Propensity scores and causal analysis of observational data

Greg Ridgeway

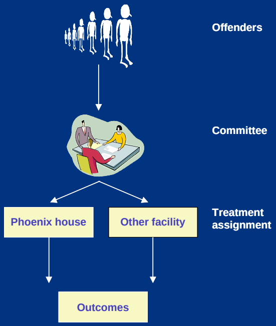
RAND Statistics Group
Santa Monica, CA
http://www.i-pensieri.com/gregr
with Dan McCaffrey, Andrew Morral, and Nelson Lim

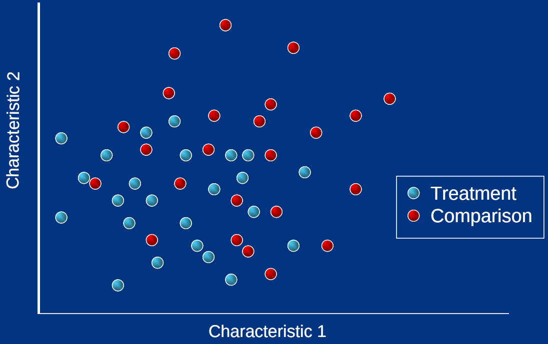
Topics of discussion

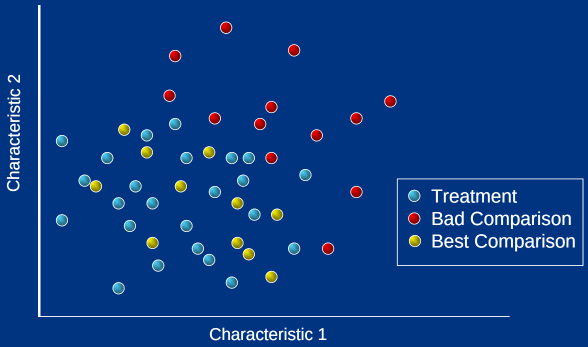
- Importance sampling and propensity scores
- Estimating propensity scores via boosted logistic regression
- Public policy examples
  - Phoenix house: Effectiveness of residential drug treatment program. Adjust treatment effect estimates for selection bias
  - Health insurance for reservists: Estimate insurance premiums reservists would be willing to pay if the DoD subsidized such a benefit

Example: Phoenix house

- The treatment assignments are non-random
- We want to estimate treatment effect
- We can reweight the individuals from the other facility to look like those from the Phoenix house




Characteristic 2
Characteristic 1
Treatment
Comparison


Characteristic 2
Characteristic 1
Treatment
Bad Comparison
Best Comparison

Estimating the causal effect of the treatment

- Each individual has a y_0 and a y_1, the outcome that would happen if they went to the control or treatment facility

$$\text{Average treatment effect of the treated} \\ = \mathrm{E}(y_1|T=1) - \mathrm{E}(y_0|T=1)$$

$$\mathrm{E}(y_1|T=1) \approx \frac{\sum_{i=T}^{N} t_i y_{1i}}{N_T}$$

$$\mathrm{E}(y_0|T=1) = \iint y_0 f(y_0, \mathbf{x}|T=1)\, d\mathbf{x}\, dy_0$$

Causal estimation

- Each individual has a y_0 and a y_1, the outcome that would happen if they went to the control or treatment facility

$$
\begin{aligned}
\mathrm{E}(y_0|T=1) &= \iint y_0 f(y_0, \mathbf{x}|T=1)\, d\mathbf{x}\, dy_0 \\
&= \iint y_0 \frac{f(y_0, \mathbf{x}|T=1)}{f(y_0, \mathbf{x}|T=0)} f(y_0, \mathbf{x}|T=0)\, d\mathbf{x}\, dy_0
\end{aligned}
$$

Causal estimation

- Each individual has a y_0 and a y_1, the outcome that would happen if they went to the control or treatment facility

$$\mathrm{E}(y_0|T=1) =$$
$$\iint y_0 \frac{f(T=1|y_0,\mathbf{x})}{f(T=0|y_0,\mathbf{x})} \frac{f(y_0,\mathbf{x})}{f(y_0,\mathbf{x})} \frac{f(T=0)}{f(T=1)} f(y_0,\mathbf{x}|T=0)\, d\mathbf{x}\, dy_0$$

