Strategies for the Analysis of Observational Data

Greg Ridgeway
gregr@rand.org

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

RAND Statistics Group, Santa Monica, CA

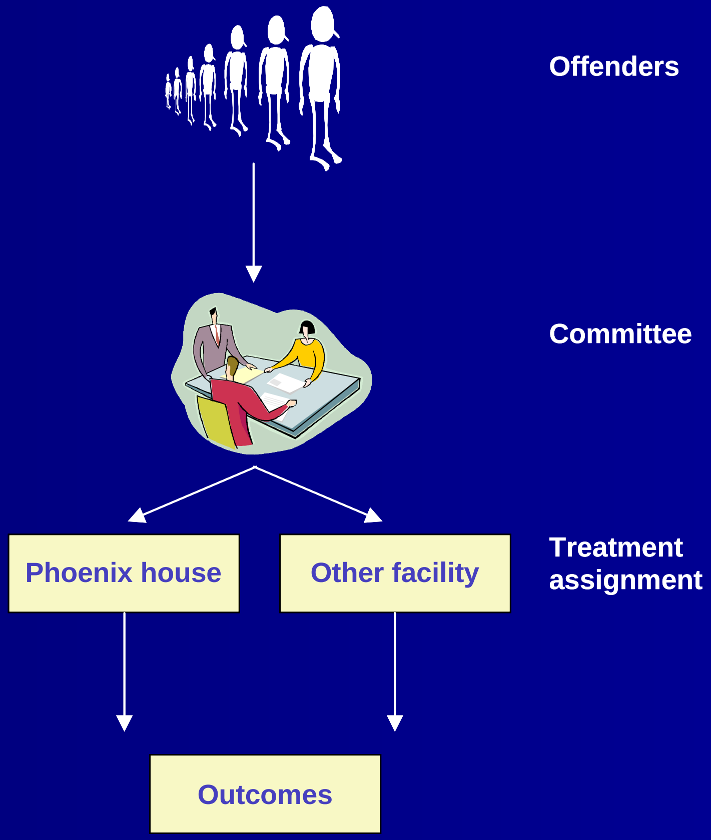
Problems I face at RAND

Assessing public policy almost always asks "what would have happened if..."

- youths sent to residential drug treatment had been sent to alternative programs
- officers treated drivers that they stopped equitably regardless of race
- military reservists were offered a DoD subsidized health plan

Example: Phoenix Academy

- The treatment assignments are non-random
- Youths in the treatment have no violent criminal history, moderate drug use
- A direct comparison ignores baseline differences



