Lecture 1: Causal Identification

POL-GA 1251
Quantitative Political Analysis II
Prof. Cyrus Samii
NYU Politics

January 22, 2018

Tue:

- Discuss overview of class and syllabus.
- Explain what "causal identification" means.
- Introduce the "potential outcomes" framework
- Relate it to linear regression model.

Thu:

- Study an idealized randomized experiment.
- Explain *estimation* concepts (estimand, estimators, bias, consistency, efficiency).
- Explain *statistical inference* concepts (sampling distribution, randomization distribution, CLT, confidence intervals, p-value).
- First homework distributed.

Model of quantitative research process:

- Theory motivates causal hypothesis or target of inference:
  - *H: manipulating X results in (...) effect on Y.*
- Hypothesis, statistical theory, and substantive theory motivate a research design:
  - Operationalize X and Y.
  - Define ways to get optimal variation in X and Y given constraints.
- Research design and statistical theory motivate analysis plan:
  - Optimal estimation strategy, given constraints.
  - Optimal testing strategy, given constraints.

Model of quantitative research process:

- Theory motivates causal hypothesis or target of inference:
  - *H: manipulating X results in (...) effect on Y.*
- Hypothesis, statistical theory, and substantive theory motivate a research design:
  - Operationalize X and Y.
  - Define ways to get optimal variation in X and Y given constraints.
- Research design and statistical theory motivate analysis plan:
  - Optimal estimation strategy, given constraints.
  - Optimal testing strategy, given constraints.

Identification refers generally to sufficiency for drawing a conclusion given *the type of data* that are available.

Identification refers generally to sufficiency for drawing a conclusion given *the type of data* that are available.

Distinguish from sampling uncertainty:

Identification refers generally to sufficiency for drawing a conclusion given *the type of data* that are available.

Distinguish from sampling uncertainty:

Manski (1995, 4) : "Studies of identification seek to characterize the conclusions that could be drawn if one could...obtain an unlimited number of observations. Identification problems cannot be solved by gathering more of the same kind of data."

Identification refers generally to sufficiency for drawing a conclusion given *the type of data* that are available.

Distinguish from sampling uncertainty:

Manski (1995, 4) : "Studies of identification seek to characterize the conclusions that could be drawn if one could...obtain an unlimited number of observations. Identification problems cannot be solved by gathering more of the same kind of data."

Informally, a parameter is "identified" if, as the sample goes to infinity, the data come to require that the parameter equals one value.

Identification refers generally to sufficiency for drawing a conclusion given *the type of data* that are available.

Distinguish from sampling uncertainty:

Manski (1995, 4) : "Studies of identification seek to characterize the conclusions that could be drawn if one could...obtain an unlimited number of observations. Identification problems cannot be solved by gathering more of the same kind of data."

Informally, a parameter is "identified" if, as the sample goes to infinity, the data come to require that the parameter equals one value.

In this class we focus on causal identification. This is a specific application of the general idea of "identification," distinct from some other applications:

Alternative application of “identification” (I):

Suppose someone says...

...they prefer Cruz over Trump, and

...they prefer Rubio over Cruz.

Does this information (data) identify the person’s preference ordering over these three candidates?

Alternative application of "identification" (II):

Suppose none of the coefficients below are equal to zero but the error terms (last ones) are iid mean zero draws. Which system identifies its coefficients?

$$x_t = \alpha_1^a + \alpha_2^a y_t + \nu_t^a$$
$$y_t = \beta_1^a + \beta_2^a x_t + \varepsilon_t^a$$

$$x_t = \alpha_1^b + \alpha_2^b y_t + \alpha_3^b w_t + \alpha_4^b v_t + \nu_t^b$$
$$y_t = \beta_1^b + \beta_2^b x_t + \beta_3^b w_t + \beta_4^b v_t + \varepsilon_t^b$$

$$x_t = \alpha_1^c + \alpha_2^c y_t + \alpha_3^c w_t + \nu_t^c$$
$$y_t = \beta_1^c + \beta_2^c x_t + \beta_4^c v_t + \varepsilon_t^c$$

Causal identification refers to sufficiency for drawing a conclusion about a *causal effect* given the type of data at hand.