- Assume $f(T|y_0,\mathbf{x}) = f(T|\mathbf{x})$

Causal estimation

- Each individual has a y_0 and a y_1, the outcome that would happen if they went to the control or treatment facility

$$\mathrm{E}(y_0|T=1) = \frac{f(T=0)}{f(T=1)} \iint y_0 \frac{p(\mathbf{x})}{1-p(\mathbf{x})} f(y_0, \mathbf{x}|T=0)\, d\mathbf{x}\, dy_0$$

$$\mathrm{E}(y_0|T=1) \quad \approx \quad \frac{\sum_{i=1}^{N} w_i(1-t_i)y_{0i}}{\sum_{i=1}^{N} w_i(1-t_i)}$$

Summary of the method

$$\mathrm{E}(y_1|T=1) \approx \frac{\sum_{i=1}^{N} t_i y_{1i}}{N_T}$$

$$\mathrm{E}(y_0|T=1) \approx \frac{\sum_{i=1}^{N} w_i(1-t_i)y_{0i}}{\sum_{i=1}^{N} w_i(1-t_i)}$$

- $w_i = \frac{p_i}{1-p_i}$, and p_i is the probability that subject i goes to the treatment group
- Derivation requires that treatment assignments depend only on $\mathbf{x}$
- $\mathbf{x}$ is high-dimensional (41) and we use the boosted logistic regression method to estimate the probabilities

Logistic log-likelihood

- Choose $p(\mathbf{x})$ to maximize

$$\mathrm{E}_{t|\mathbf{x}}\, t \log p(\mathbf{x}) + (1-t) \log(1-p(\mathbf{x}))$$

- Or on the log-odds scale, $p(\mathbf{x}) = 1/(1+e^{-F(\mathbf{x})})$, find $F(\mathbf{x})$ to maximize

$$\mathrm{E}_{t|\mathbf{x}}\, tF(\mathbf{x}) - \log\left(1 + e^{F(\mathbf{x})}\right)$$

Gradient boosting

- Initialize $F(\mathbf{x}) = 0$
- Compute the gradient of the expected log-likelihood pointwise with respect to $F(\mathbf{x})$

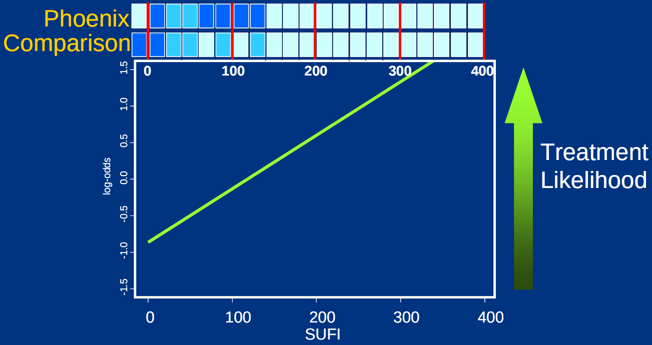
$$\frac{\partial}{\partial F(\mathbf{x})}\ell(F) = \mathrm{E}\left[t - \frac{1}{1+e^{-F(\mathbf{x})}}|\mathbf{x}\right]$$

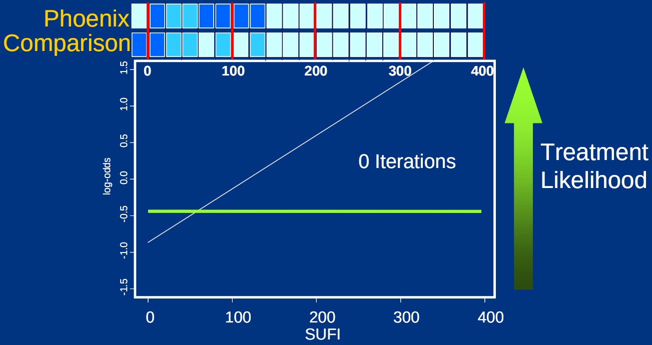
- The gradient implies that for some λ we can improve F with $F(\mathbf{x}) \leftarrow F(\mathbf{x}) + \lambda \mathrm{E}\left[t - p(\mathbf{x})|\mathbf{x}\right]$
- We will use regression trees to estimate $\mathrm{E}\left[t - p(\mathbf{x})|\mathbf{x}\right]$