Offenders

Committee

Phoenix house

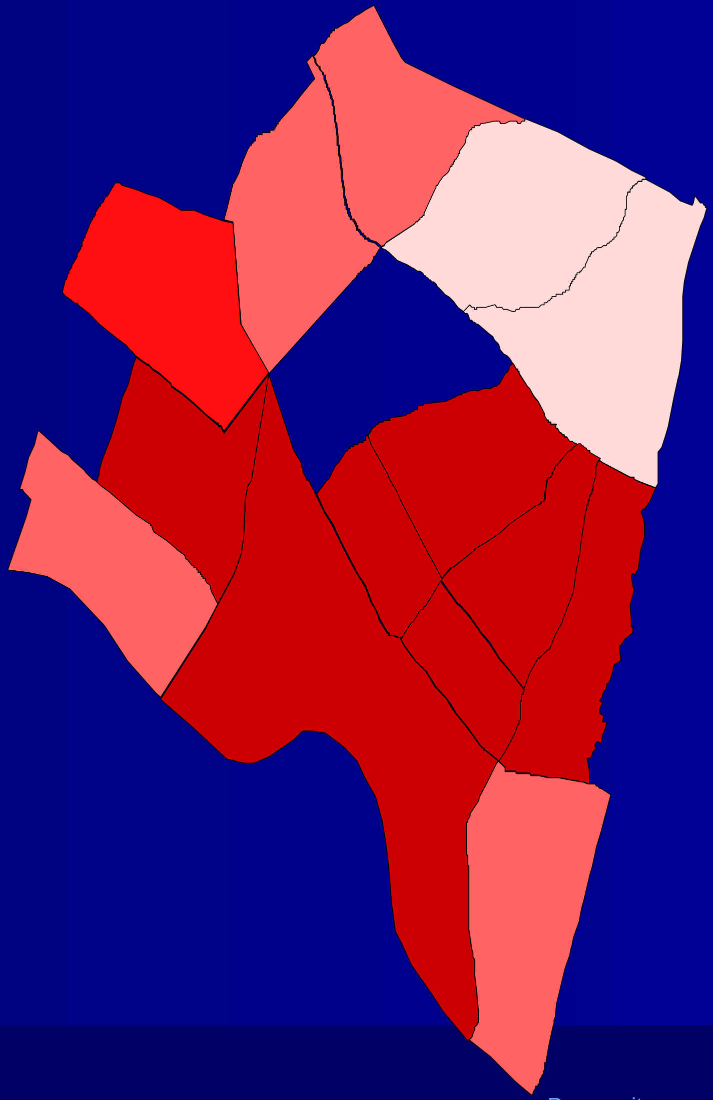
Other facility

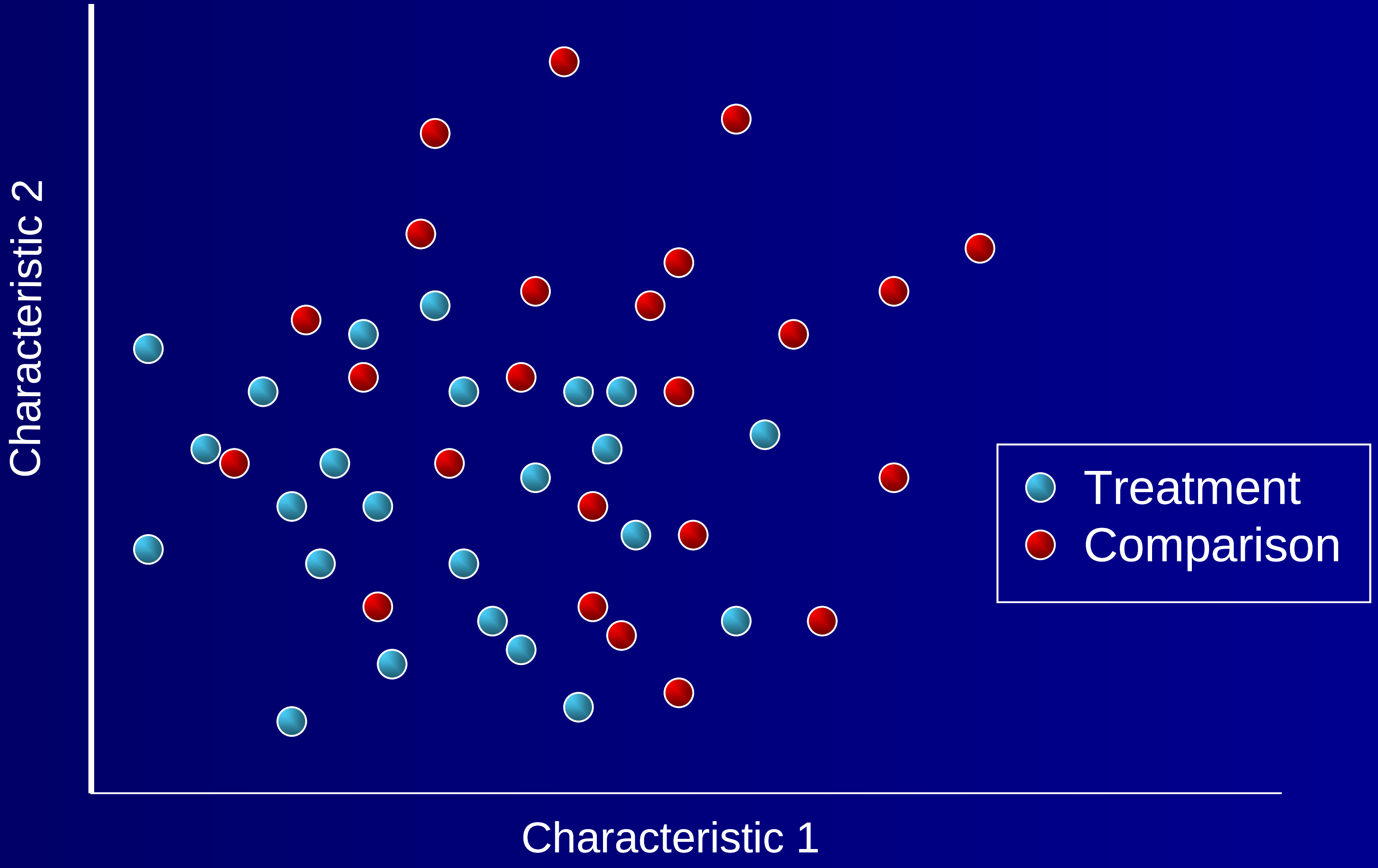
Treatment assignment

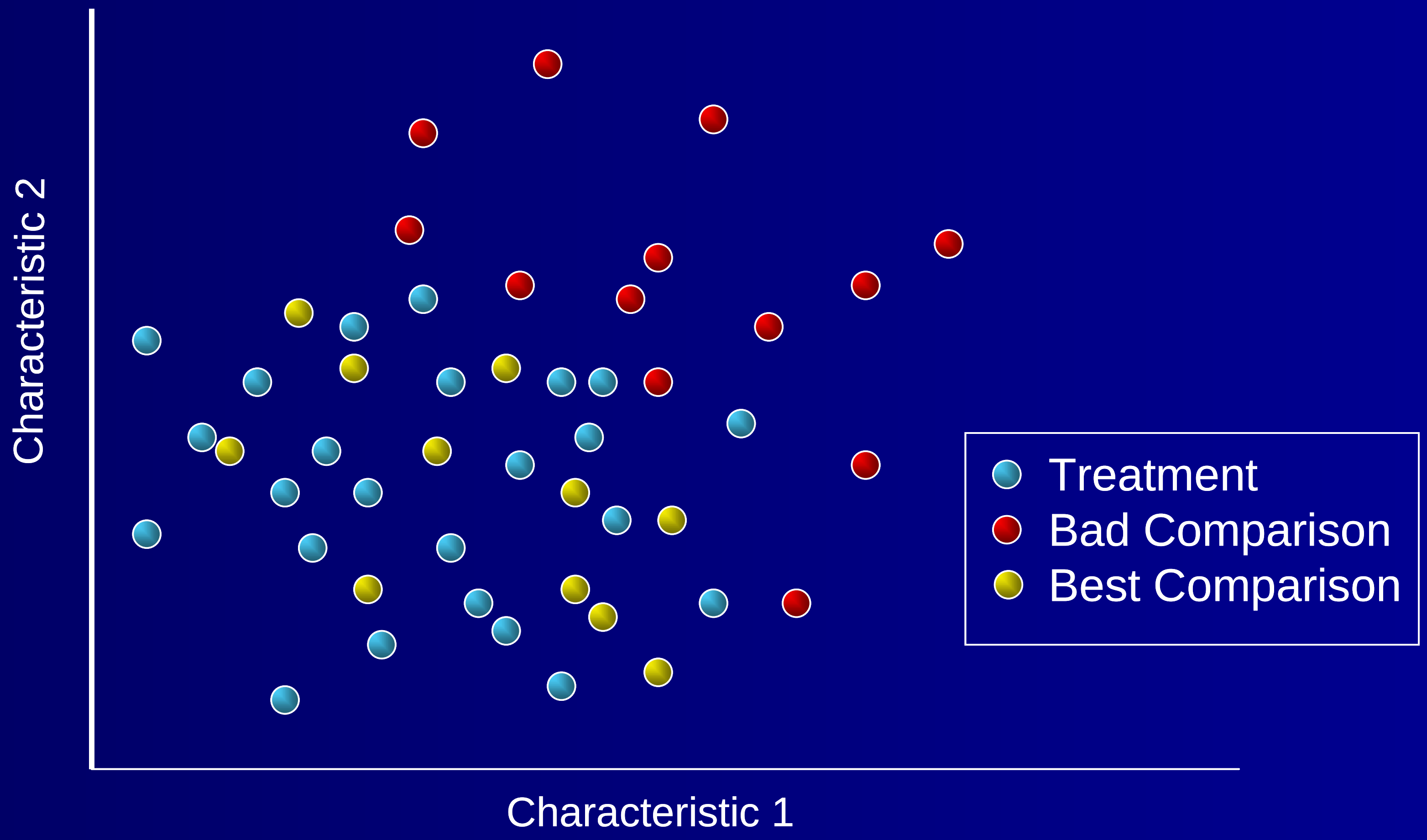
Outcomes

Example: Racially biased policing

- Drivers of different races may traverse different streets
- Policing practices can vary by neighborhood, crime patterns vary
- Direct comparisons of black drivers to white drivers ignore these differences




Characteristic 2
Characteristic 1
Treatment
Comparison


Characteristic 2
Characteristic 1
Treatment
Bad Comparison
Best Comparison

Adjusting for $\mathbf{x}$

- We must “adjust for” or “control for” $\mathbf{x}$, commonly interpreted as

$$y_{\text{obs}} = \beta_0 + \gamma T + \beta_1 x_1 + \ldots + \beta_d x_d + \epsilon$$

- γ estimates a treatment effect only if

1. the distribution of $\mathbf{x}$ is the same for the treatment and control groups **or**
2. the linear model assumptions are correct

Potential outcomes

- Each individual has a control outcome, y_0, and a treatment outcome, y_1
- Ideally we would observe both and estimate $\mathrm{E}(y_1 - y_0)$
