Instrumental Variables and Matching

EDLD 650: Week 7

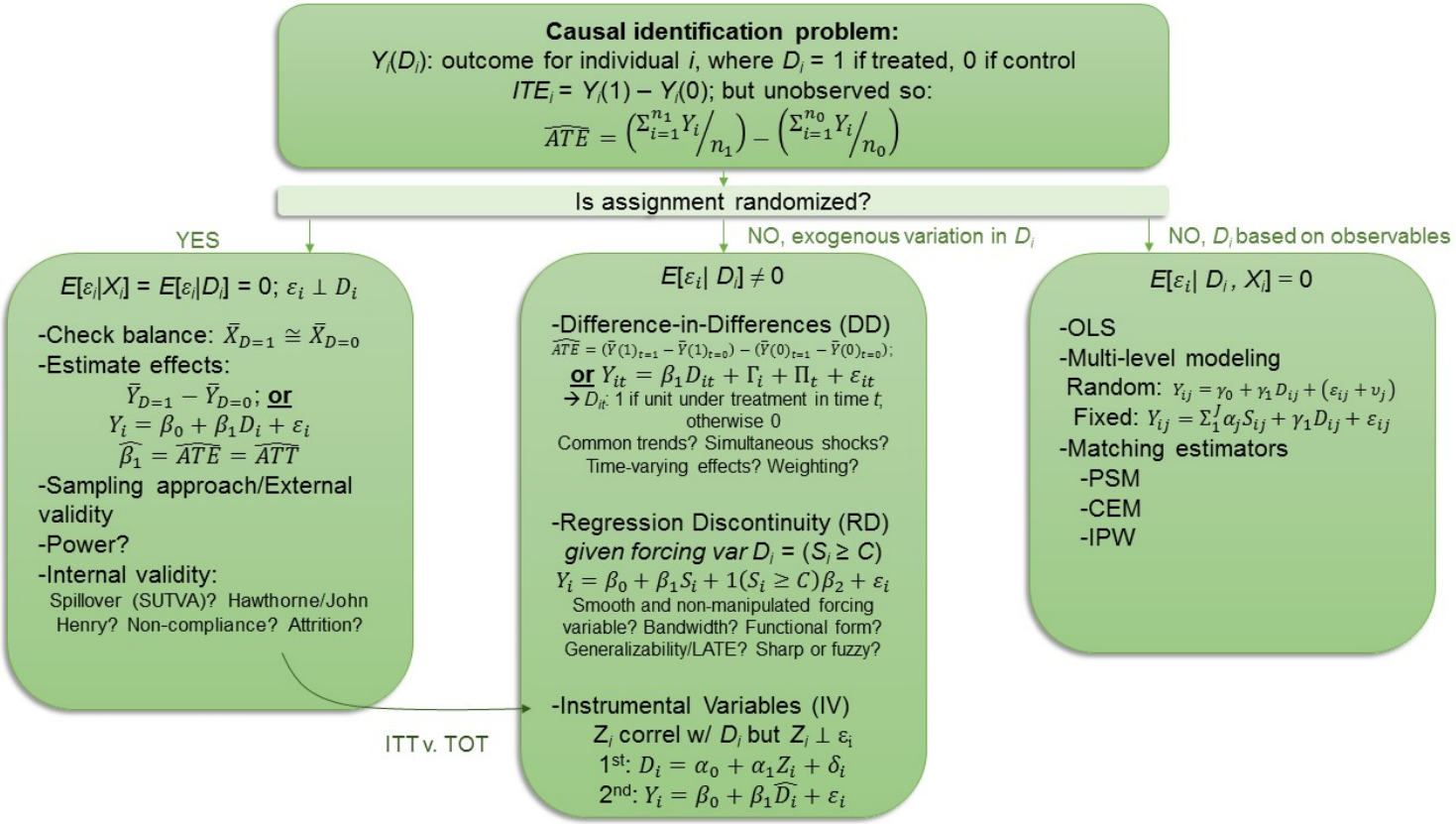
David D. Liebowitz

Agenda

1. Roadmap and goals (9:00-9:10)
2. The Kim et al paper and DARE #3 (9:10-10:20)
3. Break (10:20-10:30)
4. Marching (10:30-11:35)
5. Wrap-up (11:35-11:50)