Advantages

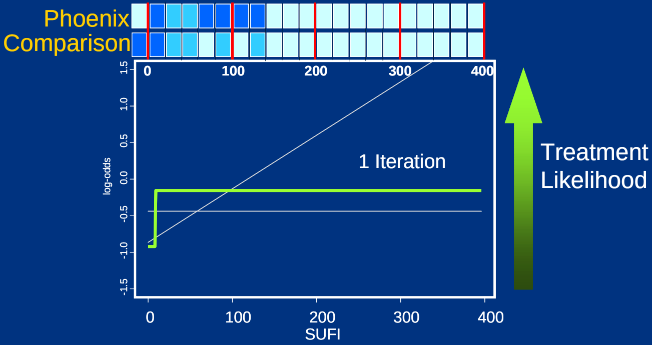
1. Boosting has a straightforward application to most prediction problems and loss functions
2. Trees handle continuous, nominal, ordinal, and missing x's
3. Invariant to one-to-one transformations of the x's
4. Model higher interaction terms with more complex regression trees
5. Use low variance models on each iteration: shrinkage, subsampling, bagging
6. Automate the selection of the number of iterations: out-of-bag estimation

Predict treatment group from abuse intensity

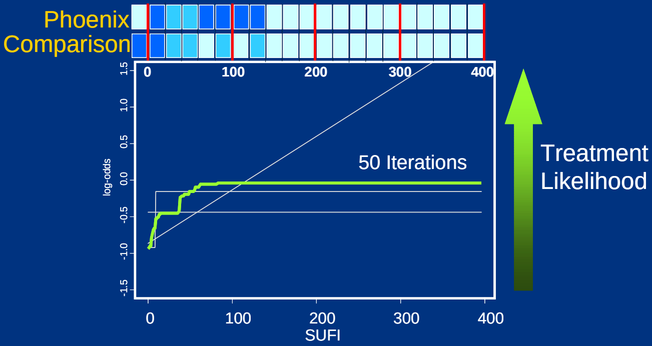




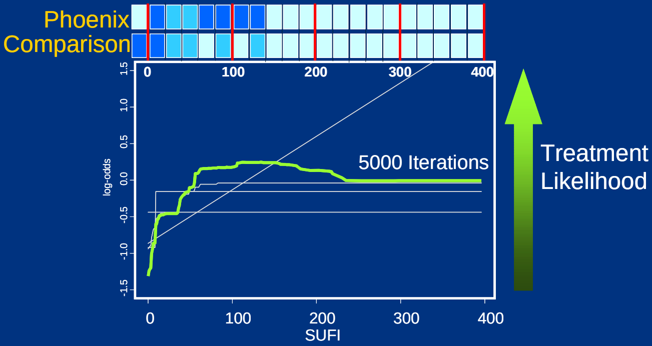
Phoenix
Comparison
0
100
200
300
400
0 Iterations
log-odds
1.5
1.0
0.5
0.0
-0.5
-1.0
-1.5
0
100
200
300
400
SUFI
Treatment
Likelihood



