Week 10: Causality with Measured Confounding

Brandon Stewart[1]

Princeton

November 2–6, 2020

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

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

- Last Week
  - frameworks for causal inference
- This Week
  - experimental ideal
  - identification with measured confounding
  - estimation via stratification, matching and regression
- Next Week
  - approaches with unmeasured confounding
- Long Run
  - causal frameworks → inference → regression → causal inference

1. The Experimental Ideal

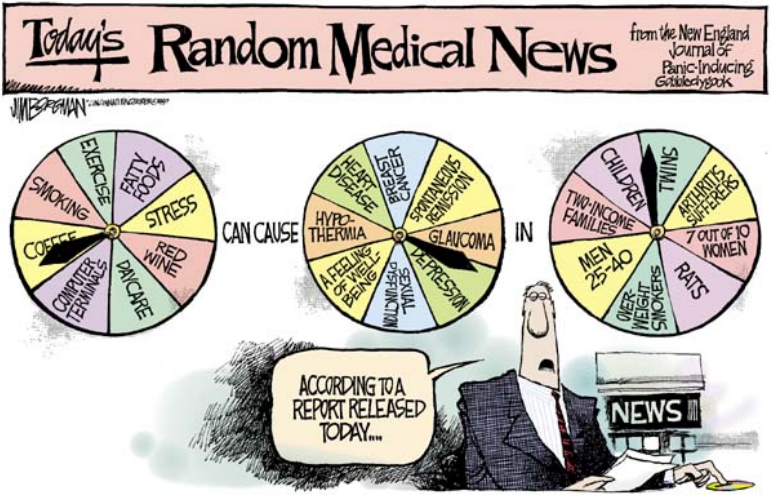
2. Identification with Measured Confounding
   - Design
   - DAGs

3. Stratification

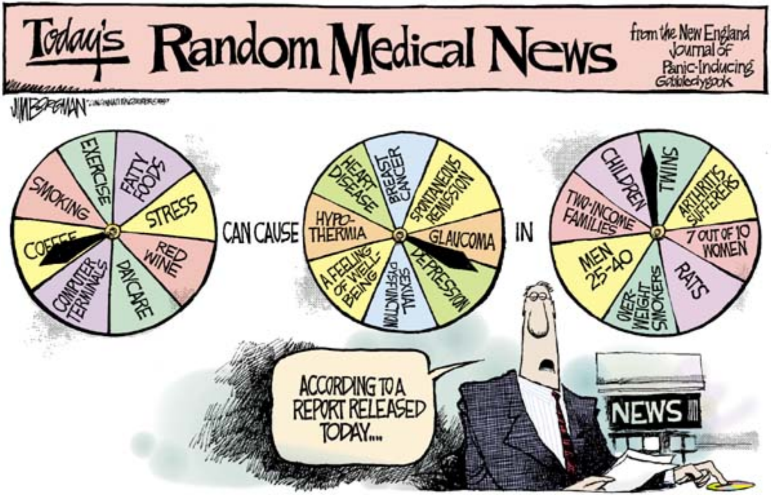
4. Matching
   - Fundamentals of Matching
   - Two Approaches to Matching

5. Regression
   - Regression with Heterogeneous Effects
   - Imputation Estimators
   - Fun With Weights

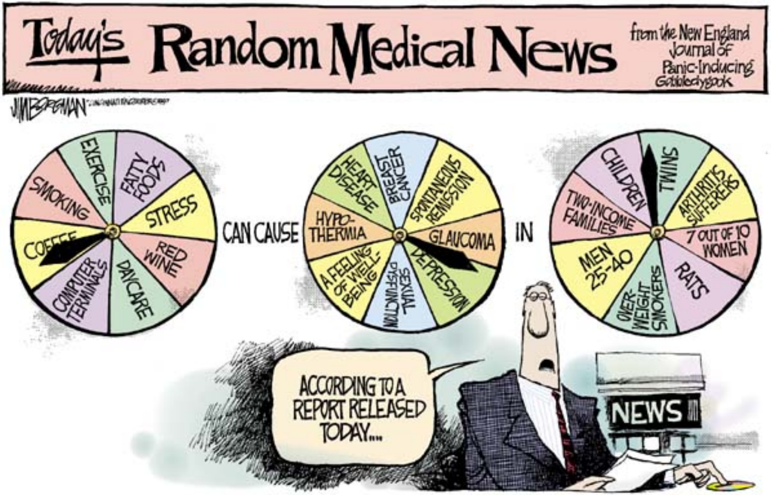
6. Estimands



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)



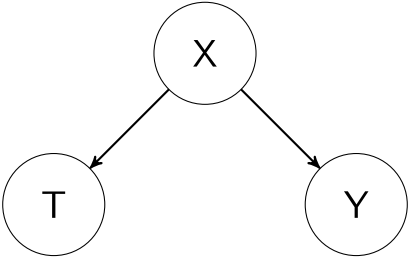
Lancet 2002: no effect of vitamin C on mortality in controlled placebo trial (controlling for nothing)



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?

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
- Better than hoping: design observational study to approximate an experiment
  - "The planner of an observational study should always ask himself [sic]: How would the study be conducted if it were possible to do it by controlled experimentation" (Cochran 1965)

Experiment review

- An experiment is a study where assignment to treatment is controlled by the researcher.
  - $P(T_i = 1)$ is controlled and known by researcher.
- In the ideal randomized experiment, two assumptions hold by design:
  1. Positivity: assignment is probabilistic: $0 < P(T_i = 1) < 1$
     - ★ No deterministic assignment.
  2. Ignorability: $P\left(T_i = 1 | \mathbf{Y}(1), \mathbf{Y}(0)\right) = P(T_i = 1)$
     - ★ Treatment assignment does not depend on any potential outcomes.
     - ★ Sometimes written as $T_i \perp\!\!\!\perp (\mathbf{Y}(1), \mathbf{Y}(0))$

Why do Experiments Help?

Remember selection bias?

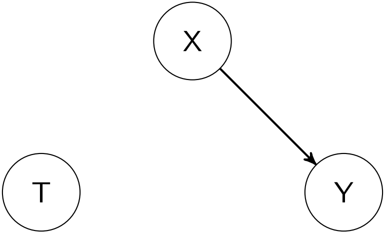
$$\begin{aligned}
&E[Y_i|T_i = 1] - E[Y_i|T_i = 0] \\
&\quad = E[Y_i(1)|T_i = 1] - E[Y_i(0)|T_i = 0] \\
&\quad = E[Y_i(1)|T_i = 1] - E[Y_i(0)|T_i = 1] + E[Y_i(0)|T_i = 1] - E[Y_i(0)|T_i = 0] \\
&\quad = \underbrace{E[Y_i(1) - Y_i(0)|T_i = 1]}_{\text{Average Treatment Effect on Treated}} + \underbrace{E[Y_i(0)|T_i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

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

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

When all goes well, an experiment eliminates selection bias.

Randomization Removes the Arrows into the Treatment



This ensures any observed or unobserved pretreatment covariates have the same distribution in the treatment and the control group. That is, they are balanced.

Stratified Designs

- Stratified randomized experiment seek to remove bad randomizations where covariates are unbalanced by chance.
- Core Procedure:
  - form J blocks, b_j, $j = 1, \ldots, J$ based on the covariates
  - completely randomized assignment within each block.
  - randomization probability depends on the block variable, B_i
  - conditional ignorability: $T_i \perp\!\!\!\perp (Y_i(1), Y_i(0)) | B_i$.
- Generally, stratified designs mean that the probability of treatment depends on a covariate, X_i: $P(T_i = 1 | X_i = x)$.
- Conditional randomization assumptions:
  - Positivity: $0 < P(T_i = 1 | X_i = x) < 1$ for all i and x.
  - Unconfoundedness: $P(T_i = 1 | \mathbf{X}, \mathbf{Y}(1), \mathbf{Y}(0)) = P(T_i = 1 | X_i)$

Identification for Stratified Random Experiments

Can we identify the ATE under these stratified designs? Yes!

$$E[Y_i(1) - Y_i(0)] = E_X\Big\{E[Y_i(1) - Y_i(0)|X_i]\Big\} \qquad \text{(iterated expectations)}$$

$$= E_X\Big\{E[Y_i(1)|X_i] - E[Y_i(0)|X_i]\Big\}$$

$$= E_X\Big\{E[Y_i(1)|T_i = 1, X_i] - E[Y_i(0)|T_i = 0, X_i]\Big\} \quad \text{(ignorability)}$$

$$= E_X\Big\{E[Y_i|T_i = 1, X_i] - E[Y_i|T_i = 0, X_i]\Big\} \qquad \text{(SUTVA)}$$

- ATE is just the weighted average of the within-strata differences in means.
- Identified because the last line is a function of observables.
- The averaging is over the distribution of the strata $\rightsquigarrow$ size of the blocks.

Experiments

- The benefit of experiments is that key decisions hold by design and that you have balance along observed AND unobserved covariates.
- We aren't really covering experiments in this class as a broader discussion would require discussion of topics like randomization schemes, blocking, power calculations, internal/external validity, pre-registration and ethics.
- We cover a bit because experiments are a dominant metaphor. Much of the work on observational studies is trying to find a circumstance where we can pretend we have a stratified experiment.
- Of course, in real experiments sometimes people don't always do what we say (compliance problems).

We Covered

- Experiments
- Stratified designs
- Identification of same

Plan for the rest of the week:

- Identification using the stratified random experiment analogy.
- Three approaches for estimation: stratification, matching and regression.
- A deeper look at the implications of a few different estimands.

Next Time: Identification with Measured Confounding!

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

- Last Week
  - frameworks for causal inference
- This Week
  - experimental ideal
  - identification with measured confounding
  - estimation via stratification, matching and regression
- Next Week
  - approaches with unmeasured confounding
- Long Run
  - causal frameworks → inference → regression → causal inference

1. The Experimental Ideal

2. Identification with Measured Confounding
   - Design
   - DAGs

3. Stratification

4. Matching
   - Fundamentals of Matching
   - Two Approaches to Matching

5. Regression
   - Regression with Heterogeneous Effects
   - Imputation Estimators
   - Fun With Weights

6. Estimands

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 (minimize snooping!).
  - Try to estimate the randomization procedure.
  - Analyze this as a stratified randomized experiment with this estimated procedure.
- This is the perspective on observational design that comes out of the potential outcomes framework.
- Core task from this perspective is to reverse engineer the treatment assignment mechanism to find the experiment.

Identification Under Selection on Observables

Identification Assumption

1. $(Y(1), Y(0)) \perp\!\!\!\perp T | X$ *(selection on observables)*
2. $0 < P(T = 1|X) < 1$ *with probability one (common support)*

Identification Result

Given selection on observables we have

$$\begin{aligned} E[Y(1) - Y(0)|X] &= E[Y(1) - Y(0)|X, T = 1] \\ &= E[Y|X, T = 1] - E[Y|X, T = 0] \end{aligned}$$

Therefore, under the positivity assumption:

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

Identification Under Selection on Observables

Identification Assumption

1. $(Y(1), Y(0)) \perp\!\!\!\perp T | X$ *(selection on observables)*
2. $0 < P(T = 1|X) < 1$ *with probability one (common support)*

Identification Result

Similarly, for the Average Treatment Effect on the Treated (ATT),

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

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

- $Y(0) \perp\!\!\!\perp T | X$ *(Selection on Observables for Controls)*
- $P(T = 1|X) < 1$ *(Weak Positivity)*

Identification Under Selection on Observables

unit	Potential Outcome under Treatment	Potential Outcome under Control		
i	$Y_i(1)$	$Y_i(0)$	T_i	X_i
1	$E[Y(1)\|X=0,T=1]$	$E[Y(0)\|X=0,T=1]$	1	0
2			1	0
3	$E[Y(1)\|X=0,T=0]$	$E[Y(0)\|X=0,T=0]$	0	0
4			0	0
5	$E[Y(1)\|X=1,T=1]$	$E[Y(0)\|X=1,T=1]$	1	1
6			1	1
7	$E[Y(1)\|X=1,T=0]$	$E[Y(0)\|X=1,T=0]$	0	1
8			0	1

Identification Under Selection on Observables

unit	Potential Outcome under Treatment	Potential Outcome under Control		
i	$Y_i(1)$	$Y_i(0)$	T_i	X_i
1	$E[Y(1)\|X=0,T=1]$	$E[Y(0)\|X=0,T=1]=$	1	0
2		$E[Y(0)\|X=0,T=0]$	1	0
3	$E[Y(1)\|X=0,T=0]$	$E[Y(0)\|X=0,T=0]$	0	0
4			0	0
5	$E[Y(1)\|X=1,T=1]$	$E[Y(0)\|X=1,T=1]=$	1	1
6		$E[Y(0)\|X=1,T=0]$	1	1
7	$E[Y(1)\|X=1,T=0]$	$E[Y(0)\|X=1,T=0]$	0	1
8			0	1

$(Y(1),Y(0))\perp\!\!\!\perp T|X$ implies that we conditioned on all confounders. The treatment is randomly assigned within each stratum of X:

$$\begin{aligned} E[Y(0)|X=0,T=1] &= E[Y(0)|X=0,T=0] \text{ and} \\ E[Y(0)|X=1,T=1] &= E[Y(0)|X=1,T=0] \end{aligned}$$

