Week 10: Causality with Measured Confounding

Brandon Stewart[1]

Princeton

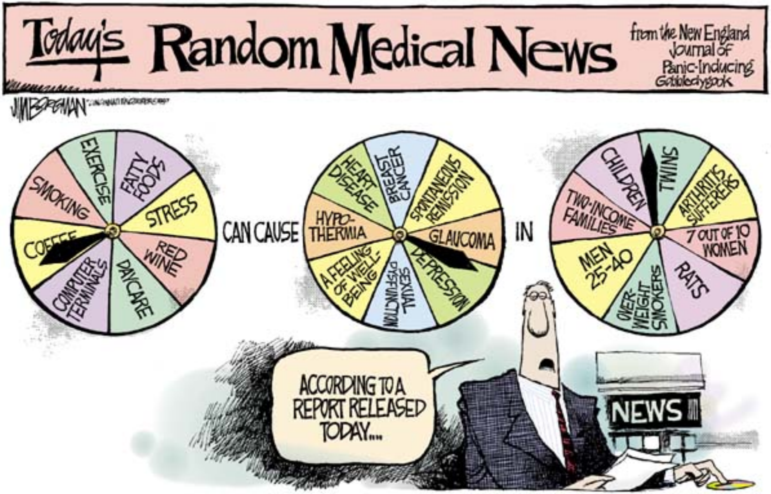
November 28 and 30, 2016

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

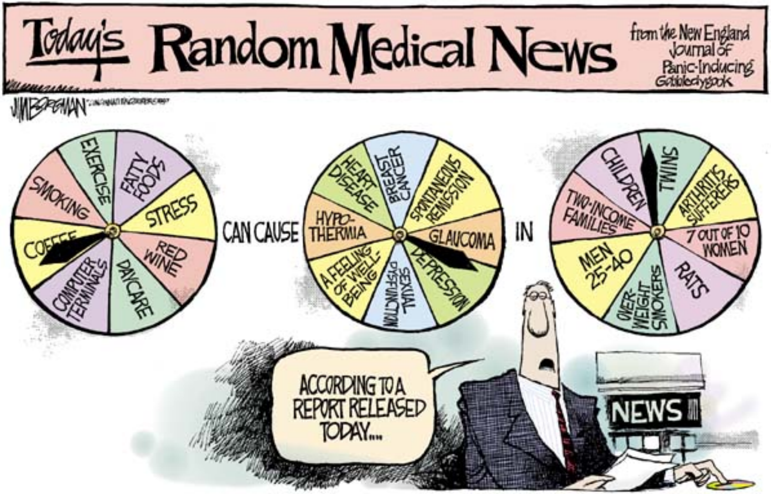
Where We've Been and Where We're Going...

- Last Week
  - regression diagnostics
- This Week
  - Monday:
    - experimental Ideal
    - identification with measured confounding
  - Wednesday:
    - regression estimation
- Next Week
  - identification with unmeasured confounding
  - instrumental variables
- Long Run
  - causality with measured confounding → unmeasured confounding → repeated data

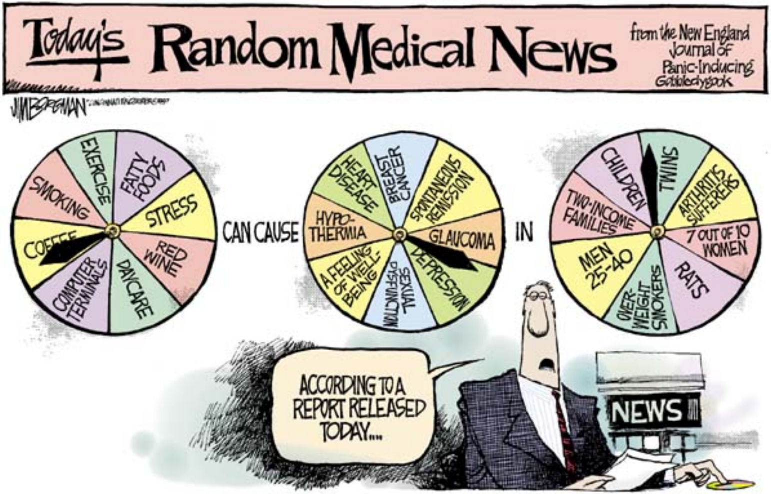
Questions?



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)

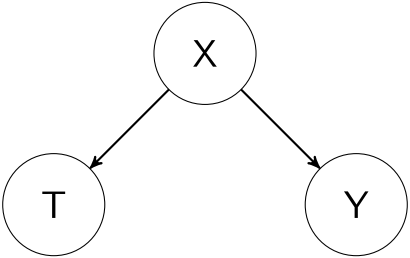


Lancet 2003: comparing among individuals with the same age, gender, blood pressure, diabetes, and smoking, those with higher vitamin C levels have lower levels of obesity, lower levels of alcohol consumption, are less likely to grow up in working class, etc.

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?

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

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:

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

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:

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

- 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	$\mathbb{E}[Y_1 \vert X = 0, D = 1]$	$\mathbb{E}[Y_0 \vert X = 0, D = 1]$	1	0
2			1	0
3	$\mathbb{E}[Y_1 \vert X = 0, D = 0]$	$\mathbb{E}[Y_0 \vert X = 0, D = 0]$	0	0
4			0	0
5	$\mathbb{E}[Y_1 \vert X = 1, D = 1]$	$\mathbb{E}[Y_0 \vert X = 1, D = 1]$	1	1
6			1	1
7	$\mathbb{E}[Y_1 \vert X = 1, D = 0]$	$\mathbb{E}[Y_0 \vert X = 1, D = 0]$	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:

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

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

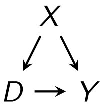
- Backdoor path: is a non-causal path from D to Y.

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.

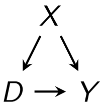
Backdoor paths and blocking paths

- 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:



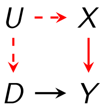
Backdoor paths and blocking paths

- 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:



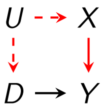
- Here there is a backdoor path $D \leftarrow X \rightarrow Y$, where X is a common cause for the treatment and the outcome.

Other types of confounding



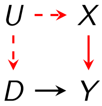
- D is enrolling in a job training program.

Other types of confounding



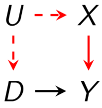
- D is enrolling in a job training program.
- Y is getting a job.

Other types of confounding



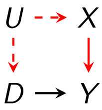
- D is enrolling in a job training program.
- Y is getting a job.
- U is being motivated

Other types of confounding



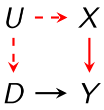
- D is enrolling in a job training program.
- Y is getting a job.
- U is being motivated
- X is number of job applications sent out.

Other types of confounding



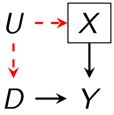
- D is enrolling in a job training program.
- Y is getting a job.
- U is being motivated
- X is number of job applications sent out.
- Big assumption here: no arrow from U to Y.

Other types of confounding



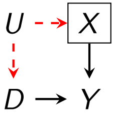
- D is exercise.
- Y is having a disease.
- U is lifestyle.
- X is smoking
- Big assumption here: no arrow from U to Y.

What's the problem with backdoor paths?



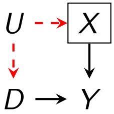
- A path is blocked if:

What's the problem with backdoor paths?



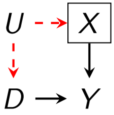
- A path is blocked if:
  1. we control for or stratify a non-collider on that path OR

What's the problem with backdoor paths?



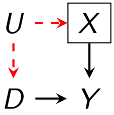
- A path is blocked if:
  1. we control for or stratify a non-collider on that path OR
  2. we do not control for a collider.

What's the problem with backdoor paths?



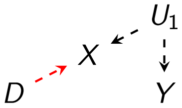
- A path is blocked if:
  1. we control for or stratify a non-collider on that path OR
  2. we do not control for a collider.
- Unblocked backdoor paths $\rightsquigarrow$ confounding.

What's the problem with backdoor paths?



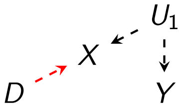
- A path is blocked if:
  1. we control for or stratify a non-collider on that path OR
  2. we do not control for a collider.
- Unblocked backdoor paths $\rightsquigarrow$ confounding.
- In the DAG here, if we condition on X, then the backdoor path is blocked.

Not all backdoor paths



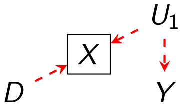
- Conditioning on the posttreatment covariates opens the non-causal path.

Not all backdoor paths



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

Not all backdoor paths



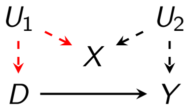
- Conditioning on the posttreatment covariates opens the non-causal path.
  - $\rightsquigarrow$ selection bias.

Don't condition on post-treatment variables

Don't condition on post-treatment variables

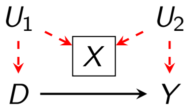
Every time you do, a puppy cries.

M-bias



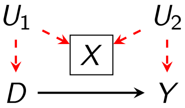
- Not all backdoor paths induce confounding.

M-bias



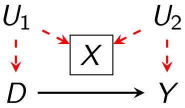
- Not all backdoor paths induce confounding.
- This backdoor path is blocked by the collider X that we don't control for.

M-bias



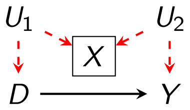
- Not all backdoor paths induce confounding.
- This backdoor path is blocked by the collider X that we don't control for.
- If we control for X $\rightsquigarrow$ opens the path and induces confounding.

M-bias



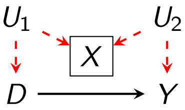
- Not all backdoor paths induce confounding.
- This backdoor path is blocked by the collider X that we don't control for.
- If we control for $X \rightsquigarrow$ opens the path and induces confounding.
  - Sometimes called M-bias.

M-bias



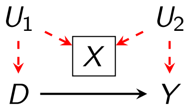
- Not all backdoor paths induce confounding.
- This backdoor path is blocked by the collider X that we don't control for.
- If we control for X $\rightsquigarrow$ opens the path and induces confounding.
  - Sometimes called M-bias.
- Controversial because of differing views on what to control for:

M-bias



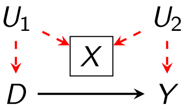
- Not all backdoor paths induce confounding.
- This backdoor path is blocked by the collider X that we don't control for.
- If we control for X $\rightsquigarrow$ opens the path and induces confounding.
  - Sometimes called M-bias.
- Controversial because of differing views on what to control for:
 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

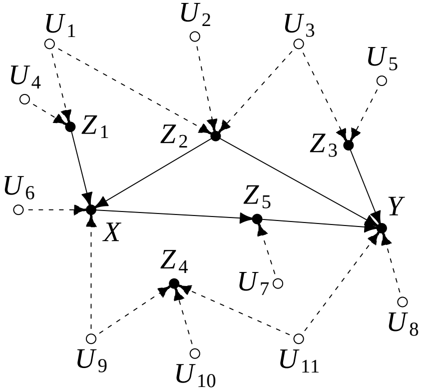
Backdoor criterion

- 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

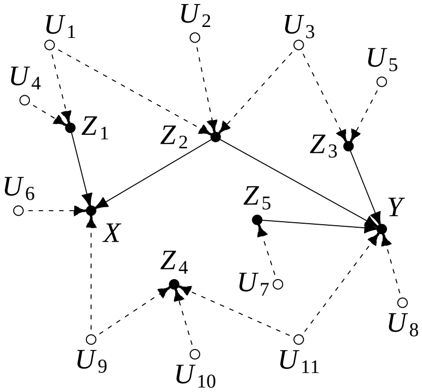
- Can we use a DAG to evaluate no unmeasured confounders?
- Pearl answered yes, with the backdoor criterion, which states that the effect of D on Y is identified if:
  1. No backdoor paths from D to Y OR
  2. Measured covariates are sufficient to block all backdoor paths from D to Y.
- First is really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



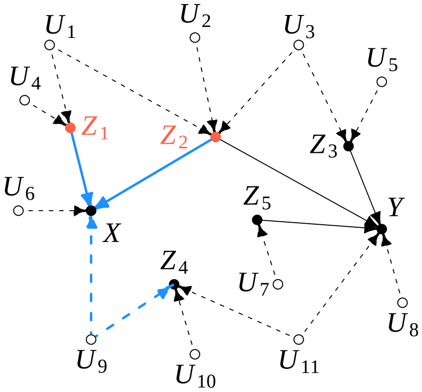
Remove arrows out of X.

Example: Sufficient Conditioning Sets



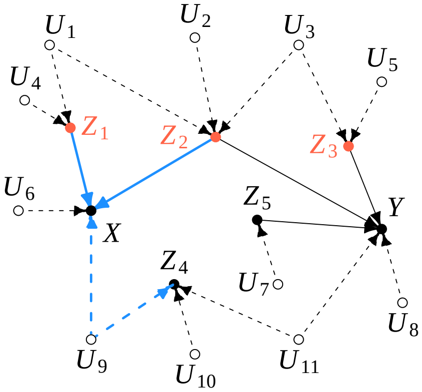
Recall that paths are blocked by "unconditioned colliders" or conditioned non-colliders

Example: Sufficient Conditioning Sets



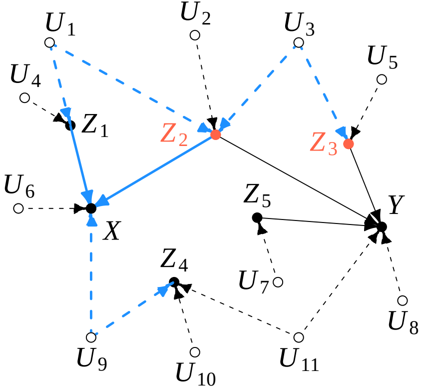
Recall that paths are blocked by "unconditioned colliders" or conditioned non-colliders

Example: Sufficient Conditioning Sets



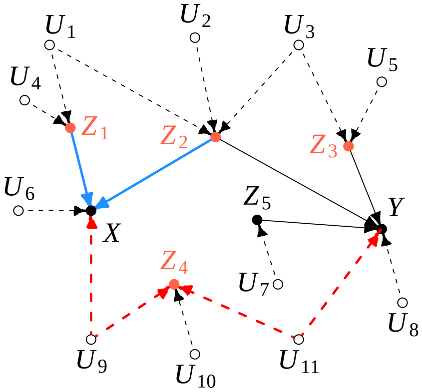
Recall that paths are blocked by "unconditioned colliders" or conditioned non-colliders

Example: Sufficient Conditioning Sets



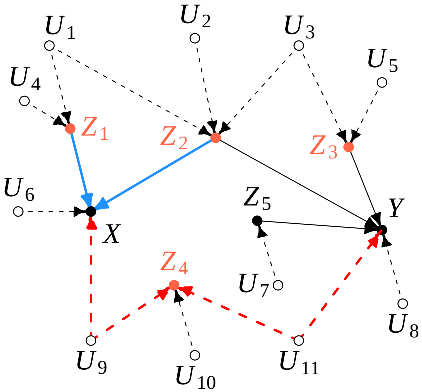
Recall that paths are blocked by "unconditioned colliders" or conditioned non-colliders

Example: Non-sufficient Conditioning Sets

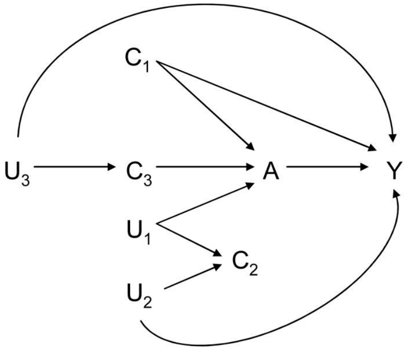


Recall that paths are blocked by "unconditioned colliders" or conditioned non-colliders

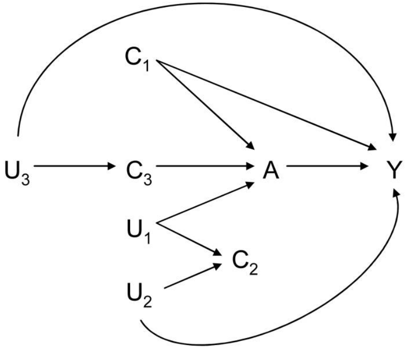
Example: Non-sufficient Conditioning Sets



Implications (via Vanderweele and Shpitser 2011)



Implications (via Vanderweele and Shpitser 2011)

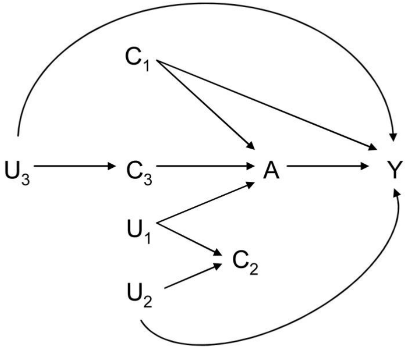


Two common criteria fail here:

1. Choose all pre-treatment covariates

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

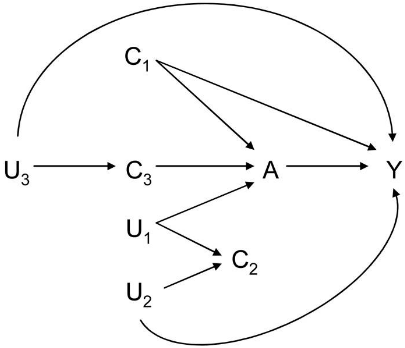
Implications (via Vanderweele and Shpitser 2011)



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

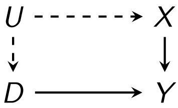
Implications (via Vanderweele and Shpitser 2011)



Two common criteria fail here:

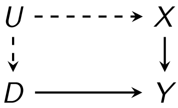
1. Choose all pre-treatment covariates
   (would condition on C_2 inducing M-bias)
2. Choose all covariates which directly cause the treatment and the outcome
   (would leave open a backdoor path $A \leftarrow C_3 \leftarrow U_3 \rightarrow Y$.)

SWIGs



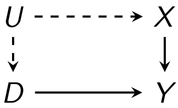
- It's a little hard to see how the backdoor criterion implies no unmeasured confounders.

SWIGs



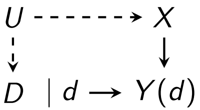
- It's a little hard to see how the backdoor criterion implies no unmeasured confounders.
  - No potential outcomes on this graph!

SWIGs



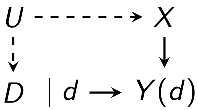
- It's a little hard to see how the backdoor criterion implies no unmeasured confounders.
  - No potential outcomes on this graph!
- Richardson and Robins: Single World Intervention Graphs

SWIGs



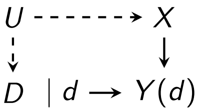
- It's a little hard to see how the backdoor criterion implies no unmeasured confounders.
  - No potential outcomes on this graph!
- Richardson and Robins: Single World Intervention Graphs
  - Split D node into natural value (D) and intervention value d.

SWIGs



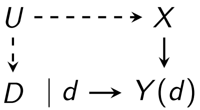
- It's a little hard to see how the backdoor criterion implies no unmeasured confounders.
  - No potential outcomes on this graph!
- Richardson and Robins: Single World Intervention Graphs
  - Split D node into natural value (D) and intervention value d.
  - Let all effects of D take their potential value under intervention $Y(d)$.

SWIGs



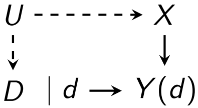
- It's a little hard to see how the backdoor criterion implies no unmeasured confounders.
  - No potential outcomes on this graph!
- Richardson and Robins: Single World Intervention Graphs
  - Split D node into natural value (D) and intervention value d.
  - Let all effects of D take their potential value under intervention $Y(d)$.
- Now can see: are D and $Y(d)$ related?

SWIGs



- It's a little hard to see how the backdoor criterion implies no unmeasured confounders.
  - No potential outcomes on this graph!
- Richardson and Robins: Single World Intervention Graphs
  - Split D node into natural value (D) and intervention value d.
  - Let all effects of D take their potential value under intervention $Y(d)$.
- Now can see: are D and $Y(d)$ related?
  - $D \leftarrow U \rightarrow X \rightarrow Y(d)$ implies not independent

SWIGs



- It's a little hard to see how the backdoor criterion implies no unmeasured confounders.
  - No potential outcomes on this graph!
- Richardson and Robins: Single World Intervention Graphs
  - Split D node into natural value (D) and intervention value d.
  - Let all effects of D take their potential value under intervention $Y(d)$.
- Now can see: are D and $Y(d)$ related?
  - $D \leftarrow U \rightarrow X \rightarrow Y(d)$ implies not independent
  - Conditioning on X blocks that backdoor path $\rightsquigarrow D \perp\!\!\!\perp Y(d) | X$

No unmeasured confounders is not testable

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

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

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 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.

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

- Here, $\overset{d}{=}$ 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	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)

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	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.
- 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 ≈ 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!

Summary

- Today we discussed issues of identification (with just a dash of estimation via stratification)

Summary

- Today we discussed issues of identification (with just a dash of estimation via stratification)
- Next class we will talk about estimation and what OLS is doing under this framework.

Summary

- Today we discussed issues of identification (with just a dash of estimation via stratification)
- Next class we will talk about estimation and what OLS is doing under this framework.
- Causal inference is hard but worth doing!

Fun with Censorship

Fun with Censorship

- Often you don't need sophisticated methods to reveal interesting findings

Fun with Censorship

- Often you don't need sophisticated methods to reveal interesting findings
- "Ansolabahere's Law": real relationship is visible in a bivariate plot and remains in a more sophisticated in a statistical model

Fun with Censorship

- Often you don't need sophisticated methods to reveal interesting findings
- "Ansolabahere's Law": real relationship is visible in a bivariate plot and remains in a more sophisticated in a statistical model
- In other words: all inferences require both visual and mathematical evidence

Fun with Censorship

- Often you don't need sophisticated methods to reveal interesting findings
- "Ansolabahere's Law": real relationship is visible in a bivariate plot and remains in a more sophisticated in a statistical model
- In other words: all inferences require both visual and mathematical evidence
- Example: King, Pan and Roberts (2013) "How Censorship in China Allows Government Criticism but Silences Collective Expression" *American Political Science Review*

Fun with Censorship

- Often you don't need sophisticated methods to reveal interesting findings
- "Ansolabahere's Law": real relationship is visible in a bivariate plot and remains in a more sophisticated in a statistical model
- In other words: all inferences require both visual and mathematical evidence
- Example: King, Pan and Roberts (2013) "How Censorship in China Allows Government Criticism but Silences Collective Expression" *American Political Science Review*
- They use very simple (statistical) methods to great effect.

Fun with Censorship

- Often you don't need sophisticated methods to reveal interesting findings
- "Ansolabahere's Law": real relationship is visible in a bivariate plot and remains in a more sophisticated in a statistical model
- In other words: all inferences require both visual and mathematical evidence
- Example: King, Pan and Roberts (2013) "How Censorship in China Allows Government Criticism but Silences Collective Expression" *American Political Science Review*
- They use very simple (statistical) methods to great effect.
- This line of work is one of my favorites.

Sequence of slides that follow courtesy of King, Pan and Roberts

Chinese Censorship

The largest selective suppression of human expression in history

Chinese Censorship

The largest selective suppression of human expression in history:

- implemented manually,
- by $\approx 200,000$ workers,
- located in government and inside social media firms

Chinese Censorship

The largest selective suppression of human expression in history:

- implemented manually,
- by $\approx 200,000$ workers,
- located in government and inside social media firms

Theories of the Goal of Censorship

Chinese Censorship

The largest selective suppression of human expression in history:

- implemented <u>manually</u>,
- by $\approx 200,000$ workers,
- located in government and inside social media firms

Theories of the Goal of Censorship

1. Stop criticism of the state

Chinese Censorship

The largest selective suppression of human expression in history:

- implemented manually,
- by $\approx 200,000$ workers,
- located in government and inside social media firms

Theories of the Goal of Censorship

1. Stop criticism of the state
2. Stop collective action

Chinese Censorship

The largest selective suppression of human expression in history:

- implemented manually,
- by $\approx 200,000$ workers,
- located in government and inside social media firms

Theories of the Goal of Censorship

1. Stop criticism of the state
2. Stop collective action

Either or both could be right or wrong.