Causal identification refers to sufficiency for drawing a conclusion about a *causal effect* given the type of data at hand.

This class focuses on strategies for non-parametric causal identification:

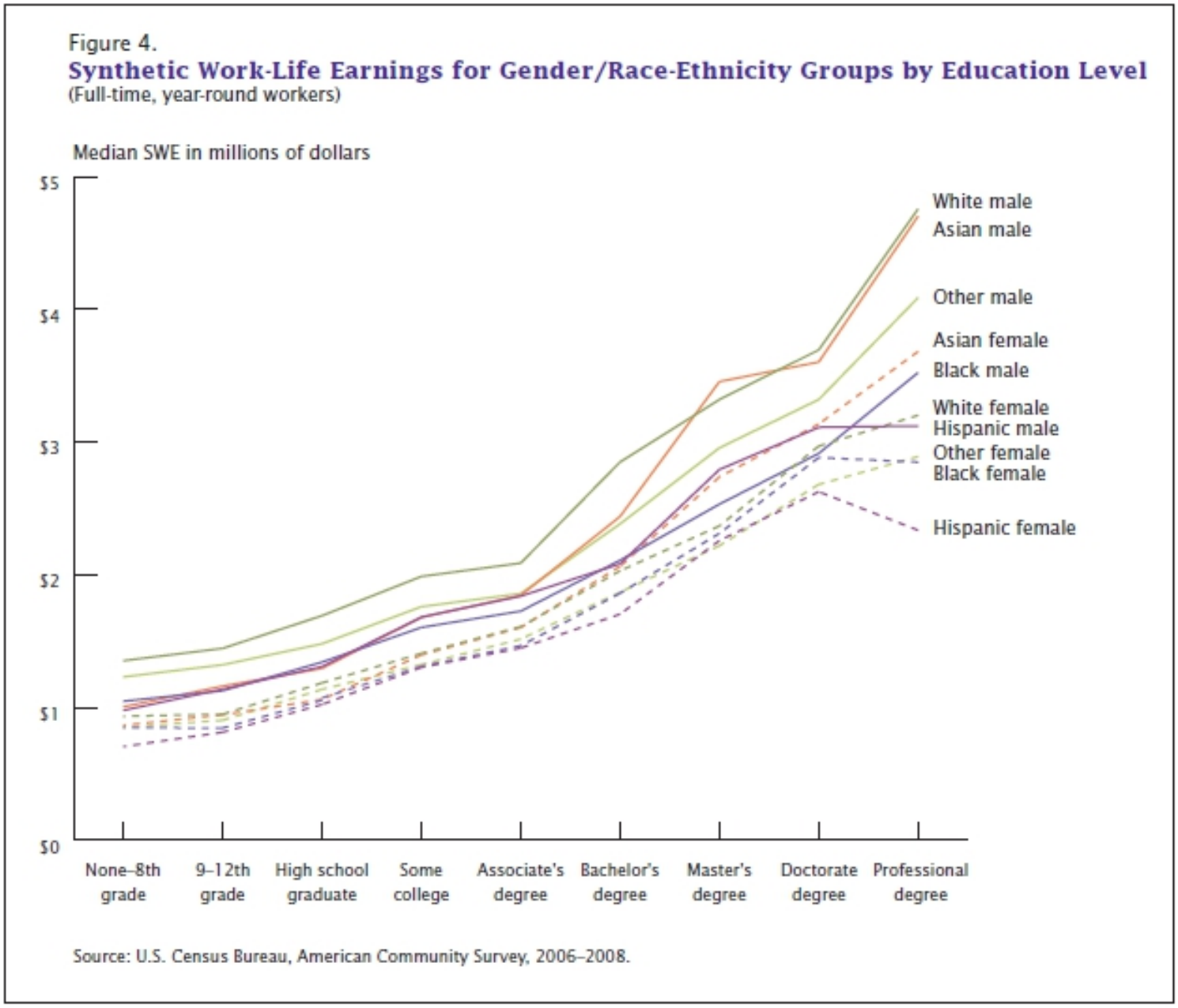
Causal identification refers to sufficiency for drawing a conclusion about a *causal effect* given the type of data at hand.

