Causal Inference with Observational Data

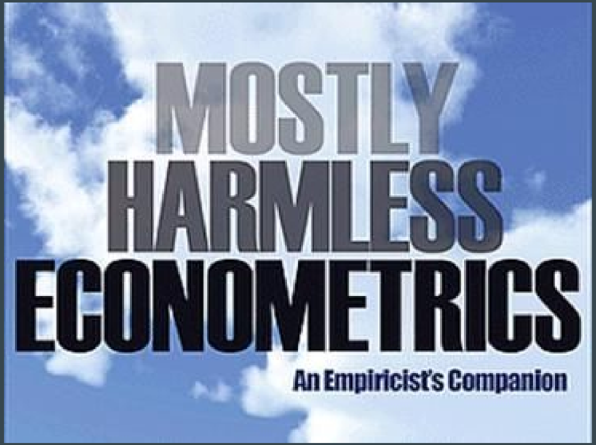
Dean and Friends
Dropkick, Monica, Zhe, and Jackie

Topics

- Why Natural Experiments
- Difference in Difference - Z
- Instrumental Variables - M
- Survival Models - M
- Propensity Score Matching - J
- Tobit Models - D
- Heckman Models - D

General Motivation

- Interventions: policy / management / new product
- Behavior-centric data
    - many, many unobserved features
- Central to all causal statements:
  (a) identification/counterfactual strategy
  (b) assumptions defending identification
    - many, many subtleties

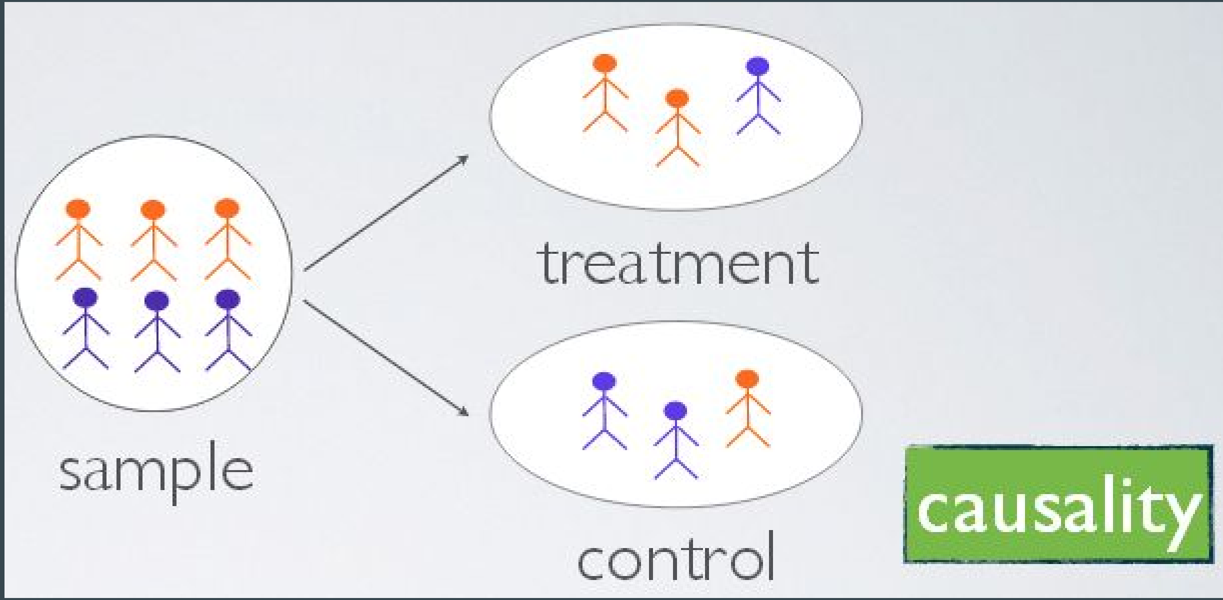


Ideally: Counterfactuals

Counterfactuals: “we only observe what actually happens”

- Naive estimate of program effect:
  E[Program] - E[None]
- With observed data:
  E[Program|D = 1] - E[None|D = 0]
- “[D = 1], [D=0]”
  - random? Probably not; related to unobservables

Simplest Approach



Next Simplest: Assume No Unobservables

Regression with controls

- e.g. Effect of a job training program
    - Basic demographics, income, education

Machine Learning

- maximizes predictive fit
- estimate of effect the same:
  $E[\ Program|D = 1\ ] - E[\ None|D = 0\ ]$