Chinese Censorship

The largest selective suppression of human expression in history:

- implemented manually,
- by $\approx 200,000$ workers,
- located in government and inside social media firms

Theories of the Goal of Censorship

1. ~~Stop criticism of the state~~ Wrong
2. Stop collective action

Chinese Censorship

The largest selective suppression of human expression in history:

- implemented manually,
- by $\approx 200,000$ workers,
- located in government and inside social media firms

Theories of the Goal of Censorship

1. ~~Stop criticism of the state~~ Wrong
2. Stop collective action Right

Chinese Censorship

The largest selective suppression of human expression in history:

- implemented manually,
- by $\approx 200,000$ workers,
- located in government and inside social media firms

Theories of the Goal of Censorship — Benefit

1. ~~Stop criticism of the state~~ Wrong
2. Stop collective action Right

Chinese Censorship

The largest selective suppression of human expression in history:

- implemented manually,
- by $\approx 200,000$ workers,
- located in government and inside social media firms

Theories of the Goal of Censorship Benefit Cost

1. ~~Stop criticism of the state~~ Wrong
2. Stop collective action Right

Chinese Censorship

The largest selective suppression of human expression in history:

- implemented manually,
- by $\approx 200,000$ workers,
- located in government and inside social media firms

Theories of the Goal of Censorship	Benefit	Cost
1. ~~Stop criticism of the state~~ Wrong	?	
2. Stop collective action Right		

Chinese Censorship

The largest selective suppression of human expression in history:

- implemented manually,
- by $\approx 200,000$ workers,
- located in government and inside social media firms

Theories of the Goal of Censorship	Benefit	Cost
1. ~~Stop criticism of the state~~ Wrong	?	Huge
2. Stop collective action Right		

Chinese Censorship

The largest selective suppression of human expression in history:

- implemented manually,
- by $\approx 200,000$ workers,
- located in government and inside social media firms

Theories of the Goal of Censorship	Benefit	Cost
1. ~~Stop criticism of the state~~ Wrong	?	Huge
2. Stop collective action Right	Huge	

Chinese Censorship

The largest selective suppression of human expression in history:

- implemented manually,
- by $\approx 200,000$ workers,
- located in government and inside social media firms

Theories of the Goal of Censorship	Benefit	Cost
1. ~~Stop criticism of the state~~ Wrong	?	Huge
2. Stop collective action Right	Huge	Small

Chinese Censorship

The largest selective suppression of human expression in history:

- implemented manually,
- by $\approx 200,000$ workers,
- located in government and inside social media firms

Theories of the Goal of Censorship	Benefit	Cost
1. ~~Stop criticism of the state~~ Wrong	?	Huge
2. Stop collective action Right	Huge	Small

(They also censor 2 other smaller categories)

Observational Study

Observational Study

- Collect 3,674,698 social media posts in 85 topic areas over 6 months

Observational Study

- Collect 3,674,698 social media posts in 85 topic areas over 6 months
- Random sample: 127,283

Observational Study

- Collect 3,674,698 social media posts in 85 topic areas over 6 months
- Random sample: 127,283
- (Repeat design; Total analyzed: 11,382,221)

Observational Study

- Collect 3,674,698 social media posts in 85 topic areas over 6 months
- Random sample: 127,283
- (Repeat design; Total analyzed: 11,382,221)
- $\rightsquigarrow$ For each post (on a timeline in one of 85 content areas):

Observational Study

- Collect 3,674,698 social media posts in 85 topic areas over 6 months
- Random sample: 127,283
- (Repeat design; Total analyzed: 11,382,221)
- $\rightsquigarrow$ For each post (on a timeline in one of 85 content areas):
  - Download content the instant it appears

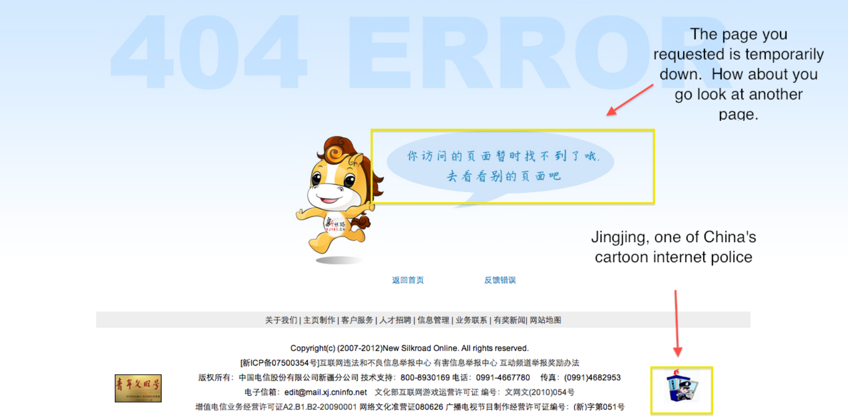
Observational Study

- Collect 3,674,698 social media posts in 85 topic areas over 6 months
- Random sample: 127,283
- (Repeat design; Total analyzed: 11,382,221)
- $\rightsquigarrow$ For each post (on a timeline in one of 85 content areas):
  - Download content the instant it appears
  - (Carefully) revisit each later to determine if it was censored

Observational Study

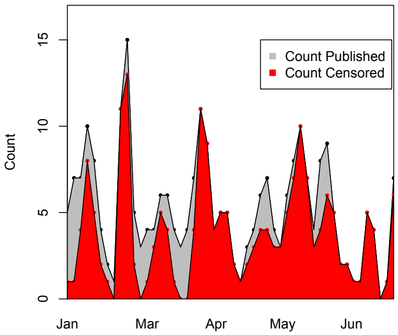
- Collect 3,674,698 social media posts in 85 topic areas over 6 months
- Random sample: 127,283
- (Repeat design; Total analyzed: 11,382,221)
- $\rightsquigarrow$ For each post (on a timeline in one of 85 content areas):
  - Download content the instant it appears
  - (Carefully) revisit each later to determine if it was censored
  - Use computer-assisted methods of text analysis (some existing, some new, all adapted to Chinese)

Censorship is not Ambiguous: BBS Error Page

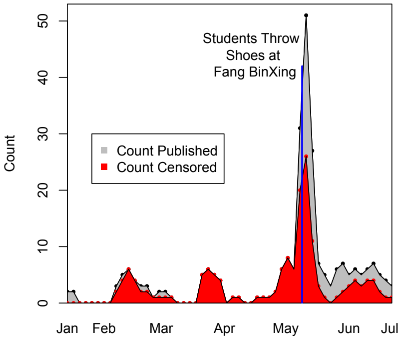


For 2 Unusual Topics: Constant Censorship Effort

For 2 Unusual Topics: Constant Censorship Effort



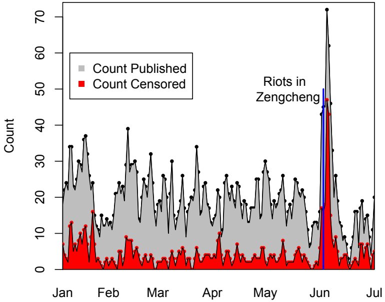
Pornography



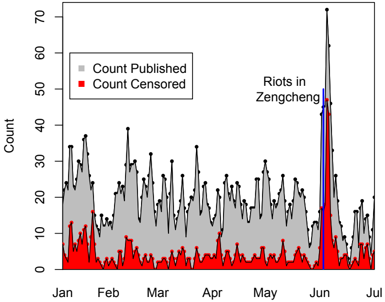
Criticism of the Censors

All other topics: Censorship & Post Volume are "Bursty"

All other topics: Censorship & Post Volume are "Bursty"

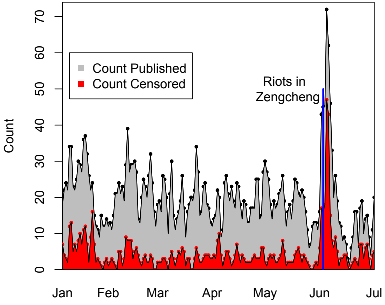


All other topics: Censorship & Post Volume are "Bursty"



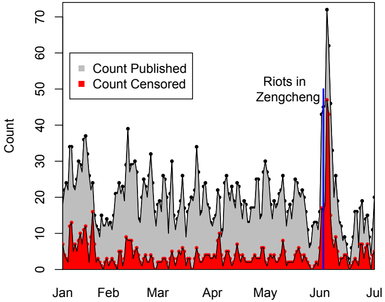
- Unit of analysis:

All other topics: Censorship & Post Volume are "Bursty"



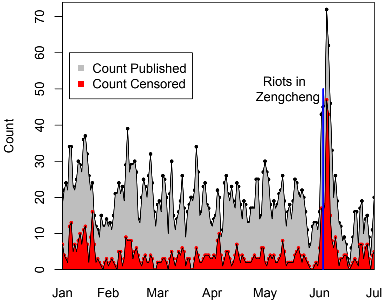
- Unit of analysis:
  - volume burst

All other topics: Censorship & Post Volume are "Bursty"



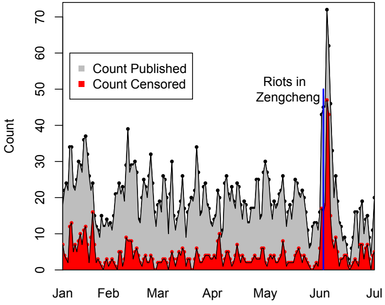
- Unit of analysis:
  - volume burst
  - ($\approx$ 3 SDs greater than baseline volume)

All other topics: Censorship & Post Volume are "Bursty"



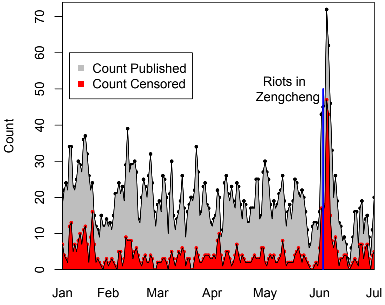
- Unit of analysis:
  - volume burst
  - ($\approx$ 3 SDs greater than baseline volume)
- They monitored 85 topic areas (Jan–July 2011)

All other topics: Censorship & Post Volume are “Bursty”



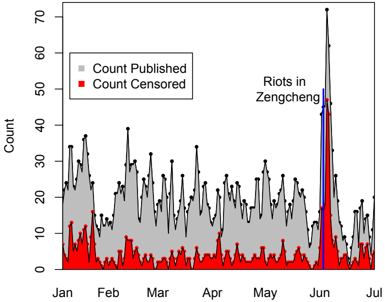
- Unit of analysis:
  - volume burst
  - ($\approx$ 3 SDs greater than baseline volume)
- They monitored 85 topic areas (Jan–July 2011)
- Found 87 volume bursts in total

All other topics: Censorship & Post Volume are "Bursty"



- Unit of analysis:
  - volume burst
  - (≈ 3 SDs greater than baseline volume)
- They monitored 85 topic areas (Jan–July 2011)
- Found 87 volume bursts in total
- Identified real world events associated with each burst

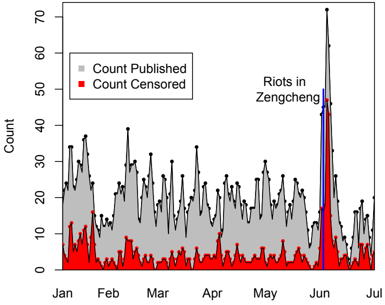
All other topics: Censorship & Post Volume are "Bursty"



- Unit of analysis:
  - volume burst
  - ($\approx$ 3 SDs greater than baseline volume)
- They monitored 85 topic areas (Jan–July 2011)
- Found 87 volume bursts in total
- Identified real world events associated with each burst

Their hypothesis: The government censors all posts in volume bursts associated with events with collective action (regardless of how critical or supportive of the state)

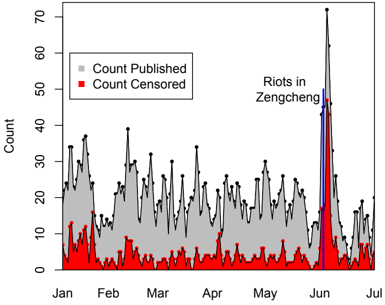
All other topics: Censorship & Post Volume are "Bursty"



- Unit of analysis:
  - volume burst
  - (≈ 3 SDs greater than baseline volume)
- They monitored 85 topic areas (Jan–July 2011)
- Found 87 volume bursts in total
- Identified real world events associated with each burst

Their hypothesis: The government censors all posts in volume bursts associated with events with collective action (regardless of how critical or supportive of the state)

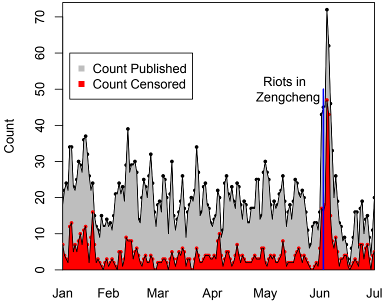
All other topics: Censorship & Post Volume are "Bursty"



- Unit of analysis:
  - volume burst
  - ($\approx$ 3 SDs greater than baseline volume)
- They monitored 85 topic areas (Jan–July 2011)
- Found 87 volume bursts in total
- Identified real world events associated with each burst

Their hypothesis: The government censors all posts in volume bursts associated with events with collective action (regardless of how critical or supportive of the state)

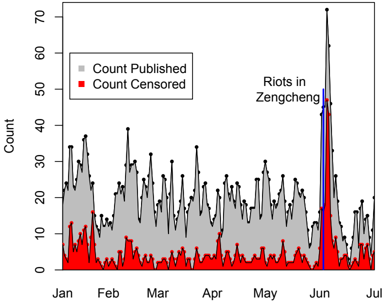
All other topics: Censorship & Post Volume are "Bursty"



- Unit of analysis:
  - volume burst
  - (≈ 3 SDs greater than baseline volume)
- They monitored 85 topic areas (Jan–July 2011)
- Found 87 volume bursts in total
- Identified real world events associated with each burst

Their hypothesis: The government censors all posts in volume bursts associated with events with collective action (regardless of how critical or supportive of the state)

All other topics: Censorship & Post Volume are "Bursty"



- Unit of analysis:
  - volume burst
  - (≈ 3 SDs greater than baseline volume)
- They monitored 85 topic areas (Jan–July 2011)
- Found 87 volume bursts in total
- Identified real world events associated with each burst

Their hypothesis: The government censors all posts in volume bursts associated with events with collective action (regardless of how critical or supportive of the state)

Observational Test 1: Post Volume

Observational Test 1: Post Volume

- Begin with 87 volume bursts in 85 topics areas

Observational Test 1: Post Volume

- Begin with 87 volume bursts in 85 topics areas
- For each burst, calculate change in % censorship inside to outside each volume burst within topic areas – censorship magnitude

Observational Test 1: Post Volume

- Begin with 87 volume bursts in 85 topics areas
- For each burst, calculate change in % censorship inside to outside each volume burst within topic areas – censorship magnitude
- If goal of censorship is to stop collective action, they expect:

Observational Test 1: Post Volume

- Begin with 87 volume bursts in 85 topics areas
- For each burst, calculate change in % censorship inside to outside each volume burst within topic areas – censorship magnitude
- If goal of censorship is to stop collective action, they expect:

1. On average, % censored should increase during volume bursts

Observational Test 1: Post Volume

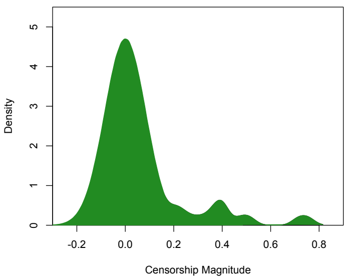
- Begin with 87 volume bursts in 85 topics areas
- For each burst, calculate change in % censorship inside to outside each volume burst within topic areas – censorship magnitude
- If goal of censorship is to stop collective action, they expect:

1. On average, % censored should increase during volume bursts
2. Some bursts (associated with politically relevant events) should have much higher censorship

Observational Test 1: Post Volume

- Begin with 87 volume bursts in 85 topics areas
- For each burst, calculate change in % censorship inside to outside each volume burst within topic areas – censorship magnitude
- If goal of censorship is to stop collective action, they expect:

1. On average, % censored should increase during volume bursts
2. Some bursts (associated with politically relevant events) should have much higher censorship



Observational Test 2: The Event Generating Volume Bursts

Observational Test 2: The Event Generating Volume Bursts

Event classification (each category can be $+$, $-$, or neutral comments about the state)

Observational Test 2: The Event Generating Volume Bursts

Event classification (each category can be $+$, $-$, or neutral comments about the state)

1. Collective Action Potential

Observational Test 2: The Event Generating Volume Bursts

Event classification (each category can be $+$, $-$, or neutral comments about the state)

1. Collective Action Potential
   - protest or organized crowd formation outside the Internet

Observational Test 2: The Event Generating Volume Bursts

Event classification (each category can be $+$, $-$, or neutral comments about the state)

1. Collective Action Potential
   - protest or organized crowd formation outside the Internet
   - individuals who have organized or incited collective action on the ground in the past;

Observational Test 2: The Event Generating Volume Bursts

Event classification (each category can be $+$, $-$, or neutral comments about the state)

1. Collective Action Potential
   - protest or organized crowd formation outside the Internet
   - individuals who have organized or incited collective action on the ground in the past;
   - topics related to nationalism or nationalist sentiment that have incited protest or collective action in the past.

Observational Test 2: The Event Generating Volume Bursts

Event classification (each category can be $+$, $-$, or neutral comments about the state)

1. Collective Action Potential
   - protest or organized crowd formation outside the Internet
   - individuals who have organized or incited collective action on the ground in the past;
   - topics related to nationalism or nationalist sentiment that have incited protest or collective action in the past.
2. Criticism of censors

Observational Test 2: The Event Generating Volume Bursts

Event classification (each category can be $+$, $-$, or neutral comments about the state)

1. Collective Action Potential
   - protest or organized crowd formation outside the Internet
   - individuals who have organized or incited collective action on the ground in the past;
   - topics related to nationalism or nationalist sentiment that have incited protest or collective action in the past.
2. Criticism of censors
3. Pornography

Observational Test 2: The Event Generating Volume Bursts

Event classification (each category can be $+$, $-$, or neutral comments about the state)

1. Collective Action Potential
   - protest or organized crowd formation outside the Internet
   - individuals who have organized or incited collective action on the ground in the past;
   - topics related to nationalism or nationalist sentiment that have incited protest or collective action in the past.
2. Criticism of censors
3. Pornography
4. (Other) News

Observational Test 2: The Event Generating Volume Bursts

Event classification (each category can be $+$, $-$, or neutral comments about the state)

1. Collective Action Potential
   - protest or organized crowd formation outside the Internet
   - individuals who have organized or incited collective action on the ground in the past;
   - topics related to nationalism or nationalist sentiment that have incited protest or collective action in the past.
2. Criticism of censors
3. Pornography
4. (Other) News
5. Government Policies

Observational Test 2: The Event Generating Volume Bursts

Event classification (each category can be $+$, $-$, or neutral comments about the state)

1. ~~Collective Action Potential~~ 

    - protest or organized crowd formation outside the Internet
    - individuals who have organized or incited collective action on the ground in the past;
    - topics related to nationalism or nationalist sentiment that have incited protest or collective action in the past.
2. Criticism of censors
3. Pornography
4. (Other) News
5. Government Policies

Observational Test 2: The Event Generating Volume Bursts

Event classification (each category can be $+$, $-$, or neutral comments about the state)

1. ~~Collective Action Potential~~ 

   - protest or organized crowd formation outside the Internet
   - individuals who have organized or incited collective action on the ground in the past;
   - topics related to nationalism or nationalist sentiment that have incited protest or collective action in the past.
2. ~~Criticism of censors~~ 

3. Pornography
4. (Other) News
5. Government Policies

Observational Test 2: The Event Generating Volume Bursts

Event classification (each category can be $+$, $-$, or neutral comments about the state)

1. ~~Collective Action Potential~~ CENSORED
    - protest or organized crowd formation outside the Internet
    - individuals who have organized or incited collective action on the ground in the past;
    - topics related to nationalism or nationalist sentiment that have incited protest or collective action in the past.
2. ~~Criticism of censors~~ CENSORED
3. ~~Pornography~~ CENSORED
4. (Other) News
5. Government Policies

Observational Test 2: The Event Generating Volume Bursts

Event classification (each category can be $+$, $-$, or neutral comments about the state)

1. ~~Collective Action Potential~~ 

   - protest or organized crowd formation outside the Internet
   - individuals who have organized or incited collective action on the ground in the past;
   - topics related to nationalism or nationalist sentiment that have incited protest or collective action in the past.
2. ~~Criticism of censors~~ 

3. ~~Pornography~~ 

4. (Other) News 
5. Government Policies

Observational Test 2: The Event Generating Volume Bursts

Event classification (each category can be $+$, $-$, or neutral comments about the state)

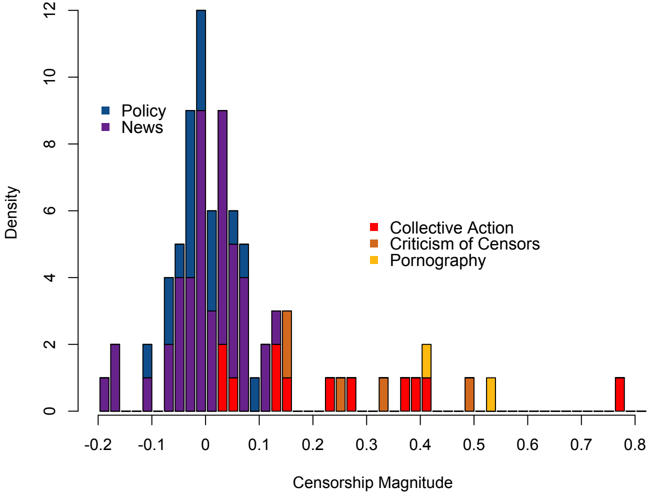
1. ~~Collective Action Potential~~ 

    - protest or organized crowd formation outside the Internet
    - individuals who have organized or incited collective action on the ground in the past;
    - topics related to nationalism or nationalist sentiment that have incited protest or collective action in the past.
2. ~~Criticism of censors~~ 

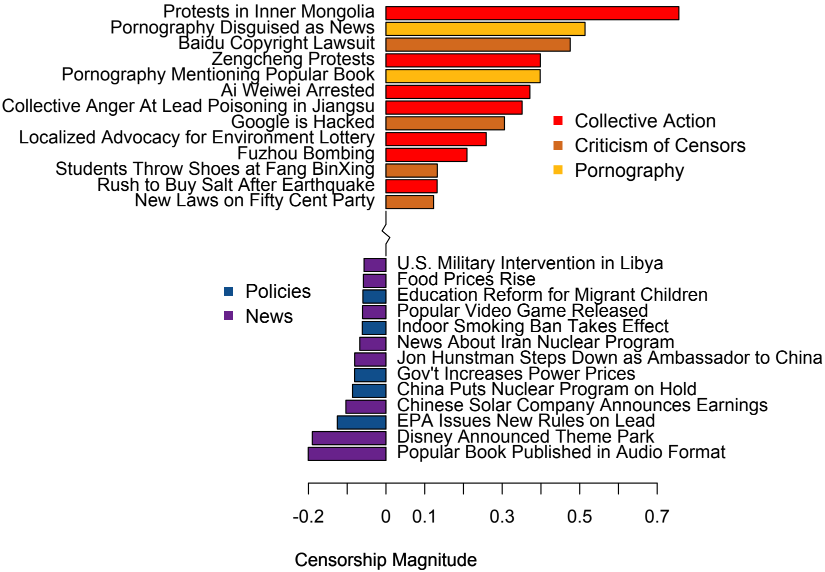
3. ~~Pornography~~ 

4. (Other) News 
5. Government Policies 

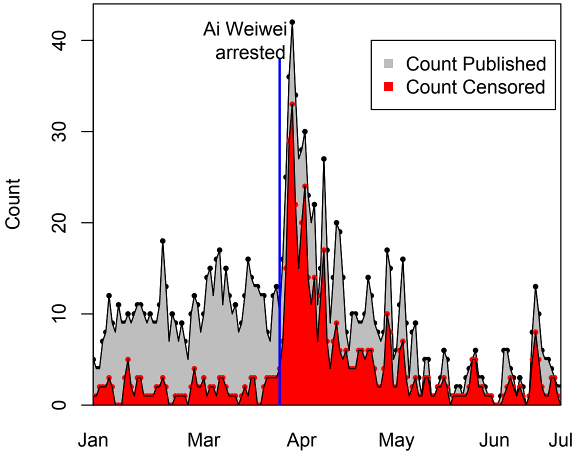
What Types of Events Are Censored?



