Causal Inference: A Tutorial

Fan Li

Department of Statistical Science
Duke University

November 27, 2018

Causality in Ancient Greek Philosophy

I would rather discover one causal law than be King of Persia.

— Democritus

We have knowledge of a thing only when we have grasped its cause.

— Aristotle, Posterior Analytics

We do not have knowledge of a thing until we have grasped its why, that is to say, its cause.

— Aristotle, Physics

Questions on Causation

- Relevant questions about causation:
  - the philosophical meaningfulness of the notion of causation
  - deducing the causes of a given effect
  - understanding the details of causal mechanism
- Here we focus on measuring the effects of causes, where statistics arguably can contribute most
- Several statistical frameworks
  - potential outcomes (J Neyman, DB Rubin)
  - causal diagrams (J Pearl)

Association versus Causation

- The research questions that motivate most studies in statistics-based sciences are causal in nature.
- The aim of standard statistical analysis is to infer associations among variables
- Causal analysis goes one step further; its aim is to infer aspects of the *data generating process*
- In most cases, *Association does not imply causation*: behind every causal conclusion there must lie some causal assumption that is not testable.

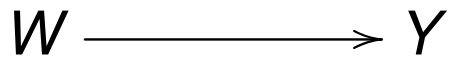
Notations

- Treatment (e.g. intervention, exposure) W: we will mostly focus on binary treatments
- Outcome (e.g. disease status) Y
- Observed covariates or confounders X
- Unobserved covariates or confounders U
- Examples of question of interest
  - Causal effect of exposure on disease
  - Comparative effectiveness research: whether one drug or medical procedure is better than the other
  - Program evaluation in economics and policy

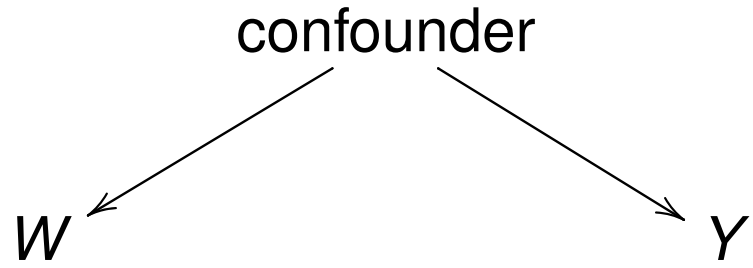
Confounding

- Confounding (or common cause) is the main complication/hurdle between association and causation
- Two Directed Acyclic Graphs (Pearl 1995)

Cause relationship:



Confounding:



A Classic Example—Smoking and Lung Cancer

Doll and Hill (1950 BMJ)

Figure: Sir Austin Bradford Hill (1897–1991)

- Smoking-cancer association
- Case-control study of lung cancer
- Risk ratio $\approx$ odds ratio, is roughly 9 even after adjusting for observed covariates:

$$\mathrm{RR}_{WY}^{\mathrm{obs}} = \frac{\Pr(Y = 1 \mid W = 1)}{\Pr(Y = 1 \mid W = 0)} \approx 9$$

- Does smoking cause lung cancer?
- Box (2013) stopped smoking after seeing Doll and Hill (1950)

A Classic Example—Smoking and Lung Cancer

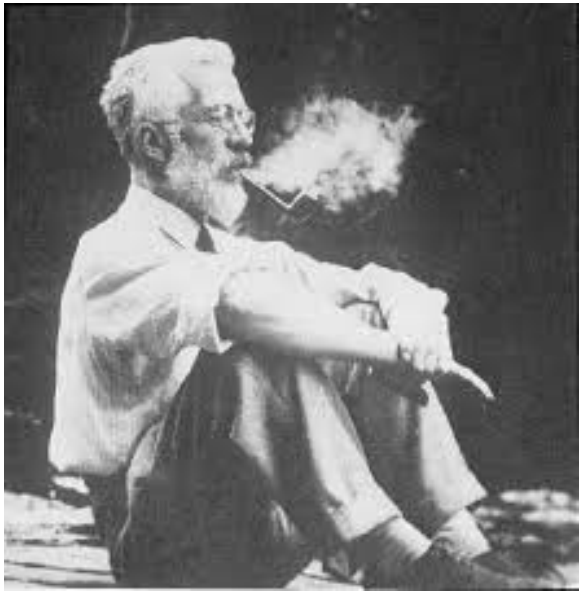


Figure: Sir Ronald Aylmer Fisher (1890–1962)

- Association does not imply causation
- "Common cause" (Reichenbach 1956, Fisher 1957 BMJ)
- Fisher (1957 BMJ):

  cigarette-smoking and lung cancer, though not mutually causative, are both influenced by a common cause, in this case the individual genotype.

Simpson's paradox: Kidney Stone Treatment

(Charig et al., BMJ, 1986)

- An extreme example of confounding is Simpson's paradox: confounder reverses the sign of the correlation between treatment and outcome
- Compare the success rates of two treatments for kidney stones
- Treatment A: open surgery; treatment B: small puncture

	Treatment A	Treatment B
Small stones	**93%** (81/87)	87% (234/270)
Large stones	**73%** (192/263)	69% (55/80)
Both	78% (273/350)	**83%** (289/350)

- What is the confounder here? Severity of the case

Potential Outcome Framework

- The Potential Outcome Framework: the most widely used framework across many disciplines
- Brief history
  - Randomized experiments: Fisher (1918, 1925), Neyman (1923)
  - Formulation (assignment mechanism and Bayesian model): Rubin (1974, 1977, 1978)
  - Observational studies and propensity scores: Rosenbaum and Rubin (1983)
  - Connecting to instrumental variables: Angrist, Imbens and Rubin (1996)

Potential Outcome Framework: Key Components

- No causation without manipulation: a “cause” must be (hypothetically) manipulatable, e.g., intervention, treatment
- Goal: estimate the effects of “cause”, not causes of effect
- Three integral components (Rubin, 1978):
  - potential outcomes: corresponding to the various levels of a treatment
  - assignment mechanisms
  - a model for the potential outcomes and covariates
- Causal effects: a comparison of the potential outcomes under treatment and control for *the same set of units*

Setup

- Data: a random sample of N units from a target population
- A treatment with two levels: $w = 0, 1$
- For each unit i, we observe the (binary) treatment status W_i, a vector of covariates X_i, and an outcome Y_i^{obs}
- For each unit i, two potential outcomes $Y_i(0), Y_i(1)$ – implicitly invoke the Stable Unit Treatment Value Assumption (SUTVA)
- Causal estimands, e.g. Average treatment effect (ATE): $\tau = \mathbb{E}[Y_i(1) - Y_i(0)]$.