- To-dos and Plus/deltas

Roadmap


Causal identification problem:
$Y_i(D_i)$: outcome for individual i, where D_i = 1 if treated, 0 if control
$ITE_i = Y_i(1) - Y_i(0)$; but unobserved so:
$\widehat{ATE} = \left(\Sigma_{i=1}^{n_1} Y_i / n_1\right) - \left(\Sigma_{i=1}^{n_0} Y_i / n_0\right)$
Is assignment randomized?
YES
NO, exogenous variation in D_i
NO, D_i based on observables
$E[\varepsilon_i|X_i] = E[\varepsilon_i|D_i] = 0$; $\varepsilon_i \perp D_i$
-Check balance: $\bar{X}_{D=1} \cong \bar{X}_{D=0}$
-Estimate effects:
$\bar{Y}_{D=1} - \bar{Y}_{D=0}$; or
$Y_i = \beta_0 + \beta_1 D_i + \varepsilon_i$
$\widehat{\beta_1} = \widehat{ATE} = \widehat{ATT}$
-Sampling approach/External validity
-Power?
-Internal validity:
Spillover (SUTVA)? Hawthorne/John Henry? Non-compliance? Attrition?
ITT v. TOT
$E[\varepsilon_i| D_i] \neq 0$
-Difference-in-Differences (DD)
$\widehat{ATE} = (\bar{Y}(1)_{t=1} - \bar{Y}(1)_{t=0}) - (\bar{Y}(0)_{t=1} - \bar{Y}(0)_{t=0})$;
or $Y_{it} = \beta_1 D_{it} + \Gamma_i + \Pi_t + \varepsilon_{it}$
→ D_{it}: 1 if unit under treatment in time t, otherwise 0
Common trends? Simultaneous shocks? Time-varying effects? Weighting?
-Regression Discontinuity (RD)
given forcing var D_i = ($S_i \geq C$)
$Y_i = \beta_0 + \beta_1 S_i + 1(S_i \geq C)\beta_2 + \varepsilon_i$
Smooth and non-manipulated forcing variable? Bandwidth? Functional form? Generalizability/LATE? Sharp or fuzzy?
-Instrumental Variables (IV)
Z_i correl w/ D_i but $Z_i \perp \varepsilon_i$
1st: $D_i = \alpha_0 + \alpha_1 Z_i + \delta_i$
2nd: $Y_i = \beta_0 + \beta_1 \widehat{D_i} + \varepsilon_i$
$E[\varepsilon_i| D_i, X_i] = 0$
-OLS
-Multi-level modeling
Random: $Y_{ij} = \gamma_0 + \gamma_1 D_{ij} + (\varepsilon_{ij} + \upsilon_j)$
Fixed: $Y_{ij} = \Sigma_1^J \alpha_j S_{ij} + \gamma_1 D_{ij} + \varepsilon_{ij}$
-Matching estimators
-PSM
-CEM
-IPW

Goals

1. Conduct IV analysis in simplified data and interpret results

2. Assess basic assumptions of IV design in an experimental setting with imperfect compliance

3. Describe the conceptual approach of using selection on observables to defend causal inferences about the effects of a treatment

Class 7 Discussion Questions

DARE-d to do it!

Student examples in class...

Break

Matching

Core causal inference challenge

What is basic problem of drawing causal inferences from non-experimental (observational) data or data from a non-random subset within an experiment?

1. Treatment and non treatment groups are not *equal in expectation*, so it is difficult to claim *variation in treatment condition* is driving *observed differences in outcomes*
2. Sample is *no longer representative* of the population (as originally defined)

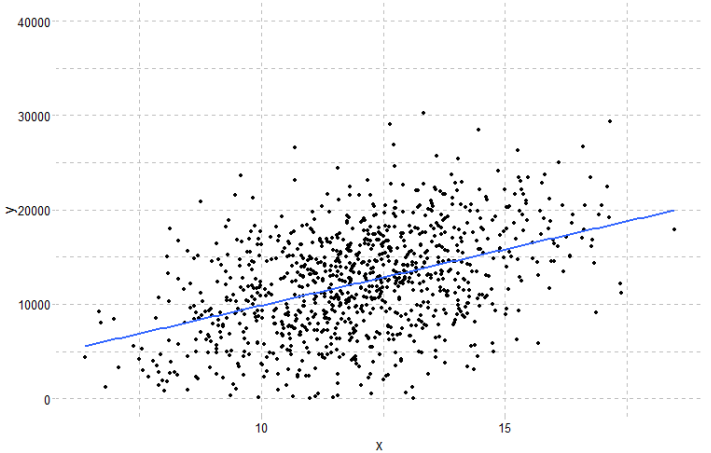
$\rightarrow$ **Biased estimate of treatment effect**

Up until now, we have relied on being able to find an arguably *exogenous* source of variation in likelihood of receiving the treatment... **but what if we can't find this???!!?**

