Week 10: Causality with Measured Confounding

Brandon Stewart[1]

Princeton

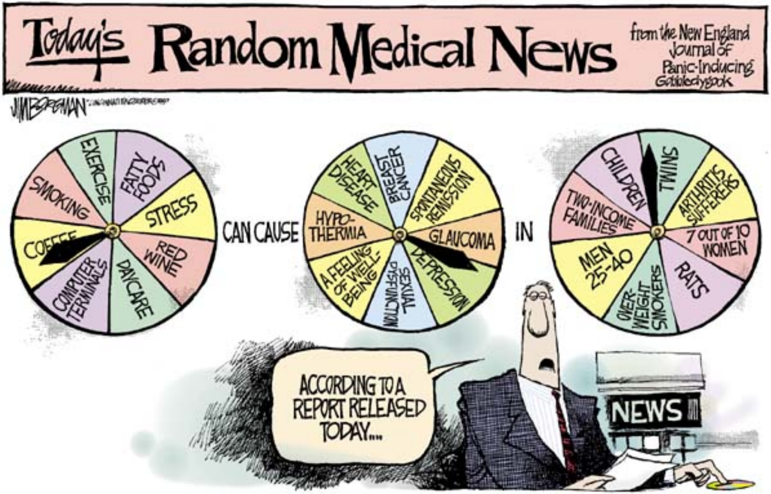
November 26 and 28, 2018

[1] These slides are heavily influenced by Matt Blackwell, Jens Hainmueller, Erin Hartman, Kosuke Imai and Gary King.

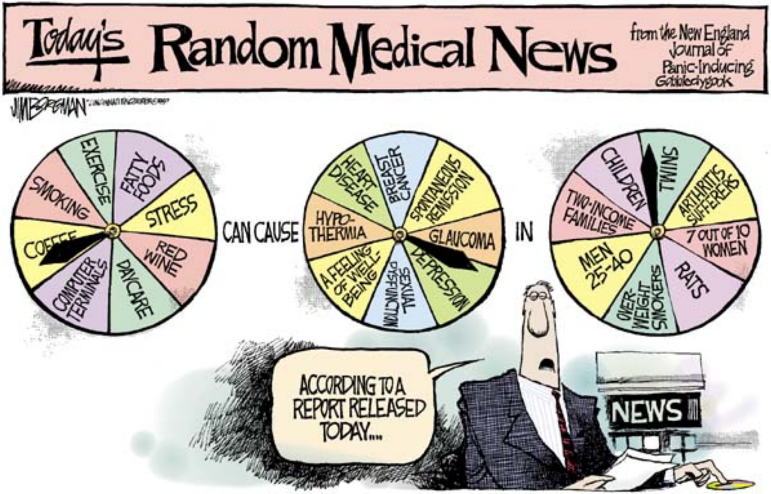
Where We've Been and Where We're Going...

- Last Week
  - intro to causal inference
- This Week
  - Monday:
    - experimental Ideal
    - identification with measured confounding
  - Wednesday:
    - regression estimation
- Next Week
  - identification with unmeasured confounding
  - instrumental variables
- Long Run
  - probability → inference → regression → causal inference

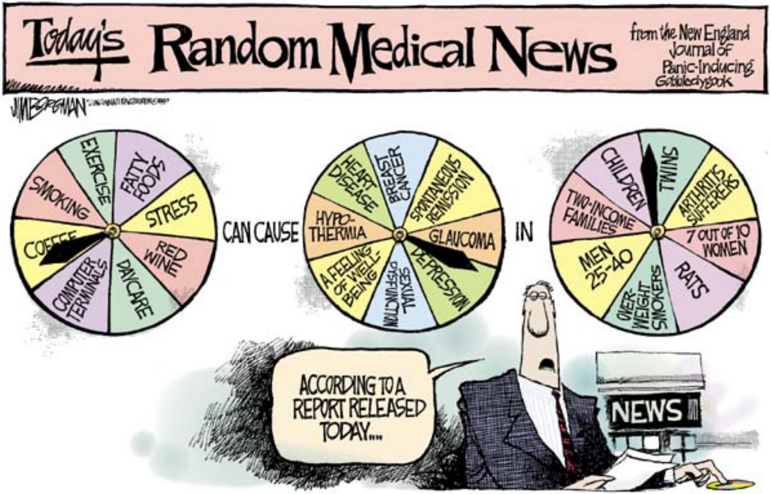
Questions?



Lancet 2001: negative correlation between coronary heart disease mortality and level of vitamin C in bloodstream (controlling for age, gender, blood pressure, diabetes, and smoking)



Lancet 2002: no effect of vitamin C on mortality in controlled placebo trial (controlling for nothing)

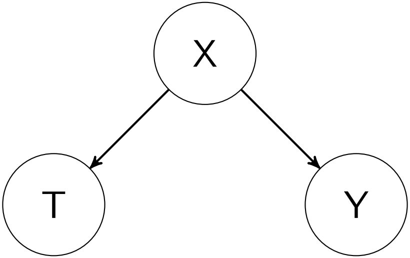


Lancet 2003: comparing among individuals with the same age, gender, blood pressure, diabetes, and smoking, those with higher vitamin C levels have lower levels of obesity, lower levels of alcohol consumption, are less likely to grow up in working class, etc.

Why So Much Variation?

Why So Much Variation?

Confounders



Observational Studies and Experimental Ideal

- Randomization forms gold standard for causal inference, because it balances observed and unobserved confounders

Observational Studies and Experimental Ideal

- Randomization forms gold standard for causal inference, because it balances observed and unobserved confounders
- Cannot always randomize so we do observational studies, where we adjust for the observed covariates and hope that unobservables are balanced

Observational Studies and Experimental Ideal

- Randomization forms gold standard for causal inference, because it balances observed and unobserved confounders
- Cannot always randomize so we do observational studies, where we adjust for the observed covariates and hope that unobservables are balanced
- Better than hoping: design observational study to approximate an experiment

Observational Studies and Experimental Ideal

- Randomization forms gold standard for causal inference, because it balances observed and unobserved confounders
- Cannot always randomize so we do observational studies, where we adjust for the observed covariates and hope that unobservables are balanced
- Better than hoping: design observational study to approximate an experiment
  - "The planner of an observational study should always ask himself: How would the study be conducted if it were possible to do it by controlled experimentation" (Cochran 1965)

Angrist and Pishke's Frequently Asked Questions

Angrist and Pishke's Frequently Asked Questions

- What is the causal relationship of interest?

Angrist and Pishke's Frequently Asked Questions

- What is the causal relationship of interest?
- What is the experiment that could ideally be used to capture the causal effect of interest?

Angrist and Pishke's Frequently Asked Questions

- What is the causal relationship of interest?
- What is the experiment that could ideally be used to capture the causal effect of interest?
- What is your identification strategy?

Angrist and Pishke's Frequently Asked Questions

- What is the causal relationship of interest?
- What is the experiment that could ideally be used to capture the causal effect of interest?
- What is your identification strategy?
- What is your mode of statistical inference?

Experiment review

- An experiment is a study where assignment to treatment is controlled by the researcher.

Experiment review

- An experiment is a study where assignment to treatment is controlled by the researcher.
  - $p_i = \mathbb{P}[D_i = 1]$ be the probability of treatment assignment probability.

Experiment review

- An experiment is a study where assignment to treatment is controlled by the researcher.
  - $p_i = \mathbb{P}[D_i = 1]$ be the probability of treatment assignment probability.
  - p_i is controlled and known by researcher in an experiment.

Experiment review

- An experiment is a study where assignment to treatment is controlled by the researcher.
  - $p_i = \mathbb{P}[D_i = 1]$ be the probability of treatment assignment probability.
  - p_i is controlled and known by researcher in an experiment.
- A randomized experiment is an experiment with the following properties:

Experiment review

- An experiment is a study where assignment to treatment is controlled by the researcher.
  - $p_i = \mathbb{P}[D_i = 1]$ be the probability of treatment assignment probability.
  - p_i is controlled and known by researcher in an experiment.
- A randomized experiment is an experiment with the following properties:

1. Positivity: assignment is probabilistic: $0 < p_i < 1$

Experiment review

- An experiment is a study where assignment to treatment is controlled by the researcher.
  - $p_i = \mathbb{P}[D_i = 1]$ be the probability of treatment assignment probability.
  - p_i is controlled and known by researcher in an experiment.
- A randomized experiment is an experiment with the following properties:

1. Positivity: assignment is probabilistic: $0 < p_i < 1$
   - No deterministic assignment.

Experiment review

- An experiment is a study where assignment to treatment is controlled by the researcher.
  - $p_i = \mathbb{P}[D_i = 1]$ be the probability of treatment assignment probability.
  - p_i is controlled and known by researcher in an experiment.
- A randomized experiment is an experiment with the following properties:

1. Positivity: assignment is probabilistic: $0 < p_i < 1$
   - No deterministic assignment.
2. Unconfoundedness: $\mathbb{P}[D_i = 1|\mathbf{Y}(1), \mathbf{Y}(0)] = \mathbb{P}[D_i = 1]$

Experiment review

- An experiment is a study where assignment to treatment is controlled by the researcher.
  - $p_i = \mathbb{P}[D_i = 1]$ be the probability of treatment assignment probability.
  - p_i is controlled and known by researcher in an experiment.
- A randomized experiment is an experiment with the following properties:

1. Positivity: assignment is probabilistic: $0 < p_i < 1$
   - No deterministic assignment.
2. Unconfoundedness: $\mathbb{P}[D_i = 1|\mathbf{Y}(1), \mathbf{Y}(0)] = \mathbb{P}[D_i = 1]$
   - Treatment assignment does not depend on any potential outcomes.

Experiment review

- An experiment is a study where assignment to treatment is controlled by the researcher.
  - $p_i = \mathbb{P}[D_i = 1]$ be the probability of treatment assignment probability.
  - p_i is controlled and known by researcher in an experiment.
- A randomized experiment is an experiment with the following properties:

1. Positivity: assignment is probabilistic: $0 < p_i < 1$
   - No deterministic assignment.
2. Unconfoundedness: $\mathbb{P}[D_i = 1|\mathbf{Y}(1), \mathbf{Y}(0)] = \mathbb{P}[D_i = 1]$
   - Treatment assignment does not depend on any potential outcomes.
   - Sometimes written as $D_i \perp\!\!\!\perp (\mathbf{Y}(1), \mathbf{Y}(0))$

Why do Experiments Help?

Remember selection bias?

$$
\begin{aligned}
&E[Y_i|D_i = 1] - E[Y_i|D_i = 0] \\
&\quad = E[Y_i(1)|D_i = 1] - E[Y_i(0)|D_i = 0] \\
&\quad = E[Y_i(1)|D_i = 1] - E[Y_i(0)|D_i = 1] + E[Y_i(0)|D_i = 1] - E[Y_i(0)|D_i = 0] \\
&\quad = \underbrace{E[Y_i(1) - Y_i(0)|D_i = 1]}_{\text{Average Treatment Effect on Treated}} + \underbrace{E[Y_i(0)|D_i = 1] - E[Y_i(0)|D_i = 0]}_{\text{selection bias}}
\end{aligned}
$$

Why do Experiments Help?

Remember selection bias?

$$
\begin{aligned}
&E[Y_i|D_i = 1] - E[Y_i|D_i = 0] \\
&\quad = E[Y_i(1)|D_i = 1] - E[Y_i(0)|D_i = 0] \\
&\quad = E[Y_i(1)|D_i = 1] - E[Y_i(0)|D_i = 1] + E[Y_i(0)|D_i = 1] - E[Y_i(0)|D_i = 0] \\
&\quad = \underbrace{E[Y_i(1) - Y_i(0)|D_i = 1]}_{\text{Average Treatment Effect on Treated}} + \underbrace{E[Y_i(0)|D_i = 1] - E[Y_i(0)|D_i = 0]}_{\text{selection bias}}
\end{aligned}
$$

In an experiment we know that treatment is randomly assigned. Thus we can do the following:

Why do Experiments Help?

Remember selection bias?

$$
\begin{aligned}
&E[Y_i|D_i=1]-E[Y_i|D_i=0]\\
&\quad = E[Y_i(1)|D_i=1]-E[Y_i(0)|D_i=0]\\
&\quad = E[Y_i(1)|D_i=1]-E[Y_i(0)|D_i=1]+E[Y_i(0)|D_i=1]-E[Y_i(0)|D_i=0]\\
&\quad = \underbrace{E[Y_i(1)-Y_i(0)|D_i=1]}_{\text{Average Treatment Effect on Treated}} + \underbrace{E[Y_i(0)|D_i=1]-E[Y_i(0)|D_i=0]}_{\text{selection bias}}
\end{aligned}
$$

In an experiment we know that treatment is randomly assigned. Thus we can do the following:

$$
\begin{aligned}
E[Y_i(1)|D_i=1]-E[Y_i(0)|D_i=0] &= E[Y_i(1)|D_i=1]-E[Y_i(0)|D_i=1]\\
&= E[Y_i(1)]-E[Y_i(0)]
\end{aligned}
$$

Why do Experiments Help?

Remember selection bias?

$$\begin{aligned} &E[Y_i|D_i=1]-E[Y_i|D_i=0] \\ &\quad = E[Y_i(1)|D_i=1]-E[Y_i(0)|D_i=0] \\ &\quad = E[Y_i(1)|D_i=1]-E[Y_i(0)|D_i=1]+E[Y_i(0)|D_i=1]-E[Y_i(0)|D_i=0] \\ &\quad = \underbrace{E[Y_i(1)-Y_i(0)|D_i=1]}_{\text{Average Treatment Effect on Treated}} + \underbrace{E[Y_i(0)|D_i=1]-E[Y_i(0)|D_i=0]}_{\text{selection bias}} \end{aligned}$$

In an experiment we know that treatment is randomly assigned. Thus we can do the following:

$$\begin{aligned} E[Y_i(1)|D_i=1]-E[Y_i(0)|D_i=0] &= E[Y_i(1)|D_i=1]-E[Y_i(0)|D_i=1] \\ &= E[Y_i(1)]-E[Y_i(0)] \end{aligned}$$

When all goes well, an experiment eliminates selection bias.

Observational studies

- Many different sets of identification assumptions that we'll cover.

Observational studies

- Many different sets of identification assumptions that we'll cover.
- To start, focus on studies that are similar to experiments, just without a known and controlled treatment assignment.

Observational studies

- Many different sets of identification assumptions that we'll cover.
- To start, focus on studies that are similar to experiments, just without a known and controlled treatment assignment.
  - No guarantee that the treatment and control groups are comparable.

Observational studies

- Many different sets of identification assumptions that we'll cover.
- To start, focus on studies that are similar to experiments, just without a known and controlled treatment assignment.
  - No guarantee that the treatment and control groups are comparable.

1. Positivity (Common Support): assignment is probabilistic: $0 < \mathbb{P}[D_i = 1|\mathbf{X}, \mathbf{Y}(1), \mathbf{Y}(0)] < 1$

Observational studies

- Many different sets of identification assumptions that we'll cover.
- To start, focus on studies that are similar to experiments, just without a known and controlled treatment assignment.
  - No guarantee that the treatment and control groups are comparable.

1. Positivity (Common Support): assignment is probabilistic: $0 < \mathbb{P}[D_i = 1|\mathbf{X}, \mathbf{Y}(1), \mathbf{Y}(0)] < 1$
2. No unmeasured confounding: $\mathbb{P}[D_i = 1|\mathbf{X}, \mathbf{Y}(1), \mathbf{Y}(0)] = \mathbb{P}[D_i = 1|\mathbf{X}]$

Observational studies

- Many different sets of identification assumptions that we'll cover.
- To start, focus on studies that are similar to experiments, just without a known and controlled treatment assignment.
  - No guarantee that the treatment and control groups are comparable.

1. Positivity (Common Support): assignment is probabilistic: $0 < \mathbb{P}[D_i = 1|\mathbf{X}, \mathbf{Y}(1), \mathbf{Y}(0)] < 1$
2. No unmeasured confounding: $\mathbb{P}[D_i = 1|\mathbf{X}, \mathbf{Y}(1), \mathbf{Y}(0)] = \mathbb{P}[D_i = 1|\mathbf{X}]$
   - For some observed $\mathbf{X}$

Observational studies

- Many different sets of identification assumptions that we'll cover.
- To start, focus on studies that are similar to experiments, just without a known and controlled treatment assignment.
  - No guarantee that the treatment and control groups are comparable.

1. Positivity (Common Support): assignment is probabilistic: $0 < \mathbb{P}[D_i = 1|\mathbf{X}, \mathbf{Y}(1), \mathbf{Y}(0)] < 1$
2. No unmeasured confounding: $\mathbb{P}[D_i = 1|\mathbf{X}, \mathbf{Y}(1), \mathbf{Y}(0)] = \mathbb{P}[D_i = 1|\mathbf{X}]$
   - For some observed $\mathbf{X}$
   - Also called: unconfoundedness

Observational studies

- Many different sets of identification assumptions that we'll cover.
- To start, focus on studies that are similar to experiments, just without a known and controlled treatment assignment.
  - No guarantee that the treatment and control groups are comparable.

1. Positivity (Common Support): assignment is probabilistic: $0 < \mathbb{P}[D_i = 1 | \mathbf{X}, \mathbf{Y}(1), \mathbf{Y}(0)] < 1$
2. No unmeasured confounding: $\mathbb{P}[D_i = 1 | \mathbf{X}, \mathbf{Y}(1), \mathbf{Y}(0)] = \mathbb{P}[D_i = 1 | \mathbf{X}]$
   - For some observed $\mathbf{X}$
   - Also called: unconfoundedness, ignorability

Observational studies

- Many different sets of identification assumptions that we'll cover.
- To start, focus on studies that are similar to experiments, just without a known and controlled treatment assignment.
  - No guarantee that the treatment and control groups are comparable.

1. Positivity (Common Support): assignment is probabilistic: $0 < \mathbb{P}[D_i = 1|\mathbf{X}, \mathbf{Y}(1), \mathbf{Y}(0)] < 1$
2. No unmeasured confounding: $\mathbb{P}[D_i = 1|\mathbf{X}, \mathbf{Y}(1), \mathbf{Y}(0)] = \mathbb{P}[D_i = 1|\mathbf{X}]$
   - For some observed $\mathbf{X}$
   - Also called: unconfoundedness, ignorability, selection on observables

Observational studies

- Many different sets of identification assumptions that we'll cover.
- To start, focus on studies that are similar to experiments, just without a known and controlled treatment assignment.
    - No guarantee that the treatment and control groups are comparable.