This class focuses on strategies for non-parametric causal identification:

Angrist and Krueger (1999):

> *The combination of a clearly labeled source of identifying variation in a causal variable and the use of a particular econometric technique to exploit this information is what we call an identification strategy.*

Does this plot identify the effect of education on income?



A taste of non-parametric causal identification analysis:

A taste of non-parametric causal identification analysis:

- Suppose a sample $\mathscr{S}$ indexed by $i = 1, ..., N$.
- We have an experiment with a binary treatment, where $Z_i = 0, 1$, is an indicator for treatment assigned.

A taste of non-parametric causal identification analysis:

- Suppose a sample $\mathscr{S}$ indexed by $i = 1, ..., N$.
- We have an experiment with a binary treatment, where $Z_i = 0, 1$, is an indicator for treatment assigned.
- M members of $\mathscr{S}$ are randomly assigned to treatment ($Z_i = 1$), the rest are assigned to control ($Z_i = 0$).

- However, compliance to treatment assignment is not perfect, and so each unit is also characterized by a "treatment received" function, $D_i(z) = 0, 1$, for treatment assignment $Z_i = z$.
- The treatment received is thus $D_i = Z_i D_i(1) + (1 - Z_i) D_i(0)$.

- However, compliance to treatment assignment is not perfect, and so each unit is also characterized by a "treatment received" function, $D_i(z) = 0, 1$, for treatment assignment $Z_i = z$.
- The treatment received is thus $D_i = Z_i D_i(1) + (1 - Z_i) D_i(0)$.
- All members in $\mathscr{S}$ must be in one out of four subgroups:

Subgroup	$D_i(0)$	$D_i(1)$
Always-takers	1	1
Compliers	0	1
Never-takers	0	0
Defiers	1	0

- However, compliance to treatment assignment is not perfect, and so each unit is also characterized by a "treatment received" function, $D_i(z) = 0, 1$, for treatment assignment $Z_i = z$.
- The treatment received is thus $D_i = Z_i D_i(1) + (1 - Z_i) D_i(0)$.
- All members in $\mathscr{S}$ must be in one out of four subgroups:

Subgroup	$D_i(0)$	$D_i(1)$
Always-takers	1	1
Compliers	0	1
Never-takers	0	0
Defiers	1	0

- Unit i's outcome depends on the treatment that unit i received, which can characterize by a "potential outcome" function, $Y_i(d)$, for treatment received, $D_i = d$.

The experiment is run and outcomes are recorded for each $i \in \mathscr{S}$.

Are this experiment and the assumptions we have made sufficient to identify the average causal effect of treatment received (D_i) for $\mathscr{S}$?

How about the average causal effect of treatment assigned (Z_i) for $\mathscr{S}$?

The experiment is run and outcomes are recorded for each $i \in \mathscr{S}$.

Are this experiment and the assumptions we have made sufficient to identify the average causal effect of treatment received (D_i) for $\mathscr{S}$?
No.

How about the average causal effect of treatment assigned (Z_i) for $\mathscr{S}$?
Yes, the "ITT."

The Road Not Taken

Two roads diverged in a yellow wood,
And sorry I could not travel both
And be one traveler, long I stood
And looked down one as far as I could
To where it bent in the undergrowth;

Then took the other, as just as fair,
And having perhaps the better claim,
Because it was grassy and wanted wear;
Though as for that the passing there
Had worn them really about the same,

And both that morning equally lay
In leaves no step had trodden black.
Oh, I kept the first for another day!
Yet knowing how way leads on to way,
I doubted if I should ever come back.

