    $$\hat{\mu}(1, x) - \hat{\mu}(0, x)$$
  - Estimate is zero for *x* where tree does not split on *w*

- Observations
  - Estimation and cross-validation not optimized for goal
  - Lots of segments in Approach 1: combining two distinct ways to partition the data

- Problems with these approaches
  1. Approaches not tailored to the goal of estimating treatment effects
  2. How do you evaluate goodness of fit for tree splitting and cross-validation?
     - $\tau_i = Y_i(1) - Y_i(0)$ is not observed and thus you don't have ground truth for any unit

Literature

Approaches in the spirit of single tree and two trees

- Beygelzimer and Langford (2009)
  - Analogous to "two trees" approach with multiple treatments; construct optimal policy
- Dudick, Langford, and Li (2011)
  - Combine inverse propensity score method with "direct methods" (analogous to single tree approach) to estimate optimal policy
- Foster, Tailor, Ruberg, *Statistics and Medicine* (2011)
  - Estimate $\mu(w,x)$ using random forests, define $\hat{\tau}_i = \hat{\mu}(1,X_i) - \hat{\mu}(0,X_i)$, and do trees on $\hat{\tau}_i$.
- Imai and Ratkovic (2013)
  - In context of randomized experiment, estimate $\mu(w,x)$ using lasso type methods, and then $\hat{\tau}(x) = \hat{\mu}(1,x) - \hat{\mu}(0,x)$.

Estimating treatment effects directly at leaves of trees

- Su, Tsai, Wang, Nickerson, Li (2009)
  - Do regular tree, but split if the t-stat for the treatment effect difference is large, rather than when the change in prediction error is large.
- Zeileis, Hothorn, and Hornick (2005)
  - "Model-based recursive partitioning": estimate a model at the leaves of a tree. In-sample splits based on prediction error, do not focus on out of sample cross-validation for tuning.

Transformed outcomes or covariates for regressions

- Tibshirani et al (2014)
- Weisberg and Pontes (2015)

None of these explore cross-validation based on treatment effect.

Another Approach: Transform the Outcome

- Suppose we have 50-50 randomization of treatment/control
  - Let $Y_i^* = \begin{cases} 2Y_i & if\ W_i = 1 \\ -2Y_i & if\ W_i = 0 \end{cases}$
  - Then $E[Y_i^*] = 2 \cdot \left(\frac{1}{2}E[Y_i(1)] - \frac{1}{2}E[Y_i(0)]\right) = E[\tau_i]$

- Suppose treatment with probability p_i
  - Let $Y_i^* = \frac{W_i - p}{p(1-p)} Y_i = \begin{cases} \frac{1}{p} Y_i & if\ W_i = 1 \\ -\frac{1}{1-p} Y_i & if\ W_i = 0 \end{cases}$
  - Then $E[Y_i^*] = \left(p\frac{1}{p}E[Y_i(1)] - (1-p)\frac{1}{1-p}E[Y_i(0)]\right) = E[\tau_i]$

- Selection on observables or stratified experiment
  - Let $Y_i^* = \frac{W_i - p(X_i)}{p(X_i)(1-p(X_i))} Y_i$
  - Estimate $\hat{p}(x)$ using traditional methods

Causal Trees:
(Conventional Tree, Transformed Outcome)

1. ## Model and Estimation
   A. Model type: Tree structure
   B. Estimator $\hat{\tau}_i^*$: sample mean of Y_i^* within leaf
   C. Set of candidate estimators *C*: correspond to different specifications of how tree is split
2. ## Criterion function (for fixed tuning parameter λ)
   A. In-sample Goodness-of-fit function:

      Q^{is} = -MSE (Mean Squared Error) $= -\frac{1}{N}\sum_{i=1}^{N}(\hat{\tau}_i^* - Y_i^*)^2$
   A. Structure and use of criterion
      i. Criterion: $Q^{crit} = Q^{is} - \lambda \times \#$ leaves
      ii. Select member of set of candidate estimators that maximizes Q^{crit}, given λ
3. ## Cross-validation approach
   A. Approach: Cross-validation on grid of tuning parameters. Select tuning parameter λ with highest Out-of-sample Goodness-of-Fit Q^{os}.
   B. Out-of-sample Goodness-of-fit function: Q^{os} = -MSE

Critique of Approach: Transform the Outcome

$$Y_i^* = \frac{W_i - p}{p(1-p)} Y_i = \begin{cases} \frac{1}{p} Y_i & if\ W_i = 1 \\ -\frac{1}{1-p} Y_i & if\ W_i = 0 \end{cases}$$

- Within a leaf, sample average of Y_i^* is not most efficient estimator of treatment effect
  - The proportion of treated units within the leaf is not the same as the overall sample proportion
- This motivates preferred approach: use sample average treatment effect in the leaf

Causal Trees:
(Causal Tree, TOT loss function)