Identification Under Selection on Observables

unit	Potential Outcome under Treatment	Potential Outcome under Control		
i	$Y_i(1)$	$Y_i(0)$	T_i	X_i
1	$E[Y(1)\|X=0,T=1]$	$E[Y(0)\|X=0,T=1]=$	1	0
2		$E[Y(0)\|X=0,T=0]$	1	0
3	$E[Y(1)\|X=0,T=0]=$	$E[Y(0)\|X=0,T=0]$	0	0
4	$E[Y(1)\|X=0,T=1]$		0	0
5	$E[Y(1)\|X=1,T=1]$	$E[Y(0)\|X=1,T=1]=$	1	1
6		$E[Y(0)\|X=1,T=0]$	1	1
7	$E[Y(1)\|X=1,T=0]=$	$E[Y(0)\|X=1,T=0]$	0	1
8	$E[Y(1)\|X=1,T=1]$		0	1

$(Y(1),Y(0)) \perp\!\!\!\perp T|X$ also implies

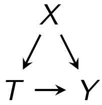
$$
\begin{aligned}
E[Y(1)|X=0,T=1] &= E[Y(1)|X=0,T=0] \text{ and} \\
E[Y(1)|X=1,T=1] &= E[Y(1)|X=1,T=0]
\end{aligned}
$$

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:
  - $P[T_i = 1 | \mathbf{X}, \mathbf{Y}(1), \mathbf{Y}(0)] = P[T_i = 1 | \mathbf{X}]$
  - 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 T to Y.
  - Would remain if we removed any arrows pointing out of T.
- Backdoor paths between T and Y $\rightsquigarrow$ common causes of T and Y:



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

Other types of confounding

$$\begin{tikzcd}
U \arrow[r, dashed] \arrow[d, dashed] & X \arrow[d] \\
T \arrow[r] & Y
\end{tikzcd}$$

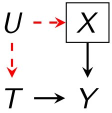
- T is enrolling in a job training program.
- Y is getting a job.
- U is being motivated
- X is number of job applications sent out.
- Big assumption here: no arrow from U to Y.

Other types of confounding

$$\begin{tikzcd}
U \arrow[r, dashed] \arrow[d, dashed] & X \arrow[d] \\
T \arrow[r] & Y
\end{tikzcd}$$

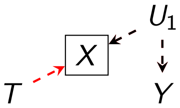
- T is exercise.
- Y is having a disease.
- U is lifestyle.
- X is smoking
- Big assumption here: no arrow from U to Y.

What's the problem with backdoor paths?



- A path is blocked if:
  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



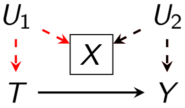
- Conditioning on the posttreatment covariates opens the non-causal path.
  - $\rightsquigarrow$ selection bias.

Don't condition on post-treatment variables

Every time you do, a puppy cries.

(just kidding. but seriously, this is one of the easiest ways to mess up your analysis if you don't know what you are doing.)

M-bias



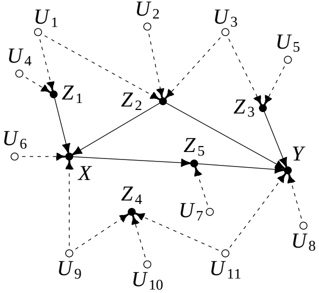
- Not all backdoor paths induce confounding.
- This backdoor path is blocked by the collider X that we don't control for.
- If we control for $X \rightsquigarrow$ opens the path and induces confounding.
  - Sometimes called M-bias.
- Controversial because of differing views on what to control for:
  - 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 argue for no unmeasured confounders?
- Pearl answered yes, with the backdoor criterion, which states that the effect of T on Y is identified if:
    1. No backdoor paths from T to Y OR
    2. Measured covariates are sufficient to block all backdoor paths from T 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.

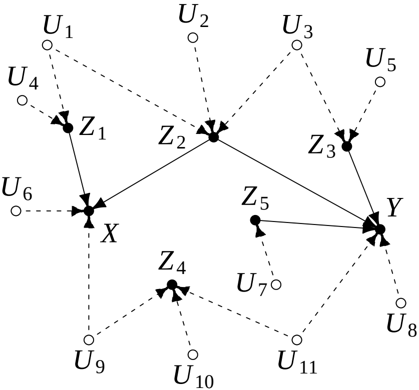
Example: Sufficient Conditioning Sets

We want to estimate the effect of X on Y for this DAG.



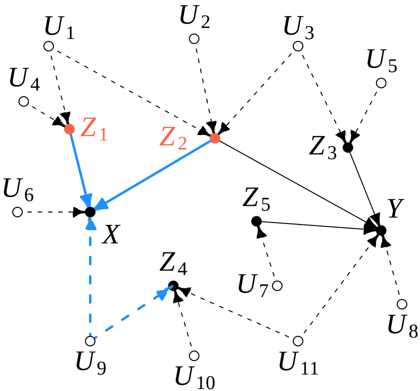
Remove arrows out of X.

Example: Sufficient Conditioning Sets



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

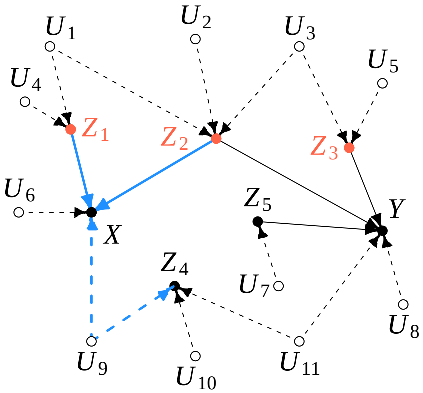
Example: Sufficient Conditioning Sets



No unblocked backdoor paths if we condition on Z_1 and Z_2

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

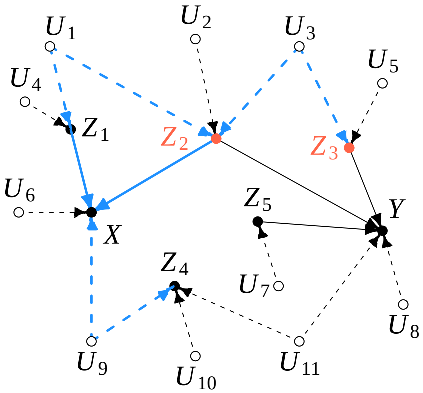
Example: Sufficient Conditioning Sets



No unblocked backdoor paths if we condition on Z_1, Z_2, and Z_3

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

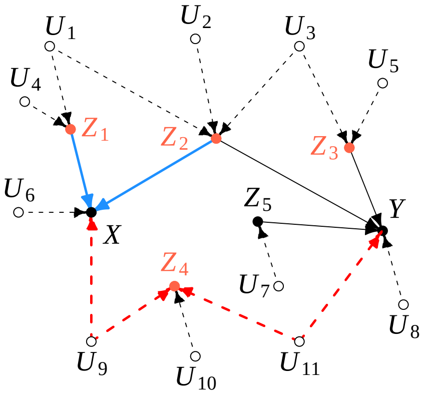
Example: Sufficient Conditioning Sets



No unblocked backdoor paths if we condition on Z_2 and Z_3

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

Example: Non-sufficient Conditioning Sets

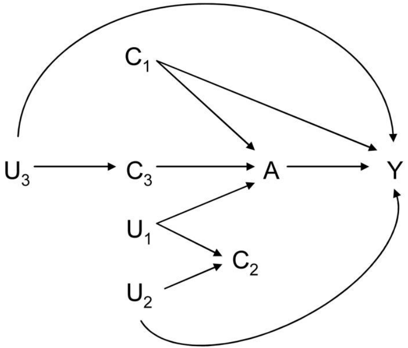


There are unblocked paths if we condition on Z_1, Z_2, Z_3, Z_4

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

More examples in Morgan and Winship

Implications (via Vanderweele and Shpitser 2011)



Two common criteria fail here:

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

No unmeasured confounders is not testable

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

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

- Recall, $\overset{d}{=}$ 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

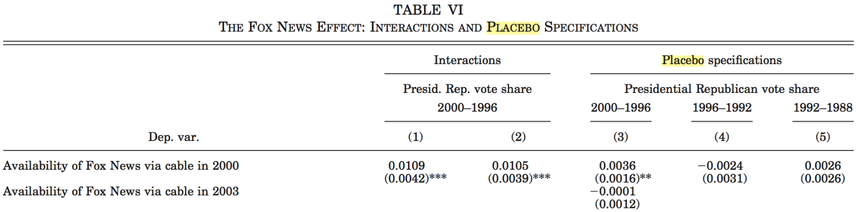


TABLE VI
THE FOX NEWS EFFECT: INTERACTIONS AND PLACEBO SPECIFICATIONS

	Interactions		Placebo specifications		
	Presid. Rep. vote share		Presidential Republican vote share		
	2000–1996		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 T_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!

So Where Does That Leave Us?

- If we can identify the right covariates **X** to get conditional ignorability we can do selection on observables.
- No unmeasured confounders $\approx$ randomized experiment.
- These variables are those that block backdoor paths and we want to be *really* sure we know what we are doing if we condition on a post-treatment variable.
- This still leaves us with a tricky estimation problem as we now need to work with distributions conditional on **X**.

We Covered

- Identification under selection on observables.
- DAGs return to help us choose what to condition on.

Next Time: Stratification!

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

- Last Week
  - frameworks for causal inference
- This Week
  - experimental ideal
  - identification with measured confounding
  - estimation via stratification, matching and regression
- Next Week
  - approaches with unmeasured confounding
- Long Run
  - causal frameworks → inference → regression → causal inference

1. The Experimental Ideal
2. Identification with Measured Confounding
   - Design
   - DAGs
3. Stratification
4. Matching
   - Fundamentals of Matching
   - Two Approaches to Matching
5. Regression
   - Regression with Heterogeneous Effects
   - Imputation Estimators
   - Fun With Weights
6. Estimands

Discrete covariates

- Suppose that we knew that T_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:

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

- Stratification is great because it makes no assumptions about parametric form.

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

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?

$15 \cdot (11/40) + 35 \cdot (13/40) + 50 \cdot (16/40) = 35.5$

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?

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

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?
- 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:

$$E[Y_i | T_i = 1, X_i = 54134] - E[Y_i | T_i = 0, X_i = 54134]$$

  - $\rightsquigarrow$ 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 did with age) but that doesn't work if there are patterns within the bins.
- Note that this is the problem of approximating conditional expectations and that's the machinery we've been building all semester!

We Covered

- Stratification!

Next Time: Matching

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

- Last Week
  - frameworks for causal inference
- This Week
  - experimental ideal
  - identification with measured confounding
  - estimation via stratification, matching and regression
- Next Week
  - approaches with unmeasured confounding
- Long Run
  - causal frameworks → inference → regression → causal inference

1. The Experimental Ideal

2. Identification with Measured Confounding
   - Design
   - DAGs

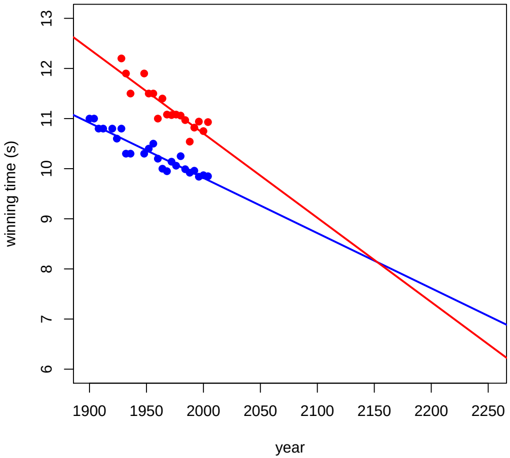
3. Stratification

4. Matching
   - Fundamentals of Matching
   - Two Approaches to Matching

5. Regression
   - Regression with Heterogeneous Effects
   - Imputation Estimators
   - Fun With Weights

6. Estimands

Remember This?



Tatem et al.'s predictions of sprint times with alternate models. Men's times are in blue, women's times are in red.

Model Dependence

Model Free Inference

To estimate $E(Y|X = x)$ at x, average many observed Y with value x

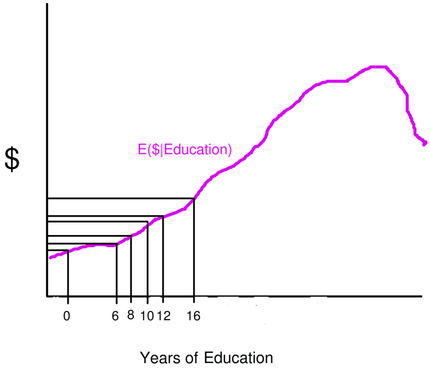
Assumptions (Model-Based Inference)

1. Definition: model dependence at x is the difference between predicted outcomes for any two models that fit about equally well.
2. The functional form follows strong continuity (think smoothness, although it is less restrictive)

Result

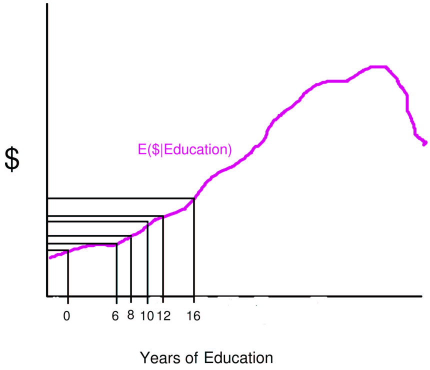
The maximum degree of model dependence: solely a function of the distance from the counterfactual to the data

What Inferences Would You Be Willing to Make?