- Instead consider

$$\mathrm{E}(y_1|T = 1) - \mathrm{E}(y_0|T = 0)$$

If the groups differ with respect to $\mathbf{x}$, we cannot determine whether differences are attributable to the treatment or $\mathbf{x}$.

Phoenix house example

Variable	treatment mean	control mean	effect size
Treatment motivation	2.52	1.35	0.89
Environmental risk	30.61	28.94	0.17
Substance use	7.61	4.59	0.69
Complex behavior	12.84	12.11	0.09
Age	15.82	15.31	0.56
⋮			⋮
N	175	274	
Average |ES|			**0.307**

- The groups differ on motivation and pre-treatment substance use
- Treatment effect or motivation effect?

Confounding

- A variable, x, is a confounder if it is related to both (y_1, y_0) and T
- In randomized studies confounders do not exist except by chance
- If the treatment and control groups differ by $\mathbf{x}$

  $$\mathrm{E}(y_1|\mathbf{x}, T = 1) - \mathrm{E}(y_0|\mathbf{x}, T = 0)$$

  and average over the distribution of $\mathbf{x}$
- If $\mathbf{x}$ is 1-3 variables, stratify by $\mathbf{x}$ and compute within strata treatment effects

Regression adjustment

- With lots of covariates we tend to use