1. Positivity (Common Support): assignment is probabilistic: $0 < \mathbb{P}[D_i = 1|\mathbf{X}, \mathbf{Y}(1), \mathbf{Y}(0)] < 1$
2. No unmeasured confounding: $\mathbb{P}[D_i = 1|\mathbf{X}, \mathbf{Y}(1), \mathbf{Y}(0)] = \mathbb{P}[D_i = 1|\mathbf{X}]$
    - For some observed $\mathbf{X}$
    - Also called: unconfoundedness, ignorability, selection on observables, no omitted variables

Observational studies

- Many different sets of identification assumptions that we'll cover.
- To start, focus on studies that are similar to experiments, just without a known and controlled treatment assignment.
  - No guarantee that the treatment and control groups are comparable.

1. Positivity (Common Support): assignment is probabilistic: $0 < \mathbb{P}[D_i = 1|\mathbf{X}, \mathbf{Y}(1), \mathbf{Y}(0)] < 1$
2. No unmeasured confounding: $\mathbb{P}[D_i = 1|\mathbf{X}, \mathbf{Y}(1), \mathbf{Y}(0)] = \mathbb{P}[D_i = 1|\mathbf{X}]$
   - For some observed $\mathbf{X}$
   - Also called: unconfoundedness, ignorability, selection on observables, no omitted variables, exogenous

Observational studies

- Many different sets of identification assumptions that we'll cover.
- To start, focus on studies that are similar to experiments, just without a known and controlled treatment assignment.
  - No guarantee that the treatment and control groups are comparable.

1. Positivity (Common Support): assignment is probabilistic: $0 < \mathbb{P}[D_i = 1|\mathbf{X}, \mathbf{Y}(1), \mathbf{Y}(0)] < 1$
2. No unmeasured confounding: $\mathbb{P}[D_i = 1|\mathbf{X}, \mathbf{Y}(1), \mathbf{Y}(0)] = \mathbb{P}[D_i = 1|\mathbf{X}]$
   - For some observed $\mathbf{X}$
   - Also called: unconfoundedness, ignorability, selection on observables, no omitted variables, exogenous, conditionally exchangeable, etc.

Designing observational studies

- Rubin (2008) argues that we should still "design" our observational studies:

Designing observational studies

- Rubin (2008) argues that we should still "design" our observational studies:
  - Pick the ideal experiment to this observational study.

Designing observational studies

- Rubin (2008) argues that we should still “design” our observational studies:
  - Pick the ideal experiment to this observational study.
  - Hide the outcome data.

Designing observational studies

- Rubin (2008) argues that we should still “design” our observational studies:
  - Pick the ideal experiment to this observational study.
  - Hide the outcome data.
  - Try to estimate the randomization procedure.

Designing observational studies

- Rubin (2008) argues that we should still "design" our observational studies:
  - Pick the ideal experiment to this observational study.
  - Hide the outcome data.
  - Try to estimate the randomization procedure.
  - Analyze this as an experiment with this estimated procedure.

Designing observational studies

- Rubin (2008) argues that we should still “design” our observational studies:
  - Pick the ideal experiment to this observational study.
  - Hide the outcome data.
  - Try to estimate the randomization procedure.
  - Analyze this as an experiment with this estimated procedure.
- Tries to minimize “snooping” by picking the best modeling strategy before seeing the outcome.

Discrete covariates

- Suppose that we knew that D_i was unconfounded within levels of a binary X_i.

Discrete covariates

- Suppose that we knew that D_i was unconfounded within levels of a binary X_i.
- Then we could always estimate the causal effect using iterated expectations as in a stratified randomized experiment:

Discrete covariates

- Suppose that we knew that D_i was unconfounded within levels of a binary X_i.
- Then we could always estimate the causal effect using iterated expectations as in a stratified randomized experiment:

Discrete covariates

- Suppose that we knew that D_i was unconfounded within levels of a binary X_i.
- Then we could always estimate the causal effect using iterated expectations as in a stratified randomized experiment:

$$\mathbb{E}_X\left\{\mathbb{E}[Y_i|D_i = 1, X_i] - \mathbb{E}[Y_i|D_i = 0, X_i]\right\}$$

Discrete covariates

- Suppose that we knew that D_i was unconfounded within levels of a binary X_i.
- Then we could always estimate the causal effect using iterated expectations as in a stratified randomized experiment:

$$\begin{aligned}
&\mathbb{E}_X\Big\{\mathbb{E}[Y_i|D_i=1,X_i]-\mathbb{E}[Y_i|D_i=0,X_i]\Big\}\\
&=\underbrace{\Big(\mathbb{E}[Y_i|D_i=1,X_i=1]-\mathbb{E}[Y_i|D_i=0,X_i=1]\Big)}_{\text{diff-in-means for } X_i=1}\underbrace{\mathbb{P}[X_i=1]}_{\text{share of } X_i=1}\\
&\quad+\underbrace{\Big(\mathbb{E}[Y_i|D_i=1,X_i=0]-\mathbb{E}[Y_i|D_i=0,X_i=0]\Big)}_{\text{diff-in-means for } X_i=0}\underbrace{\mathbb{P}[X_i=0]}_{\text{share of } X_i=0}
\end{aligned}$$

Discrete covariates

- Suppose that we knew that D_i was unconfounded within levels of a binary X_i.
- Then we could always estimate the causal effect using iterated expectations as in a stratified randomized experiment:

$$\mathbb{E}_X\Big\{\mathbb{E}[Y_i|D_i=1,X_i]-\mathbb{E}[Y_i|D_i=0,X_i]\Big\}$$
$$=\underbrace{\Big(\mathbb{E}[Y_i|D_i=1,X_i=1]-\mathbb{E}[Y_i|D_i=0,X_i=1]\Big)}_{\text{diff-in-means for } X_i=1}\underbrace{\mathbb{P}[X_i=1]}_{\text{share of } X_i=1}$$
$$+\underbrace{\Big(\mathbb{E}[Y_i|D_i=1,X_i=0]-\mathbb{E}[Y_i|D_i=0,X_i=0]\Big)}_{\text{diff-in-means for } X_i=0}\underbrace{\mathbb{P}[X_i=0]}_{\text{share of } X_i=0}$$

- Never used our knowledge of the randomization for this quantity.

Stratification Example: Smoking and Mortality (Cochran, 1968)

TABLE 1

DEATH RATES PER 1,000 PERSON-YEARS

Smoking group	Canada	U.K.	U.S.
Non-smokers	20.2	11.3	13.5
Cigarettes	20.5	14.1	13.5
Cigars/pipes	35.5	20.7	17.4

Stratification Example: Smoking and Mortality (Cochran, 1968)

TABLE 2

MEAN AGES, YEARS

Smoking group	Canada	U.K.	U.S.
Non-smokers	54.9	49.1	57.0
Cigarettes	50.5	49.8	53.2
Cigars/pipes	65.9	55.7	59.7

Stratification

To control for differences in age, we would like to compare different smoking-habit groups with the same age distribution

Stratification

To control for differences in age, we would like to compare different smoking-habit groups with the same age distribution

One possibility is to use stratification:

Stratification

To control for differences in age, we would like to compare different smoking-habit groups with the same age distribution

One possibility is to use stratification:

- for each country, divide each group into different age subgroups

Stratification

To control for differences in age, we would like to compare different smoking-habit groups with the same age distribution

One possibility is to use stratification:

- for each country, divide each group into different age subgroups
- calculate death rates within age subgroups

Stratification

To control for differences in age, we would like to compare different smoking-habit groups with the same age distribution

One possibility is to use stratification:

- for each country, divide each group into different age subgroups
- calculate death rates within age subgroups
- average within age subgroup death rates using fixed weights (e.g. number of cigarette smokers)

Stratification: Example

	Death Rates Pipe Smokers	# Pipe-Smokers	# Non-Smokers
Age 20 - 50	15	11	29
Age 50 - 70	35	13	9
Age + 70	50	16	2
Total		40	40

What is the average death rate for Pipe Smokers?

Stratification: Example

	Death Rates Pipe Smokers	# Pipe-Smokers	# Non-Smokers
Age 20 - 50	15	11	29
Age 50 - 70	35	13	9
Age + 70	50	16	2
Total		40	40

What is the average death rate for Pipe Smokers?
$15 \cdot (11/40) + 35 \cdot (13/40) + 50 \cdot (16/40) = 35.5$

Stratification: Example

	Death Rates Pipe Smokers	# Pipe-Smokers	# Non-Smokers
Age 20 - 50	15	11	29
Age 50 - 70	35	13	9
Age + 70	50	16	2
Total		40	40

What is the average death rate for Pipe Smokers if they had same age distribution as Non-Smokers?

Stratification: Example

	Death Rates Pipe Smokers	# Pipe-Smokers	# Non-Smokers
Age 20 - 50	15	11	29
Age 50 - 70	35	13	9
Age + 70	50	16	2
Total		40	40

What is the average death rate for Pipe Smokers if they had same age distribution as Non-Smokers?

$15 \cdot (29/40) + 35 \cdot (9/40) + 50 \cdot (2/40) = 21.2$

Smoking and Mortality (Cochran, 1968)

TABLE 3

ADJUSTED DEATH RATES USING 3 AGE GROUPS

Smoking group	Canada	U.K.	U.S.
Non-smokers	20.2	11.3	13.5
Cigarettes	28.3	12.8	17.7
Cigars/pipes	21.2	12.0	14.2

Continuous covariates

- So, great, we can stratify. Why not do this all the time?

Continuous covariates

- So, great, we can stratify. Why not do this all the time?
- What if X_i = income for unit i?

Continuous covariates

- So, great, we can stratify. Why not do this all the time?
- What if X_i = income for unit i?
  - Each unit has its own value of X_i: \$54,134, \$123,043, \$23,842.

Continuous covariates

- So, great, we can stratify. Why not do this all the time?
- What if X_i = income for unit i?
  - Each unit has its own value of X_i: \$54,134, \$123,043, \$23,842.
  - If $X_i = 54134$ is unique, will only observe 1 of these:

$$\mathbb{E}[Y_i | D_i = 1, X_i = 54134] - \mathbb{E}[Y_i | D_i = 0, X_i = 54134]$$

Continuous covariates

- So, great, we can stratify. Why not do this all the time?
- What if X_i = income for unit i?
  - Each unit has its own value of X_i: \$54,134, \$123,043, \$23,842.
  - If $X_i = 54134$ is unique, will only observe 1 of these:

$$\mathbb{E}[Y_i|D_i = 1, X_i = 54134] - \mathbb{E}[Y_i|D_i = 0, X_i = 54134]$$

  - $\rightsquigarrow$ cannot stratify to each unique value of X_i:

Continuous covariates

- So, great, we can stratify. Why not do this all the time?
- What if X_i = income for unit i?
  - Each unit has its own value of X_i: \$54,134, \$123,043, \$23,842.
  - If $X_i = 54134$ is unique, will only observe 1 of these:

$$\mathbb{E}[Y_i|D_i = 1, X_i = 54134] - \mathbb{E}[Y_i|D_i = 0, X_i = 54134]$$

  - $\rightsquigarrow$ cannot stratify to each unique value of X_i:
- Practically, this is massively important: almost always have data with unique values.

Continuous covariates

- So, great, we can stratify. Why not do this all the time?
- What if X_i = income for unit i?
  - Each unit has its own value of X_i: \$54,134, \$123,043, \$23,842.
  - If $X_i = 54134$ is unique, will only observe 1 of these:

$$\mathbb{E}[Y_i|D_i = 1, X_i = 54134] - \mathbb{E}[Y_i|D_i = 0, X_i = 54134]$$

  - $\rightsquigarrow$ cannot stratify to each unique value of X_i:
- Practically, this is massively important: almost always have data with unique values.

One option is to discretize as we discussed with age, we will discuss more later this week!

Identification Under Selection on Observables

Identification Assumption

1. $(Y_1, Y_0) \perp\!\!\!\perp D | X$ *(selection on observables)*
2. $0 < \Pr(D = 1|X) < 1$ *with probability one (common support)*

Identification Result

Given selection on observables we have

$$\begin{aligned} \mathbb{E}[Y_1 - Y_0|X] &= \mathbb{E}[Y_1 - Y_0|X, D = 1] \\ &= \mathbb{E}[Y|X, D = 1] - \mathbb{E}[Y|X, D = 0] \end{aligned}$$

Therefore, under the common support condition:

$$\begin{aligned} \tau_{ATE} &= \mathbb{E}[Y_1 - Y_0] = \int \mathbb{E}[Y_1 - Y_0|X]\, dP(X) \\ &= \int \big(\mathbb{E}[Y|X, D = 1] - \mathbb{E}[Y|X, D = 0]\big)\, dP(X) \end{aligned}$$

Identification Under Selection on Observables

Identification Assumption

1. $(Y_1, Y_0) \perp\!\!\!\perp D|X$ *(selection on observables)*
2. $0 < \Pr(D = 1|X) < 1$ *with probability one (common support)*

Identification Result

Similarly,

$$\begin{aligned} \tau_{ATT} &= \mathbb{E}[Y_1 - Y_0|D = 1] \\ &= \int \left(\mathbb{E}[Y|X, D = 1] - \mathbb{E}[Y|X, D = 0]\right) dP(X|D = 1) \end{aligned}$$

To identify τ_{ATT} the selection on observables and common support conditions can be relaxed to:

- $Y_0 \perp\!\!\!\perp D|X$ *(SOO for Controls)*
- $\Pr(D = 1|X) < 1$ *(Weak Overlap)*

Identification Under Selection on Observables

unit	Potential Outcome under Treatment	Potential Outcome under Control		
i	Y_{1i}	Y_{0i}	D_i	X_i
1	$\mathbb{E}[Y_1 \mid X=0, D=1]$	$\mathbb{E}[Y_0 \mid X=0, D=1]$	1	0
2			1	0
3	$\mathbb{E}[Y_1 \mid X=0, D=0]$	$\mathbb{E}[Y_0 \mid X=0, D=0]$	0	0
4			0	0
5	$\mathbb{E}[Y_1 \mid X=1, D=1]$	$\mathbb{E}[Y_0 \mid X=1, D=1]$	1	1
6			1	1
7	$\mathbb{E}[Y_1 \mid X=1, D=0]$	$\mathbb{E}[Y_0 \mid X=1, D=0]$	0	1
8			0	1

Identification Under Selection on Observables

unit	Potential Outcome under Treatment	Potential Outcome under Control		
i	Y_{1i}	Y_{0i}	D_i	X_i
1	$\mathbb{E}[Y_1 \mid X=0, D=1]$	$\mathbb{E}[Y_0 \mid X=0, D=1]=$	1	0
2		$\mathbb{E}[Y_0 \mid X=0, D=0]$	1	0
3	$\mathbb{E}[Y_1 \mid X=0, D=0]$	$\mathbb{E}[Y_0 \mid X=0, D=0]$	0	0
4			0	0
5	$\mathbb{E}[Y_1 \mid X=1, D=1]$	$\mathbb{E}[Y_0 \mid X=1, D=1]=$	1	1
6		$\mathbb{E}[Y_0 \mid X=1, D=0]$	1	1
7	$\mathbb{E}[Y_1 \mid X=1, D=0]$	$\mathbb{E}[Y_0 \mid X=1, D=0]$	0	1
8			0	1

$(Y_1, Y_0) \perp\!\!\!\perp D \mid X$ implies that we conditioned on all confounders. The treatment is randomly assigned within each stratum of X:

$$\mathbb{E}[Y_0 \mid X=0, D=1] = \mathbb{E}[Y_0 \mid X=0, D=0] \text{ and}$$
$$\mathbb{E}[Y_0 \mid X=1, D=1] = \mathbb{E}[Y_0 \mid X=1, D=0]$$

Identification Under Selection on Observables

unit	Potential Outcome under Treatment	Potential Outcome under Control		
i	Y_{1i}	Y_{0i}	D_i	X_i
1	$\mathbb{E}[Y_1 \mid X=0, D=1]$	$\mathbb{E}[Y_0 \mid X=0, D=1]=$	1	0
2		$\mathbb{E}[Y_0 \mid X=0, D=0]$	1	0
3	$\mathbb{E}[Y_1 \mid X=0, D=0] =$	$\mathbb{E}[Y_0 \mid X=0, D=0]$	0	0
4	$\mathbb{E}[Y_1 \mid X=0, D=1]$		0	0
5	$\mathbb{E}[Y_1 \mid X=1, D=1]$	$\mathbb{E}[Y_0 \mid X=1, D=1]=$	1	1
6		$\mathbb{E}[Y_0 \mid X=1, D=0]$	1	1
7	$\mathbb{E}[Y_1 \mid X=1, D=0] =$	$\mathbb{E}[Y_0 \mid X=1, D=0]$	0	1
8	$\mathbb{E}[Y_1 \mid X=1, D=1]$		0	1

$(Y_1, Y_0) \perp\!\!\!\perp D \mid X$ also implies

$$\begin{aligned} \mathbb{E}[Y_1 \mid X=0, D=1] &= \mathbb{E}[Y_1 \mid X=0, D=0] \text{ and} \\ \mathbb{E}[Y_1 \mid X=1, D=1] &= \mathbb{E}[Y_1 \mid X=1, D=0] \end{aligned}$$

What is confounding?

- Confounding is the bias caused by common causes of the treatment and outcome.

What is confounding?

- Confounding is the bias caused by common causes of the treatment and outcome.
  - Leads to "spurious correlation."

What is confounding?

- Confounding is the bias caused by common causes of the treatment and outcome.
  - Leads to "spurious correlation."
- In observational studies, the goal is to avoid confounding inherent in the data.

What is confounding?

- Confounding is the bias caused by common causes of the treatment and outcome.
    - Leads to "spurious correlation."
- In observational studies, the goal is to avoid confounding inherent in the data.
- Pervasive in the social sciences:

What is confounding?

- Confounding is the bias caused by common causes of the treatment and outcome.
  - Leads to "spurious correlation."
- In observational studies, the goal is to avoid confounding inherent in the data.
- Pervasive in the social sciences:
  - effect of income on voting (confounding: age)

What is confounding?

