# Week 10: Causality with Measured Confounding

Brandon Stewart[1]

Princeton

November 14–16, 2022

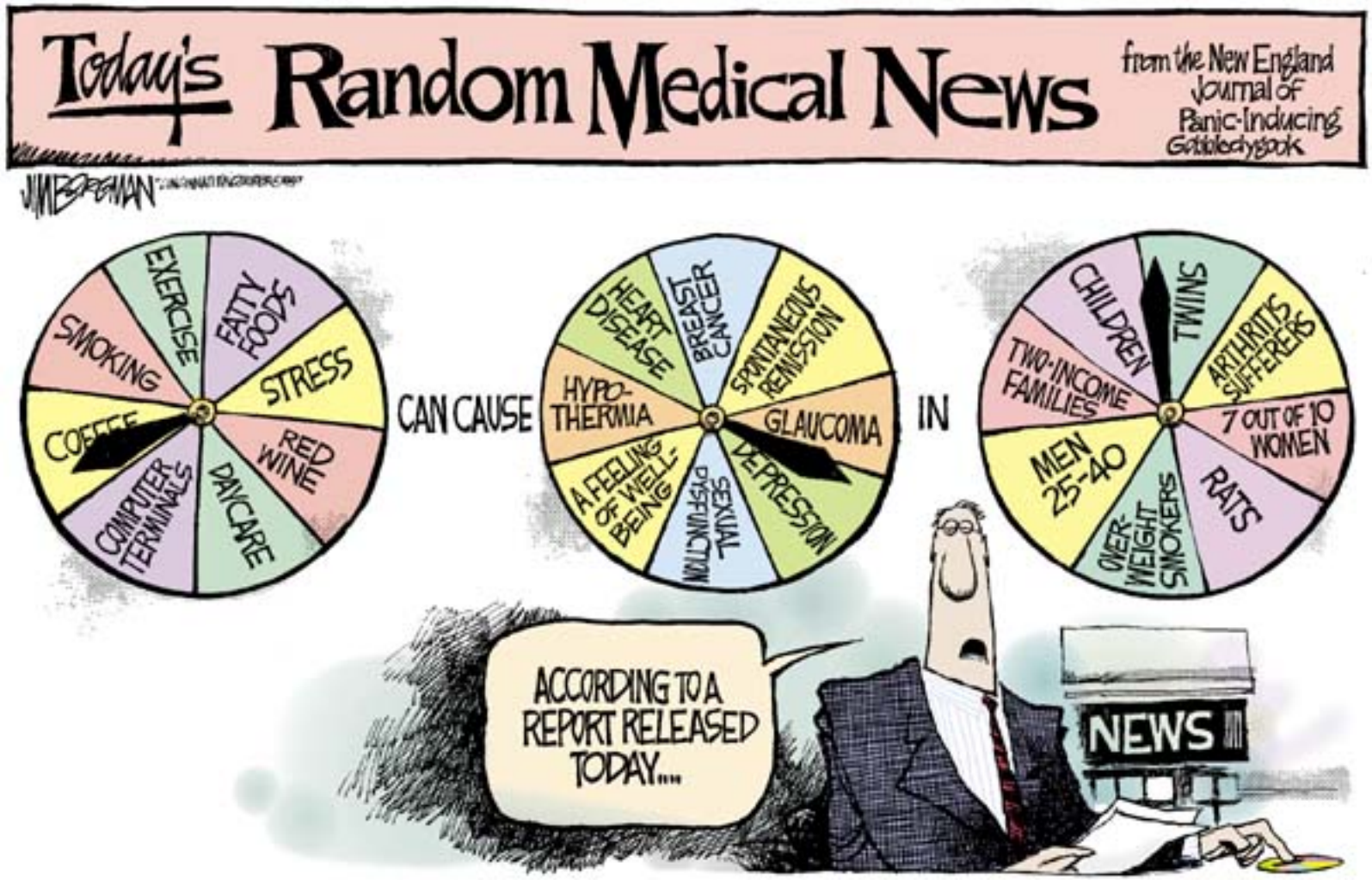
[1]These slides are heavily influenced by Matt Blackwell, Jens Hainmueller, Erin Hartman, Kosuke Imai, and Ian Lundberg.

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

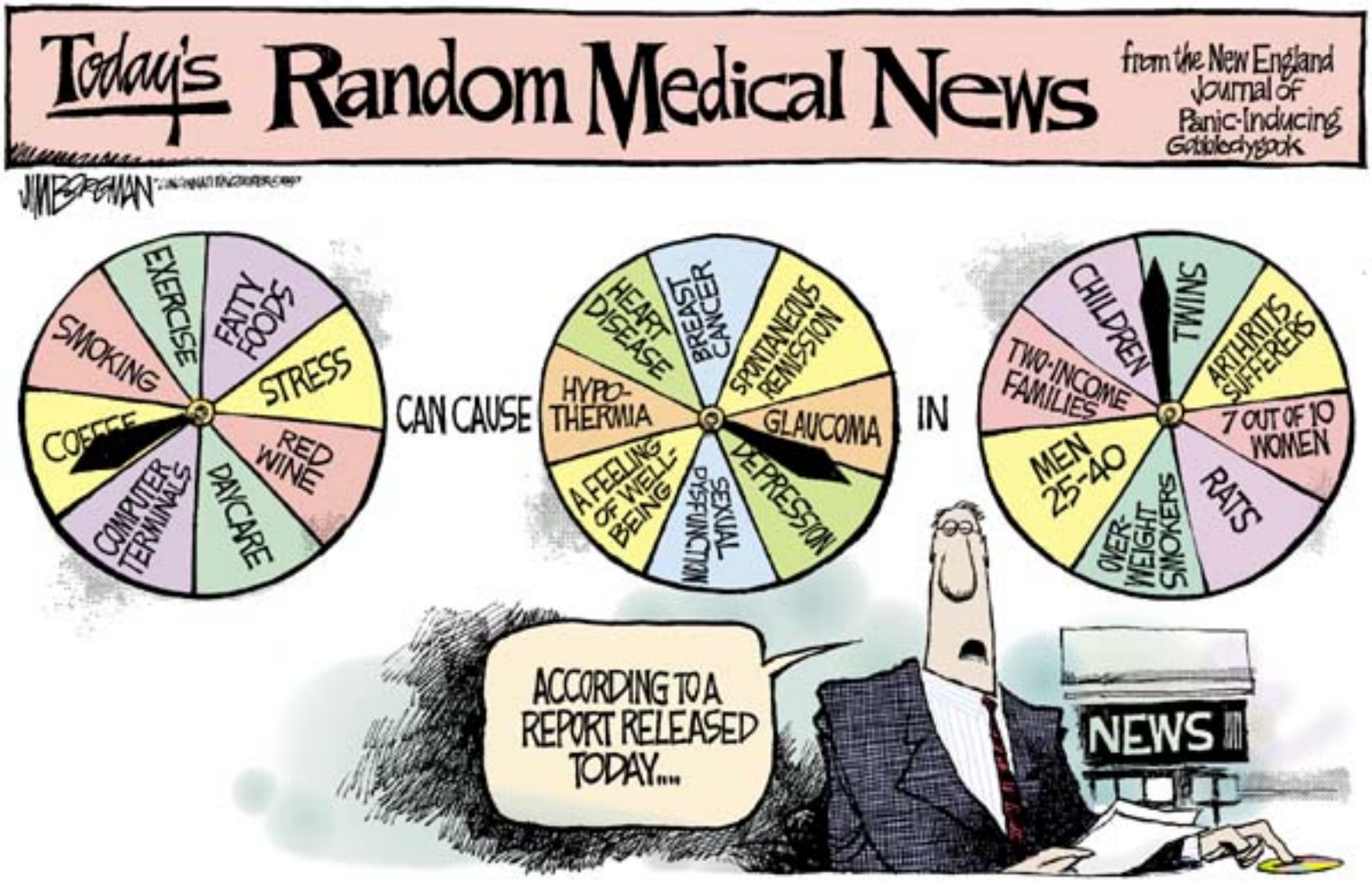
- Last Week
  - frameworks for causal inference
- This Week
  - experimental ideal
  - identification with measured confounding
  - estimation via stratification and regression
- Next Week
  - approaches with unmeasured confounding
- Long Run
  - causal frameworks $\rightarrow$ inference $\rightarrow$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