The Fundamental Problem of Causal Inference

Holland, 1986, JASA

- For each unit, we can observe at most one of the two potential outcomes, the other is missing (counterfactual?)
- Causal inference under the potential outcome framework is essentially a missing data problem
- To identify causal effects from observed data, one must make additional (structural or/and stochastic) assumptions
- Key identifying assumptions are on assignment mechanism: the probabilistic rule that decides which unit gets assigned to which treatment

Perfect Doctor

	Potential Outcomes			Observed Data		
	$Y(0)$	$Y(1)$		W	$Y(0)$	$Y(1)$
	13	14		1	?	14
	6	0		0	6	?
	4	1		0	4	?
	5	2		0	5	?
	6	3		0	6	?
	6	1		0	6	?
	8	10		1	?	10
	8	9		1	?	9
True averages	7	5	Observed averages		5.4	11

Two key assumptions

Rosenbaum and Rubin, 1983, Biometrika

- Strong ignorability is the key assumption, consisting of
  - Assumption 1 (Positivity (a.k.a. overlap)): each unit has no zero probability of receiving either treatment
  - Assumption 2 (Unconfoundedness (a.k.a. ignorability)): no unmeasured confounders; if two groups have the same distribution of observed covariates, the treatment assignment is random
- Positivity is testable, but unconfoundedness is generally not

Overlap and Balance

- Under unconfoundedness, the causal effects are identified from the observed data:
  1. First conditional on subpopulations with covariate balance (via e.g., randomization, or matching, stratification), calculate the difference between treatment and control groups
  2. Average over all such subpopulations (X)
- The key is to obtain covariate overlap and balance between groups
- Balance of confounders (observed and unobserved) play a central role in causal inference
- Observed difference in outcomes might be purely due to the imbalance of confounders between groups

Classification of assignment mechanisms

- Randomized experiments:
  - strong ignorability automatically holds
  - good balance is (in large samples) guaranteed
- Unconfounded observational studies
  - strong ignorability is assumed
  - balance need to be achieved
- Quasi-experiments: looking for “natural" experiments (under assumptions)

Randomized Experiments

- In randomized experiments, assignment mechanism is known and controlled by investigators
- Strong ignorability automatically holds
- Randomization does:
  - balance observed covariates
  - balance unobserved covariates
  - balance potential outcomes, i.e. guarantee unconfoundedness

Role of Randomization

- Under randomization, causal effects are identified by the difference in the outcome between the treatment and control groups
- Under randomization, *association does imply causation* (of course within the potential outcome framework with assumptions)

Chance Imbalance in Randomized Experiments

- Randomization "should" balance all covariates (observed and unobserved) on average...
- But covariates may be imbalanced by random chance
- Why is covariate balance important in randomized experiments?
- Because better balance
  - Provides more meaningful estimates of the causal effect
  - Increases power, particularly if imbalanced covariates correlated with outcome

Covariate Balance in Randomized Experiments

- Option 1: force better balance on important covariates *by design* –"Block what you can; randomize what you cannot" (George Box)
  - stratified randomized experiments
  - paired randomized experiments
  - rerandomization
- Option 2: correct imbalance in covariates by *analysis*
  - outcome: gain scores
  - separate analysis within subgroups
  - covariate adjustment via regression or weighting

Randomized Experiments: Complications

- Noncompliance
- Loss to follow-up
- Truncation due to “death", e.g. patients died before end of study on life quality
- Generalize to wider population
- Ethical and practical constraints: clinical equipoise, sequential trials, pragmatic trials