1. Model and Estimation
   A. Model type: Tree structure
   B. Estimator $\hat{\tau}_i^{CT}$: sample average treatment effect within leaf (w/ PSW)
   C. Set of candidate estimators C: correspond to different specifications of how tree is split
2. Criterion function (for fixed tuning parameter λ)
   A. In-sample Goodness-of-fit function:

      Q^{is} = -MSE (Mean Squared Error) $= -\frac{1}{N}\sum_{i=1}^{N}(\hat{\tau}_i^{CT} - Y_i^*)^2$
   A. Structure and use of criterion
      i. Criterion: $Q^{crit} = Q^{is} - \lambda \times \#$ leaves
      ii. Select member of set of candidate estimators that maximizes Q^{crit}, given λ
3. Cross-validation approach
   A. Approach: Cross-validation on grid of tuning parameters. Select tuning parameter λ with highest Out-of-sample Goodness-of-Fit Q^{os}.
   B. Out-of-sample Goodness-of-fit function: Q^{os} = -MSE

Causal Trees

- What are you estimating? Within a leaf estimate treatment effect rather than a mean
  - Difference in average outcomes for treated and control group
  - Weight by normalized inverse propensity score in observational studies
- What is your goal? MSE of *treatment effects:* $-E_{S^T}\left[\sum_{i \in S^T}(\tau_i - \hat{\tau}(X_i))^2\right]$
- Problem: this is infeasible (true treatment effect unobserved)
  - We show we can estimate the criteria
- We also modify existing methods to be "honest." We decouple model selection from model estimation.
  - Split sample, one sample to build tree, second to estimate effects.
  - This changes criteria—novel idea for the literature.

$$-E_{S^T,S^E}\left[\sum_{i \in S^T}(\tau_i - \hat{\tau}(X_i; S^E))^2\right]$$

  - Tradeoff:
    - COST: sample splitting means build shallower tree, less personalized predictions, and lower MSE of treatment effects.
    - BENEFIT: Valid confidence intervals with coverage rates that do not deteriorate as data generating process gets more complex or more covariates are added.

Honest Causal Trees

- Honest estimation changes expected criterion
  - Criterion anticipates that we will re-estimate effects in the leaves.
    - The bias due to "dishonest" selection of tree structure will be eliminated.
    - Eliminating the bias was the main purpose of cross-validation in standard method.
  - We face uncertainty in what honest sample will estimate
    - Small leaves will create noise.
    - Splitting on variables that don't affect treatment effect can reduce variance

- Criterion for splitting and cross-validation changes
  - Given set of leaves, MSE on test set taking into account re-estimation.
  - Uncertainty over estimation set and test set at time of evaluation.

Standard *Prediction* Trees

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

$$\mathrm{EMSE}_\mu(\Pi) \equiv \mathbb{E}_{\mathcal{S}^{\mathrm{te}}, \mathcal{S}^{\mathrm{est}}} \left[\mathrm{MSE}_\mu(\mathcal{S}^{\mathrm{te}}, \mathcal{S}^{\mathrm{est}}, \Pi)\right]$$

Conventional CART uses for training and CV, respectively:

$$-\mathrm{MSE}_\mu(\mathcal{S}^{\mathrm{tr}}, \mathcal{S}^{\mathrm{tr}}, \Pi) = \frac{1}{N^{\mathrm{tr}}} \sum_{i \in \mathcal{S}^{\mathrm{tr}}} \hat{\mu}^2(X_i; \mathcal{S}^{\mathrm{tr}}, \Pi)$$

$$-\mathrm{MSE}_\mu(\mathcal{S}^{\mathrm{tr,cv}}, \mathcal{S}^{\mathrm{tr,tr}}, \Pi)$$

$$= \frac{1}{N^{tr,cv}} \sum_{i \in \mathcal{S}^{tr,cv}} \left(\left(\hat{\mu}(X_i; \mathcal{S}^{tr,tr})\right)^2 - 2\hat{\mu}(X_i; \mathcal{S}^{tr,cv})\hat{\mu}(X_i; \mathcal{S}^{tr,tr}) \right)$$

Honest *Prediction* Trees

$$-\mathrm{EMSE}_\mu(\Pi) = -\mathbb{E}_{(Y_i,X_i),\mathcal{S}^{\mathrm{est}}}\left[(Y_i - \mu(X_i;\Pi))^2 - Y_i^2\right]$$

$$-\mathbb{E}_{X_i,\mathcal{S}^{\mathrm{est}}}\left[\left(\hat{\mu}(X_i;\mathcal{S}^{\mathrm{est}},\Pi) - \mu(X_i;\Pi)\right)^2\right] =$$

$$\mathbb{E}_{X_i}[\mu^2(X_i;\Pi)] - \mathbb{E}_{\mathcal{S}^{\mathrm{est}},X_i}\left[\mathbb{V}(\hat{\mu}^2(X_i;\mathcal{S}^{\mathrm{est}},\Pi)\right],$$

