MACHINE LEARNING & PUBLIC POLICY

LECTURE V: MACHINE LEARNING FOR CAUSAL INFERENCE

Professor Edward McFowland III

Information Systems and Decision Sciences
Carlson School of Managment

University of Minnesota

PREDICTION FOR POLICY

Observation 1: Sometimes correlation is valuable on its own

ANOMALY DETECTION

Observation 11: Sometimes goal is one of detection or discovery

CAUSAL INFERENCE

SETUP AND NOTATION

- We follow the Neyman-Rubin Causal Model
- We have an iid sample $\mathcal{N}$ of units $(1, \ldots, N)$ from a population $\mathcal{P}$
- Every unit i is described by the tuple $(Y_i(0), Y_i(1), X_i, W_i)$
  - $\left(Y_i(0), Y_i(1)\right)$ are the set of potential outcomes under (binary) treatment conditions (e.g., test scores)
  - X_i is the vector of pre-treatment covariates (e.g., demographics)
  - W_i is the (binary) treatment indicator (e.g., if received tutoring or not)
  - $Y_i^{obs} = Y_i(W_i)$ (e.g., the test score we observe)

CHALLENGES TO CASUAL INFERENCE

- Estimands of interest are?
  - $\tau_{ATE} \coloneqq \mathbb{E}[Y_i(1) - Y_i(0)]$
- What would be a good (unbiased and consistent) estimator?

$$\tau_{ATE} \coloneqq \frac{1}{N}\sum Y_i(1) - \frac{1}{N}\sum Y_i(0)$$

  - Is this feasible? What estimator this feasible?

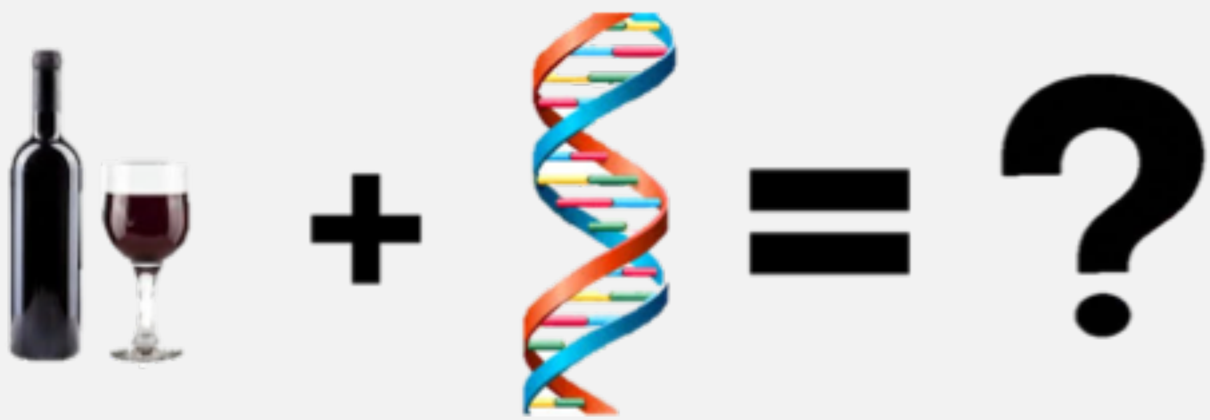
$$\hat{\tau}_{ATE} \coloneqq \frac{1}{N_T}\sum_{W_i=1} Y_i^{obs} - \frac{1}{N_C}\sum_{W_i=0} Y_i^{obs}$$

- What is the problem with what is feasible?
  - Selection: $\big(Y_i(0), Y_i(1)\big) \perp\!\!\!\perp W_i$
  - Students who have educated parents, may be more likely to have higher test scores, and are very motivated so they select to get tutoring
  - Athletes may be more likely to have lower test scores, and are forced to get tutoring

ARE WE AT AN IMPASSE?

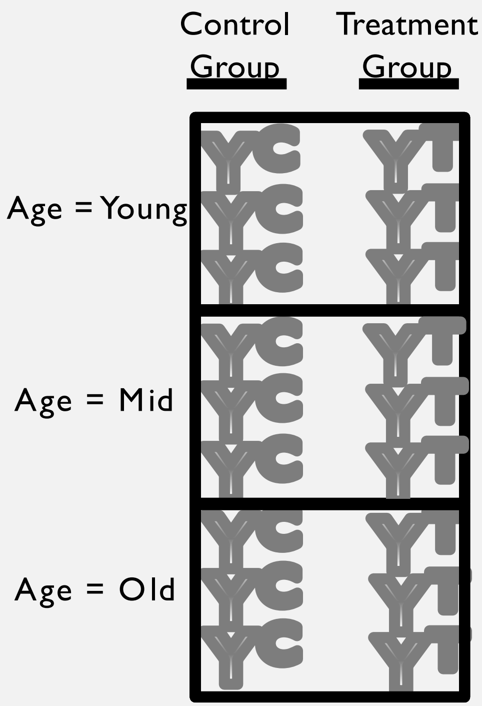
- We need additional assumptions, e.g., selection on X_i (unconfoundedness)
  - $(Y_i(0), Y_i(1)) \perp\!\!\!\perp W_i \mid X_i$
- Students who have educated parents, may be more likely to have higher test scores, and are very motivated so they select to get tutoring
  - If X_i contains the educational level of parents we are ok
- Athletes may be more likely to have lower test scores, and are forced to get tutoring
  - If X_i contains an indicator of the student if an athlete we are ok
- Also, X_i can contain information highly predictive of $P(W_i)$
- Randomized Experiments are the gold standard, if possible, because
  - $(Y_i(0), Y_i(1)) \perp\!\!\!\perp W_i$
  - $\hat{\tau}_{ATE} := \frac{1}{N_T} \sum_{W_i=1} Y_i^{obs} - \frac{1}{N_C} \sum_{W_i=0} Y_i^{obs}$ therefore, unbiased & consistent estimate of τ_{ATE}