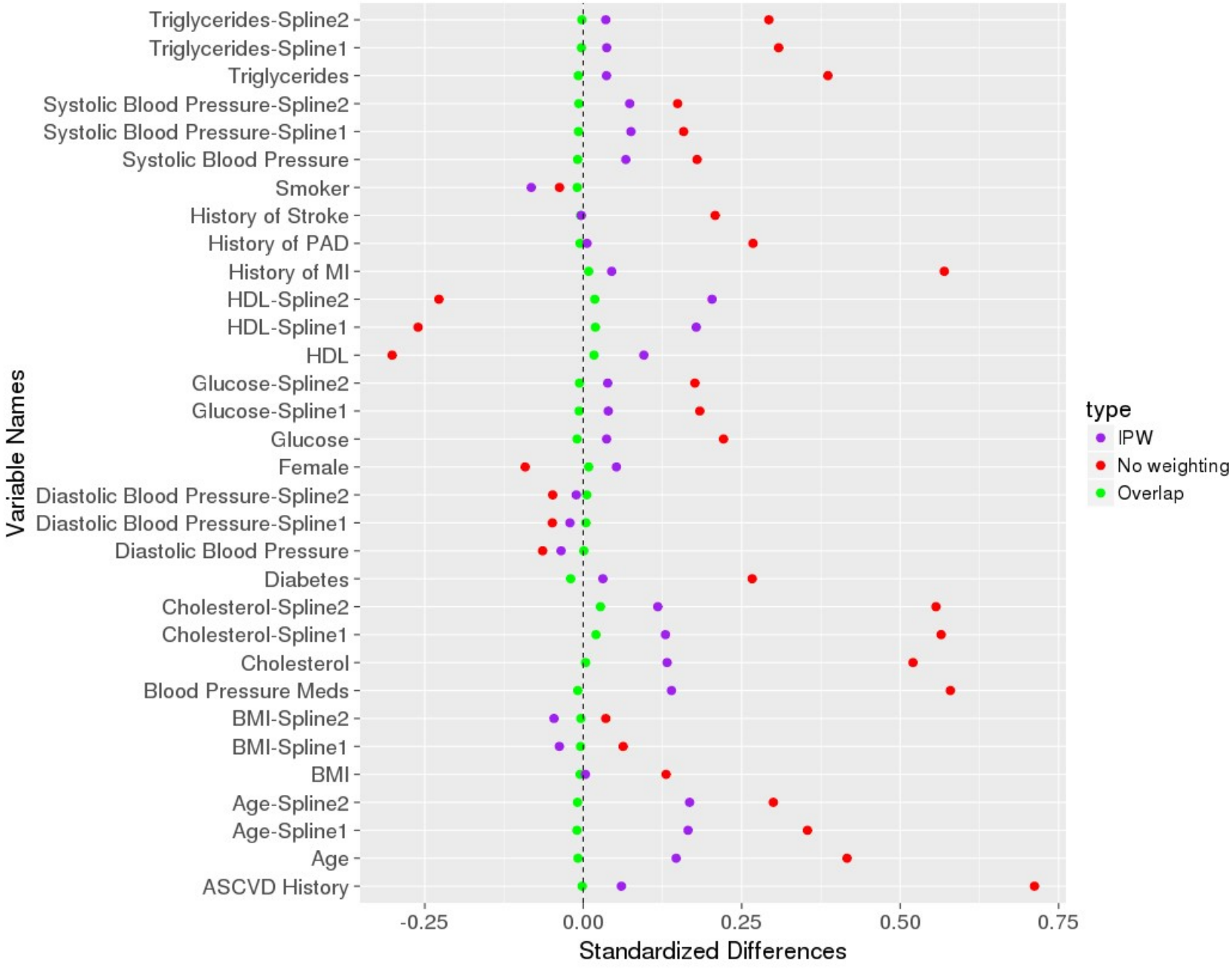
Observational Studies

- In observational studies, we do not control or know the treatment assignment mechanism
- Measured and unmeasured confounders: usually unbalanced between groups
- Self-selection to treatment is prevalent
- Must make (often untestable) structural assumptions on assignment mechanism to identify causal effects
- Strong Ignorability is not guaranteed but usually assumed in the vast majority of observational studies

Example: Framingham Heart Study

(Thomas, Lorenzi, et al. 2018)

- **Goal:** evaluate the effect of statins on health outcomes
- **Patients:** cross-sectional population from the offspring cohort with a visit 6 (1995-1998)
- **Treatment:** statin use at visit 6 vs. no statin use
- **Outcomes:** CV death, myocardial infarction (MI), stroke
- **Confounders:** sex, age, body mass index, diabetes, history of MI, history of PAD, history of stroke...
- Significant imbalance between treatment and control groups in covariates


Triglycerides-Spline2
Triglycerides-Spline1
Triglycerides
Systolic Blood Pressure-Spline2
Systolic Blood Pressure-Spline1
Systolic Blood Pressure
Smoker
History of Stroke
History of PAD
History of MI
HDL-Spline2
HDL-Spline1
HDL
Glucose-Spline2
Glucose-Spline1
Glucose
Female
Diastolic Blood Pressure-Spline2
Diastolic Blood Pressure-Spline1
Diastolic Blood Pressure
Diabetes
Cholesterol-Spline2
Cholesterol-Spline1
Cholesterol
Blood Pressure Meds
BMI-Spline2
BMI-Spline1
BMI
Age-Spline2
Age-Spline1
Age
ASCVD History
Variable Names
-0.25
0.00
0.25
0.50
0.75
Standardized Differences
type
IPW
No weighting
Overlap

Regression Adjustment

- Need to adjust difference in the outcomes due to the differences in covariates
- Most commonly via a regression model:

$$Y \sim a + bW + cX + dW \cdot X$$

- Potential problems
  - Regression itself does not take care of lack of overlap or balance
  - In regions where the groups do not have covariate overlap, causal estimation is purely based on extrapolation
  - Sensitivity to model-specification

Strategies to Reduce Model Sensitivity

- To mitigate model dependence, two strategies: (1) design - balance covariates, (2) analysis- flexible models
- Best strategy is to use both jointly: first balance covariates in the design stage, then use flexible models in the analysis stage
- Balance covariates
  - Stratification or matching
  - Propensity score methods
- Flexible models
  - Semiparametric models (e.g., power series)
  - Machine learning methods (e.g., tree-based methods (CART, random forest), boosting)
  - Bayesian non- and semi parametric models (e.g., Gaussian Processes, BART, Dirichlet Processes mixtures)

Balancing covariates: small number of covariates

- When the number of covariates is small, the adjustment can be achieved by exact matching or stratification
- Exact matching: for each treated subject, get a control with exact same value of the covariate
- Exact matching ensures distributions of covariates in treatment and control groups are exactly the same, thus eliminate bias due to difference in X
- Exact matching is usually infeasible, even with low-dimensional covariates

Matching

- Regression estimators impute the missing potential outcomes using the estimated regression function
- Matching estimators also impute the missing potential outcomes, but do so using only the outcomes of nearest neighbours of the opposite treatment group (similar to nonparametric kernel regression methods)
- Matching is often (but not exclusively) been applied in settings where there is a large reservoir of potential controls

Matching: Dimensional Reduction

- Matching is good, but...
- What if there is a large number of covariates? With just 20 binary covariates, there are 2^{20} or about a million covariate patterns
- Direct matching or stratification is nearly impossible
- Need dimensional reduction: propensity score

Propensity score

Rosenbaum and Rubin, 1983, Biometrika

- The propensity score $e(x)$: the probability of a unit receiving a treatment given covariates
- Two key properties
  1. The propensity score $e(X)$ balances the distribution of all observed covariates X between the treatment groups
  2. If the treatment is unconfounded given X, then the treatment is unconfounded given $e(X)$