Selection bias

Imagine: outcome Y is a measure of later life success that depends on an earlier education attainment, X; *AND* that this is the actual, causal relationship between X and Y:

Hypothetical causal relation



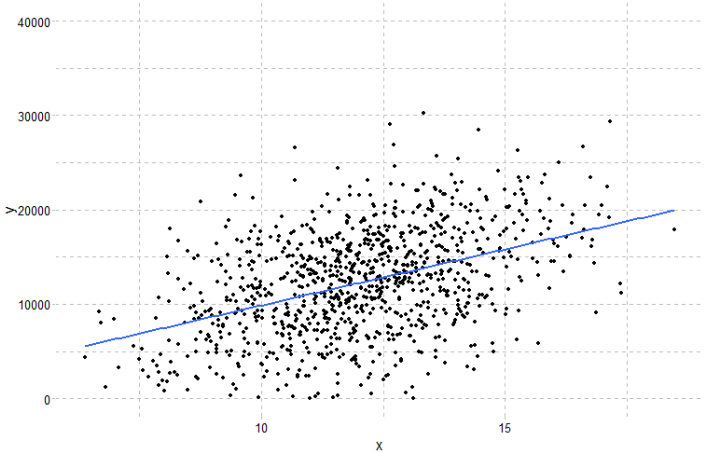
Observed relationship

- but...in addition to the underlying causal relationship between X and Y, society consistently favors one group of individuals over another and in-so-doing, constrains some individuals' ability to access higher levels of educational attainment X.
- so...one group of individuals would consistently experience higher levels of attainment X *and* later life success Y

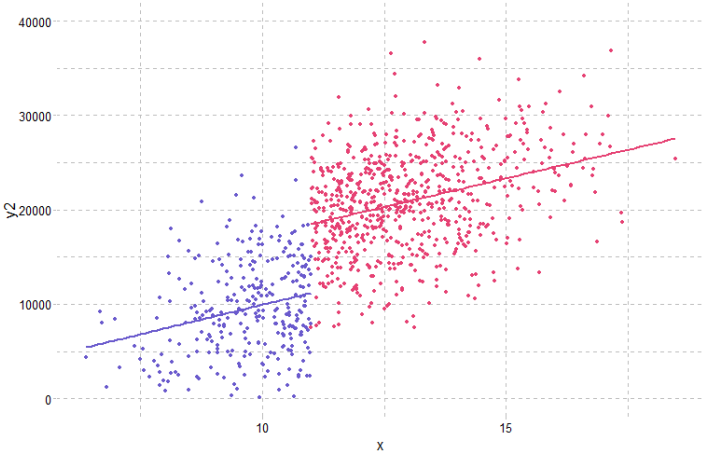
Selection bias

Imagine: outcome Y is a measure of later life success that depends on an earlier education attainment, X; *AND* that this is the actual, causal relationship between X and Y:

Hypothetical causal relation



Observed relationship

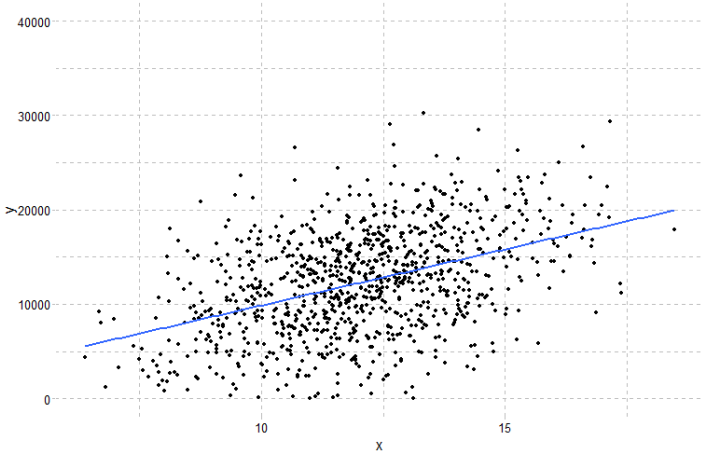


- but...in addition to the underlying causal relationship between X and Y, society consistently favors one group of individuals over another and in-so-doing, constrains some individuals' ability to access higher levels of educational attainment X.
- so...one group of individuals would consistently experience higher levels of attainment X *and* later life success Y

