Retooling propensity score techniques with machine learning

evaluation of a solution to the Los Angeles drug epidemic

Greg Ridgeway

with Dan McCaffrey and Andrew Morral

http://www.i-pensieri.com/gregr

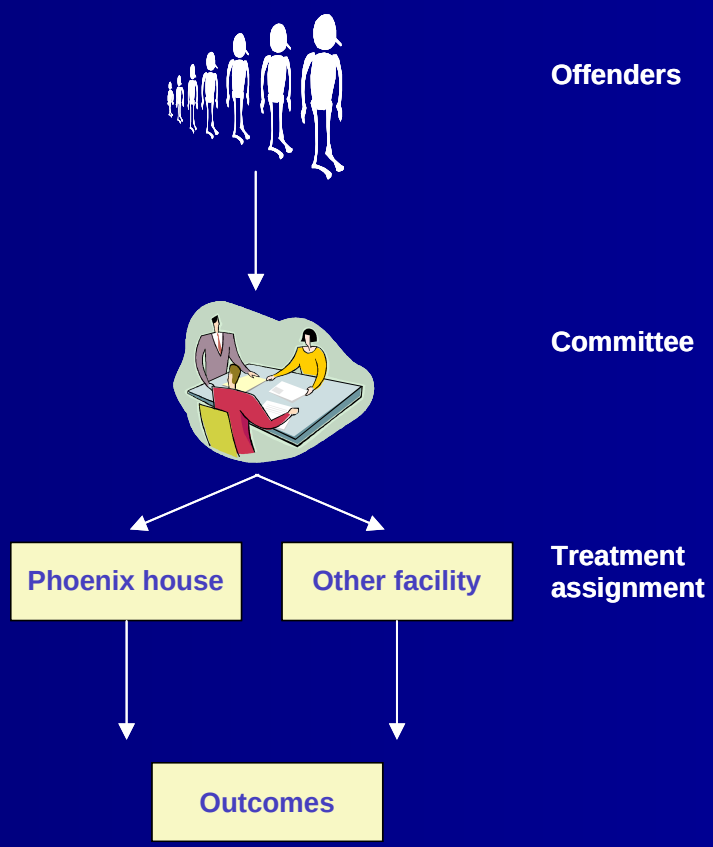
RAND Statistics Group, Santa Monica, CA

Outline

- Importance sampling and propensity scores
- Estimating propensity scores via boosted logistic regression
- Phoenix house: Effectiveness of a residential drug treatment program. Adjust treatment effect estimates for selection bias

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



Causal estimation

- Each individual has a control outcome, y_0, and a treatment outcome, y_1.

Average treatment effect of the treated

$$= \mathrm{E}(y_1|T=1) - \mathrm{E}(y_0|T=1)$$

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

Causal estimation

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

Causal estimation

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

- Apply Bayes Theorem to $f(y_0, \mathbf{x}|T)$.

Causal estimation

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

- Assume $f(T|y_0,\mathbf{x}) = f(T|\mathbf{x})$
- This the "strong ignorability asumption." If $\mathbf{x}$ contains all the information used in assigning treatments, then this assumption holds.

Causal estimation

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

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

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. This is the problem on which machine learning has focused for years.