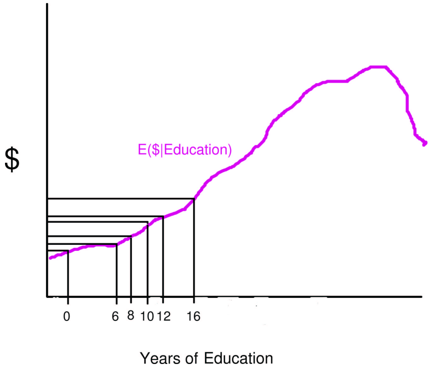
- A Factual Question: How much salary would someone receive with 12 years of education (a high school degree)?
- The model-free estimate: mean(Y) among those with $X = 12$.
- The model-based estimate: $\hat{Y} = X\hat{\beta} = 12 \times \$1,000 = \$12,000$

Counterfactual Inferences with Interpolation



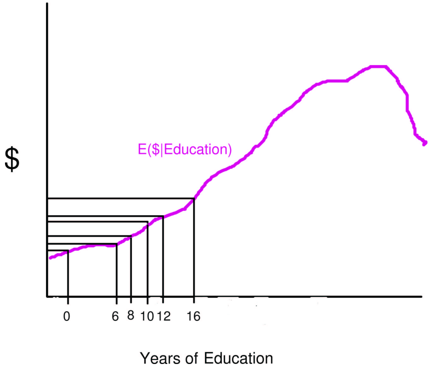
- How much salary would someone receive with 14 years of education (an Associates Degree)?
- Model free estimate: impossible
- Model-based estimate: $\hat{Y} = X\hat{\beta} = 14 \times \$1,000 = \$14,000$

Counterfactual Inference with Extrapolation



- How much salary would someone receive with 24 years of education (a Ph.D.)?
- $\hat{Y} = X\hat{\beta} = 24 \times \$1,000 = \$24,000$

Another Counterfactual Inference with Extrapolation



- How much salary would someone receive with 53 years of education?
- $\hat{Y} = X\hat{\beta} = 53 \times \$1,000 = \$53,000$
- What's changed? How would we recognize it when the example is less extreme or multidimensional?

Model Dependence with One Explanatory Variable

- Suppose Y is starting salary; X is education in 10 categories.
- To estimate $E(Y|X)$: we need 10 parameters, $E(Y|X = x_j)$, $j = 1, \ldots, 10$.
- Model-free method: average observations on Y for each value of X
- Model-based method: regress Y on X, summarizing 10 parameters with 2 (intercept and slope).
- The difference between the 10 we need and the 2 we estimate with regression is pure assumption.
- (If X were continuous, we would be reducing ∞ to 2, also by assumption)

Model Dependence with Two Explanatory Variables

Variables: X (education) and Z, parent's income, both with 10 categories

- How many parameters do we now need to estimate? 20? Nope. Its $10 \times 10 = 100$. This is the curse of dimensionality: the number of parameters goes up geometrically, not additively.
- If we run a regression, we are summarizing 100 parameters with 3 (an intercept and two slopes).
- But what about including an interaction? Right, so now we're summarizing 100 parameters with 4.
- The difference: an enormous assumption based on convenience, not evidence or theory.

Model Dependence with Many Explanatory Variables

- Suppose: 15 explanatory variables, with 10 categories each.
  - need to estimate 10^{15} (a quadrillion) parameters with how many observations?
  - Regression reduces this to 16 parameters; quite an assumption!
- Suppose: 80 explanatory variables.
  - 10^{80} is more than the number of atoms in the universe.
  - Yet, with a few simple assumptions, we can still run a regression and estimate only 81 parameters.
- The curse of dimensionality introduces huge assumptions, often unrecognized.

Overview of Matching

- Goal: reduce model dependence in our matching approach
- Makes parametric models work better rather than substitute for them (i.e,. matching is not an estimator; its a preprocessing method).
- It also prunes away data where we don't have common support (both treatment and control at same level of x),
- Apply model to preprocessed (pruned) rather than raw data
- Overall idea:
  - If each treated unit exactly matches a control unit w.r.t. X, then: (1) treated and control groups are identical, (2) X is no longer a confounder, (3) no need to worry about the functional form ($\bar{Y}_T - \bar{Y}_C$ is good enough).
  - If treated and control groups are better balanced than when you started, due to pruning, model dependence is reduced

Matching as Preprocessing

- Y_i dep var, T_i (1=treated, 0=control), X_i confounders
- Treatment Effect for treated observation i:

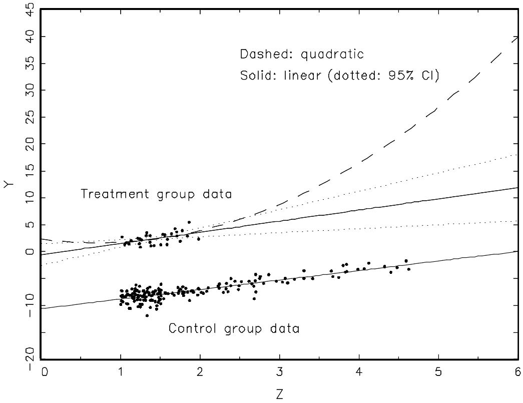
$$\begin{aligned}\text{TE}_i &= Y_i(1) - Y_i(0)\\ &= \text{observed} - \textit{unobserved}\end{aligned}$$

- Estimate $Y_i(0)$ with Y_j from matched ($X_i \approx X_j$) control
- Prune nonmatches: reduces imbalance & model dependence
- Follow preprocessing with whatever statistical method you'd have used without matching

(Warning: Pruning nonmatches can change your feasible estimand.)

How Matching Helps with Model Dependence

(King and Zeng, 2006: fig.4 Political Analysis)

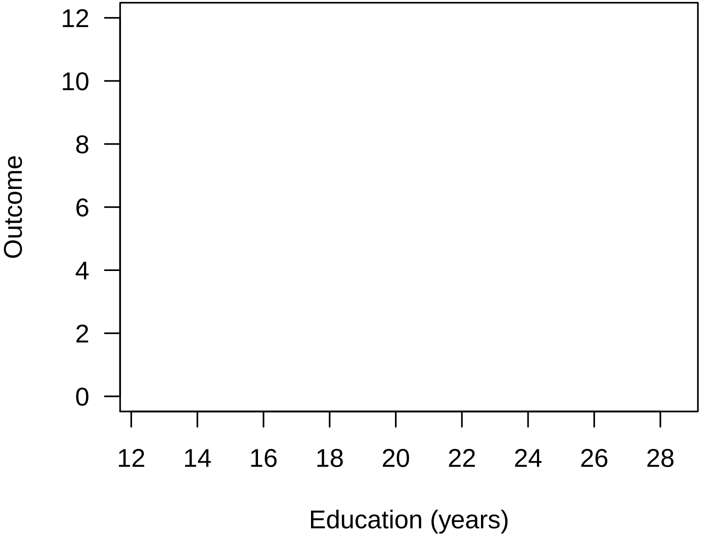


What to do?

- Preprocess I: Eliminate extrapolation region
- Preprocess II: Match (prune) within interpolation region
- Model remaining imbalance (as you would w/o matching)

Matching within the Interpolation Region

(Ho, Imai, King, Stuart, 2007: fig.1, Political Analysis)



Empirical Illustration: Carpenter, AJPS, 2002

- Hypothesis: Democratic senate majorities slow FDA drug approval time
- $n = 408$ new drugs (262 approved, 146 pending).
- lognormal survival model.
- seven oversight variables (median adjusted ADA scores for House and Senate Committees as well as for House and Senate floors, Democratic Majority in House and Senate, and Democratic Presidency).
- 18 control variables (clinical factors, firm characteristics, media variables, etc.)

Evaluating Reduction in Model Dependence

- Focus on the causal effect of a Democratic majority in the Senate (identified by Carpenter as not robust).
- Match: prune 49 units (2 treated, 17 control units).
- run 262,143 possible specifications and calculates ATE for each.
- Look at variability in ATE estimate across specifications.
- (Normal applications would only use one or a few specifications.)

Reducing Model Dependence

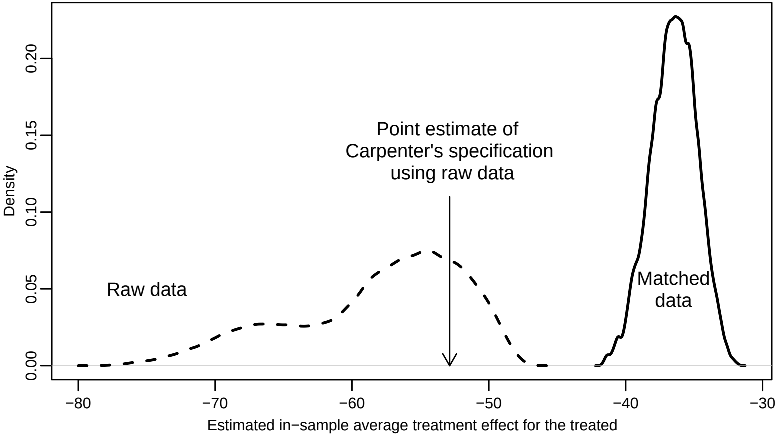


Figure: SATT Histogram: Effect of Democratic Senate majority on FDA drug approval time, across 262, 143 specifications.

1. The Experimental Ideal

2. Identification with Measured Confounding
   - Design
   - DAGs

3. Stratification

4. Matching
   - Fundamentals of Matching
   - Two Approaches to Matching

5. Regression
   - Regression with Heterogeneous Effects
   - Imputation Estimators
   - Fun With Weights

6. Estimands

Exact matching

- Let X_i take on a finite number of values, $x \in \mathcal{X}$.
- Let $\mathbb{I}_t = \{1, 2, \ldots, N_t\}$ be the set of treated units.
- Exact matching. For each treated unit, $i \in \mathbb{I}_t$:
  - Find the set of unmatched control units j such that $X_i = X_j$
  - Randomly select one of these control units to be the match, indicated $j(i)$.
- Let $\mathbb{I}_c = \{j(1), \ldots, j(N_t)\}$ be the set of matched controls.
- Last, discard all unmatched control units.
- The distribution of X_i will be exactly the same for treated and matched control:

$$P(X_i = x | T_i = 1) = P(X_i = x | T_i = 0, \mathbb{I}_c)$$

Weakening the identification assumptions

- No unmeasured confounders, SUTVA, and exact matches $\rightsquigarrow$ identifying the ATT.
- Can weaken no unmeasured confounders to conditional mean independence (CMI):

$$E[Y_i(0)|X_i, T_i = 1] = E[Y_i(0)|X_i, T_i = 0]$$

- Two nice features of CMI:
  1. Only have to make assumptions about $Y_i(0)$ not $Y_i(1)$
  2. Only places restrictions on the means, not other parts of the distribution (variance, skew, kurtosis, etc)

Analyzing exactly matched data

- How do we analyze the exactly matched data?
- Dead simple difference in means:

$$\widehat{\tau}_m = \frac{1}{N_t}\sum_{i=1}^{N_t} Y_i - \frac{1}{N_c}\sum_{j\in\mathbb{I}_c} Y_j$$

- Notice that we matched 1 treated to 1 control exactly, so we have:

$$\widehat{\tau}_m = \frac{1}{N_t}\sum_{i=1}^{N_t}(Y_i - Y_{j(i)})$$

- $\rightsquigarrow$ average of the within matched-pair differences.

Beyond exact matching

- With high-dimensional X_i, not feasible to exact match.
- Let S be a matching solution: a subset of the data produced by the matching procedure: $(\mathbb{I}_t, \mathbb{I}_c)$.
- Suppose that this procedure produces balance:

$$T_i \perp\!\!\!\perp X_i | S$$

- With no unmeasured confounders we have:

$$\big(Y_i(0), Y_i(1)\big) \perp\!\!\!\perp T_i | S$$

- Balance is checkable $\rightsquigarrow$ are T_i and X_i related in the matched data?

The matching procedure

1. Choose a number of matches
2. Choose a distance metric
3. Find matches (drop non-matches)
4. Check balance
5. Repeat (1)-(4) until balance is acceptable
6. Calculate the effect of the treatment on the outcome in the matched dataset.

More than 1 control match

- What if we have enough controls to have M matched controls per treated?
  - $P(X_i = x | T_i = 1) = P(X_i = x | T_i = 0, \mathbb{I}_c)$ because M is constant across treated units.
- Now, $J_M(i)$ is a set of M control matches. Use these to "impute" missing potential outcome.
- For $i \in \mathbb{I}_t$ define:

$$\widehat{Y}_i(0) = \frac{1}{M} \sum_{j \in J_M(i)} Y_j$$

- New estimator for the effect:

$$\widehat{\tau}_m = \frac{1}{N_t} \sum_{i=1}^{N_t} (Y_i - \widehat{Y}_i(0))$$

- Under no unmeasured confounding, $\widehat{Y}_i(0)$ is a good predictor of the true potential outcome under control, Y_i.