TREATMENT EFFECT HETEROGENEITY

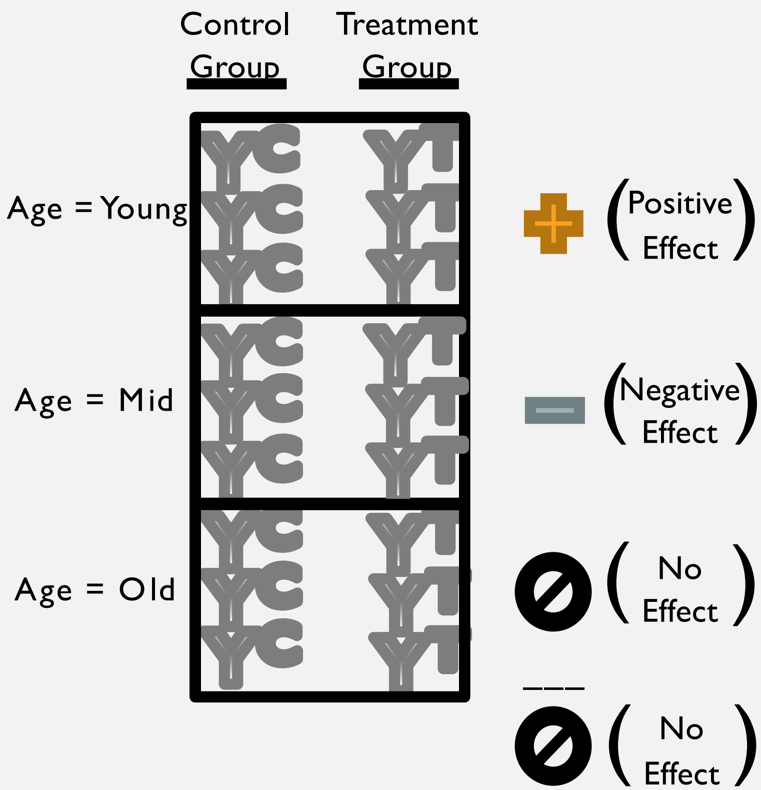


TREATMENT EFFECT HETEROGENEITY

TREATMENT EFFECT HETEROGENEITY

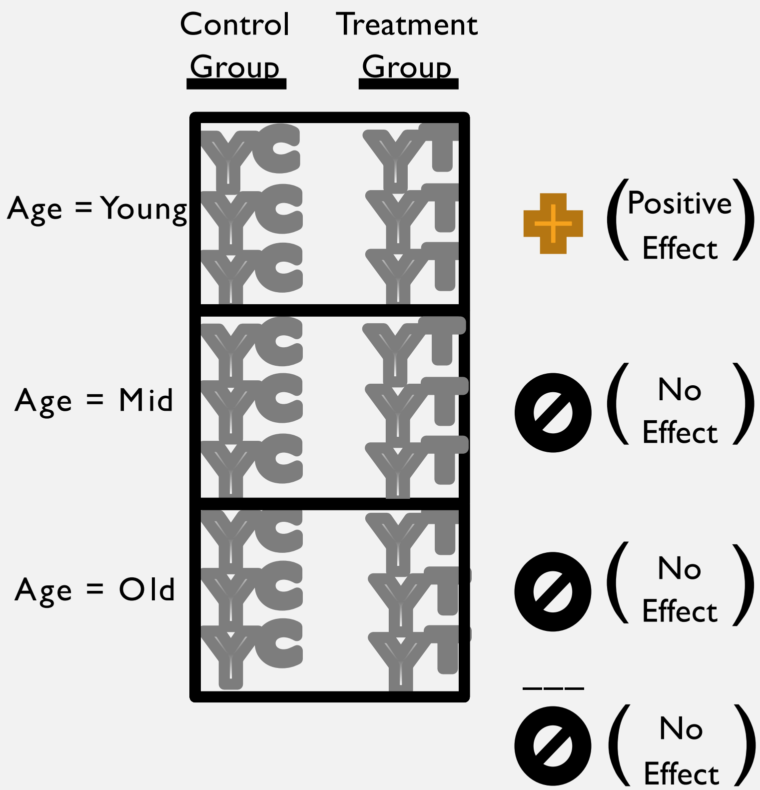


TREATMENT EFFECT HETEROGENEITY



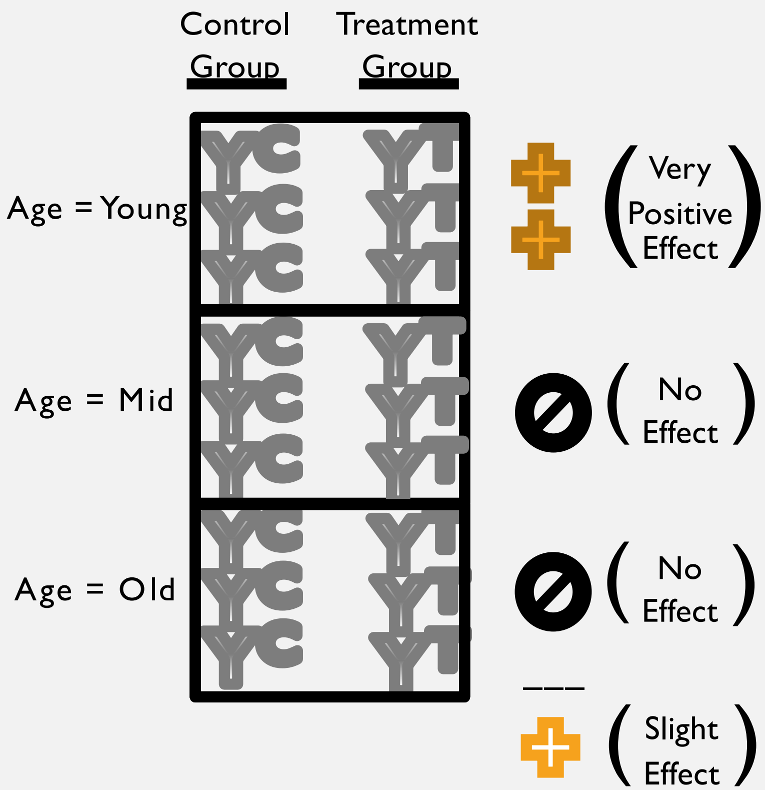
- Positive and negative effects can cancel

TREATMENT EFFECT HETEROGENEITY



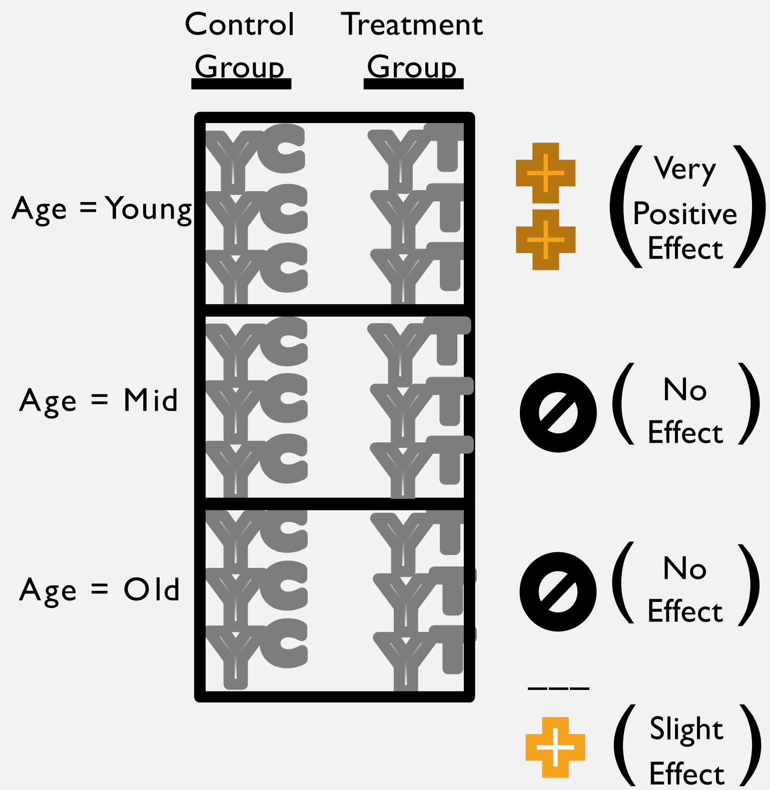
- Positive and negative effects can cancel
- True effect can be masked

TREATMENT EFFECT HETEROGENEITY



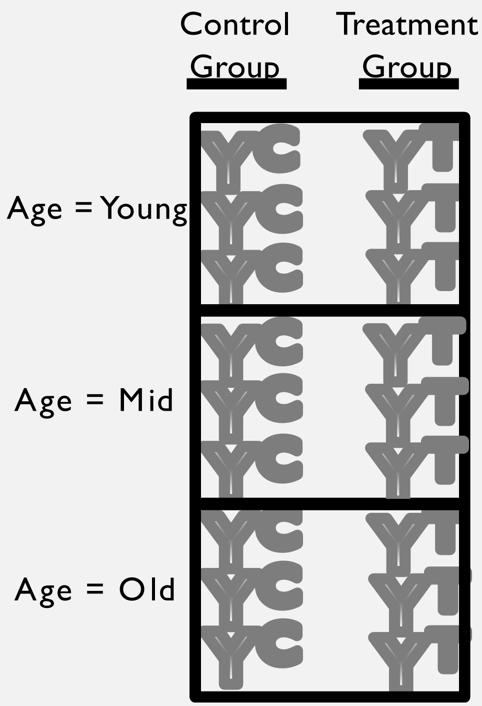
- Positive and negative effects can cancel
- True effect can be masked
- Effects could really be driven by a subpopulation

TREATMENT EFFECT HETEROGENEITY



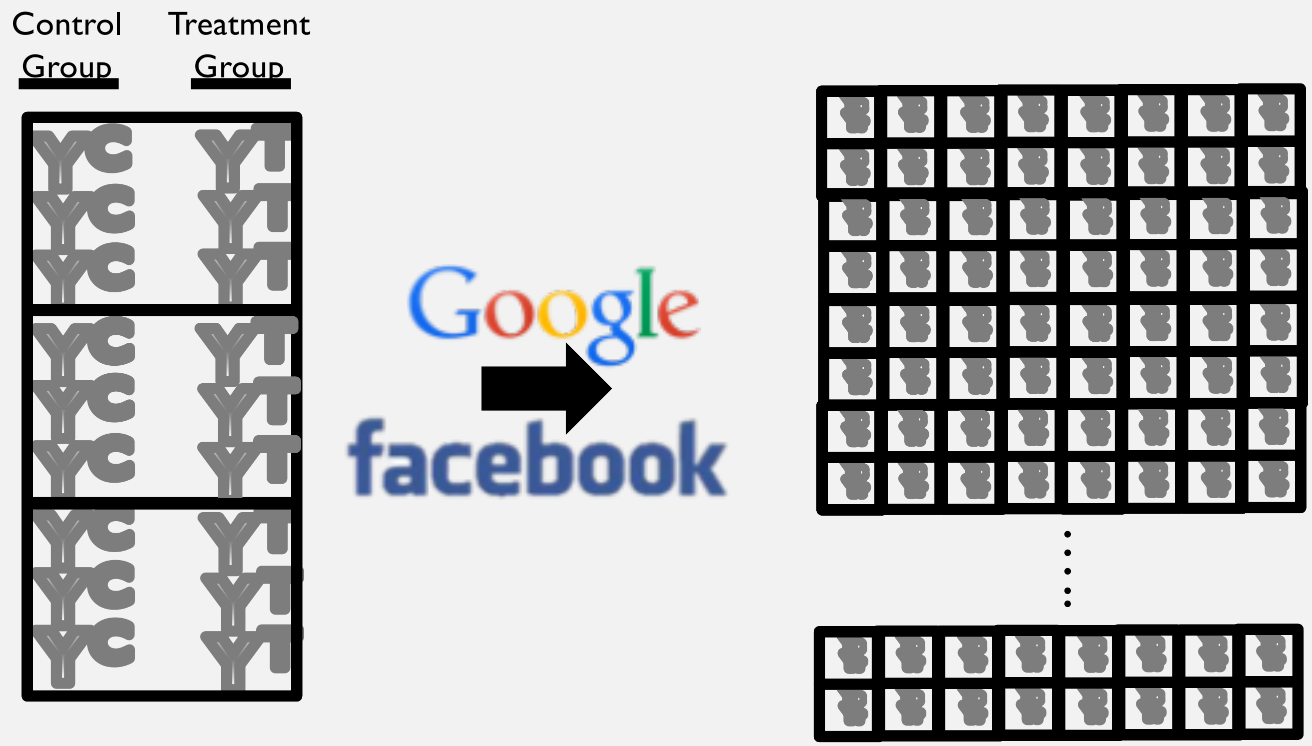
- Positive and negative effects can cancel
- True effect can be masked
  - Ex: FDA Approved BiDil Drug
- Effects could really be driven by a subpopulation
  - Ex: Perry Preschool

TREATMENT EFFECT HETEROGENEITY



- We need richer estimands of interest
  - $\tau_{CATE}(x) := \mathbb{E}[Y_i(1) - Y_i(0) | X_i = x]$
- With a good (unbiased and consistent) estimator
  - $\hat{\tau}_{CATE}(x) := \frac{1}{N_{T,x}} \sum_{\substack{W_i=1 \\ X_i=x}} Y_i^{obs} - \frac{1}{N_{C,x}} \sum_{\substack{W_i=0 \\ X_i=x}} Y_i^{obs}$