- Confounding is the bias caused by common causes of the treatment and outcome.
  - Leads to "spurious correlation."
- In observational studies, the goal is to avoid confounding inherent in the data.
- Pervasive in the social sciences:
  - effect of income on voting (confounding: age)
  - effect of job training program on employment (confounding: motivation)

What is confounding?

- Confounding is the bias caused by common causes of the treatment and outcome.
  - Leads to "spurious correlation."
- In observational studies, the goal is to avoid confounding inherent in the data.
- Pervasive in the social sciences:
  - effect of income on voting (confounding: age)
  - effect of job training program on employment (confounding: motivation)
  - effect of political institutions on economic development (confounding: previous economic development)

What is confounding?

- Confounding is the bias caused by common causes of the treatment and outcome.
    - Leads to "spurious correlation."
- In observational studies, the goal is to avoid confounding inherent in the data.
- Pervasive in the social sciences:
    - effect of income on voting (confounding: age)
    - effect of job training program on employment (confounding: motivation)
    - effect of political institutions on economic development (confounding: previous economic development)
- No unmeasured confounding assumes that we've measured all sources of confounding.

Big problem

- How can we determine if no unmeasured confounding holds if we didn't assign the treatment?

Big problem

- How can we determine if no unmeasured confounding holds if we didn't assign the treatment?
- Put differently:

Big problem

- How can we determine if no unmeasured confounding holds if we didn't assign the treatment?
- Put differently:
    - What covariates do we need to condition on?

Big problem

- How can we determine if no unmeasured confounding holds if we didn't assign the treatment?
- Put differently:
  - What covariates do we need to condition on?
  - What covariates do we need to include in our regressions?

Big problem

- How can we determine if no unmeasured confounding holds if we didn't assign the treatment?
- Put differently:
  - What covariates do we need to condition on?
  - What covariates do we need to include in our regressions?
- One way, from the assumption itself:

Big problem

- How can we determine if no unmeasured confounding holds if we didn't assign the treatment?
- Put differently:
  - What covariates do we need to condition on?
  - What covariates do we need to include in our regressions?
- One way, from the assumption itself:
  - $\mathbb{P}[D_i = 1|\mathbf{X}, \mathbf{Y}(1), \mathbf{Y}(0)] = \mathbb{P}[D_i = 1|\mathbf{X}]$

Big problem

- How can we determine if no unmeasured confounding holds if we didn't assign the treatment?
- Put differently:
  - What covariates do we need to condition on?
  - What covariates do we need to include in our regressions?
- One way, from the assumption itself:
  - $\mathbb{P}[D_i = 1|\mathbf{X}, \mathbf{Y}(1), \mathbf{Y}(0)] = \mathbb{P}[D_i = 1|\mathbf{X}]$
  - Include covariates such that, conditional on them, the treatment assignment does not depend on the potential outcomes.

Big problem

- How can we determine if no unmeasured confounding holds if we didn't assign the treatment?
- Put differently:
  - What covariates do we need to condition on?
  - What covariates do we need to include in our regressions?
- One way, from the assumption itself:
  - $\mathbb{P}[D_i = 1|\mathbf{X}, \mathbf{Y}(1), \mathbf{Y}(0)] = \mathbb{P}[D_i = 1|\mathbf{X}]$
  - Include covariates such that, conditional on them, the treatment assignment does not depend on the potential outcomes.
- Another way: use DAGs and look at back-door paths.

Backdoor paths and blocking paths

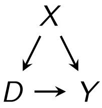
- Backdoor path: is a non-causal path from D to Y.

Backdoor paths and blocking paths

- Backdoor path: is a non-causal path from D to Y.
  - Would remain if we removed any arrows pointing out of D.

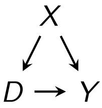
Backdoor paths and blocking paths

- Backdoor path: is a non-causal path from D to Y.
  - Would remain if we removed any arrows pointing out of D.
- Backdoor paths between D and Y $\rightsquigarrow$ common causes of D and Y:



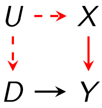
Backdoor paths and blocking paths

- Backdoor path: is a non-causal path from D to Y.
  - Would remain if we removed any arrows pointing out of D.
- Backdoor paths between D and Y $\rightsquigarrow$ common causes of D and Y:



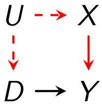
- Here there is a backdoor path $D \leftarrow X \rightarrow Y$, where X is a common cause for the treatment and the outcome.

Other types of confounding



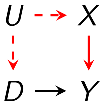
- D is enrolling in a job training program.

Other types of confounding



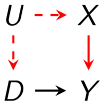
- D is enrolling in a job training program.
- Y is getting a job.

Other types of confounding



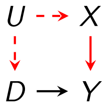
- D is enrolling in a job training program.
- Y is getting a job.
- U is being motivated

Other types of confounding



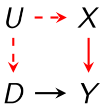
- D is enrolling in a job training program.
- Y is getting a job.
- U is being motivated
- X is number of job applications sent out.

Other types of confounding



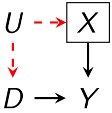
- D is enrolling in a job training program.
- Y is getting a job.
- U is being motivated
- X is number of job applications sent out.
- Big assumption here: no arrow from U to Y.

Other types of confounding



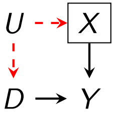
- D is exercise.
- Y is having a disease.
- U is lifestyle.
- X is smoking
- Big assumption here: no arrow from U to Y.

What's the problem with backdoor paths?



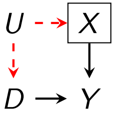
- A path is blocked if:

What's the problem with backdoor paths?



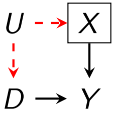
- A path is blocked if:
  1. we control for or stratify a non-collider on that path OR

What's the problem with backdoor paths?



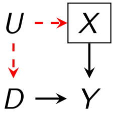
- A path is blocked if:
  1. we control for or stratify a non-collider on that path OR
  2. we do not control for a collider.

What's the problem with backdoor paths?



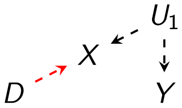
- A path is blocked if:
  1. we control for or stratify a non-collider on that path OR
  2. we do not control for a collider.
- Unblocked backdoor paths $\rightsquigarrow$ confounding.

What's the problem with backdoor paths?



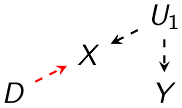
- A path is blocked if:
  1. we control for or stratify a non-collider on that path OR
  2. we do not control for a collider.
- Unblocked backdoor paths $\rightsquigarrow$ confounding.
- In the DAG here, if we condition on X, then the backdoor path is blocked.

Not all backdoor paths



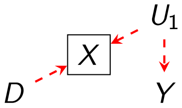
- Conditioning on the posttreatment covariates opens the non-causal path.

Not all backdoor paths



- Conditioning on the posttreatment covariates opens the non-causal path.
    - $\rightsquigarrow$ selection bias.

Not all backdoor paths



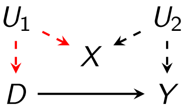
- Conditioning on the posttreatment covariates opens the non-causal path.
  - $\rightsquigarrow$ selection bias.

Don't condition on post-treatment variables

Don't condition on post-treatment variables

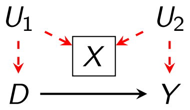
Every time you do, a puppy cries.

M-bias



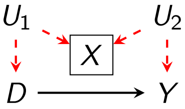
- Not all backdoor paths induce confounding.

M-bias



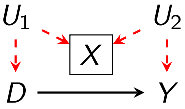
- Not all backdoor paths induce confounding.
- This backdoor path is blocked by the collider X that we don't control for.

M-bias



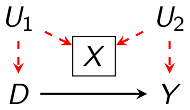
- Not all backdoor paths induce confounding.
- This backdoor path is blocked by the collider X that we don't control for.
- If we control for $X \rightsquigarrow$ opens the path and induces confounding.

M-bias



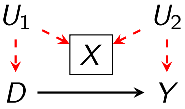
- Not all backdoor paths induce confounding.
- This backdoor path is blocked by the collider X that we don't control for.
- If we control for X $\rightsquigarrow$ opens the path and induces confounding.
    - Sometimes called M-bias.

M-bias



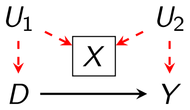
- Not all backdoor paths induce confounding.
- This backdoor path is blocked by the collider X that we don't control for.
- If we control for X $\rightsquigarrow$ opens the path and induces confounding.
  - Sometimes called M-bias.
- Controversial because of differing views on what to control for:

M-bias



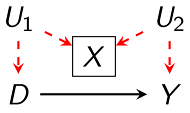
- Not all backdoor paths induce confounding.
- This backdoor path is blocked by the collider X that we don't control for.
- If we control for $X \rightsquigarrow$ opens the path and induces confounding.
  - Sometimes called M-bias.
- Controversial because of differing views on what to control for:
  - Rubin thinks that M-bias is a "mathematical curiosity" and we should control for all pretreatment variables

M-bias



- Not all backdoor paths induce confounding.
- This backdoor path is blocked by the collider X that we don't control for.
- If we control for X $\rightsquigarrow$ opens the path and induces confounding.
  - Sometimes called M-bias.
- Controversial because of differing views on what to control for:
  - Rubin thinks that M-bias is a "mathematical curiosity" and we should control for all pretreatment variables
  - Pearl and others think M-bias is a real threat.

M-bias



- Not all backdoor paths induce confounding.
- This backdoor path is blocked by the collider X that we don't control for.
- If we control for X $\rightsquigarrow$ opens the path and induces confounding.
  - Sometimes called M-bias.
- Controversial because of differing views on what to control for:
  - Rubin thinks that M-bias is a "mathematical curiosity" and we should control for all pretreatment variables
  - Pearl and others think M-bias is a real threat.
  - See the Elwert and Winship piece for more!

Backdoor criterion

- Can we use a DAG to evaluate no unmeasured confounders?

Backdoor criterion

- Can we use a DAG to evaluate no unmeasured confounders?
- Pearl answered yes, with the backdoor criterion, which states that the effect of D on Y is identified if:

Backdoor criterion

- Can we use a DAG to evaluate no unmeasured confounders?
- Pearl answered yes, with the backdoor criterion, which states that the effect of D on Y is identified if:
  1. No backdoor paths from D to Y OR

Backdoor criterion

- Can we use a DAG to evaluate no unmeasured confounders?
- Pearl answered yes, with the backdoor criterion, which states that the effect of D on Y is identified if:
  1. No backdoor paths from D to Y OR
  2. Measured covariates are sufficient to block all backdoor paths from D to Y.

Backdoor criterion

- Can we use a DAG to evaluate no unmeasured confounders?
- Pearl answered yes, with the backdoor criterion, which states that the effect of D on Y is identified if:
  1. No backdoor paths from D to Y OR
  2. Measured covariates are sufficient to block all backdoor paths from D to Y.
- First is really only valid for randomized experiments.

Backdoor criterion

- Can we use a DAG to evaluate no unmeasured confounders?
- Pearl answered yes, with the backdoor criterion, which states that the effect of D on Y is identified if:
  1. No backdoor paths from D to Y OR
  2. Measured covariates are sufficient to block all backdoor paths from D to Y.
- First is really only valid for randomized experiments.
- The backdoor criterion is fairly powerful. Tells us:

Backdoor criterion

- Can we use a DAG to evaluate no unmeasured confounders?
- Pearl answered yes, with the backdoor criterion, which states that the effect of D on Y is identified if:
  1. No backdoor paths from D to Y OR
  2. Measured covariates are sufficient to block all backdoor paths from D to Y.
- First is really only valid for randomized experiments.
- The backdoor criterion is fairly powerful. Tells us:
  - if there is confounding given this DAG,

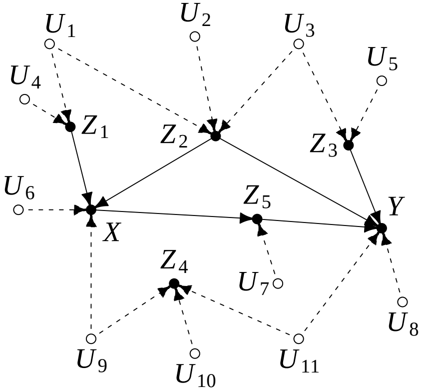
Backdoor criterion

- Can we use a DAG to evaluate no unmeasured confounders?
- Pearl answered yes, with the backdoor criterion, which states that the effect of D on Y is identified if:
  1. No backdoor paths from D to Y OR
  2. Measured covariates are sufficient to block all backdoor paths from D to Y.
- First is really only valid for randomized experiments.
- The backdoor criterion is fairly powerful. Tells us:
  - if there is confounding given this DAG,
  - if it is possible to remove the confounding, and

Backdoor criterion

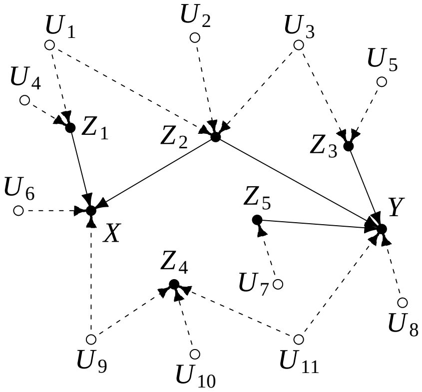
- Can we use a DAG to evaluate no unmeasured confounders?
- Pearl answered yes, with the backdoor criterion, which states that the effect of D on Y is identified if:
  1. No backdoor paths from D to Y OR
  2. Measured covariates are sufficient to block all backdoor paths from D to Y.
- First is really only valid for randomized experiments.
- The backdoor criterion is fairly powerful. Tells us:
  - if there is confounding given this DAG,
  - if it is possible to remove the confounding, and
  - what variables to condition on to eliminate the confounding.

Example: Sufficient Conditioning Sets



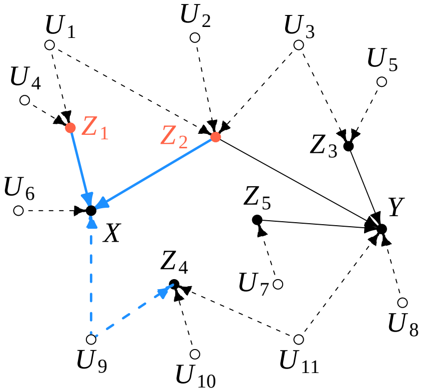
Remove arrows out of X.

Example: Sufficient Conditioning Sets



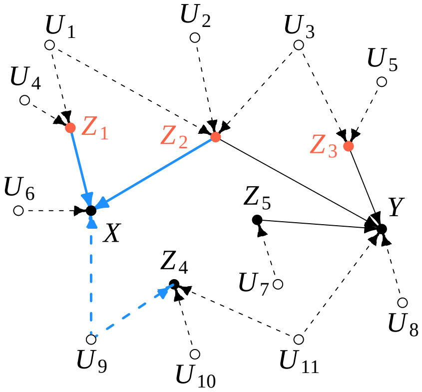
Recall that paths are blocked by "unconditioned colliders" or conditioned non-colliders

Example: Sufficient Conditioning Sets



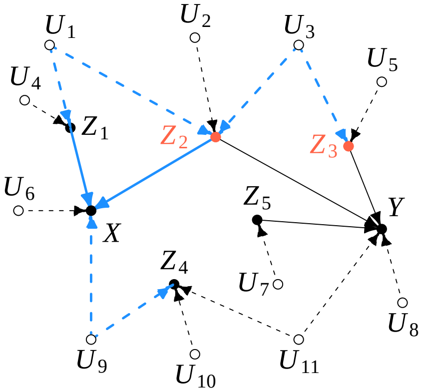
Recall that paths are blocked by "unconditioned colliders" or conditioned non-colliders

Example: Sufficient Conditioning Sets



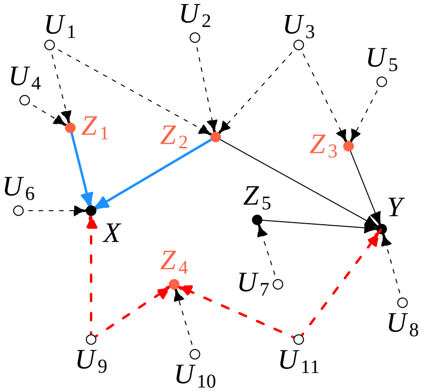
Recall that paths are blocked by "unconditioned colliders" or conditioned non-colliders

Example: Sufficient Conditioning Sets



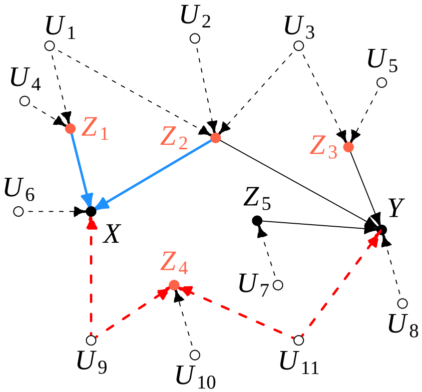
Recall that paths are blocked by "unconditioned colliders" or conditioned non-colliders

Example: Non-sufficient Conditioning Sets

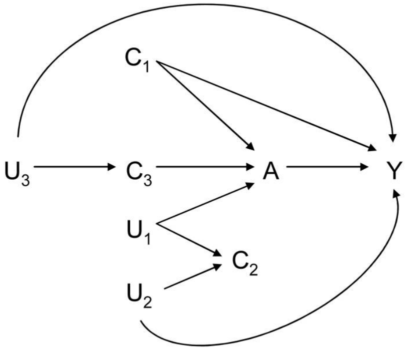


Recall that paths are blocked by "unconditioned colliders" or conditioned non-colliders

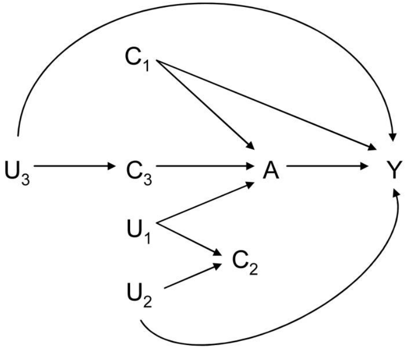
Example: Non-sufficient Conditioning Sets



Implications (via Vanderweele and Shpitser 2011)



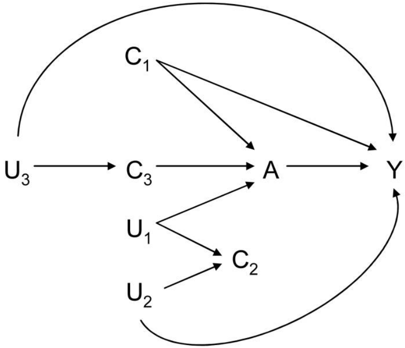
Implications (via Vanderweele and Shpitser 2011)



Two common criteria fail here:

1. Choose all pre-treatment covariates
2. Choose all covariates which directly cause the treatment and the outcome

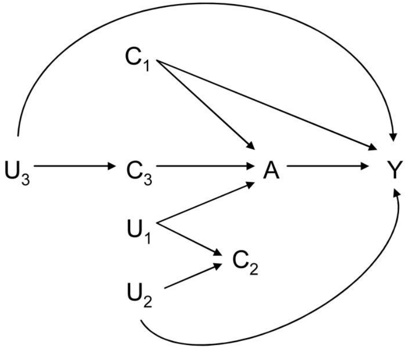
Implications (via Vanderweele and Shpitser 2011)



Two common criteria fail here:

1. Choose all pre-treatment covariates
   (would condition on C_2 inducing M-bias)
2. Choose all covariates which directly cause the treatment and the outcome

Implications (via Vanderweele and Shpitser 2011)



Two common criteria fail here:

1. Choose all pre-treatment covariates
   (would condition on C_2 inducing M-bias)
2. Choose all covariates which directly cause the treatment and the outcome
   (would leave open a backdoor path $A \leftarrow C_3 \leftarrow U_3 \rightarrow Y$.)

