Robustness

Professor Susan Athey

ML and Causal Inference

Robustness of Causal Estimates
Athey and Imbens (AER P&P, 2015)

- General nonlinear models/estimation methods
- Causal effect is defined as a function of model parameters
  - Simple case with binary treatment, effect is $\tau_i = Y_i(1) - Y_i(0)$
- Consider other variables/features as "attributes"
- Proposed metric for robustness:
  - Use a series of "tree" models to partition the sample by attributes
    - Simple case: take each attribute one by one
  - Re-estimate model within each partition
  - For each tree, calculate overall sample average effect as a weighted average of effects within each partition
  - This yields a set of sample average effects
  - Propose the standard deviation of effects as robustness measure

Robustness of Causal Estimates
Athey and Imbens (AER P&P, 2015)

- Four Applications:
  - Randomly assigned training program
  - Treated individuals with artificial control group from census data (Lalonde)
  - Lottery data (Imbens, Rubin & Sacerdote (2001))
  - Regression of earnings on education from NLSY
- Findings
  - Robustness measure better for randomized experiments, worse in observational studies

Lalonde data

Variable	Exper $\hat{\theta}_B$	Exper s.e.	Non-exper $\hat{\theta}_B$	Non-exper s.e.
Base Model	1.67	(0.67)	1.07	(0.63)
$\hat{\sigma}_\theta$		[0.13]		[2.13]
Split on	est.	$\mathcal{X}^2(10)$	est.	$\mathcal{X}^2(10)$
treatment	1.58	22.8	-4.26	45.0
age	1.55	10.1	1.97	144.5
black	1.71	11.4	1.38	26.2
hispanic	1.61	7.2	1.54	59.7
married	1.87	11.0	1.06	10.4
education	1.77	14.7	1.25	74.5
nodegree	1.33	18.6	1.77	46.1
re74	1.64	11.2	0.58	71.0
re75	1.63	8.9	-0.88	94.1
u74	1.64	11.2	-2.44	88.6
u75	1.71	7.0	-0.83	81.9

Comparing Robustness

Variation of $\widehat{\theta}$ over Model Specifications

	Lalonde Exper	Non-exp	IRS	NLS
Est	1.67	1.07	-0.44	0.059
(s.e.)	(0.67)	(0.63)	(0.012)	(0.010)
σ_θ	[0.13]	[2.13]	[0.10]	[0.004]
ratio	0.20	3.38	0.83	0.40

Robustness Metrics: Desiderata

- Invariant to:
  - Scaling of explanatory variables
  - Transformations of vector of explanatory variables
  - Adding irrelevant variables

- Each member model must be somehow distinct to create variance, yet we want to allow lots of interactions
  - Need to add lots of rich but different models

- Well-grounded way to weight models
  - This paper had equal weighting

Robustness Metrics: Work In Progress

Std Deviation versus Worst-Case

- Desire for set of alternative models that grows richer
  - New additions are similar to previous ones, lower std dev
- Standard dev metric:
  - Need to weight models to put more weight on *distinct* alternative models
- "Worst-case" or "bounds":
  - Find the lowest and highest parameter estimates from a set of models
  - Ok to add more models that are similar to existing ones.
  - But worst-case is very sensitive to outliers—how do you rule out "bad" models?

Theoretical underpinnings

- Subjective versus objective uncertainty
  - Subjective uncertainty: correct model
  - Objective uncertainty: distribution of model estimates given correct model
- What are the preferences of the "decision-maker" who values robustness?
  - "Variational preferences"
    - "Worst-case" in set of possible beliefs, allow for a "cost" of beliefs that captures beliefs that are "less likely." (see Strzalecki, 2011)
  - Our approach for exog. covariate case:
    - Convex cost to models that perform poorly out of sample from a predictive perspective
- Good model
  - Low obj. uncertainty: tightly estimated
  - Other models that predict outcomes well also yield similar parameter estimates