TREATMENT EFFECT HETEROGENEITY



ML FOR HTE

Observation III: Sometimes correlation can be (forced into) causation

MACHINE LEARNING'S CONTRIBUTIONS

- Regression Methods
  - OLS and Regularized Regression (e.g., LASSO)
  - Imai and Ratkovic (2013)
- Single Tree Methods
  - Athey and Imbens (2017)
- Ensemble Methods
  - Wager and Athey (2017)
- Anomalous Pattern Detection
  - McFowland et al. (2018)

(SPARSE) REGRESSION METHOD

- OLS
$$y = \gamma_0 + \gamma_1 X_1 + \gamma_2 X_2 + \cdots + \beta_1 W + \beta_2 X_1 W + \beta_3 X_1^2 W + \cdots + \varepsilon$$
- Difficult to estimate, so formulate as a problem of variable selection
- LASSO:
$$\min_{\alpha} \|y - X\alpha\|_2^2 + \lambda\|\alpha\|_1 \text{ for } \alpha = (\beta, \gamma)$$
- The goals of β and γ are fundamentally different
  - β meant to capture the (heterogeneity in) treatment effect that exist
  - γ meant to explain how the outcome in the control condition behaves
- Imai and Ratkovic (2013) (R-package: FindIt)1
$$\min_{\alpha=(\beta,\gamma)} \|y - X\alpha\|_2^2 + \lambda_\beta \|\beta\|_1 + \lambda_\gamma \|\gamma\|_1$$
- See Belloni et al, 2014 for how to preform inference after sparse selection
- Issues with Regression?
  - Have to pre-specify the model (and the dimensions of heterogeneity)

TREE METHODS

- Trees offer a flexible means to estimate $Y_i = f(X_i) + \epsilon_i$
  - Where Y_i is an outcome of interest and X_i are features along which Y_i can vary
- What is the outcome of interest in HTE?

$$\begin{aligned}\tau_i &= Y_i(1) - Y_i(0)\\ &= f(X_i)\end{aligned}$$

  - Cannot learning a tree will directly because τ_i is not observed for any unit in the data
- Recall

$$\begin{aligned}\tau_{CATE}(x) &= \mathbb{E}[Y_i(1) - Y_i(0) |\, X_i = x]\\ &= \mu(1,x) - \mu(1,x)\end{aligned}$$

$$\mu(w,x) = \mathbb{E}\left[Y_i^{obs} | W_i = w, X_i = x\right]$$

TREE METHODS

$$\tau_{CATE}(x) = \mu(1,x) - \mu(1,x)$$

$$\mu(w,x) = \mathbb{E}\left[Y_i^{obs} | W_i = w, X_i = x\right]$$

- Approach I: Use 1 tree to estimate $\hat{\mu}(w,x)$, including W as feature
- Approach II: Use 2 trees to estimate $\hat{\mu}(1,x)$ and $\hat{\mu}(0,x)$ separately

$$\hat{\tau}_{CATE}(x) = \hat{\mu}(1,x) - \hat{\mu}(0,x)$$

- $Y_i^{obs} \perp\!\!\!\perp W_i$, therefore $\hat{\mu}(w,x)$ is unbiased & consistent (for A.I and A.II)
- Künzel et al. (2018) explains when each Approach is better
  - Also proposes a way to combine them to combine their strengths and avoid their weaknesses.
- Issues?
  - Neither approach is actually optimizing for fitting $\hat{\tau}_{CATE}(x)$
  - Tree may never actually split on W (A.I) and lose interpretable subpopulations (A.II)

TREE METHODS

- Athey and Imbens (2017) proposes Causal Tree
  - Directly optimize tree splitting criteria to minimize $MSE(\hat{\tau}, \tau)$
    - Proposes using separate data when splitting for "Honest Estimation"
  - Key assumption: $\mathbb{E}_{Test}[\tau_i | X_i] = \mathbb{E}_{Test}[\hat{\tau}(x_i) | X_i]$
  - Provides valid asymptotic confidence intervals
- Benefits of Trees?
  - Does not require pre-specification
- Issues with Trees?
  - Greedy, Unstable, Piece-wise constant function estimation

ENSEMBLE METHODS

- Many methods propose to simply use ensemble methods, instead of trees, for Approaches I and II
  - Ex: Bayesian Additive Regression Trees (Hill, 2012), Combing supper learnings (Grimmer, 2018)
  - Same issues still remain: not actually optimizing for fitting $\hat{\tau}_{CATE}(x)$
- Wager and Athey (2017) propose CausalForest
  - Uses random forest for honest treatment effect estimation for individual units
  - Directly optimize tree splitting criteria to minimize $MSE(\hat{\tau}, \tau)$
  - Proves asymptotic normality of random forest, providing valid confidence intervals
- Benefits?
  - Stable and smooth function estimation
- Issues?
  - No natural subpopulations or groups.

LIMITATIONS

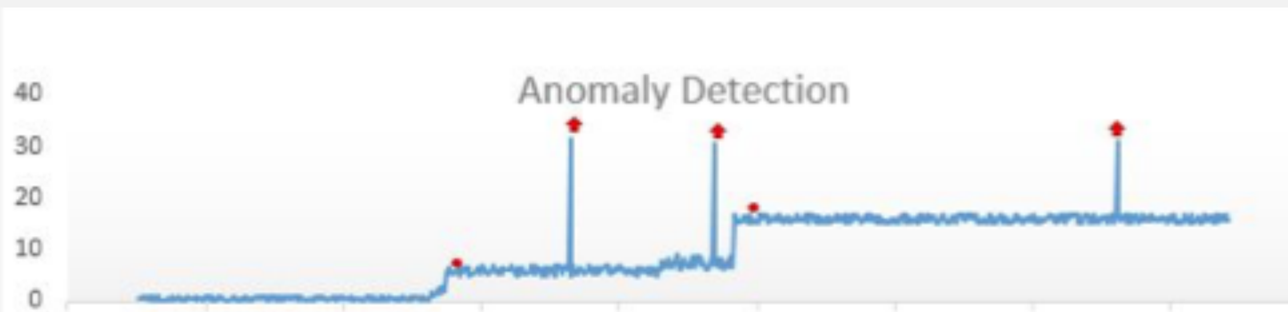
- Regression Methods
  - Pre-specification of the model
- Single Tree Methods
  - Greedy and unstable
- Ensemble Methods
  - Fairly uninterpretable/no natural subpopulations
- General Limitations
  - The mean and only the mean
    - Other moments can be effected
    - Simpsons Paradox
  - Risk minimization not effect maximization
    - Small number of subpopulations considered
    - No guarantee on their "interestingness"
  - No "discovery", only model inspection

ANOMALY DETECTION FOR HTE

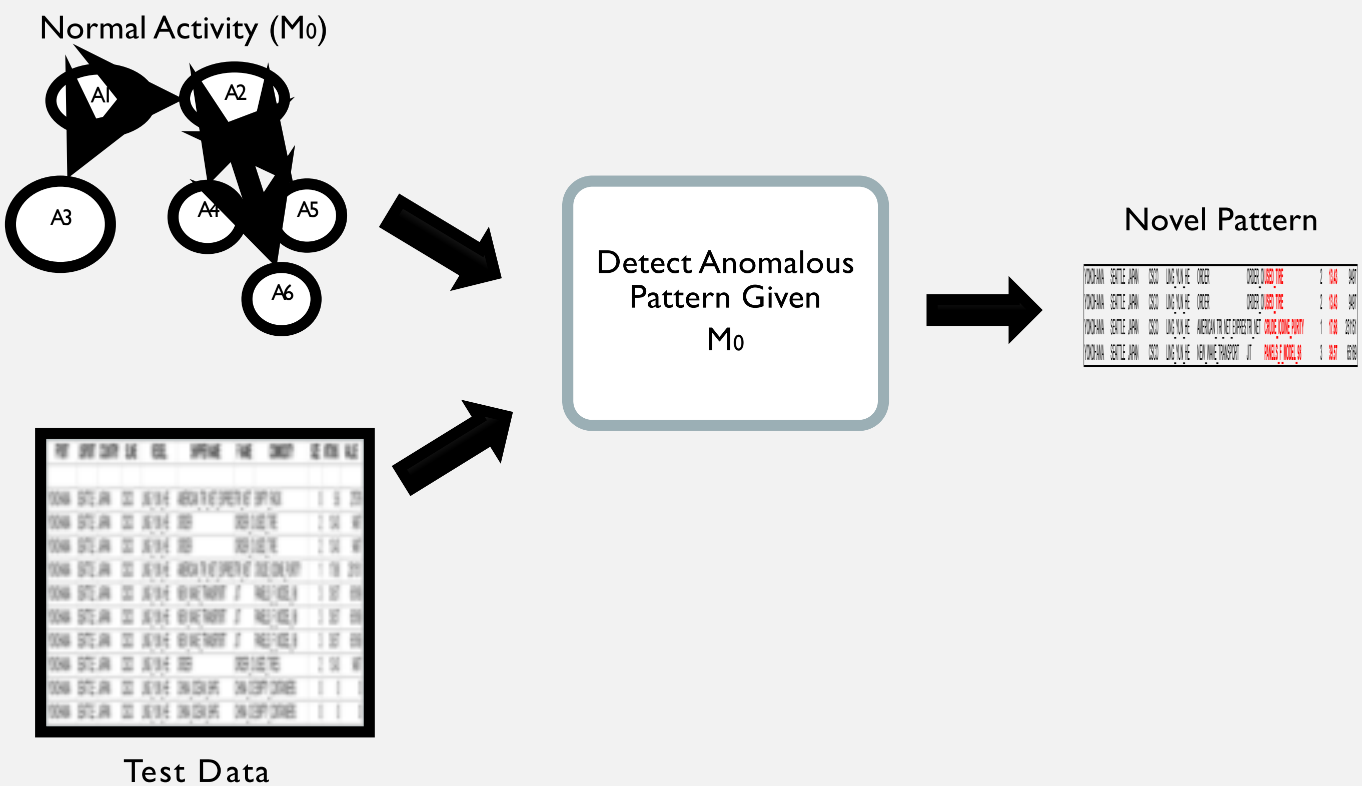
Observation III: Sometimes correlation can be (forced into) causation

ANOMALY DETECTION PARADIGM

- Identifying when a "system" deviates away from its expected behavior.