Number of matches

- How many control matches should we include?
  - Small $M \rightsquigarrow$ small sample sizes
  - Large $M \rightsquigarrow$ worse matches (each additional match is further away).
- If M varies by treated unit, need to weight observations to ensure balance.

With or without replacement

- Matching with replacement: a single control unit can be matched to multiple treated units
- Benefits:
  - Better matches!
  - Order of matching does not matter.
- Drawbacks:
  - Inference is more complicated.
  - $\rightsquigarrow$ need to account for multiple appearances with weights.
  - Potentially higher uncertainty (using the same data multiple times = relying on less data).

Defining closeness

- We want to find control observations that are similar to the treated unit on X_i.
- How do we define distance/similarity on X_i if it is high dimensional?
- We need a distance metric which maps two covariates vectors into a single number.
  - Lower values $\rightsquigarrow$ more similar values of X_i.
  - Choice of distance metric will lead to different matches.

Exact distance metric

- Exact: only match units to other units that have the same exact values of X_i.

$$D_{ij} = \begin{cases} 0 & \text{if } X_i = X_j \\ \infty & \text{if } X_i \neq X_j \end{cases}$$

Euclidean distance

- The normalized Euclidean distance metric just uses the sum of the normalized distances for each covariate.
  - "Closeness" is standardized across covariates.
- Suppose that $X_i = (X_{i1}, \dots, X_{iK})'$, so that there are K covariates.
- Then the Euclidean distance metric is:

$$D_{ij} = \sqrt{\sum_{k=1}^{K} \frac{(X_{ik} - X_{jk})^2}{\widehat{\sigma}_k^2}}$$

- Here, $\widehat{\sigma}_k^2$ is the variance of the kth variable:

$$\widehat{\sigma}_k^2 = \frac{1}{N-1} \sum_{i=1}^{N} (X_{ik} - \bar{X}_k)$$

Mahalanobis distance

- Mahalanobis distance: Euclidean distance adjusted for covariance in the data.
- Intuition: if X_{ik} and $X_{ik'}$ are two covariates that are highly correlated, then their contribution to the distances should be lower.
  - Easy to get close on correlated covariates $\rightsquigarrow$ downweight.
  - Harder to get close on uncorrelated covariates $\rightsquigarrow$ upweight.
- Metric:

$$D_{ij} = \sqrt{(X_i - X_j)'\widehat{\Sigma}^{-1}(X_i - X_j)}$$

- $\widehat{\Sigma}$ is the estimated variance-covariance matrix of the observations:

$$\widehat{\Sigma} = \frac{1}{N}\sum_{i=1}^{N}(X_i - \bar{X})(X_i - \bar{X})^T$$

Complications

- Combining distance metrics:
  - Exact on race/gender, Mahalanobis on the rest.
- Some matches are too far on the distance metric.
  - Dropping those matches (treated and control) improves balance.
  - Dropping treated units changes the quantity of interest.
- Implementation: a caliper, which is the maximum distance we would accept

Estimands

- Matching easiest to justify for the Average Treatment Effect on the Treated.
  - Dropping control units doesn't affect this identification.
- Can also identify the Average Treatment Effect on Control by finding matched treated units for the controls.
- Combine the two to get the ATE:

$$\tau = \tau_{ATT} P(T_i = 1) + \tau_{ATC} P(T_i = 0)$$

- Estimated:

$$\widehat{\tau} = \widehat{\tau}_{ATT} \left(\frac{N_t}{N} \right) + \widehat{\tau}_{ATC} \left(\frac{N_c}{N} \right)$$

Moving the goalposts

- Common support: finding the subspace of X_i where there is overlap between the treated and control groups.
  - Hard to extrapolate outside region.
  - Theoretical: effect of voting for those under 18 $(P(D_i = 1|X_i < 18) = 0)$.
  - Empirical: no/extremely few treated units in a sea of controls.
  - Solution: restrict analysis to common support (dropping treated and controls).
- Moving the goalposts: dropping treated units.
  - We move away from being able to identify the Average Treatment Effect on the Treated (ATT).
  - Now it's the ATT in the matched subsample (sometimes called the feasible ATT).
  - Good to be clear about this.

Matching methods

- Now that we have distances between all units, we just need to match!
- For a particular unit, easy:

$$j(i) = \arg\min_{j \in \mathbb{J}_c} D_{ij}$$

  - $\mathbb{J}_c$ are the available controls for matching.
- This is nearest neighbor: "Find the control unit with the smallest distance metric."
- Do the same for all treated units.
- What about ties?
  - Randomly choose between them.
- Note: in nearest neighbor without replacement the order matters!

Assessing balance

- All matching methods seek to maximize balance:

$$P(X_i = x | T_i = 1, S) = P(X_i = x | T_i = 0, S)$$

- Choice of balance metric will determine which matching method does better.
  - If you use Mahalanobis distance as the balance metric, then matching on the Mahalanobis score will do well because that's what it's designed to do.
- Options:
  - Differences-in-means/medians, standardized.
  - Quantile-quantile plots/KS statistics for comparing the entire distribution of X_i.
  - L_1: multivariate histogram.

1. The Experimental Ideal
2. Identification with Measured Confounding
   - Design
   - DAGs
3. Stratification
4. Matching
   - Fundamentals of Matching
   - Two Approaches to Matching
5. Regression
   - Regression with Heterogeneous Effects
   - Imputation Estimators
   - Fun With Weights
6. Estimands

Two Approaches to Matching

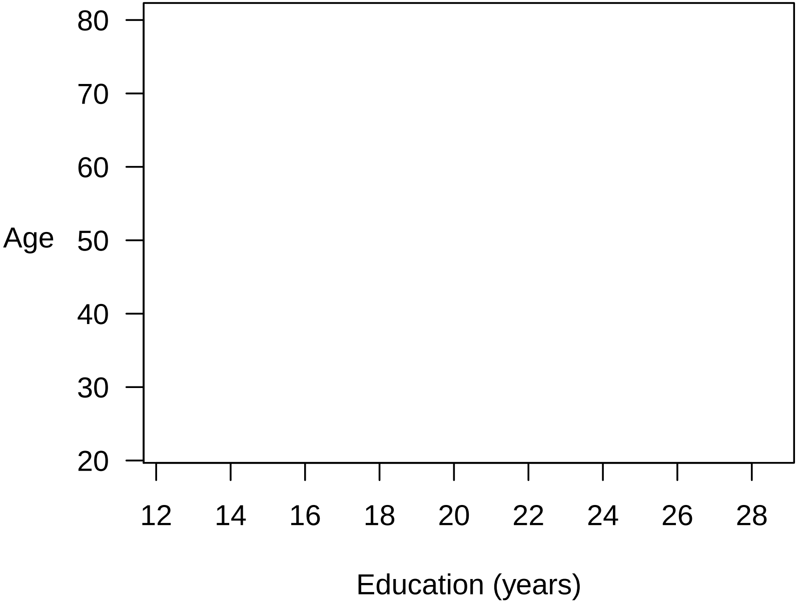
- There are many approaches to matching. We will cover just two for the sake of time.
- This isn't a statement that these are the best two, just a set which are straightforward to learn.
- Which is the best method? The one that produces the best balance!

Method 1: Mahalanobis Distance Matching

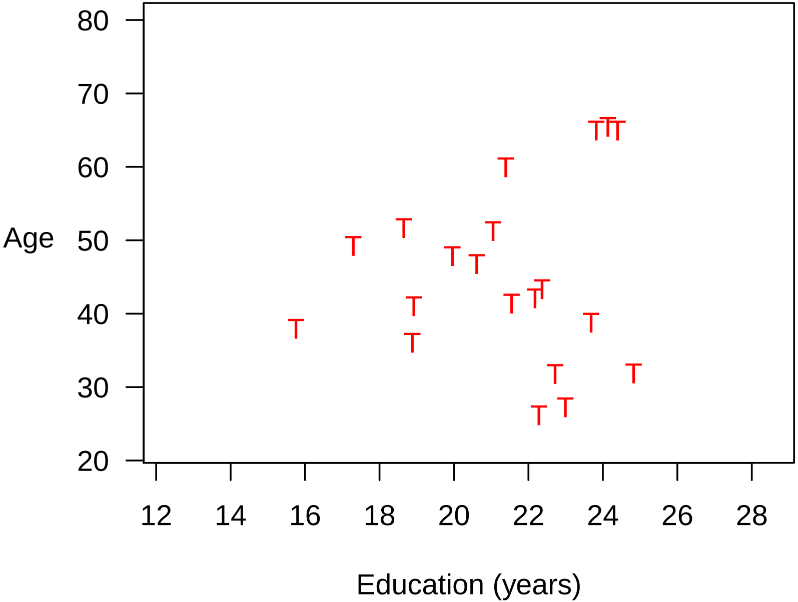
(Approximates Fully Blocked Experiment)

1. Preprocess (Matching)
   - Distance$(X_i, X_j) = \sqrt{(X_i - X_j)' S^{-1} (X_i - X_j)}$
   - Match each treated unit to the nearest control unit
   - Control units: not reused; pruned if unused
   - Prune matches if Distance>caliper
2. Checking Measure imbalance, tweak, repeat, . . .
3. Estimation Difference in means or a model