This uses fact that estimator on independent sample is unbiased

$$\widehat{\mathbb{V}}(\hat{\mu}(x;\mathcal{S}^{\mathrm{est}},\Pi)) \equiv \frac{S^2_{\mathcal{S}^{\mathrm{tr}}}(\ell(x;\Pi))}{N^{\mathrm{est}}(\ell(x;\Pi))}$$

$$\widehat{\mathbb{E}}\left[\mathbb{V}(\hat{\mu}^2(X_i;\mathcal{S}^{\mathrm{est}},\Pi)|i \in \mathcal{S}^{\mathrm{te}}\right] \equiv \frac{1}{N^{\mathrm{est}}} \cdot \sum_{\ell\in\Pi} S^2_{\mathcal{S}^{\mathrm{tr}}}(\ell)$$

$$-\widehat{\mathrm{EMSE}}_\mu(\mathcal{S}^{\mathrm{tr}},N^{\mathrm{est}},\Pi) \equiv$$

$$\frac{1}{N^{\mathrm{tr}}}\sum_{i\in\mathcal{S}^{\mathrm{tr}}}\hat{\mu}^2(X_i;\mathcal{S}^{\mathrm{tr}},\Pi) - \left(\frac{1}{N^{\mathrm{tr}}} + \frac{1}{N^{\mathrm{est}}}\right)\cdot\sum_{\ell\in\Pi} S^2_{\mathcal{S}^{\mathrm{tr}}}(\ell(x;\Pi))$$

Standard *Causal* Trees

$$\mathrm{MSE}_\tau(\mathcal{S}^{\mathrm{te}}, \mathcal{S}^{\mathrm{est}}, \Pi) \equiv \frac{1}{\#(\mathcal{S}^{\mathrm{te}})} \sum_{i \in \mathcal{S}^{\mathrm{te}}} \left\{ \left(\tau_i - \hat{\tau}(X_i; \mathcal{S}^{\mathrm{est}}, \Pi)\right)^2 - \tau_i^2 \right\}$$

$$\mathrm{EMSE}_\tau(\Pi) \equiv \mathbb{E}_{\mathcal{S}^{\mathrm{te}}, \mathcal{S}^{\mathrm{est}}} \left[\mathrm{MSE}_\tau(\mathcal{S}^{\mathrm{te}}, \mathcal{S}^{\mathrm{est}}, \Pi) \right]$$

$$\widehat{\mathrm{MSE}}_\tau(\mathcal{S}^{\mathrm{te}}, \mathcal{S}^{\mathrm{tr}}, \Pi) \equiv -\frac{2}{N^{\mathrm{tr}}} \sum_{i \in \mathcal{S}^{\mathrm{te}}} \hat{\tau}(X_i; \mathcal{S}^{\mathrm{te}}, \Pi) \cdot \hat{\tau}(X_i; \mathcal{S}^{\mathrm{tr}}, \Pi)$$

$$+ \frac{1}{N^{\mathrm{tr}}} \sum_{i \in \mathcal{S}^{\mathrm{te}}} \hat{\tau}^2(X_i; \mathcal{S}^{\mathrm{tr}}, \Pi).$$

For training and CV, respectively:

$$-\widehat{\mathrm{MSE}}_\tau(\mathcal{S}^{\mathrm{tr}}, \mathcal{S}^{\mathrm{tr}}, \Pi) = \frac{1}{N^{\mathrm{tr}}} \sum_{i \in \mathcal{S}^{\mathrm{tr}}} \hat{\tau}^2(X_i; \mathcal{S}^{\mathrm{tr}}, \Pi)$$

$$-\widehat{\mathrm{MSE}}_\tau(\mathcal{S}^{\mathrm{tr,cv}}, \mathcal{S}^{\mathrm{tr,tr}}, \Pi)$$

Honest *Causal* Trees

$$-\text{EMSE}_\tau(\Pi) = \mathbb{E}_{X_i}[\tau^2(X_i;\Pi)] - \mathbb{E}_{\mathcal{S}^{\text{est}},X_i}\left[\mathbb{V}(\hat{\tau}^2(X_i;\mathcal{S}^{\text{est}},\Pi)\right]$$

This uses fact that estimator on independent sample is unbiased

For training and CV, respectively:

$$-\widehat{\text{EMSE}}_\tau(\mathcal{S}^{\text{tr}}, N^{\text{est}}, \Pi) \equiv \frac{1}{N^{\text{tr}}} \sum_{i\in\mathcal{S}^{\text{tr}}} \hat{\tau}^2(X_i;\mathcal{S}^{\text{tr}},\Pi)$$

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

$$-\widehat{\text{EMSE}}_\tau(\mathcal{S}^{\text{tr,cv}}, N^{\text{est}}, \Pi)$$

Inference

- Attractive feature of trees:
  - Can easily separate tree construction from treatment effect estimation
  - Tree constructed on training sample is indep. of sampling variation in test sample
  - Holding tree from training sample fixed, can 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.