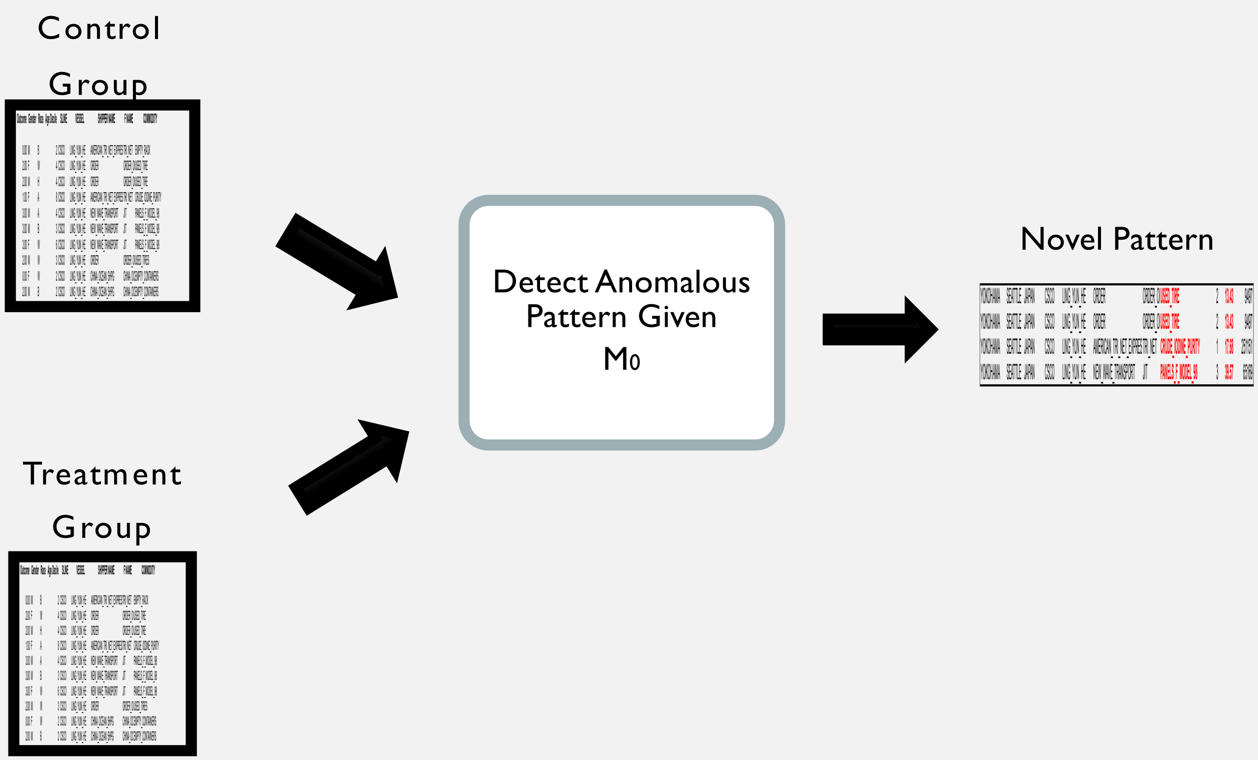


Anomalous Pattern Detection Procedure

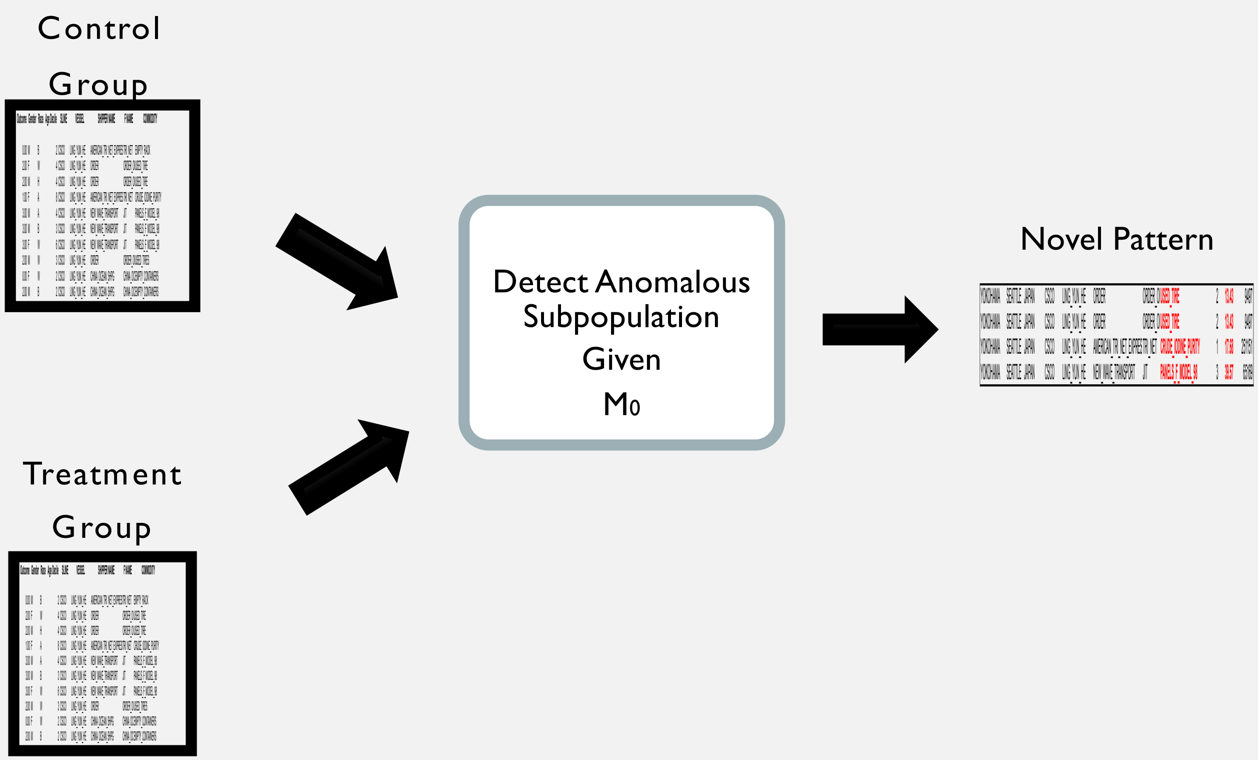


McFowland et al., 2013

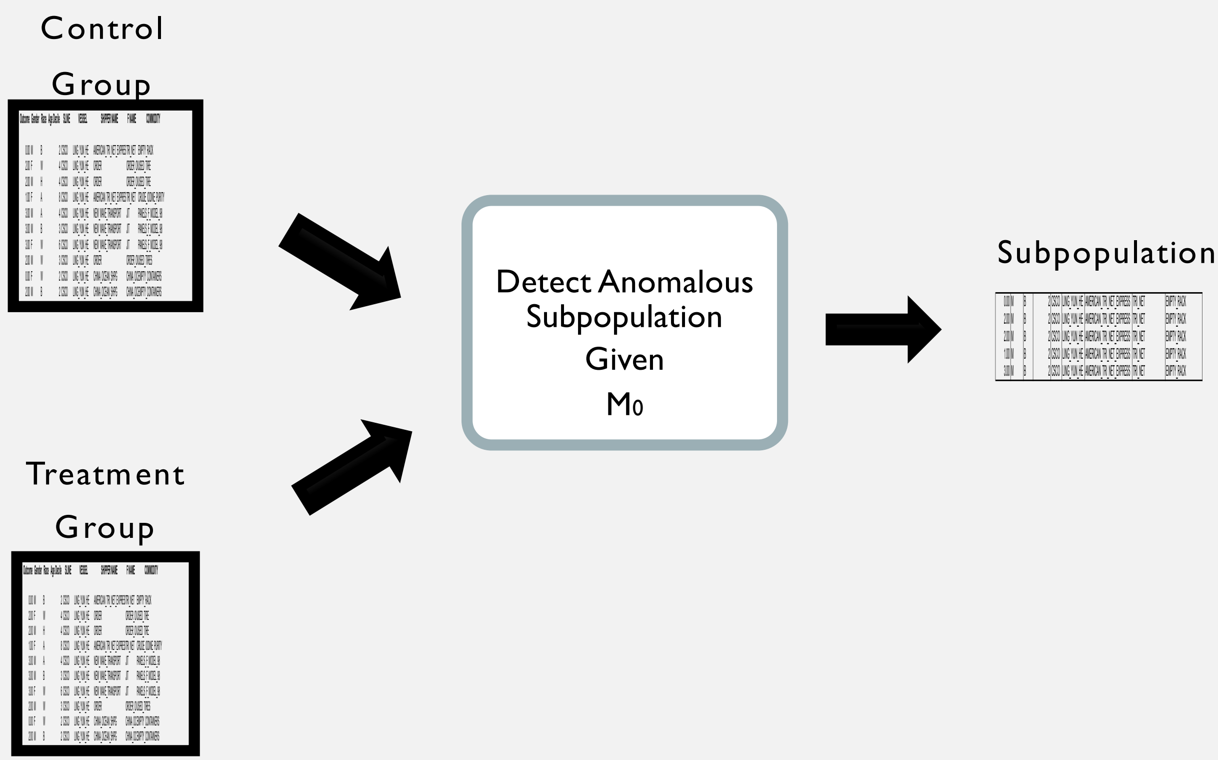
HTE Pattern Detection



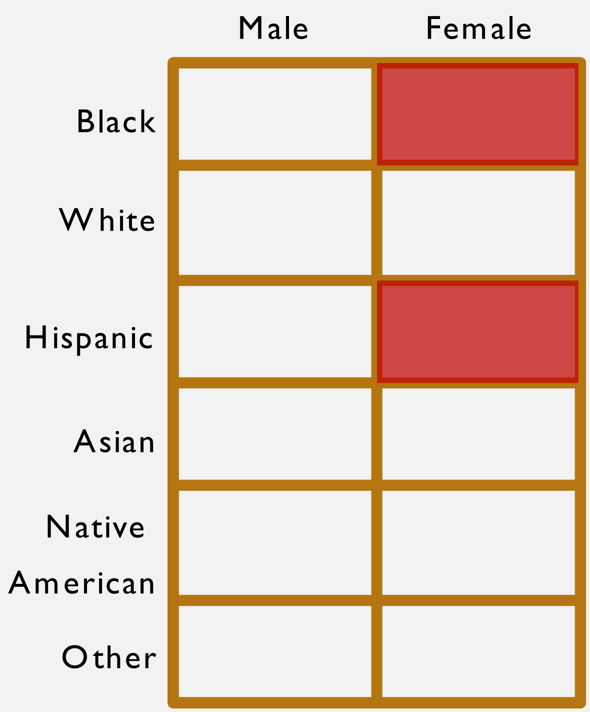
HTE Pattern Detection



HTE Pattern Detection

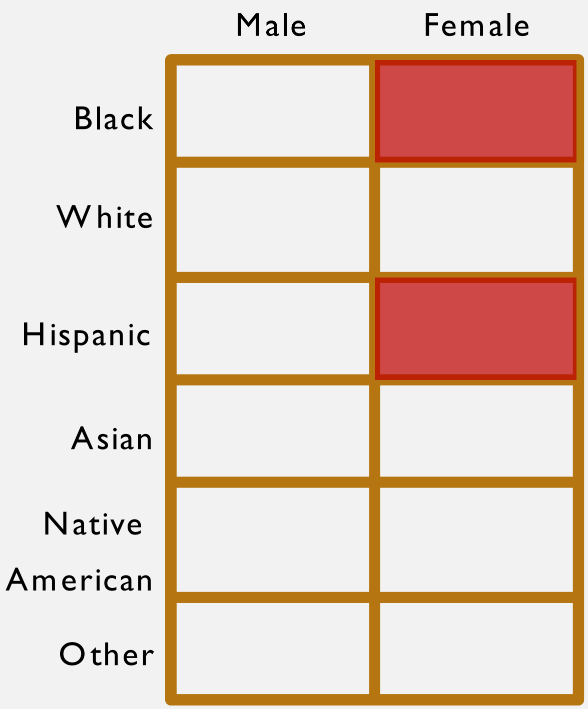


THE GOAL



Detect a subpopulation (subsets of attribute values), which correspond to anomalous outcomes for subjects in the treatment group

THE GOAL

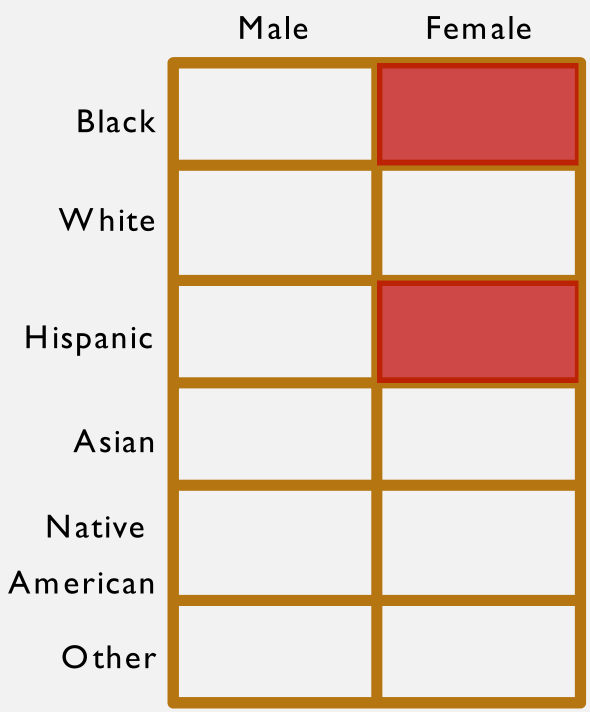


Detect a subpopulation (subsets of attribute values), which correspond to anomalous outcomes for subjects in the treatment group

<u>The Optimization</u>

$$s_1 \subseteq \{a_1 \ldots a_t\}, \ldots, s_M \subseteq \{a_1 \ldots a_p\}$$

THE GOAL



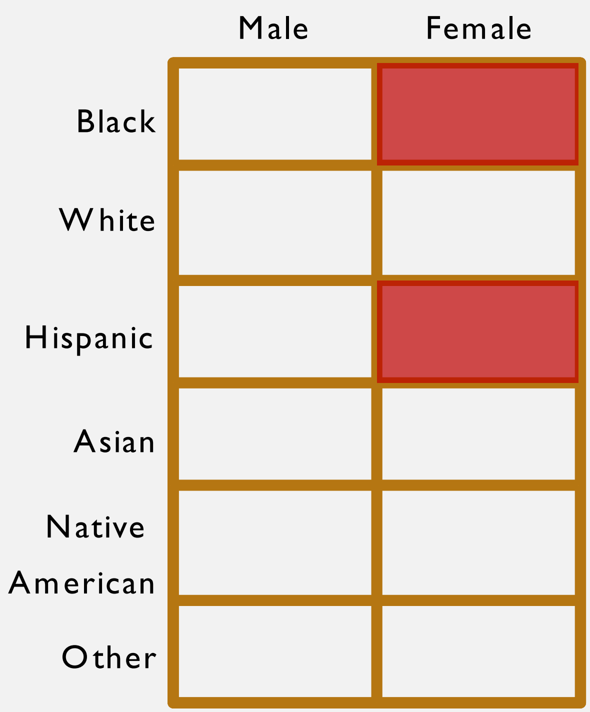
Detect a subpopulation (subsets of attribute values), which correspond to anomalous outcomes for subjects in the treatment group

The Optimization

$$s_1 \subseteq \{a_1 \ldots a_t\}, \ldots, s_M \subseteq \{a_1 \ldots a_p\}$$

$$S = s_1 \times \ldots \times s_M$$

THE GOAL



Detect a subpopulation (subsets of attribute values), which correspond to anomalous outcomes for subjects in the treatment group

The Optimization

$$s_1 \subseteq \{a_1 \ldots a_t\}, \ldots, s_M \subseteq \{a_1 \ldots a_p\}$$

$$S = s_1 \times \ldots \times s_M$$

$$S^* = \mathit{argmax}_S\, F(S)$$

Treatment Effects Subset Scan (TESS)

	Male	Female
Black	Y^{BM}	Y^{BF}
White	Y^{WM}	Y^{WF}
Hispanic	Y^{HM}	Y^{HF}
Asian	Y^{AM}	Y^{AF}