Propensity score

- Propensity score is a scalar summary (summary statistic) of the covariates w.r.t. the assignment mechanism
- Propensity score is central to ensure balance and overlap
- The propensity score balances the **observed** covariates, but does not generally balance **unobserved** covariates
- In most observational studies, the propensity score $e(X)$ is unknown and thus needs to be estimated

Propensity score: analysis procedure

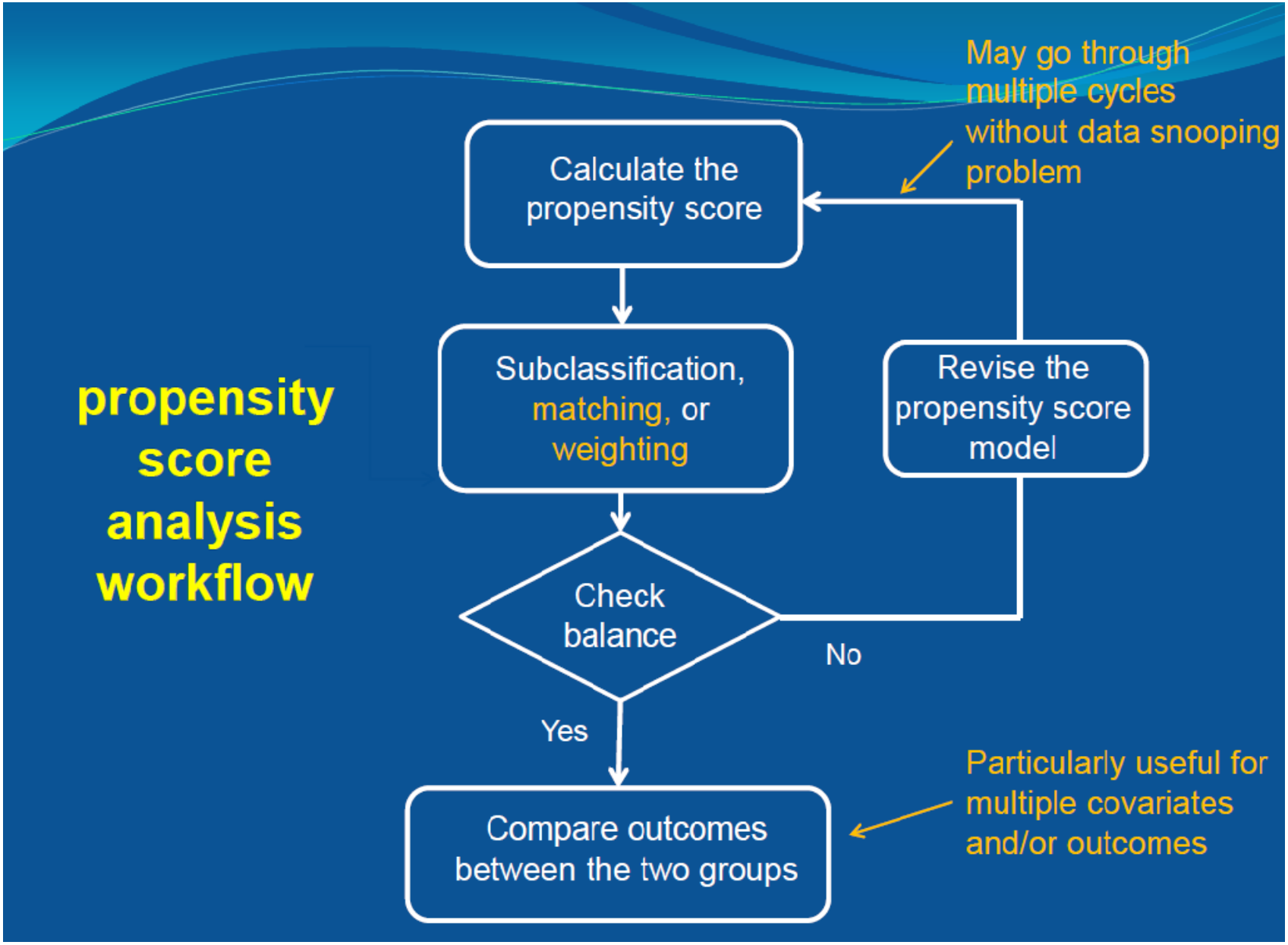
Propensity score analysis typically involves two stages:

Stage 1 Estimate the propensity score, by e.g. a logistic regression or a machine learning method

Stage 2 Given the estimated propensity score, estimate the causal effects through one of these methods:

- Stratification
- Weighting
- Matching
- Regression
- Mixed procedure of the above

Propensity score analysis workflow



Propensity score matching

- Special case of matching: the distance metric is the (estimated) propensity score
- 1-to-n nearest neighbor matching is common when the control group is large compared to treatment group
- Pros: intuitive, robust, matched pairs, balance distributions in directions uncorrelated to estimated PS
- Cons
  - much tuning: with or without replacement, 1-to-1 or 1-to-n, caliper, ties
  - programming is hard
  - difficult to extend to complex situations: sequential treatments, multi-valued treatments