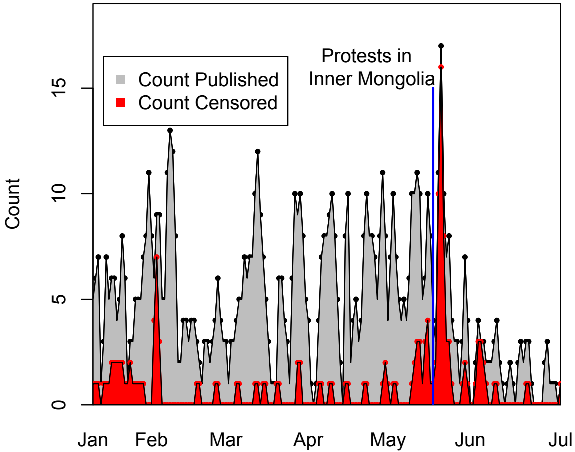
What Types of Events Are Censored?



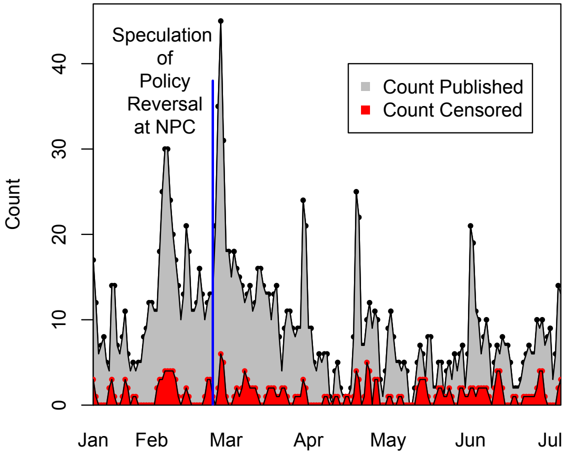
Censoring Collective Action: Ai Weiwei's Arrest



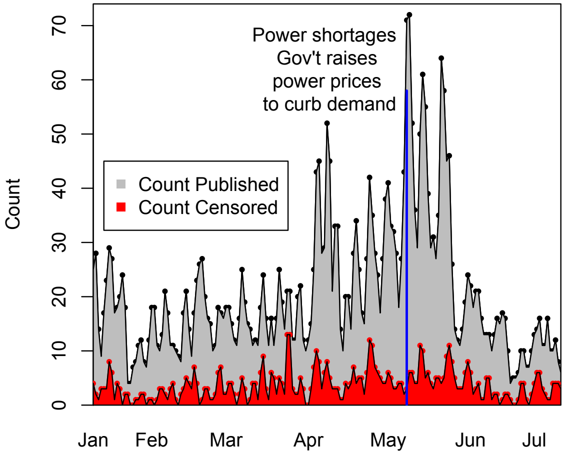
Censoring Collective Action: Protests in Inner Mongolia



Low Censorship on One Child Policy



Low Censorship on News: Power Prices



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

- Last Week
  - regression diagnostics
- This Week
  - Monday:
    - experimental Ideal
    - identification with measured confounding
  - Wednesday:
    - regression estimation
- Next Week
  - identification with unmeasured confounding
  - instrumental variables
- Long Run
  - causality with measured confounding $\rightarrow$ unmeasured confounding $\rightarrow$ repeated data

Questions?

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 average is often 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 is an excellent introduction to the agnostic view of regression and I recommend checking it out. Here I will give you just a flavor of it.

Regression as parametric modeling

Let's start with the parametric view we have taken thus far.

Regression as parametric modeling

Let's start with the parametric view we have taken thus far.

- Gauss-Markov assumptions:
  - linearity, i.i.d. sample, full rank X_i, zero conditional mean error, homoskedasticity.

Regression as parametric modeling

Let's start with the parametric view we have taken thus far.

- Gauss-Markov assumptions:
  - linearity, i.i.d. sample, full rank X_i, zero conditional mean error, homoskedasticity.
- $\rightsquigarrow$ OLS is BLUE, plus normality of the errors and we get small sample SEs.

Regression as parametric modeling

Let's start with the parametric view we have taken thus far.

- Gauss-Markov assumptions:
  - linearity, i.i.d. sample, full rank X_i, zero conditional mean error, homoskedasticity.
- $\rightsquigarrow$ OLS is BLUE, plus normality of the errors and we get small sample SEs.
- What is the basic approach here? It is a model for the conditional distribution of Y_i given X_i:

$$[Y_i|X_i] \sim N(X_i'\beta, \sigma^2)$$

Agnostic views on regression

$$[Y_i|X_i] \sim N(X_i'\beta, \sigma^2)$$

- Above parametric view has strong distributional assumption on Y_i.

Agnostic views on regression

$$[Y_i|X_i] \sim N(X_i'\beta, \sigma^2)$$

- Above parametric view has strong distributional assumption on Y_i.
- Properties like BLUE or BUE depend on these assumptions holding.

Agnostic views on regression

$$[Y_i | X_i] \sim N(X_i'\beta, \sigma^2)$$

- Above parametric view has strong distributional assumption on Y_i.
- Properties like BLUE or BUE depend on these assumptions holding.
- Alternative: take an agnostic view on regression.

Agnostic views on regression

$$[Y_i|X_i] \sim N(X_i'\beta, \sigma^2)$$

- Above parametric view has strong distributional assumption on Y_i.
- Properties like BLUE or BUE depend on these assumptions holding.
- Alternative: take an agnostic view on regression.
  - Use OLS without believing these assumptions.

Agnostic views on regression

$$[Y_i|X_i] \sim N(X_i'\beta, \sigma^2)$$

- Above parametric view has strong distributional assumption on Y_i.
- Properties like BLUE or BUE depend on these assumptions holding.
- Alternative: take an agnostic view on regression.
  - Use OLS without believing these assumptions.
- Lose the distributional assumptions, focus on the conditional expectation function (CEF):

$$\mu(x) = \mathbb{E}[Y_i|X_i = x] = \sum_y y \cdot \mathbb{P}[Y_i = y|X_i = x]$$

Justifying linear regression

- Define linear regression:

$$\beta = \arg\min_b \mathbb{E}[(Y_i - X_i' b)^2]$$

Justifying linear regression

- Define linear regression:

$$\beta = \arg\min_b \mathbb{E}[(Y_i - X_i' b)^2]$$

- The solution to this is the following:

$$\beta = \mathbb{E}[X_i X_i']^{-1} \mathbb{E}[X_i Y_i]$$

Justifying linear regression

- Define linear regression:

$$\beta = \arg\min_{b} \mathbb{E}[(Y_i - X_i' b)^2]$$

- The solution to this is the following:

$$\beta = \mathbb{E}[X_i X_i']^{-1} \mathbb{E}[X_i Y_i]$$

- Note that the is the population coefficient vector, not the estimator yet.

Justifying linear regression

- Define linear regression:

$$\beta = \arg\min_b \mathbb{E}[(Y_i - X_i' b)^2]$$

- The solution to this is the following:

$$\beta = \mathbb{E}[X_i X_i']^{-1} \mathbb{E}[X_i Y_i]$$

- Note that the is the population coefficient vector, not the estimator yet.
- In other words, even a non-linear CEF has a "true" linear approximation, even though that approximation may not be great.

Regression anatomy

- Consider simple linear regression:

$$(\alpha, \beta) = \arg\min_{a,b} \mathbb{E}\left[(Y_i - a - bX_i)^2\right]$$

Regression anatomy

- Consider simple linear regression:

$$(\alpha, \beta) = \arg\min_{a,b} \mathbb{E}\left[(Y_i - a - bX_i)^2\right]$$

- In this case, we can write the population/true slope β as:

$$\beta = \mathbb{E}[X_i X_i']^{-1}\mathbb{E}[X_i Y_i] = \frac{\text{Cov}(Y_i, X_i)}{\text{Var}[X_i]}$$

Regression anatomy

- Consider simple linear regression:

$$(\alpha, \beta) = \arg\min_{a,b} \mathbb{E}\left[(Y_i - a - bX_i)^2\right]$$

- In this case, we can write the population/true slope β as:

$$\beta = \mathbb{E}[X_i X_i']^{-1}\mathbb{E}[X_i Y_i] = \frac{\text{Cov}(Y_i, X_i)}{\text{Var}[X_i]}$$

- With more covariates, β is more complicated, but we can still write it like this.

Regression anatomy

- Consider simple linear regression:

$$(\alpha, \beta) = \arg\min_{a,b} \mathbb{E}\left[(Y_i - a - bX_i)^2\right]$$

- In this case, we can write the population/true slope β as:

$$\beta = \mathbb{E}[X_i X_i']^{-1}\mathbb{E}[X_i Y_i] = \frac{\text{Cov}(Y_i, X_i)}{\mathrm{Var}[X_i]}$$

- With more covariates, β is more complicated, but we can still write it like this.
- Let $\tilde{X}_{ki}$ be the residual from a regression of X_{ki} on all the other independent variables. Then, β_k, the coefficient for X_{ki} is:

$$\beta_k = \frac{\text{Cov}(Y_i, \tilde{X}_{ki})}{\mathrm{Var}(\tilde{X}_{ki})}$$

Justification 1: Linear CEFs

Justification 1: Linear CEFs

- Justification 1: if the CEF is linear, the population regression function is it. That is, if $E[Y_i|X_i] = X_i'b$, then $b = \beta$.

Justification 1: Linear CEFs

- Justification 1: if the CEF is linear, the population regression function is it. That is, if $E[Y_i|X_i] = X_i'b$, then $b = \beta$.
- When would we expect the CEF to be linear? Two cases.

Justification 1: Linear CEFs

- Justification 1: if the CEF is linear, the population regression function is it. That is, if $E[Y_i|X_i] = X_i'b$, then $b = \beta$.
- When would we expect the CEF to be linear? Two cases.
  1. Outcome and covariates are multivariate normal.
  2. Linear regression model is saturated.

Justification 1: Linear CEFs

- Justification 1: if the CEF is linear, the population regression function is it. That is, if $E[Y_i|X_i] = X_i'b$, then $b = \beta$.
- When would we expect the CEF to be linear? Two cases.
  1. Outcome and covariates are multivariate normal.
  2. Linear regression model is saturated.
- A model is saturated if there are as many parameters as there are possible combination of the X_i variables.

Saturated model example

- Two binary variables, X_{1i} for marriage status and X_{2i} for having children.

Saturated model example

- Two binary variables, X_{1i} for marriage status and X_{2i} for having children.

Saturated model example

- Two binary variables, X_{1i} for marriage status and X_{2i} for having children.
- Four possible values of X_i, four possible values of $\mu(X_i)$:

$$E[Y_i|X_{1i} = 0, X_{2i} = 0]$$
$$E[Y_i|X_{1i} = 1, X_{2i} = 0]$$
$$E[Y_i|X_{1i} = 0, X_{2i} = 1]$$
$$E[Y_i|X_{1i} = 1, X_{2i} = 1]$$

Saturated model example

- Two binary variables, X_{1i} for marriage status and X_{2i} for having children.
- Four possible values of X_i, four possible values of $\mu(X_i)$:

$$E[Y_i|X_{1i}=0, X_{2i}=0] = \alpha$$
$$E[Y_i|X_{1i}=1, X_{2i}=0]$$
$$E[Y_i|X_{1i}=0, X_{2i}=1]$$
$$E[Y_i|X_{1i}=1, X_{2i}=1]$$

- We can write the CEF as follows:

$$E[Y_i|X_{1i}, X_{2i}] = \alpha + \beta X_{1i} + \gamma X_{2i} + \delta(X_{1i}X_{2i})$$

Saturated model example

- Two binary variables, X_{1i} for marriage status and X_{2i} for having children.
- Four possible values of X_i, four possible values of $\mu(X_i)$:

$$
\begin{aligned}
E[Y_i|X_{1i} = 0, X_{2i} = 0] &= \alpha \\
E[Y_i|X_{1i} = 1, X_{2i} = 0] &= \alpha + \beta \\
E[Y_i|X_{1i} = 0, X_{2i} = 1] \\
E[Y_i|X_{1i} = 1, X_{2i} = 1]
\end{aligned}
$$

- We can write the CEF as follows:

$$E[Y_i|X_{1i}, X_{2i}] = \alpha + \beta X_{1i} + \gamma X_{2i} + \delta(X_{1i}X_{2i})$$

Saturated model example

- Two binary variables, X_{1i} for marriage status and X_{2i} for having children.
- Four possible values of X_i, four possible values of $\mu(X_i)$:

$$E[Y_i|X_{1i} = 0, X_{2i} = 0] = \alpha$$
$$E[Y_i|X_{1i} = 1, X_{2i} = 0] = \alpha + \beta$$
$$E[Y_i|X_{1i} = 0, X_{2i} = 1] = \alpha + \gamma$$
$$E[Y_i|X_{1i} = 1, X_{2i} = 1]$$

- We can write the CEF as follows:

$$E[Y_i|X_{1i}, X_{2i}] = \alpha + \beta X_{1i} + \gamma X_{2i} + \delta(X_{1i}X_{2i})$$

Saturated model example

- Two binary variables, X_{1i} for marriage status and X_{2i} for having children.
- Four possible values of X_i, four possible values of $\mu(X_i)$:

$$E[Y_i|X_{1i}=0, X_{2i}=0] = \alpha$$
$$E[Y_i|X_{1i}=1, X_{2i}=0] = \alpha + \beta$$
$$E[Y_i|X_{1i}=0, X_{2i}=1] = \alpha + \gamma$$
$$E[Y_i|X_{1i}=1, X_{2i}=1] = \alpha + \beta + \gamma + \delta$$

- We can write the CEF as follows:

$$E[Y_i|X_{1i}, X_{2i}] = \alpha + \beta X_{1i} + \gamma X_{2i} + \delta(X_{1i}X_{2i})$$

Saturated models example

$$E[Y_i|X_{1i}, X_{2i}] = \alpha + \beta X_{1i} + \gamma X_{2i} + \delta(X_{1i}X_{2i})$$

- Basically, each value of $\mu(X_i)$ is being estimated separately.

Saturated models example

$$E[Y_i|X_{1i}, X_{2i}] = \alpha + \beta X_{1i} + \gamma X_{2i} + \delta(X_{1i}X_{2i})$$

- Basically, each value of $\mu(X_i)$ is being estimated separately.
  - $\rightsquigarrow$ within-strata estimation.

Saturated models example

$$E[Y_i|X_{1i}, X_{2i}] = \alpha + \beta X_{1i} + \gamma X_{2i} + \delta(X_{1i}X_{2i})$$

- Basically, each value of $\mu(X_i)$ is being estimated separately.
  - $\rightsquigarrow$ within-strata estimation.
  - No borrowing of information from across values of X_i.

Saturated models example

$$E[Y_i|X_{1i}, X_{2i}] = \alpha + \beta X_{1i} + \gamma X_{2i} + \delta(X_{1i}X_{2i})$$

- Basically, each value of $\mu(X_i)$ is being estimated separately.
  - $\rightsquigarrow$ within-strata estimation.
  - No borrowing of information from across values of X_i.
- Requires a set of dummies for each categorical variable plus all interactions.

Saturated models example

$$E[Y_i|X_{1i}, X_{2i}] = \alpha + \beta X_{1i} + \gamma X_{2i} + \delta(X_{1i}X_{2i})$$

- Basically, each value of $\mu(X_i)$ is being estimated separately.
  - $\rightsquigarrow$ within-strata estimation.
  - No borrowing of information from across values of X_i.
- Requires a set of dummies for each categorical variable plus all interactions.
- Or, a series of dummies for each unique combination of X_i.

Saturated models example

$$E[Y_i | X_{1i}, X_{2i}] = \alpha + \beta X_{1i} + \gamma X_{2i} + \delta(X_{1i} X_{2i})$$

- Basically, each value of $\mu(X_i)$ is being estimated separately.
  - $\rightsquigarrow$ within-strata estimation.
  - No borrowing of information from across values of X_i.
- Requires a set of dummies for each categorical variable plus all interactions.
- Or, a series of dummies for each unique combination of X_i.
- This makes linearity hold mechanically and so linearity is not an assumption.

Saturated model example

- Washington (AER) data on the effects of daughters.

Saturated model example

- Washington (AER) data on the effects of daughters.
- We'll look at the relationship between voting and number of kids (causal?).

Saturated model example

- Washington (AER) data on the effects of daughters.
- We'll look at the relationship between voting and number of kids (causal?).

```
girls <- foreign::read.dta("girls.dta")
head(girls[, c("name", "totchi", "aauw")])
```

```
##                     name totchi aauw
## 1   ABERCROMBIE, NEIL      0  100
## 2   ACKERMAN, GARY L.      3   88
## 3 ADERHOLT, ROBERT B.      0    0
## 4     ALLEN, THOMAS H.     2  100
## 5  ANDREWS, ROBERT E.      2  100
## 6          ARCHER, W.R.    7    0
```

Linear model

```
summary(lm(aauw ~ totchi, data = girls))
```

```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    61.31       1.81   33.81   <2e-16 ***
## totchi         -5.33       0.62   -8.59   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 42 on 1733 degrees of freedom
##   (5 observations deleted due to missingness)
## Multiple R-squared:  0.0408, Adjusted R-squared:  0.0403
## F-statistic: 73.8 on 1 and 1733 DF,  p-value: <2e-16
```

Saturated model

```
summary(lm(aauw ~ as.factor(totchi), data = girls))
```

```
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)            56.41       2.76   20.42  < 2e-16 ***
## as.factor(totchi)1      5.45       4.11    1.33   0.1851
## as.factor(totchi)2     -3.80       3.27   -1.16   0.2454
## as.factor(totchi)3    -13.65       3.45   -3.95  8.1e-05 ***
## as.factor(totchi)4    -19.31       4.01   -4.82  1.6e-06 ***
## as.factor(totchi)5    -15.46       4.85   -3.19   0.0015 **
## as.factor(totchi)6    -33.59      10.42   -3.22   0.0013 **
## as.factor(totchi)7    -17.13      11.41   -1.50   0.1336
## as.factor(totchi)8    -55.33      12.28   -4.51  7.0e-06 ***
## as.factor(totchi)9    -50.41      24.08   -2.09   0.0364 *
## as.factor(totchi)10   -53.41      20.90   -2.56   0.0107 *
## as.factor(totchi)12   -56.41      41.53   -1.36   0.1745
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 41 on 1723 degrees of freedom
##   (5 observations deleted due to missingness)
## Multiple R-squared:  0.0506, Adjusted R-squared:  0.0446
## F-statistic: 8.36 on 11 and 1723 DF,  p-value: 1.84e-14
```

Saturated model minus the constant

```
summary(lm(aauw ~ as.factor(totchi) - 1, data = girls))
```

```
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## as.factor(totchi)0      56.41       2.76   20.42   <2e-16 ***
## as.factor(totchi)1      61.86       3.05   20.31   <2e-16 ***
## as.factor(totchi)2      52.62       1.75   30.13   <2e-16 ***
## as.factor(totchi)3      42.76       2.07   20.62   <2e-16 ***
## as.factor(totchi)4      37.11       2.90   12.79   <2e-16 ***
## as.factor(totchi)5      40.95       3.99   10.27   <2e-16 ***
## as.factor(totchi)6      22.82      10.05    2.27   0.0233 *
## as.factor(totchi)7      39.29      11.07    3.55   0.0004 ***
## as.factor(totchi)8       1.08      11.96    0.09   0.9278
## as.factor(totchi)9       6.00      23.92    0.25   0.8020
## as.factor(totchi)10      3.00      20.72    0.14   0.8849
## as.factor(totchi)12      0.00      41.43    0.00   1.0000
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 41 on 1723 degrees of freedom
##   (5 observations deleted due to missingness)
## Multiple R-squared:  0.587,  Adjusted R-squared:  0.584
## F-statistic:  204 on 12 and 1723 DF,  p-value: <2e-16
```

Compare to within-strata means

- The saturated model makes no assumptions about the between-strata relationships.

Compare to within-strata means

- The saturated model makes no assumptions about the between-strata relationships.
- Just calculates within-strata means:

Compare to within-strata means

- The saturated model makes no assumptions about the between-strata relationships.
- Just calculates within-strata means:

```
c1 <- coef(lm(aauw ~ as.factor(totchi) - 1, data = girls))
c2 <- with(girls, tapply(aauw, totchi, mean, na.rm = TRUE))
rbind(c1, c2)
```

```
##      0  1  2  3  4  5  6  7   8 9 10 12
## c1 56 62 53 43 37 41 23 39 1.1 6  3  0
## c2 56 62 53 43 37 41 23 39 1.1 6  3  0
```

Other justifications for OLS

- **Justification 2**: $X_i'\beta$ is the best linear predictor (in a mean-squared error sense) of Y_i.

Other justifications for OLS

- **Justification 2**: $X_i'\beta$ is the best linear predictor (in a mean-squared error sense) of Y_i.
  - Why?

Other justifications for OLS

- **Justification 2**: $X_i'\beta$ is the best linear predictor (in a mean-squared error sense) of Y_i.
  - Why? $\beta = \arg\min_b \mathbb{E}[(Y_i - X_i'b)^2]$

Other justifications for OLS

- **Justification 2**: $X_i'\beta$ is the best linear predictor (in a mean-squared error sense) of Y_i.
    - Why? $\beta = \arg\min_b \mathbb{E}[(Y_i - X_i'b)^2]$
- **Justification 3**: $X_i'\beta$ provides the minimum mean squared error linear approximation to $E[Y_i|X_i]$.

Other justifications for OLS

- **Justification 2**: $X_i'\beta$ is the best linear predictor (in a mean-squared error sense) of Y_i.
  - Why? $\beta = \arg\min_b \mathbb{E}[(Y_i - X_i'b)^2]$
- **Justification 3**: $X_i'\beta$ provides the minimum mean squared error linear approximation to $E[Y_i|X_i]$.
- Even if the CEF is not linear, a linear regression provides the best linear approximation to that CEF.