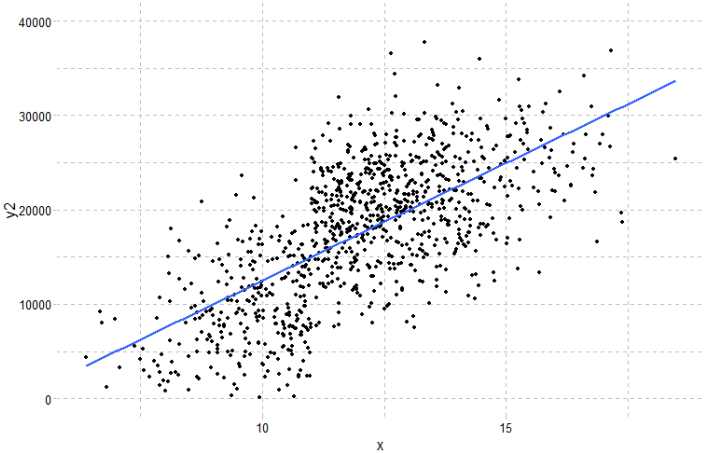
Selection bias

Imagine: outcome Y is a measure of later life success that depends on an earlier education attainment, X; *AND* that this is the actual, causal relationship between X and Y:

Hypothetical causal relation



Observed relationship



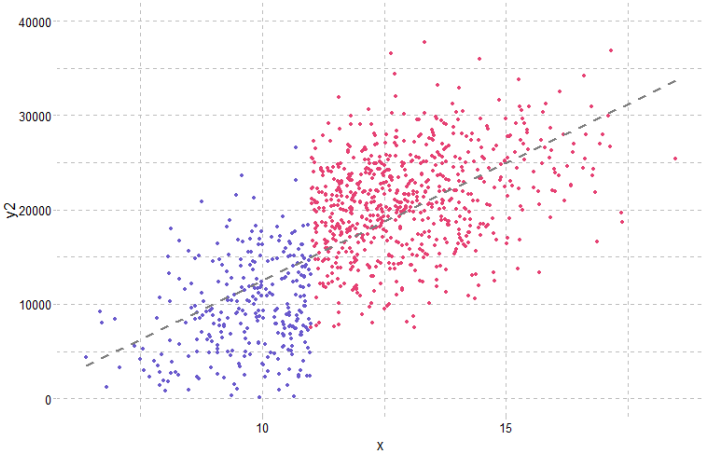
- however...we would not necessarily observe this constraint or know what this group is, and so we would only observe a *biased* relationship between X and Y

A possible solution?

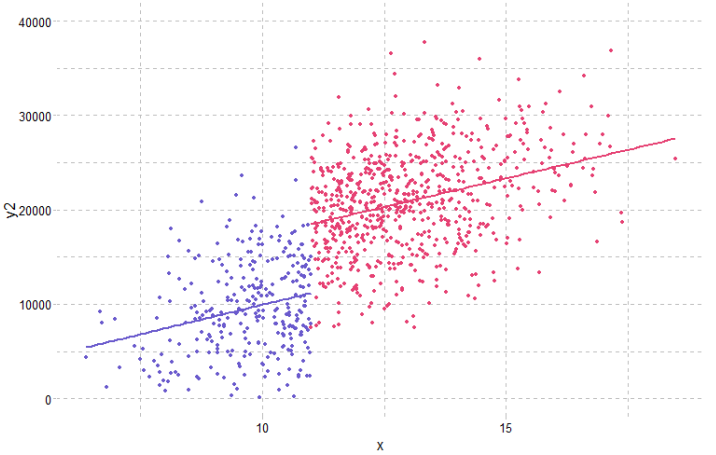
Big idea: if we were sure we knew that the only factor driving selection into treatment was individuals' membership in this group:

- We can ignore overall point cloud and refuse to estimate the biased Y|X slope
- Instead, conduct analysis within each subsidiary point clouds
  - Obtain estimates of treatment effect within each point cloud
  - Average to obtain overall *unbiased* estimate of treatment effect of more educational attainment

Ignore biased observed relationship



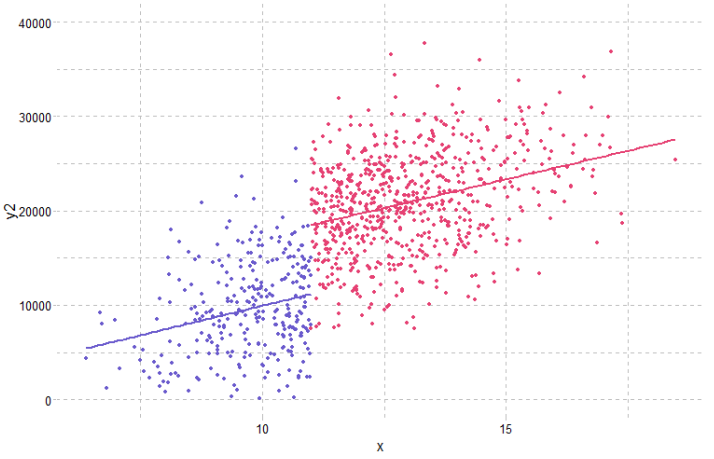
Estimate treatment effect absent bias



Selection on observables

This is the key conceptual basis for approaches such as: **stratification, weighting and matching**. They are used to remove "observed bias" from treatment effects estimated in observational data.