Propensity score weighting

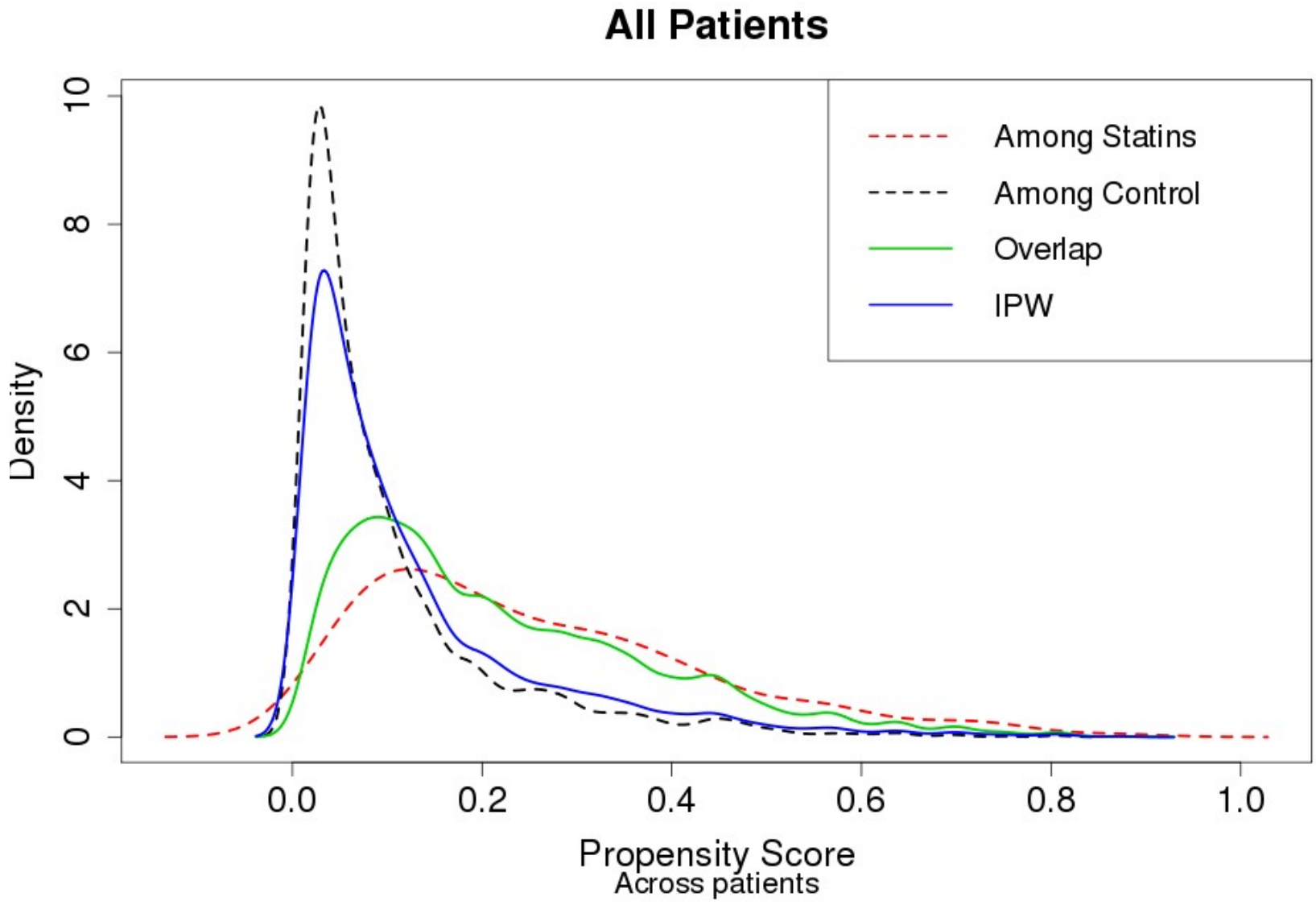
Li, Morgan, Zaslavsky, 2018, JASA

- Another popular approach is (propensity score) weighting
- Main idea: re-weigh the treatment and control groups to create a pseudo-population—the target population—where the two groups are balanced, in expectation
- A general class of balancing weights
- Different weighting schemes: different target population and causal estimands
- One should choose the target population *a priori*

Propensity score weighting: two schemes

- Inverse probability weights (IPW)
  - Weigh each unit by the inverse of its probability of being assigned to the current group
  - Target population: the population that the study sample is presentative of
  - But what if the sample is a convenience sample?
- Overlap weights (Li, Morgan, Zaslavsky, 2018, JASA)
  - Weigh each unit by its probability of being assigned to the opposite group
  - Target population: the population with the most overlap in characteristics between groups (clinical equipoise)
  - Overlap weights give exact balance of covariates