- We do not require ANY assumptions about sparsity of true data-generating process. *Coverage* does not deteriorate at all as you increase number of covariates.
  - But we do not attempt to make fully personalized estimates

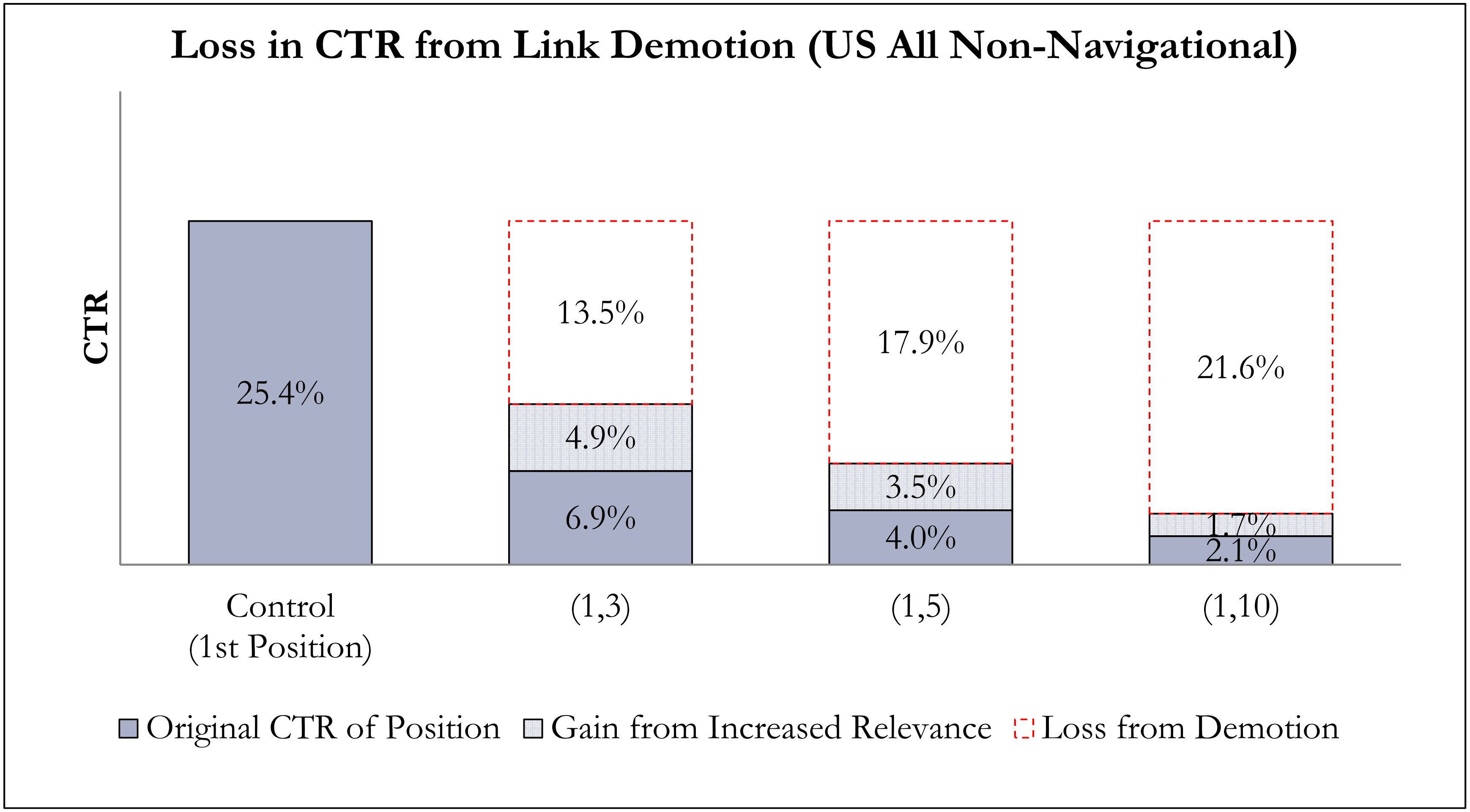
Comparing Alternative Approaches to Preferred Honest Causal Tree

- Dishonest with double the sample
  - Does worse if true model is sparse (also the case where bias is less severe)
  - Has similar or better MSE in many cases, but poor coverage of confidence intervals
- Splitting on statistical criteria of model fit
  - Paper shows formally how these methods differ (proposed in a small related literature, one that doesn't consider honesty and cross-validation issues)
  - Splitting on T-statistic on treatment effect ignores variance reduction from reducing imbalance on covariates
  - Splitting on overall model fit prioritizes level heterogeneity above treatment effects

Application: Treatment Effect Heterogeneity in Estimating Position Effects in Search