$$y_{\text{obs}} = \beta_0 + \gamma T + \beta_1 x_1 + \ldots + \beta_d x_d + \epsilon$$

- This works fine if

$$\begin{aligned} y_0 &= \beta_0 + \beta_1 x_1 + \ldots + \beta_d x_d + \epsilon \\ y_1 &= \beta_0 + \gamma + \beta_1 x_1 + \ldots + \beta_d x_d + \epsilon \end{aligned}$$

- This method is terribly non-robust to model misspecification
- With many covariates, estimates can be unstable

Regression adjustment

- Fisher developed such analysis to adjust for *chance discrepancies* in the treatment and control groups
- When the two groups have little overlap in terms of $\mathbf{x}$, the model assumptions completely drive the result. This situation is difficult to detect
- One strategy is to fit more flexible models: splines, decision trees, kernel regression

Balance on $\mathbf{x}$

- Idea: reweight so that the distribution of the control group’s features matches the treatment group’s features

$$\begin{aligned} f(\mathbf{x}|T=1) &\propto w(\mathbf{x})f(\mathbf{x}|T=0) \\ w(\mathbf{x}) &\propto \frac{p(\mathbf{x})}{1-p(\mathbf{x})} \end{aligned}$$

- Weighting comparison subjects with $p/(1-p)$ replicates the effect of randomization.