After That: Natural & Quasi-Experiments

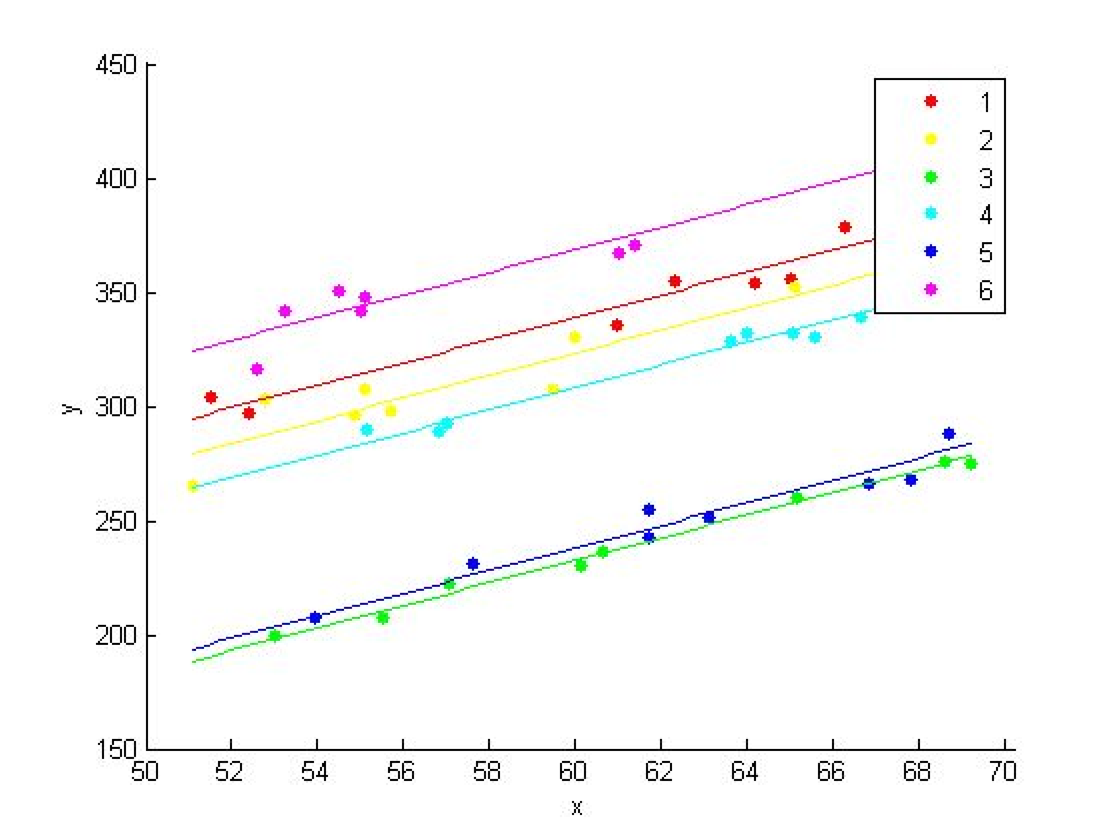
Natural Separation of Groups

US Military Draft on Random Social Security Numbers

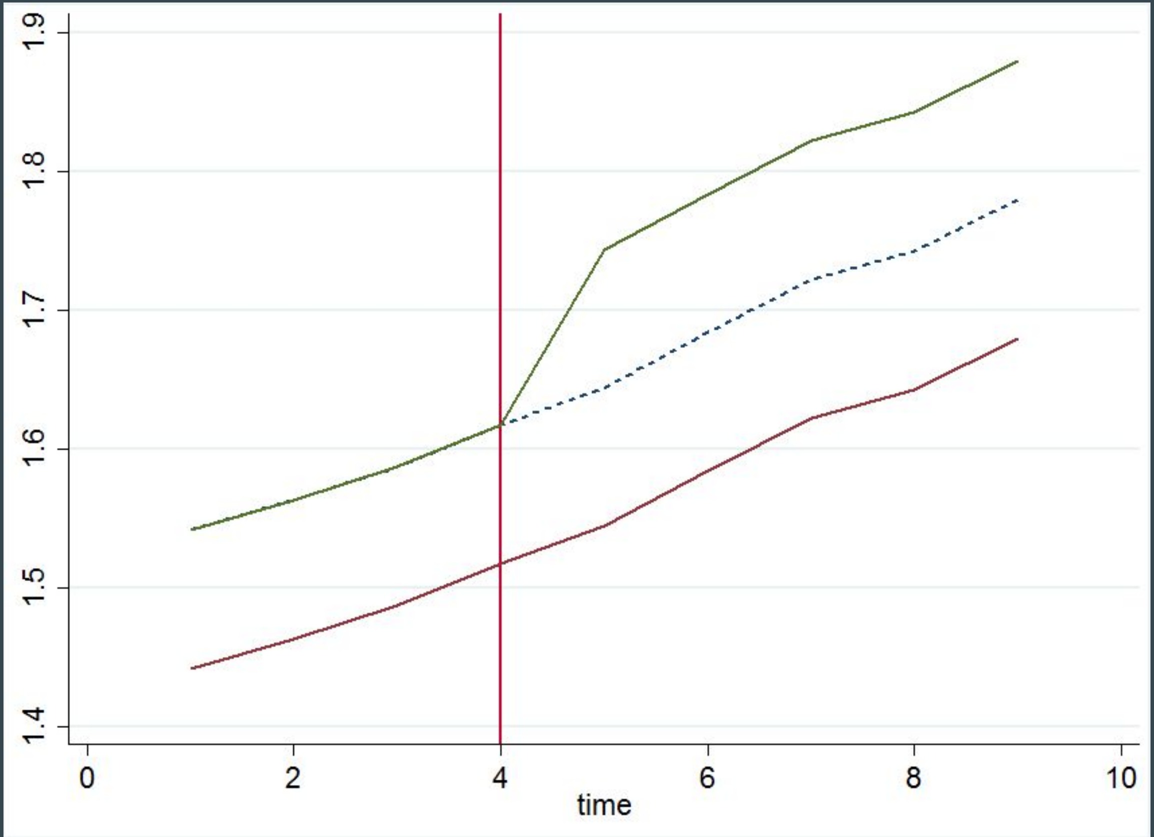
odd/even?