- This family of approaches relies on **selection on observables** into treatment (*more on this later*)
- It is not a magical way of getting causal estimates when you don't have an identification strategy
- Like all the other methods we have studied, it requires a deep substantive understanding of why some are treated and others aren't



Stratification

Stratum		Treatment frequency		Outcome (voting %)		
8th grade test	Motivation	No college	College	No college	College	Diff
Low perf	Favored group	325	295	0.24	0.38	0.14
	Disfavored group	480	90	0.12	0.18	0.06
Middle perf	Favored group	180	165	0.44	0.48	0.04
	Disfavored group	245	110	0.29	0.39	0.10
High perf	Favored group	28	380	0.47	0.87	0.40
	Disfavored group	140	200	0.55	0.66	0.11
					Weighted ATE	0.14

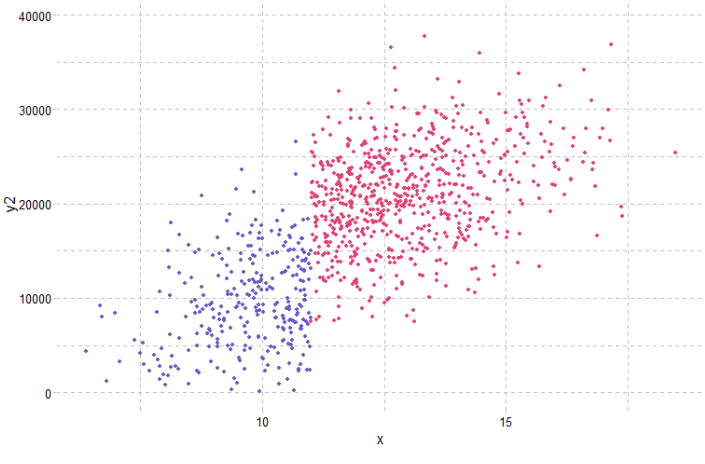
Emerging issues:

Diminishing sample size within stratum

- Imprecise estimates
- Reduced power

In extreme, group may have no observations in a strata

- Lack of common support
- Can't estimate treatment effect



Matching approach

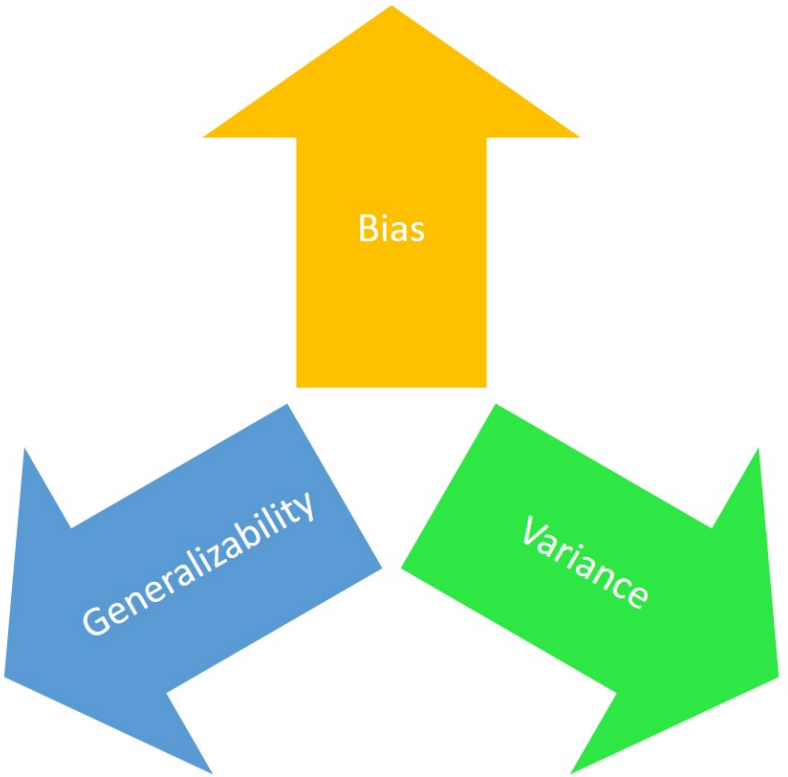
The **conditional independence assumption (CIA)** (Rosenbaum & Rubin, 1983) states that treatment is as-good-as random conditional on a known set of covariates.