Native American	Y^{NM}	Y^{NF}
Other	Y^{OM}	Y^{OF}

I. Compute the statistical anomalousness of each treatment group subject

II. Detect subpopulation that is collectively the most anomalous

McFowland et al., 2018

Treatment Effects Subset Scan (TESS)

	Male	Female
Black	P^{BM}	P^{BF}
White	P^{WM}	P^{WF}
Hispanic	P^{HM}	P^{HF}
Asian	P^{AM}	P^{AF}
Native American	P^{NM}	P^{NF}
Other	P^{OM}	P^{OF}

I. Compute the statistical anomalousness of each treatment group subject
-- This measurement will be a p-value

II. Detect subpopulation that is collectively the most anomalous
-- Many subjects with significant p-values

Treatment Effects Subset Scan (TESS)

	Male	Female
Black	Y^{BM}	Y^{BF}
White	Y^{WM}	Y^{WF}
Hispanic	Y^{HM}	Y^{HF}
Asian	Y^{AM}	Y^{AF}
Native American	Y^{NM}	Y^{NF}
Other	Y^{OM}	Y^{OF}

Control Group

1. Compute the statistical anomalousness of each treatment group subject
   1. Estimate Conditional Distribution Under H_0

Treatment Effects Subset Scan (TESS)

	Male	Female
Black	Y^{BM}	Y^{BF}
White	Y^{WM}	Y^{WF}
Hispanic	Y^{HM}	Y^{HF}
Asian	Y^{AM}	Y^{AF}
Native American	Y^{NM}	Y^{NF}
Other	Y^{OM}	Y^{OF}

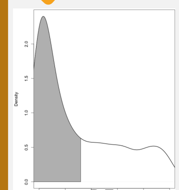
Treatment Group

1. Compute the statistical anomalousness of each treatment group subject
   1. Estimate Conditional Distribution Under H_0
   2. Compute empirical p-values

Treatment Effects Subset Scan (TESS)

	Male	Female
Black	Y^{BM}	Y^{BF}
White	Y^{WM}	Y^{WF}
Hispanic	Y^{HM}	Y^{HF}
Asian	Y^{AM}	Y^{AF}
Native American	Y^{NM}	Y^{NF}
Other	Y^{OM}	Y^{OF}