→ estimate effect of military on career outcomes

Fixed Effects & Differences-in-Differences



Fixed Effects & Differences-in-Differences



Fixed Effects & Differences-in-Differences

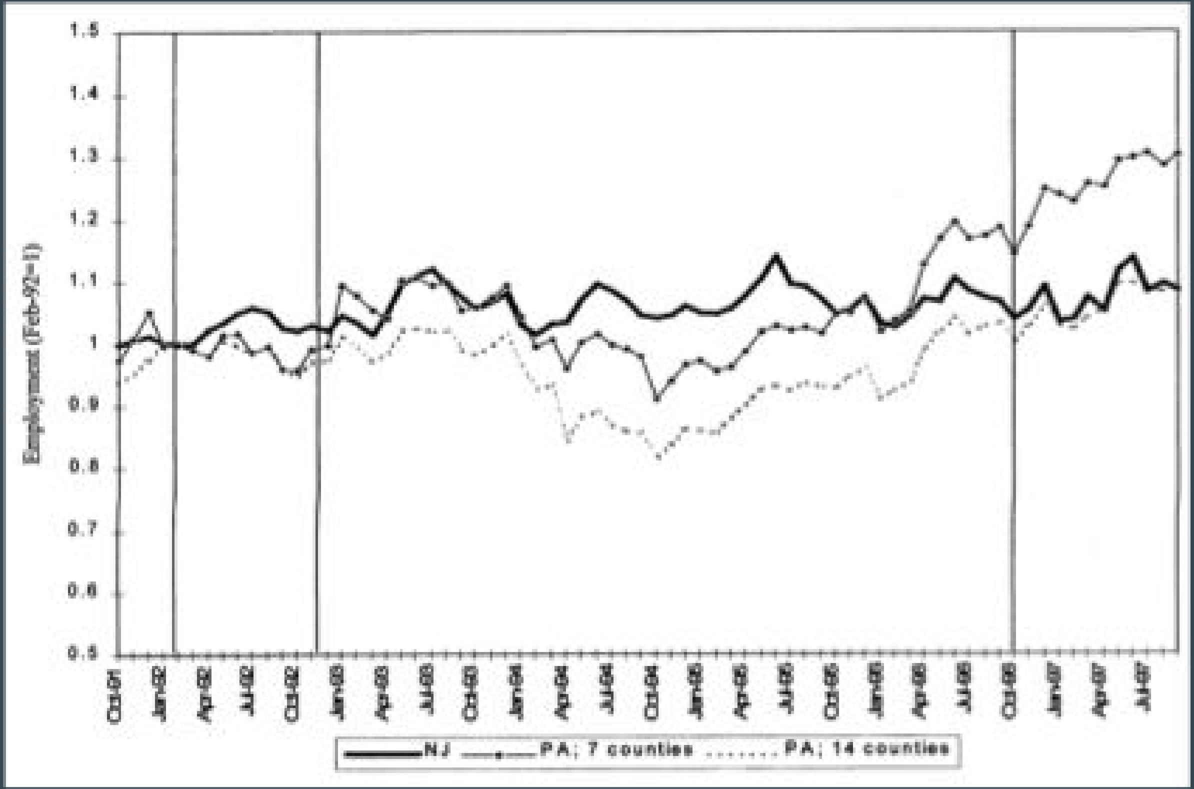
Not usually that simple.