If "selection on observables" in fact happens, matching estimators take this literally.

The basic idea: estimate a treatment effect only using observations with (nearly?) identical values of $\mathbf{X}_i$. The CIA allows us to make a claim about causality within these groups.

We match untreated observations to treated observations using $\mathbf{X}_i$ and calculate the average treatment effect by comparing $Y_i(1)$ to outcomes for "matched" untreated individuals $Y_i(0)$.

The classic tradeoff



We want to minimize bias in our estimates by finding a match that most closely approximates each treated unit but we don't want to overly restrict the definition of matching so as to require excluding too many units or producing a sample that does not reflect our originally defined population.

Propensity scores (I)

Phase I:

1. Investigate the selection process explicitly by fitting a "selection model":
    - Fit a logistic model, with treatment group membership as outcome, and predictors you believe describe the process of selection explicitly:

$$D_i = \frac{1}{1 + e^{-\mathbf{X}_i \theta_i}}$$

2. Use selection model to estimate fitted probability of selection into treatment $(\hat{p})$ for each participant

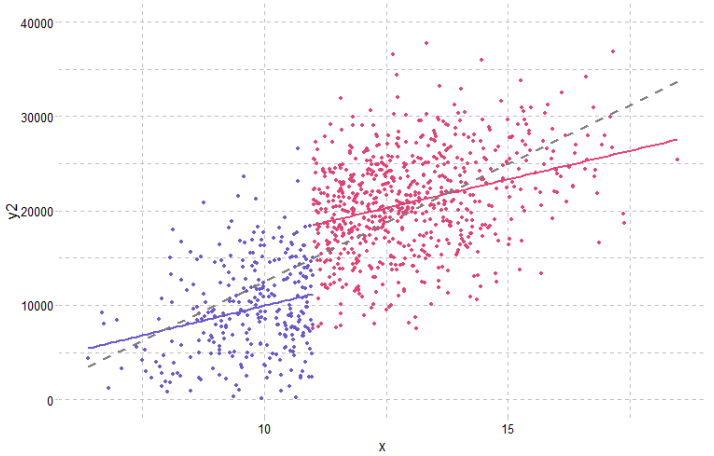
- Output these “propensity scores” into your dataset
- They summarize the probability that each participant will be selected into the treatment, given their values on the covariates.
- They are the best non-linear composite of the covariates for describing selection into experimental conditions, given your choice of covariates.

Propensity scores (II)

Phase II:

1. Stratify the sample using the propensity scores:
    - Enforce overlap:
        - Drop control units with $\hat{p}$ below the minimum propensity score in the treatment group
        - Drop treated units with $\hat{p}$ above the maximum propensity score in the control group
    - Rule of thumb: as few as five strata may remove up to 90% of the observed bias
2. Within each stratum, check the balancing condition has been satisfied:
    - On the propensity scores themselves
    - On each of the covariates separately
3. If the balancing condition has not been met:
    - Re-stratify, combining or splitting strata, until balancing condition is met
    - If this fails, re-specify the selection model (nonlinear terms, interactions?) and start again
4. Once you have achieved balance, estimate treatment effect within each stratum, and average up

Difference from OLS?



1. It's not that different
2. Regression approaches make strong assumptions about the equivalence of treatment effects across different groups (can be solved with interactions or non-parametric approaches)...
3. including groups **for which there is no common support** (a case of predicting outside of data range and cannot be solved for with interactions)
4. Each additional covariate makes additional assumptions about equivalence of effects across groups (results in X-factorial potential interactions)
5. We've assumed that errors are homoscedastic across groups and pooled all of that variation to obtain a common standard error

Good/bad candidates?

Remember, the only reason any of this is worth doing is if there is a clear case where (a) selection on the observables has occurred; and (b) the source of the selection can not be modeled via exogenous variation in treatment

With a partner:

- Identify 2-3 examples of situations in which selection into treatment or the sample can be addressed via an approach from the matching family
- Identify 2-3 examples in which a matching approach would be suspect to persistent bias in estimates

Two princes

credit: not sure where original is from?