# Administrative Notes

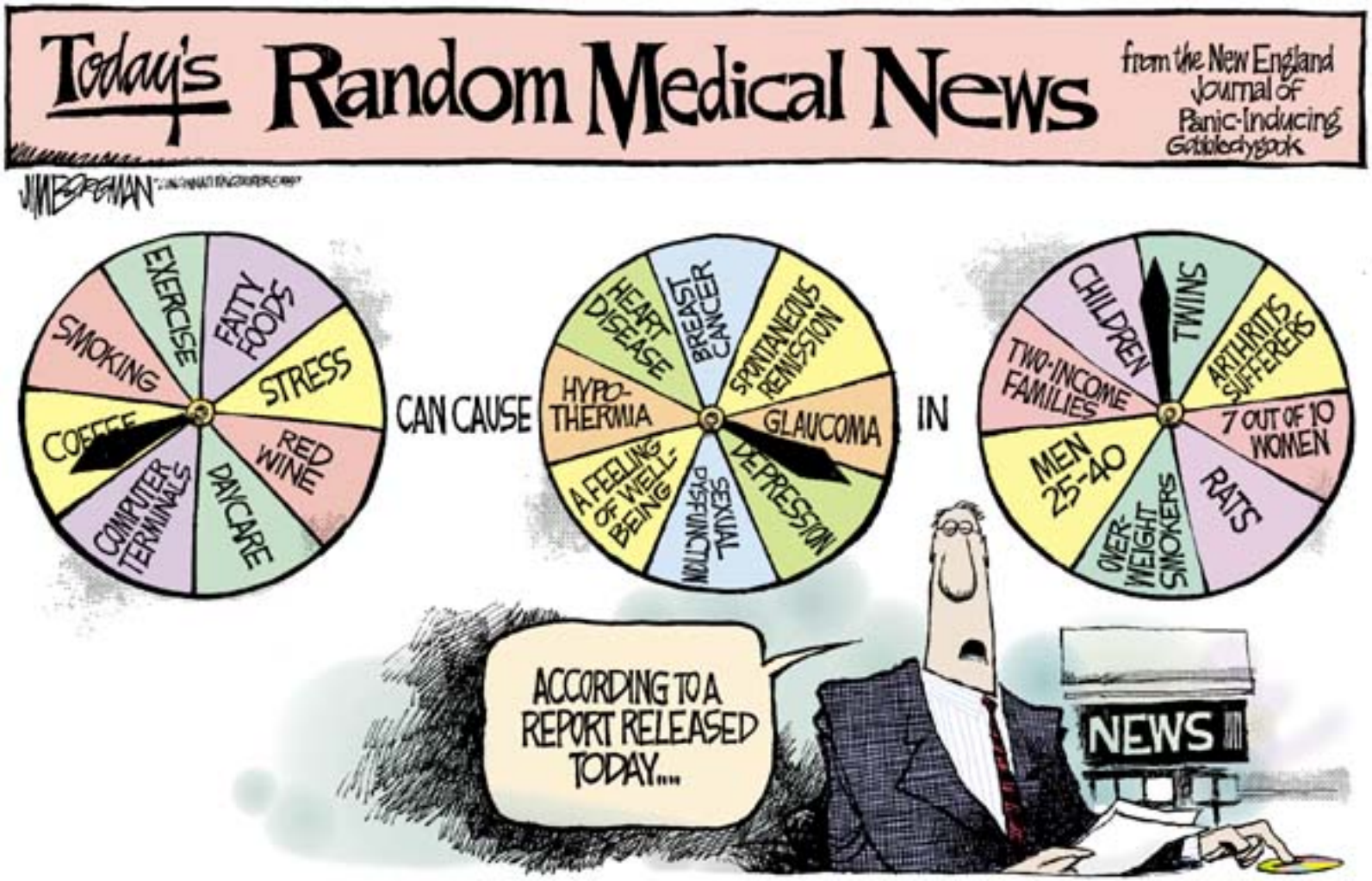
- Summary of the Rest of Class:
  - Nov 14–16: Measured Confounding (Week 10), Pset 9 due 11/16, Precept 11/17
  - Nov 21: Discussion and No Bad Question Q&A. No precept this week
  - Nov 28–30: Unmeasured Confounding (Week 11), Pset 10 due 11/30, Precept 12/1
  - Dec 5–7: Fixed Effects Plus Wrap-Up (Week 12), Pset 11 due 12/7, Precept 12/8
  - Dec 15: Review Session during Normal Precept Time
  - Dec 16–23: Final Exam Period
- Nov 21: Read Lundberg, Ian, Rebecca Johnson, and Brandon M. Stewart. "What is your estimand? Defining the target quantity connects statistical evidence to theory." *American Sociological Review* 86.3 (2021): 532-565.
- Some guest lectures planned after Thanksgiving