Framingham revisited: weighted distribution



Results: composite of non-death endpoints

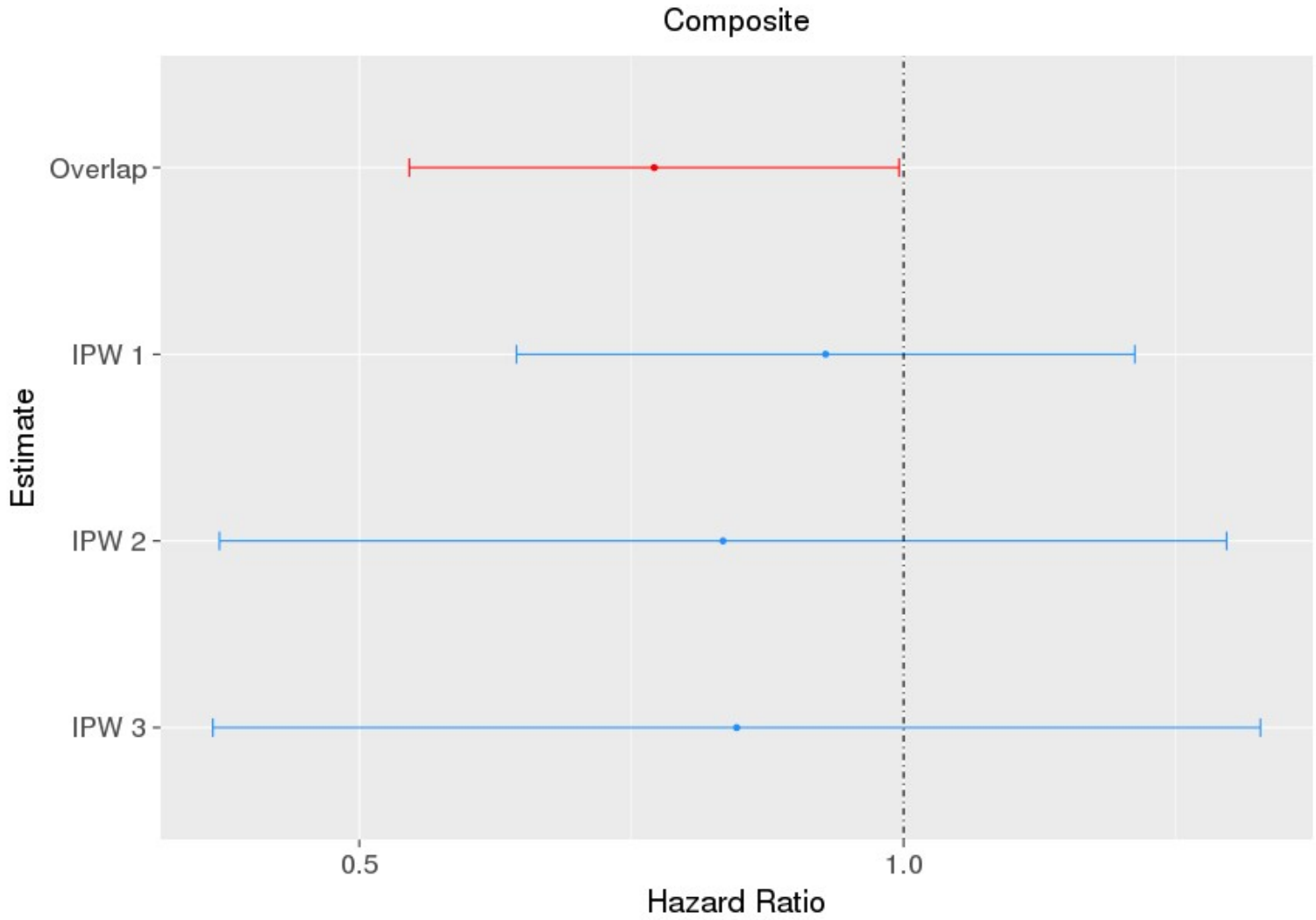


Figure: IPW 1: No trimming; IPW 2: trimming ps between (.10, 0.90); IPW 3: asymmetric trimming 5th% ps of trt, 95th% of ps for control

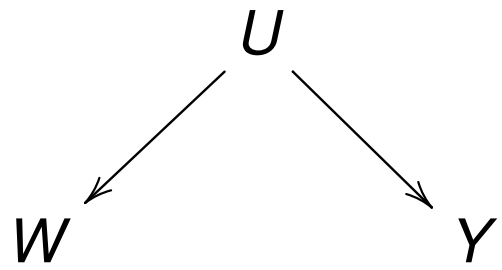
Sensitivity analysis

- Unconfoundedness is inherently untestable (unknown unknowns)
- One should always perform sensitivity analysis to assess how sensitive the causal analysis is to violation to unconfoundedness
- Sensitivity is different from testing, more of a “insurance” check
- Sensitivity analysis in causal inference dates back to the Hill-Fisher debate on causation between smoking and lung cancer, and first formalized in Cornfield (1959, JNCI)

Smoking and Lung Cancer: Revisited

Cornfield et al., 1959, JNCI

Common cause hypothesis



- Smoking W
- Lung cancer Y
- Genetic factor U

- Fisher argued the association between smoking and lung cancer may be due to a common gene that causes both
- Cornfield showed: assuming Fisher is right, the smoking-gene association must satisfy: $\mathrm{RR}_{WU} \geq \mathrm{RR}_{WY} \approx 9$
- Such a genetic confounder is too strong to be realistic
- Thus, here association must be due to causal

Sensitivity analysis

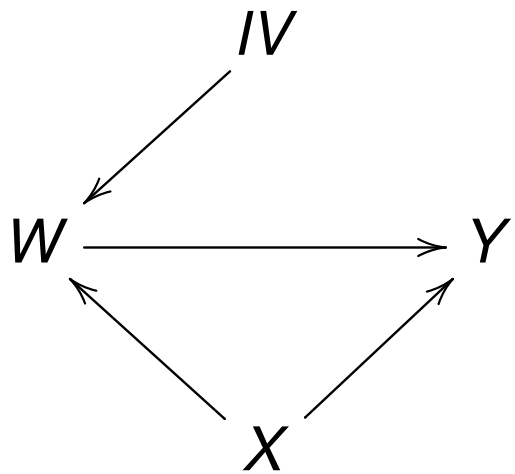
- Fundamental ideas
  - Check what would happen to the same analysis had there was an unmeasured confounder? (Rosenbaum and Rubin, 1983b)
  - Or, how strong an unmeasured confounder has to be to explain away the observed effects? (E-value) (Ding and VanderWeele, 2014, 2016, 2016)
- Seldom done in substantive field, but should always be checked

Quasi-Experiments

- Leverage the variation in treatment assignment resulted from nature or policy
- Three main categories
  - Instrumental variables (IV)
  - Regression discontinuity designs (RDD)
  - Difference-in-Differences (DiD)

Instrumental Variables

- An instrumental variable (IV): a variable that has a causal effect on the treatment, but (is assumed to) have no "direct" causal effect on the outcome



IV Example 1: Season of Birth

Angrist and Krueger, 1991, Quarterly Journal of Economics

- Goal: evaluate the effect of schooling on earnings
- Challenge: Relationship between year of schooling and earnings is highly confounded by factors like family social-economics status
- IV: quarter of the year of birth
- Main idea
  - When one was born in the year is largely randomized, by nature; it should not affect one's later earnings directly
  - It does directly affects when a child first attended school, and in combination with the compulsory education requirement, this can create up to one year of difference in schooling

IV Example 2: Distance to Hospitals

McClellan, McNeil, Newhouse, 1994, JAMA

- Goal: evaluate the effect of intensive treatment of acute MI on mortality
- Challenge: Relationship between receiving intensive treatment and mortality among AMI patients is highly confounded by factors like case severity
- IV: distance to the closest hospital
- Main idea
  - Where one lives is largely randomized; it should not directly affect one's survival following AMI
  - It does directly affect what type of hospitals (high vs. low volume and treatment availability) the patient first went, this in turns affects which treatment the patient received