A family affair

Matching approaches include:

- Stratification
- Propensity-Score Matching (PSM)
  - Nearest neighbor (Euclidian or Mahalanobis distance)
  - Kernel matching
  - Machine learning assisted matching
  - Calipers
  - With or without replacement
- Inverse Probability Weighting (IPW)
- Coarsened Exact Matching (CEM)
- Inexact Matching
- Synthetic controls in DD strategies
- Doubly-robust (e.g., matching and weighted) estimates
- And combinations of these and more...

...**the approach matters and requires close attention to procedure**

Strengths/limitations of approaches

Approach	Strengths	Limitations
Propensity-score nearest neighbor matching w/ calipers and replacement	- Simulates ideal randomized experiment - Limits dimensionality problem - Calipers restrict poor matches - Replacement takes maximal advantage of available data	- May generate poor matches - Model dependent Lacks transparency; PS in aribtrary units - Potential for bias (King & Nielsen, 2019)
Propensity-score stratification	- Simulates block-randomized experiment - Limits dimensionality problem	- May produce worse matches than nearest neighbor - Lacks transparency; stratum arbitrary
Inverse probablity (PS) matching	- Retains all original sample data - Corrects bias of estimate with greater precision than matching/stratification	- Non-transparent/a-theoretical
Coarsened Exact Matching	- Matching variables can be pre-specified (and pre-registered) - Matching substantively driven Transparent matching process - Eliminates same bias as propensity score if SOO occurs	- May generate poor matches depending on how coarsened variables are - May lead to disgarding large portions of sample