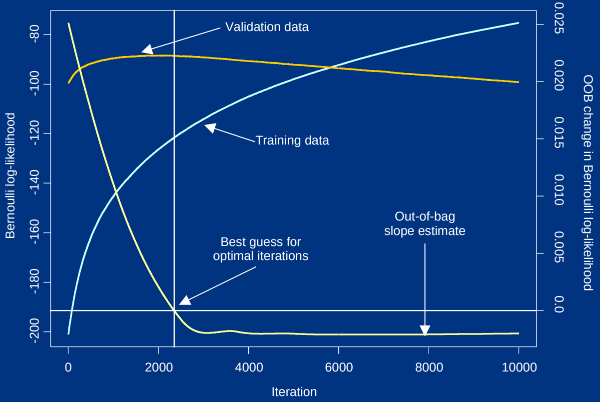
Phoenix
Comparison
0
100
200
300
400
1 Iteration
log-odds
1.5
1.0
0.5
0.0
-0.5
-1.0
-1.5
0
100
200
300
400
SUFI
Treatment
Likelihood



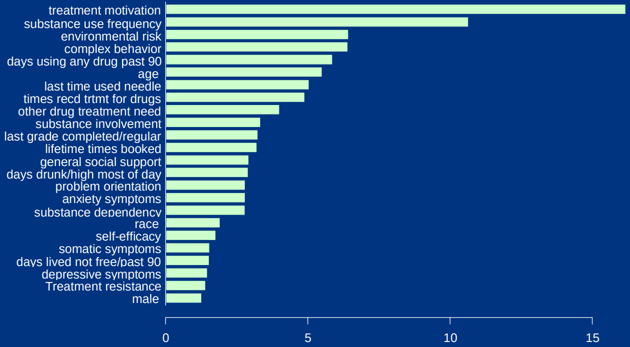
Phoenix
Comparison
0
100
200
300
400
1.5
1.0
0.5
0.0
-0.5
-1.0
-1.5
log-odds
50 Iterations
0
100
200
300
400
SUFI
Treatment
Likelihood


Phoenix
Comparison
0
100
200
300
400
5000 Iterations
log-odds
1.5
1.0
0.5
0.0
-0.5
-1.0
-1.5
0
100
200
300
400
SUFI
Treatment
Likelihood

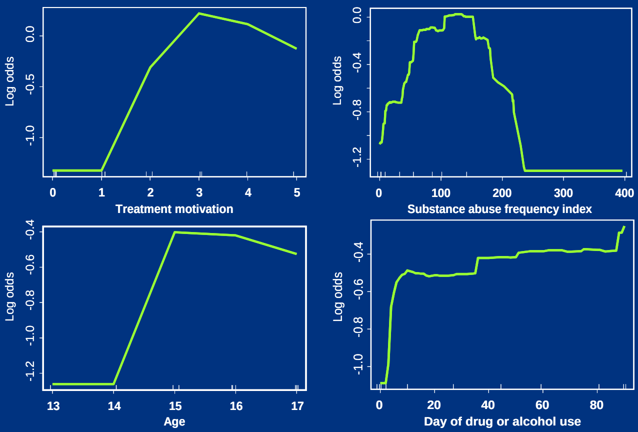
Estimating the optimal number of iterations