Other justifications for OLS

- **Justification 2**: $X_i'\beta$ is the best linear predictor (in a mean-squared error sense) of Y_i.
  - Why? $\beta = \arg\min_b \mathbb{E}[(Y_i - X_i'b)^2]$
- **Justification 3**: $X_i'\beta$ provides the minimum mean squared error linear approximation to $E[Y_i|X_i]$.
- Even if the CEF is not linear, a linear regression provides the best linear approximation to that CEF.
- Don't need to believe the assumptions (linearity) in order to use regression as a good approximation to the CEF.

Other justifications for OLS

- **Justification 2**: $X_i'\beta$ is the best linear predictor (in a mean-squared error sense) of Y_i.
  - Why? $\beta = \arg\min_b \mathbb{E}[(Y_i - X_i'b)^2]$
- **Justification 3**: $X_i'\beta$ provides the minimum mean squared error linear approximation to $E[Y_i|X_i]$.
- Even if the CEF is not linear, a linear regression provides the best linear approximation to that CEF.
- Don't need to believe the assumptions (linearity) in order to use regression as a good approximation to the CEF.
- Warning if the CEF is very nonlinear then this approximation could be terrible!!

The error terms

- Let's define the error term: $e_i \equiv Y_i - X_i'\beta$ so that:

The error terms

- Let's define the error term: $e_i \equiv Y_i - X_i'\beta$ so that:

$$Y_i = X_i'\beta + [Y_i - X_i'\beta] = X_i'\beta + e_i$$

The error terms

- Let's define the error term: $e_i \equiv Y_i - X_i'\beta$ so that:

$$Y_i = X_i'\beta + [Y_i - X_i'\beta] = X_i'\beta + e_i$$

- Note the residual e_i is uncorrelated with X_i:

$$\begin{aligned}
\mathbb{E}[X_i e_i] &= \mathbb{E}[X_i(Y_i - X_i'\beta)] \\
&= \mathbb{E}[X_i Y_i] - \mathbb{E}[X_i X_i'\beta] \\
&= \mathbb{E}[X_i Y_i] - \mathbb{E}\left[X_i X_i' \mathbb{E}[X_i X_i']^{-1}\mathbb{E}[X_i Y_i]\right] \\
&= \mathbb{E}[X_i Y_i] - \mathbb{E}[X_i X_i']\mathbb{E}[X_i X_i']^{-1}\mathbb{E}[X_i Y_i] \\
&= \mathbb{E}[X_i Y_i] - \mathbb{E}[X_i Y_i] = 0
\end{aligned}$$

The error terms

- Let's define the error term: $e_i \equiv Y_i - X_i'\beta$ so that:

$$Y_i = X_i'\beta + [Y_i - X_i'\beta] = X_i'\beta + e_i$$

- Note the residual e_i is uncorrelated with X_i:

$$\begin{aligned}
\mathbb{E}[X_i e_i] &= \mathbb{E}[X_i(Y_i - X_i'\beta)] \\
&= \mathbb{E}[X_i Y_i] - \mathbb{E}[X_i X_i'\beta] \\
&= \mathbb{E}[X_i Y_i] - \mathbb{E}\left[X_i X_i' \mathbb{E}[X_i X_i']^{-1} \mathbb{E}[X_i Y_i]\right] \\
&= \mathbb{E}[X_i Y_i] - \mathbb{E}[X_i X_i'] \mathbb{E}[X_i X_i']^{-1} \mathbb{E}[X_i Y_i] \\
&= \mathbb{E}[X_i Y_i] - \mathbb{E}[X_i Y_i] = 0
\end{aligned}$$

The error terms

- Let's define the error term: $e_i \equiv Y_i - X_i'\beta$ so that:

$$Y_i = X_i'\beta + [Y_i - X_i'\beta] = X_i'\beta + e_i$$

- Note the residual e_i is uncorrelated with X_i:

$$\begin{aligned}
\mathbb{E}[X_i e_i] &= \mathbb{E}[X_i(Y_i - X_i'\beta)] \\
&= \mathbb{E}[X_i Y_i] - \mathbb{E}[X_i X_i'\beta] \\
&= \mathbb{E}[X_i Y_i] - \mathbb{E}\left[X_i X_i' \mathbb{E}[X_i X_i']^{-1} \mathbb{E}[X_i Y_i]\right] \\
&= \mathbb{E}[X_i Y_i] - \mathbb{E}[X_i X_i'] \mathbb{E}[X_i X_i']^{-1} \mathbb{E}[X_i Y_i] \\
&= \mathbb{E}[X_i Y_i] - \mathbb{E}[X_i Y_i] = 0
\end{aligned}$$

- No assumptions on the linearity of $\mathbb{E}[Y_i | X_i]$.

OLS estimator

- We know the population value of β is:

$$\beta = \mathbb{E}[X_i X_i']^{-1}\mathbb{E}[X_i Y_i]$$

OLS estimator

- We know the population value of β is:

$$\beta = \mathbb{E}[X_i X_i']^{-1}\mathbb{E}[X_i Y_i]$$

- How do we get an estimator of this?

OLS estimator

- We know the population value of β is:

$$\beta = \mathbb{E}[X_i X_i']^{-1} \mathbb{E}[X_i Y_i]$$

- How do we get an estimator of this?
- **Plug-in principle** $\rightsquigarrow$ replace population expectation with sample versions:

$$\hat{\beta} = \left[\frac{1}{N} \sum_i X_i X_i' \right]^{-1} \frac{1}{N} \sum_i X_i Y_i$$

OLS estimator

- We know the population value of β is:

$$\beta = \mathbb{E}[X_i X_i']^{-1}\mathbb{E}[X_i Y_i]$$

- How do we get an estimator of this?
- **Plug-in principle** $\rightsquigarrow$ replace population expectation with sample versions:

$$\hat{\beta} = \left[\frac{1}{N}\sum_i X_i X_i'\right]^{-1} \frac{1}{N}\sum_i X_i Y_i$$

- If you work through the matrix algebra, this turns out to be:

$$\hat{\beta} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y}$$

Asymptotic OLS inference

- With this representation in hand, we can write the OLS estimator as follows:

$$\hat{\beta} = \beta + \left[\sum_i X_i X_i'\right]^{-1} \sum_i X_i e_i$$

Asymptotic OLS inference

- With this representation in hand, we can write the OLS estimator as follows:

$$\hat{\beta} = \beta + \left[\sum_i X_i X_i'\right]^{-1} \sum_i X_i e_i$$

- Core idea: $\sum_i X_i e_i$ is the sum of r.v.s so the CLT applies.

Asymptotic OLS inference

- With this representation in hand, we can write the OLS estimator as follows:

$$\hat{\beta} = \beta + \left[\sum_i X_i X_i'\right]^{-1} \sum_i X_i e_i$$

- Core idea: $\sum_i X_i e_i$ is the sum of r.v.s so the CLT applies.
- That, plus some simple asymptotic theory allows us to say:

$$\sqrt{N}(\hat{\beta} - \beta) \rightsquigarrow N(0, \Omega)$$

Asymptotic OLS inference

- With this representation in hand, we can write the OLS estimator as follows:

$$\hat{\beta} = \beta + \left[\sum_i X_i X_i'\right]^{-1} \sum_i X_i e_i$$

- Core idea: $\sum_i X_i e_i$ is the sum of r.v.s so the CLT applies.
- That, plus some simple asymptotic theory allows us to say:

$$\sqrt{N}(\hat{\beta} - \beta) \rightsquigarrow N(0, \Omega)$$

- Converges in distribution to a Normal distribution with mean vector 0 and covariance matrix, Ω:

$$\Omega = \mathbb{E}[X_i X_i']^{-1} \mathbb{E}[X_i X_i' e_i^2] \mathbb{E}[X_i X_i']^{-1}.$$

Asymptotic OLS inference

- With this representation in hand, we can write the OLS estimator as follows:

$$\hat{\beta} = \beta + \left[\sum_i X_i X_i'\right]^{-1} \sum_i X_i e_i$$

- Core idea: $\sum_i X_i e_i$ is the sum of r.v.s so the CLT applies.
- That, plus some simple asymptotic theory allows us to say:

$$\sqrt{N}(\hat{\beta} - \beta) \rightsquigarrow N(0, \Omega)$$

- Converges in distribution to a Normal distribution with mean vector 0 and covariance matrix, Ω:

$$\Omega = \mathbb{E}[X_i X_i']^{-1} \mathbb{E}[X_i X_i' e_i^2] \mathbb{E}[X_i X_i']^{-1}.$$

- No linearity assumption needed!

Estimating the variance

- In large samples then:

$$\sqrt{N}(\hat{\beta} - \beta) \sim N(0, \Omega)$$

Estimating the variance

- In large samples then:

$$\sqrt{N}(\hat{\beta} - \beta) \sim N(0, \Omega)$$

- How to estimate Ω? **Plug-in principle** again!

$$\widehat{\Omega} = \left[\sum_i X_i X_i'\right]^{-1} \left[\sum_i X_i X_i' \hat{e}_i^2\right] \left[\sum_i X_i X_i'\right]^{-1}.$$

Estimating the variance

- In large samples then:

$$\sqrt{N}(\hat{\beta} - \beta) \sim N(0, \Omega)$$

- How to estimate Ω? **Plug-in principle** again!

$$\widehat{\Omega} = \left[\sum_i X_i X_i'\right]^{-1} \left[\sum_i X_i X_i' \hat{e}_i^2\right] \left[\sum_i X_i X_i'\right]^{-1}.$$

- Replace e_i with its emprical counterpart (residuals) $\hat{e}_i = Y_i - X_i'\hat{\beta}$.

Estimating the variance

- In large samples then:

$$\sqrt{N}(\hat{\beta} - \beta) \sim N(0, \Omega)$$

- How to estimate Ω? **Plug-in principle** again!

$$\widehat{\Omega} = \left[\sum_i X_i X_i'\right]^{-1} \left[\sum_i X_i X_i' \hat{e}_i^2\right] \left[\sum_i X_i X_i'\right]^{-1}.$$

- Replace e_i with its emprical counterpart (residuals) $\hat{e}_i = Y_i - X_i'\hat{\beta}$.
- Replace the population moments of X_i with their sample counterparts.

Estimating the variance

- In large samples then:

$$\sqrt{N}(\hat{\beta} - \beta) \sim N(0, \Omega)$$

- How to estimate Ω? **Plug-in principle** again!

$$\widehat{\Omega} = \left[\sum_i X_i X_i'\right]^{-1} \left[\sum_i X_i X_i' \hat{e}_i^2\right] \left[\sum_i X_i X_i'\right]^{-1}.$$

- Replace e_i with its emprical counterpart (residuals) $\hat{e}_i = Y_i - X_i'\hat{\beta}$.
- Replace the population moments of X_i with their sample counterparts.
- The square root of the diagonals of this covariance matrix are the "robust" or Huber-White standard errors.

Agnostic Statistics

- The key insight here is that we can derive estimators under somewhat weaker assumptions

Agnostic Statistics

- The key insight here is that we can derive estimators under somewhat weaker assumptions
- They still rely heavily on large samples (asymptotic results) and independent samples.

Agnostic Statistics

- The key insight here is that we can derive estimators under somewhat weaker assumptions
- They still rely heavily on large samples (asymptotic results) and independent samples.
- See Aronow and Miller for much more.

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:
$$\tau_R = \frac{\text{Cov}(Y_i, D_i - E[D_i|X_i])}{\text{Var}(D_i - E[D_i|X_i])}$$
- $D_i - \mathbb{E}[D_i|X_i]$ is the residual from a regression of D_i on the full set of dummies.

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:

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

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

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

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:

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

- 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
$\mathbb{P}[X_i = 1] = 0.75$	$\mathbb{P}[X_i = 0] = 0.25$
$\mathbb{P}[D_i = 1 \vert X_i = 1] = 0.9$	$\mathbb{P}[D_i = 1 \vert X_i = 0] = 0.5$

OLS weighting example

- Binary covariate:

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

OLS weighting example

- Binary covariate:

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

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 $\tau = 0.5$

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: $\mathbb{E}[\sigma_d^2(X_i)] = 0.13$

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 $X_i = 0$: $0.25/0.13 = 1.92$.

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 $X_i = 0$: $0.25/0.13 = 1.92$.

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

OLS weighting example

- Binary covariate:

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

- 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 $X_i = 0$: $0.25/0.13 = 1.92$.

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

OLS weighting example

- Binary covariate:

$$\begin{aligned} &\text{Group 1} & &\text{Group 2} \\ \mathbb{P}[X_i = 1] &= 0.75 & \mathbb{P}[X_i = 0] &= 0.25 \\ \mathbb{P}[D_i = 1 | X_i = 1] &= 0.9 & \mathbb{P}[D_i = 1 | X_i = 0] &= 0.5 \\ \sigma^2_d(1) &= 0.09 & \sigma^2_d(0) &= 0.25 \\ \tau(1) &= 1 & \tau(0) &= -1 \end{aligned}$$

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

- 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 $X_i = 0$: $0.25/0.13 = 1.92$.

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

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:

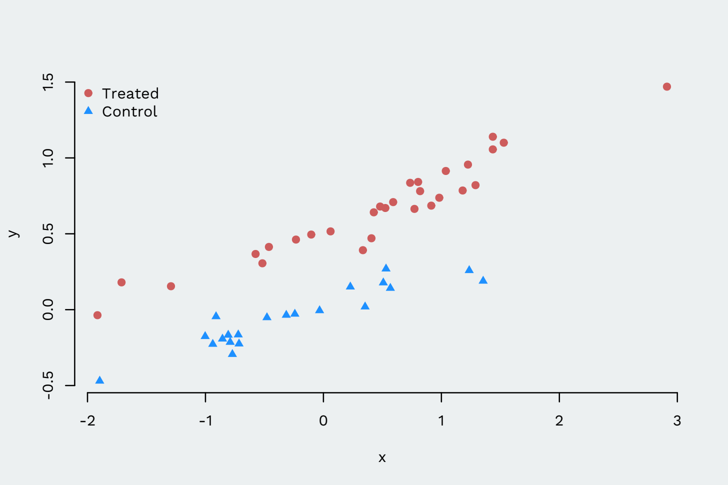
Notes on imputation estimators

- 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

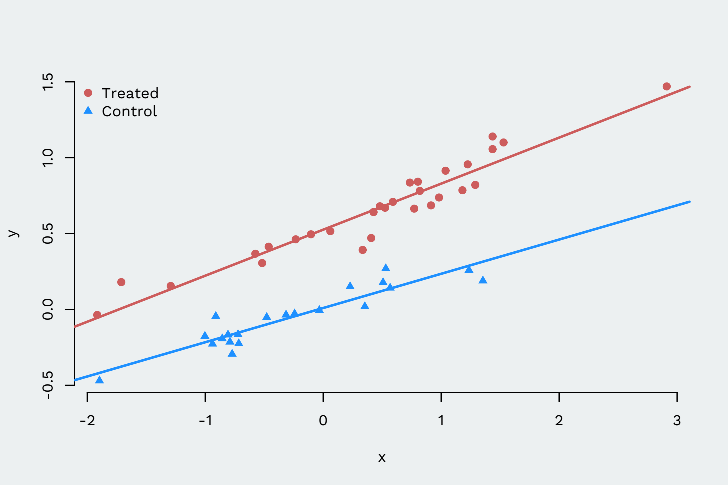
Notes on imputation estimators

- 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)

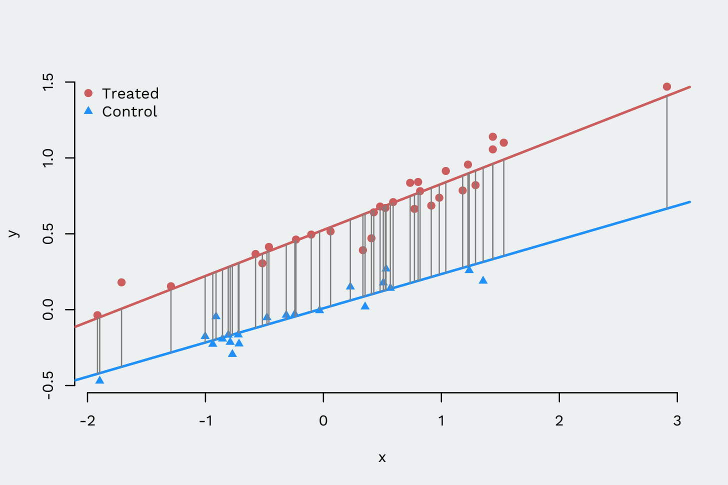
Imputation estimator visualization



Imputation estimator visualization

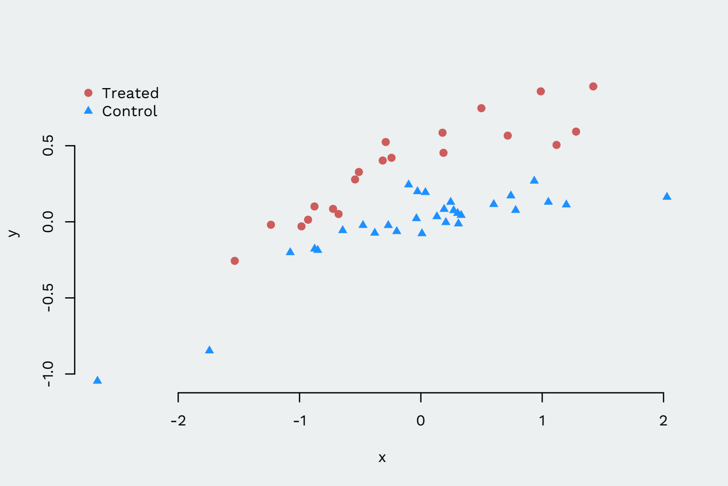


Imputation estimator visualization



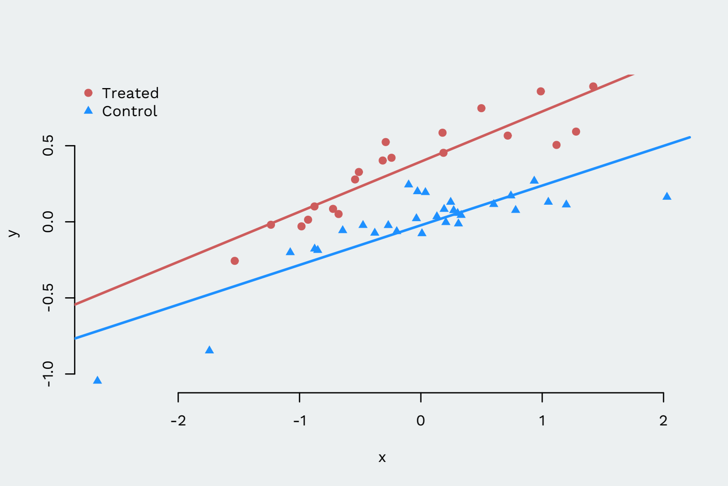
Nonlinear relationships

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



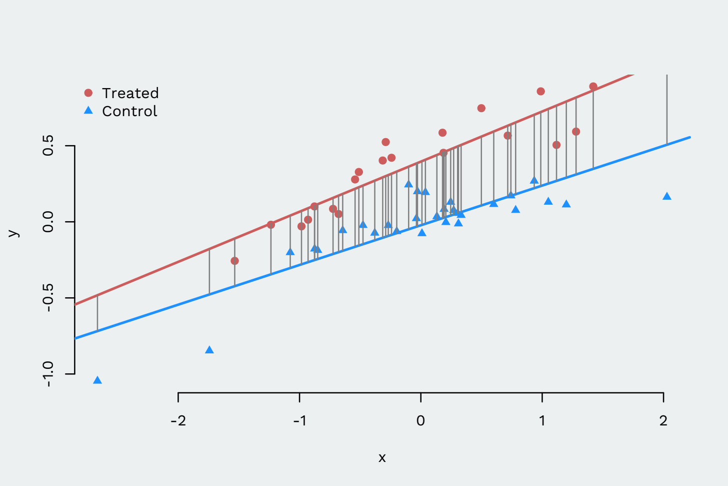
Nonlinear relationships

- 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
##
```

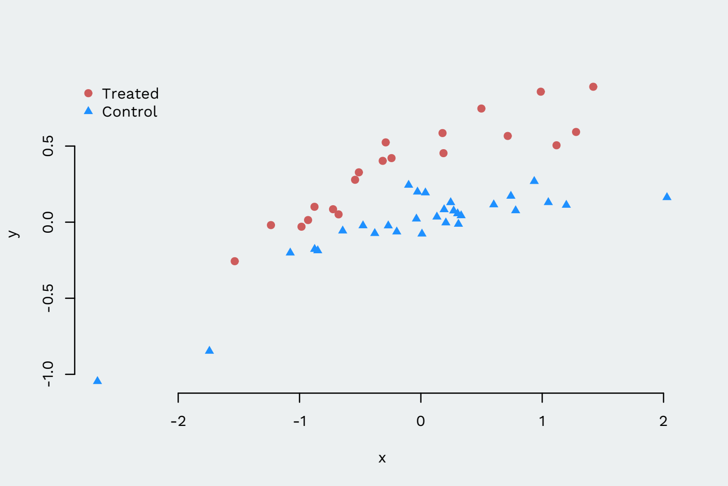
Using semiparametric regression

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

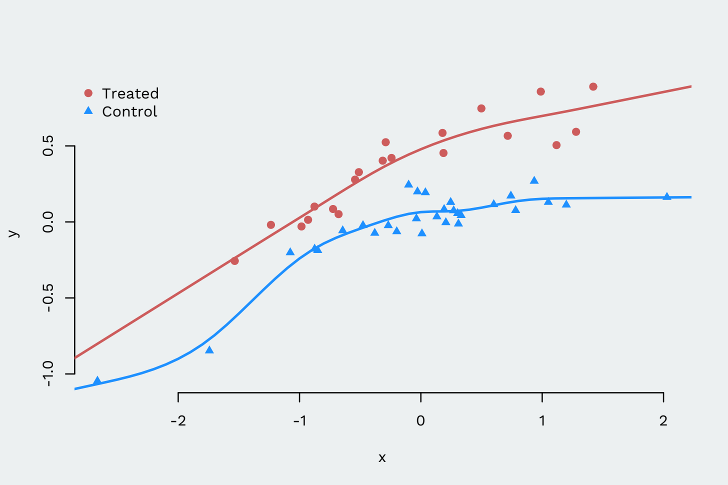
```
library(mgcv)
mod0 <- gam(y ~ s(x), subset = d == 0)
summary(mod0)
```