No unmeasured confounders is not testable

- No unmeasured confounding places no restrictions on the observed data.

$$\underbrace{\big(Y_i(0)\big|D_i = 1, X_i\big)}_{\text{unobserved}} \overset{d}{=} \underbrace{\big(Y_i(0)\big|D_i = 0, X_i\big)}_{\text{observed}}$$

No unmeasured confounders is not testable

- No unmeasured confounding places no restrictions on the observed data.

$$\underbrace{\left(Y_i(0)\middle|D_i = 1, X_i\right)}_{\text{unobserved}} \overset{d}{=} \underbrace{\left(Y_i(0)\middle|D_i = 0, X_i\right)}_{\text{observed}}$$

- Here, $\overset{d}{=}$ means equal in distribution.

No unmeasured confounders is not testable

- No unmeasured confounding places no restrictions on the observed data.

$$\underbrace{\left(Y_i(0)\middle|D_i = 1, X_i\right)}_{\text{unobserved}} \stackrel{d}{=} \underbrace{\left(Y_i(0)\middle|D_i = 0, X_i\right)}_{\text{observed}}$$

- Here, $\stackrel{d}{=}$ means equal in distribution.
- No way to directly test this assumption without the counterfactual data, which is missing by definition!

No unmeasured confounders is not testable

- No unmeasured confounding places no restrictions on the observed data.

$$\underbrace{\left(Y_i(0)\middle|D_i = 1, X_i\right)}_{\text{unobserved}} \stackrel{d}{=} \underbrace{\left(Y_i(0)\middle|D_i = 0, X_i\right)}_{\text{observed}}$$

- Here, $\stackrel{d}{=}$ means equal in distribution.
- No way to directly test this assumption without the counterfactual data, which is missing by definition!
- With backdoor criterion, you must have the correct DAG.

Assessing no unmeasured confounders

TABLE VI
THE FOX NEWS EFFECT: INTERACTIONS AND PLACEBO SPECIFICATIONS

	Interactions		Placebo specifications		
	Presid. Rep. vote share 2000–1996		Presidential Republican vote share 2000–1996	Presidential Republican vote share 1996–1992	Presidential Republican vote share 1992–1988
Dep. var.	(1)	(2)	(3)	(4)	(5)
Availability of Fox News via cable in 2000	0.0109 (0.0042)***	0.0105 (0.0039)***	0.0036 (0.0016)**	−0.0024 (0.0031)	0.0026 (0.0026)
Availability of Fox News via cable in 2003			−0.0001 (0.0012)		

- Can do "placebo" tests, where D_i cannot have an effect (lagged outcomes, etc)

Assessing no unmeasured confounders

TABLE VI
THE FOX NEWS EFFECT: INTERACTIONS AND PLACEBO SPECIFICATIONS

	Interactions		Placebo specifications		
	Presid. Rep. vote share 2000–1996		Presidential Republican vote share		
			2000–1996	1996–1992	1992–1988
Dep. var.	(1)	(2)	(3)	(4)	(5)
Availability of Fox News via cable in 2000	0.0109 (0.0042)***	0.0105 (0.0039)***	0.0036 (0.0016)**	−0.0024 (0.0031)	0.0026 (0.0026)
Availability of Fox News via cable in 2003			−0.0001 (0.0012)		

- Can do "placebo" tests, where D_i cannot have an effect (lagged outcomes, etc)
- Della Vigna and Kaplan (2007, QJE): effect of Fox News availability on Republican vote share

Assessing no unmeasured confounders

TABLE VI
THE FOX NEWS EFFECT: INTERACTIONS AND PLACEBO SPECIFICATIONS

	Interactions		Placebo specifications		
	Presid. Rep. vote share 2000–1996		Presidential Republican vote share 2000–1996	1996–1992	1992–1988
Dep. var.	(1)	(2)	(3)	(4)	(5)
Availability of Fox News via cable in 2000	0.0109 (0.0042)***	0.0105 (0.0039)***	0.0036 (0.0016)**	−0.0024 (0.0031)	0.0026 (0.0026)
Availability of Fox News via cable in 2003			−0.0001 (0.0012)		

- Can do "placebo" tests, where D_i cannot have an effect (lagged outcomes, etc)
- Della Vigna and Kaplan (2007, QJE): effect of Fox News availability on Republican vote share
    - Availability in 2000/2003 can't affect past vote shares.

Assessing no unmeasured confounders

TABLE VI
THE FOX NEWS EFFECT: INTERACTIONS AND PLACEBO SPECIFICATIONS

	Interactions		Placebo specifications		
	Presid. Rep. vote share 2000–1996		Presidential Republican vote share 2000–1996	Presidential Republican vote share 1996–1992	Presidential Republican vote share 1992–1988
Dep. var.	(1)	(2)	(3)	(4)	(5)
Availability of Fox News via cable in 2000	0.0109 (0.0042)***	0.0105 (0.0039)***	0.0036 (0.0016)**	−0.0024 (0.0031)	0.0026 (0.0026)
Availability of Fox News via cable in 2003			−0.0001 (0.0012)		

- Can do "placebo" tests, where D_i cannot have an effect (lagged outcomes, etc)
- Della Vigna and Kaplan (2007, QJE): effect of Fox News availability on Republican vote share
    - Availability in 2000/2003 can't affect past vote shares.
- Unconfoundedness could still be violated even if you pass this test!

Alternatives to no unmeasured confounding

- Without explicit randomization, we need some way of identifying causal effects.

Alternatives to no unmeasured confounding

- Without explicit randomization, we need some way of identifying causal effects.
- No unmeasured confounders $\approx$ randomized experiment.

Alternatives to no unmeasured confounding

- Without explicit randomization, we need some way of identifying causal effects.
- No unmeasured confounders $\approx$ randomized experiment.
    - Identification results very similar to experiments.

Alternatives to no unmeasured confounding

- Without explicit randomization, we need some way of identifying causal effects.
- No unmeasured confounders $\approx$ randomized experiment.
  - Identification results very similar to experiments.
- With unmeasured confounding are we doomed? Maybe not!

Alternatives to no unmeasured confounding

- Without explicit randomization, we need some way of identifying causal effects.
- No unmeasured confounders $\approx$ randomized experiment.
  - Identification results very similar to experiments.
- With unmeasured confounding are we doomed? Maybe not!
- Other approaches rely on finding plausibly exogenous variation in assignment of D_i:

Alternatives to no unmeasured confounding

- Without explicit randomization, we need some way of identifying causal effects.
- No unmeasured confounders $\approx$ randomized experiment.
  - Identification results very similar to experiments.
- With unmeasured confounding are we doomed? Maybe not!
- Other approaches rely on finding plausibly exogenous variation in assignment of D_i:
  - Instrumental variables (randomization + exclusion restriction)

Alternatives to no unmeasured confounding

- Without explicit randomization, we need some way of identifying causal effects.
- No unmeasured confounders $\approx$ randomized experiment.
  - Identification results very similar to experiments.
- With unmeasured confounding are we doomed? Maybe not!
- Other approaches rely on finding plausibly exogenous variation in assignment of D_i:
  - Instrumental variables (randomization + exclusion restriction)
  - Over-time variation (diff-in-diff, fixed effects)

Alternatives to no unmeasured confounding

- Without explicit randomization, we need some way of identifying causal effects.
- No unmeasured confounders $\approx$ randomized experiment.
  - Identification results very similar to experiments.
- With unmeasured confounding are we doomed? Maybe not!
- Other approaches rely on finding plausibly exogenous variation in assignment of D_i:
  - Instrumental variables (randomization + exclusion restriction)
  - Over-time variation (diff-in-diff, fixed effects)
  - Arbitrary thresholds for treatment assignment (RDD)

Alternatives to no unmeasured confounding

- Without explicit randomization, we need some way of identifying causal effects.
- No unmeasured confounders $\approx$ randomized experiment.
  - Identification results very similar to experiments.
- With unmeasured confounding are we doomed? Maybe not!
- Other approaches rely on finding plausibly exogenous variation in assignment of D_i:
  - Instrumental variables (randomization + exclusion restriction)
  - Over-time variation (diff-in-diff, fixed effects)
  - Arbitrary thresholds for treatment assignment (RDD)
  - All discussed in the next couple of weeks!

Where We've Been and Where We're Going...

- Last Week
  - intro to causal inference
- This Week
  - Monday:
    - experimental Ideal
    - identification with measured confounding
  - Wednesday:
    - regression estimation
- Next Week
  - identification with unmeasured confounding
  - instrumental variables
- Long Run
  - probability → inference → regression → causal inference

Questions?

Identification vs. Estimation

Identification vs. Estimation

- An approximately ordered causal workflow:

Identification vs. Estimation

- An approximately ordered causal workflow:
  1) Question $\leftarrow$ the thing we care about

Identification vs. Estimation

- An approximately ordered causal workflow:
  1) Question $\leftarrow$ the thing we care about
  2) Ideal Experiment $\leftarrow$ what's the counterfactual we care about

Identification vs. Estimation

- An approximately ordered causal workflow:
  1) Question $\leftarrow$ the thing we care about
  2) Ideal Experiment $\leftarrow$ what's the counterfactual we care about
  3) Estimand $\leftarrow$ the causal quantity of interest

Identification vs. Estimation

- An approximately ordered causal workflow:
  1) Question $\leftarrow$ the thing we care about
  2) Ideal Experiment $\leftarrow$ what's the counterfactual we care about
  3) Estimand $\leftarrow$ the causal quantity of interest
  4) Identification Strategy $\leftarrow$ how we connect features of a probability distribution of observed data to causal estimand.

Identification vs. Estimation

- An approximately ordered causal workflow:
  1) Question ← the thing we care about
  2) Ideal Experiment ← what's the counterfactual we care about
  3) Estimand ← the causal quantity of interest
  4) Identification Strategy ← how we connect features of a probability distribution of observed data to causal estimand.
  5) Estimation ← how we estimate a feature of a probability distribution from observed data.

Identification vs. Estimation

- An approximately ordered causal workflow:
  1) Question $\leftarrow$ the thing we care about
  2) Ideal Experiment $\leftarrow$ what's the counterfactual we care about
  3) Estimand $\leftarrow$ the causal quantity of interest
  4) Identification Strategy $\leftarrow$ how we connect features of a probability distribution of observed data to causal estimand.
  5) Estimation $\leftarrow$ how we estimate a feature of a probability distribution from observed data.
  6) Inference/Uncertainty $\leftarrow$ what would have happened if we observed a different treatment assignment? (and possibly sampled a different population)

Identification vs. Estimation

- An approximately ordered causal workflow:
  1) Question $\leftarrow$ the thing we care about
  2) Ideal Experiment $\leftarrow$ what's the counterfactual we care about
  3) Estimand $\leftarrow$ the causal quantity of interest
  4) Identification Strategy $\leftarrow$ how we connect features of a probability distribution of observed data to causal estimand.
  5) Estimation $\leftarrow$ how we estimate a feature of a probability distribution from observed data.
  6) Inference/Uncertainty $\leftarrow$ what would have happened if we observed a different treatment assignment? (and possibly sampled a different population)
- 'Whats your identification strategy?' means 'what are the assumptions that allow you to claim that the association you've estimated has a causal interpretation?'

Identification vs. Estimation

- An approximately ordered causal workflow:
  1) Question $\leftarrow$ the thing we care about
  2) Ideal Experiment $\leftarrow$ what's the counterfactual we care about
  3) Estimand $\leftarrow$ the causal quantity of interest
  4) Identification Strategy $\leftarrow$ how we connect features of a probability distribution of observed data to causal estimand.
  5) Estimation $\leftarrow$ how we estimate a feature of a probability distribution from observed data.
  6) Inference/Uncertainty $\leftarrow$ what would have happened if we observed a different treatment assignment? (and possibly sampled a different population)
- 'Whats your identification strategy?' means 'what are the assumptions that allow you to claim that the association you've estimated has a causal interpretation?'
- Selection on observables is an identification strategy

Identification vs. Estimation

- An approximately ordered causal workflow:
  1) Question $\leftarrow$ the thing we care about
  2) Ideal Experiment $\leftarrow$ what's the counterfactual we care about
  3) Estimand $\leftarrow$ the causal quantity of interest
  4) Identification Strategy $\leftarrow$ how we connect features of a probability distribution of observed data to causal estimand.
  5) Estimation $\leftarrow$ how we estimate a feature of a probability distribution from observed data.
  6) Inference/Uncertainty $\leftarrow$ what would have happened if we observed a different treatment assignment? (and possibly sampled a different population)
- 'Whats your identification strategy?' means 'what are the assumptions that allow you to claim that the association you've estimated has a causal interpretation?'
- Selection on observables is an identification strategy
- Identification depends on assumptions not statistical models.

Estimation

Estimation

- Estimation is secondary to identification.

Estimation

- Estimation is secondary to identification.
- Selection on observables generally requires estimating at least one conditional expectation function and there are many ways to do that.

Estimation

- Estimation is secondary to identification.
- Selection on observables generally requires estimating at least one conditional expectation function and there are many ways to do that.
- An incomplete list of strategies:
  - matching
  - weighting
  - regression
  - combinations of the above

Estimation

- Estimation is secondary to identification.
- Selection on observables generally requires estimating at least one conditional expectation function and there are many ways to do that.
- An incomplete list of strategies:
  - matching
  - weighting
  - regression
  - combinations of the above
- Today we will talk about regression because that's the subject of the class.

Estimation

- Estimation is secondary to identification.
- Selection on observables generally requires estimating at least one conditional expectation function and there are many ways to do that.
- An incomplete list of strategies:
  - matching
  - weighting
  - regression
  - combinations of the above
- Today we will talk about regression because that's the subject of the class.
- A big topic I'm skipping over as outside the scope of class is the propensity score (conditional expectation of the treatment given the covariates).

Regression

David Freedman:

> *I sometimes have a nightmare about Kepler. Suppose a few of us were transported back in time to the year 1600, and were invited by the Emperor Rudolph II to set up an Imperial Department of Statistics in the court at Prague. Despairing of those circular orbits, Kepler enrolls in our department. We teach him the general linear model, least squares, dummy variables, everything. He goes back to work, fits the best circular orbit for Mars by least squares, puts in a dummy variable for the exceptional observation - and publishes. And that's the end, right there in Prague at the beginning of the 17th century.*

Regression and Causality

Regression and Causality

- Regression is an estimation strategy that can be used with an identification strategy to estimate a causal effect

Regression and Causality

- Regression is an estimation strategy that can be used with an identification strategy to estimate a causal effect
- When is regression causal?

Regression and Causality

- Regression is an estimation strategy that can be used with an identification strategy to estimate a causal effect
- When is regression causal? When the CEF is causal.

Regression and Causality

- Regression is an estimation strategy that can be used with an identification strategy to estimate a causal effect
- When is regression causal? When the CEF is causal.
- This means that the question of whether regression has a causal interpretation is a question about identification

Identification under Selection on Observables: Regression

Consider the linear regression of $Y_i = \beta_0 + \tau D_i + X_i'\beta + \epsilon_i$.

Identification under Selection on Observables: Regression

Consider the linear regression of $Y_i = \beta_0 + \tau D_i + X_i' \beta + \epsilon_i$.

Given selection on observables, there are mainly three identification scenarios:

Identification under Selection on Observables: Regression

Consider the linear regression of $Y_i = \beta_0 + \tau D_i + X_i' \beta + \epsilon_i$.

Given selection on observables, there are mainly three identification scenarios:

1. Constant treatment effects and outcomes are linear in X
   - τ will provide unbiased and consistent estimates of ATE.
2. Constant treatment effects and unknown functional form
   - τ will provide well-defined linear approximation to the average causal response function $\mathbb{E}[Y|D=1,X] - \mathbb{E}[Y|D=0,X]$. Approximation may be very poor if $\mathbb{E}[Y|D,X]$ is misspecified and then τ may be biased for the ATE.

Identification under Selection on Observables: Regression

Consider the linear regression of $Y_i = \beta_0 + \tau D_i + X_i'\beta + \epsilon_i$.

Given selection on observables, there are mainly three identification scenarios:

1. Constant treatment effects and outcomes are linear in X
   - τ will provide unbiased and consistent estimates of ATE.
2. Constant treatment effects and unknown functional form
   - τ will provide well-defined linear approximation to the average causal response function $\mathbb{E}[Y|D=1,X] - \mathbb{E}[Y|D=0,X]$. Approximation may be very poor if $\mathbb{E}[Y|D,X]$ is misspecified and then τ may be biased for the ATE.
3. Heterogeneous treatment effects (τ differs for different values of X)
   - If outcomes are linear in X, τ is unbiased and consistent estimator for conditional-variance-weighted average of the underlying causal effects. This averagecan be different from the ATE.