Synthesis and wrap-up

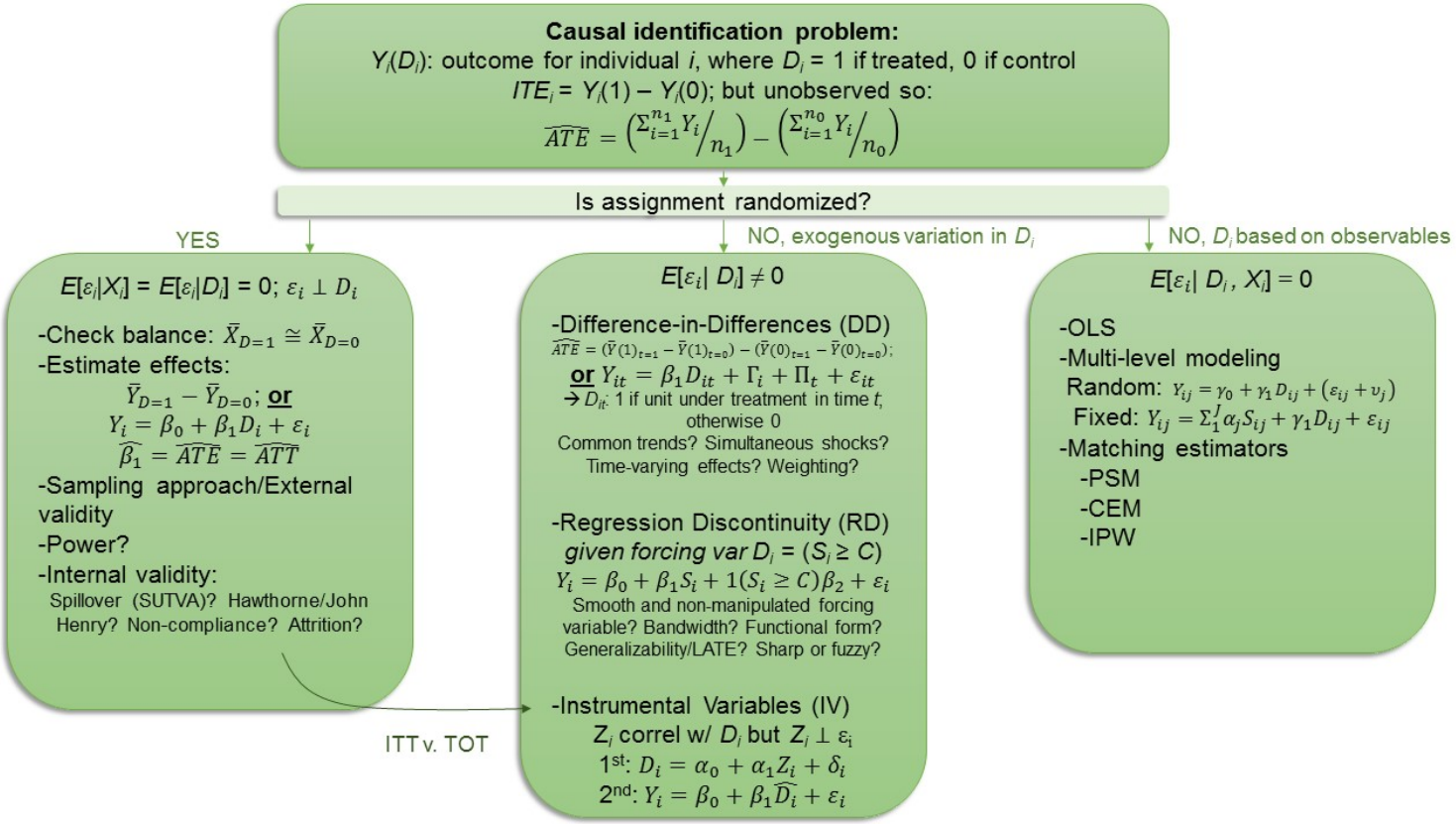
Goals

1. Conduct IV analysis in simplified data and interpret results

2. Assess basic assumptions of IV design in an experimental setting with imperfect compliance

3. Describe the conceptual approach of using selection on observables to defend causal inferences about the effects of a treatment

Roadmap


Causal identification problem:
$Y_i(D_i)$: outcome for individual i, where D_i = 1 if treated, 0 if control
$ITE_i = Y_i(1) - Y_i(0)$; but unobserved so:
$\widehat{ATE} = \left(\Sigma_{i=1}^{n_1} Y_i / n_1\right) - \left(\Sigma_{i=1}^{n_0} Y_i / n_0\right)$
Is assignment randomized?
YES
NO, exogenous variation in D_i
NO, D_i based on observables
$E[\varepsilon_i|X_i] = E[\varepsilon_i|D_i] = 0$; $\varepsilon_i \perp D_i$
-Check balance: $\bar{X}_{D=1} \cong \bar{X}_{D=0}$
-Estimate effects:
$\bar{Y}_{D=1} - \bar{Y}_{D=0}$; or
$Y_i = \beta_0 + \beta_1 D_i + \varepsilon_i$
$\widehat{\beta_1} = \widehat{ATE} = \widehat{ATT}$
-Sampling approach/External validity
-Power?
-Internal validity:
Spillover (SUTVA)? Hawthorne/John Henry? Non-compliance? Attrition?
$E[\varepsilon_i| D_i] \neq 0$
-Difference-in-Differences (DD)
$\widehat{ATE} = (\bar{Y}(1)_{t=1} - \bar{Y}(1)_{t=0}) - (\bar{Y}(0)_{t=1} - \bar{Y}(0)_{t=0})$;
or $Y_{it} = \beta_1 D_{it} + \Gamma_i + \Pi_t + \varepsilon_{it}$
→ D_{it}: 1 if unit under treatment in time t, otherwise 0
Common trends? Simultaneous shocks? Time-varying effects? Weighting?
-Regression Discontinuity (RD)
given forcing var D_i = ($S_i \geq C$)
$Y_i = \beta_0 + \beta_1 S_i + 1(S_i \geq C)\beta_2 + \varepsilon_i$
Smooth and non-manipulated forcing variable? Bandwidth? Functional form? Generalizability/LATE? Sharp or fuzzy?
-Instrumental Variables (IV)
Z_i correl w/ D_i but $Z_i \perp \varepsilon_i$
1st: $D_i = \alpha_0 + \alpha_1 Z_i + \delta_i$
2nd: $Y_i = \beta_0 + \beta_1 \widehat{D_i} + \varepsilon_i$
ITT v. TOT
$E[\varepsilon_i| D_i, X_i] = 0$
-OLS
-Multi-level modeling
Random: $Y_{ij} = \gamma_0 + \gamma_1 D_{ij} + (\varepsilon_{ij} + \upsilon_j)$
Fixed: $Y_{ij} = \Sigma_1^J \alpha_j S_{ij} + \gamma_1 D_{ij} + \varepsilon_{ij}$
-Matching estimators
-PSM
-CEM
-IPW

To-Dos

Week 8: Matching

Readings:

- Murnane and Willett, Chapter 12
- Diaz & Handa - evaluation of Mexico's PROGRESA program
- Additional readings: Cunningham, Ch. 5; Dehejia & Wahba (2002); Iacus, King & Porro (2011); King et al. (2011); King & Nielsen (2019)

Assignments Due

- **DARE 4** (last one!!!)
    - Due 9:00am, Feb. 28
- **Final Research Project**
    - Presentation, March 8
    - Paper, March 18 (submit early [March 10] for feedback)

Feedback

Plus/Deltas

Front side of index card

Clear/Murky

On back