Two-stage Least Square (TSLS) Estimator

- Traditional model:

$$Y_i = \beta_0 + \beta_1 W_i + \beta_2 X_i + \epsilon_i.$$

  where β_1 is the causal effect
- Direct OLS estimate of β_1 is biased
- With IV, we can fit a two-stage least square (2SLS) regression to estimate β_1:

$$M1: \quad Y_i = \pi_{10} + \pi_{11} Z_i + \pi_{12} X_i + u_i$$

$$M2: \quad W_i = \pi_{20} + \pi_{21} Z_i + \pi_{22} X_i + v_i$$

- The 2SLS estimate of β_1 is a ratio:

$$\hat{\beta}_1^{2sls} = \hat{\pi}_{11} / \hat{\pi}_{21}$$

Instrumental Variables: Assumptions

- IVs, when available, are extremely useful tools to draw causal inference
- But, good IVs are hard to come by
- A good IV must satisfy two conditions (assumptions)
  1. Have a strong effect on the treatment, o.w. the estimate will have large variance
  2. Not have any direct effect on the outcome, o.w. has the same endogenous problem as the treatment
- Still one of the most popular causal inference methods

Regression discontinuity design (RDD)

- Regression discontinuity designs: the treatment status changes discontinuously according to some underlying pre-treatment variable – the *running variable*
- Basic idea: comparing units with similar values of the running variable, but different levels of treatment would lead to causal effect of the treatment at the threshold
- The discontinuity is often created by a pre-fixed, artificial threshold of a policy
- Treatment among the subjects around the threshold can be viewed as locally randomized

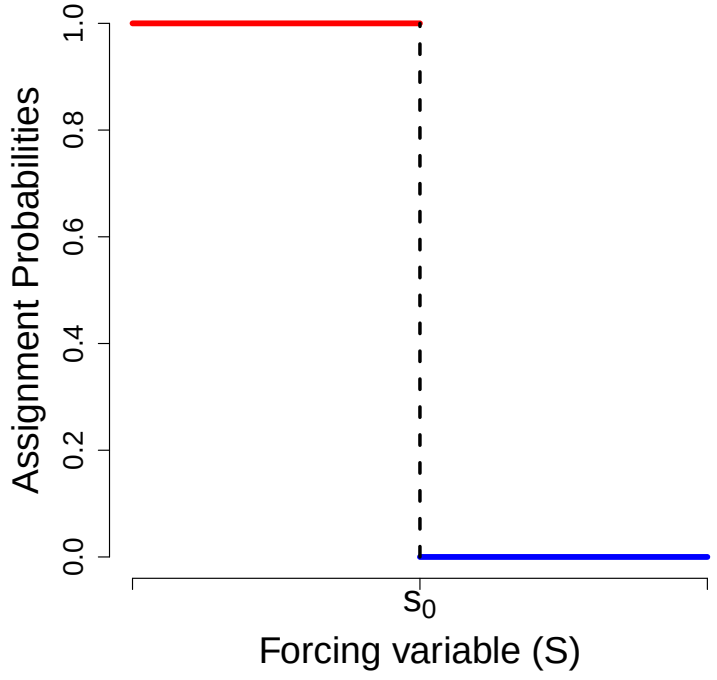
RDD Example: Financial Aid and Dropout

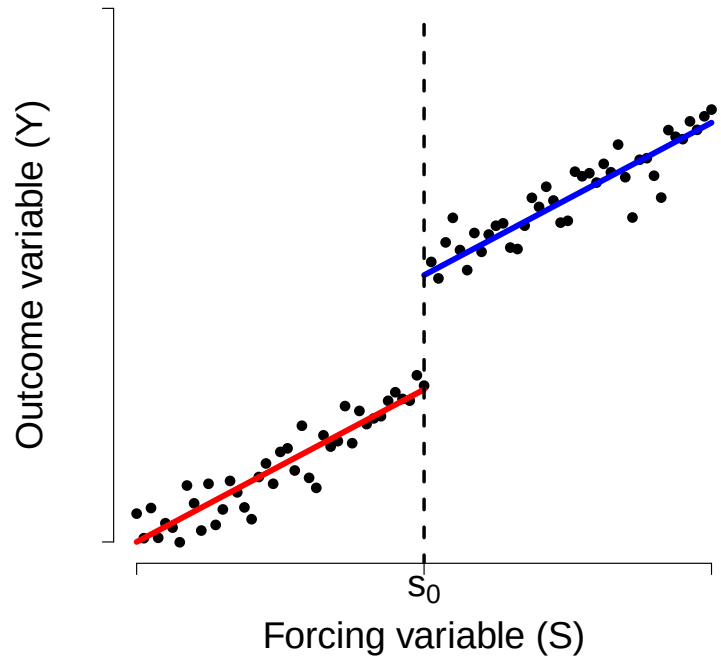
Li, Mattei, Mealli, 2015, AOAS

- Goal: evaluate the effect of financial aid on preventing dropout in Italian colleges
- Challenge: students who received aids and who did not are different in observed and unobserved ways
- Running variable: family wealth – eligibility to financial aid depends solely on whether the family wealth is above or below a fixed threshold
- Main idea
  - Arguably students whose family wealth are just above and just below the threshold are comparable in their background
  - The artificial threshold set by the administration creates a “local randomization" of treatment around the threshold