- Queries highly heterogeneous
  - Tens of millions of unique search phrases each month
  - Query mix changes month to month for a variety of reasons
  - Behavior conditional on query is fairly stable
- Desire for segments.
  - Want to understand heterogeneity and make decisions based on it
  - "Tune" algorithms separately by segment
  - Want to predict outcomes if query mix changes
    - For example, bring on new syndication partner with more queries of a certain type

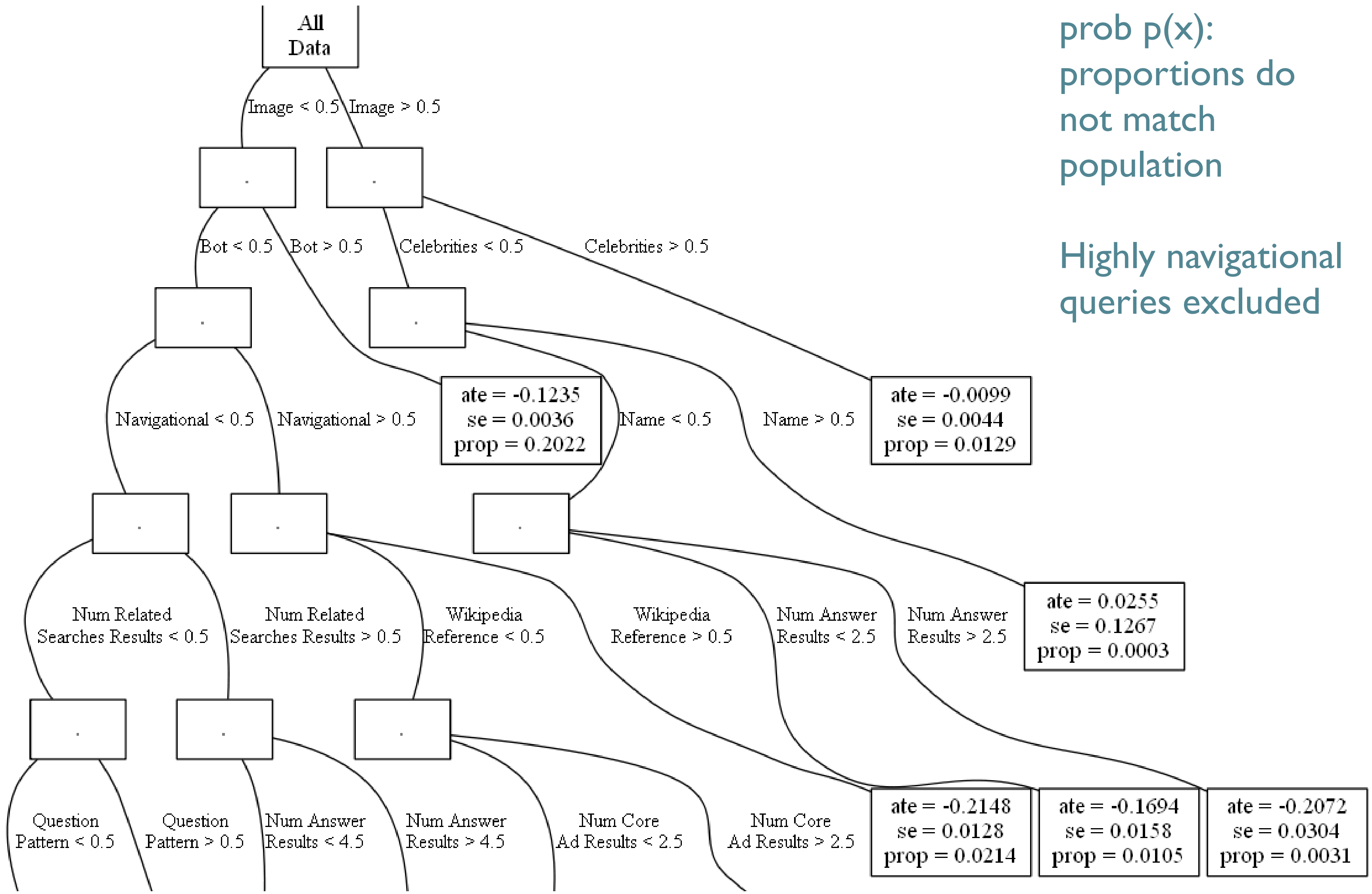
Relevance v. Position

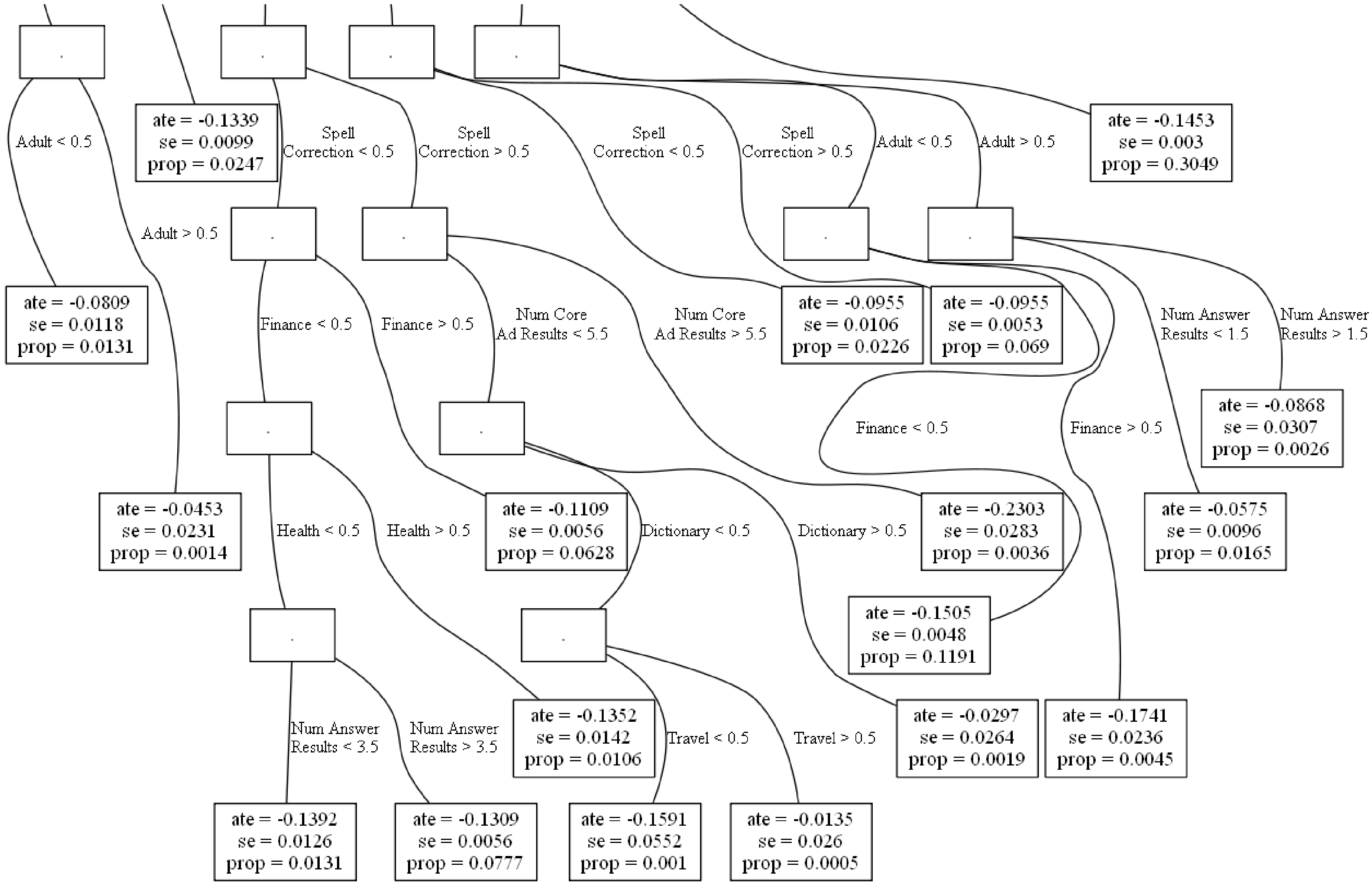


Search Experiment Tree: Effect of Demoting Top Link (Test Sample Effects)

Some data excluded with prob p(x): proportions do not match population

Highly navigational queries excluded





ate = -0.1339
se = 0.0099
prop = 0.0247
ate = -0.1453
se = 0.003
prop = 0.3049
Adult < 0.5
Adult > 0.5
Spell Correction < 0.5
Spell Correction > 0.5
Spell Correction < 0.5
Spell Correction > 0.5