Examples:

Effect of minimum-wage increase in NJ
(uses eastern PA as counterfactual)

Effect of Uber/Lyft on drunk driving homicides
(uses time-based diff-in-diff)

Main Task: Defending Identification Strategy



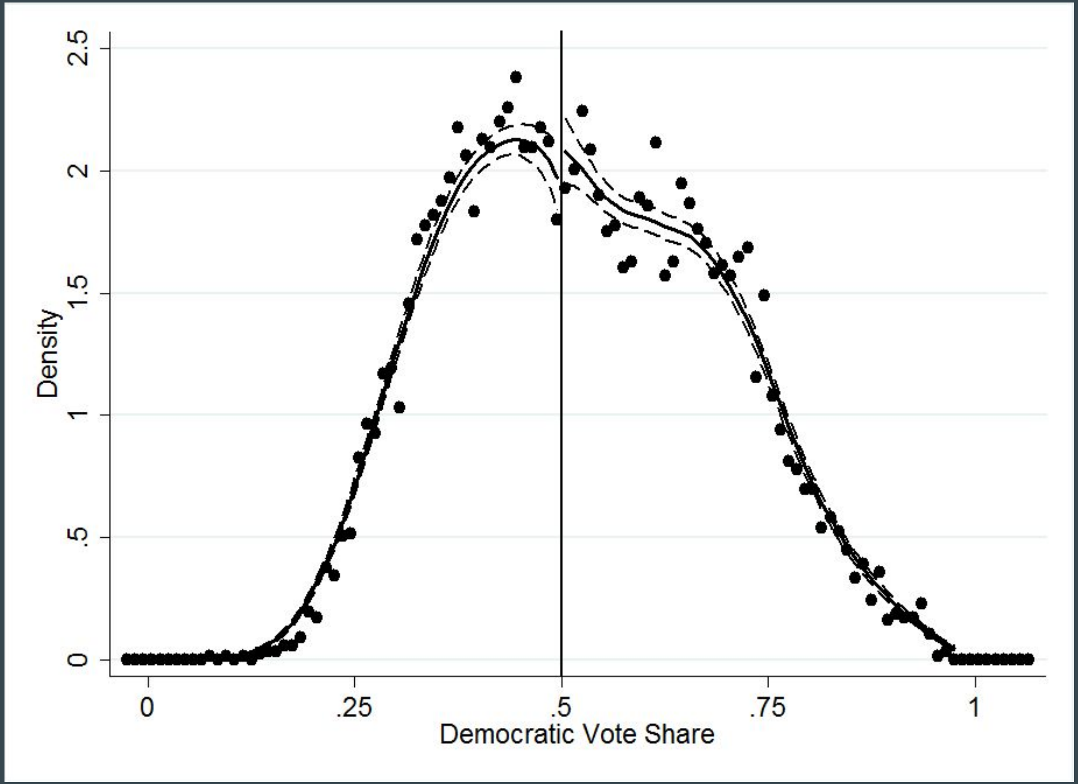
Discontinuity/Threshold Design

Scholarship Effect

Vote Effect

Class Size Effect

Flaw: ***local*** estimate



Instrumental Variables

- Want to look at the effect of treatment on outcome
  - Controlled experiments often not viable in social sciences
  - Usually working with observational data
- Potential issue with classical regression: endogeneity (explanatory variables correlated with error term)
- To try and avoid this, use an instrument for treatment explanatory variable of interest.
- An instrument must be:
  - Correlated with explanatory variable of interest
  - Uncorrelated with error term