Treatment Group



y_i^{BF}

1. Compute the statistical anomalousness of each treatment group subject
   1. Estimate Conditional Distribution Under H_0
   2. Compute empirical p-values

Treatment Effects Subset Scan (TESS)

	Male	Female
Black	P^{BM}	P^{BF}
White	P^{WM}	P^{WF}
Hispanic	P^{HM}	P^{HF}
Asian	P^{AM}	P^{AF}
Native American	P^{NM}	P^{NF}
Other	P^{OM}	P^{OF}

Treatment Group

1. Compute the statistical anomalousness of each treatment group subject
   1. Estimate Conditional Distribution Under H_0
   2. Compute empirical p-values

Treatment Effects Subset Scan (TESS)

	Male	Female
Black	P^{BM}	P^{BF}
White	P^{WM}	P^{WF}
Hispanic	P^{HM}	P^{HF}
Asian	P^{AM}	P^{AF}
Native American	P^{NM}	P^{NF}
Other	P^{OM}	P^{OF}

Treatment Group

1. Compute the statistical anomalousness of each treatment group subject
   1. Estimate Conditional Distribution Under H_0
   2. Compute empirical p-values
      i. Maps each bin's distribution to the same interval

Treatment Effects Subset Scan (TESS)

	Male	Female
Black	P^{BM}	P^{BF}
White	P^{WM}	P^{WF}
Hispanic	P^{HM}	P^{HF}
Asian	P^{AM}	P^{AF}
Native American	P^{NM}	P^{NF}
Other	P^{OM}	P^{OF}

Treatment Group

1. Compute the statistical anomalousness of each treatment group subject
   1. Estimate Conditional Distribution Under H_0
   2. Compute empirical p-values
      i. Maps each bin's distribution to the same interval
      ii. P^{ij} ~ Uniform[0,1] under H_0

Treatment Effects Subset Scan (TESS)

	Male	Female
Black	P^{BM}	P^{BF}
White	P^{WM}	P^{WF}
Hispanic	P^{HM}	P^{HF}
Asian	P^{AM}	P^{AF}
Native American	P^{NM}	P^{NF}
Other	P^{OM}	P^{OF}

Treatment Group

I. Compute the statistical anomalousness of each treatment group subject
   1. Estimate Conditional Distribution Under H_0
   2. Compute empirical p-values
      i. Maps each bin's distribution to the same interval
      ii. $P^{ij} \sim$ Uniform[0,1] under H_0
      iii. For any N p-values, we expect $N*\alpha$ to be significant at level α

Treatment Effects Subset Scan (TESS)

	Male	Female
Black	P^{BM}	P^{BF}
White	P^{WM}	P^{WF}
Hispanic	P^{HM}	P^{HF}
Asian	P^{AM}	P^{AF}
Native American	P^{NM}	P^{NF}
Other	P^{OM}	P^{OF}

Treatment Group

1. Compute the statistical anomalousness of each treatment group subject
    1. Estimate Conditional Distribution Under H_0
    2. Compute empirical p-values
        i. Maps each bin's distribution to the same interval
        ii. $P^{ij} \sim$ Uniform[0,1] under H_0
        iii. For any N p-values, we expect $N^*\alpha$ to be significant at level α

Higher Criticism:

$$F(S) = \max_{\alpha} \frac{N_{\alpha} - N\alpha}{\sqrt{N\alpha(1-\alpha)}}$$

Treatment Effects Subset Scan (TESS)

	Male	Female
Black	P^{BM}	P^{BF}
White	P^{WM}	P^{WF}
Hispanic	P^{HM}	P^{HF}
Asian	P^{AM}	P^{AF}
Native American	P^{NM}	P^{NF}
Other	P^{OM}	P^{OF}

Treatment Group

I. Compute the statistical anomalousness of each treatment group subject

II. Discover subsets of attribute values that define the most anomalous outcomes

1. Maximize F(S) over all subsets of $s_1 \times \ldots \times s_M$
   - Naïve search is infeasible $O(2^{\sum|A_i|})$

Treatment Effects Subset Scan (TESS)

Nonparametric Scan Statistic (NPSS)

Have: $S \subseteq \{A_1 \times \ldots \times A_M\}$

$= \{s_1 \times \ldots \times s_M\}$

I. Compute the statistical anomalousness of each treatment group subject

II. Discover subsets of attribute values that define the most anomalous outcomes

1. Maximize F(S) over all subsets of $s_1 \times \ldots \times s_M$
   - Naïve search is infeasible $O(2^{\Sigma|A_i|})$

Treatment Effects Subset Scan (TESS)

Nonparametric Scan Statistic (NPSS)

Have: $S \subseteq \{A_1 \times \ldots \times A_M\}$
$= \{s_1 \times \ldots \times s_M\}$

Select: $F(S) = max_\alpha\ \phi(\alpha, N_\alpha(S), N(S))$

Want: $max_S\ F(S)$

Assume: $\phi \uparrow$ w.r.t N_α
$\phi \downarrow$ w.r.t N and α
ϕ is convex

I. Compute the statistical anomalousness of each treatment group subject

II. Discover subsets of attribute values that define the most anomalous outcomes

1. Maximize F(S) over all subsets of $s_1 \times \ldots \times s_M$
   - Naïve search is infeasible $O(2^{\Sigma|A_i|})$

Treatment Effects Subset Scan (TESS)

Nonparametric Scan Statistic (NPSS)

Have: $S \subseteq \{A_1 \times \dots \times A_M\}$
$= \{s_1 \times \dots \times s_M\}$

Select: $F(S) = max_\alpha\ \phi(\alpha, N_\alpha(S), N(S))$

Want: $max_S\ F(S)$

Assume: $\phi \uparrow$ w.r.t N_α
$\phi \downarrow$ w.r.t N and α
ϕ is convex

There Exist: $G(a_i)$

Such That: $\max_{s_j \subseteq \{a_1, \dots, a_t\}} F(s_j \mid A_{-j}) = \max_{i=1\dots t} F\left(\{a_{(1)} \dots a_{(t)}\} \mid A_{-j}\right)$

Only Consider: $\{a_{(1)}\}$
$\{a_{(1)}, a_{(2)}\}$
$\vdots$
$\{a_{(1)}, \dots, a_{(M)}\}$

I. Compute the statistical anomalousness of each treatment group subject
II. Discover subsets of attribute values that define the most anomalous outcomes
  1. Maximize F(S) over all subsets of $s_1 \times \dots \times s_M$
     - Naïve search is infeasible $O(2^{\Sigma|A_i|})$

Treatment **E**ffects **S**ubset **S**can (TESS)

Nonparametric Scan Statistic (NPSS)

Have: $S \subseteq \{A_1 \times \ldots \times A_M\}$
$= \{s_1 \times \ldots \times s_M\}$

Select: $F(S) = max_\alpha\, \phi(\alpha, N_\alpha(S), N(S))$

Want: $max_S\, F(S)$

Assume: $\phi \uparrow$ w.r.t N_α
$\phi \downarrow$ w.r.t N and α
ϕ is convex

There Exist: $G(a_i)$

Such That: $\max_{s_j \subseteq \{a_1, \ldots, a_t\}} F(s_j \mid A_{-j}) = \max_{i=1\ldots t} F\left(\left\{a_{(1)} \ldots a_{(t)}\right\} \mid A_{-j}\right)$

Only Consider: $\{$Black$\}$
$\{$Black, Hispanic$\}$
$\vdots$
$\{$Black, Hispanic, Asian, …, White $\}$

I. Compute the statistical anomalousness of each treatment group subject
II. Discover subsets of attribute values that define the most anomalous outcomes
   1. Maximize F(S) over all subsets of $s_1 \times \ldots \times s_M$
      - Naïve search is infeasible $O(2^{\Sigma|A_i|})$

Treatment Effects Subset Scan (TESS)

Nonparametric Scan Statistic (NPSS)

Have: $S \subseteq \{A_1 \times \ldots \times A_M\}$
$= \{s_1 \times \ldots \times s_M\}$

Select: $F(S) = max_\alpha\ \phi(\alpha, N_\alpha(S), N(S))$

Want: $max_S\ F(S)$

Assume: $\phi \uparrow$ w.r.t N_α
$\phi \downarrow$ w.r.t N and α
ϕ is convex

There Exist: $G(a_i) = \frac{1}{n(a_i)} \sum_{A_{-j}} I(p_{ij} \leq \alpha)$

Such That: $\max_{s_j \subseteq \{a_1, \ldots, a_t\}} F(s_j | A_{-j}) = \max_{i=1\ldots t} F\left(\{a_{(1)} \ldots a_{(t)}\} | A_{-j}\right)$

Only Consider: $\{$Black$\}$
$\{$Black, Hispanic$\}$
$\vdots$
$\{$Black, Hispanic, Asian, …, White $\}$

I. Compute the statistical anomalousness of each treatment group subject

II. Discover subsets of attribute values that define the most anomalous outcomes

1. Maximize F(S) over all subsets of $s_1 \times \ldots \times s_M$
   - Naïve search is infeasible $O(2^{\sum|A_i|})$

Treatment Effects Subset Scan (TESS)

Nonparametric Scan Statistic (NPSS)

Have: $S \subseteq \{A_1 \times \ldots \times A_M\}$
$= \{s_1 \times \ldots \times s_M\}$