Relative influence



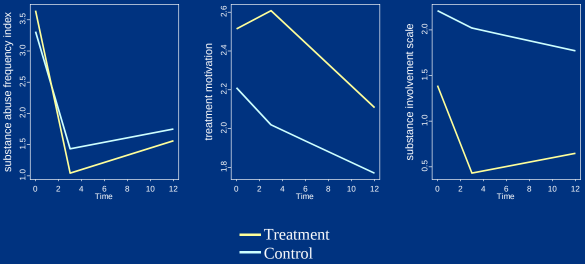
Marginal effects



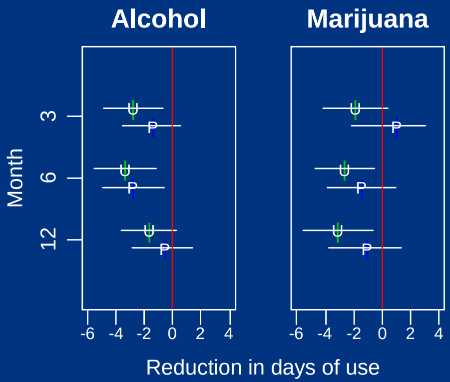
Balance of subject features

Variable	treatment mean	weighted control mean	unweighted control mean	t
treatment motivation	2.52	2.22	1.35	1.84
environmental risk	30.61	30.68	28.94	-0.07
substance abuse	76.85	67.59	43.34	1.16
complex behavior	12.84	12.77	12.11	0.07
age	15.82	15.77	15.31	0.45
l5a124	0.62	0.55	0.38	1.13
withdrawal index	2.42	2.34	2.27	0.75
days in detention	44.37	52.37	54.11	-0.74
substance problem	9.91	9.26	6.64	1.27
age of first use	12.55	12.27	11.97	1.04
ESS	175	106	274	

Results: Phoenix house



Results: Phoenix house

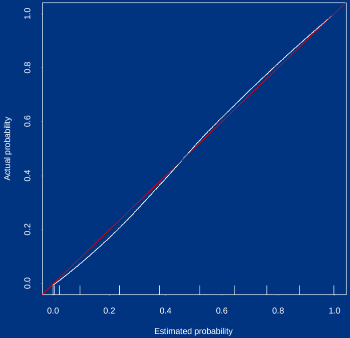


Example: Military reservist and health coverage

- Reservists often have difficulty maintaining employer sponsored health insurance
- The DoD wants to determine the price that reservists would be willing to pay if offered health coverage
- Problem: Survey of reservists did not ask how much they would be willing to pay
- Premiums paid is an item on the national health survey of the general US population (NHIS)

Calibration: Reservists example

- We use boosted logistic regression to estimate $P(\text{reservist}|\mathbf{x})$



Sanity check: Reservist example

- In both the national sample and reservist sample we observe indicators of having health insurance or not
- NHIS unweighted: 16%
- NHIS weighted: 22.3%
- Reservist sample: 21.6%

Results: Reservist example

Estimated annual premium (SD): $814 ($21)

	HMO	PPO
Married	$1233	$1230
Single	$576	$577

Summary

- Causal questions are the norm in public policy ... as is observational data
- Propensity scoring via importance sampling is a coherent framework to understand and develop propensity score methods
- Boosting methods offer flexible modeling strategies when faced with many features, features of different types, redundant features
- Public policy is a ripe area for the intersection of statistical methodology and data mining

Propensity scores and causal analysis of observational data

Greg Ridgeway

RAND Statistics Group
Santa Monica, CA
http://www.i-pensieri.com/gregr
with Dan McCaffrey, Andrew Morral, and Nelson Lim

LogitBoost: Logistic regression

1. Initialize $\hat{F}(\mathbf{x}) = \log \frac{\bar{y}}{1-\bar{y}}$
2. Let $z_i = y_i - \frac{1}{1-\exp\left(-\hat{F}(\mathbf{x}_i)\right)}$
3. Construct a tree structured predictor of z_i
4. The tree assigns each observation to a terminal node $g(T_k) = \arg\max_\lambda \sum_{i \in T_k} L(y_i, \hat{F}(\mathbf{x}_i) + \lambda)$
5. Update our guess as

$$\hat{F}(\mathbf{x}) \leftarrow \hat{F}(\mathbf{x}) + g(\mathbf{x})$$

6. Return to step (2) for M iterations

LogitBoost: Logistic regression

1. Initialize $\hat{F}(\mathbf{x}) = \log \frac{\bar{y}}{1-\bar{y}}$
2. Let $z_i = y_i - \frac{1}{1-\exp\left(-\hat{F}(\mathbf{x}_i)\right)}$
3. Construct a tree structured predictor of z_i
4. The tree assigns each observation to a terminal node $g(T_k) = \arg\max_\lambda \sum_{i \in T_k} y_i(\hat{F}(\mathbf{x}_i) + \lambda) - \log\left(1 + \exp(\hat{F}(\mathbf{x}_i) + \lambda)\right)$
5. Update our guess as

$$\hat{F}(\mathbf{x}) \leftarrow \hat{F}(\mathbf{x}) + g(\mathbf{x})$$

6. Return to step (2) for M iterations