```
##
## 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
##
```

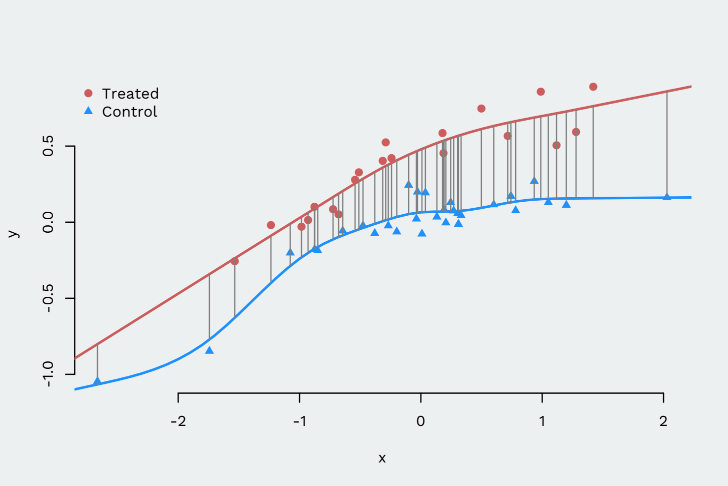
Using GAMs



Using GAMs



Using GAMs



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

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.

Next Week

- Causality with Unmeasured Confounding
- Reading:
  - Fox Chapter 9.8 Instrumental Variables and TSLS
  - Angrist and Pishke Chapter 4 Instrumental Variables
  - Morgan and Winship Chapter 9 Instrumental Variable Estimators of Causal Effects
  - Optional: Hernan and Robins Chapter 16 Instrumental Variable Estimation
  - Optional: Sovey, Allison J. and Green, Donald P. 2011. "Instrumental Variables Estimation in Political Science: A Readers' Guide." *American Journal of Political Science*

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

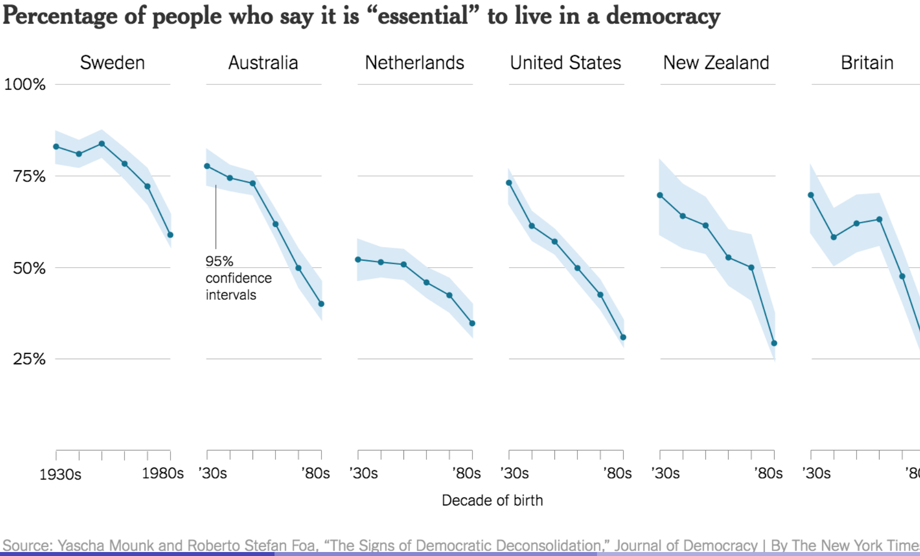
The Danger of Deconsolidation

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.

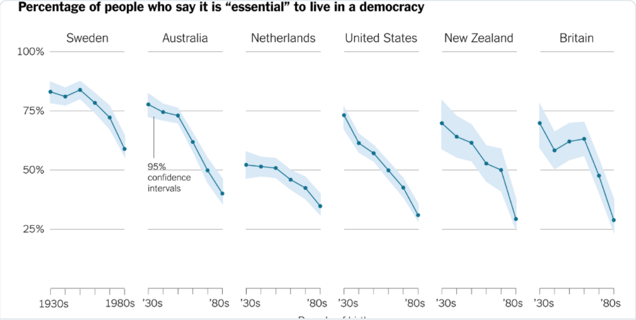
Visualization in the New York Times



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/

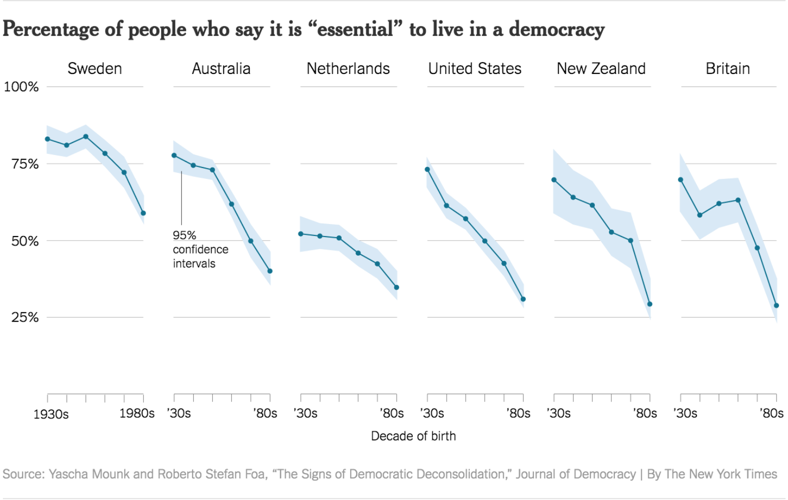


How Stable Are Democracies? 'Warning Signs Are Flashing Red'

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

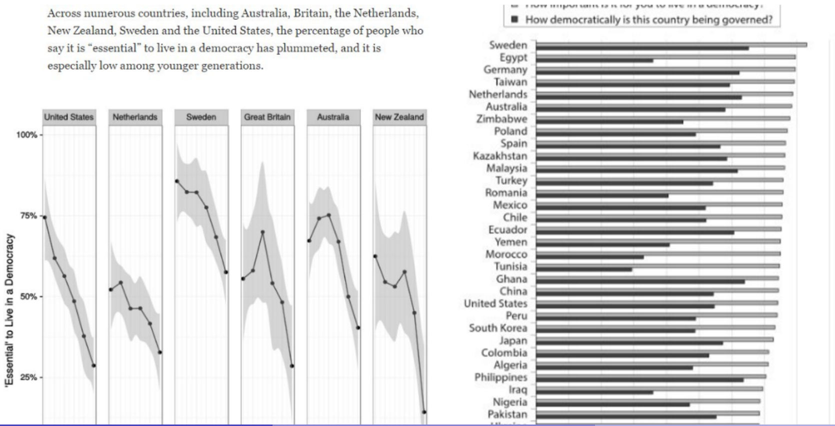
nytimes.com

Alternate Graphs



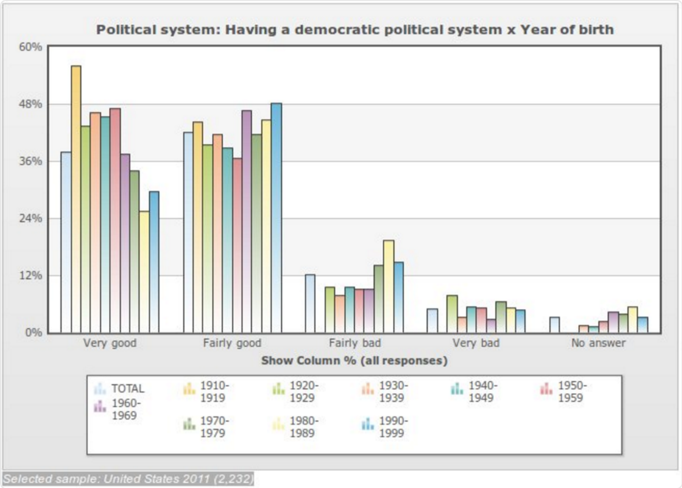
Alternate Graphs

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

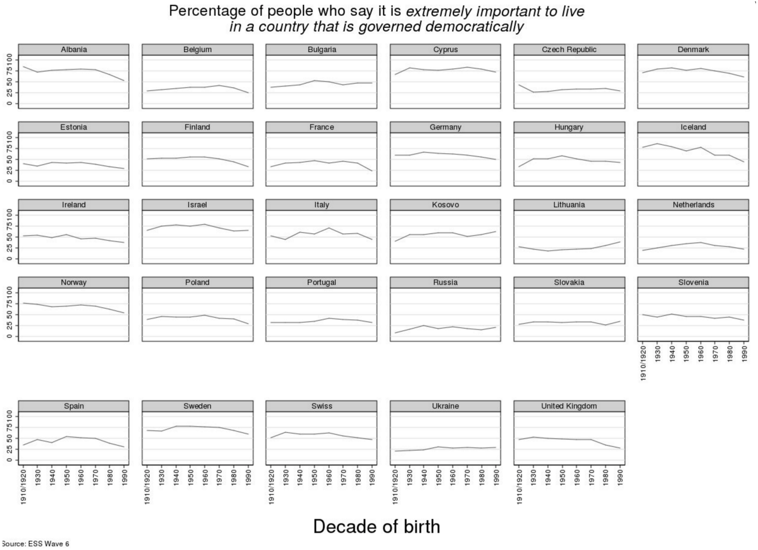


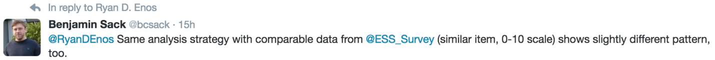
Alternate Graphs

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

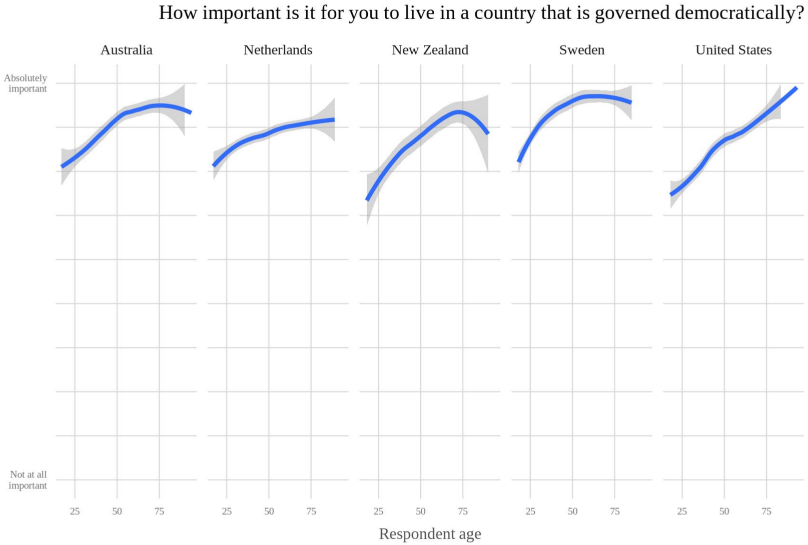


Alternate Graphs



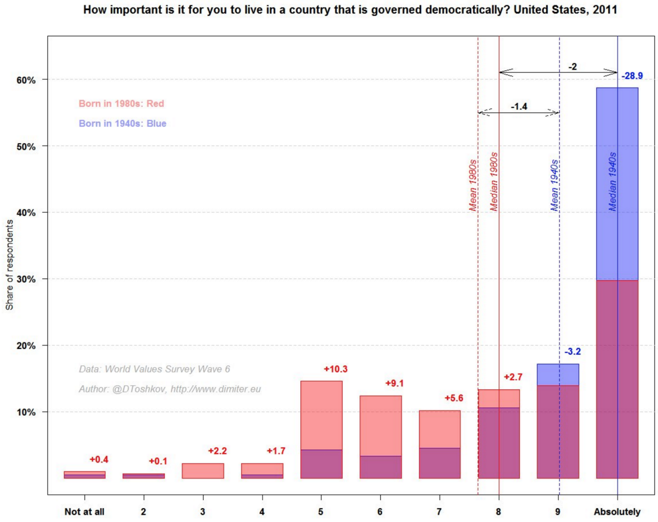


Alternate Graphs



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!)

This Appendix

- The main lecture slides have glossed over some of the details and assumptions for identification
- This appendix contains mathematical results and conditions necessary to estimate causal effects.
- I have also included a section with more details on blocking

Subclassification Estimator

Identification Result

$$\tau_{ATE} = \int \left(\mathbb{E}[Y|X, D=1] - \mathbb{E}[Y|X, D=0] \right) dP(X)$$

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

Assume X takes on K different cells $\{X^1, ..., X^k, ..., X^K\}$. Then the analogy principle suggests estimators:

Subclassification Estimator

Identification Result

$$\tau_{ATE} = \int \left(\mathbb{E}[Y|X, D=1] - \mathbb{E}[Y|X, D=0]\right) dP(X)$$

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

Assume X takes on K different cells $\{X^1, ..., X^k, ..., X^K\}$. Then the analogy principle suggests estimators:

$$\widehat{\tau}_{ATE} = \sum_{k=1}^{K} \left(\bar{Y}_1^k - \bar{Y}_0^k\right) \cdot \left(\frac{N^k}{N}\right); \ \widehat{\tau}_{ATT} = \sum_{k=1}^{K} \left(\bar{Y}_1^k - \bar{Y}_0^k\right) \cdot \left(\frac{N_1^k}{N_1}\right)$$

- N^k is # of obs. and N_1^k is # of treated obs. in cell k
- $\bar{Y}_1^k$ is mean outcome for the treated in cell k
- $\bar{Y}_0^k$ is mean outcome for the untreated in cell k

Subclassification by Age ($K = 2$)

X_k	Death Rate Smokers	Death Rate Non-Smokers	Diff.	# Smokers	# Obs.
Old	28	24	4	3	10
Young	22	16	6	7	10
Total				10	20

What is $\widehat{\tau}_{ATE} = \sum_{k=1}^{K} \left(\bar{Y}_1^k - \bar{Y}_0^k\right) \cdot \left(\frac{N^k}{N}\right)$?

Subclassification by Age ($K = 2$)

X_k	Death Rate Smokers	Death Rate Non-Smokers	Diff.	# Smokers	# Obs.
Old	28	24	4	3	10
Young	22	16	6	7	10
Total				10	20

What is $\widehat{\tau}_{ATE} = \sum_{k=1}^{K} \left(\bar{Y}_1^k - \bar{Y}_0^k\right) \cdot \left(\frac{N^k}{N}\right)$?

$\widehat{\tau}_{ATE} = 4 \cdot (10/20) + 6 \cdot (10/20) = 5$

Subclassification by Age ($K = 2$)

X_k	Death Rate Smokers	Death Rate Non-Smokers	Diff.	# Smokers	# Obs.
Old	28	24	4	3	10
Young	22	16	6	7	10
Total				10	20

What is $\widehat{\tau}_{ATT} = \sum_{k=1}^{K} \left(\bar{Y}_1^k - \bar{Y}_0^k \right) \cdot \left(\frac{N_1^k}{N_1} \right)$?

Subclassification by Age ($K = 2$)

X_k	Death Rate Smokers	Death Rate Non-Smokers	Diff.	# Smokers	# Obs.
Old	28	24	4	3	10
Young	22	16	6	7	10
Total				10	20

What is $\widehat{\tau}_{ATT} = \sum_{k=1}^{K} \left(\bar{Y}_1^k - \bar{Y}_0^k \right) \cdot \left(\frac{N_1^k}{N_1} \right)$?

$\widehat{\tau}_{ATT} = 4 \cdot (3/10) + 6 \cdot (7/10) = 5.4$

Subclassification by Age and Gender ($K = 4$)

X_k	Death Rate Smokers	Death Rate Non-Smokers	Diff.	# Smokers	# Obs.
Old, Male	28	22	4	3	7
Old, Female		24		0	3
Young, Male	21	16	5	3	4
Young, Female	23	17	6	4	6
Total				10	20

What is $\widehat{\tau}_{ATE} = \sum_{k=1}^{K} \left(\bar{Y}_1^k - \bar{Y}_0^k\right) \cdot \left(\frac{N^k}{N}\right)$?

Subclassification by Age and Gender ($K = 4$)

X_k	Death Rate Smokers	Death Rate Non-Smokers	Diff.	# Smokers	# Obs.
Old, Male	28	22	4	3	7
Old, Female		24		0	3
Young, Male	21	16	5	3	4
Young, Female	23	17	6	4	6
Total				10	20

What is $\widehat{\tau}_{ATE} = \sum_{k=1}^{K} \left(\bar{Y}_1^k - \bar{Y}_0^k\right) \cdot \left(\frac{N^k}{N}\right)$?

Not identified!

Subclassification by Age and Gender ($K = 4$)

X_k	Death Rate Smokers	Death Rate Non-Smokers	Diff.	# Smokers	# Obs.
Old, Male	28	22	4	3	7
Old, Female		24		0	3
Young, Male	21	16	5	3	4
Young, Female	23	17	6	4	6
Total				10	20

What is $\widehat{\tau}_{ATT} = \sum_{k=1}^{K} \left(\bar{Y}_1^k - \bar{Y}_0^k \right) \cdot \left(\frac{N_1^k}{N_1} \right)$?

Subclassification by Age and Gender ($K = 4$)

X_k	Death Rate Smokers	Death Rate Non-Smokers	Diff.	# Smokers	# Obs.
Old, Male	28	22	4	3	7
Old, Female		24		0	3
Young, Male	21	16	5	3	4
Young, Female	23	17	6	4	6
Total				10	20

What is $\widehat{\tau}_{ATT} = \sum_{k=1}^{K} \left(\bar{Y}_1^k - \bar{Y}_0^k \right) \cdot \left(\frac{N_1^k}{N_1} \right)$?

$\widehat{\tau}_{ATT} = 4 \cdot (3/10) + 5 \cdot (3/10) + 6 \cdot (4/10) = 5.1$

Selection Bias

Recall the selection problem when comparing the mean outcomes for the treated and the untreated:

Problem

$$\underbrace{\mathbb{E}[Y|D=1] - \mathbb{E}[Y|D=0]}_{\textit{Difference in Means}} = \mathbb{E}[Y_1|D=1] - \mathbb{E}[Y_0|D=0]$$

$$= \underbrace{\mathbb{E}[Y_1 - Y_0|D=1]}_{ATT} + \underbrace{\{\mathbb{E}[Y_0|D=1] - \mathbb{E}[Y_0|D=0]\}}_{BIAS}$$

How can we eliminate the bias term?

Selection Bias

Recall the selection problem when comparing the mean outcomes for the treated and the untreated:

Problem

$$\underbrace{\mathbb{E}[Y|D=1]-\mathbb{E}[Y|D=0]}_{\textit{Difference in Means}} = \mathbb{E}[Y_1|D=1]-\mathbb{E}[Y_0|D=0]$$
$$= \underbrace{\mathbb{E}[Y_1 - Y_0|D=1]}_{ATT} + \underbrace{\{\mathbb{E}[Y_0|D=1]-\mathbb{E}[Y_0|D=0]\}}_{BIAS}$$

How can we eliminate the bias term?

- As a result of randomization, the selection bias term will be zero
- The treatment and control group will tend to be similar along all characteristics (identical in expectation), including the potential outcomes under the control condition

Identification Under Random Assignment

Identification Assumption

$(Y_1, Y_0) \perp\!\!\!\perp D$ *(random assignment)*

Identification Under Random Assignment

Identification Assumption

$(Y_1, Y_0) \perp\!\!\!\perp D$ *(random assignment)*

Identification Result

Problem: $\tau_{ATE} = \mathbb{E}[Y_1 - Y_0]$ *is unobserved. But given random assignment*

$$\begin{aligned} \mathbb{E}[Y|D=1] &= \mathbb{E}[D \cdot Y_1 + (1-D) \cdot Y_0 | D=1] \\ &= \mathbb{E}[Y_1|D=1] \end{aligned}$$

Identification Under Random Assignment

Identification Assumption

$(Y_1, Y_0) \perp\!\!\!\perp D$ *(random assignment)*

Identification Result

Problem: $\tau_{ATE} = \mathbb{E}[Y_1 - Y_0]$ *is unobserved. But given random assignment*

$$\begin{aligned}
\mathbb{E}[Y|D=1] &= \mathbb{E}[D \cdot Y_1 + (1-D) \cdot Y_0 | D=1] \\
&= \mathbb{E}[Y_1 | D=1] \\
&= \mathbb{E}[Y_1]
\end{aligned}$$

$$\mathbb{E}[Y|D=0] =$$

Identification Under Random Assignment

Identification Assumption

$(Y_1, Y_0) \perp\!\!\!\perp D$ *(random assignment)*

Identification Result

Problem: $\tau_{ATE} = \mathbb{E}[Y_1 - Y_0]$ *is unobserved. But given random assignment*

$$\begin{aligned} \mathbb{E}[Y|D=1] &= \mathbb{E}[D \cdot Y_1 + (1-D) \cdot Y_0 | D=1] \\ &= \mathbb{E}[Y_1|D=1] \\ &= \mathbb{E}[Y_1] \end{aligned}$$

$$\begin{aligned} \mathbb{E}[Y|D=0] &= \mathbb{E}[D \cdot Y_1 + (1-D) \cdot Y_0 | D=0] \\ &= \mathbb{E}[Y_0|D=0] \\ &= \mathbb{E}[Y_0] \end{aligned}$$

$$\tau_{ATE} = \mathbb{E}[Y_1 - Y_0] = \mathbb{E}[Y_1] - \mathbb{E}[Y_0] = \underbrace{\mathbb{E}[Y|D=1] - \mathbb{E}[Y|D=0]}_{\textit{Difference in Means}}$$

Average Treatment Effect (ATE)

Imagine a population with 4 units:

i	Y_{1i}	Y_{0i}	Y_i	D_i
1	3	0	3	1
2	1	1	1	1
3	2	0	0	0
4	2	1	1	0

What is $\tau_{ATE} = \mathbb{E}[Y_1] - \mathbb{E}[Y_0]$?

Average Treatment Effect (ATE)

Imagine a population with 4 units:

i	Y_{1i}	Y_{0i}	Y_i	D_i
1	3	0	3	1
2	1	1	1	1
3	2	0	0	0
4	2	1	1	0
$\mathbb{E}[Y_1]$	2			
$\mathbb{E}[Y_0]$		.5		

$$\tau_{ATE} = \mathbb{E}[Y_1] - \mathbb{E}[Y_0] = 2 - .5 = 1.5$$

Average Treatment Effect (ATE)

Imagine a population with 4 units:

i	Y_{1i}	Y_{0i}	Y_i	D_i
1	3	?	3	1
2	1	?	1	1
3	?	0	0	0
4	?	1	1	0
$\mathbb{E}[Y_1]$	?			
$\mathbb{E}[Y_0]$		?		

What is $\tau_{ATE} = \mathbb{E}[Y_1] - \mathbb{E}[Y_0]$?

Average Treatment Effect (ATE)

Imagine a population with 4 units:

i	Y_{1i}	Y_{0i}	Y_i	D_i	$P(D_i = 1)$
1	3	?	3	1	?
2	1	?	1	1	?
3	?	0	0	0	?
4	?	1	1	0	?
$\mathbb{E}[Y_1]$	?				
$\mathbb{E}[Y_0]$		?			

What is $\tau_{ATE} = \mathbb{E}[Y_1] - \mathbb{E}[Y_0]$? In an experiment, the researcher controls the probability of assignment to treatment for all units $P(D_i = 1)$ and by imposing equal probabilities we ensure that treatment assignment is independent of the potential outcomes, i.e. $(Y_1, Y_0) \perp\!\!\!\perp D$.

Average Treatment Effect (ATE)

Imagine a population with 4 units:

i	Y_{1i}	Y_{0i}	Y_i	D_i	$P(D_i = 1)$
1	3	0	3	1	2/4
2	1	1	1	1	2/4
3	2	0	0	0	2/4
4	2	1	1	0	2/4
$\mathbb{E}[Y_1]$	2				
$\mathbb{E}[Y_0]$		.5			

What is $\tau_{ATE} = \mathbb{E}[Y_1] - \mathbb{E}[Y_0]$? Given that D_i is randomly assigned with probability 1/2, we have $\mathbb{E}[Y|D = 1] = \mathbb{E}[Y_1|D = 1] = \mathbb{E}[Y_1]$.

All possible randomizations with two treated units:

Treated Units:	1 & 2	1 & 3	1 & 4	2 & 3	2 & 4	3 & 4
Average $Y|D = 1$:	2	2.5	2.5	1.5	1.5	2

So $\mathbb{E}[Y|D = 1] = \mathbb{E}[Y_1] = 2$

Average Treatment Effect (ATE)

Imagine a population with 4 units:

i	Y_{1i}	Y_{0i}	Y_i	D_i	$P(D_i = 1)$
1	3	0	3	1	$2/4$
2	1	1	1	1	$2/4$
3	2	0	0	0	$2/4$
4	2	1	1	0	$2/4$
$\mathbb{E}[Y_1]$	2				
$\mathbb{E}[Y_0]$		.5			

By the same logic, we have: $\mathbb{E}[Y|D=0] = \mathbb{E}[Y_0|D=0] = \mathbb{E}[Y_0] = .5$.

Therefore the average treatment effect is identified:

$$\tau_{ATE} = \mathbb{E}[Y_1] - \mathbb{E}[Y_0] = \underbrace{\mathbb{E}[Y|D=1] - \mathbb{E}[Y|D=0]}_{\text{Difference in Means}}$$

Average Treatment Effect (ATE)

Imagine a population with 4 units:

i	Y_{1i}	Y_{0i}	Y_i	D_i	$P(D_i = 1)$
1	3	0	3	1	$2/4$
2	1	1	1	1	$2/4$
3	2	0	0	0	$2/4$
4	2	1	1	0	$2/4$
$\mathbb{E}[Y_1]$	2				
$\mathbb{E}[Y_0]$		.5			

Also since $\mathbb{E}[Y|D=0] = \mathbb{E}[Y_0|D=0] = \mathbb{E}[Y_0|D=1] = \mathbb{E}[Y_0]$
we have that

$$\begin{aligned} \tau_{ATT} &= \mathbb{E}[Y_1 - Y_0|D=1] = \mathbb{E}[Y_1|D=1] - \mathbb{E}[Y_0|D=0] \\ &= \mathbb{E}[Y_1] - \mathbb{E}[Y_0] = \mathbb{E}[Y_1 - Y_0] \\ &= \tau_{ATE} \end{aligned}$$

Identification under Random Assignment

Identification Assumption

$(Y_1, Y_0) \perp\!\!\!\perp D$ *(random assignment)*

Identification Result

We have that

$$\mathbb{E}[Y_0|D=0] = \mathbb{E}[Y_0] = \mathbb{E}[Y_0|D=1]$$

and therefore

$$\begin{aligned}\underbrace{E[Y|D=1] - \mathbb{E}[Y|D=0]}_{\text{Difference in Means}} &= \underbrace{E[Y_1 - Y_0|D=1]}_{ATET} + \underbrace{\{\mathbb{E}[Y_0|D=1] - \mathbb{E}[Y_0|D=0]\}}_{BIAS} \\ &= \underbrace{E[Y_1 - Y_0|D=1]}_{ATET}\end{aligned}$$

As a result,

$$\underbrace{E[Y|D=1] - \mathbb{E}[Y|D=0]}_{\text{Difference in Means}} = \tau_{ATE} = \tau_{ATET}$$

Identification in Randomized Experiments

Identification Assumption

Given random assignment $(Y_1, Y_0) \perp\!\!\!\perp D$

Identification Result

Let $F_{Y_d}(y)$ *be the cumulative distribution function (CDF) of* Y_d*, then*

$$\begin{aligned} F_{Y_0}(y) &= \Pr(Y_0 \leq y) = \Pr(Y_0 \leq y | D = 0) \\ &= \Pr(Y \leq y | D = 0). \end{aligned}$$

Similarly,

$$F_{Y_1}(y) = \Pr(Y \leq y | D = 1).$$

So the effect of the treatment at any quantile $\theta \in [0, 1]$ *is identified:*

$$\alpha_\theta = Q_\theta(Y_1) - Q_\theta(Y_0) = Q_\theta(Y | D = 1) - Q_\theta(Y | D = 0)$$

where $F_{Y_d}(Q_\theta(Y_d)) = \theta$.

Estimation Under Random Assignment

Consider a randomized trial with N individuals.

Estimand

$\tau_{ATE} = \mathbb{E}[Y_1 - Y_0] = \mathbb{E}[Y|D=1] - \mathbb{E}[Y|D=0]$

Estimator

Estimation Under Random Assignment

Consider a randomized trial with N individuals.

Estimand

$\tau_{ATE} = \mathbb{E}[Y_1 - Y_0] = \mathbb{E}[Y|D=1] - \mathbb{E}[Y|D=0]$

Estimator

By the analogy principle we use

$$\widehat{\tau} = \bar{Y}_1 - \bar{Y}_0$$

$$\bar{Y}_1 = \frac{\sum Y_i \cdot D_i}{\sum D_i} = \frac{1}{N_1} \sum_{D_i=1} Y_i\,;$$

$$\bar{Y}_0 = \frac{\sum Y_i \cdot (1 - D_i)}{\sum (1 - D_i)} = \frac{1}{N_0} \sum_{D_i=0} Y_i$$

with $N_1 = \sum_i D_i$ *and* $N_0 = N - N_1$.

Under random assignment, $\widehat{\tau}$ *is an unbiased and consistent estimator of* τ_{ATE} *(*$\mathbb{E}[\widehat{\tau}] = \tau_{ATE}$ *and* $\widehat{\tau}_N \xrightarrow{p} \tau_{ATE}$*.)*

Unbiasedness Under Random Assignment

One way of showing that $\widehat{\tau}$ is unbiased is to exploit the fact that under independence of potential outcomes and treatment status, $\mathbb{E}[D] = \frac{N_1}{N}$ and $\mathbb{E}[1 - D] = \frac{N_0}{N}$

Unbiasedness Under Random Assignment