Adult < 0.5
Adult > 0.5
ate = -0.0809
se = 0.0118
prop = 0.0131
Finance < 0.5
Finance > 0.5
Num Core Ad Results < 5.5
Num Core Ad Results > 5.5
ate = -0.0955
se = 0.0106
prop = 0.0226
ate = -0.0955
se = 0.0053
prop = 0.069
Num Answer Results < 1.5
Num Answer Results > 1.5
ate = -0.0868
se = 0.0307
prop = 0.0026
Finance < 0.5
Finance > 0.5
ate = -0.0453
se = 0.0231
prop = 0.0014
Health < 0.5
Health > 0.5
ate = -0.1109
se = 0.0056
prop = 0.0628
Dictionary < 0.5
Dictionary > 0.5
ate = -0.2303
se = 0.0283
prop = 0.0036
ate = -0.0575
se = 0.0096
prop = 0.0165
ate = -0.1505
se = 0.0048
prop = 0.1191
Num Answer Results < 3.5
Num Answer Results > 3.5
ate = -0.1352
se = 0.0142
prop = 0.0106
Travel < 0.5
Travel > 0.5
ate = -0.0297
se = 0.0264
prop = 0.0019
ate = -0.1741
se = 0.0236
prop = 0.0045
ate = -0.1392
se = 0.0126
prop = 0.0131
ate = -0.1309
se = 0.0056
prop = 0.0777
ate = -0.1591
se = 0.0552
prop = 0.001
ate = -0.0135
se = 0.026
prop = 0.0005