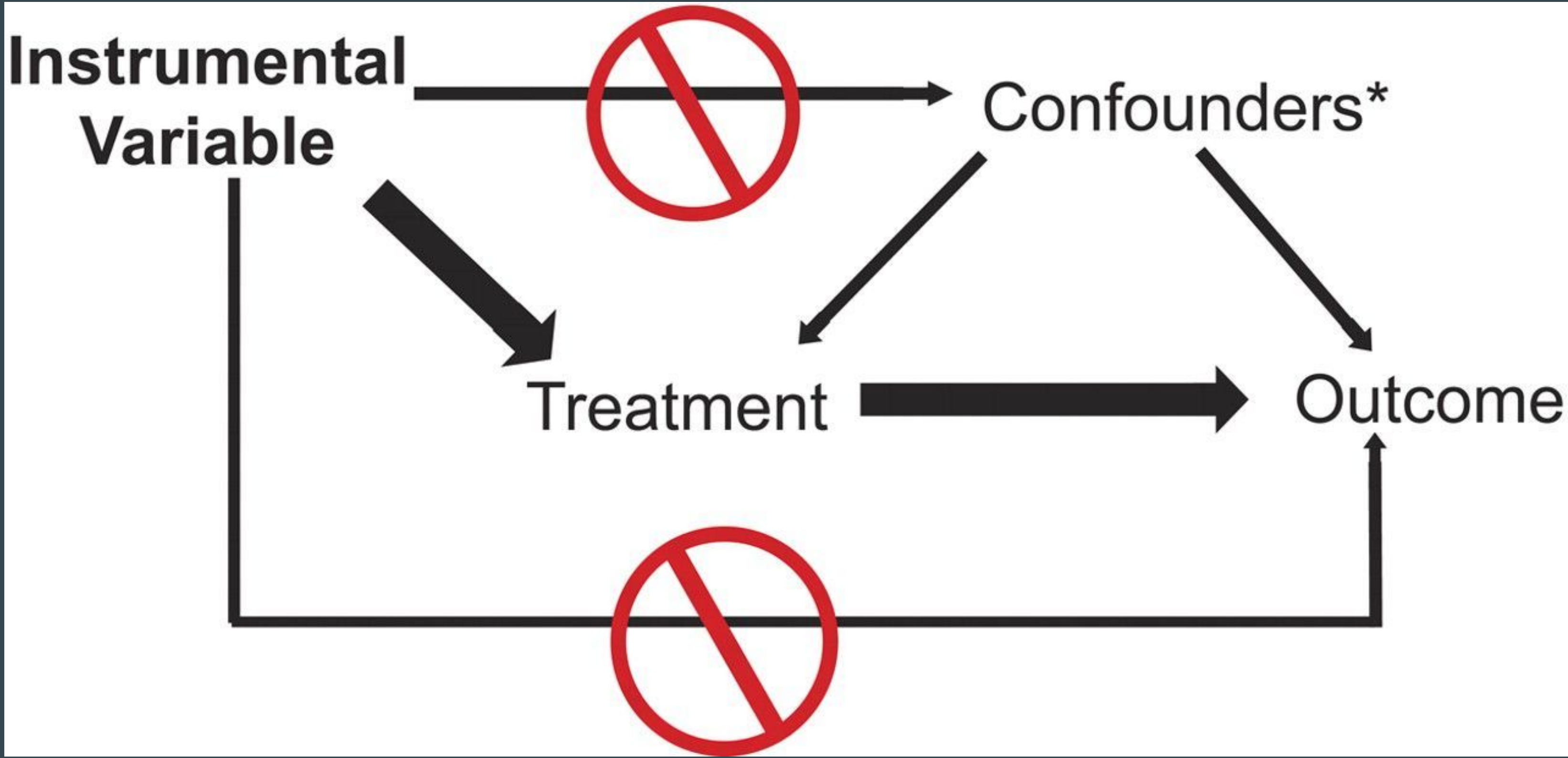
Instrumental Variables

$$y = X\boldsymbol{\beta} + \varepsilon$$

- Replace X with predicted values of X that are
  - Related to actual X
  - Uncorrelated with ε
- Estimation: most commonly 2SLS

$$X = Z\gamma + u$$
$$\rightarrow y = X\boldsymbol{\beta} + \varepsilon$$

- Where to find instruments: policy reforms, geographic differences
- Problems with IV: exclusion restriction untestable, weak instruments cause problems


Instrumental Variable
Confounders*
Treatment
Outcome

Survival Models

- What is it?
  - Analysis of waiting times until an event occurs
  - Usually used when event only occurs once
  - E.g. time until death, first marriage, first birth, first divorce...
  - (multiple occurrences: see event history analysis)
- Stuff we are interested in estimating
  - Survival function S(t)
    - Probability that time of event T is greater than t
    - $S(t) = P(T>=t) = 1 - F(t)$
  - Hazard function h(t)
    - Instantaneous death/failure rate
    - (-) slope of the log of S(t)

Survival Models

General form:

$$\log(h(t)) = \log(h(0)) + X\beta$$
$$h(t) = h(0)\exp(X\beta)$$

How to estimate S(t) / h(t):