Select: $F(S) = \mathit{max}_\alpha\ \phi(\alpha, N_\alpha(S), N(S))$

Want: $\mathit{max}_S\ F(S)$

Assume: $\phi \uparrow$ w.r.t N_α
$\phi \downarrow$ w.r.t N and α
ϕ is convex

There Exist: $G(a_i) = \frac{1}{n(a_i)} \sum_{A_{-j}} I(p_{ij} \leq \alpha)$

Such That: $\max_{s_j \subseteq \{a_1, \ldots, a_t\}} F(s_j \mid A_{-j}) = \max_{i=1\ldots t} F\left(\{a_{(1)} \ldots a_{(t)}\} \mid A_{-j}\right)$

Intuitively: $F(\{a_{(1)}, a_{(3)}\}) \leq F(\{a_{(1)}, a_{(2)}\})$
$G(a_{(3)}) \leq G(a_{(2)})$

I. Compute the statistical anomalousness of each treatment group subject

II. Discover subsets of attribute values that define the most anomalous outcomes

1. Maximize F(S) over all subsets of $s_1 \times \ldots \times s_M$
   - Naïve search is infeasible $O(2^{\Sigma|A_i|})$

Treatment Effects Subset Scan (TESS)

Nonparametric Scan Statistic (NPSS)

Have: $S \subseteq \{A_1 \times \ldots \times A_M\}$
$= \{s_1 \times \ldots \times s_M\}$

Select: $F(S) = max_\alpha\, \phi(\alpha, N_\alpha(S), N(S))$

Want: $max_S\, F(S)$

Assume: $\phi \uparrow$ w.r.t N_α
$\phi \downarrow$ w.r.t N and α
ϕ is convex

There Exist: $G(a_i) = \frac{1}{n(a_i)} \sum_{A_{-j}} I(p_{ij} \leq \alpha)$

Such That: $\max_{s_j \subseteq \{a_1, \ldots, a_t\}} F(s_j | A_{-j}) = \max_{i=1\ldots t} F\left(\{a_{(1)} \ldots a_{(t)}\} | A_{-j}\right)$

Intuitively: $F(\{a_{(1)}, a_{(3)}\}) \leq F(\{a_{(1)}, a_{(2)}\})$
$G(a_{(3)}) \leq G(a_{(2)})$

I. Compute the statistical anomalousness of each treatment group subject

II. Discover subsets of attribute values that define the most anomalous outcomes

1. Maximize F(S) over all subsets of $s_1 \times \ldots \times s_M$
   - Naïve search is infeasible $O(2^{\Sigma|A_i|})$

Higher Criticism:

$$F(S) = \max_\alpha \frac{N_\alpha - N\alpha}{\sqrt{N\alpha(1-\alpha)}}$$

Treatment Effects Subset Scan (TESS)

Nonparametric Scan Statistic (NPSS)

Have: $S \subseteq \{A_1 \times \ldots \times A_M\}$
$= \{s_1 \times \ldots \times s_M\}$

Select: $F(S) = max_\alpha\ \phi(\alpha, N_\alpha(S), N(S))$

Want: $max_S\ F(S)$

Assume: $\phi \uparrow$ w.r.t N_α
$\phi \downarrow$ w.r.t N and α
ϕ is convex

There Exist: $G(a_i) = \frac{1}{n(a_i)} \sum_{A_{-j}} I(p_{ij} \leq \alpha)$

Such That: $\max_{s_j \subseteq \{a_1, \ldots, a_t\}} F(s_j \mid A_{-j}) = \max_{i=1\ldots t} F\left(\{a_{(1)} \ldots a_{(t)}\} \mid A_{-j}\right)$

Intuitively: $F(\{a_{(1)}, a_{(3)}\}) \leq F(\{a_{(1)}, a_{(2)}\})$
$G(a_{(3)}) \leq G(a_{(2)})$

I. Compute the statistical anomalousness of each treatment group subject

II. Discover subsets of attribute values that define the most anomalous outcomes

1. Maximize F(S) over all subsets of $s_1 \times \ldots \times s_M$
   - NPSS over an attribute in O(t log t)

Higher Criticism:

$$F(S) = \max_\alpha \frac{N_\alpha - N\alpha}{\sqrt{N\alpha(1-\alpha)}}$$

TESS Search Procedure

	Male	Female
Black	P^{BM}	P^{BF}
White	P^{WM}	P^{WF}
Hispanic	P^{HM}	P^{HF}
Asian	P^{AM}	P^{AF}
Native American	P^{NM}	P^{NF}
Other	P^{OM}	P^{OF}

Treatment Group

I. Compute the statistical anomalousness of each treatment group subject
   1. Estimate Conditional Distribution Under H_0
   2. Compute empirical p-values

II. Discover subsets of attribute values that define the most anomalous outcomes
   1. Maximize F(S) over all subsets of $s_1 \times \ldots \times s_M$
      - NPSS over an attribute in O(t log t)

TESS Search Procedure

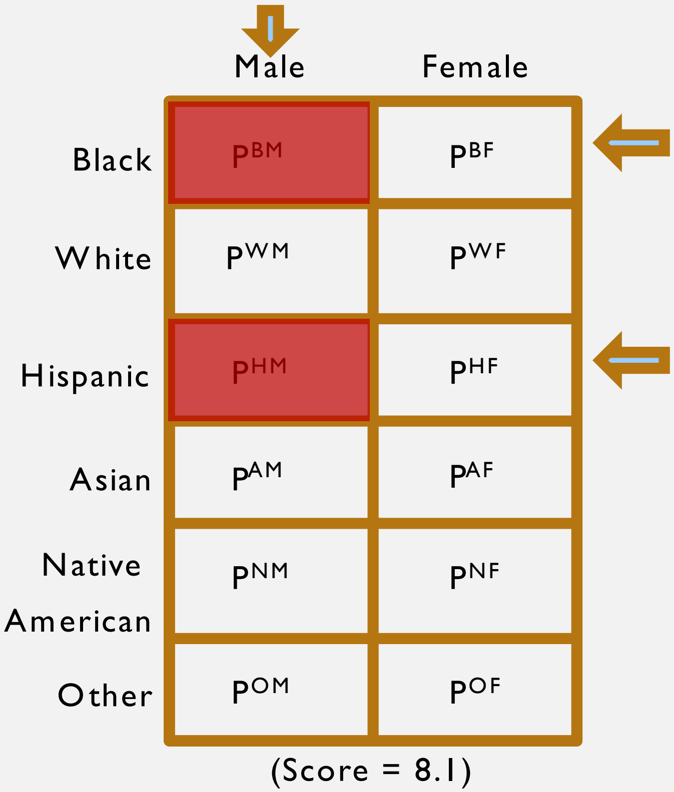
	Male	Female
Black	P^{BM}	P^{BF}
White	P^{WM}	P^{WF}
Hispanic	P^{HM}	P^{HF}
Asian	P^{AM}	P^{AF}
Native American	P^{NM}	P^{NF}
Other	P^{OM}	P^{OF}

(Score = 7.5)

I. Compute the statistical anomalousness of each treatment group subject
   1. Estimate Conditional Distribution Under H_0
   2. Compute empirical p-values

II. Discover subsets of attribute values that define the most anomalous outcomes
   1. Maximize F(S) over all subsets of $s_1 \times \ldots \times s_M$
      - NPSS over an attribute in O(t log t)

TESS Search Procedure



I. Compute the statistical anomalousness of each treatment group subject

1. Estimate Conditional Distribution Under H_0
2. Compute empirical p-values

II. Discover subsets of attribute values that define the most anomalous outcomes

1. Maximize F(S) over all subsets of $s_1 \times \ldots \times s_M$
   - NPSS over an attribute in O(t log t)

TESS Search Procedure

	Male	Female
Black	P^{BM}	P^{BF}
White	P^{WM}	P^{WF}
Hispanic	P^{HM}	P^{HF}
Asian	P^{AM}	P^{AF}
Native American	P^{NM}	P^{NF}
Other	P^{OM}	P^{OF}

(Score = 9.3)

I. Compute the statistical anomalousness of each treatment group subject

1. Estimate Conditional Distribution Under H_0
2. Compute empirical p-values

II. Discover subsets of attribute values that define the most anomalous outcomes

1. Maximize F(S) over all subsets of $s_1 \times \ldots \times s_M$
   - NPSS over an attribute in $O(t \log t)$

TESS Search Procedure

	Male	Female
Black	P^{BM}	P^{BF}
White	P^{WM}	P^{WF}
Hispanic	P^{HM}	P^{HF}
Asian	P^{AM}	P^{AF}
Native American	P^{NM}	P^{NF}
Other	P^{OM}	P^{OF}

(Score = 9.3)

I. Compute the statistical anomalousness of each treatment group subject
   1. Estimate Conditional Distribution Under H_0
   2. Compute empirical p-values

II. Discover subsets of attribute values that define the most anomalous outcomes
   1. Maximize F(S) over all subsets of $s_1 \times \ldots \times s_M$
      - NPSS over an attribute in O(t log t)

<u>Significance of our subpopulation</u>

Compare subpopulation score to maximum scores of simulated datasets under H_0

CASE STUDY: TENNESSEE STAR

TENNESSEE STAR ANALYSIS (1985)

- Effect of classrooms size on achievement (test scores)
- 4 year panel (kindergarten to 3rd grade)
- 6,500 students, 330 classrooms, 80 schools
  - Total of over 11,000 records
- Treatment Conditions (randomized within school)
  - Regular Size Class (20-25 students)
  - Regular Size + Aide Class (20-25 students)
  - Small Size Class (13-17 students)