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 $F(\mathbf{x})$

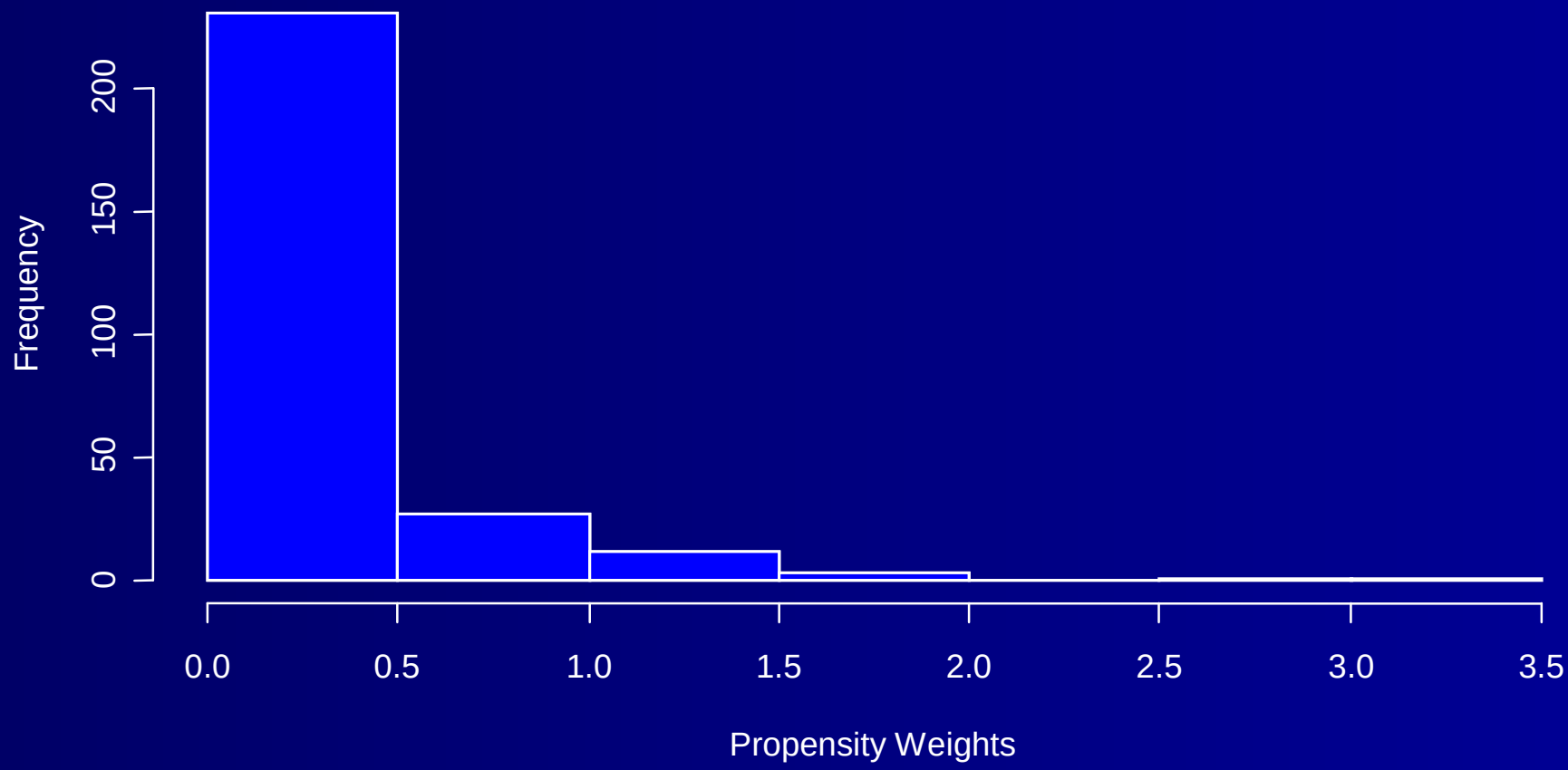
$$\frac{\partial}{\partial F(\mathbf{x})}\ell(F) = \mathrm{E}\left[t - \frac{1}{1+e^{-F(\mathbf{x})}}|\mathbf{x}\right]$$

- 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. Handles 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

Observed control group weights



Balance of subject features

Variable	treatment mean	weighted control mean	unweighted control mean	effect size weighted	effect size unweighted
Treatment motivation	2.52	2.22	1.35	0.23	0.89
Environmental risk	30.61	31.09	28.94	-0.05	0.17
Substance use	7.61	6.94	4.59	0.16	0.69
Complex behavior	12.84	13.00	12.11	-0.02	0.09
Age	15.82	15.76	15.31	0.07	0.56
⋮				⋮	⋮
ESS	175	107.5	274		
Average |ES|				**0.107**	**0.307**

Results: Phoenix house

	Unweighted	GBM	Logit, 0.05	Logit, 0.20
Estimated Treatment Effect (confidence interval)				
Marijuana	-11.8	-5.9	-1.9	-5.2
	(-19.7,-3.8)	(-16.2, 4.3)	(-12.7, 8.8)	(-24.4, 14.1)
Alcohol	-1.2	2.8	1.5	3.1
	(-5.5,3.0)	(-3.6,9.3)	(-10.2, 13.3)	(-10.5, 16.7)
Measures of model fit				
Deviance	NA	466.4	539.2	511.4
ASAM	0.31	0.11	0.14	0.20
SE, Marijuana	4.0	5.2	6.6	11.8
SE, Alcohol	2.2	3.3	7.2	8.3

Remaining questions

- Selecting the optimal number of iterations
  - Cross-validation, out-of-bag estimation, or minimize imbalance in pretreatment characteristics
  - There must be a bias/variance tradeoff but it is difficult to optimize
- Sensitivity to the independence assumption

Assessing sensitivity

- Let $G \geq 1$, for an observed weight of w the range for true weight is in $[w/G, wG]$
- Maximize $S = \sum_{i \in C} a_i w_i y_i / \sum_{i \in C} a_i w_i$ subject to $1/G < a_i < G$
- Repeat, this time minimizing S

G	Maximum	Minimum
1.24	0.00	-11.32
2.00	13.78	-20.58
3.00	23.19	-26.52
4.00	28.06	-29.87

Central epidemiological questions

- Epidemiology discerns whether various groups are at greater risk
  - detecting racial biases,
  - estimating the number of uninsured reservists,
  - assess a gun violence suppression program,
  - others?
- How can we best create suitable comparison groups?
- Outcomes can censored (death, incarceration) but the treatment exposure can influence censoring.