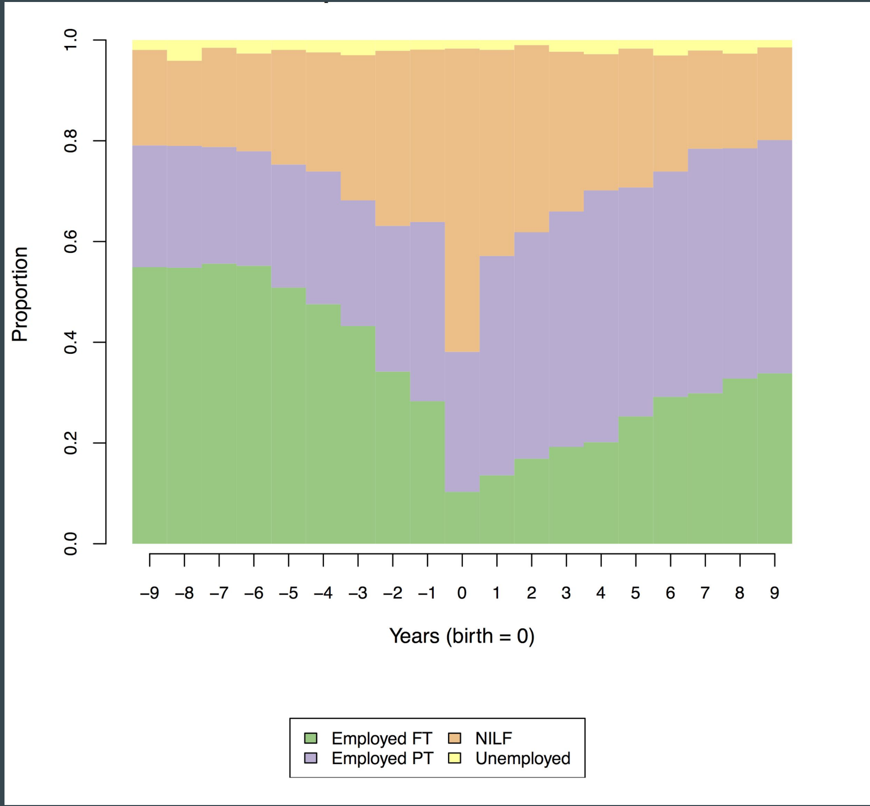
- Non-parametric (Kaplein-Meier)
- Semi- parametric (Cox proportional hazards)
- Parametric (Poisson regression)

Censoring: often observations are censored i.e. T> t(observation)

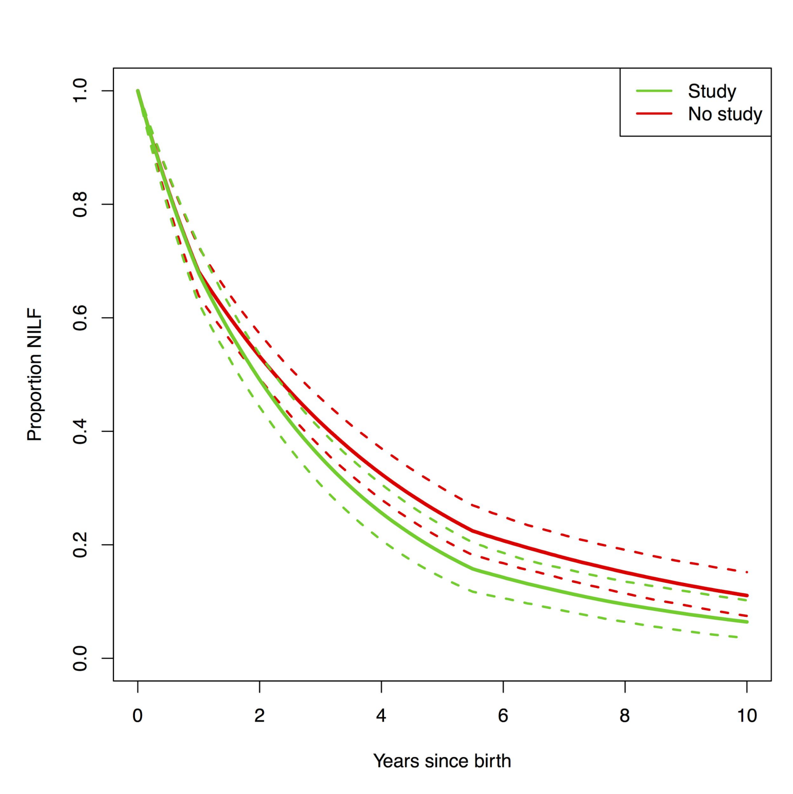
- Can still use to get info about population exposure, but not occurences

Mothers returning to study

Work patterns before and after birth



Proportion not in labor force by study group



What if the treatment and control groups look very different?