Propensity score estimation

- $p(\mathbf{x})$ is known as the propensity score
- If T is independent of y_1 given $\mathbf{x}$ then the reweighting will yield the correct treatment effect

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

Summary of the method

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

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

- $w_i = \frac{p_i}{1-p_i}$, and p_i is the probability that subject i goes to the treatment group
- Need to estimate $p(\mathbf{x})$

Logistic regression

- Model the log-odds $\log \frac{p(\mathbf{x})}{1-p(\mathbf{x})} = f(\mathbf{x})$
- Often $f(\mathbf{x})$ is set to be linear, i.e. linear logistic regression
- Seems to just shift the problem to an earlier modeling stage
- Suggest generalized boosted models (GBM)

Advantages of GBM

1. Excellent estimation of $p(\mathbf{x})$
2. The resulting model 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. Implemented in R in the `gbm` library

Balance of subject features

		weighted	unweighted		
	treatment	control	control	effect size	
Variable	mean	mean	mean	weighted	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**

$$\mathrm{ESS} = \left(\sum w_i\right)^2 / \sum w_i^2$$

Weighting balances the groups

- Now that the weighted data looks like a randomized study
  - analyses involve simple comparisons of means and percentages
  - any remaining discrepancies can be convariate adjusted

Results: Phoenix house

	Unweighted	GBM
Estimated Treatment Effect (confidence interval)		
Marijuana	-11.8	-5.9
	(-19.7, -3.8)	(-16.2, 4.3)
Alcohol	-1.2	2.8
	(-5.5, 3.0)	(-3.6, 9.3)

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)

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

Balance of driver features

	% Black drivers	% Non-black drivers	% Non-black drivers
		(weighted)	(unweighted)
	N=3,703	ESS=2,089	N=3,033
Region			
A	31%		27%
B	32%		14%
C	1%		3%
D	11%		21%
E	9%		8%
F	3%		6%
G	14%		21%

Balance of driver features

	% Black drivers	% Non-black drivers	% Non-black drivers
		(weighted)	(unweighted)
	N=3,703	ESS=2,089	N=3,033
Region			
A	31%	30%	27%
B	32%	33%	14%
C	1%	1%	3%
D	11%	12%	21%
E	9%	9%	8%
F	3%	3%	6%
G	14%	14%	21%

Balance of driver features

	% Black drivers N=3,703	% Non-black drivers (weighted) ESS=2,089	% Non-black drivers (unweighted) N=3,033
Time			
12am-4am	16%	16%	7%
4am-8am	4%	4%	4%
8am-12pm	17%	17%	21%
12pm-4pm	20%	20%	28%
4pm-8pm	24%	24%	26%
8pm-12am	20%	21%	13%
Age			
Under 18	3%	3%	3%
18-29	47%	48%	38%
30-39	22%	22%	26%
40+	28%	27%	33%

Stop outcomes

	Black drivers	Non-black drivers	
		weighted	unweighted
Citation rate	68%	70%	79%
0-9 minutes	47%	53%	66%
Pat search	2.6%	2.9%	1.9%
Consent search	2.2%	1.7%	0.9%
Probable cause	3.2%	1.4%	1.0%

Remaining issues

- The bias/variance tradeoff is difficult to optimize. Aggressively trying to balance on subject features costs power
- Subject features associated with group assignment but not outcomes can greatly increase variance without offering any reduction in bias
- Detecting insufficient overlap between the groups is fairly easy using ESS or histograms of estimated propensity scores
- There still may be other unobserved confounders

Conclusions

- Abandon linear models for non-randomized studies. They do not give you what you think they give you
- Reweight so that the data look like a randomized study
- The reweighting can be challenging but it is easy to diagnose
- The final analysis is trivial to calculate and explain