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)

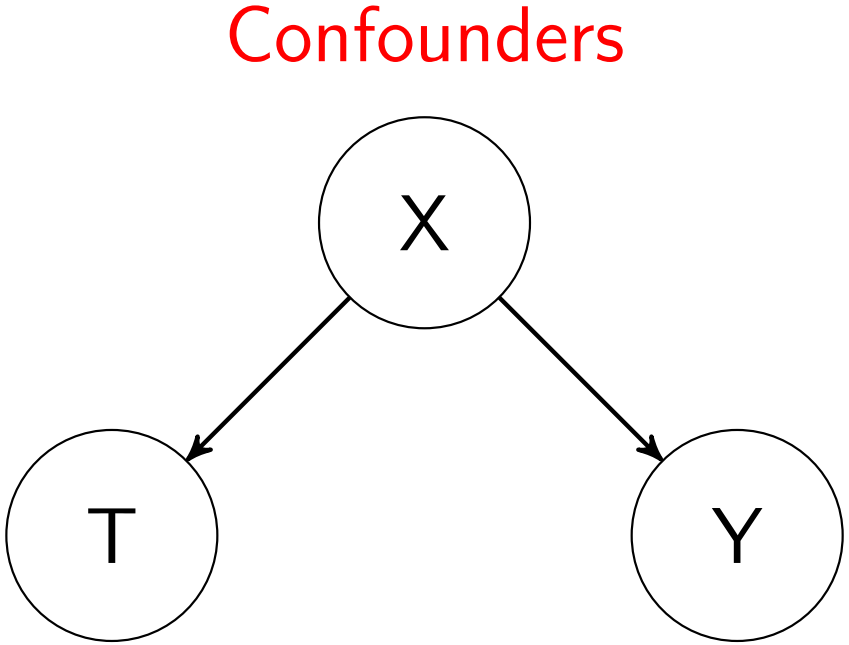



Lancet 2002: no effect of vitamin C on mortality in controlled placebo trial (controlling for nothing)




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?



# 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: $P(T_i = 1 | \mathbf{Y}(1), \mathbf{Y}(0)) = P(T_i = 1)$
     - ★ 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?

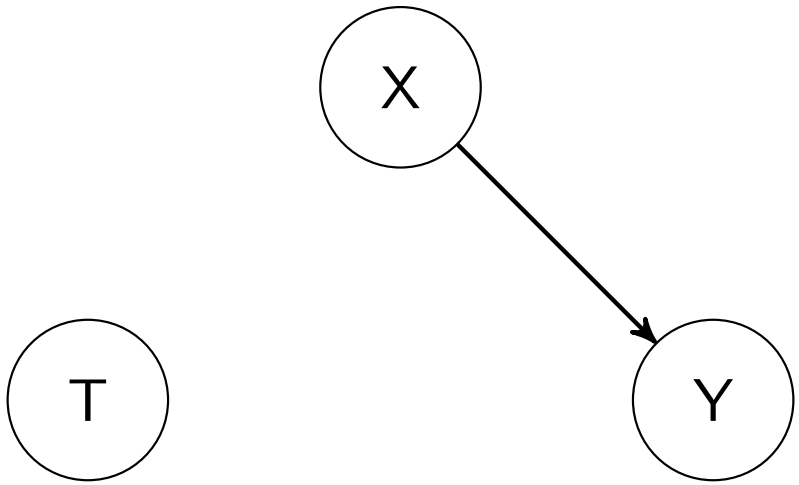
$$
\begin{aligned}
& E[Y_i|T_i=1] - E[Y_i|T_i=0] \\
&\quad = E[Y_i(1)|T_i=1] - E[Y_i(0)|T_i=0] \\
&\quad = E[Y_i(1)|T_i=1] - E[Y_i(0)|T_i=1] + E[Y_i(0)|T_i=1] - E[Y_i(0)|T_i=0] \\
&\quad = \underbrace{E[Y_i(1) - Y_i(0)|T_i=1]}_{\text{Average Treatment Effect on Treated}} + \underbrace{E[Y_i(0)|T_i=1]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

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

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

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!

$$
\begin{aligned}
E[Y_i(1) - Y_i(0)] &= E_X\Big\{E[Y_i(1) - Y_i(0)|X_i]\Big\} && \text{(iterated expectations)} \\
&= E_X\Big\{E[Y_i(1)|X_i] - E[Y_i(0)|X_i]\Big\} \\
&= E_X\Big\{E[Y_i(1)|T_i = 1, X_i] - E[Y_i(0)|T_i = 0, X_i]\Big\} && \text{(ignorability)} \\
&= E_X\Big\{E[Y_i|T_i = 1, X_i] - E[Y_i|T_i = 0, X_i]\Big\} && \text{(SUTVA)}
\end{aligned}
$$

- 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).

# Avoiding Common Areas of Confusion

- contribution not attribution: we care about a difference which doesn't make it the main reason, nor does it imply a morality claim, it doesn't make $T$ the reason it happened, it doesn't mean that $T$ is "responsible" for Y
- T can 'cause' Y if it is neither necessary nor sufficient
- If you know that on average A causes B and B causes C this doesn't mean you know that A causes C (example A$\rightarrow$B for one subgroup, B$\rightarrow$C for second subgroup, still no A$\rightarrow$C)
- estimation of causal effects does not require identical treatment and control groups
- you need a clear counterfactual to have a well-defined causal effect. For example of 'the recession was caused by Wall Street' may make intuitive sense but is it well-defined?

http://egap.org/methods-guides/10-things-you-need-know-about-causal-inference

# 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 and regression
- Next Week
  - approaches with unmeasured confounding
- Long Run
  - causal frameworks $\rightarrow$ inference $\rightarrow$ regression $\rightarrow$ causal inference