What we observe

- The average outcome of the treated individuals ***conditional on them receiving*** the treatment or intervention
- The average outcome of the untreated individuals ***conditional on them not receiving*** the treatment or intervention

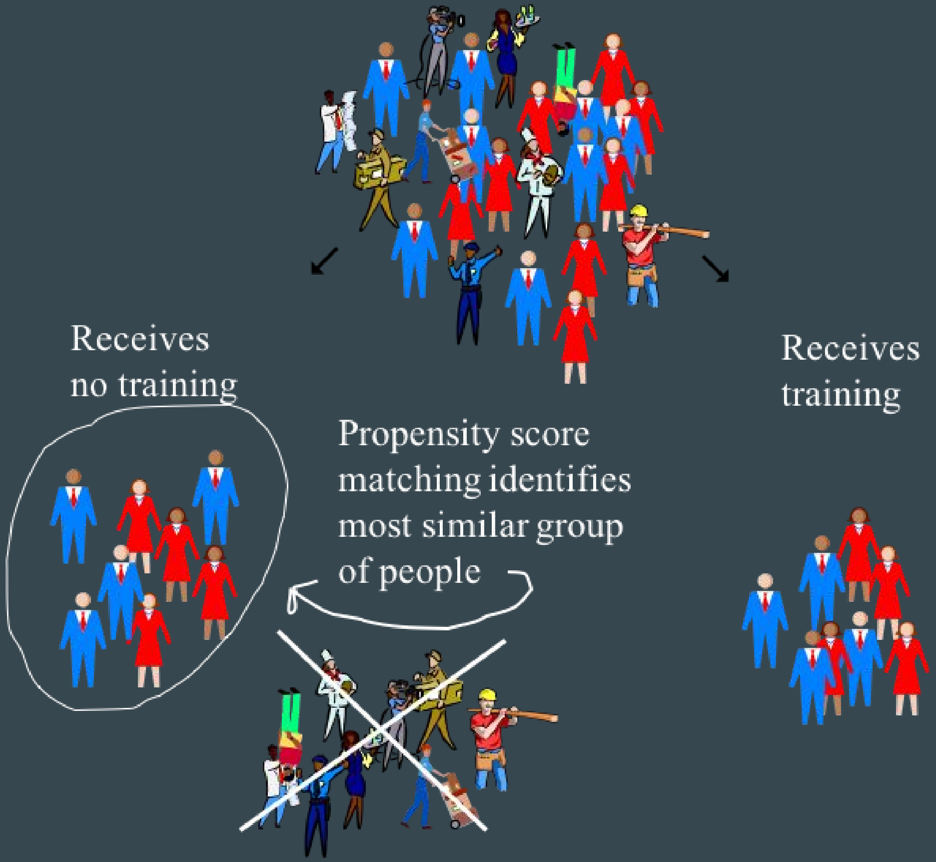
These are not directly comparable!

What we want

- The ***average difference in potential outcomes*** for each individual if they did versus did not receive treatment

Slide credit: Jennifer Hill

If we think we know how these groups differ, we can match them

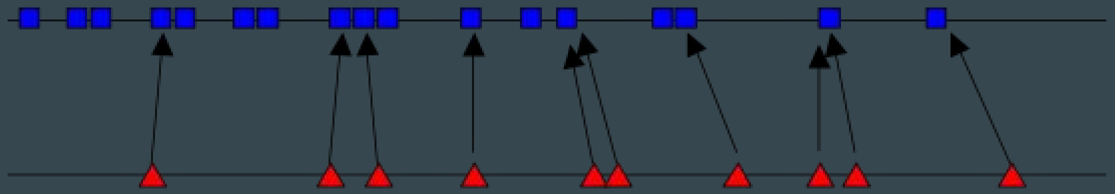


- This requires assuming ***ignorability***: that we have measured all the covariates that we need to predict how likely an individual is to receive the treatment
- Build a classifier on the likelihood of receiving treatment
  - ***Match*** individuals from the treatment to similar individuals in the control group
  - “One-number summary”
- We never really believe we have measured all the relevant covariates!