I shall be telling this with a sigh
Somewhere ages and ages hence:
Two roads diverged in a wood, and I—
I took the one less traveled by,
And that has made all the difference.

Potential outcomes model of causality
(Neyman, 1923; Rubin, 1974, 1978)

A causal effect is a contrast between “potential outcomes.”

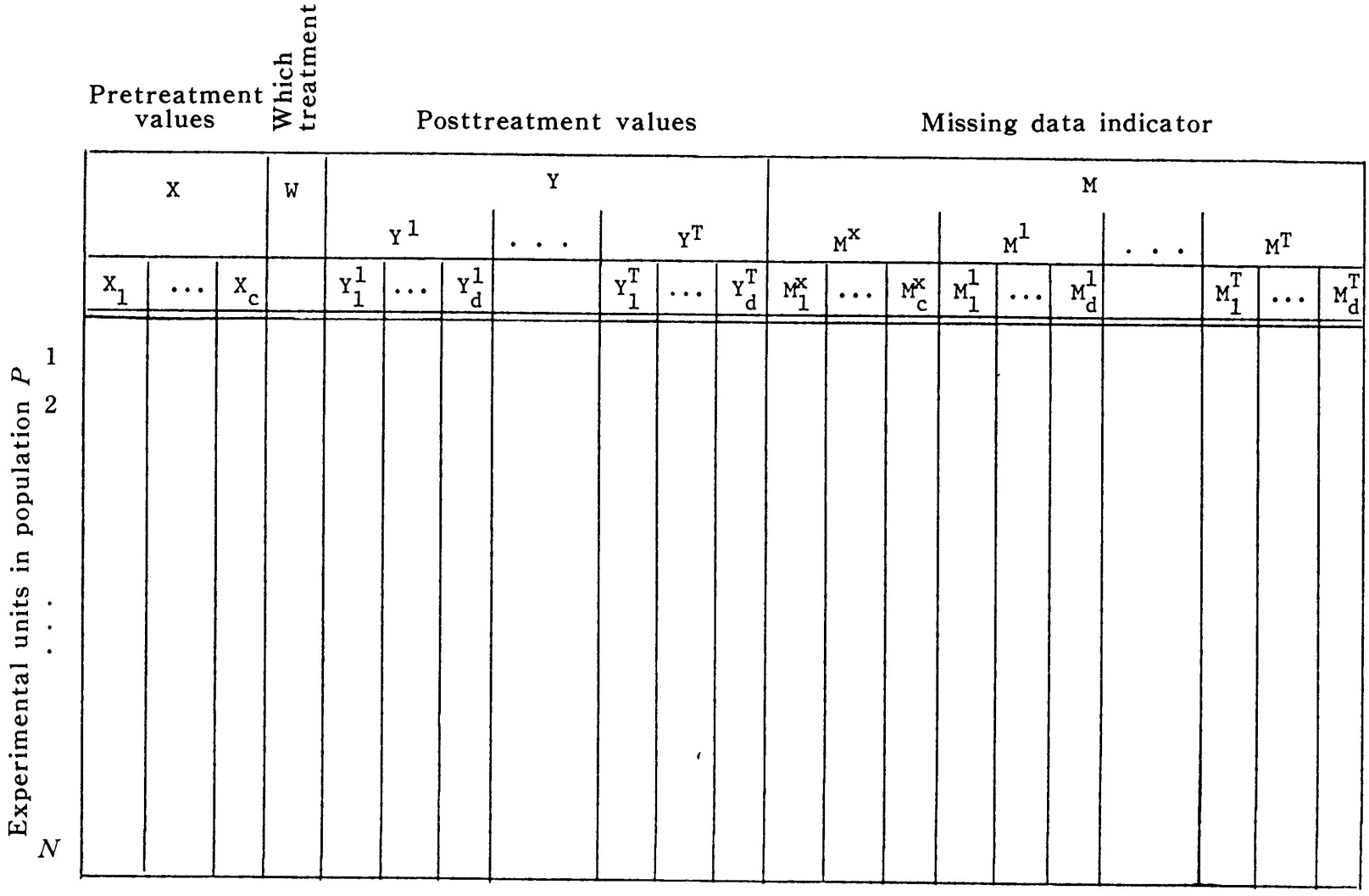


FIG. 1. All values in a study of T treatments.