Use Test Sample for Segment Means & Std Errors to Avoid Bias

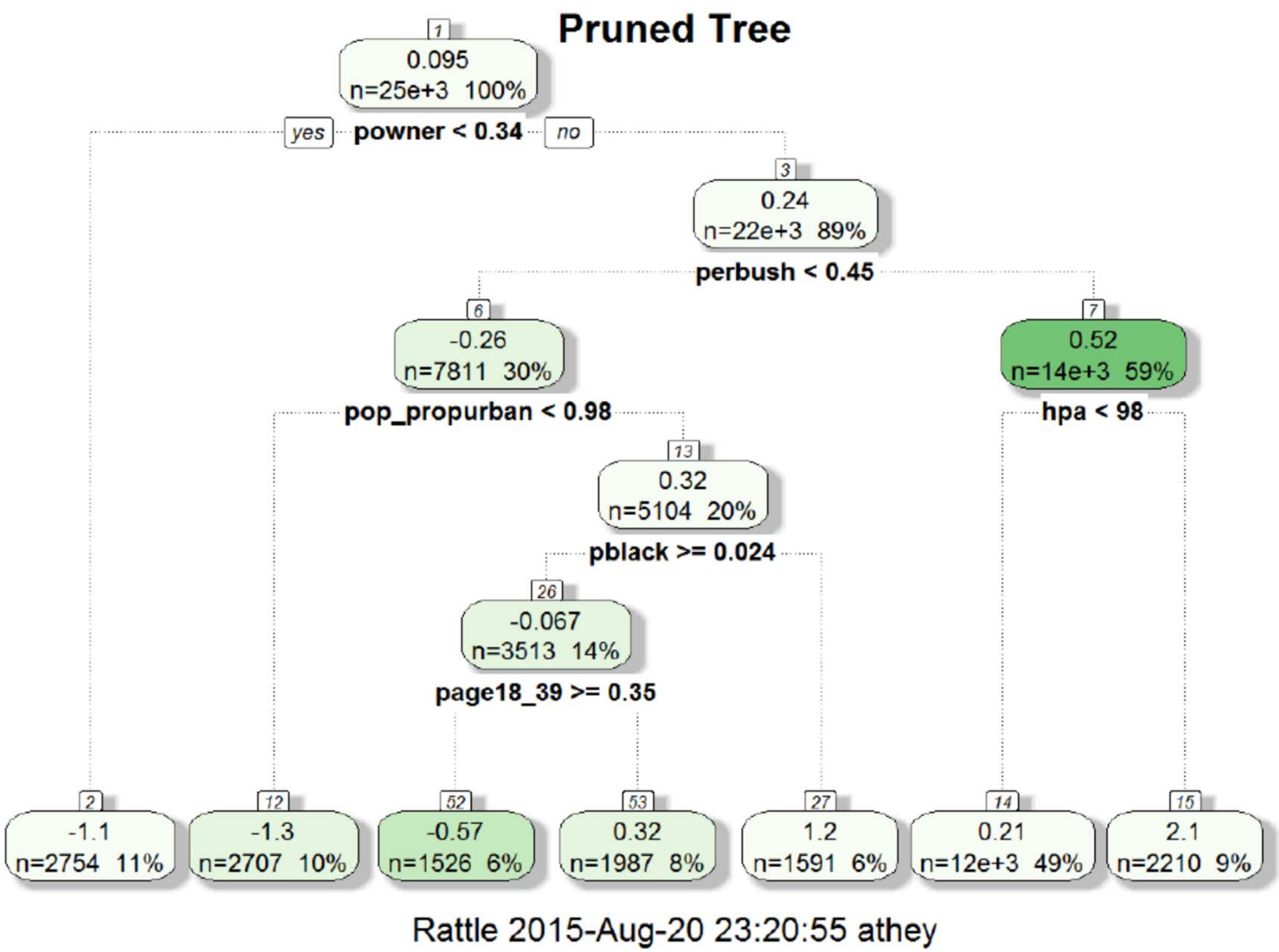
Variance of estimated treatment effects in training sample 2.5 times that in test sample (adaptive estimates biased)