# 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 idea: imagine that your data was generated by a stratified randomized 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:*

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

# 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 $\tau_{ATT}$ 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 $X$:

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

# 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

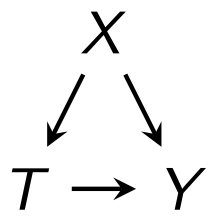
$$E[Y(1)|X=0,T=1] = E[Y(1)|X=0,T=0] \text{ and}$$
$$E[Y(1)|X=1,T=1] = E[Y(1)|X=1,T=0]$$

# 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.

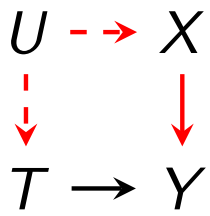
# Backdoor paths and blocking paths

- 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 $Y$:



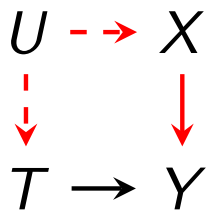

- Here there is a backdoor path $T \leftarrow X \rightarrow Y$, where $X$ is a common cause for the treatment and the outcome.

# Other types of confounding



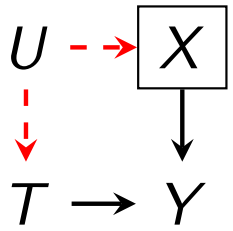

- $T$ is enrolling in a job training program.
- $Y$ is getting a job.
- $U$ is being motivated
- $X$ is number of job applications sent out.
- Big assumption here: no arrow from $U$ to $Y$.

# Other types of confounding



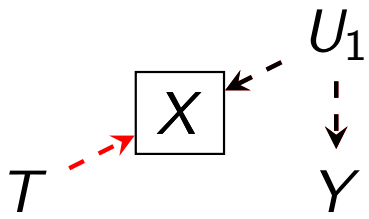

- $T$ is exercise.
- $Y$ is having a disease.
- $U$ is lifestyle.
- $X$ is smoking
- Big assumption here: no arrow from $U$ to $Y$.

# 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



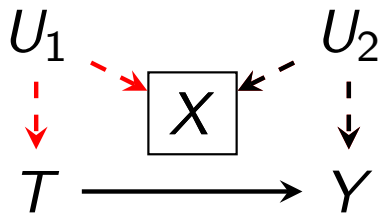

- Conditioning on the posttreatment covariates opens the non-causal path.
  - $\rightsquigarrow$ selection bias.

# 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



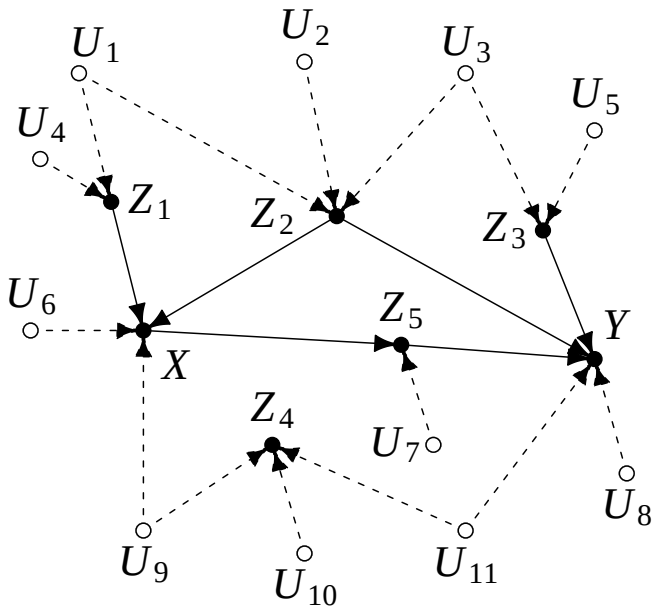

- Not all backdoor paths induce confounding.
- This backdoor path is blocked by the collider $X$ that we don't control for.
- If we control for $X$ $\rightsquigarrow$ opens the path and induces confounding.
  - Sometimes called M-bias.
- Controversial because of differing views on what to control for: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

