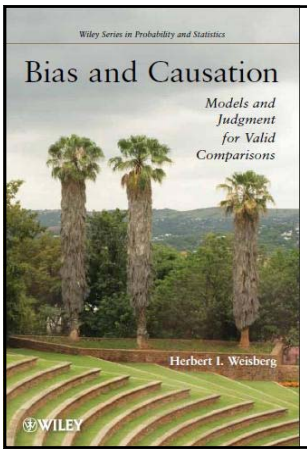


Selection Bias and Predictive Modeling

A Causal Perspective

Herbert I. Weisberg, Ph.D.
Correlation Research, Inc.

Typical Insurance Applications

Target:	Action:	Success:
Claims	Investigate	Reduce payment
Applicants	Guidelines	Reject bad risk
Prospects	Solicit	Acquire prospect
Policyholders	Audit	Increase premium
Policyholders	Service	Prevent attrition

Typical Approach

1. Measure *outcomes* in a sample of population
2. Build model to predict outcome *value* or *probability*
3. *Score* and *rank* individuals in sample
4. Select *cutoff* as criterion for *selection*
5. Conduct RCT to evaluate improvement

Predictive Accuracy

		Outcome: Good	Outcome: Bad	
Predict	Bad	100	400	500
	Good	1400	600	2000

Sensitivity = 400/1000 = .40
Specificity = 1400/1500 = .93
Positive Predictive Value (PPV) = 400/500 = .80
Negative Predictive Value (NPV) = 1400/2000 = .70

Causal Effect in Selected Subset

		Outcome: Good	Outcome: Bad	
Condition:	T	400	100	500
	C	100	400	500

$RD = .2 - .8 = -.6$

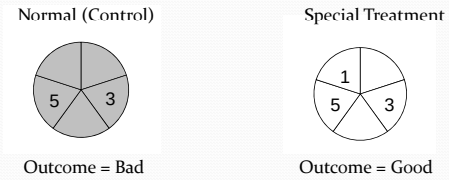
Problem with Usual Approach

- Goal: Maximize improvement
- Ideal: Select only those who would change (counterfactual)

BUT:
- Model: Targets those normally "Bad" but not necessarily correctable
- Accuracy: Measure (e.g., sensitivity) used is not necessarily appropriate

THEREFORE:
- Selected Subset based on Model may not be optimal (Selection Bias)
- Causal effect may improve performance very little

Underlying Causal Model



Response Patterns

	Treated	"Control"
Doomed:	Bad	Bad
Causal:	Bad	Good
Preventive:	Good	Bad
Immune:	Good	Good

Distributions of Response Patterns

Response Pattern	Selected	Unselected
Doomed	p_1	q_1
Causal	p_2	q_2
Preventive	p_3	q_3
Immune	p_4	q_4

Example

Response Pattern:	Selected	Unselected
Doomed	100	200
Causal	0	0
Preventive	300	400
Immune	100	1400
Total	500	2000

Causal Effect in Selected Subset

		Outcome: Good	Outcome: Bad	
Condition:	T	400	100	500
	C	100	400	500

$RD = .2 - .8 = -.6$

Causal Effect in Selected Subset

		Outcome: Good	Outcome: Bad	
Condition	T	$N_T(p_3 + p_4)$	$N_T(p_1 + p_2)$	N_T
	C	$N_C(p_2 + p_4)$	$N_C(p_1 + p_3)$	N_C

$RD = p_2 - p_3$

Causal Effect in Selected Subset
(No "Causals")

Condition		Outcome: Good	Outcome: Bad	
	T	$N_T(p_3 + p_4)$	$N_T p_1$	N_T
	C	$N_C p_4$	$N_C(p_1 + p_3)$	N_C

$$RD = -p_3$$

Example

Response Pattern:	**Selected**	**Unselected**
Doomed	300	200
Causal	0	0
Preventive	100	400
Immune	100	1400
Total	500	2000

Causal Effect in Selected Subset

Condition		Outcome: Good	Outcome: Bad	
	T	200	300	500
	C	100	400	500

$$RD = .6 - .8 = -.2$$

The Cutting Edge

1. Attempt to predict "success" not just outcome
2. Uplift (a.k.a. Incremental, True Lift, Net) Modeling
3. Derive models under Treatment and Control conditions
4. Select targets based on "difference score"

Marketing Research Terminology

	Treated	**"Control"**
"Lost Causes"	--------	--------
"Do Not Disturbs"	--------	$$$$$
"Persuadables"	$$$$$	--------
"Sure Things"	$$$$$	$$$$$

Caveats

1. Still have issues of *selection bias*
2. Each of two models predicts *outcome*, not *success*
3. Individual differences are highly *variable*
4. Some applications in insurance may differ