Honest Estimates			Adaptive Estimates		
Treatment Effect	Standard Error	Proportion	Treatment Effect	Standard Error	Proportion
-0.124	0.004	0.202	-0.124	0.004	0.202
-0.134	0.010	0.025	-0.135	0.010	0.024
-0.010	0.004	0.013	-0.007	0.004	0.013
-0.215	0.013	0.021	-0.247	0.013	0.022
-0.145	0.003	0.305	-0.148	0.003	0.304
-0.111	0.006	0.063	-0.110	0.006	0.064
-0.230	0.028	0.004	-0.268	0.028	0.004
-0.058	0.010	0.017	-0.032	0.010	0.017
-0.087	0.031	0.003	-0.056	0.029	0.003
-0.151	0.005	0.119	-0.169	0.005	0.119
-0.174	0.024	0.005	-0.168	0.024	0.005
0.026	0.127	0.000	0.286	0.124	0.000
-0.030	0.026	0.002	-0.009	0.025	0.002
-0.135	0.014	0.011	-0.114	0.015	0.010
-0.159	0.055	0.001	-0.143	0.053	0.001
-0.014	0.026	0.001	0.008	0.050	0.000
-0.081	0.012	0.013	-0.050	0.012	0.013
-0.045	0.023	0.001	-0.045	0.021	0.001
-0.169	0.016	0.011	-0.200	0.016	0.011
-0.207	0.030	0.003	-0.279	0.031	0.003
-0.096	0.011	0.023	-0.083	0.011	0.022
-0.096	0.005	0.069	-0.096	0.005	0.070
-0.139	0.013	0.013	-0.159	0.013	0.013
-0.131	0.006	0.078	-0.128	0.006	0.078

Analyzing Field Experiments: Revisit Karlan and List (AER)

- Field experiment on charitable giving
  - Solicitations: treatment groups are offered match of various sizes, example gifts, and limits
  - Finding: match increases gift amount, but larger sizes do not have incremental effect
  - Some exploration of heterogeneity
- Apply causal trees:
  - Explore heterogeneity more systematically
  - Highlight the risks of data mining
  - Unlike the search application, it appears there isn't a lot of treatment effect heterogeneity

Effect of All Treatments (Pooled) Training Data



Training Sample

```
Coefficients:
                           Estimate Std. Error t value Pr(>|t|)    
leavesr-1.342               1.9395     0.2768   7.006 2.51e-12 ***
leavesr-1.083               1.6189     0.2845   5.691 1.28e-08 ***
leavesr-0.573               0.9498     0.3696   2.570  0.01018 *  
leavesr0.213                0.3260     0.1350   2.415  0.01574 *  
leavesr0.321                0.6566     0.3257   2.016  0.04379 *  
leavesr1.171                0.2839     0.3669   0.774  0.43907    
leavesr2.086                1.6176     0.3288   4.921 8.68e-07 ***
leavesr-1.342:treatment    -1.3417     0.3445  -3.895 9.86e-05 ***
leavesr-1.083:treatment    -1.0826     0.3475  -3.116  0.00184 ** 
leavesr-0.573:treatment    -0.5733     0.4593  -1.248  0.21201    
leavesr0.213:treatment      0.2131     0.1648   1.294  0.19581    
leavesr0.321:treatment      0.3210     0.4035   0.796  0.42632    
leavesr1.171:treatment      1.1707     0.4527   2.586  0.00972 ** 
leavesr2.086:treatment      2.0856     0.3951   5.279 1.31e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Estimation Sample

```
Coefficients:
                         Estimate Std. Error t value Pr(>|t|)
leavesr-1.342             0.62900    0.28579   2.201  0.02775 *
leavesr-1.083             0.51463    0.28810   1.786  0.07407 .
leavesr-0.573             0.75687    0.40245   1.881  0.06003 .
leavesr0.213              0.34395    0.13808   2.491  0.01275 *
leavesr0.321              1.07704    0.34707   3.103  0.00192 **
leavesr1.171              1.09966    0.36281   3.031  0.00244 **
leavesr2.086              3.39262    0.31770  10.679  < 2e-16 ***
leavesr-1.342:treatment   0.37260    0.34983   1.065  0.28685
leavesr-1.083:treatment   0.29186    0.35475   0.823  0.41067
leavesr-0.573:treatment   0.10772    0.48868   0.220  0.82553
leavesr0.213:treatment    0.17869    0.16888   1.058  0.29002
leavesr0.321:treatment   -0.28147    0.42071  -0.669  0.50347
leavesr1.171:treatment   -0.08332    0.44850  -0.186  0.85262
leavesr2.086:treatment    0.89775    0.39055   2.299  0.02153 *
```

Interpretation: Sample Splitting is Key

- Training set makes it look like there is lots of heterogeneity
- Estimation sample accurately shows most of that is spurious
- You get the right, if disappointing, answer out of the estimation sample

Summary

- Key to approach
  - Distinguish between causal and predictive parts of model

- Combining two literatures
  - Combining very well established tools from different literatures
  - Systematic model selection with many covariates
  - Optimized for problem of causal effects
    - In terms of tradeoff between granular prediction and overfitting
  - With valid inference
    - While sacrificing fully personalized predictions, but gaining…
  - Easy to communicate method and interpret results
    - Output is a partition of sample, treatment effects and standard errors