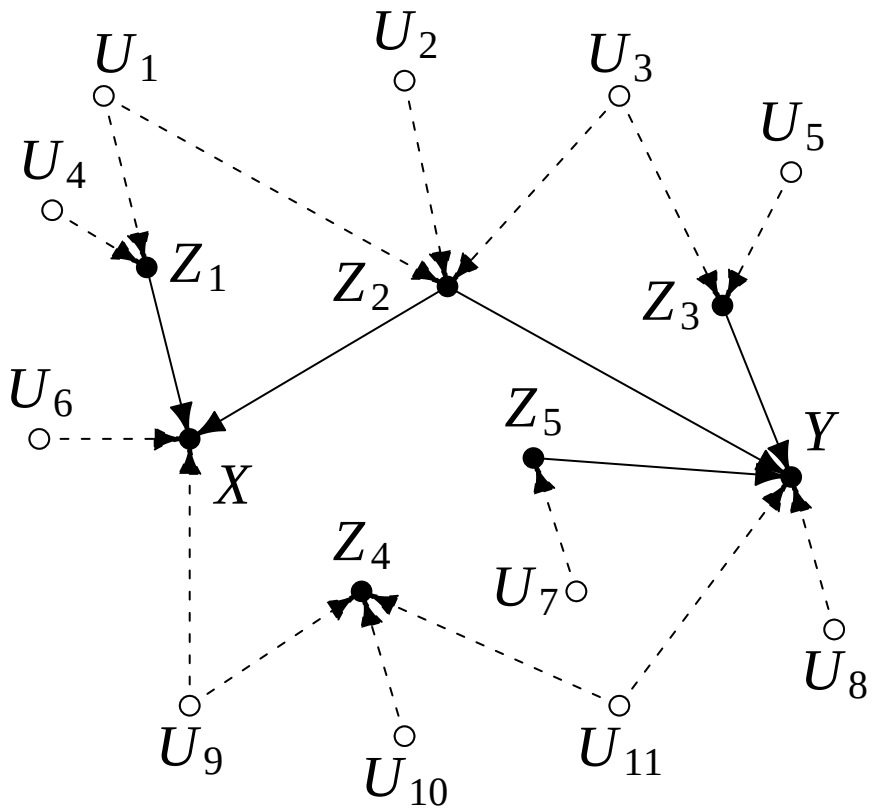
# Example: Sufficient Conditioning Sets

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



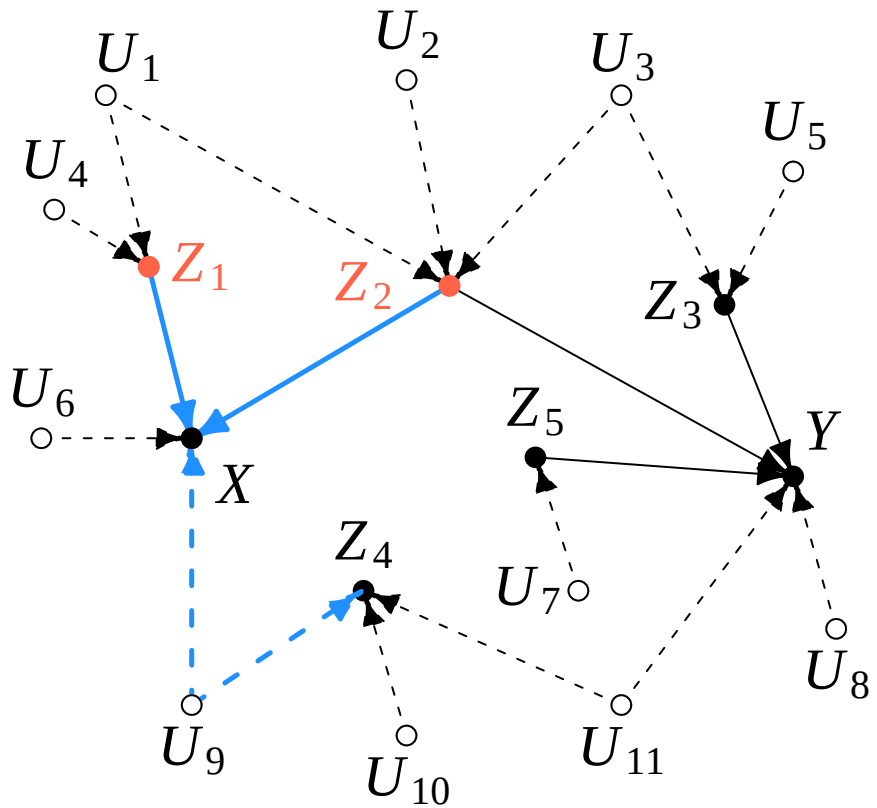

Remove arrows out of $X$.