- A population $\mathscr{P}$ indexed by $i = 1, \ldots, N$.

- A population $\mathscr{P}$ indexed by $i = 1, \ldots, N$.
- Covariates, x_i, for each $i \in \mathscr{P}$.

- A population $\mathscr{P}$ indexed by $i = 1, \ldots, N$.
- Covariates, x_i, for each $i \in \mathscr{P}$.
- Potential outcomes, y_{wi}, fixed for each $i \in \mathscr{P}$ given treatment $w \in \mathscr{W}$, the support of W_i.

- A population $\mathscr{P}$ indexed by $i = 1, \ldots, N$.
- Covariates, x_i, for each $i \in \mathscr{P}$.
- Potential outcomes, y_{wi}, fixed for each $i \in \mathscr{P}$ given treatment $w \in \mathscr{W}$, the support of W_i.
- Random treatment variable, W_i, assigned to $i \in \mathscr{P}$.

- A population $\mathscr{P}$ indexed by $i = 1, \ldots, N$.
- Covariates, x_i, for each $i \in \mathscr{P}$.
- Potential outcomes, y_{wi}, fixed for each $i \in \mathscr{P}$ given treatment $w \in \mathscr{W}$, the support of W_i.
- Random treatment variable, W_i, assigned to $i \in \mathscr{P}$.
- Potential outcomes for i depend only on treatment values for i and no one else — "stable unit treatment value assumption" (SUTVA), a.k.a. "no interference."

- A population $\mathscr{P}$ indexed by $i = 1, \ldots, N$.
- Covariates, x_i, for each $i \in \mathscr{P}$.
- Potential outcomes, y_{wi}, fixed for each $i \in \mathscr{P}$ given treatment $w \in \mathscr{W}$, the support of W_i.
- Random treatment variable, W_i, assigned to $i \in \mathscr{P}$.
- Potential outcomes for i depend only on treatment values for i and no one else — "stable unit treatment value assumption" (SUTVA), a.k.a. "no interference."
- Missing data indicators, $M_{i,j}^{(k)}$ for characteristic j of characteristic type k for all $i \in \mathscr{P}$. Refers to sampling or missing data.

- A population $\mathscr{P}$ indexed by $i = 1, \ldots, N$.
- Covariates, x_i, for each $i \in \mathscr{P}$.
- Potential outcomes, y_{wi}, fixed for each $i \in \mathscr{P}$ given treatment $w \in \mathscr{W}$, the support of W_i.
- Random treatment variable, W_i, assigned to $i \in \mathscr{P}$.
- Potential outcomes for i depend only on treatment values for i and no one else — "stable unit treatment value assumption" (SUTVA), a.k.a. "no interference."
- Missing data indicators, $M_{i,j}^{(k)}$ for characteristic j of characteristic type k for all $i \in \mathscr{P}$. Refers to sampling or missing data.
- *Unit level causal effects* for members of $\mathscr{P}$ are fixed a priori, and they compare y_{wi} to $y_{\tilde{w}i}$ for $w \neq \tilde{w}$ and $i \in \mathscr{P}$.