Slide credit: Jennifer Hill

Better than doing nothing, worse than a field test

matched to	Person	Treat	Educ.	Age	Y0	Y1	Y	linear pscore	pscore
11	1	1	1	26	10	14	14	0.77	0.68
12	2	1	1	21	8	12	12	0.60	0.65
11	3	1	1	30	12	16	16	0.91	0.71
12	4	1	1	19	8	12	12	0.53	0.63
8	5	1	0	25	6	10	10	-0.71	0.33
7	6	1	0	22	4	8	8	-1.14	0.24
	7	0	0	21	4	8	4	-1.29	0.22
	8	0	0	26	6	10	6	-0.56	0.36
	~~9~~	~~0~~	~~0~~	~~28~~	~~8~~	~~12~~	~~8~~	~~-0.27~~	~~0.43~~
	~~10~~	~~0~~	~~0~~	~~20~~	~~4~~	~~8~~	~~4~~	~~-1.43~~	~~0.19~~
	11	0	1	26	10	14	10	0.77	0.68
	12	0	1	21	8	12	8	0.60	0.65
	~~13~~	~~0~~	~~0~~	~~16~~	~~2~~	~~6~~	~~2~~	~~-2.01~~	~~0.12~~
	~~14~~	~~0~~	~~0~~	~~15~~	~~1~~	~~5~~	~~1~~	~~-2.15~~	~~0.10~~



- Makes fewer (parametric) assumptions than controlling for the covariates in a standard regression
- Needs at least some overlap between the treatment and control groups or matching will fail (But at least you will know that it failed!)
- Can try many different model specifications to predict propensity of receiving treatment--use the one that gives you the most balance
- Great for inverse probability of treatment weighting!

Slide credit: Jennifer Hill

Tobit Models (Corner solution models)

- When do you use them?
    - When your response or ‘y’ variable is zero for a nontrivial fraction of the population but is roughly continuously distributed over positive values.
    - An example is the amount an individual spends on alcohol in a given month
- Why does a linear regression not work?
    - Negative values
    - Bunching around zero - conditional distribution not normal
    - The x’s don't really have a constant marginal effect on y

Tobit Models

$$y^* = \beta_0 + \mathbf{x}\boldsymbol{\beta} + u, \quad u|\mathbf{x} \sim \text{Normal}(0, \sigma^2)$$

$$y = \max(0, y^*).$$

Tobit Models - Interpretation

- It's really hard!
- We care about two things in particular
  - E(y|y >0,x) - for the subpopulation which is positive
  - E(y|x) - for the entire population
- We then take the partial derivatives

$$
\begin{aligned}
E(y|y>0,\mathbf{x}) &= \mathbf{x}\beta + E(u|u>-\mathbf{x}\beta) \\
&= \mathbf{x}\beta + \sigma E[(u/\sigma|u/\sigma > -\mathbf{x}\beta/\sigma)] \\
&= \mathbf{x}\beta + \sigma\phi(\mathbf{x}\beta/\sigma)/\Phi(\mathbf{x}\beta/\sigma) \\
&= \mathbf{x}\beta + \sigma\lambda(\mathbf{x}\beta/\sigma)
\end{aligned}
$$

$$
\begin{aligned}
\mathrm{E}(y|\mathbf{x}) &= P(y>0|\mathbf{x})\mathrm{E}(y|y>0,\mathbf{x}) \\
&= \Phi(\mathbf{x}\beta/\sigma)\mathrm{E}(y|y>0,\mathbf{x}).
\end{aligned}
$$

$$
\mathrm{E}(y|y>0,\mathbf{x}) = \mathbf{x}\beta + \sigma\lambda(\mathbf{x}\beta/\sigma),
$$

Tobit Models - Examples

- 753 women in sample - annual hours worked
  - 428 worked for a wage
  - 325 stayed at home and worked 0 hours
- Amount spent on healthcare annually
  - Some (very healthy) people do not visit hospitals or doctors in certain years
- Amount of alcohol consumed monthly

Heckman Models

- Deals with truncated data (incidental truncation)
  - We restrict attention to a subset of the population before sampling
  - The ‘omitted variable’ in this case is how people were selected into the sample
  - Ie: It is NOT a random sample
- Assume that the underlying population satisfies some linear regression model
- Example: wage of married women

$$y = \beta_0 + \mathbf{x}\boldsymbol{\beta} + u, \quad u|\mathbf{x} \sim \text{Normal}(0, \sigma^2).$$

The Two stages

- 1st Stage:
  - Estimate probability of being included in sample (logistic/probit)
  - For wages of working women.... Education?
  - Must include a variable that causes selection in sample but does not explain your ‘y’
  - Compute what is called the ‘inverse Mills ratio’ for each observation
- 2nd Stage:
  - Estimate the regression you would have run but add the ‘inverse Mills ratio’ as a predictor in the model
  - If the coefficient on this ‘inverse mills ratio’ is 0 then you ‘can’ say that there is no sample selection bias and can use a standard linear regression.

Questions?