One way of showing that $\widehat{\tau}$ is unbiased is to exploit the fact that under independence of potential outcomes and treatment status, $\mathbb{E}[D] = \frac{N_1}{N}$ and $\mathbb{E}[1 - D] = \frac{N_0}{N}$
Rewrite the estimators as follows:

$$\widehat{\tau} = \frac{1}{N} \sum_{i=1}^{N} \left(\frac{D \cdot Y_1}{N_1/N} - \frac{(1 - D) \cdot Y_0}{N_0/N} \right)$$

Take expectations with respect to the sampling distribution given by the design. Under the Neyman model, Y_1 and Y_0 are fixed and only D_i is random.

Unbiasedness Under Random Assignment

One way of showing that $\widehat{\tau}$ is unbiased is to exploit the fact that under independence of potential outcomes and treatment status, $\mathbb{E}[D] = \frac{N_1}{N}$ and $\mathbb{E}[1-D] = \frac{N_0}{N}$

Rewrite the estimators as follows:

$$\widehat{\tau} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{D \cdot Y_1}{N_1/N} - \frac{(1-D)\cdot Y_0}{N_0/N}\right)$$

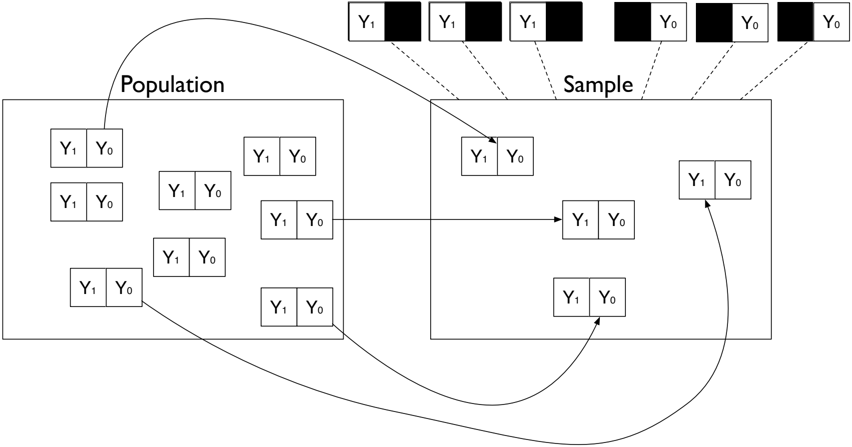
Take expectations with respect to the sampling distribution given by the design. Under the Neyman model, Y_1 and Y_0 are fixed and only D_i is random.

$$\mathbb{E}[\widehat{\tau}] = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{\mathbb{E}[D]\cdot Y_1}{N_1/N} - \frac{\mathbb{E}[(1-D)]\cdot Y_0}{N_0/N}\right) = \frac{1}{N}\sum_{i=1}^{N}(Y_1 - Y_0) = \tau$$

What is the Estimand?

- So far we have emphasized effect estimation, but what about uncertainty?
- In the design based literature, variability in our estimates can arise from two sources:
    1. Sampling variation induced by the procedure that selected the units into our sample.
    2. Variation induced by the particular realization of the treatment variable.
- This distinction is important, but often ignored

What is the Estimand?



SATE and PATE

- Typically we focus on estimating the average causal effect in a particular sample: **S**ample **A**verage **T**reatment **E**ffect (SATE)
  - Uncertainty arises only from hypothetical randomizations.
  - Inferences are limited to the sample in our study.
- Might care about the **P**opulation **A**verage **T**reatment **E**ffect (PATE)
  - Requires precise knowledge about the sampling process that selected units from the population into the sample.
  - Need to account for two sources of variation:
    - ⋆ Variation from the sampling process
    - ⋆ Variation from treatment assignment.
- Thus, in general, $\text{Var}(\widehat{\text{PATE}}) > \text{Var}(\widehat{\text{SATE}})$.

Standard Error for Sample ATE

The standard error is the standard deviation of a sampling distribution:
$SE_{\widehat{\theta}} \equiv \sqrt{\frac{1}{J}\sum_1^J(\widehat{\theta}_j - \overline{\widehat{\theta}})^2}$ (with J possible random assignments).

i	Y_{1i}	Y_{0i}	Y_i	D_i	$P(D_i = 1)$
1	3	0	3	1	2/4
2	1	1	1	1	2/4
3	2	0	0	0	2/4
4	2	1	1	0	2/4

ATE estimates given all possible random assignments with two treated units:

Treated Units:	1 & 2	1 & 3	1 & 4	2 & 3	2 & 4	3 & 4
$\widehat{ATE}$:	1.5	1.5	2	1	1.5	1.5

The average $\widehat{ATE}$ is 1.5 and therefore the true standard error is
$SE_{\widehat{ATE}} = \sqrt{\frac{1}{6}[(1.5-1.5)^2 + (1.5-1.5)^2 + (2-1.5)^2 + (1-1.5)^2 + (1.5-1.5)^2 + (1.5-1.5)^2]} \approx .28$

Standard Error for Sample ATE

Standard Error for Sample ATE

Given complete randomization of N units with N_1 assigned to treatment and $N_0 = N - N_1$ to control, the true standard error of the estimated sample ATE is given by

$$SE_{\widehat{ATE}} = \sqrt{\left(\frac{N-N_1}{N-1}\right)\frac{Var[Y_{1i}]}{N_1} + \left(\frac{N-N_0}{N-1}\right)\frac{Var[Y_{0i}]}{N_0} + \left(\frac{1}{N-1}\right)2Cov[Y_{1i}, Y_{0i}]}$$

with population variances and covariance

$$Var[Y_{di}] \equiv \frac{1}{N}\sum_{1}^{N}\left(Y_{di} - \frac{\sum_1^N Y_{di}}{N}\right)^2 = \sigma^2_{Y_d|D_i=d}$$

$$Cov[Y_{1i}, Y_{0i}] \equiv \frac{1}{N}\sum_{1}^{N}\left(Y_{1i} - \frac{\sum_1^N Y_{1i}}{N}\right)\left(Y_{0i} - \frac{\sum_1^N Y_{0i}}{N}\right) = \sigma^2_{Y_1,Y_0}$$

Standard Error for Sample ATE

Standard Error for Sample ATE

Given complete randomization of N units with N_1 assigned to treatment and $N_0 = N - N_1$ to control, the true standard error of the estimated sample ATE is given by

$$SE_{\widehat{ATE}} = \sqrt{\left(\frac{N-N_1}{N-1}\right)\frac{Var[Y_{1i}]}{N_1} + \left(\frac{N-N_0}{N-1}\right)\frac{Var[Y_{0i}]}{N_0} + \left(\frac{1}{N-1}\right)2Cov[Y_{1i}, Y_{0i}]}$$

with population variances and covariance

$$Var[Y_{di}] \equiv \frac{1}{N}\sum_1^N \left(Y_{di} - \frac{\sum_1^N Y_{di}}{N}\right)^2 = \sigma^2_{Y_d|D_i=d}$$

$$Cov[Y_{1i}, Y_{0i}] \equiv \frac{1}{N}\sum_1^N \left(Y_{1i} - \frac{\sum_1^N Y_{1i}}{N}\right)\left(Y_{0i} - \frac{\sum_1^N Y_{0i}}{N}\right) = \sigma^2_{Y_1,Y_0}$$

Plugging in, we obtain the true standard error of the estimated sample ATE

$$SE_{\widehat{ATE}} = \sqrt{\left(\frac{4-2}{4-1}\right)\frac{.25}{2} + \left(\frac{4-2}{4-1}\right)\frac{.5}{2} + \left(\frac{1}{4-1}\right)2(-.25)} \approx .28$$

Standard Error for Sample ATE

Standard Error for Sample ATE

Given complete randomization of N units with N_1 assigned to treatment and $N_0 = N - N_1$ to control, the true standard error of the estimated sample ATE is given by

$$SE_{\widehat{ATE}} = \sqrt{\left(\frac{N-N_1}{N-1}\right)\frac{Var[Y_{1i}]}{N_1} + \left(\frac{N-N_0}{N-1}\right)\frac{Var[Y_{0i}]}{N_0} + \left(\frac{1}{N-1}\right)2Cov[Y_{1i}, Y_{0i}]}$$

with population variances and covariance

$$Var[Y_{di}] \equiv \frac{1}{N}\sum_1^N \left(Y_{di} - \frac{\sum_1^N Y_{di}}{N}\right)^2 = \sigma^2_{Y_d|D_i=d}$$

$$Cov[Y_{1i}, Y_{0i}] \equiv \frac{1}{N}\sum_1^N \left(Y_{1i} - \frac{\sum_1^N Y_{1i}}{N}\right)\left(Y_{0i} - \frac{\sum_1^N Y_{0i}}{N}\right) = \sigma^2_{Y_1,Y_0}$$

Standard error decreases if:

- N grows
- $Var[Y_1]$, $Var[Y_0]$ decrease
- $Cov[Y_1, Y_0]$ decreases

Conservative Estimator $\hat{SE}_{\widehat{ATE}}$

Conservative Estimator for Standard Error for Sample ATE

$$\widehat{SE}_{\widehat{ATE}} = \sqrt{\frac{\widehat{Var[Y_{1i}]}}{N_1} + \frac{\widehat{Var[Y_{0i}]}}{N_0}}$$

with estimators of the sample variances given by

$$\widehat{Var[Y_{1i}]} \equiv \frac{1}{N_1 - 1} \sum_{i|D_i=1}^{N} \left(Y_{1i} - \frac{\sum_{i|D_i=1}^{N} Y_{1i}}{N_1} \right)^2 = \widehat{\sigma}^2_{Y|D_i=1}$$

$$\widehat{Var[Y_{0i}]} \equiv \frac{1}{N_0 - 1} \sum_{i|D_i=0}^{N} \left(Y_{0i} - \frac{\sum_{i|D_i=0}^{N} Y_{0i}}{N_0} \right)^2 = \widehat{\sigma}^2_{Y|D_i=0}$$

Conservative Estimator $\hat{SE}_{\widehat{ATE}}$

Conservative Estimator for Standard Error for Sample ATE

$$\widehat{SE}_{\widehat{ATE}} = \sqrt{\frac{\widehat{Var[Y_{1i}]}}{N_1} + \frac{\widehat{Var[Y_{0i}]}}{N_0}}$$

with estimators of the sample variances given by

$$\widehat{Var[Y_{1i}]} \equiv \frac{1}{N_1 - 1} \sum_{i|D_i=1}^{N} \left(Y_{1i} - \frac{\sum_{i|D_i=1}^{N} Y_{1i}}{N_1} \right)^2 = \widehat{\sigma}^2_{Y|D_i=1}$$

$$\widehat{Var[Y_{0i}]} \equiv \frac{1}{N_0 - 1} \sum_{i|D_i=0}^{N} \left(Y_{0i} - \frac{\sum_{i|D_i=0}^{N} Y_{0i}}{N_0} \right)^2 = \widehat{\sigma}^2_{Y|D_i=0}$$

What about the covariance?

Conservative Estimator $\hat{SE}_{\widehat{ATE}}$

Conservative Estimator for Standard Error for Sample ATE

$$\widehat{SE}_{\widehat{ATE}} = \sqrt{\frac{\widehat{Var[Y_{1i}]}}{N_1} + \frac{\widehat{Var[Y_{0i}]}}{N_0}}$$

with estimators of the sample variances given by

$$\widehat{Var[Y_{1i}]} \equiv \frac{1}{N_1 - 1} \sum_{i|D_i=1}^{N} \left(Y_{1i} - \frac{\sum_{i|D_i=1}^{N} Y_{1i}}{N_1} \right)^2 = \widehat{\sigma}^2_{Y|D_i=1}$$

$$\widehat{Var[Y_{0i}]} \equiv \frac{1}{N_0 - 1} \sum_{i|D_i=0}^{N} \left(Y_{0i} - \frac{\sum_{i|D_i=0}^{N} Y_{0i}}{N_0} \right)^2 = \widehat{\sigma}^2_{Y|D_i=0}$$

- Conservative compared to the true standard error, i.e. $SE_{\widehat{ATE}} < \widehat{SE}_{\widehat{ATE}}$
- Asymptotically unbiased in two special cases:
- if τ_i is constant (i.e. $Cor[Y_1, Y_0] = 1$)
- if we estimate standard error of population average treatment effect ($Cov[Y_1, Y_0]$ is negligible when we sample from a large population)
- Equivalent to standard error for two sample t-test with unequal variances or "robust" standard error in regression of Y on D

Proof: $SE_{\widehat{ATE}} \leq \hat{SE}_{\widehat{ATE}}$

Upper bound for standard error is when $Cor[Y_1, Y_0] = 1$:

$$Cor[Y_1, Y_0] = \frac{Cov[Y_1, Y_0]}{\sqrt{Var[Y_1]Var[Y_0]}} \leq 1 \iff Cov[Y_1, Y_0] \leq \sqrt{Var[Y_1]Var[Y_0]}$$

$$\begin{aligned}
SE_{\widehat{ATE}} &= \sqrt{\left(\frac{N-N_1}{N-1}\right)\frac{Var[Y_1]}{N_1} + \left(\frac{N-N_0}{N-1}\right)\frac{Var[Y_0]}{N_0} + \left(\frac{1}{N-1}\right)2Cov[Y_1, Y_0]} \\
&= \sqrt{\frac{1}{N-1}\left(\frac{N_0}{N_1}Var[Y_1] + \frac{N_1}{N_0}Var[Y_0] + 2Cov[Y_1, Y_0]\right)} \\
&\leq \sqrt{\frac{1}{N-1}\left(\frac{N_0}{N_1}Var[Y_1] + \frac{N_1}{N_0}Var[Y_0] + 2\sqrt{Var[Y_1]Var[Y_0]}\right)} \\
&\leq \sqrt{\frac{1}{N-1}\left(\frac{N_0}{N_1}Var[Y_1] + \frac{N_1}{N_0}Var[Y_0] + Var[Y_1] + Var[Y_0]\right)}
\end{aligned}$$

Last step follows from the following inequality

$$\begin{aligned}
(\sqrt{Var[Y_1]} - \sqrt{Var[Y_0]})^2 &\geq 0 \\
Var[Y_1] - 2\sqrt{Var[Y_1]Var[Y_0]} + Var[Y_0] &\geq 0 \iff Var[Y_1] + Var[Y_0] \geq 2\sqrt{Var[Y_1]Var[Y_0]}
\end{aligned}$$

Proof: $SE_{\widehat{ATE}} \leq \hat{SE}_{\widehat{ATE}}$

$$
\begin{aligned}
SE_{\widehat{ATE}} &\leq \sqrt{\frac{1}{N-1}\left(\frac{N_0}{N_1}Var[Y_1] + \frac{N_1}{N_0}Var[Y_0] + Var[Y_1] + Var[Y_0]\right)} \\
&\leq \sqrt{\frac{N_0^2 Var[Y_1] + N_1^2 Var[Y_0] + N_1 N_0 (Var[Y_1] + Var[Y_0])}{(N-1)N_1 N_0}} \\
&\leq \sqrt{\frac{(N_0^2 + N_1 N_0) Var[Y_1] + (N_1^2 + N_1 N_0) Var[Y_0]}{(N-1)N_1 N_0}} \\
&\leq \sqrt{\frac{(N_0 + N_1) N_0 Var[Y_1]}{(N-1)N_1 N_0} + \frac{(N_1 + N_0) N_1 Var[Y_0]}{(N-1)N_1 N_0}} \\
&\leq \sqrt{\frac{N\, Var[Y_1]}{(N-1)N_1} + \frac{N\, Var[Y_0]}{(N-1)N_0}} \\
&\leq \sqrt{\frac{N}{N-1}\left(\frac{Var[Y_1]}{N_1} + \frac{Var[Y_0]}{(N_0)}\right)} \\
&\leq \sqrt{\frac{N}{N-1}\left(\frac{\widehat{Var[Y_1]}}{N_1} + \frac{\widehat{Var[Y_0]}}{(N_0)}\right)}
\end{aligned}
$$

So the estimator for the standard error is conservative.

Standard Error for Sample ATE

i	Y_{1i}	Y_{0i}	Y_i
1	3	0	3
2	1	1	1
3	2	0	0
4	2	1	1

$\widehat{SE}_{\widehat{ATE}}$ estimates given all possible assignments with two treated units:

Treated Units:	1 & 2	1 & 3	1 & 4	2 & 3	2 & 4	3 & 4
$\widehat{ATE}$:	1.5	1.5	2	1	1.5	1.5
$\widehat{SE}_{\widehat{ATE}}$:	1.11	.5	.71	.71	.5	.5

The average $\widehat{SE}_{\widehat{ATE}}$ is $\approx .67$ compared to the true standard error of $SE_{\widehat{ATE}} \approx .28$

Example: Effect of Training on Earnings

- Treatment Group:
  - $N_1 = 7,487$
  - Estimated Average Earnings $\bar{Y}_1$: \$16,199
  - Estimated Sample Standard deviation $\widehat{\sigma}_{Y|D_i=1}$: \$17,038
- Control Group :
  - $N_0 = 3,717$
  - Estimated Average Earnings $\bar{Y}_0$: \$15,040
  - Estimated Sample deviation $\widehat{\sigma}_{Y|D_i=0}$: \$16,180
- Estimated average effect of training:

Example: Effect of Training on Earnings

- Treatment Group:
  - $N_1 = 7,487$
  - Estimated Average Earnings $\bar{Y}_1$: \$16,199
  - Estimated Sample Standard deviation $\widehat{\sigma}_{Y|D_i=1}$: \$17,038
- Control Group :
  - $N_0 = 3,717$
  - Estimated Average Earnings $\bar{Y}_0$: \$15,040
  - Estimated Sample deviation $\widehat{\sigma}_{Y|D_i=0}$: \$16,180
- Estimated average effect of training:
  - $\widehat{\tau}_{ATE} = \bar{Y}_1 - \bar{Y}_0 = 16,199 - 15,040 = \$1,159$
- Estimated standard error for effect of training:

Example: Effect of Training on Earnings

- Treatment Group:
  - $N_1 = 7,487$
  - Estimated Average Earnings $\bar{Y}_1$: \$16,199
  - Estimated Sample Standard deviation $\widehat{\sigma}_{Y|D_i=1}$: \$17,038
- Control Group :
  - $N_0 = 3,717$
  - Estimated Average Earnings $\bar{Y}_0$: \$15,040
  - Estimated Sample deviation $\widehat{\sigma}_{Y|D_i=0}$: \$16,180
- Estimated average effect of training:
  - $\widehat{\tau}_{ATE} = \bar{Y}_1 - \bar{Y}_0 = 16,199 - 15,040 = \$1,159$
- Estimated standard error for effect of training:
  - $\widehat{SE}_{\widehat{ATE}} = \sqrt{\frac{\widehat{\sigma}^2_{Y|D_i=1}}{N_1} + \frac{\widehat{\sigma}^2_{Y|D_i=0}}{(N_0)}} = \sqrt{\frac{17{,}038^2}{7{,}487} + \frac{16{,}180^2}{3{,}717}} \approx \$330$
- Is this consistent with a zero average treatment effect $\alpha_{ATE} = 0$?

Testing the Null Hypothesis of Zero Average Effect

- Under the null hypothesis H_0: $\tau_{ATE} = 0$, the average potential outcomes in the population are the same for treatment and control: $\mathbb{E}[Y_1] = \mathbb{E}[Y_0]$.
- Since units are randomly assigned, both the treatment and control groups should therefore have the same sample average earnings
- However, we in fact observe a difference in mean earnings of $\$1,159$
- What is the probability of observing a difference this large if the true average effect of the training were zero (i.e. the null hypothesis were true)?

Testing the Null Hypothesis of Zero Average Effect

- Use a two-sample t-test with unequal variances:

$$t = \frac{\widehat{\tau}}{\sqrt{\frac{\widehat{\sigma}^2_{Y_i|D_i=1}}{N_1} + \frac{\widehat{\sigma}^2_{Y_i|D_i=0}}{N_0}}} = \frac{\$1,159}{\sqrt{\frac{\$17,038^2}{7,487} + \frac{\$16,180^2}{3,717}}} \approx 3.5$$

  - From basic statistical theory, we know that $t_N \xrightarrow{d} \mathcal{N}(0,1)$
  - And for a standard normal distribution, the probability of observing a value of t that is larger than $|t| > 1.96$ is $< .05$
  - So obtaining a value as high as $t = 3.5$ is very unlikely under the null hypothesis of a zero average effect
  - We reject the null hypothesis H_0: $\tau_0 = 0$ against the alternative H_1: $\tau_0 \neq 0$ at asymptotic 5% significance level whenever $|t| > 1.96$.
  - Inverting the test statistic we can construct a 95% confidence interval

$$\widehat{\tau}_{ATE} \pm 1.96 \cdot \widehat{SE}_{\widehat{ATE}}$$

Testing the Null Hypothesis of Zero Average Effect

R Code