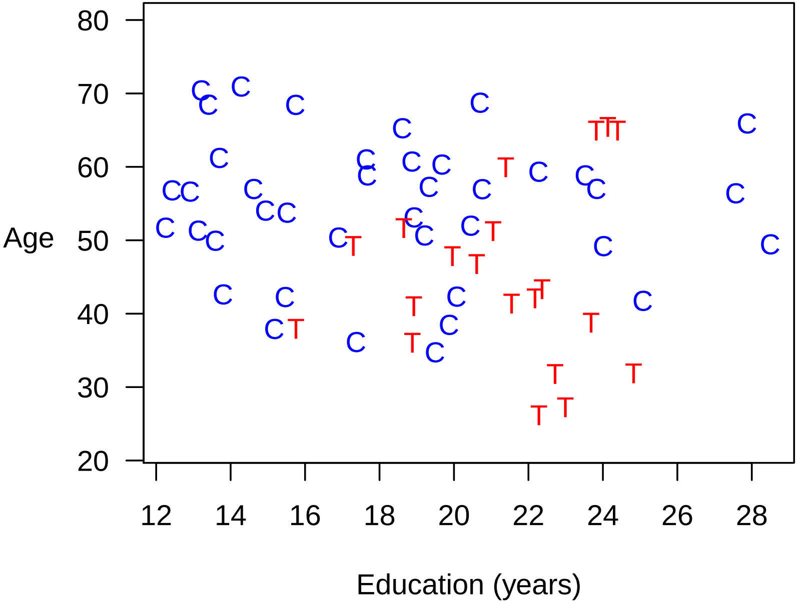
Mahalanobis Distance Matching



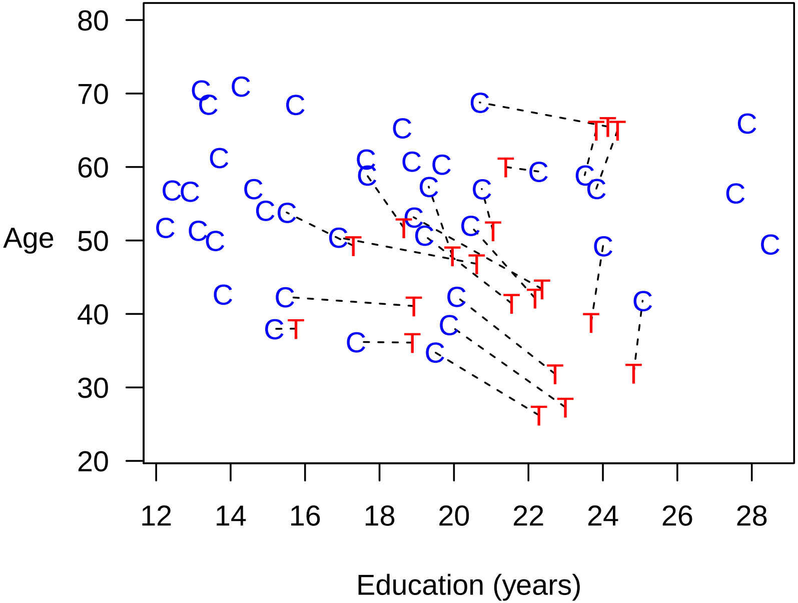
Mahalanobis Distance Matching



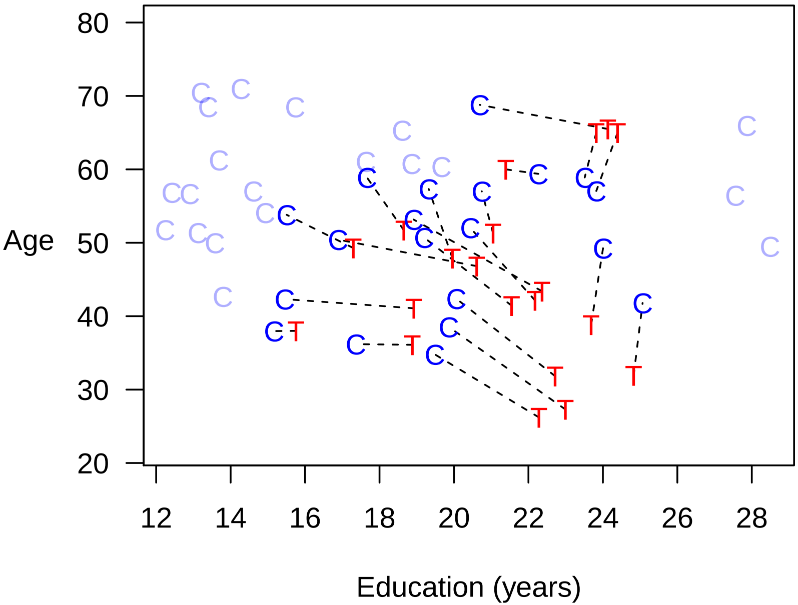
Mahalanobis Distance Matching



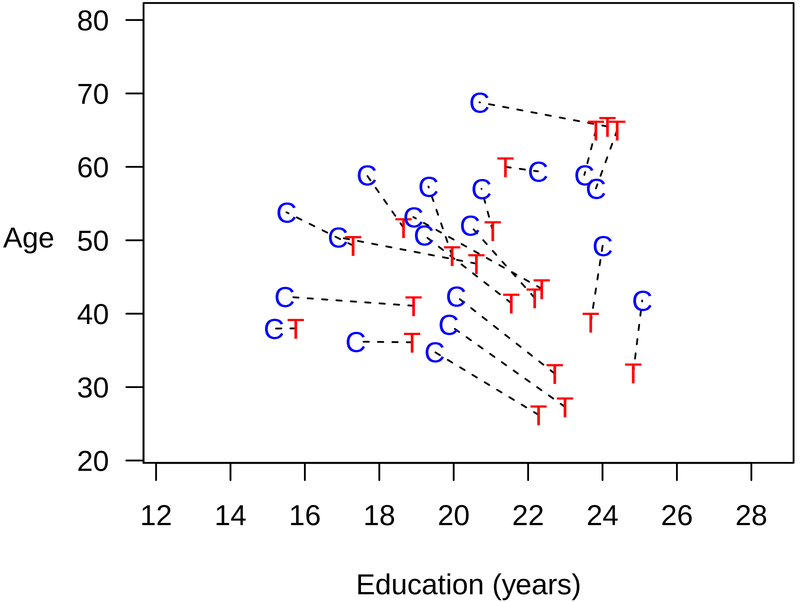
Mahalanobis Distance Matching



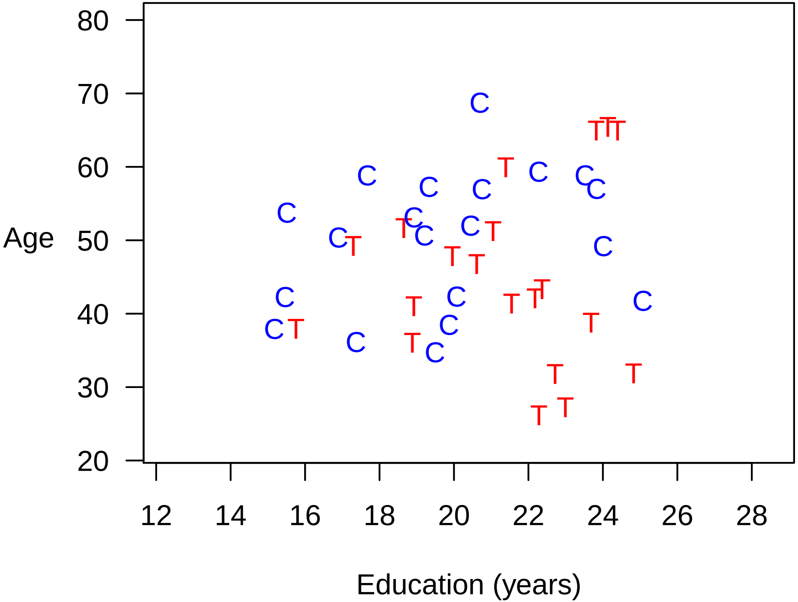
Mahalanobis Distance Matching



Mahalanobis Distance Matching



Mahalanobis Distance Matching



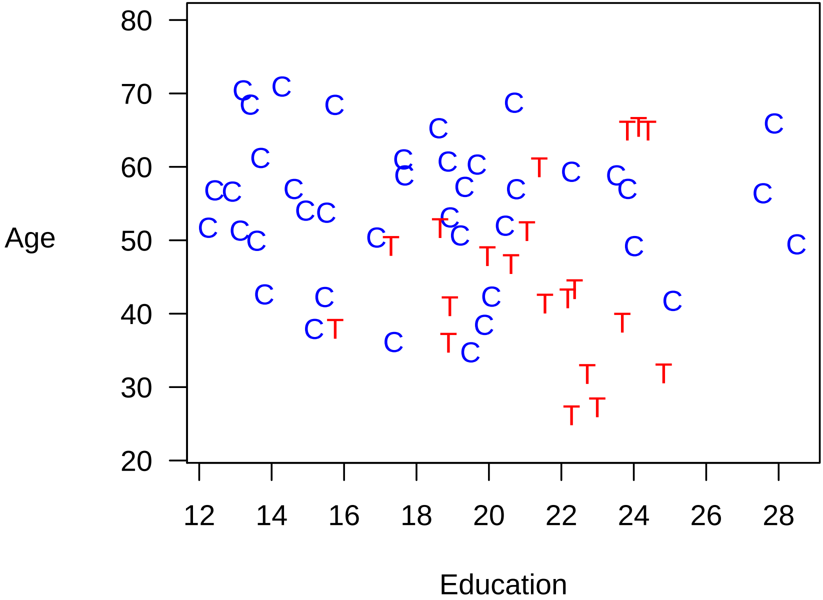
Method 2: Coarsened Exact Matching

(Approximates Fully Blocked Experiment)

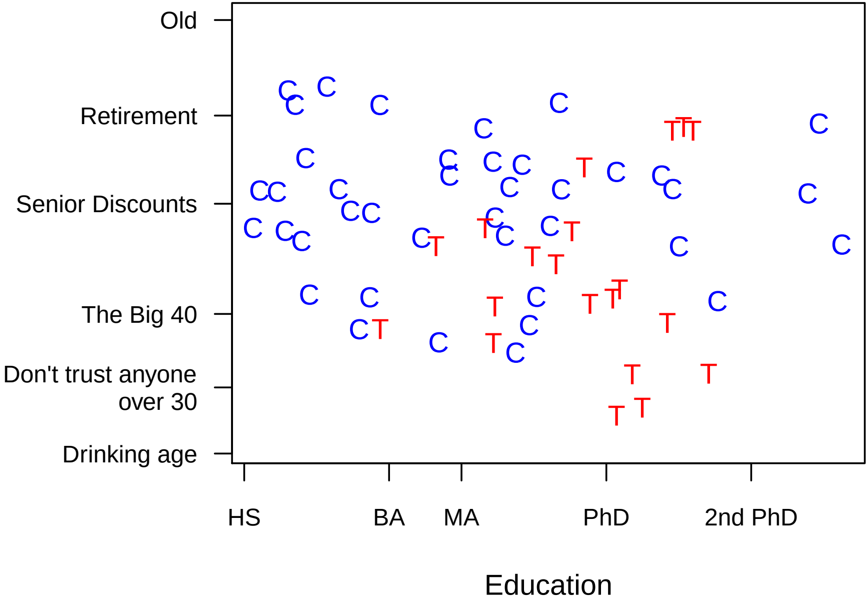
1. Preprocess (Matching)
   - Temporarily coarsen X as much as you're willing
     - ★ e.g., Education (grade school, high school, college, graduate)
   - Apply exact matching to the coarsened X, $C(X)$
     - ★ Sort observations into strata, each with unique values of $C(X)$
     - ★ Prune any stratum with 0 treated or 0 control units
   - Pass on original (uncoarsened) units except those pruned
2. Checking Determine matched sample size, tweak, repeat, . . .
   - Easier, but still iterative
3. Estimation Difference in means or a model
   - Need to weight controls in each stratum to equal treateds