# Example: Sufficient Conditioning Sets




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

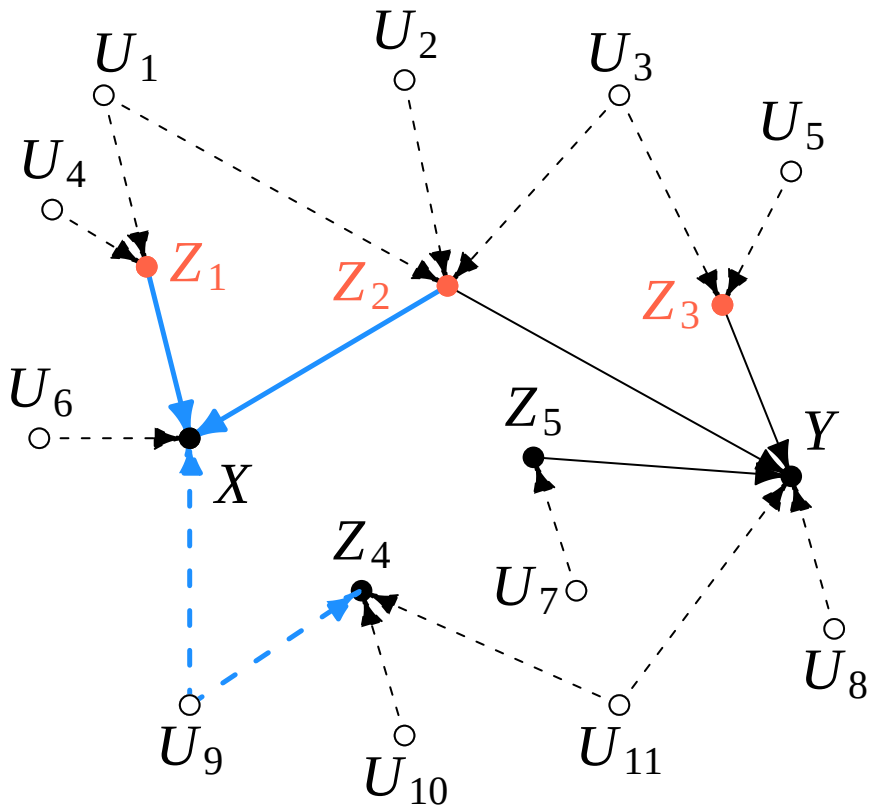
# Example: Sufficient Conditioning Sets




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

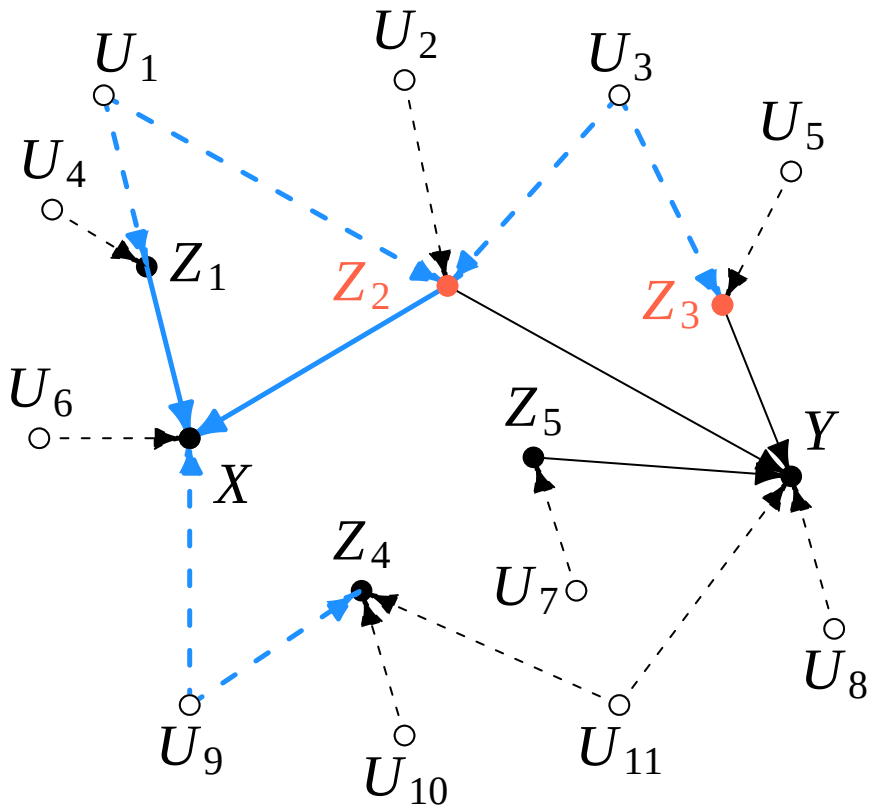
# Example: Sufficient Conditioning Sets




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

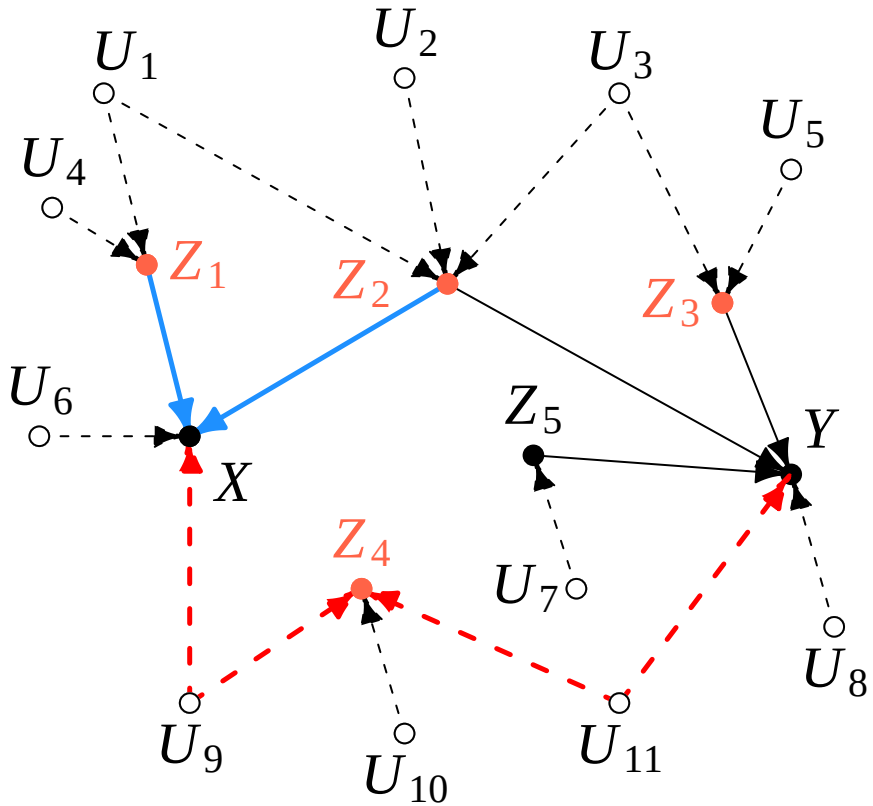
# Example: Sufficient Conditioning Sets




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

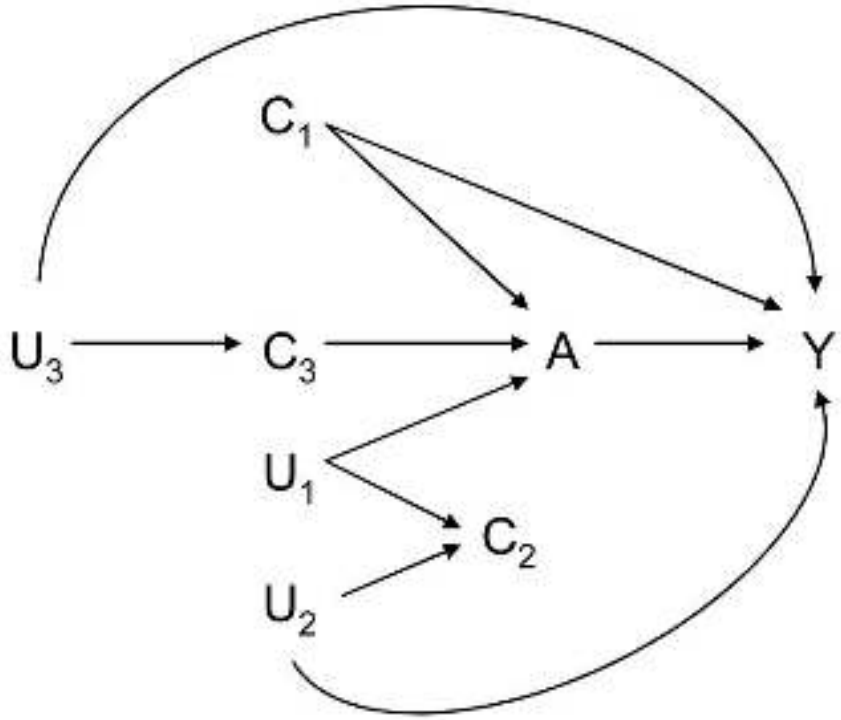



There are unblocked paths if we condition on $Z_1$, $Z_2$, $Z_3$, $Z_4$

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 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 $T_i$ 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 $\mathbf{X}$ 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 $\mathbf{X}$.

1. The Experimental Ideal
2. Identification with Measured Confounding
   - Design
   - DAGs
3. Stratification
4. Regression
   - Regression with Constant Effects
   - Additive Regression with Heterogeneous Effects
   - Imputation Estimators

# 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:

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

- Stratification is great because it makes no assumptions about parametric form.
- More generally stratification just involves a weighted average of subgroup means.

# 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 |

# Limits to Stratification

- So, great, we can stratify. Why not do this all the time?
- There are three core problems:
  1. Empty cells (combinations of covariates $X_i$ which have either no treatment or control)
  2. Continuous covariates (empty cells are inevitable!)
  3. Estimation variance (even if we didn't have empty cells, we might have very few observations to estimate the mean making it noisy).
- One option is to discretize into bins or regularize our estimates in various ways, but for many settings we may want another approach.