- A population $\mathscr{P}$ indexed by $i = 1, \ldots, N$.
- Covariates, x_i, for each $i \in \mathscr{P}$.
- Potential outcomes, y_{wi}, fixed for each $i \in \mathscr{P}$ given treatment $w \in \mathscr{W}$, the support of W_i.
- Random treatment variable, W_i, assigned to $i \in \mathscr{P}$.
- Potential outcomes for i depend only on treatment values for i and no one else — "stable unit treatment value assumption" (SUTVA), a.k.a. "no interference."
- Missing data indicators, $M_{i,j}^{(k)}$ for characteristic j of characteristic type k for all $i \in \mathscr{P}$. Refers to sampling or missing data.
- *Unit level causal effects* for members of $\mathscr{P}$ are fixed a priori, and they compare y_{wi} to $y_{\tilde{w}i}$ for $w \neq \tilde{w}$ and $i \in \mathscr{P}$.
- *Population causal effects* for compare aggregates of unit level causal effects for members of $\mathscr{P}$.

Things to note:

- Effects are defined in an "agnostic" or "non-parametric" way.
- Potential outcomes and covariates are fixed for each i. Treatments and response indicators are stochastic.
- Effects are defined by letting only treatments vary, holding units fixed.
- Thus, causal effects are defined *only* for units that can conceivably receive different treatment values.
- This is analogous to Pearl's (2009) definition of causal effects, based on how outcomes change when you intervene (cf. pp. 98-102).
- The test for the above is "manipulation" (Holland, 1986).

Holland (1986) : “For causal inference, it is critical that each unit be potentially exposable to any one of the causes.”

Angrist and Krueger (1999) : “The problem of ambiguous counterfactuals is typically resolved by focusing on hypothetic manipulations in the world as is.”

Recall, a unit level causal effect compares y_{wi} to $y_{\tilde{w}i}$ for $w \neq \tilde{w}$.

"Fundamental problem of causal inference" (Holland, 1986) : For each i potential outcomes for all w exist, but we only observe the potential outcome for the treatment value that i receives.

- "Scientific solution": Use theory to determine when units are interchangeable.
- "Statistical solution": Study features of conditional distributions, such as averages.

Causal identification under the potential outcomes model

Basic statistical setting (adjusting notation to conform to MHE):

- Consider a random draw, i, from $\mathscr{P}$, countable but large.
- Each draw is characterized by
  - a covariate vector, X_i,
  - potential outcomes that under SUTVA are characterized as Y_{di} for all $d \in \mathscr{D}$, as well as
  - treatment assignments, $D_i \in \mathscr{D}$.
- By random sampling, for arbitrary characteristics A_i and B_i,

$$E[A_i] = |\mathscr{P}|^{-1} \sum_{j \in \mathscr{P}} A_j$$

$$Var(A_i) = |\mathscr{P}|^{-1} \sum_{j \in \mathscr{P}} (A_i - E[A_i])^2$$

$$Cov(A_i, B_i) = |\mathscr{P}|^{-1} \sum_{j \in \mathscr{P}} (A_i - E[A_i])(B_i - E[B_i]).$$

- Suppose $\mathscr{D} = \{0, 1\}$. Then under SUTVA, potential outcomes for an arbitrary draw from $\mathscr{P}$ are Y_{1i} and Y_{0i}.
- A unit level treatment effect for an arbitrary draw from $\mathscr{P}$ is,

$$\rho_i = Y_{1i} - Y_{0i},$$

for which,

$$E[\rho_i] = E[Y_{1i} - Y_{0i}] = \rho, \tag{1}$$

the average treatment effect (ATE).

- For an arbitrary draw from $\mathscr{P}$, we observe,

$$Y_i = D_i Y_{1i} + (1 - D_i) Y_{i0}.$$

Consider a simple difference in outcomes between arbitrary units for which $D_i = 1$ versus an arbitrary unit for which $D_i = 0$.