Coarsened Exact Matching

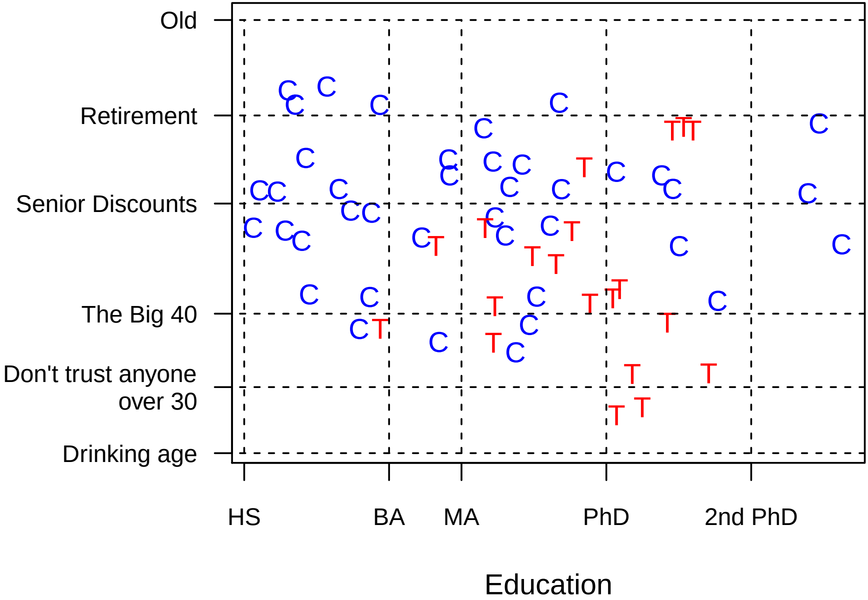
Coarsened Exact Matching



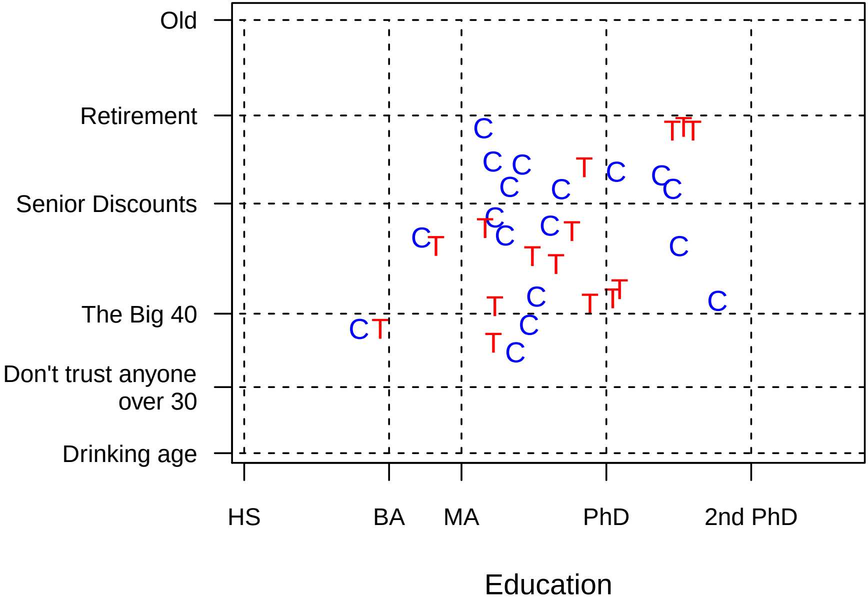
Coarsened Exact Matching



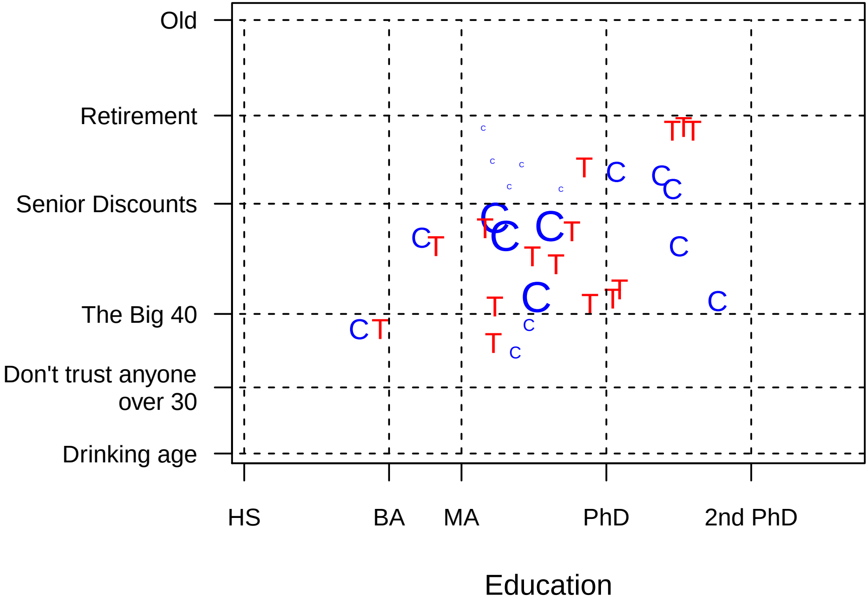
Coarsened Exact Matching



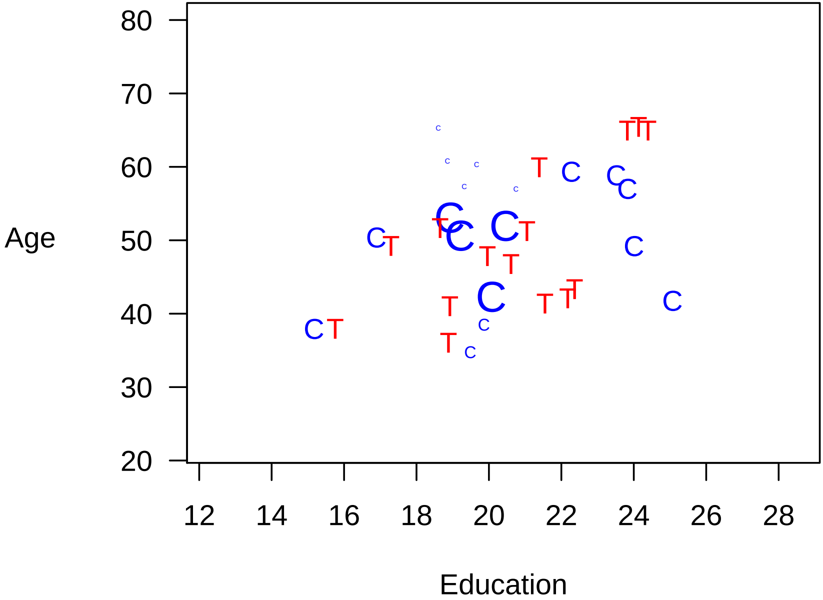
Coarsened Exact Matching



Coarsened Exact Matching



Coarsened Exact Matching



Final Thoughts on Matching

- Matching is an entire literature unto itself (you could probably teach a whole class just on that). There is so much we didn't cover here including bias corrections and how to get variance estimators.
- The central advantage of matching is that it is transparent about where the counterfactual estimates are coming from (you can look at the matched unit!).
- Technically the thing it is buying you relative to regression methods we will talk about next is that it limits extrapolation.
- But there is no need for these techniques to compete—we can match and then use regression!
- Importantly, there is nothing magic about matching, it is just another way of conditioning.

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

- Last Week
  - frameworks for causal inference
- This Week
  - experimental ideal
  - identification with measured confounding
  - estimation via stratification, matching and regression
- Next Week
  - approaches with unmeasured confounding
- Long Run
  - causal frameworks → inference → regression → causal inference

1. The Experimental Ideal

2. Identification with Measured Confounding
   - Design
   - DAGs

3. Stratification

4. Matching
   - Fundamentals of Matching
   - Two Approaches to Matching

5. Regression
   - Regression with Heterogeneous Effects
   - Imputation Estimators
   - Fun With Weights

6. Estimands

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 $Y_i = \beta_0 + \tau T_i + X_i'\beta + u_i$.

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 $E[Y|T=1,X] - E[Y|T=0,X]$. Approximation may be very poor if $E[Y|T,X]$ is misspecified and then τ may be biased for the ATE.
3. Heterogeneous treatment effects (τ differs for different values of X)
   - even If outcomes are linear in X, τ converges to the conditional-variance-weighted average of the underlying causal effects rather than the ATE.

Ideal Case: Constant Effects Model With Linearly Separable Confounding

Assume a data generating process with constant effects and linearly separable confounding:

$$Y_i(t) = Y_i = \beta_0 + \tau T_i + \eta_i$$

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

$$\begin{aligned} Y_i &= \beta_0 + \tau T_i + \eta_i \\ &= \beta_0 + \tau T_i + X_i' \beta + u_i \end{aligned}$$

- Thus, a regression where T_i and X_i are entered linearly can recover the ATE. (The regression model matches the data generating process)

Heterogeneous effects, binary treatment

- Completely randomized experiment:

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

- 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, $E[\varepsilon_i | T_i] = 0$
- Thus, OLS estimates the ATE with no covariates.

Adding covariates

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

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

- 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:

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

- 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:
$$\tau_R = \frac{\text{Cov}(Y_i, T_i - E[T_i|X_i])}{\mathrm{Var}(T_i - E[T_i|X_i])}$$
- $T_i - E[T_i|X_i]$ is the residual from a regression of T_i on the full set of dummies.
- With a little work we can show:
$$\tau_R = \frac{E\left[\tau(X_i)(T_i - E[T_i|X_i])^2\right]}{E[(T_i - E[T_i|X_i])^2]} = \frac{E[\tau(X_i)\sigma_t^2(X_i)]}{E[\sigma_t^2(X_i)]}$$
- $\sigma_t^2(x) = \mathrm{Var}[T_i|X_i = x]$ is the conditional variance of treatment assignment.

ATE versus OLS

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

- Compare to the ATE:

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

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

Regression weighting

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

- 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 T_i is binary, then we know the conditional variance will be:

$$\sigma_t^2(x) = P[T_i = 1|X_i = x]\,(1 - P[T_i = 1|X_i = x])$$

- Maximum variance with $P[T_i = 1|X_i = x] = 1/2$.

OLS weighting example

- Binary covariate:

$$\begin{array}{rr} \text{Group 1} & \text{Group 2} \\ P[X_i = 1] = 0.75 & P[X_i = 0] = 0.25 \\ P[T_i = 1 | X_i = 1] = 0.9 & P[T_i = 1 | X_i = 0] = 0.5 \\ \sigma_t^2(1) = 0.09 & \sigma_t^2(0) = 0.25 \\ \tau(1) = 1 & \tau(0) = -1 \end{array}$$

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

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

When will OLS estimate the ATE?

- When does $\tau = \tau_R$? (where τ_R is the estimate from a regression)
- Constant treatment effects: $\tau(x) = \tau = \tau_R$
- Constant probability of treatment: $e(x) = P[T_i = 1 | X_i = x] = e$.
  - 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?
  - Accept the bias (might be relatively small with saturated models)
  - Use a different regression approach
- Let $\mu_t(x) = E[Y_i(t)|X_i = x]$ be the CEF for the potential outcome under $T_i = t$.
- By SUTVA and no unmeasured confounders, we have $\mu_t(x) = E[Y_i|T_i = t, X_i = x]$.
- Estimate a regression of Y_i on X_i among the $T_i = t$ group.
- Then, $\hat{\mu}_d(x)$ is just a predicted value from the regression for $X_i = x$.
- How can we use this?

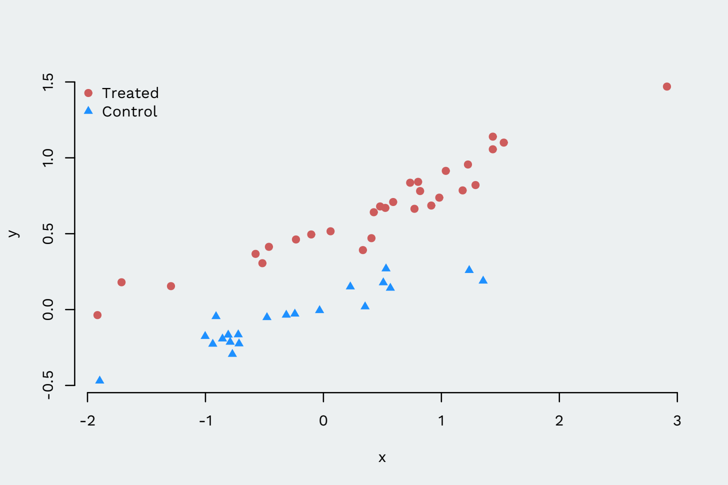
Imputation estimators

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

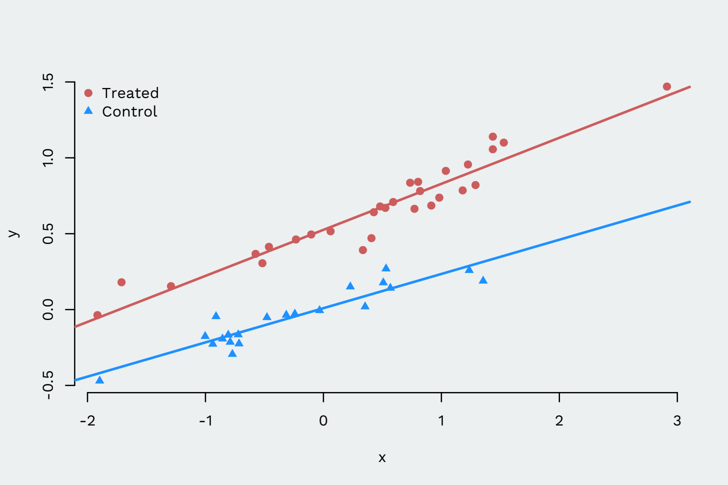
$$\tau_{imp} = \frac{1}{N} \sum_i \hat{\mu}_1(X_i) - \hat{\mu}_0(X_i)$$

- Sometimes called an **imputation estimator**.

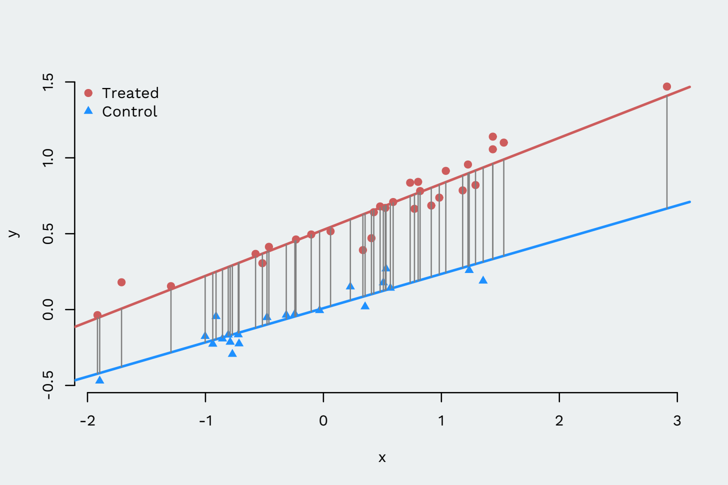
Imputation estimator visualization



Imputation estimator visualization

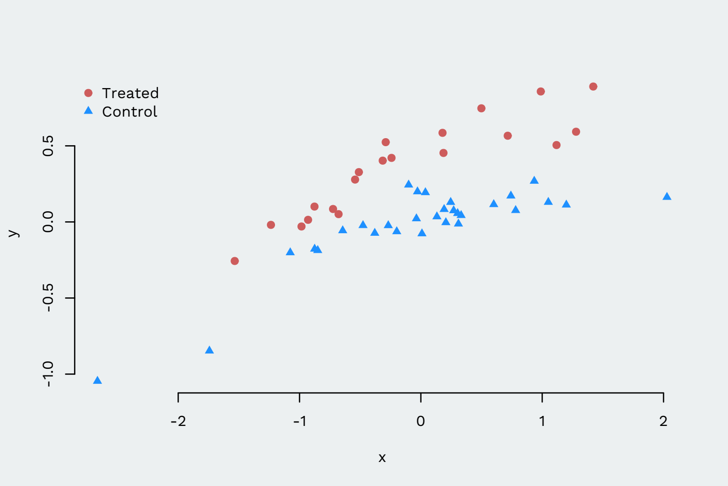


Imputation estimator visualization



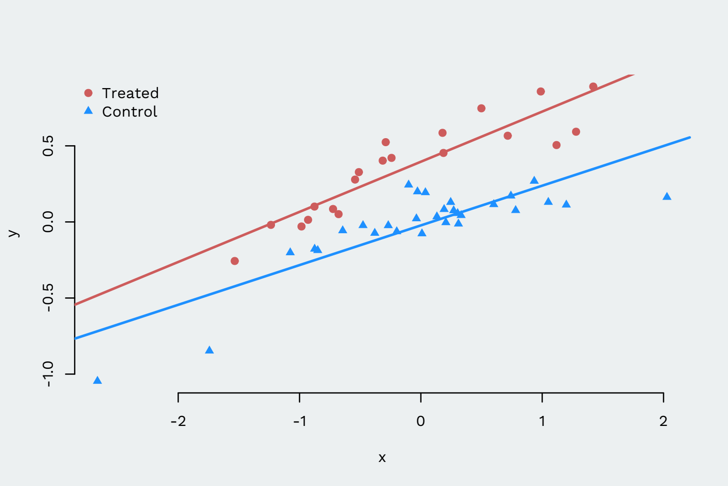
Nonlinear relationships

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



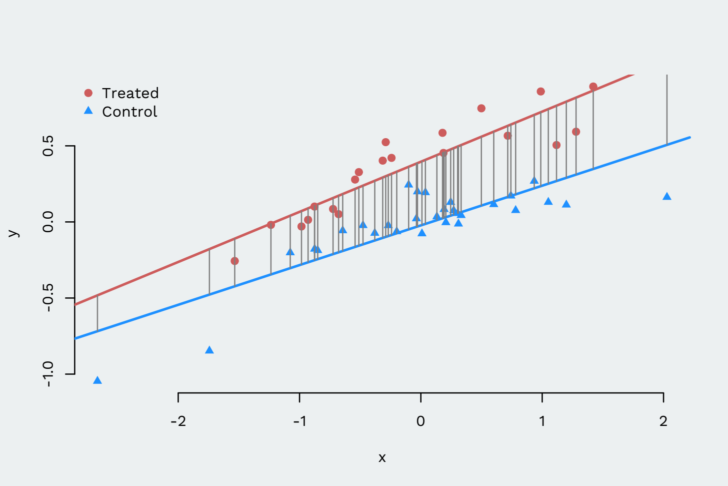
Nonlinear relationships

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



Nonlinear relationships

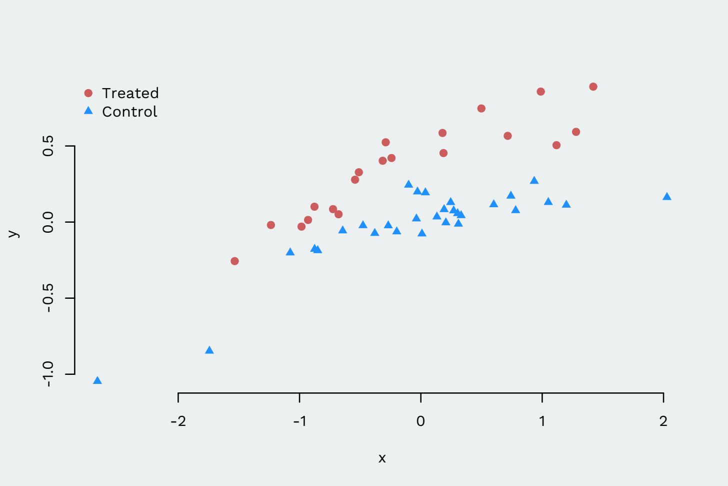
- Same idea but with nonlinear relationship between Y_i and X_i:



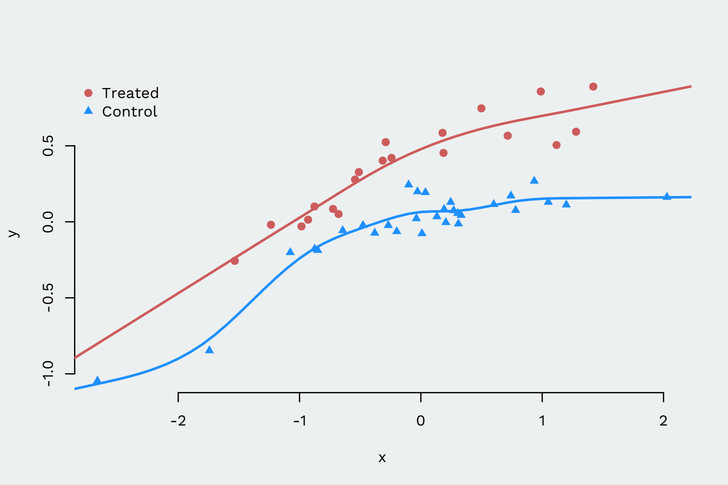
Using semiparametric regression

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

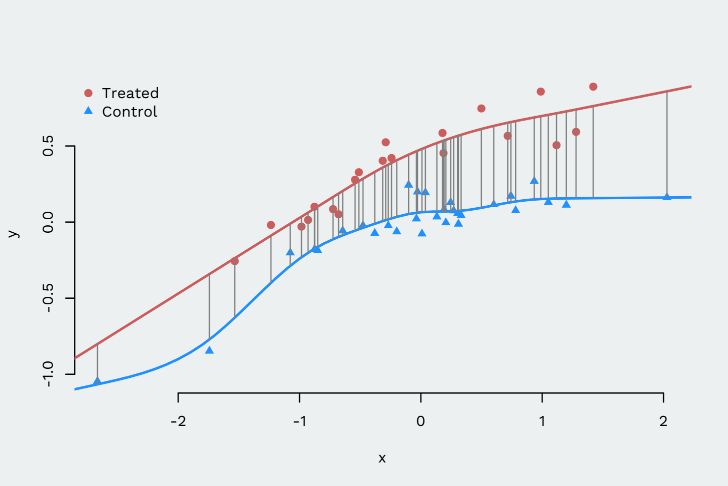
Using GAMs



Using GAMs



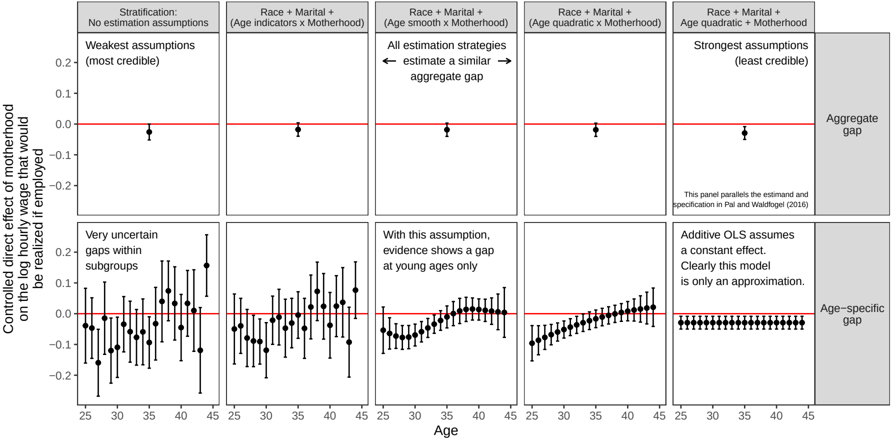
Using GAMs



Imputation Estimators

- Imputation estimators (also called the parametric g-formula) are a great and underutilized technique (to learn more, check out Naimi, Cole, and Kennedy, 2017).
- It is harder to implement than vanilla OLS particularly for uncertainty estimation, but you can always bootstrap!
- If $\hat{\mu}_t(x)$ are consistent estimators, then τ_{imp} is consistent for the ATE.
- To be flexible, people are increasingly using machine learning techniques like: kernel regression, neural networks, regression trees, etc.
- As we just saw, GAMs are a nice trade-off of the ease vs. flexibility side.
- These kinds of things will tend to matter a lot more for conditional treatment effects than the overall aggregate treatment effect, but you also don't know for sure until you try.

Example from Lundberg, Johnson and Stewart



All the Steps Together

1) Set the target. Define a theoretical estimand. Requires substantive **argument**.

Average difference in the **potential outcome** each woman *i* would realize

$$\tau = \frac{1}{n}\sum_{i=1}^{n}\left(\overset{\text{if she were an employed mother}}{Y_i(\text{Mother, Employed})} \quad \overset{\text{versus}}{-} \quad \overset{\text{if she were an employed non-mother}}{Y_i(\text{Non-mother, Employed})} \right)$$

2) Link to observables. Define an empirical estimand. Requires conceptual **assumptions**.

Average difference in the **realized outcomes** of women with the covariates $\vec{x}_i$ of women *i* who

$$\theta = \frac{1}{n}\sum_{i=1}^{n}\left(\overset{\text{actually are mothers}}{E(Y \mid \vec{X} = \vec{x}_i, \text{Motherhood} = \text{Mother})} \quad \overset{\text{versus}}{-} \quad \overset{\text{actually are not mothers}}{E(Y \mid \vec{X} = \vec{x}_i, \text{Motherhood} = \text{Non-mother})} \right)$$

3) Learn from data. Select an estimation strategy. Requires statistical **evidence**.

Average difference in the **regression prediction** at the covariates $\vec{x}_i$ of women *i* if we

$$\underset{\substack{\uparrow \\ \text{estimate of} \\ \text{the estimand}}}{\hat{\theta}} = \frac{1}{n}\sum_{i=1}^{n}\left(\overset{\text{recode as a mother}}{\underbrace{\hat{E}(Y \mid \vec{X} = \vec{x}_i, \text{Motherhood} = \text{Mother})}_{\substack{\uparrow \\ \text{estimated } \hat{Y}_i(\text{Mother})}}} \quad \overset{\text{versus}}{-} \quad \overset{\text{recode as not a mother}}{\underbrace{\hat{E}(Y \mid \vec{X} = \vec{x}_i, \text{Motherhood} = \text{Non-mother})}_{\substack{\uparrow \\ \text{estimated } \hat{Y}_i(\text{Non-mother})}}} \right)$$

Connections to the Coefficient In Regression

$$\hat{E}(Y \mid \vec{X} = \vec{x}_i, \text{Motherhood}) = \begin{cases} \hat\alpha \qquad + \vec{x}_i'\hat{\vec\gamma} & \text{if Motherhood} = \text{Non-Mother} \\ \hat\alpha + \hat\beta + \vec{x}_i'\hat{\vec\gamma} & \text{if Motherhood} = \text{Mother} \end{cases}$$

Intercept

Coefficient on motherhood

Coefficients on other covariates

By parametric approximation

$$\hat\theta = \frac{1}{n}\sum_{i=1}^{n} \left(\overbrace{\left(\hat\alpha + \hat\beta + \vec{x}_i'\hat{\vec\gamma}\right)}^{\hat{Y}_i(\text{Mother})} - \overbrace{\left(\hat\alpha + \vec{x}_i'\hat{\vec\gamma}\right)}^{\hat{Y}_i(\text{Non-mother})} \right)$$

$$= \frac{1}{n}\sum_{i=1}^{n} \underbrace{\left(\hat\alpha + \vec{x}_i'\hat{\vec\gamma} - \hat\alpha - \vec{x}_i'\hat{\vec\gamma}\right)}_{\substack{\text{Cancels because model} \\ \text{assumes no interactions}}} + \frac{1}{n}\sum_{i=1}^{n} \hat\beta$$

$$= \hat\beta \qquad \leftarrow \text{coefficient on motherhood}$$

1. The Experimental Ideal
2. Identification with Measured Confounding
   - Design
   - DAGs
3. Stratification
4. Matching
   - Fundamentals of Matching
   - Two Approaches to Matching
5. Regression
   - Regression with Heterogeneous Effects
   - Imputation Estimators
   - Fun With Weights
6. Estimands

Fun With Weights

Aronow, Peter M., and Cyrus Samii. "Does Regression Produce Representative Estimates of Causal Effects?." *American Journal of Political Science* (2015).[2]

- Imagine we care about the possibly heterogeneous causal effect of a treatment T and we control for some covariates X?
- We can express the regression as a weighting over individual observation treatment effects where the weight depends only on X.
- Useful technology for understanding what our models are identifying off of by showing us our effective sample.

[2] I'm grateful to Peter Aronow for sharing his slides, several of which are used here.

How this works

We start by asking what the estimate of the average causal effect of interest converges to in a large sample:

$$\hat{\beta} \xrightarrow{p} \frac{E[w_i \tau_i]}{E[w_i]} \text{ where } w_i = (T_i - E[T_i|X])^2 \,,$$

so that $\hat{\beta}$ converges to a reweighted causal effect. As $E[w_i|X_i] = \text{Var}[T_i|X_i]$, we obtain an average causal effect reweighted by conditional variance of the treatment.

Estimation

A simple, consistent plug-in estimator of w_i is available: $\hat{w}_i = \tilde{T}_i^2$ where $\tilde{T}_i$ is the residualized treatment. (the proof is connected to the partialing out strategy)

Easily implemented in R:

```
wts <- (t - predict(lm(t~x)))^2
```

Implications

- Unpacking the black box of regression gives us substantive insight
- When some observations have no weight, this means that the covariates completely explain their treatment condition.
- This is a feature, not a bug, of regression: we can't learn anything from those cases anyway (i.e. it is automatically handling issues of common support).
- The downside is that we have to be aware of what happened!

Application

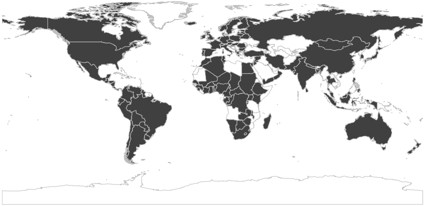
Jensen (2003), "Democratic Governance and Multinational Corporations: Political Regimes and Inflows of Foreign Direct Investment."

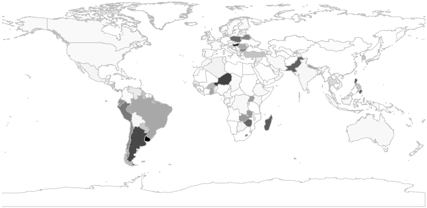
Jensen presents a large-N TSCS-analysis of the causal effects of governance (as measured by the Polity III score) on Foreign Direct Investment (FDI).

The nominal sample: 114 countries from 1970 to 1997.

Jensen estimates that a 1 unit increase in polity score corresponds to a 0.020 increase in net FDI inflows as a percentage of GDP ($p < 0.001$).

Nominal and Effective Samples





Over 50% of the weight goes to just 12 (out of 114) countries.

Broader Implications

When causal effects are heterogeneous, we can draw a distinction between "internally valid" and "externally valid" estimates of an Average Treatment Effect (ATE).

- "Internally valid": reliable estimates of ATEs, but perhaps not for the population you care about
  - randomized (lab, field, survey) experiments, instrumental variables, regression discontinuity designs, other natural experiments
- "Externally valid": perhaps unreliable estimates of ATEs, but for the population of interest
  - large-N analyses, representative surveys

Broader Implications

Aronow and Samii argue that analyses which use regression, even with a representative sample, have no greater claim to external validity than do [natural] experiments.

- When a treatment is “as-if” randomly assigned conditional on covariates, regression distorts the sample by implicitly applying weights.
- The effective sample (upon which causal effects are estimated) may have radically different properties than the nominal sample.
- When there is an underlying natural experiment in the data, a properly specified regression model may reproduce the internally valid estimate associated with the natural experiment.