1. The Experimental Ideal

2. Identification with Measured Confounding
   - Design
   - DAGs

3. Stratification

4. Regression
   - Regression with Constant Effects
   - Additive Regression with Heterogeneous Effects
   - Imputation Estimators

# From Stratification to Regression

- Nonparametric identification yielded this estimand:

$$E[Y_i(1) - Y_i(0)] = E\Big[E[Y_i \mid T_i = 1, \mathbf{X}_i] - E[Y_i \mid T_i = 0, \mathbf{X}_i]\Big]$$

- What if we estimated the inner conditional expectations with a regression model like we have been doing all semester?
- We will sometimes write the empirical quantity of interest as

$$\mu_t(\mathbf{x}) = E[Y_i(t) \mid T = t, \mathbf{X}_i = \mathbf{x}]$$

- Ignorability tell us that the counterfactual quantities can be written in terms of these quantities based on observed data, but we still need to estimate those conditional expectations!

# The parametric g-formula: Simple case


| | |
|---|---|
| Nonparametric identification | $E[Y(t)] = E[E[Y \mid X, T = t]]$ |
| Parametric model for $\mu(\mathbf{x})$ | $\alpha + \gamma X + \beta T$ |

Regression estimator for the ATE:

$$\hat{\tau} = \frac{1}{n} \sum_{i=1}^{n} \hat{\mu}_1(\mathbf{X}_i) - \hat{\mu}_0(\mathbf{X}_i)$$

Estimator in words:

1. Estimate the regression model(s)
2. Change all treatment values to 1 and 0
3. Predict for everyone
4. Take the sample mean of the difference.

# Connection to $\hat{\beta}$ for Additive Regression

Estimator for the effect $\tau = E[Y(1)] - E[Y(0)]$:

$$\begin{aligned}
\hat{\tau} &= \left( \frac{1}{n} \sum_{i=1}^{n} \left( \hat{\alpha} + \hat{\gamma} \mathbf{X}_i + \hat{\beta} \times 1 \right) \right) \\
&\qquad - \left( \frac{1}{n} \sum_{i=1}^{n} \left( \hat{\alpha} + \hat{\gamma} \mathbf{X}_i + \hat{\beta} \times 0 \right) \right) \\
&= \frac{1}{n} \sum_{i=1}^{n} \hat{\beta} \\
&= \hat{\beta}
\end{aligned}$$

With additive regression, the parametric g-formula estimator simplifies to $\hat{\beta}$.

# The parametric g-formula is more general

Suppose we add the $T \times \mathbf{X}$ interaction to the model.

$$E(Y \mid T, \mathbf{X}) = \alpha + \gamma \mathbf{X} + \beta T + \eta T \mathbf{X}$$

Estimator for the effect $E(Y^1) - E(Y^0)$:

$$\begin{aligned}\hat{\tau} = &\left( \frac{1}{n} \sum_{i=1}^{n} \left( \hat{\alpha} + \hat{\gamma} \mathbf{X}_i + \hat{\beta} \times 1 + \hat{\eta} \times 1 \times \mathbf{X}_i) \right) \right) \\ &- \left( \frac{1}{n} \sum_{i=1}^{n} \left( \hat{\alpha} + \hat{\gamma} \mathbf{X}_i + \hat{\beta} \times 0 + \hat{\eta} \times 0 \times \mathbf{X}_i \right) \right) \\ = &\frac{1}{n} \sum_{i=1}^{n} \left( \hat{\beta} + \hat{\eta} \mathbf{X}_i \right)\end{aligned}$$

This no longer collapses to a coefficient!

# The parametric g-formula: Word of warning

The parametric g-formula does not replace any assumptions.

You need:

Identification Assumptions (causal)

| | |
|---|---|
| Consistency | $Y_i = Y_i(T_i)$ |
| Ignorability | $T \perp\!\!\!\perp \{Y_i(0), Y_i(1)\} \mid \mathbf{X}$ |
| Positivity | $P(T = t \mid \mathbf{X} = \mathbf{x}) > 0$ |

Estimation Assumptions (statistical)

| | |
|---|---|
| Estimation | $E(Y \mid T, \mathbf{X})$ follows the assumed model |

To use a parametric model, we added the bottom assumption.

- Estimation assumptions have implications in the data and thus can be somewhat checked
- Pragmatically you need common support (i.e. treated data where control units are and vice-versa).

Let's consider the implications of estimation with regression under different assumptions about the model.

# Identification under Selection on Observables: Regression

Consider the estimation model: $Y_i = \beta_0 + \tau T_i + X_i'\beta + u_i$.

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

1. Constant treatment effects and outcomes are linear in $X$
   - $\tau$ will provide unbiased and consistent estimates of ATE.
2. Constant treatment effects and unknown functional form
   - $\tau$ 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 $\tau$ may be biased for the ATE.
3. Heterogeneous treatment effects ($\tau$ differs for different values of $X$)
   - even If outcomes are linear in $X$, $\tau$ 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, $\tau_R$? 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 $\tau_R$? Well, we can rely on the regression anatomy:

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

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

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

- $\sigma_t^2(x) = \text{Var}[T_i|X_i = x]$ 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:

| Group 1 | Group 2 |
|---|---|
| $P[X_i = 1] = 0.75$ | $P[X_i = 0] = 0.25$ |
| $P[T_i = 1 \mid X_i = 1] = 0.9$ | $P[T_i = 1 \mid X_i = 0] = 0.5$ |
| $\sigma_t^2(1) = 0.09$ | $\sigma_t^2(0) = 0.25$ |
| $\tau(1) = 1$ | $\tau(0) = -1$ |

- 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 $\tau_R$ 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.

# Implications

- Knowing the weights 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!