Identification under Selection on Observables: Regression

Identification Assumption

1. *Constant treatment effect:* $\tau = Y_{1i} - Y_{0i}$ *for all* i
2. *Control outcome is linear in* X: $Y_{0i} = \beta_0 + X_i'\beta + \epsilon_i$ *with* $\epsilon_i \perp\!\!\!\perp X_i$ *(no omitted variables and linearly separable confounding)*

Identification Result

Then $\tau_{ATE} = \mathbb{E}[Y_1 - Y_0]$ *is identified by a regression of the observed outcome on the covariates and the treatment indicator*
$Y_i = \beta_0 + \tau D_i + X_i'\beta + \epsilon_i$

Ideal Case: Linear Constant Effects Model

Assume constant linear effects and linearly separable confounding:

$$Y_i(d) = Y_i = \beta_0 + \tau D_i + \eta_i$$

Ideal Case: Linear Constant Effects Model

Assume constant linear effects and linearly separable confounding:

$$Y_i(d) = Y_i = \beta_0 + \tau D_i + \eta_i$$

- **Linearly separable confounding:** assume that $\mathbb{E}[\eta_i | X_i] = X_i'\beta$, which means that $\eta_i = X_i'\beta + \epsilon_i$ where $\mathbb{E}[\epsilon_i | X_i] = 0$.

Ideal Case: Linear Constant Effects Model

Assume constant linear effects and linearly separable confounding:

$$Y_i(d) = Y_i = \beta_0 + \tau D_i + \eta_i$$

- **Linearly separable confounding:** assume that $\mathbb{E}[\eta_i|X_i] = X_i'\beta$, which means that $\eta_i = X_i'\beta + \epsilon_i$ where $\mathbb{E}[\epsilon_i|X_i] = 0$.
- Under this model, $(Y_1, Y_0) \perp\!\!\!\perp D|X$ implies $\epsilon_i|X \perp\!\!\!\perp D$

Ideal Case: Linear Constant Effects Model

Assume constant linear effects and linearly separable confounding:

$$Y_i(d) = Y_i = \beta_0 + \tau D_i + \eta_i$$

- **Linearly separable confounding:** assume that $\mathbb{E}[\eta_i | X_i] = X_i' \beta$, which means that $\eta_i = X_i'\beta + \epsilon_i$ where $\mathbb{E}[\epsilon_i | X_i] = 0$.
- Under this model, $(Y_1, Y_0) \perp\!\!\!\perp D | X$ implies $\epsilon_i | X \perp\!\!\!\perp D$
- As a result,

$$\begin{aligned} Y_i &= \beta_0 + \tau D_i + \mathbb{E}[\eta_i] \\ &= \beta_0 + \tau D_i + X_i'\beta + \mathbb{E}[\epsilon_i] \\ &= \beta_0 + \tau D_i + X_i'\beta \end{aligned}$$

Ideal Case: Linear Constant Effects Model

Assume constant linear effects and linearly separable confounding:

$$Y_i(d) = Y_i = \beta_0 + \tau D_i + \eta_i$$

- **Linearly separable confounding:** assume that $\mathbb{E}[\eta_i|X_i] = X_i'\beta$, which means that $\eta_i = X_i'\beta + \epsilon_i$ where $\mathbb{E}[\epsilon_i|X_i] = 0$.
- Under this model, $(Y_1, Y_0) \perp\!\!\!\perp D|X$ implies $\epsilon_i|X \perp\!\!\!\perp D$
- As a result,

$$\begin{aligned} Y_i &= \beta_0 + \tau D_i + \mathbb{E}[\eta_i] \\ &= \beta_0 + \tau D_i + X_i'\beta + \mathbb{E}[\epsilon_i] \\ &= \beta_0 + \tau D_i + X_i'\beta \end{aligned}$$

- Thus, a regression where D_i and X_i are entered linearly can recover the ATE.

Implausible $\rightsquigarrow$ Plausible

Implausible $\rightsquigarrow$ Plausible

- Constant effects and linearly separable confounding aren't very appealing or plausible assumptions

Implausible $\rightsquigarrow$ Plausible

- Constant effects and linearly separable confounding aren't very appealing or plausible assumptions
- To understand what happens when they don't hold, we need to understand the properties of regression with minimal assumptions: this is often called an agnostic view of regression.

Implausible $\rightsquigarrow$ Plausible

- Constant effects and linearly separable confounding aren't very appealing or plausible assumptions
- To understand what happens when they don't hold, we need to understand the properties of regression with minimal assumptions: this is often called an agnostic view of regression.
- The Aronow and Miller book (and lecture 7) provide some context but essentially as long as we have iid sampling, we will asymptotically obtain the best linear approximation to the CEF.

1. The Experimental Ideal
2. Assumption of No Unmeasured Confounding
3. Estimation Under No Unmeasured Confounding
4. Regression Estimators
5. Regression and Causality
6. Regression Under Heterogeneous Effects
7. Fun with Visualization, Replication and the NYT

Regression and causality

- Most econometrics textbooks: regression defined without respect to causality.

Regression and causality

- Most econometrics textbooks: regression defined without respect to causality.
- But then when is $\hat{\beta}$ "biased"? What does this even mean?

Regression and causality

- Most econometrics textbooks: regression defined without respect to causality.
- But then when is $\hat{\beta}$ "biased"? What does this even mean?
- The question, then, is when does knowing the CEF tell us something about causality?

Regression and causality

- Most econometrics textbooks: regression defined without respect to causality.
- But then when is $\hat{\beta}$ "biased"? What does this even mean?
- The question, then, is when does knowing the CEF tell us something about causality?
- Angrist and Pishke argues that a regression is causal when the CEF it approximates is causal. Identification is king.

Regression and causality

- Most econometrics textbooks: regression defined without respect to causality.
- But then when is $\hat{\beta}$ "biased"? What does this even mean?
- The question, then, is when does knowing the CEF tell us something about causality?
- Angrist and Pishke argues that a regression is causal when the CEF it approximates is causal. Identification is king.
- We will show that under certain conditions, a regression of the outcome on the treatment and the covariates can recover a causal parameter, but perhaps not the one in which we are interested.

Linear constant effects model, binary treatment

Now with the benefit of covering agnostic regression, let's review again the simple case.

- Experiment: with a simple experiment, we can rewrite the consistency assumption to be a regression formula:

Linear constant effects model, binary treatment

Now with the benefit of covering agnostic regression, let's review again the simple case.

- Experiment: with a simple experiment, we can rewrite the consistency assumption to be a regression formula:

$$Y_i = D_i Y_i(1) + (1 - D_i) Y_i(0)$$

Linear constant effects model, binary treatment

Now with the benefit of covering agnostic regression, let's review again the simple case.

- Experiment: with a simple experiment, we can rewrite the consistency assumption to be a regression formula:

$$\begin{aligned} Y_i &= D_i Y_i(1) + (1 - D_i) Y_i(0) \\ &= Y_i(0) + (Y_i(1) - Y_i(0)) D_i \end{aligned}$$

Linear constant effects model, binary treatment

Now with the benefit of covering agnostic regression, let's review again the simple case.

- Experiment: with a simple experiment, we can rewrite the consistency assumption to be a regression formula:

$$\begin{aligned} Y_i &= D_i Y_i(1) + (1 - D_i) Y_i(0) \\ &= Y_i(0) + (Y_i(1) - Y_i(0)) D_i \\ &= \mathbb{E}[Y_i(0)] + \tau D_i + (Y_i(0) - \mathbb{E}[Y_i(0)]) \end{aligned}$$

Linear constant effects model, binary treatment

Now with the benefit of covering agnostic regression, let's review again the simple case.

- Experiment: with a simple experiment, we can rewrite the consistency assumption to be a regression formula:

$$\begin{aligned} Y_i &= D_i Y_i(1) + (1 - D_i) Y_i(0) \\ &= Y_i(0) + (Y_i(1) - Y_i(0)) D_i \\ &= \mathbb{E}[Y_i(0)] + \tau D_i + (Y_i(0) - \mathbb{E}[Y_i(0)]) \\ &= \mu^0 + \tau D_i + v_i^0 \end{aligned}$$

Linear constant effects model, binary treatment

Now with the benefit of covering agnostic regression, let's review again the simple case.

- Experiment: with a simple experiment, we can rewrite the consistency assumption to be a regression formula:

$$\begin{aligned} Y_i &= D_i Y_i(1) + (1 - D_i) Y_i(0) \\ &= Y_i(0) + (Y_i(1) - Y_i(0)) D_i \\ &= \mathbb{E}[Y_i(0)] + \tau D_i + (Y_i(0) - \mathbb{E}[Y_i(0)]) \\ &= \mu^0 + \tau D_i + v_i^0 \end{aligned}$$

- Note that if ignorability holds (as in an experiment) for $Y_i(0)$, then it will also hold for v_i^0, since $\mathbb{E}[Y_i(0)]$ is constant. Thus, this satifies the usual assumptions for regression.

Now with covariates

- Now assume no unmeasured confounders: $Y_i(d) \perp\!\!\!\perp D_i | X_i$.

Now with covariates

- Now assume no unmeasured confounders: $Y_i(d) \perp\!\!\!\perp D_i | X_i$.
- We will assume a linear model for the potential outcomes:

$$Y_i(d) = \alpha + \tau \cdot d + \eta_i$$

Now with covariates

- Now assume no unmeasured confounders: $Y_i(d) \perp\!\!\!\perp D_i | X_i$.
- We will assume a linear model for the potential outcomes:

$$Y_i(d) = \alpha + \tau \cdot d + \eta_i$$

- Remember that linearity isn't an assumption if D_i is binary

Now with covariates

- Now assume no unmeasured confounders: $Y_i(d) \perp\!\!\!\perp D_i | X_i$.
- We will assume a linear model for the potential outcomes:

$$Y_i(d) = \alpha + \tau \cdot d + \eta_i$$

- Remember that linearity isn't an assumption if D_i is binary
- Effect of D_i is constant here, the η_i are the only source of individual variation and we have $E[\eta_i] = 0$.

Now with covariates

- Now assume no unmeasured confounders: $Y_i(d) \perp\!\!\!\perp D_i | X_i$.
- We will assume a linear model for the potential outcomes:

$$Y_i(d) = \alpha + \tau \cdot d + \eta_i$$

- Remember that linearity isn't an assumption if D_i is binary
- Effect of D_i is constant here, the η_i are the only source of individual variation and we have $E[\eta_i] = 0$.
- Consistency assumption allows us to write this as:

$$Y_i = \alpha + \tau D_i + \eta_i.$$

Covariates in the error

- Let's assume that η_i is linear in X_i: $\eta_i = X_i'\gamma + \nu_i$

Covariates in the error

- Let's assume that η_i is linear in X_i: $\eta_i = X_i'\gamma + \nu_i$
- New error is uncorrelated with X_i: $\mathbb{E}[\nu_i | X_i] = 0$.

Covariates in the error

- Let's assume that η_i is linear in X_i: $\eta_i = X_i'\gamma + \nu_i$
- New error is uncorrelated with X_i: $\mathbb{E}[\nu_i|X_i] = 0$.
- This is an assumption! Might be false!

Covariates in the error

- Let's assume that η_i is linear in X_i: $\eta_i = X_i'\gamma + \nu_i$
- New error is uncorrelated with X_i: $\mathbb{E}[\nu_i|X_i] = 0$.
- This is an assumption! Might be false!
- Plug into the above:

Covariates in the error

- Let's assume that η_i is linear in X_i: $\eta_i = X_i'\gamma + \nu_i$
- New error is uncorrelated with X_i: $\mathbb{E}[\nu_i|X_i] = 0$.
- This is an assumption! Might be false!
- Plug into the above:

$$\mathbb{E}[Y_i(d)|X_i] = E[Y_i|D_i, X_i]$$

Covariates in the error

- Let's assume that η_i is linear in X_i: $\eta_i = X_i'\gamma + \nu_i$
- New error is uncorrelated with X_i: $\mathbb{E}[\nu_i|X_i] = 0$.
- This is an assumption! Might be false!
- Plug into the above:

$$\mathbb{E}[Y_i(d)|X_i] = E[Y_i|D_i, X_i] = \alpha + \tau D_i + E[\eta_i|X_i]$$

Covariates in the error

- Let's assume that η_i is linear in X_i: $\eta_i = X_i'\gamma + \nu_i$
- New error is uncorrelated with X_i: $\mathbb{E}[\nu_i|X_i] = 0$.
- This is an assumption! Might be false!
- Plug into the above:

$$\begin{aligned}\mathbb{E}[Y_i(d)|X_i] = E[Y_i|D_i, X_i] &= \alpha + \tau D_i + E[\eta_i|X_i] \\ &= \alpha + \tau D_i + X_i'\gamma + E[\nu_i|X_i]\end{aligned}$$

Covariates in the error

- Let's assume that η_i is linear in X_i: $\eta_i = X_i'\gamma + \nu_i$
- New error is uncorrelated with X_i: $\mathbb{E}[\nu_i | X_i] = 0$.
- This is an assumption! Might be false!
- Plug into the above:

$$\begin{aligned}
\mathbb{E}[Y_i(d)|X_i] = E[Y_i|D_i, X_i] &= \alpha + \tau D_i + E[\eta_i|X_i] \\
&= \alpha + \tau D_i + X_i'\gamma + E[\nu_i|X_i] \\
&= \alpha + \tau D_i + X_i'\gamma
\end{aligned}$$

Summing up regression with constant effects

- Reviewing the assumptions we've used:

Summing up regression with constant effects

- Reviewing the assumptions we've used:
  - no unmeasured confounders

Summing up regression with constant effects

- Reviewing the assumptions we've used:
  - no unmeasured confounders
  - constant treatment effects

Summing up regression with constant effects

- Reviewing the assumptions we've used:
  - no unmeasured confounders
  - constant treatment effects
  - linearity of the treatment/covariates

Summing up regression with constant effects

- Reviewing the assumptions we've used:
  - no unmeasured confounders
  - constant treatment effects
  - linearity of the treatment/covariates
- Under these, we can run the following regression to estimate the ATE, τ:

$$Y_i = \alpha + \tau D_i + X_i' \gamma + \nu_i$$

Summing up regression with constant effects

- Reviewing the assumptions we've used:
  - no unmeasured confounders
  - constant treatment effects
  - linearity of the treatment/covariates
- Under these, we can run the following regression to estimate the ATE, τ:

$$Y_i = \alpha + \tau D_i + X_i'\gamma + \nu_i$$

- Works with continuous or ordinal D_i if effect of these variables is truly linear.

Heterogeneous effects, binary treatment

- Completely randomized experiment:

Heterogeneous effects, binary treatment

- Completely randomized experiment:

$$Y_i = D_i Y_i(1) + (1 - D_i) Y_i(0)$$

Heterogeneous effects, binary treatment

- Completely randomized experiment:

$$\begin{aligned} Y_i &= D_i Y_i(1) + (1 - D_i) Y_i(0) \\ &= Y_i(0) + (Y_i(1) - Y_i(0)) D_i \end{aligned}$$

Heterogeneous effects, binary treatment

- Completely randomized experiment:

$$\begin{aligned} Y_i &= D_i Y_i(1) + (1 - D_i) Y_i(0) \\ &= Y_i(0) + (Y_i(1) - Y_i(0)) D_i \\ &= \mu_0 + \tau_i D_i + (Y_i(0) - \mu_0) \end{aligned}$$

Heterogeneous effects, binary treatment

- Completely randomized experiment:

$$\begin{aligned} Y_i &= D_i Y_i(1) + (1 - D_i) Y_i(0) \\ &= Y_i(0) + (Y_i(1) - Y_i(0)) D_i \\ &= \mu_0 + \tau_i D_i + (Y_i(0) - \mu_0) \\ &= \mu_0 + \tau D_i + (Y_i(0) - \mu_0) + (\tau_i - \tau) \cdot D_i \end{aligned}$$

Heterogeneous effects, binary treatment

- Completely randomized experiment:

$$\begin{aligned}
Y_i &= D_i Y_i(1) + (1 - D_i) Y_i(0) \\
&= Y_i(0) + (Y_i(1) - Y_i(0)) D_i \\
&= \mu_0 + \tau_i D_i + (Y_i(0) - \mu_0) \\
&= \mu_0 + \tau D_i + (Y_i(0) - \mu_0) + (\tau_i - \tau) \cdot D_i \\
&= \mu_0 + \tau D_i + \varepsilon_i
\end{aligned}$$

Heterogeneous effects, binary treatment

- Completely randomized experiment:

$$\begin{aligned} Y_i &= D_i Y_i(1) + (1 - D_i) Y_i(0) \\ &= Y_i(0) + (Y_i(1) - Y_i(0)) D_i \\ &= \mu_0 + \tau_i D_i + (Y_i(0) - \mu_0) \\ &= \mu_0 + \tau D_i + (Y_i(0) - \mu_0) + (\tau_i - \tau) \cdot D_i \\ &= \mu_0 + \tau D_i + \varepsilon_i \end{aligned}$$

- Error term now includes two components:

Heterogeneous effects, binary treatment

- Completely randomized experiment:

$$\begin{aligned} Y_i &= D_i Y_i(1) + (1 - D_i) Y_i(0) \\ &= Y_i(0) + (Y_i(1) - Y_i(0)) D_i \\ &= \mu_0 + \tau_i D_i + (Y_i(0) - \mu_0) \\ &= \mu_0 + \tau D_i + (Y_i(0) - \mu_0) + (\tau_i - \tau) \cdot D_i \\ &= \mu_0 + \tau D_i + \varepsilon_i \end{aligned}$$

- Error term now includes two components:
  1. "Baseline" variation in the outcome: $(Y_i(0) - \mu_0)$

Heterogeneous effects, binary treatment

- Completely randomized experiment:

$$\begin{aligned} Y_i &= D_i Y_i(1) + (1 - D_i) Y_i(0) \\ &= Y_i(0) + (Y_i(1) - Y_i(0)) D_i \\ &= \mu_0 + \tau_i D_i + (Y_i(0) - \mu_0) \\ &= \mu_0 + \tau D_i + (Y_i(0) - \mu_0) + (\tau_i - \tau) \cdot D_i \\ &= \mu_0 + \tau D_i + \varepsilon_i \end{aligned}$$

- Error term now includes two components:
  1. "Baseline" variation in the outcome: $(Y_i(0) - \mu_0)$
  2. Variation in the treatment effect, $(\tau_i - \tau)$

Heterogeneous effects, binary treatment

- Completely randomized experiment:

$$\begin{aligned} Y_i &= D_i Y_i(1) + (1 - D_i) Y_i(0) \\ &= Y_i(0) + (Y_i(1) - Y_i(0)) D_i \\ &= \mu_0 + \tau_i D_i + (Y_i(0) - \mu_0) \\ &= \mu_0 + \tau D_i + (Y_i(0) - \mu_0) + (\tau_i - \tau) \cdot D_i \\ &= \mu_0 + \tau D_i + \varepsilon_i \end{aligned}$$

- Error term now includes two components:
  1. "Baseline" variation in the outcome: $(Y_i(0) - \mu_0)$
  2. Variation in the treatment effect, $(\tau_i - \tau)$
- We can verify that under experiment, $\mathbb{E}[\varepsilon_i | D_i] = 0$

Heterogeneous effects, binary treatment

- Completely randomized experiment:

$$\begin{aligned} Y_i &= D_i Y_i(1) + (1 - D_i) Y_i(0) \\ &= Y_i(0) + (Y_i(1) - Y_i(0)) D_i \\ &= \mu_0 + \tau_i D_i + (Y_i(0) - \mu_0) \\ &= \mu_0 + \tau D_i + (Y_i(0) - \mu_0) + (\tau_i - \tau) \cdot D_i \\ &= \mu_0 + \tau D_i + \varepsilon_i \end{aligned}$$

- Error term now includes two components:
  1. "Baseline" variation in the outcome: $(Y_i(0) - \mu_0)$
  2. Variation in the treatment effect, $(\tau_i - \tau)$
- We can verify that under experiment, $\mathbb{E}[\varepsilon_i | D_i] = 0$
- Thus, OLS estimates the ATE with no covariates.