TENNESSEE STAR ANALYSIS

read	math	gender	ethnicity	lunch	grade	school	experience	degree	tethnicity	schoolid
439	463	male	afam	free	kindergarten	inner-city	0	bachelor	cauc	19
448	559	male	cauc	non-free	kindergarten	rural	16	bachelor	cauc	69
431	454	male	cauc	free	kindergarten	rural	8	bachelor	cauc	5
395	423	female	afam	free	kindergarten	inner-city	17	master	cauc	16
451	500	female	cauc	non-free	kindergarten	rural	3	bachelor	afam	56
430	473	male	cauc	non-free	kindergarten	rural	13	master	cauc	38
437	468	male	cauc	non-free	kindergarten	rural	6	master	cauc	69
490	528	male	cauc	non-free	kindergarten	suburban	18	bachelor	cauc	52
439	484	male	cauc	non-free	kindergarten	suburban	13	master	cauc	54
424	459	female	cauc	free	kindergarten	rural	12	bachelor	cauc	12
437	528	female	afam	free	kindergarten	suburban	1	bachelor	afam	21
424	559	male	cauc	free	kindergarten	rural	13	bachelor	cauc	79
431	454	male	cauc	non-free	kindergarten	rural	13	master	cauc	8
451	473	male	cauc	non-free	kindergarten	rural	3	bachelor	cauc	66
421	459	female	afam	free	kindergarten	inner-city	11	bachelor	cauc	31

TENNESSEE STAR ANALYSIS

	(1)	(2)
Treatment	3.4791	-0.2909
	(2.547)	(2.277)
Sample	All 2nd Grade	All 3rd Grade
R-squared	0.000	0.000
Observations	4263	4063

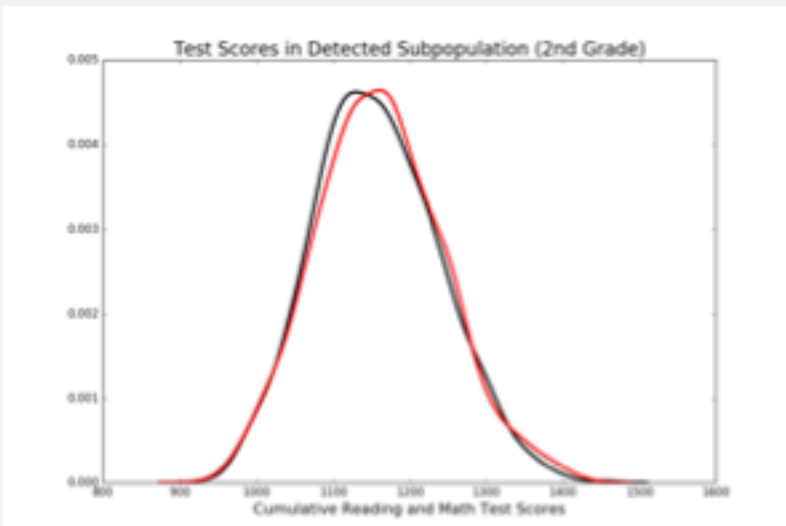
Notes: All estimates are from OLS models.

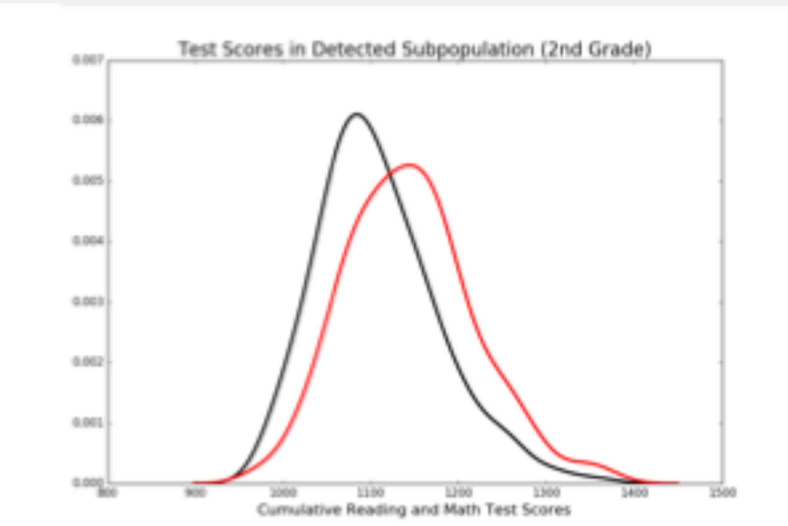
Standard errors are in parentheses.
*** p<0.01, ** p<0.05, * p<0.1

TENNESSEE STAR ANALYSIS

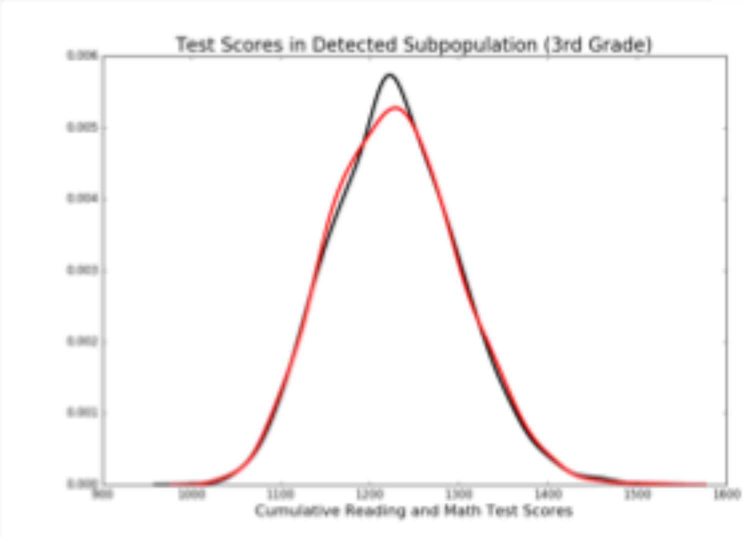
- Detected Subpopulation
  - grade:
    - 2nd or 3rd
  - school:
    - inner-city or urban
  - experience:
    - [10, infinity)
  - other features do not have differential effects

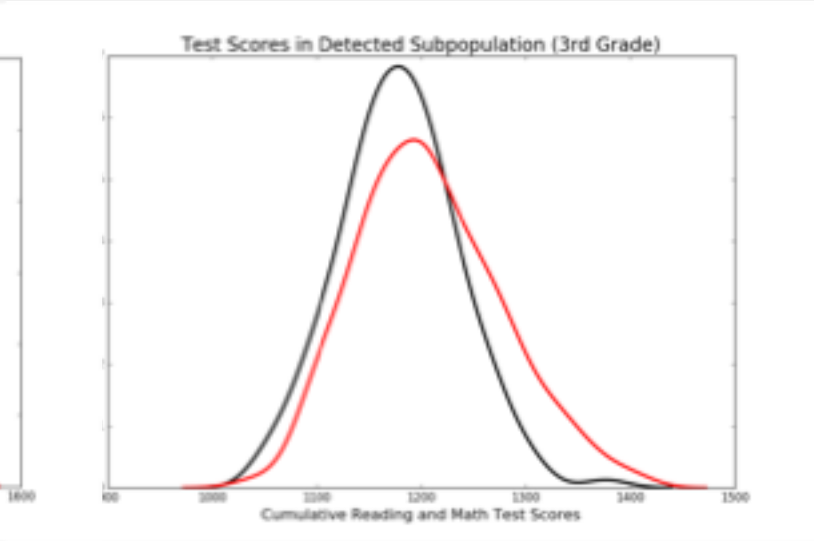
TENNESSEE STAR ANALYSIS





TENNESSEE STAR ANALYSIS





TENNESSEE STAR ANALYSIS

	(1)	(2)	(3)
Treatment	3.4791	36.066***	1.309
	(2.547)	(6.055)	(2.772)
Sample	All 2^{nd} Grade	Detected Group (2^{nd} Grade)	Undetected Group (2^{nd} Grade)
P-value	0.172	<0.001	0.0637
Observations	4263	620	3643

Notes: All estimates are from OLS models.

Standard errors are in parentheses.

*** p<0.001, ** p<0.05, * p<0.1

TENNESSEE STAR ANALYSIS

	(1)	(2)	(3)
Treatment	-0.291	18.703***	0.1
	(2.277)	(5.18)	(2.478)
Sample	All 3rd Grade	Detected Group (3rd Grade)	Undetected Group (3rd Grade)
P-value	0.898	<0.001	0.968
Observations	4063	706	3357

Notes: All estimates are from OLS models.

Standard errors are in parentheses.
*** p<0.001, ** p<0.05, * p<0.1

THE END!

REFERENCES

- Sima Sajjadiani, Aaron J. Sojourner, John D. Kammeyer-Mueller & Elton Mykerezi. (2018). Using Machine Learning to Translate Applicant Work History into Predictors of Performance and Turnover. Working Paper.
- Jean, N., Burke, M., Xie, M., Davis, W. M., Lobell, D. B., & Ermon, S. (2016). Combining satellite imagery and machine learning to predict poverty. *Science*, *353*(6301), 790–794. http://doi.org/10.1126/science.aaf7894
- Boris Babenko, Jonathan Hersh, David Newhouse, Anusha Ramakrishnan, Tom Swartz. (2017).. Poverty Mapping Using Convolutional Neural Networks Trained on High and Medium Resolution Satellite Images, With an Application in Mexico. arXiv.org.
- Blumenstock, JE, Cadamuro, G, On, R (2015). Predicting Poverty and Wealth from Mobile Phone Metadata, Science, 350(6264), 1073-1076.

OTHER PAPERS/IDEAS

- Generalized Adversarial Method of Moments
- High-dimensional regression adjustments in randomized experiments
- Lasso adjustments of treatment effect estimates in randomized experiments
- A Simple Method for Estimating Interactions Between a Treatment and a Large Number of Covariates
- A Nonparametric Bayesian Analysis of Heterogenous Treatment Effects in Digital Experimentation
- Bayesian Nonparametric Modeling for Causal Inference
- Deep IV: A Flexible Approach for Counterfactual Prediction