```
> d <- read.dta("jtpa.dta")
> head(d[,c("earnings","assignmt")])
  earnings assignmt
1     1353        1
2     4984        1
3    27707        1
4    31860        1
5    26615        0
>
> meanAsd <- function(x){
+   out <- c(mean(x),sd(x))
+   names(out) <- c("mean","sd")
+   return(out)
+ }
>
> aggregate(earnings~assignmt,data=d,meanAsd)
  assignmt earnings.mean earnings.sd
1        0      15040.50    16180.25
2        1      16199.94    17038.85
```

Testing the Null Hypothesis of Zero Average Effect

R Code

```
> t.test(earnings~assignmt,data=d,var.equal=FALSE)

        Welch Two Sample t-test

data:  earnings by assignmt
t = -3.5084, df = 7765.599, p-value = 0.0004533
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1807.2427  -511.6239
sample estimates:
mean in group 0 mean in group 1
       15040.50        16199.94
```

Regression to Estimate the Average Treatment Effect

Estimator (Regression)

The ATE can be expressed as a regression equation:

$$
\begin{aligned}
Y_i &= D_i\, Y_{1i} + (1 - D_i)\, Y_{0i} \\
&= Y_{0i} + (Y_{1i} - Y_{0i})\, D_i \\
&= \underbrace{\bar{Y}_0}_{\alpha} + \underbrace{(\bar{Y}_1 - \bar{Y}_0)}_{\tau_{Reg}} D_i + \underbrace{\{(Y_{i0} - \bar{Y}_0) + D_i \cdot [(Y_{i1} - \bar{Y}_1) - (Y_{i0} - \bar{Y}_0)]\}}_{\epsilon} \\
&= \alpha + \tau_{Reg} D_i + \epsilon_i
\end{aligned}
$$

- τ_{Reg} could be biased for τ_{ATE} in two ways:

Regression to Estimate the Average Treatment Effect

Estimator (Regression)

The ATE can be expressed as a regression equation:

$$\begin{aligned} Y_i &= D_i\, Y_{1i} + (1 - D_i)\, Y_{0i} \\ &= Y_{0i} + (Y_{1i} - Y_{0i})\, D_i \\ &= \underbrace{\bar{Y}_0}_{\alpha} + \underbrace{(\bar{Y}_1 - \bar{Y}_0)}_{\tau_{Reg}} D_i + \underbrace{\{(Y_{i0} - \bar{Y}_0) + D_i \cdot [(Y_{i1} - \bar{Y}_1) - (Y_{i0} - \bar{Y}_0)]\}}_{\epsilon} \\ &= \alpha + \tau_{Reg} D_i + \epsilon_i \end{aligned}$$

- τ_{Reg} could be biased for τ_{ATE} in two ways:
  - Baseline difference in potential outcomes under control that is correlated with D_i.
  - Individual treatment effects τ_i are correlated with D_i
  - Under random assignment, both correlations are zero in expectation

Regression to Estimate the Average Treatment Effect

Estimator (Regression)

The ATE can be expressed as a regression equation:

$$\begin{aligned}
Y_i &= D_i\, Y_{1i} + (1 - D_i)\, Y_{0i} \\
&= Y_{0i} + (Y_{1i} - Y_{0i})\, D_i \\
&= \underbrace{\bar{Y}_0}_{\alpha} + \underbrace{(\bar{Y}_1 - \bar{Y}_0)}_{\tau_{Reg}} D_i + \underbrace{\{(Y_{i0} - \bar{Y}_0) + D_i \cdot [(Y_{i1} - \bar{Y}_1) - (Y_{i0} - \bar{Y}_0)]\}}_{\epsilon} \\
&= \alpha + \tau_{Reg}\, D_i + \epsilon_i
\end{aligned}$$

- τ_{Reg} could be biased for τ_{ATE} in two ways:
  - Baseline difference in potential outcomes under control that is correlated with D_i.
  - Individual treatment effects τ_i are correlated with D_i
  - Under random assignment, both correlations are zero in expectation
- Effect heterogeneity implies "heteroskedasticity", i.e. error variance differs by values of D_i.
  - Neyman model imples "robust" standard errors.
- Can use regression in experiments without assuming constant effects.

Regression to Estimate the Average Treatment Effect

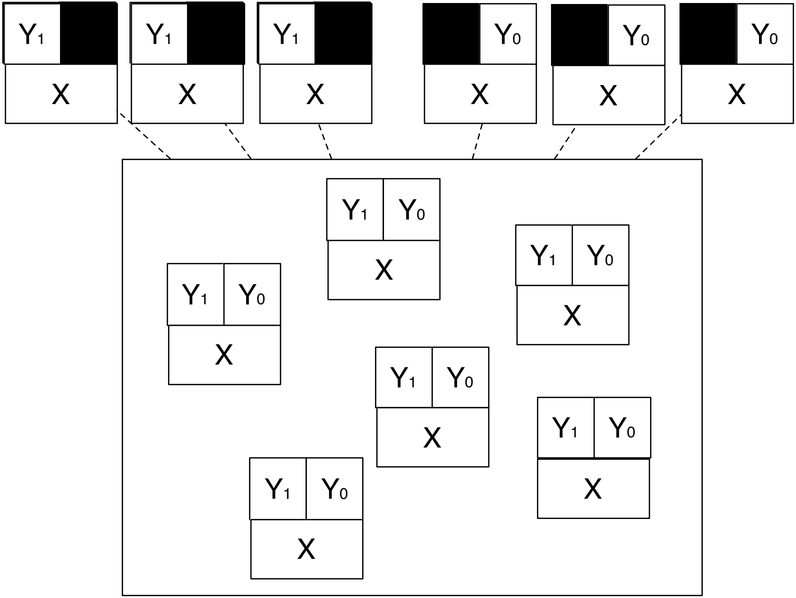
R Code

```
> library(sandwich)
> library(lmtest)
>
> lout <- lm(earnings~assignmt,data=d)
> coeftest(lout,vcov = vcovHC(lout, type = "HC1")) # matches Stata

t test of coefficients:

            Estimate Std. Error t value  Pr(>|t|)
(Intercept) 15040.50     265.38 56.6752 < 2.2e-16 ***
assignmt     1159.43     330.46  3.5085 0.0004524 ***
---
```

Covariates and Experiments



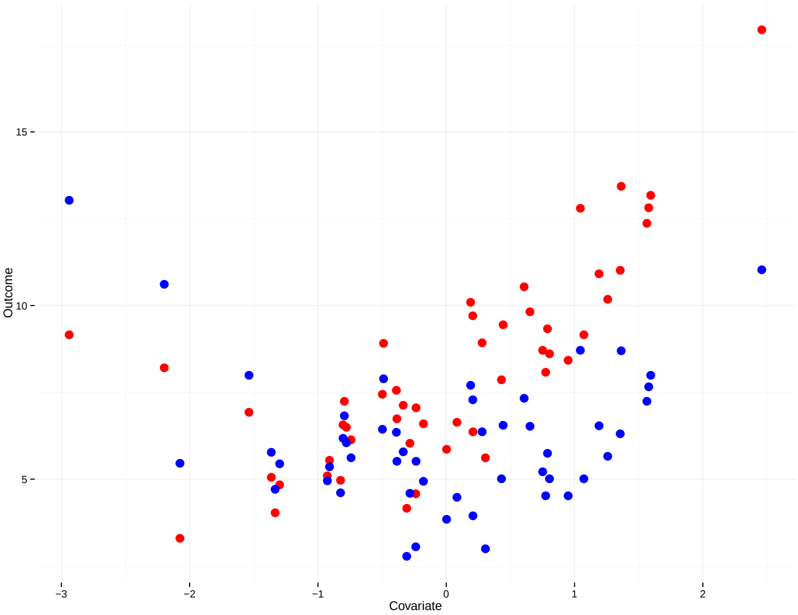
Covariates

- Randomization is gold standard for causal inference because in expectation it balances observed but also unobserved characteristics between treatment and control group.
- Unlike potential outcomes, you observe baseline covariates for all units. Covariate values are predetermined with respect to the treatment and do not depend on D_i.
- Under randomization, $f_{X|D}(X|D=1) \stackrel{d}{=} f_{X|D}(X|D=0)$ (equality in distribution).
- Similarity in distributions of covariates is known as covariate balance.
- If this is not the case, then one of two possibilities:

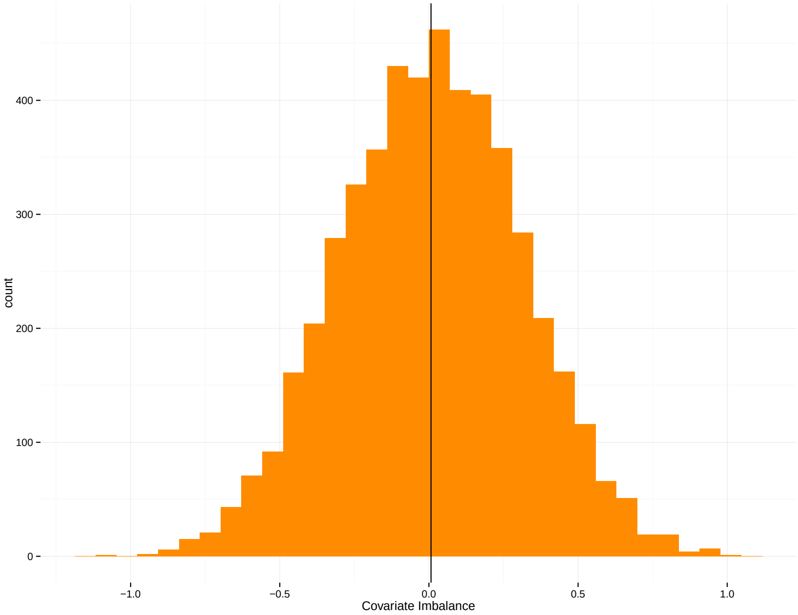
Covariates

- Randomization is gold standard for causal inference because in expectation it balances observed but also unobserved characteristics between treatment and control group.
- Unlike potential outcomes, you observe baseline covariates for all units. Covariate values are predetermined with respect to the treatment and do not depend on D_i.
- Under randomization, $f_{X|D}(X|D=1) \stackrel{d}{=} f_{X|D}(X|D=0)$ (equality in distribution).
- Similarity in distributions of covariates is known as covariate balance.
- If this is not the case, then one of two possibilities:
  - Randomization was compromised.
  - Sampling error (bad luck)
- One should always test for covariate balance on important covariates, using so called "balance checks" (eg. t-tests, F-tests, etc.)

Covariates and Experiments



Covariates and Experiments



Regression with Covariates

- Practioners often run some variant of the following model with experimental data:

$$Y_i = \alpha + \tau D_i + X_i\beta + \epsilon_i$$

- Why include X_i when experiments "control" for covariates by design?

Regression with Covariates

- Practioners often run some variant of the following model with experimental data:

$$Y_i = \alpha + \tau D_i + X_i\beta + \epsilon_i$$

- Why include X_i when experiments “control” for covariates by design?
  - Correct for chance covariate imbalances that indicate that $\hat{\tau}$ may be far from τ_{ATE}.

Regression with Covariates

- Practioners often run some variant of the following model with experimental data:

$$Y_i = \alpha + \tau D_i + X_i\beta + \epsilon_i$$

- Why include X_i when experiments "control" for covariates by design?
    - Correct for chance covariate imbalances that indicate that $\hat{\tau}$ may be far from τ_{ATE}.
    - Increase precision: remove variation in the outcome accounted for by pre-treatment characteristics, thus making it easier to attribute remaining differences to the treatment.

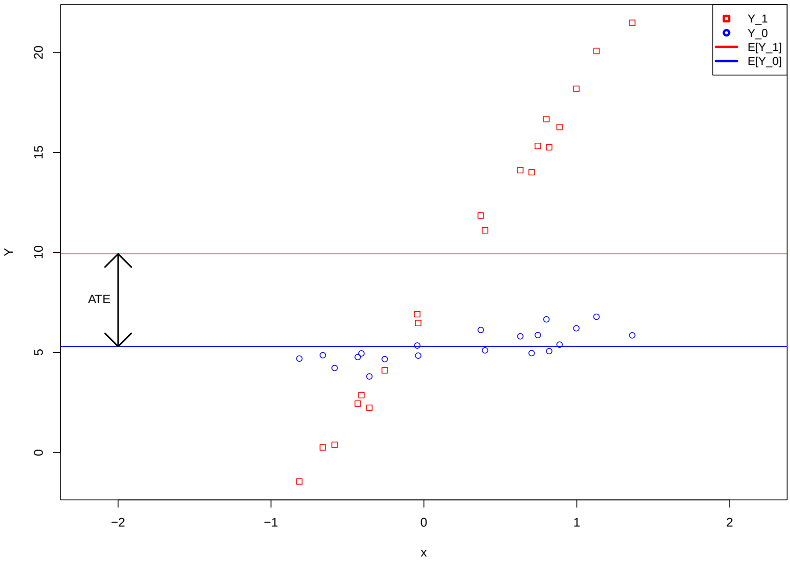
Regression with Covariates

- Practioners often run some variant of the following model with experimental data:

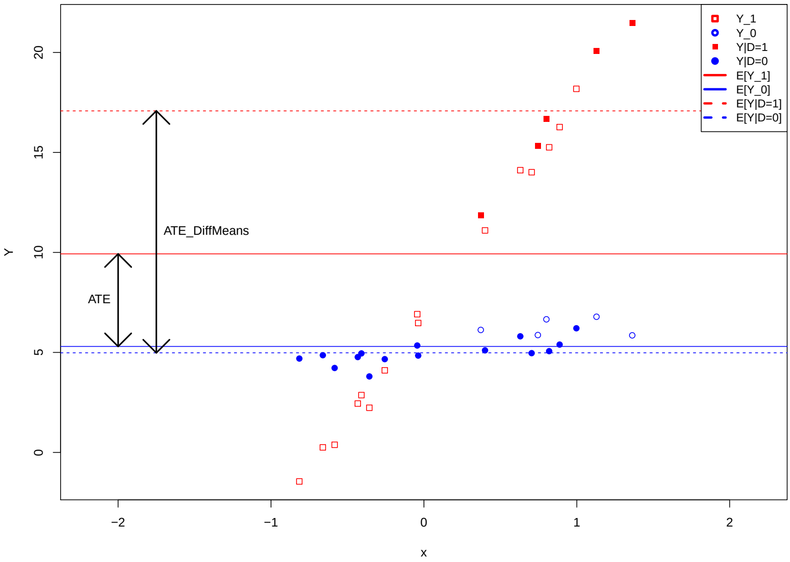
$$Y_i = \alpha + \tau D_i + X_i\beta + \epsilon_i$$

- Why include X_i when experiments "control" for covariates by design?
  - Correct for chance covariate imbalances that indicate that $\hat{\tau}$ may be far from τ_{ATE}.
  - Increase precision: remove variation in the outcome accounted for by pre-treatment characteristics, thus making it easier to attribute remaining differences to the treatment.
- ATE estimates are robust to model specification (with sufficient N).
  - Never control for **post-treatment** covariates!

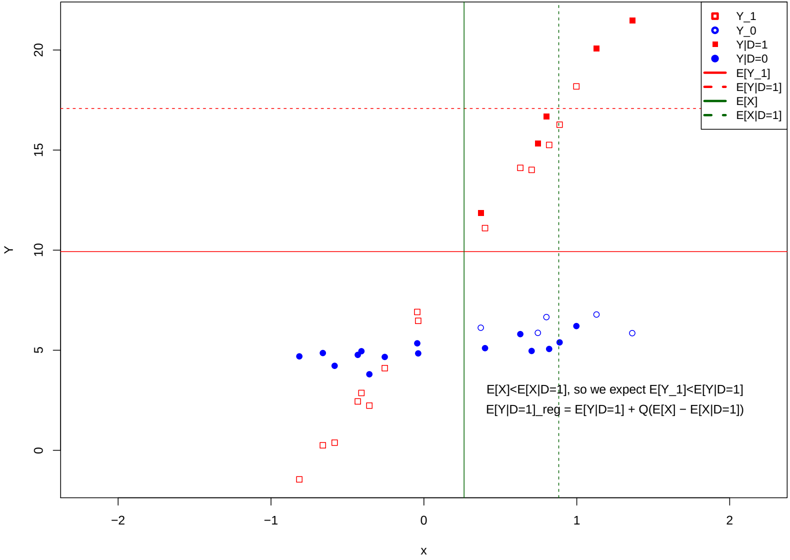
True ATE



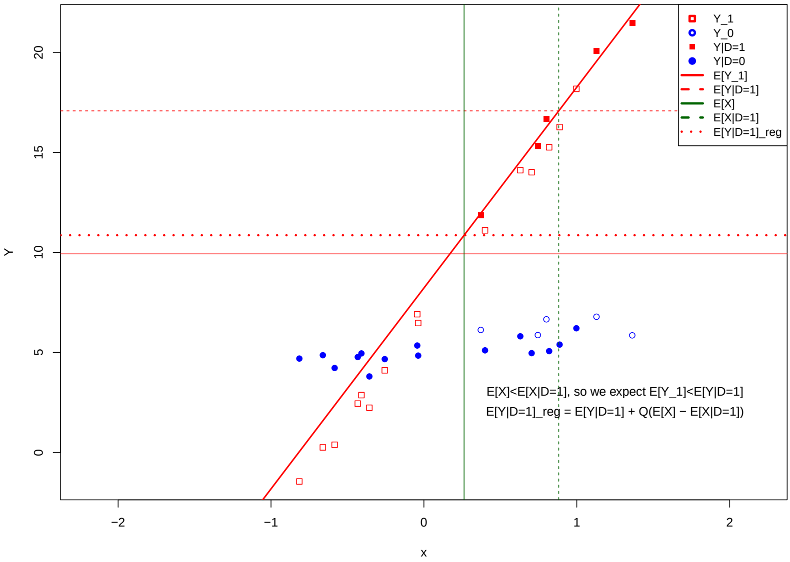
True ATE and Unadjusted Regression Estimator



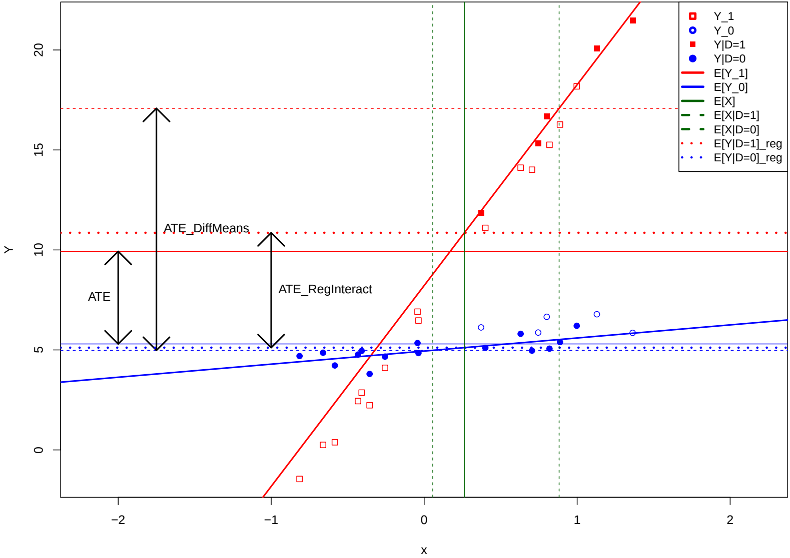
Adjusted Regression Estimator



Adjusted Regression Estimator



Adjusted Regression Estimator



Covariate Adjustment with Regression

Freedman (2008) shows that regression of the form:

$$Y_i = \alpha + \tau_{reg} D_i + \beta_1 X_i + \epsilon_i$$

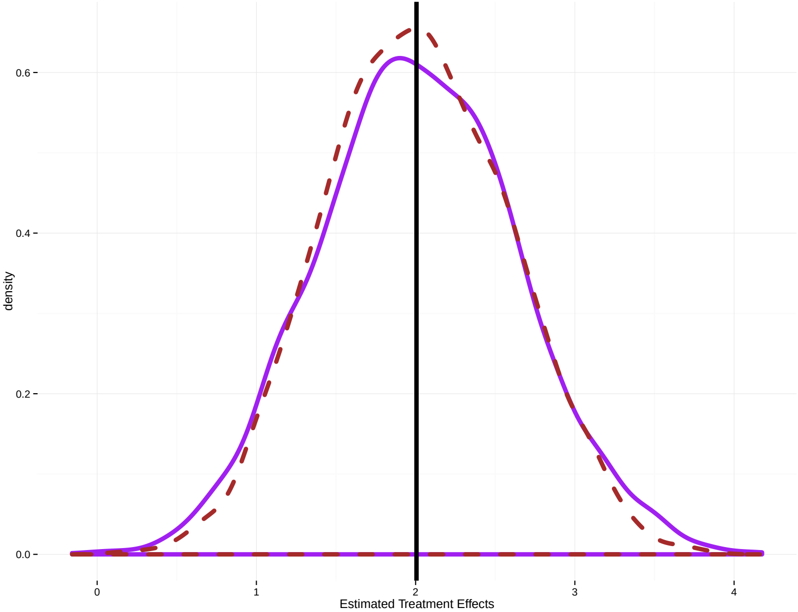
- $\hat{\tau}_{reg}$ is consistent for ATE and has small sample bias (unless model is true)
  - bias is on the order of $1/n$ and diminishes rapidly as N increases
- $\hat{\tau}_{reg}$ will not necessarily improve precision if model is incorrect
  - But harmful to precision only if more than 3/4 of units are assigned to one treatment condition or $\mathrm{Cov}(D_i, Y_1 - Y_0)$ larger than $\mathrm{Cov}(D_i, Y)$.

Lin (2013) shows that regression of the form:

$$Y_i = \alpha + \tau_{interact} D_i + \beta_1 \cdot (X_i - \bar{X}) + \beta_2 \cdot D_i \cdot (X_i - \bar{X}) + \epsilon_i$$

- $\hat{\tau}_{interact}$ is consistent for ATE and has the same small sample bias
- Cannot hurt asymptotic precision even if model is incorrect and will likely increase precision if covariates are predictive of the outcomes.
- Results hold for multiple covariates

Covariate Adjustment with Regression



Why are Experimental Findings Robust to Alternative Specifications?

Note the following important property of OLS known as the Frisch-Waugh-Lovell (FWL) theorem or Anatomy of Regression:

$$\beta_k = \frac{Cov(Y_i, \tilde{x}_{ki})}{Var(\tilde{x}_{ki})}$$

where $\tilde{x}_{ki}$ is the residual from a regression of x_{ki} on all other covariates.

Why are Experimental Findings Robust to Alternative Specifications?

Note the following important property of OLS known as the Frisch-Waugh-Lovell (FWL) theorem or Anatomy of Regression:

$$\beta_k = \frac{Cov(Y_i, \tilde{x}_{ki})}{Var(\tilde{x}_{ki})}$$

where $\tilde{x}_{ki}$ is the residual from a regression of x_{ki} on all other covariates. Any multivariate regression coefficient can be expressed as the coefficient on a bivariate regression between the outcome and the regressor, after "partialling out" other variables in the model.

Why are Experimental Findings Robust to Alternative Specifications?

Note the following important property of OLS known as the Frisch-Waugh-Lovell (FWL) theorem or Anatomy of Regression:

$$\beta_k = \frac{Cov(Y_i, \tilde{x}_{ki})}{Var(\tilde{x}_{ki})}$$

where $\tilde{x}_{ki}$ is the residual from a regression of x_{ki} on all other covariates.
Any multivariate regression coefficient can be expressed as the coefficient on a bivariate regression between the outcome and the regressor, after "partialling out" other variables in the model.
Let $\tilde{D}_i$ be the residuals after regressing D_i on X_i. For experimental data, on average, what will $\tilde{D}_i$ be equal to?

Why are Experimental Findings Robust to Alternative Specifications?

Note the following important property of OLS known as the Frisch-Waugh-Lovell (FWL) theorem or Anatomy of Regression:

$$\beta_k = \frac{Cov(Y_i, \tilde{x}_{ki})}{Var(\tilde{x}_{ki})}$$

where $\tilde{x}_{ki}$ is the residual from a regression of x_{ki} on all other covariates.
Any multivariate regression coefficient can be expressed as the coefficient on a bivariate regression between the outcome and the regressor, after "partialling out" other variables in the model.
Let $\tilde{D}_i$ be the residuals after regressing D_i on X_i. For experimental data, on average, what will $\tilde{D}_i$ be equal to?
Since $\tilde{D}_i \approx D_i$, multivariate regressions will yield similar results to bivariate regressions.

Summary: Covariate Adjustment with Regression

- One does not need to believe in the classical linear model (linearity and constant treatment effects) to tolerate or even advocate OLS covariate adjustment in randomized experiments (agnostic view of regression)
- Covariate adjustment can buy you power (and thus allows for a smaller sample).
- Small sample bias might be a concern in small samples, but usually swamped by efficiency gains.
- Since covariates are controlled for by design, results are typically not model dependent
- Best if covariate adjustment strategy is pre-specified as this rules out fishing.
- Always show the unadjusted estimate for transparency.

Testing in Small Samples: Fisher's Exact Test

- Test of differences in means with large N:

$$H_0 : \mathbb{E}[Y_1] = \mathbb{E}[Y_0], \quad H_1 : \mathbb{E}[Y_1] \neq \mathbb{E}[Y_0] \text{ (weak null)}$$

Testing in Small Samples: Fisher's Exact Test

- Test of differences in means with large N:

$$H_0 : \mathbb{E}[Y_1] = \mathbb{E}[Y_0], \quad H_1 : \mathbb{E}[Y_1] \neq \mathbb{E}[Y_0] \text{ (weak null)}$$

- Fisher's Exact Test with small N:

$$H_0 : Y_1 = Y_0,$$

Testing in Small Samples: Fisher's Exact Test

- Test of differences in means with large N:

$$H_0 : \mathbb{E}[Y_1] = \mathbb{E}[Y_0], \quad H_1 : \mathbb{E}[Y_1] \neq \mathbb{E}[Y_0] \text{ (weak null)}$$

- Fisher's Exact Test with small N:

$$H_0 : Y_1 = Y_0, \quad H_1 : Y_1 \neq Y_0 \qquad \text{(sharp null of no effect)}$$

Testing in Small Samples: Fisher's Exact Test

- Test of differences in means with large N:

$$H_0 : \mathbb{E}[Y_1] = \mathbb{E}[Y_0], \quad H_1 : \mathbb{E}[Y_1] \neq \mathbb{E}[Y_0] \text{ (weak null)}$$

- Fisher's Exact Test with small N:

$$H_0 : Y_1 = Y_0, \quad H_1 : Y_1 \neq Y_0 \qquad \text{(sharp null of no effect)}$$

- Let Ω be the set of all possible randomization realizations.
- We only observe the outcomes, Y_i, for one realization of the experiment. We calculate $\hat{\tau} = \bar{Y}_1 - \bar{Y}_0$.

Testing in Small Samples: Fisher's Exact Test

- Test of differences in means with large N:

$$H_0 : \mathbb{E}[Y_1] = \mathbb{E}[Y_0], \quad H_1 : \mathbb{E}[Y_1] \neq \mathbb{E}[Y_0] \text{ (weak null)}$$

- Fisher's Exact Test with small N:

$$H_0 : Y_1 = Y_0, \quad H_1 : Y_1 \neq Y_0 \qquad \text{(sharp null of no effect)}$$

- Let Ω be the set of all possible randomization realizations.
- We only observe the outcomes, Y_i, for one realization of the experiment. We calculate $\hat{\tau} = \bar{Y}_1 - \bar{Y}_0$.
- Under the sharp null hypothesis, we can compute the value that the difference in means estimator would have taken under any other realization, $\hat{\tau}(\omega)$, for $\omega \in \Omega$.

Testing in Small Samples: Fisher's Exact Test

i	Y_{1i}	Y_{0i}	D_i
1	3	?	1
2	1	?	1
3	?	0	0
4	?	1	0
$\widehat{\tau}_{ATE}$			1.5

What do we know given the sharp null $H_0 : Y_1 = Y_0$?

Testing in Small Samples: Fisher's Exact Test

i	Y_{1i}	Y_{0i}	D_i
1	3	3	1
2	1	1	1
3	0	0	0
4	1	1	0
$\widehat{\tau}_{ATE}$			1.5
$\hat{\tau}(\omega)$			1.5

Given the full schedule of potential outcomes under the sharp null, we can compute the null distribution of ATE_{H_0} across all possible randomization.

Testing in Small Samples: Fisher's Exact Test

i	Y_{1i}	Y_{0i}	D_i	D_i
1	3	3	1	1
2	1	1	1	0
3	0	0	0	1
4	1	1	0	0
$\widehat{\tau}_{ATE}$			1.5	
$\hat{\tau}(\omega)$			1.5	0.5

Testing in Small Samples: Fisher's Exact Test

i	Y_{1i}	Y_{0i}	D_i	D_i	D_i
1	3	3	1	1	1
2	1	1	1	0	0
3	0	0	0	1	0
4	1	1	0	0	1
$\widehat{\tau}_{ATE}$			1.5		
$\hat{\tau}(\omega)$			1.5	0.5	1.5

Testing in Small Samples: Fisher's Exact Test

i	Y_{1i}	Y_{0i}	D_i	D_i	D_i	D_i
1	3	3	1	1	1	0
2	1	1	1	0	0	1
3	0	0	0	1	0	1
4	1	1	0	0	1	0
$\widehat{\tau}_{ATE}$			1.5			
$\hat{\tau}(\omega)$			1.5	0.5	1.5	-1.5

Testing in Small Samples: Fisher's Exact Test

i	Y_{1i}	Y_{0i}	D_i	D_i	D_i	D_i	D_i
1	3	3	1	1	1	0	0
2	1	1	1	0	0	1	1
3	0	0	0	1	0	1	0
4	1	1	0	0	1	0	1
$\widehat{\tau}_{ATE}$			1.5				
$\hat{\tau}(\omega)$			1.5	0.5	1.5	-1.5	-.5

Testing in Small Samples: Fisher's Exact Test

i	Y_{1i}	Y_{0i}	D_i	D_i	D_i	D_i	D_i	D_i
1	3	3	1	1	1	0	0	0
2	1	1	1	0	0	1	1	0
3	0	0	0	1	0	1	0	1
4	1	1	0	0	1	0	1	1
$\widehat{\tau}_{ATE}$			1.5					
$\hat{\tau}(\omega)$			1.5	0.5	1.5	-1.5	-.5	-1.5

So $\Pr(\hat{\tau}(\omega) \geq \widehat{\tau}_{ATE}) = 2/6 \approx .33$.

Which assumptions are needed?

Testing in Small Samples: Fisher's Exact Test

i	Y_{1i}	Y_{0i}	D_i	D_i	D_i	D_i	D_i	D_i
1	3	3	1	1	1	0	0	0
2	1	1	1	0	0	1	1	0
3	0	0	0	1	0	1	0	1
4	1	1	0	0	1	0	1	1
$\widehat{\tau}_{ATE}$			1.5					
$\hat{\tau}(\omega)$			1.5	0.5	1.5	-1.5	-.5	-1.5

So $\Pr(\hat{\alpha}(\omega) \geq \widehat{\tau}_{ATE}) = 2/6 \approx .33$.

Which assumptions are needed? None! Randomization as "reasoned basis for causal inference" (Fisher 1935)

Blocking

- Imagine you have data on the units that you are about to randomly assign. Why leave it to "pure" chance to balance the observed characteristics?
- Idea in blocking is to pre-stratify the sample and then to randomize separately within each stratum to ensure that the groups start out with identical observable characteristics on the blocked factors.
- You effectively run a separate experiment within each stratum, randomization will balance the unobserved attributes
- Why is this helpful?
  - Four subjects with pre-treatment outcomes of $\{2,2,8,8\}$
  - Divided evenly into treatment and control groups and treatment effect is zero
  - Simple random assignment will place $\{2,2\}$ and $\{8,8\}$ together in the same treatment or control group $1/3$ of the time

Blocking

Imagine you run an experiment where you block on gender. It's possible to think about an ATE composed of two seperate block-specific ATEs:

$$\tau = \frac{N_f}{N_f + N_m} \cdot \tau_f + \frac{N_m}{N_f + N_m} \cdot \tau_m$$

Blocking

Imagine you run an experiment where you block on gender. It's possible to think about an ATE composed of two seperate block-specific ATEs:

$$\tau = \frac{N_f}{N_f + N_m} \cdot \tau_f + \frac{N_m}{N_f + N_m} \cdot \tau_m$$

An unbiased estimator for this quantity will be

$$\hat{\tau}_B = \frac{N_f}{N_f + N_m} \cdot \hat{\tau}_f + \frac{N_m}{N_f + N_m} \cdot \hat{\tau}_m$$

Blocking

Imagine you run an experiment where you block on gender. It's possible to think about an ATE composed of two seperate block-specific ATEs:

$$\tau = \frac{N_f}{N_f + N_m} \cdot \tau_f + \frac{N_m}{N_f + N_m} \cdot \tau_m$$

An unbiased estimator for this quantity will be

$$\hat{\tau}_B = \frac{N_f}{N_f + N_m} \cdot \hat{\tau}_f + \frac{N_m}{N_f + N_m} \cdot \hat{\tau}_m$$

or more generally, if there are J strata or blocks, then

$$\hat{\tau}_B = \sum_{j=1}^{J} \frac{N_j}{N} \hat{\tau}_j$$

Blocking

Because the randomizations in each block are independent, the variance of the blocking estimator is simply $(\mathrm{Var}(aX + bY) = a^2\mathrm{Var}(X) + b^2\mathrm{Var}(Y))$
:

$$\mathrm{Var}(\hat{\tau}_B) = \left(\frac{N_f}{N_f + N_m}\right)^2 \mathrm{Var}(\hat{\tau}_f) + \left(\frac{N_m}{N_f + N_m}\right)^2 \mathrm{Var}(\hat{\tau}_m)$$

or more generally

$$Var(\hat{\tau}_B) = \sum_{j=1}^{J} \left(\frac{N_j}{N}\right)^2 \mathrm{Var}(\hat{\tau}_j)$$

Blocking with Regression

When analyzing a blocked randomized experiment with OLS and the probability of receiving treatment is equal across blocks, then OLS with block "fixed effects" will result in a valid estimator of the ATE:

$$y_i = \tau D_i + \sum_{j=2}^{J} \beta_j \cdot B_{ij} + \epsilon_i$$

where B_j is a dummy for the j-th block (one omitted as reference category).

Blocking with Regression

When analyzing a blocked randomized experiment with OLS and the probability of receiving treatment is equal across blocks, then OLS with block "fixed effects" will result in a valid estimator of the ATE:

$$y_i = \tau D_i + \sum_{j=2}^{J} \beta_j \cdot B_{ij} + \epsilon_i$$

where B_j is a dummy for the j-th block (one omitted as reference category).

If probabilites of treatment, $p_{ij} = P(D_{ij} = 1)$, vary by block, then weight each observation:

$$w_{ij} = \left(\frac{1}{p_{ij}}\right) D_i + \left(\frac{1}{1 - p_{ij}}\right) (1 - D_i)$$

Blocking with Regression

When analyzing a blocked randomized experiment with OLS and the probability of receiving treatment is equal across blocks, then OLS with block “fixed effects” will result in a valid estimator of the ATE:

$$y_i = \tau D_i + \sum_{j=2}^{J} \beta_j \cdot B_{ij} + \epsilon_i$$

where B_j is a dummy for the j-th block (one omitted as reference category).

If probabilites of treatment, $p_{ij} = P(D_{ij} = 1)$, vary by block, then weight each observation:

$$w_{ij} = \left(\frac{1}{p_{ij}}\right) D_i + \left(\frac{1}{1 - p_{ij}}\right) (1 - D_i)$$

Why do this? When treatment probabilities vary by block, then OLS will weight blocks by the variance of the treatment variable in each block. Without correcting for this, OLS will result in biased estimates of ATE!

When Does Blocking Help?

Imagine a model for a complete and blocked randomized design:

$$Y_i = \alpha + \tau_{CR} D_i + \varepsilon_i$$

When Does Blocking Help?

Imagine a model for a complete and blocked randomized design:

$$Y_i = \alpha + \tau_{CR} D_i + \varepsilon_i \quad (1)$$

$$Y_i = \alpha + \tau_{BR} D_i + \sum_{j=2}^{J} \beta_j B_{ij} + \varepsilon_i^* \quad (2)$$

where B_j is a dummy for the j-th block. Then given iid sampling:

When Does Blocking Help?

Imagine a model for a complete and blocked randomized design:

$$Y_i = \alpha + \tau_{CR} D_i + \varepsilon_i \quad (1)$$

$$Y_i = \alpha + \tau_{BR} D_i + \sum_{j=2}^{J} \beta_j B_{ij} + \varepsilon_i^* \quad (2)$$

where B_j is a dummy for the j-th block. Then given iid sampling:

$$Var[\widehat{\tau}_{CR}] = \frac{\sigma_\varepsilon^2}{\sum_{i=1}^{n}(D_i - \bar{D})^2} \qquad \text{with } \widehat{\sigma}_\varepsilon^2 =$$

When Does Blocking Help?

Imagine a model for a complete and blocked randomized design:

$$Y_i = \alpha + \tau_{CR} D_i + \varepsilon_i \quad (1)$$

$$Y_i = \alpha + \tau_{BR} D_i + \sum_{j=2}^{J} \beta_j B_{ij} + \varepsilon_i^* \quad (2)$$

where B_j is a dummy for the j-th block. Then given iid sampling:

$$Var[\widehat{\tau}_{CR}] = \frac{\sigma_\varepsilon^2}{\sum_{i=1}^n (D_i - \bar{D})^2} \quad \text{with } \widehat{\sigma}_\varepsilon^2 = \frac{\sum_{i=1}^n \widehat{\varepsilon}_i^2}{n-2} = \frac{SSR_{\widehat{\varepsilon}}}{n-2}$$

$$Var[\widehat{\tau}_{BR}] = \frac{\sigma_{\varepsilon^*}^2}{\sum_{i=1}^n (D_i - \bar{D})^2 (1 - R_j^2)} \quad \text{with } \widehat{\sigma}_{\varepsilon^*}{}^2 = \frac{\sum_{i=1}^n \widehat{\varepsilon}_i^{*2}}{n-k-1} = \frac{SSR_{\widehat{\varepsilon}^*}}{n-k-1}$$

where R_j^2

When Does Blocking Help?

Imagine a model for a complete and blocked randomized design:

$$Y_i = \alpha + \tau_{CR} D_i + \varepsilon_i \quad (1)$$

$$Y_i = \alpha + \tau_{BR} D_i + \sum_{j=2}^{J} \beta_j B_{ij} + \varepsilon_i^* \quad (2)$$

where B_j is a dummy for the j-th block. Then given iid sampling:

$$Var[\widehat{\tau}_{CR}] = \frac{\sigma_\varepsilon^2}{\sum_{i=1}^n (D_i - \bar{D})^2} \quad \text{with } \widehat{\sigma}_\varepsilon^2 = \frac{\sum_{i=1}^n \widehat{\varepsilon}_i^2}{n-2} = \frac{SSR_{\widehat{\varepsilon}}}{n-2}$$

$$Var[\widehat{\tau}_{BR}] = \frac{\sigma_{\varepsilon^*}^2}{\sum_{i=1}^n (D_i - \bar{D})^2 (1 - R_j^2)} \quad \text{with } \widehat{\sigma}_{\varepsilon^*}{}^2 = \frac{\sum_{i=1}^n \widehat{\varepsilon}_i^{*2}}{n-k-1} = \frac{SSR_{\widehat{\varepsilon}^*}}{n-k-1}$$

where R_j^2 is R^2 from regression of D on all B_j variables and a constant.

When Does Blocking Help?

$$Y_i = \alpha + \tau_{CR} D_i + \varepsilon_i \quad (3)$$

$$Y_i = \alpha + \tau_{BR} D_i + \sum_{j=2}^{J} \beta_j B_{ij} + \varepsilon_i^* \quad (4)$$

where B_k is a dummy for the k-th block. Then given iid sampling:

$$V[\widehat{\tau}_{CR}] = \frac{\sigma_\varepsilon^2}{\sum_{i=1}^n (D_i - \bar{D})^2} \quad \text{with } \widehat{\sigma}_\varepsilon^2 = \frac{\sum_{i=1}^n \widehat{\varepsilon}_i^2}{n-2} = \frac{SSR_{\widehat{\varepsilon}}}{n-2}$$

$$V[\widehat{\tau}_{BR}] = \frac{\sigma_{\varepsilon^*}^2}{\sum_{i=1}^n (D_i - \bar{D})^2 (1 - R_j^2)} \quad \text{with } \widehat{\sigma}_{\varepsilon^*}{}^2 = \frac{\sum_{i=1}^n \widehat{\varepsilon}_i^{*2}}{n-k-1} = \frac{SSR_{\widehat{\varepsilon}^*}}{n-k-1}$$

where R_j^2 is R^2 from regression of D on the B_k dummies and a constant.
So when is $Var[\widehat{\tau}_{BR}] < Var[\widehat{\tau}_{CR}]$?

When Does Blocking Help?

$$
\begin{aligned}
Y_i &= \alpha + \tau_{CR} D_i + \varepsilon_i && (5)\\
Y_i &= \alpha + \tau_{BR} D_i + \sum_{j=2}^{J} \beta_j B_{ij} + \varepsilon_i^* && (6)
\end{aligned}
$$

where B_k is a dummy for the k-th block. Then given iid sampling:

$$
\begin{aligned}
V[\widehat{\tau}_{CR}] &= \frac{\sigma_\varepsilon^2}{\sum_{i=1}^n (D_i - \bar{D})^2} && \text{with } \widehat{\sigma}_\varepsilon^2 = \frac{\sum_{i=1}^n \widehat{\varepsilon}_i^2}{n-2} = \frac{SSR_{\widehat{\varepsilon}}}{n-2}\\
V[\widehat{\tau}_{BR}] &= \frac{\sigma_{\varepsilon^*}^2}{\sum_{i=1}^n (D_i - \bar{D})^2 (1 - R_j^2)} && \text{with } \widehat{\sigma}_{\varepsilon^*}{}^2 = \frac{\sum_{i=1}^n \widehat{\varepsilon_i^*}^2}{n-k-1} = \frac{SSR_{\widehat{\varepsilon^*}}}{n-k-1}
\end{aligned}
$$

where R_j^2 is R^2 from regression of D on the B_k dummies and a constant.
Since $R_j^2 \approx 0$ $V[\widehat{\tau}_{BR}] < V[\widehat{\tau}_{CR}]$ if $\frac{SSR_{\widehat{\varepsilon^*}}}{n-k-1} < \frac{SSR_{\widehat{\varepsilon}}}{n-2}$

Blocking

- How does blocking help?
  - Increases efficiency if the blocking variables predict outcomes (i.e. they "remove" the variation that is driven by nuisance factors)
  - Blocking on irrelevant predictors can burn up degrees of freedom.
  - Can help with small sample bias due to "bad" randomization
  - Is powerful especially in small to medium sized samples.

Blocking

- How does blocking help?
  - Increases efficiency if the blocking variables predict outcomes (i.e. they "remove" the variation that is driven by nuisance factors)
  - Blocking on irrelevant predictors can burn up degrees of freedom.
  - Can help with small sample bias due to "bad" randomization
  - Is powerful especially in small to medium sized samples.
- What to block on?
  - "Block what you can, randomize what you can't"
  - The baseline of the outcome variable and other main predictors.
  - Variables desired for subgroup analysis
- How to block?
  - Stratification
  - Pair-matching
  - Check: `blockTools` library.

Analysis with Blocking

- **"As ye randomize, so shall ye analyze"** (Senn 2004): Need to account for the method of randomization when performing statistical analysis.
- If using OLS, strata dummies should be included when analyzing results of stratified randomization.
  - If probability of treatment assignment varies across blocks, then weight treated units by probability of being in treatment and controls by the probability of being a control.
- Failure to control for the method of randomization can result in incorrect test size.