Adding covariates

- What happens with no unmeasured confounders? Need to condition on X_i now.

Adding covariates

- What happens with no unmeasured confounders? Need to condition on X_i now.
- Remember identification of the ATE/ATT using iterated expectations.

Adding covariates

- What happens with no unmeasured confounders? Need to condition on X_i now.
- Remember identification of the ATE/ATT using iterated expectations.
- ATE is the weighted sum of Conditional Average Treatment Effects (CATEs):

$$\tau = \sum_x \tau(x) \Pr[X_i = x]$$

Adding covariates

- What happens with no unmeasured confounders? Need to condition on X_i now.
- Remember identification of the ATE/ATT using iterated expectations.
- ATE is the weighted sum of Conditional Average Treatment Effects (CATEs):

$$\tau = \sum_x \tau(x) \Pr[X_i = x]$$

- ATE/ATT are weighted averages of CATEs.

Adding covariates

- What happens with no unmeasured confounders? Need to condition on X_i now.
- Remember identification of the ATE/ATT using iterated expectations.
- ATE is the weighted sum of Conditional Average Treatment Effects (CATEs):

$$\tau = \sum_x \tau(x) \Pr[X_i = x]$$

- ATE/ATT are weighted averages of CATEs.
- What about the regression estimand, τ_R? How does it relate to the ATE/ATT?

Heterogeneous effects and regression

- Let's investigate this under a saturated regression model:

$$Y_i = \sum_x B_{xi}\alpha_x + \tau_R D_i + e_i.$$

Heterogeneous effects and regression

- Let's investigate this under a saturated regression model:

$$Y_i = \sum_x B_{xi}\alpha_x + \tau_R D_i + e_i.$$

- Use a dummy variable for each unique combination of X_i: $B_{xi} = \mathbb{I}(X_i = x)$

Heterogeneous effects and regression

- Let's investigate this under a saturated regression model:

$$Y_i = \sum_x B_{xi}\alpha_x + \tau_R D_i + e_i.$$

- Use a dummy variable for each unique combination of X_i: $B_{xi} = \mathbb{I}(X_i = x)$
- Linear in X_i by construction!

Investigating the regression coefficient

- How can we investigate τ_R? Well, we can rely on the regression anatomy:
$$\tau_R = \frac{\text{Cov}(Y_i, D_i - E[D_i|X_i])}{\mathrm{Var}(D_i - E[D_i|X_i])}$$

Investigating the regression coefficient

- How can we investigate τ_R? Well, we can rely on the regression anatomy:

$$\tau_R = \frac{\text{Cov}(Y_i, D_i - E[D_i|X_i])}{\mathrm{Var}(D_i - E[D_i|X_i])}$$

- $D_i - \mathbb{E}[D_i|X_i]$ is the residual from a regression of D_i on the full set of dummies.

Investigating the regression coefficient

- How can we investigate τ_R? Well, we can rely on the regression anatomy:
$$\tau_R = \frac{\text{Cov}(Y_i, D_i - E[D_i|X_i])}{\mathrm{Var}(D_i - E[D_i|X_i])}$$
- $D_i - \mathbb{E}[D_i|X_i]$ is the residual from a regression of D_i on the full set of dummies.
- With a little work we can show:

$$\tau_R = \frac{\mathbb{E}\left[\tau(X_i)(D_i - \mathbb{E}[D_i|X_i])^2\right]}{\mathbb{E}[(D_i - E[D_i|X_i])^2]} = \frac{\mathbb{E}[\tau(X_i)\sigma_d^2(X_i)]}{\mathbb{E}[\sigma_d^2(X_i)]}$$

Investigating the regression coefficient

- How can we investigate τ_R? Well, we can rely on the regression anatomy:

$$\tau_R = \frac{\text{Cov}(Y_i, D_i - E[D_i|X_i])}{\text{Var}(D_i - E[D_i|X_i])}$$

- $D_i - \mathbb{E}[D_i|X_i]$ is the residual from a regression of D_i on the full set of dummies.
- With a little work we can show:

$$\tau_R = \frac{\mathbb{E}\left[\tau(X_i)(D_i - \mathbb{E}[D_i|X_i])^2\right]}{\mathbb{E}[(D_i - E[D_i|X_i])^2]} = \frac{\mathbb{E}[\tau(X_i)\sigma_d^2(X_i)]}{\mathbb{E}[\sigma_d^2(X_i)]}$$

- $\sigma_d^2(x) = \text{Var}[D_i|X_i = x]$ is the conditional variance of treatment assignment.

ATE versus OLS

$$\tau_R = \mathbb{E}[\tau(X_i)W_i] = \sum_x \tau(x)\frac{\sigma_d^2(x)}{\mathbb{E}[\sigma_d^2(X_i)]}\mathbb{P}[X_i = x]$$

ATE versus OLS

$$\tau_R = \mathbb{E}[\tau(X_i)W_i] = \sum_x \tau(x)\frac{\sigma_d^2(x)}{\mathbb{E}[\sigma_d^2(X_i)]}\mathbb{P}[X_i = x]$$

- Compare to the ATE:

$$\tau = \mathbb{E}[\tau(X_i)] = \sum_x \tau(x)\mathbb{P}[X_i = x]$$

ATE versus OLS

$$\tau_R = \mathbb{E}[\tau(X_i)W_i] = \sum_x \tau(x)\frac{\sigma_d^2(x)}{\mathbb{E}[\sigma_d^2(X_i)]}\mathbb{P}[X_i = x]$$

- Compare to the ATE:

$$\tau = \mathbb{E}[\tau(X_i)] = \sum_x \tau(x)\mathbb{P}[X_i = x]$$

- Both weight strata relative to their size ($\mathbb{P}[X_i = x]$)

ATE versus OLS

$$\tau_R = \mathbb{E}[\tau(X_i)W_i] = \sum_x \tau(x)\frac{\sigma_d^2(x)}{\mathbb{E}[\sigma_d^2(X_i)]}\mathbb{P}[X_i = x]$$

- Compare to the ATE:

$$\tau = \mathbb{E}[\tau(X_i)] = \sum_x \tau(x)\mathbb{P}[X_i = x]$$

- Both weight strata relative to their size ($\mathbb{P}[X_i = x]$)
- OLS weights strata higher if the treatment variance in those strata ($\sigma_d^2(x)$) is higher in those strata relative to the average variance across strata ($\mathbb{E}[\sigma_d^2(X_i)]$).

ATE versus OLS

$$\tau_R = \mathbb{E}[\tau(X_i)W_i] = \sum_x \tau(x)\frac{\sigma_d^2(x)}{\mathbb{E}[\sigma_d^2(X_i)]}\mathbb{P}[X_i = x]$$

- Compare to the ATE:

$$\tau = \mathbb{E}[\tau(X_i)] = \sum_x \tau(x)\mathbb{P}[X_i = x]$$

- Both weight strata relative to their size ($\mathbb{P}[X_i = x]$)
- OLS weights strata higher if the treatment variance in those strata ($\sigma_d^2(x)$) is higher in those strata relative to the average variance across strata ($\mathbb{E}[\sigma_d^2(X_i)]$).
- The ATE weights only by their size.

Regression weighting

$$W_i = \frac{\sigma_d^2(X_i)}{\mathbb{E}[\sigma_d^2(X_i)]}$$

- Why does OLS weight like this?

Regression weighting

$$W_i = \frac{\sigma_d^2(X_i)}{\mathbb{E}[\sigma_d^2(X_i)]}$$

- Why does OLS weight like this?
- OLS is a minimum-variance estimator $\rightsquigarrow$ more weight to more precise within-strata estimates.

Regression weighting

$$W_i = \frac{\sigma_d^2(X_i)}{\mathbb{E}[\sigma_d^2(X_i)]}$$

- Why does OLS weight like this?
- OLS is a minimum-variance estimator $\rightsquigarrow$ more weight to more precise within-strata estimates.
- Within-strata estimates are most precise when the treatment is evenly spread and thus has the highest variance.

Regression weighting

$$W_i = \frac{\sigma_d^2(X_i)}{\mathbb{E}[\sigma_d^2(X_i)]}$$

- Why does OLS weight like this?
- OLS is a minimum-variance estimator $\rightsquigarrow$ more weight to more precise within-strata estimates.
- Within-strata estimates are most precise when the treatment is evenly spread and thus has the highest variance.
- If D_i is binary, then we know the conditional variance will be:

Regression weighting

$$W_i = \frac{\sigma_d^2(X_i)}{\mathbb{E}[\sigma_d^2(X_i)]}$$

- Why does OLS weight like this?
- OLS is a minimum-variance estimator $\rightsquigarrow$ more weight to more precise within-strata estimates.
- Within-strata estimates are most precise when the treatment is evenly spread and thus has the highest variance.
- If D_i is binary, then we know the conditional variance will be:

$$\sigma_d^2(x) = \mathbb{P}[D_i = 1 | X_i = x] \left(1 - \mathbb{P}[D_i = 1 | X_i = x]\right)$$

Regression weighting

$$W_i = \frac{\sigma_d^2(X_i)}{\mathbb{E}[\sigma_d^2(X_i)]}$$

- Why does OLS weight like this?
- OLS is a minimum-variance estimator $\rightsquigarrow$ more weight to more precise within-strata estimates.
- Within-strata estimates are most precise when the treatment is evenly spread and thus has the highest variance.
- If D_i is binary, then we know the conditional variance will be:

$$\sigma_d^2(x) = \mathbb{P}[D_i = 1|X_i = x]\,(1 - \mathbb{P}[D_i = 1|X_i = x])$$

- Maximum variance with $\mathbb{P}[D_i = 1|X_i = x] = 1/2$.

OLS weighting example

- Binary covariate:

OLS weighting example

- Binary covariate:

Group 1	Group 2
$\mathbb{P}[X_i = 1] = 0.75$	$\mathbb{P}[X_i = 0] = 0.25$

OLS weighting example

- Binary covariate:

Group 1	Group 2
$\mathbb{P}[X_i = 1] = 0.75$	$\mathbb{P}[X_i = 0] = 0.25$
$\mathbb{P}[D_i = 1 \| X_i = 1] = 0.9$	$\mathbb{P}[D_i = 1 \| X_i = 0] = 0.5$

OLS weighting example

- Binary covariate:

Group 1	Group 2
$\mathbb{P}[X_i = 1] = 0.75$	$\mathbb{P}[X_i = 0] = 0.25$
$\mathbb{P}[D_i = 1 \mid X_i = 1] = 0.9$	$\mathbb{P}[D_i = 1 \mid X_i = 0] = 0.5$
$\sigma^2_d(1) = 0.09$	$\sigma^2_d(0) = 0.25$

OLS weighting example

- Binary covariate:

Group 1	Group 2
$\mathbb{P}[X_i = 1] = 0.75$	$\mathbb{P}[X_i = 0] = 0.25$
$\mathbb{P}[D_i = 1 \vert X_i = 1] = 0.9$	$\mathbb{P}[D_i = 1 \vert X_i = 0] = 0.5$
$\sigma^2_d(1) = 0.09$	$\sigma^2_d(0) = 0.25$
$\tau(1) = 1$	$\tau(0) = -1$

OLS weighting example

- Binary covariate:

Group 1	Group 2
$\mathbb{P}[X_i = 1] = 0.75$	$\mathbb{P}[X_i = 0] = 0.25$
$\mathbb{P}[D_i = 1 \vert X_i = 1] = 0.9$	$\mathbb{P}[D_i = 1 \vert X_i = 0] = 0.5$
$\sigma_d^2(1) = 0.09$	$\sigma_d^2(0) = 0.25$
$\tau(1) = 1$	$\tau(0) = -1$

- Implies the ATE is $\tau = 0.5$

OLS weighting example

- Binary covariate:

Group 1	Group 2
$\mathbb{P}[X_i = 1] = 0.75$	$\mathbb{P}[X_i = 0] = 0.25$
$\mathbb{P}[D_i = 1 \vert X_i = 1] = 0.9$	$\mathbb{P}[D_i = 1 \vert X_i = 0] = 0.5$
$\sigma_d^2(1) = 0.09$	$\sigma_d^2(0) = 0.25$
$\tau(1) = 1$	$\tau(0) = -1$

- Implies the ATE is $\tau = 0.5$
- Average conditional variance: $\mathbb{E}[\sigma_d^2(X_i)] = 0.13$

OLS weighting example

- Binary covariate:

Group 1	Group 2
$\mathbb{P}[X_i = 1] = 0.75$	$\mathbb{P}[X_i = 0] = 0.25$
$\mathbb{P}[D_i = 1 \mid X_i = 1] = 0.9$	$\mathbb{P}[D_i = 1 \mid X_i = 0] = 0.5$
$\sigma_d^2(1) = 0.09$	$\sigma_d^2(0) = 0.25$
$\tau(1) = 1$	$\tau(0) = -1$

- Implies the ATE is $\tau = 0.5$
- Average conditional variance: $\mathbb{E}[\sigma_d^2(X_i)] = 0.13$
- $\rightsquigarrow$ weights for $X_i = 1$ are: $0.09/0.13 = 0.692$, for $X_i = 0$: $0.25/0.13 = 1.92$.

OLS weighting example

- Binary covariate:

Group 1	Group 2
$\mathbb{P}[X_i = 1] = 0.75$	$\mathbb{P}[X_i = 0] = 0.25$
$\mathbb{P}[D_i = 1 \mid X_i = 1] = 0.9$	$\mathbb{P}[D_i = 1 \mid X_i = 0] = 0.5$
$\sigma_d^2(1) = 0.09$	$\sigma_d^2(0) = 0.25$
$\tau(1) = 1$	$\tau(0) = -1$

- Implies the ATE is $\tau = 0.5$
- Average conditional variance: $\mathbb{E}[\sigma_d^2(X_i)] = 0.13$
- $\rightsquigarrow$ weights for $X_i = 1$ are: $0.09/0.13 = 0.692$, for $X_i = 0$: $0.25/0.13 = 1.92$.

$$\tau_R = \mathbb{E}[\tau(X_i)W_i]$$

OLS weighting example

- Binary covariate:

Group 1	Group 2
$\mathbb{P}[X_i = 1] = 0.75$	$\mathbb{P}[X_i = 0] = 0.25$
$\mathbb{P}[D_i = 1 \mid X_i = 1] = 0.9$	$\mathbb{P}[D_i = 1 \mid X_i = 0] = 0.5$
$\sigma_d^2(1) = 0.09$	$\sigma_d^2(0) = 0.25$
$\tau(1) = 1$	$\tau(0) = -1$

- Implies the ATE is $\tau = 0.5$
- Average conditional variance: $\mathbb{E}[\sigma_d^2(X_i)] = 0.13$
- $\rightsquigarrow$ weights for $X_i = 1$ are: $0.09/0.13 = 0.692$, for $X_i = 0$: $0.25/0.13 = 1.92$.

$$\begin{aligned}\tau_R &= \mathbb{E}[\tau(X_i)W_i] \\ &= \tau(1)W(1)\mathbb{P}[X_i = 1] + \tau(0)W(0)\mathbb{P}[X_i = 0]\end{aligned}$$

OLS weighting example

- Binary covariate:

	Group 1	Group 2
	$\mathbb{P}[X_i = 1] = 0.75$	$\mathbb{P}[X_i = 0] = 0.25$
	$\mathbb{P}[D_i = 1\vert X_i = 1] = 0.9$	$\mathbb{P}[D_i = 1\vert X_i = 0] = 0.5$
	$\sigma^2_d(1) = 0.09$	$\sigma^2_d(0) = 0.25$
	$\tau(1) = 1$	$\tau(0) = -1$

- Implies the ATE is $\tau = 0.5$
- Average conditional variance: $\mathbb{E}[\sigma^2_d(X_i)] = 0.13$
- $\rightsquigarrow$ weights for $X_i = 1$ are: $0.09/0.13 = 0.692$, for $X_i = 0$: $0.25/0.13 = 1.92$.

$$\begin{aligned}\tau_R &= \mathbb{E}[\tau(X_i)W_i] \\ &= \tau(1)W(1)\mathbb{P}[X_i = 1] + \tau(0)W(0)\mathbb{P}[X_i = 0] \\ &= 1 \times 0.692 \times 0.75 + -1 \times 1.92 \times 0.25\end{aligned}$$

OLS weighting example

- Binary covariate:

Group 1	Group 2
$\mathbb{P}[X_i = 1] = 0.75$	$\mathbb{P}[X_i = 0] = 0.25$
$\mathbb{P}[D_i = 1 \vert X_i = 1] = 0.9$	$\mathbb{P}[D_i = 1 \vert X_i = 0] = 0.5$
$\sigma_d^2(1) = 0.09$	$\sigma_d^2(0) = 0.25$
$\tau(1) = 1$	$\tau(0) = -1$

- Implies the ATE is $\tau = 0.5$
- Average conditional variance: $\mathbb{E}[\sigma_d^2(X_i)] = 0.13$
- $\rightsquigarrow$ weights for $X_i = 1$ are: $0.09/0.13 = 0.692$, for $X_i = 0$: $0.25/0.13 = 1.92$.

$$
\begin{aligned}
\tau_R &= \mathbb{E}[\tau(X_i)W_i] \\
&= \tau(1)W(1)\mathbb{P}[X_i = 1] + \tau(0)W(0)\mathbb{P}[X_i = 0] \\
&= 1 \times 0.692 \times 0.75 + -1 \times 1.92 \times 0.25 \\
&= 0.039
\end{aligned}
$$

When will OLS estimate the ATE?