What is the expected value of this difference?

$$
\begin{aligned}
E[Y_i|D_i = 1] - E[Y_i|D_i = 0] &= E[Y_{1i}|D_i = 1] - E[Y_{0i}|D_i = 0] \\
&= \underbrace{E[Y_{1i}|D_i = 1] - E[Y_{0i}|D_i = 1]}_{\text{Average treatment effect on the treated (ATT)}} \\
&\quad + \underbrace{E[Y_{0i}|D_i = 1] - E[Y_{0i}|D_i = 0]}_{\text{Selection bias}}. \qquad (2)
\end{aligned}
$$

Identifying assumption 1 (random assignment):

$$D_i \perp\!\!\!\perp (Y_{1i}, Y_{0i}) \text{ and } 0 < Pr[D_i = 1] < 1 \tag{3}$$

Recall $A \perp\!\!\!\perp B$ implies $E[A|B] = E[A]$. And so, under (3),

$$\underbrace{E[Y_{0i}|D_i = 1] - E[Y_{0i}|D_i = 0]}_{\text{Selection bias}} = 0$$

and

$$\underbrace{E[Y_{1i}|D_i = 1] - E[Y_{0i}|D_i = 1]}_{\text{ATT}} = E[Y_{1i} - Y_{0i}],$$

in which case the simple difference, (2), equals ρ.

Identifying assumption 2 (conditionally independent/unconfounded/strongly ignorable assignment):

$$D_i \perp\!\!\!\perp (Y_{1i}, Y_{0i}) | X_i \text{ and } 0 < Pr[D_i = 1 | X_i = x] < 1 \text{ for all } x \in \mathscr{X} \quad (4)$$

The conditional average treatment effect (CATE) is given by,

$$\rho(x) = E[Y_{1i} - Y_{0i} | X_i = x]$$

By the same logic as before, assumption (4) implies,

$$E[Y_i | D_i = 1, X_i = x] - E[Y_i | D_i = 0, X_i = x] = \rho(x).$$

Marginalization over $\mathscr{X}$, the support of X_i, yields,

$$\int_{\mathscr{X}} \rho(x) dF(x) = \rho.$$

- Randomization and conditionally independent assignment are examples of identifying assumptions.
- Either is sufficient for identification of the average treatment effect.
- Through the semester we will be looking at these and other identifying assumptions and then defining statistical methods that make use of them.

Relating this to to the regression framework

$$Y_i = \beta_0 + D_i\beta_1 + \varepsilon_i$$

Let D_i have support $\{0,1\}$.

Define Y_i as before,

$$Y_i = D_i Y_{1i} + (1 - D_i) Y_{0i}$$

Then,

$$Y_i = D_i Y_{1i} + (1 - D_i) Y_{0i}$$

(add and subtract $E[Y_{0i}]$ and $E[Y_{1i}]$, rearrange)

$$= \underbrace{E[Y_{0i}]}_{\beta_0} + \underbrace{D_i E[Y_{1i} - Y_{0i}]}_{+ \quad D_i\beta_1} + \underbrace{D_i(Y_{1i} - E[Y_{1i}]) + (1 - D_i)(Y_{0i} - E[Y_{0i}])}_{+ \quad \varepsilon_i}$$

or

$$= \underbrace{E[Y_{0i}]}_{\beta_0} + \underbrace{D_i E[Y_{1i} - Y_{0i}]}_{+ \quad D_i\beta_1} + \underbrace{(Y_{0i} - E[Y_{0i}]) + D_i[(Y_{1i} - Y_{0i}) - (E[Y_{1i}] - E[Y_{0i}])]}_{+ \quad \varepsilon_i}$$

- ε_i can be interpreted as (i) heterogeneity in potential outcomes, or (ii) heterogeneity in baseline potential outcomes plus effect heterogeneity.
- D_i is random (different than classical regression).
- Effect heterogeneity implies heteroskedasticity assumption needed on ε_i, because error variance differs over D_i.
- Equivalency means that we can retain much regression theory and intuitions while being "agnostic" about the nature of causal effects (e.g. we don't have to assume homogenous effects).
- Generalizations to multivalued treatments are straightforward (either dose-response functions or a bunch of binary contrasts).