We Covered

- Regression based estimation of causal effects under measured confounding.
- Imputation estimators which generalize to broader class of machine learning estimators for the conditional expectation function.
- A fun with about how the weighting interpretation can tell you about your sample.

Next Time: Estimands

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

- Last Week
  - frameworks for causal inference
- This Week
  - experimental ideal
  - identification with measured confounding
  - estimation via stratification, matching and regression
- Next Week
  - approaches with unmeasured confounding
- Long Run
  - causal frameworks → inference → regression → causal inference

1. The Experimental Ideal

2. Identification with Measured Confounding
   - Design
   - DAGs

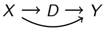
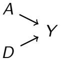
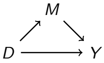
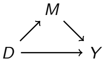
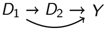
3. Stratification

4. Matching
   - Fundamentals of Matching
   - Two Approaches to Matching

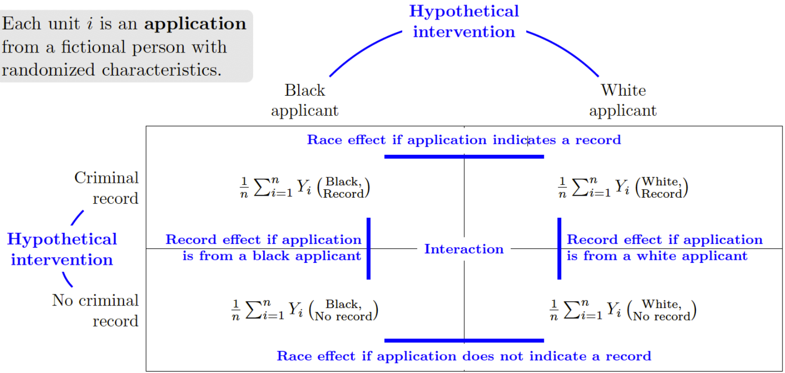
5. Regression
   - Regression with Heterogeneous Effects
   - Imputation Estimators
   - Fun With Weights

6. Estimands

Estimands

Estimand name	Mathematical statement	DAG	Reference	Colloquial terms
Average treatment effect	$\frac{1}{n}\sum_i Y_i(d') - Y_i(d)$	$D \longrightarrow Y$	Morgan and Winship (2015)	Effect
Conditional average treatment effect	$\frac{1}{n_x}\sum_{i:X_i=x}(Y_i(d') - Y_i(d))$		Athey and Imbens (2016)	Effect heterogeneity or moderation
Causal interaction	$\frac{1}{n}\sum_i \left(\left(Y_i(a',d') - Y_i(a',d) \right) - \left(Y_i(a,d') - Y_i(a,d) \right) \right)$		Vanderweele 2015	Joint treatment effect
Controlled direct effect	$\frac{1}{n}\sum_i \left(Y_i(d',m) - Y_i(d,m) \right)$		Acharya Blackwell and Sen (2016)	Mediation
Natural direct effect	$\frac{1}{n}\sum_i \left(Y_i(d',M_i(d)) - Y_i(d,M_i(d)) \right)$		Imai et al 2011	Mediation
Effect of dynamic treatment regime	$\frac{1}{n}\sum_i Y_i(d_1',d_2') - Y_i(d_1,d_2)$		Wodtke et al 2011	Cumulative effect

Each unit i is an **application** from a fictional person with randomized characteristics.

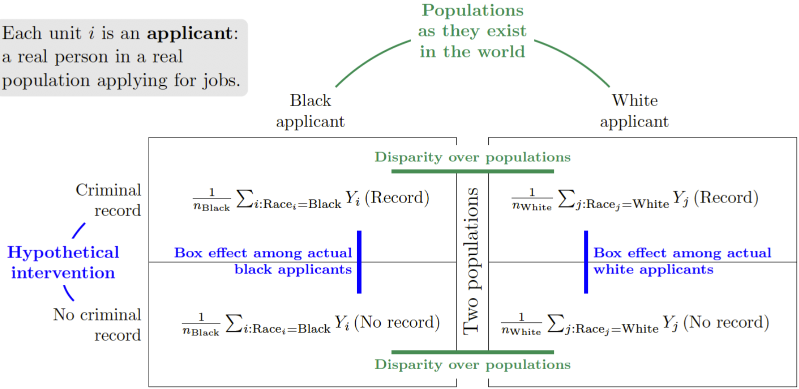


Hypothetical intervention

	Black applicant	White applicant
Criminal record	$\frac{1}{n}\sum_{i=1}^{n} Y_i\binom{\text{Black,}}{\text{Record}}$	$\frac{1}{n}\sum_{i=1}^{n} Y_i\binom{\text{White,}}{\text{Record}}$
No criminal record	$\frac{1}{n}\sum_{i=1}^{n} Y_i\binom{\text{Black,}}{\text{No record}}$	$\frac{1}{n}\sum_{i=1}^{n} Y_i\binom{\text{White,}}{\text{No record}}$

Hypothetical intervention
Race effect if application indicates a record
Record effect if application is from a black applicant
Interaction
Record effect if application is from a white applicant
Race effect if application does not indicate a record

Each unit i is an **applicant**: a real person in a real population applying for jobs.



Populations as they exist in the world

	Black applicant	White applicant
Criminal record	$\frac{1}{n_{\text{Black}}}\sum_{i:\text{Race}_i=\text{Black}} Y_i\,(\text{Record})$	$\frac{1}{n_{\text{White}}}\sum_{j:\text{Race}_j=\text{White}} Y_j\,(\text{Record})$
No criminal record	$\frac{1}{n_{\text{Black}}}\sum_{i:\text{Race}_i=\text{Black}} Y_i\,(\text{No record})$	$\frac{1}{n_{\text{White}}}\sum_{j:\text{Race}_j=\text{White}} Y_j\,(\text{No record})$

Hypothetical intervention

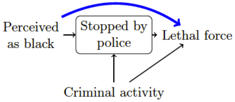
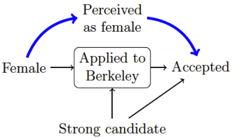
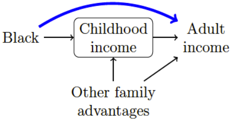
Disparity over populations

Box effect among actual black applicants

Two populations

Box effect among actual white applicants

Disparity over populations

Study	Empirical regularity	Misleading conclusion	Directed Acyclic Graph
Fryer (2019)	Among those they stop, police shoot the same proportion of black individuals as white individuals.	Police do not discriminate against black individuals when using lethal force.	 Perceived as black Stopped by police Lethal force Criminal activity
Bickel et al. (1975)	Among those who apply, Berkeley departments admit a higher proportion of women than of men.	Admissions committees do not discriminate against women.	 Perceived as female Female Applied to Berkeley Accepted Strong candidate
Chetty et al. (2020)	Among those with equal childhood incomes, black and white women earn similar amounts as adults.	Equalizing childhood incomes would eliminate the racial gap in women's adult incomes.	 Black Childhood income Adult income Other family advantages

Misunderstandings Between Experimentalists and Observationalists (Imai, King, Stuart 2008)

- Despite a common framework, there are still disagreements between experimental and observational design approaches
- Mostly related to the debate between internal and external validity of estimates
- Most researchers are inherently interested in Population Average Treatment Effects (PATE)

Decomposition of Causal Effect Estimation Error

- Difference in means estimator:

$$D \equiv \left(\frac{1}{n/2}\sum_{i\in\{I_i=1,T_i=1\}} Y_i\right) - \left(\frac{1}{n/2}\sum_{i\in\{I_i=1,T_i=0\}} Y_i\right).$$

- Estimation Error:

$$\Delta \equiv \text{PATE} - D$$

- Pretreatment confounders: X are observed and U are unobserved
- Decomposition:

$$\begin{aligned}\Delta &= \Delta_S + \Delta_T \\ &= (\Delta_{S_X} + \Delta_{S_U}) + (\Delta_{T_X} + \Delta_{T_U})\end{aligned}$$

Error due to Δ_S (sample selection), Δ_T (treatment imbalance), and each due to observed (X_i) and unobserved (U_i) covariates

Note: Analogous decompositions hold for other estimands of interest.

Selection Error

- Definition:

$$\Delta_S \equiv \text{PATE} - \text{SATE} = \frac{N-n}{N}(\text{NATE} - \text{SATE}),$$

  where NATE is the nonsample average treatment effect.

- Δ_S vanishes if:
  1. The sample is a census ($I_i = 1$ for all observations and $n = N$);
  2. SATE $=$ NATE; or
  3. Switch quantity of interest from PATE to SATE

Decomposing Selection Error

- Decomposition:
$$\Delta_S = \Delta_{S_X} + \Delta_{S_U}$$
- $\Delta_{S_X} = 0$ when empirical distribution of (observed) X is identical in population and sample: $\widetilde{F}(X \mid I = 0) = \widetilde{F}(X \mid I = 1)$.
- $\Delta_{S_U} = 0$ when empirical distribution of (unobserved) U is identical in population and sample: $\widetilde{F}(U \mid I = 0) = \widetilde{F}(U \mid I = 1)$.
- conditions are unverifiable: X is observed only in sample and U is not observed at all.
- Δ_{S_X} vanishes if weighting on X
- Δ_{S_U} cannot be corrected after the fact

Decomposing Treatment Imbalance

- Decomposition:

$$\Delta_T = \Delta_{T_X} + \Delta_{T_U}$$

- $\Delta_{T_X} = 0$ when X is balanced between treateds and controls:

$$\widetilde{F}(X \mid T = 1, I = 1) \quad = \quad \widetilde{F}(X \mid T = 0, I = 1).$$

  Verifiable from data; can be generated ex ante by blocking or enforced ex post via matching or parametric adjustment

- $\Delta_{T_U} = 0$ when U is balanced between treateds and controls:

$$\widetilde{F}(U \mid T = 1, I = 1) \quad = \quad \widetilde{F}(U \mid T = 0, I = 1).$$

  Unverifiable. Achieved only by assumption or, on average, by random treatment assignment

Effects of Design Components on Estimation Error

Design Choice	Δ_{S_X}	Δ_{S_U}	Δ_{T_X}	Δ_{T_U}
Random sampling	$\overset{\text{avg}}{=} 0$	$\overset{\text{avg}}{=} 0$		
Complete stratified random sampling	$= 0$	$\overset{\text{avg}}{=} 0$		
Focus on SATE rather than PATE	$= 0$	$= 0$		
Weighting for nonrandom sampling	$= 0$	$=?$		
Large sample size	$\to ?$	$\to ?$	$\to ?$	$\to ?$
Random treatment assignment			$\overset{\text{avg}}{=} 0$	$\overset{\text{avg}}{=} 0$
Complete blocking			$= 0$	$=?$
Exact matching			$= 0$	$=?$
By Assumption				
No selection bias	$\overset{\text{avg}}{=} 0$	$\overset{\text{avg}}{=} 0$		
Ignorability				$\overset{\text{avg}}{=} 0$
No omitted variables				$= 0$

The Benefits of Major Research Designs: Overview

	Δ_{S_X}	Δ_{S_U}	Δ_{T_X}	Δ_{T_U}
Ideal experiment	$\to 0$	$\to 0$	$= 0$	$\to 0$
Randomized trials (Limited or no blocking)	$\neq 0$	$\neq 0$	$\overset{\text{avg}}{=} 0$	$\overset{\text{avg}}{=} 0$
Randomized trials (Full blocking)	$\neq 0$	$\neq 0$	$= 0$	$\overset{\text{avg}}{=} 0$
Survey Experiment (Limited or no blocking)	$\to ?$	$\to ?$	$\to 0$	$\to 0$
Observational Study (Representative data set, Well-matched)	≈ 0	≈ 0	≈ 0	$\neq 0$
Observational Study (Unrepresentative but partially, correctable data, well-matched)	≈ 0	$\neq 0$	≈ 0	$\neq 0$
Observational Study (Unrepresentative data set, Well-matched)	$\neq 0$	$\neq 0$	≈ 0	$\neq 0$

The Benefits of Major Research Designs: Overview

	Δ_{S_X}	Δ_{S_U}	Δ_{T_X}	Δ_{T_U}
Ideal experiment	$\to 0$	$\to 0$	$= 0$	$\to 0$
Randomized trials (Limited or no blocking)	$\neq 0$	$\neq 0$	$\overset{\text{avg}}{=} 0$	$\overset{\text{avg}}{=} 0$
Randomized trials (Full blocking)	$\neq 0$	$\neq 0$	$= 0$	$\overset{\text{avg}}{=} 0$
Survey Experiment (Limited or no blocking, no non-response)	$\to 0$	$\to 0$	$\to 0$	$\to 0$
Observational Study (Representative data set, Well-matched)	≈ 0	≈ 0	≈ 0	$\neq 0$
Observational Study (Unrepresentative but partially, correctable data, well-matched)	≈ 0	$\neq 0$	≈ 0	$\neq 0$
Observational Study (Unrepresentative data set, Well-matched)	$\neq 0$	$\neq 0$	≈ 0	$\neq 0$

This Week in Review

- The Experimental Ideal
- Identification with Measured Confounding
- Estimation by Stratification, Matching and Regression

Next week: Selection with Unmeasured Confounding!