- When does $\tau = \tau_R$?

When will OLS estimate the ATE?

- When does $\tau = \tau_R$?
- Constant treatment effects: $\tau(x) = \tau = \tau_R$

When will OLS estimate the ATE?

- When does $\tau = \tau_R$?
- Constant treatment effects: $\tau(x) = \tau = \tau_R$
- Constant probability of treatment: $e(x) = \mathbb{P}[D_i = 1 | X_i = x] = e$.

When will OLS estimate the ATE?

- When does $\tau = \tau_R$?
- Constant treatment effects: $\tau(x) = \tau = \tau_R$
- Constant probability of treatment: $e(x) = \mathbb{P}[D_i = 1 | X_i = x] = e$.
  - Implies that the OLS weights are 1.

When will OLS estimate the ATE?

- When does $\tau = \tau_R$?
- Constant treatment effects: $\tau(x) = \tau = \tau_R$
- Constant probability of treatment: $e(x) = \mathbb{P}[D_i = 1 | X_i = x] = e$.
  - Implies that the OLS weights are 1.
- Incorrect linearity assumption in X_i will lead to more bias.

Other ways to use regression

- What's the path forward?

Other ways to use regression

- What's the path forward?
  - Accept the bias (might be relatively small with saturated models)

Other ways to use regression

- What's the path forward?
  - Accept the bias (might be relatively small with saturated models)
  - Use a different regression approach

Other ways to use regression

- What's the path forward?
  - Accept the bias (might be relatively small with saturated models)
  - Use a different regression approach
- Let $\mu_d(x) = \mathbb{E}[Y_i(d)|X_i = x]$ be the CEF for the potential outcome under $D_i = d$.

Other ways to use regression

- What's the path forward?
  - Accept the bias (might be relatively small with saturated models)
  - Use a different regression approach
- Let $\mu_d(x) = \mathbb{E}[Y_i(d)|X_i = x]$ be the CEF for the potential outcome under $D_i = d$.
- By consistency and n.u.c., we have $\mu_d(x) = \mathbb{E}[Y_i|D_i = d, X_i = x]$.

Other ways to use regression

- What's the path forward?
  - Accept the bias (might be relatively small with saturated models)
  - Use a different regression approach
- Let $\mu_d(x) = \mathbb{E}[Y_i(d)|X_i = x]$ be the CEF for the potential outcome under $D_i = d$.
- By consistency and n.u.c., we have $\mu_d(x) = \mathbb{E}[Y_i|D_i = d, X_i = x]$.
- Estimate a regression of Y_i on X_i among the $D_i = d$ group.

Other ways to use regression

- What's the path forward?
  - Accept the bias (might be relatively small with saturated models)
  - Use a different regression approach
- Let $\mu_d(x) = \mathbb{E}[Y_i(d)|X_i = x]$ be the CEF for the potential outcome under $D_i = d$.
- By consistency and n.u.c., we have $\mu_d(x) = \mathbb{E}[Y_i|D_i = d, X_i = x]$.
- Estimate a regression of Y_i on X_i among the $D_i = d$ group.
- Then, $\hat{\mu}_d(x)$ is just a predicted value from the regression for $X_i = x$.

Other ways to use regression

- What's the path forward?
  - Accept the bias (might be relatively small with saturated models)
  - Use a different regression approach
- Let $\mu_d(x) = \mathbb{E}[Y_i(d)|X_i = x]$ be the CEF for the potential outcome under $D_i = d$.
- By consistency and n.u.c., we have $\mu_d(x) = \mathbb{E}[Y_i|D_i = d, X_i = x]$.
- Estimate a regression of Y_i on X_i among the $D_i = d$ group.
- Then, $\hat{\mu}_d(x)$ is just a predicted value from the regression for $X_i = x$.
- How can we use this?

Imputation estimators

- Impute the treated potential outcomes with $\widehat{Y}_i(1) = \hat{\mu}_1(X_i)$!

Imputation estimators

- Impute the treated potential outcomes with $\widehat{Y}_i(1) = \hat{\mu}_1(X_i)$!
- Impute the control potential outcomes with $\widehat{Y}_i(0) = \hat{\mu}_0(X_i)$!

Imputation estimators

- Impute the treated potential outcomes with $\widehat{Y}_i(1) = \hat{\mu}_1(X_i)$!
- Impute the control potential outcomes with $\widehat{Y}_i(0) = \hat{\mu}_0(X_i)$!
- Procedure:

Imputation estimators

- Impute the treated potential outcomes with $\widehat{Y}_i(1) = \hat{\mu}_1(X_i)$!
- Impute the control potential outcomes with $\widehat{Y}_i(0) = \hat{\mu}_0(X_i)$!
- Procedure:
  - Regress Y_i on X_i in the treated group and get predicted values for all units (treated or control).

Imputation estimators

- Impute the treated potential outcomes with $\widehat{Y}_i(1) = \hat{\mu}_1(X_i)$!
- Impute the control potential outcomes with $\widehat{Y}_i(0) = \hat{\mu}_0(X_i)$!
- Procedure:
  - Regress Y_i on X_i in the treated group and get predicted values for all units (treated or control).
  - Regress Y_i on X_i in the control group and get predicted values for all units (treated or control).

Imputation estimators

- Impute the treated potential outcomes with $\widehat{Y}_i(1) = \hat{\mu}_1(X_i)$!
- Impute the control potential outcomes with $\widehat{Y}_i(0) = \hat{\mu}_0(X_i)$!
- Procedure:
  - Regress Y_i on X_i in the treated group and get predicted values for all units (treated or control).
  - Regress Y_i on X_i in the control group and get predicted values for all units (treated or control).
  - Take the average difference between these predicted values.

Imputation estimators

- Impute the treated potential outcomes with $\widehat{Y}_i(1) = \hat{\mu}_1(X_i)$!
- Impute the control potential outcomes with $\widehat{Y}_i(0) = \hat{\mu}_0(X_i)$!
- Procedure:
    - Regress Y_i on X_i in the treated group and get predicted values for all units (treated or control).
    - Regress Y_i on X_i in the control group and get predicted values for all units (treated or control).
    - Take the average difference between these predicted values.
- More mathematically, look like this:

$$\tau_{imp} = \frac{1}{N} \sum_i \hat{\mu}_1(X_i) - \hat{\mu}_0(X_i)$$

Imputation estimators

- Impute the treated potential outcomes with $\widehat{Y}_i(1) = \hat{\mu}_1(X_i)$!
- Impute the control potential outcomes with $\widehat{Y}_i(0) = \hat{\mu}_0(X_i)$!
- Procedure:
  - Regress Y_i on X_i in the treated group and get predicted values for all units (treated or control).
  - Regress Y_i on X_i in the control group and get predicted values for all units (treated or control).
  - Take the average difference between these predicted values.
- More mathematically, look like this:

$$\tau_{imp} = \frac{1}{N} \sum_i \hat{\mu}_1(X_i) - \hat{\mu}_0(X_i)$$

- Sometimes called an **imputation estimator**.

Simple imputation estimator

- Use `predict()` from the within-group models on the data from the entire sample.

Simple imputation estimator

- Use `predict()` from the within-group models on the data from the entire sample.
- Useful trick: use a model on the entire data and `model.frame()` to get the right design matrix:

Simple imputation estimator

- Use `predict()` from the within-group models on the data from the entire sample.
- Useful trick: use a model on the entire data and `model.frame()` to get the right design matrix:

```
## heterogeneous effects
y.het <- ifelse(d == 1, y + rnorm(n, 0, 5), y)

mod <- lm(y.het ~ d + X)
mod1 <- lm(y.het ~ X, subset = d == 1)
mod0 <- lm(y.het ~ X, subset = d == 0)
y1.imps <- predict(mod1, model.frame(mod))
y0.imps <- predict(mod0, model.frame(mod))
mean(y1.imps - y0.imps)
```

```
## [1] 0.61
```

Notes on imputation estimators

- If $\hat{\mu}_d(x)$ are consistent estimators, then τ_{imp} is consistent for the ATE.

Notes on imputation estimators

- If $\hat{\mu}_d(x)$ are consistent estimators, then τ_{imp} is consistent for the ATE.
- Why don't people use this?

Notes on imputation estimators

- If $\hat{\mu}_d(x)$ are consistent estimators, then τ_{imp} is consistent for the ATE.
- Why don't people use this?
  - Most people don't know the results we've been talking about.

Notes on imputation estimators

- If $\hat{\mu}_d(x)$ are consistent estimators, then τ_{imp} is consistent for the ATE.
- Why don't people use this?
  - Most people don't know the results we've been talking about.
  - Harder to implement than vanilla OLS.

Notes on imputation estimators

- If $\hat{\mu}_d(x)$ are consistent estimators, then τ_{imp} is consistent for the ATE.
- Why don't people use this?
  - Most people don't know the results we've been talking about.
  - Harder to implement than vanilla OLS.
- Can use linear regression to estimate $\hat{\mu}_d(x) = x'\beta_d$

Notes on imputation estimators

- If $\hat{\mu}_d(x)$ are consistent estimators, then τ_{imp} is consistent for the ATE.
- Why don't people use this?
  - Most people don't know the results we've been talking about.
  - Harder to implement than vanilla OLS.
- Can use linear regression to estimate $\hat{\mu}_d(x) = x'\beta_d$
- Recent trend is to estimate $\hat{\mu}_d(x)$ via non-parametric methods such as:

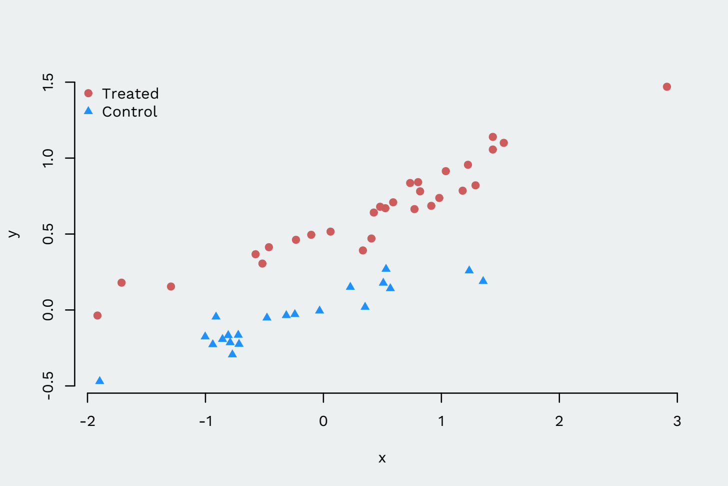
Notes on imputation estimators

- If $\hat{\mu}_d(x)$ are consistent estimators, then τ_{imp} is consistent for the ATE.
- Why don't people use this?
  - Most people don't know the results we've been talking about.
  - Harder to implement than vanilla OLS.
- Can use linear regression to estimate $\hat{\mu}_d(x) = x'\beta_d$
- Recent trend is to estimate $\hat{\mu}_d(x)$ via non-parametric methods such as:
  - Kernel regression, local linear regression, regression trees, etc

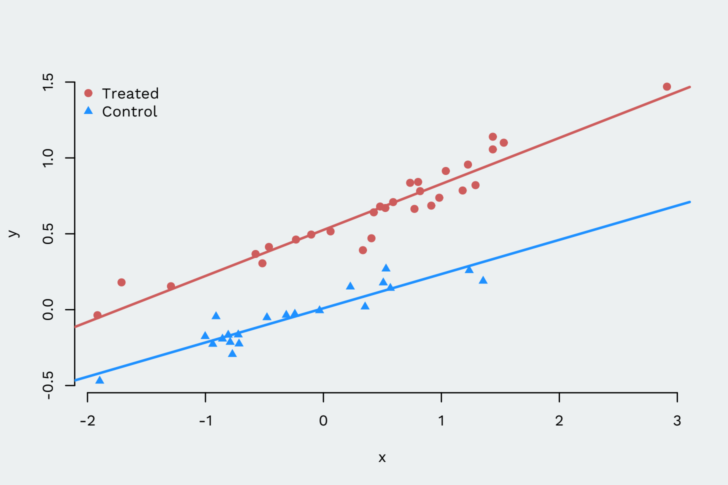
Notes on imputation estimators

- If $\hat{\mu}_d(x)$ are consistent estimators, then τ_{imp} is consistent for the ATE.
- Why don't people use this?
  - Most people don't know the results we've been talking about.
  - Harder to implement than vanilla OLS.
- Can use linear regression to estimate $\hat{\mu}_d(x) = x'\beta_d$
- Recent trend is to estimate $\hat{\mu}_d(x)$ via non-parametric methods such as:
  - Kernel regression, local linear regression, regression trees, etc
  - Easiest is generalized additive models (GAMs)

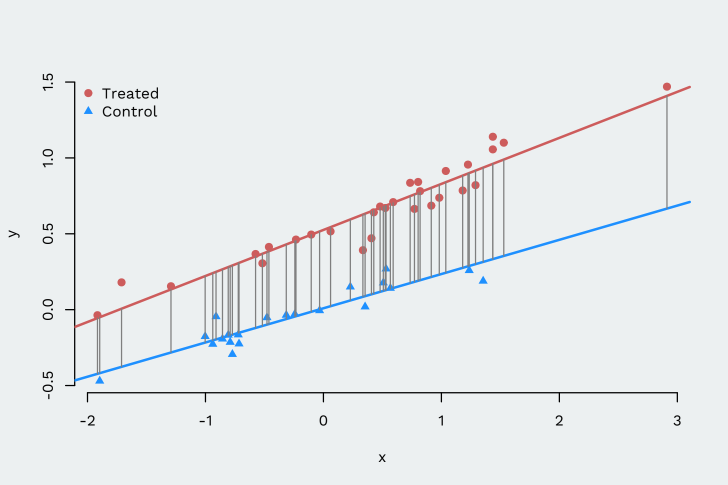
Imputation estimator visualization



Imputation estimator visualization

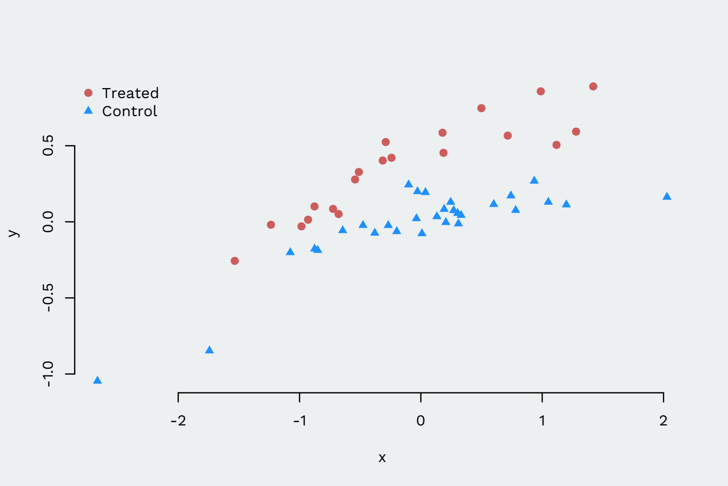


Imputation estimator visualization



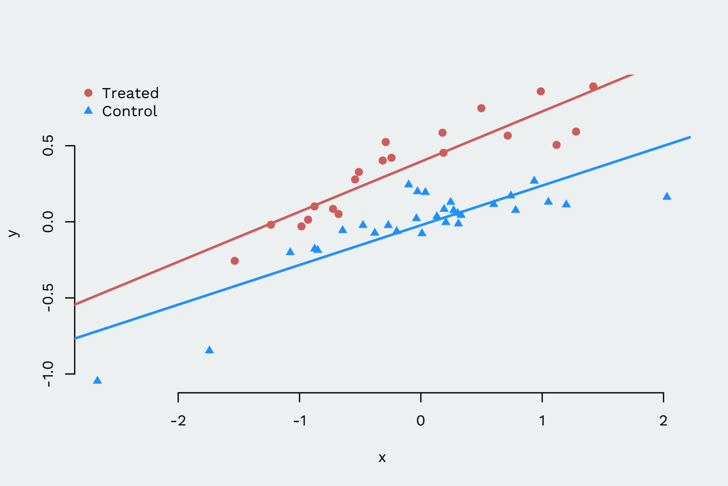
Nonlinear relationships

- Same idea but with nonlinear relationship between Y_i and X_i:



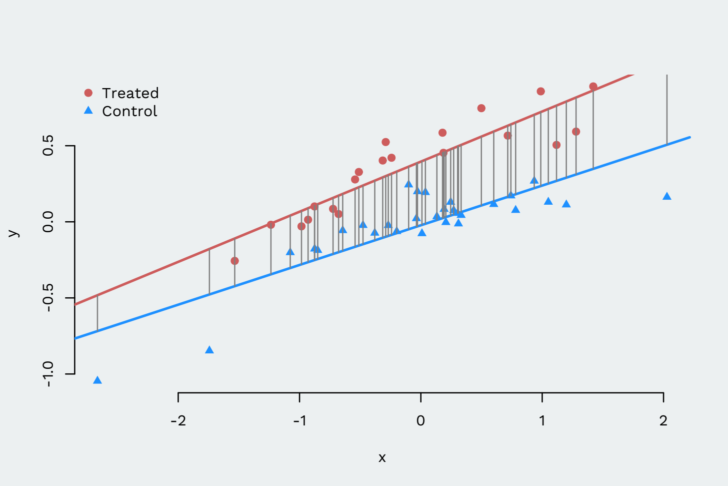
Nonlinear relationships

- Same idea but with nonlinear relationship between Y_i and X_i:



Nonlinear relationships

- Same idea but with nonlinear relationship between Y_i and X_i:



Using semiparametric regression

- Here, CEFs are nonlinear, but we don't know their form.

```r
library(mgcv)
mod0 <- gam(y ~ s(x), subset = d == 0)
summary(mod0)
```

```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## y ~ s(x)
##
## Parametric coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -0.0225     0.0154   -1.46     0.16
##
## Approximate significance of smooth terms:
##        edf Ref.df    F p-value
## s(x) 6.03   7.08 41.3  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

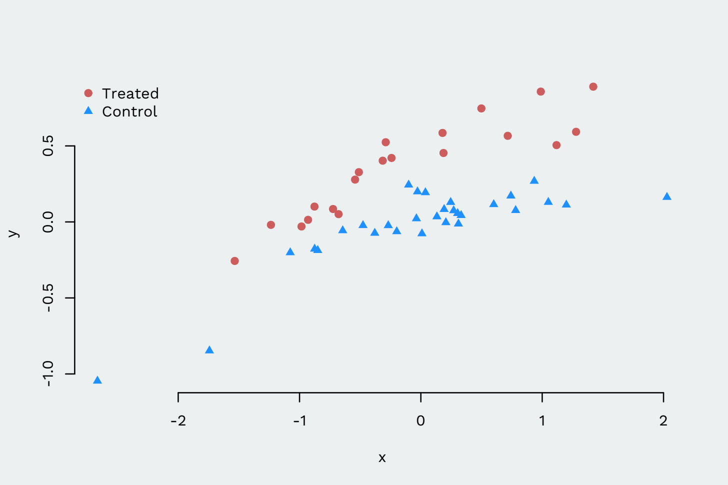
Using semiparametric regression

- Here, CEFs are nonlinear, but we don't know their form.
- We can use GAMs from the `mgcv` package to for flexible estimate:

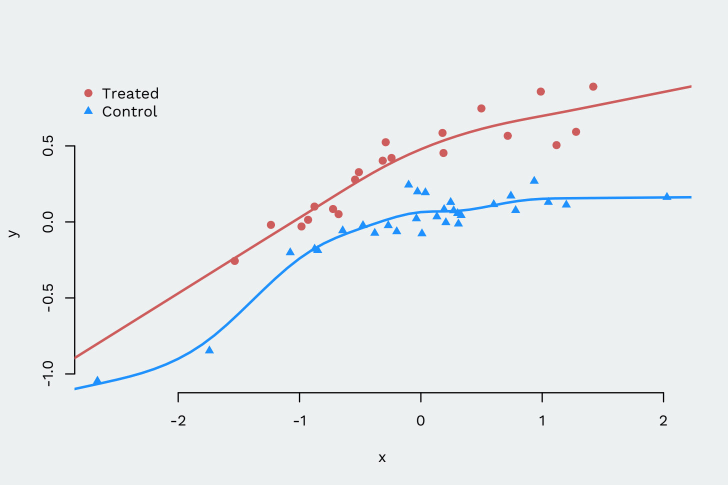
```
library(mgcv)
mod0 <- gam(y ~ s(x), subset = d == 0)
summary(mod0)
```

```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## y ~ s(x)
##
## Parametric coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -0.0225     0.0154   -1.46     0.16
##
## Approximate significance of smooth terms:
##        edf Ref.df    F p-value
## s(x) 6.03   7.08 41.3  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

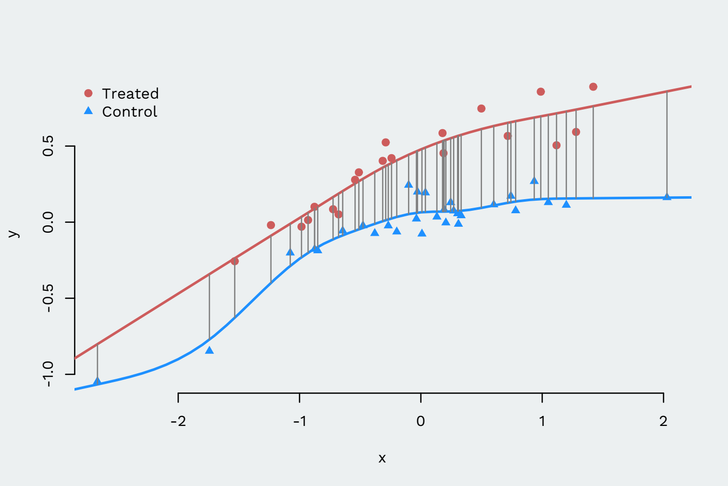
Using GAMs



Using GAMs



Using GAMs



‘Wait...so what are we actually doing most of the time?’

A Discussion

Conclusions

- Regression is mechanically very simple, but philosophically somewhat complicated

Conclusions

- Regression is mechanically very simple, but philosophically somewhat complicated
- It is a useful descriptive tool for approximating a conditional expectation function

Conclusions

- Regression is mechanically very simple, but philosophically somewhat complicated
- It is a useful descriptive tool for approximating a conditional expectation function
- Once again though, the estimand of interest isn't necessarily the regression coefficient.

Conclusions

- Regression is mechanically very simple, but philosophically somewhat complicated
- It is a useful descriptive tool for approximating a conditional expectation function
- Once again though, the estimand of interest isn't necessarily the regression coefficient.
- There are many other approaches to estimation, but identification is key.

Next Week

- Causality with Unmeasured Confounding
- Reading:
  - Angrist and Pishke Chapter 4 Instrumental Variables and Chapter 6 on Regression Discontinuity Designs
  - Morgan and Winship Chapter 9 Instrumental Variable Estimators of Causal Effects
  - Optional: Hernan and Robins Chapter 16 Instrumental Variable Estimation

Visualization in the New York Times

AMERICAS

How Stable Are Democracies? ‘Warning Signs Are Flashing

The Interpreter

By AMANDA TAUB NOV. 29, 2016

WASHINGTON — Yascha Mounk is used to being the most pessimistic person in the room. Mr. Mounk, a lecturer in government at Harvard, has spent the past few years challenging one of the bedrock assumptions of Western politics: that once a country becomes a liberal democracy, it will stay that way.

His research suggests something quite different: that liberal democracies around the world may be at serious risk of decline.

Mr. Mounk’s interest in the topic began rather unusually. In 2014, he published a book, “Stranger in My Own Country.” It started as a memoir of his experiences growing up as a Jew in Germany, but became a broader investigation of how contemporary European nations were struggling to construct new, multicultural national identities.

Visualization in the New York Times

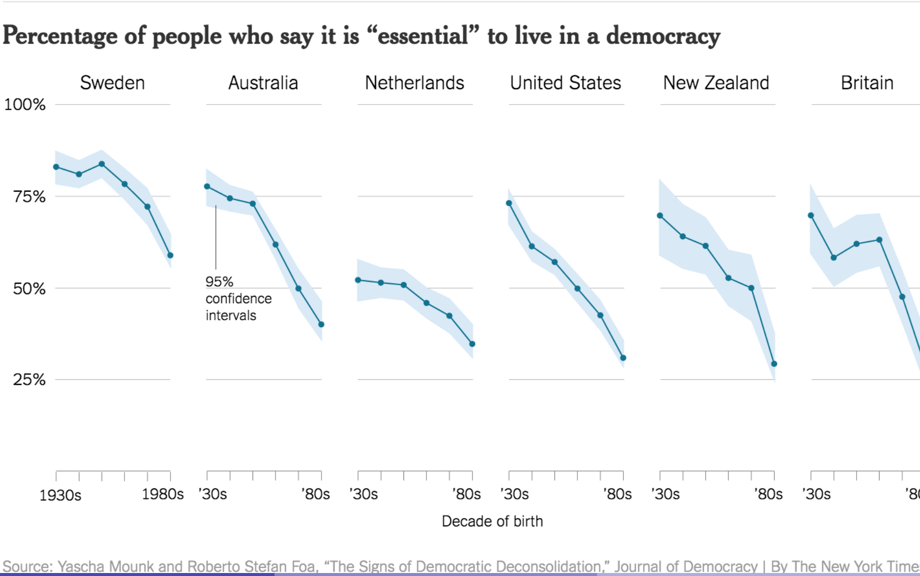
The Danger of Deconsolidation

THE DEMOCRATIC DISCONNECT

Roberto Stefan Foa and Yascha Mounk

***Roberto Stefan Foa** is a principal investigator of the World Values Survey and fellow of the Laboratory for Comparative Social Research. His writing has appeared in a wide range of journals, books, and publications by the UN, OECD, and World Bank. **Yascha Mounk** is a lecturer on political theory in Harvard University's Government Department and a Carnegie Fellow at New America, a Washington, D.C.–based think tank. His dissertation on the role of personal responsibility in contemporary politics and philosophy will be published by Harvard University Press, and his essays have appeared in* Foreign Affairs, *the* New York Times, *and the* Wall Street Journal.

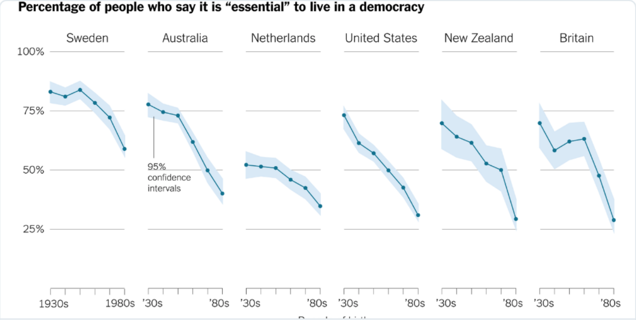
Visualization in the New York Times



Visualization in the New York Times

Ryan D. Enos @RyanDEnos · 19h

Lots of worried chatter a/b @amandataub article on work of @Yascha_Mounk. Important, but want to raise cautions 1/

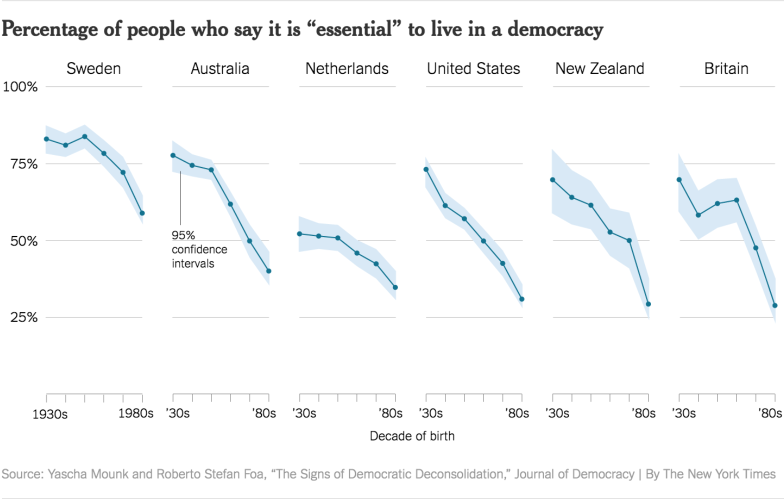


How Stable Are Democracies? 'Warning Signs Are Flashing Red'

New research tries to spot the collapse of liberal democracies before they happen, and it suggests that Western democracy may be seriously ill.

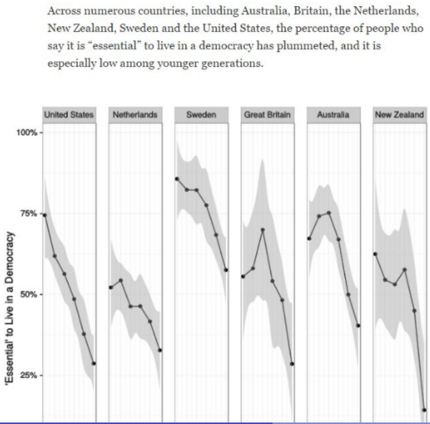
nytimes.com

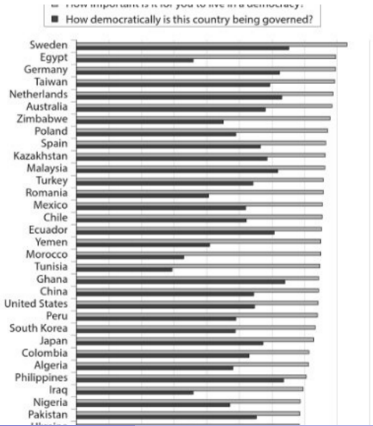
Alternate Graphs



Alternate Graphs

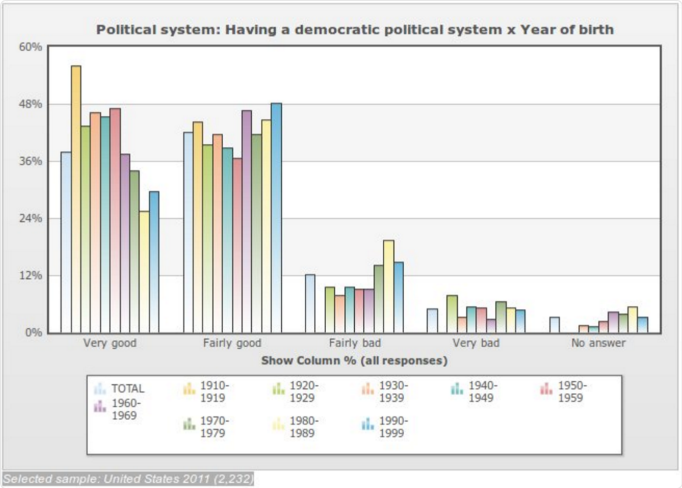
.@RyanDEnos Compare NYT/JoD (left) to the very same data analysed differently by Bartels and Achen (2016) (right). Extreme score vs means.



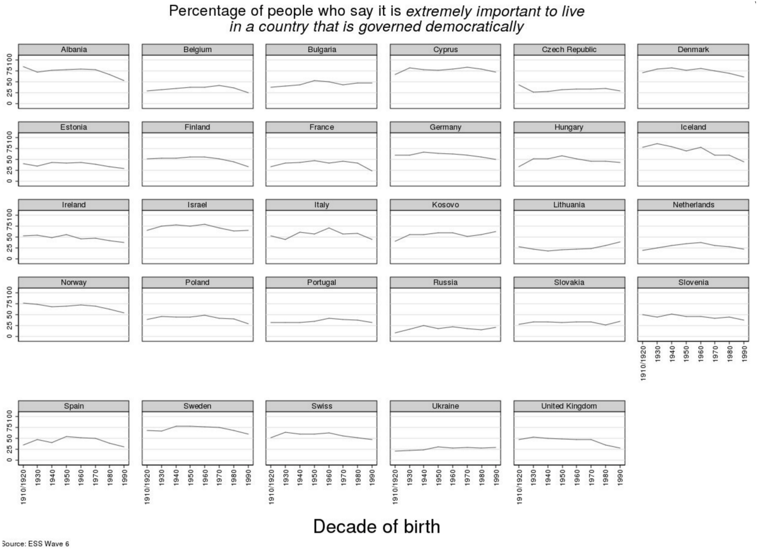


Alternate Graphs

@RyanDEnos They also stop at the 80s cohort. The data has the 90's as well. I wonder why they would stop there...



Alternate Graphs

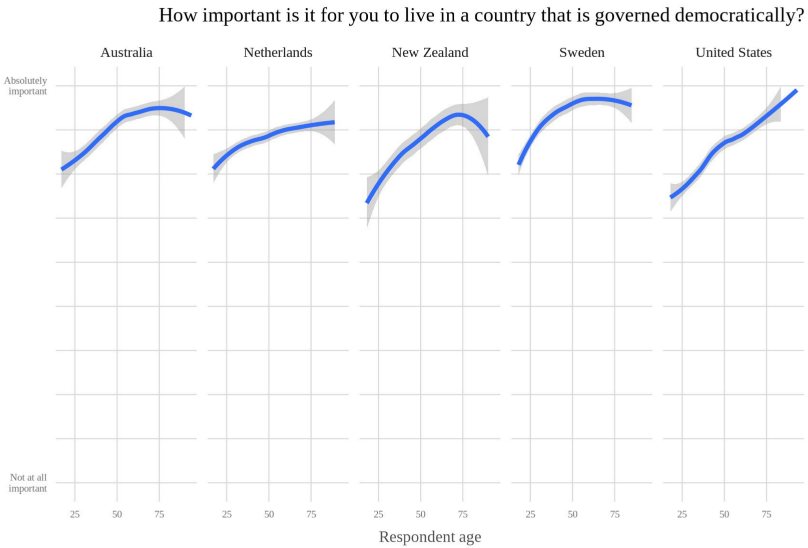


In reply to Ryan D. Enos

Benjamin Sack @bcsack · 15h

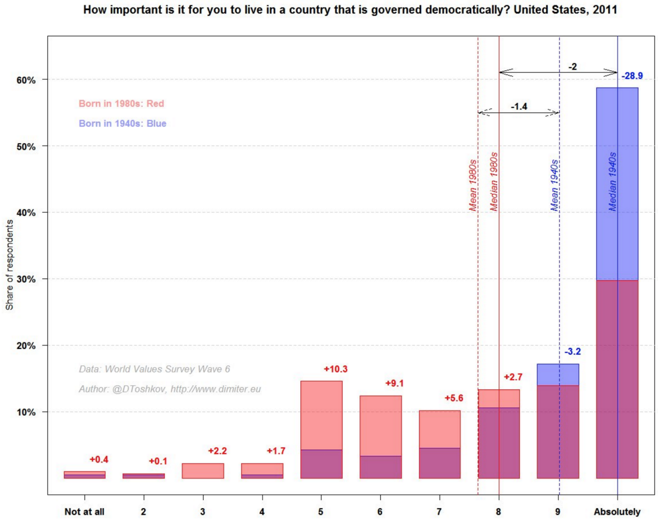
@RyanDEnos Same analysis strategy with comparable data from @ESS_Survey (similar item, 0-10 scale) shows slightly different pattern, too.

Alternate Graphs



614 Bantam @jpbach · 15h
@RyanDEnos @bshor @nataliemjb @TomWGvdMeer this is a "quick and dirty" plot I did with WVS wave 6. Not quite so terrifying.

Alternate Graphs



Dimiter Toshkov @DToshkov · 31m
my take on the democratic deconsolidation graph that scared everyone yesterday. Blue is 1940s cohort, red is 1980s. First, United States

Thoughts

Two stories here:

Thoughts

Two stories here:

1. Visualization and data coding choices are important

Thoughts

Two stories here:

1. Visualization and data coding choices are important
2. The internet is amazing (especially with replication data being available!)