Sharp RDD

- The treatment status is a deterministic step function of a running variable

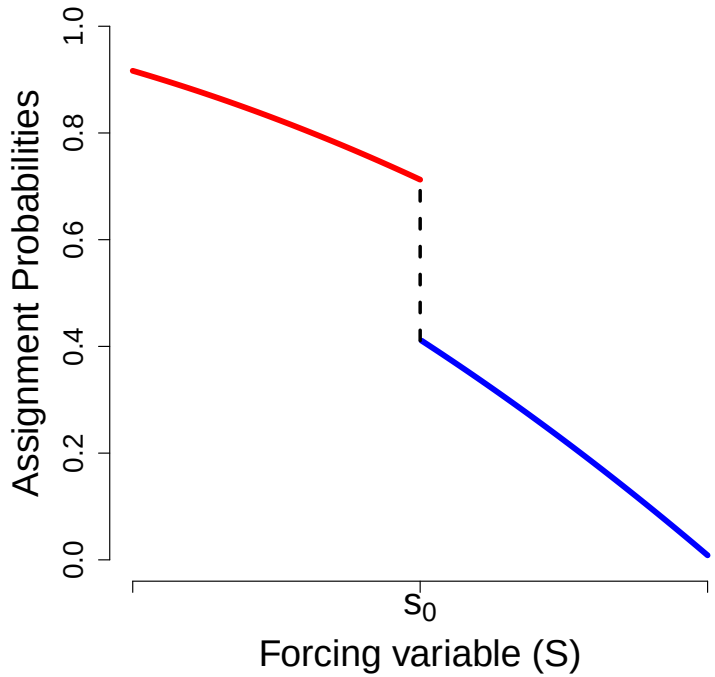




- Focus on causal effects of the treatment at the threshold
- A jump in s_0 is interpreted as causal effect

Fuzzy RDD

- A value of the running variable falling above or below the threshold acts as encouragement to take the treatment



- In fuzzy RDDs, the receipt of the treatment depends also on individual choices, raising non-ignorability issues
- Fuzzy RDDs are related to IV: falling above or below the threshold can be viewed as an instrument

RDD: assumption and limitationss

- Key assumption: continuity at the threshold or local randomization
- Key to analysis: identify a small window around the threshold where local randomization is reasonable
- Limitations
  - Treatment effect local to the threshold, how generalizable?
  - Manipulation of the running variable

Difference-in-Differences (DiD)

- A treatment-control comparison is not necessarily a causal comparison because of the potential systematic differences between two groups
- A unit is arguably the “best match” for itself
- A before-after comparison (of the same units) is not necessarily a causal comparison because of the potential change in time
- Difference-in-Differences (DiD) design combines both: before-after treatment-control comparison
- Setup: two or more groups, with units observed in two or more periods. In some periods and some groups are exposed to the treatment

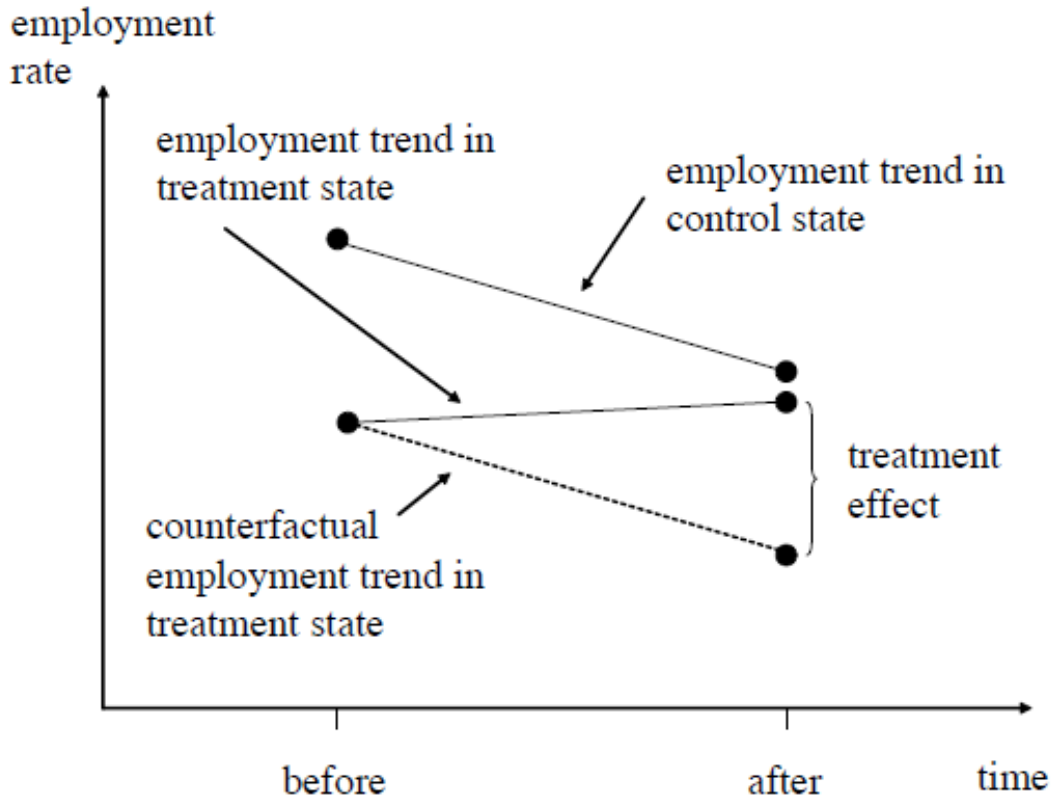
DiD Example: Minimum wages and Employment

Card and Krueger, 1994, American Economic Review

- Goal: study the effect of increase in minimum wage on employment
- Units: fast-food restaurants in New Jersey and adjacent eastern PA
- Intervention: raise of the state minimum wage; NJ raised the minimum on April 1, 1992, but PA not
- Outcome: number of FTE per restaurant, observed in both areas, and both right-before and after the change
- Main idea:
  - The restaurants near the state border are arguably similar
  - Variation in treatment created by discontinuity in time (policy change) and space (state border)

DiD: Parallel Trend Assumption

- Key assumption: Parallel trend – treatment and the control group experience the same trends in the absence of treatment



DiD: Limitations and Alternatives

- Analysis is usually done via a fixed-effects regression model (time-specific and unit-specific effects)
- Limitations
  - Parallel trend is untestable and may be implausible
  - Scale-dependent: parallel trend for Y does not transfer to log Y
  - Serial correlation between observations
- Alternatives: unconfoundedness conditional on past outcomes and covariates
- Uncounfoundedness is also untestable

Final words

- Causal inference is hard, but fundamental and exciting
- The potential outcome framework
- Make and check assumptions
- Design is the key
- Many open questions