# Example Replication from Aronow and Samii (2015)

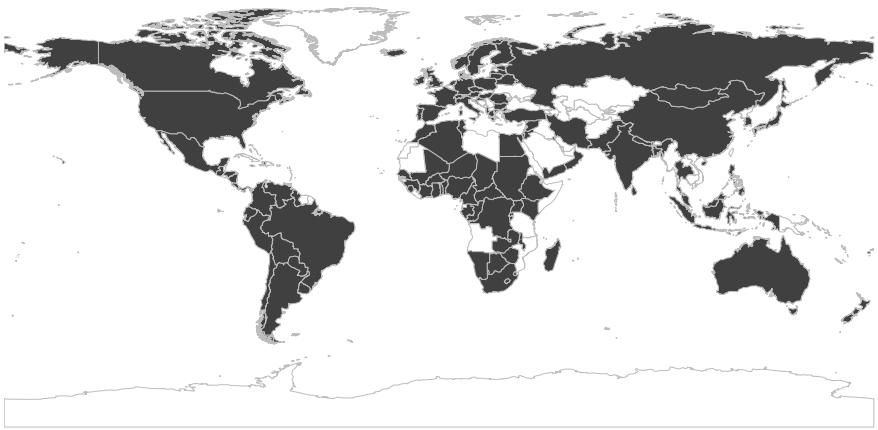
Jensen (2003), "Democratic Governance and Multinational Corporations: Political Regimes and Inflows of Foreign Direct Investment."

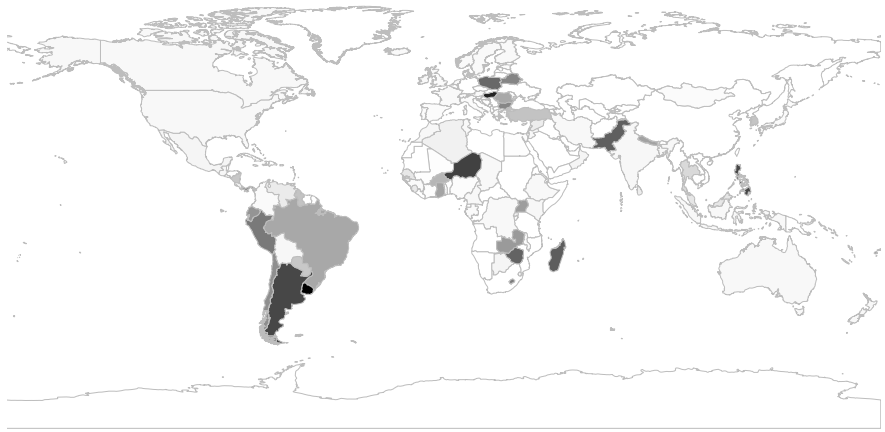
Jensen presents a large-$N$ TSCS-analysis of the causal effects of governance (as measured by the Polity III score) on Foreign Direct Investment (FDI).

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

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.

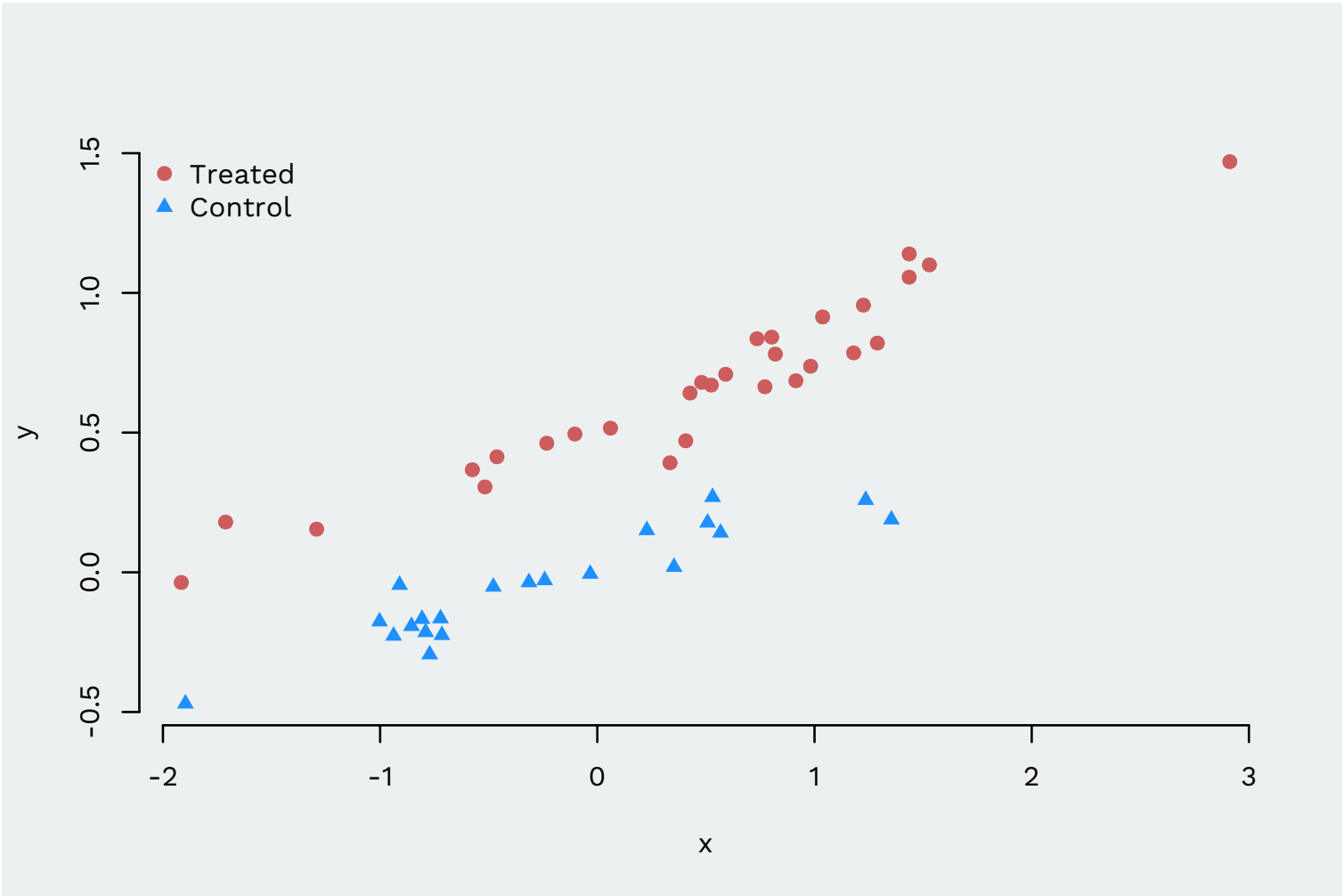
# Parametric g-Formula With Flexible Models

- 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) using a flexible model.
  - Regress $Y_i$ on $X_i$ in the control group and get predicted values for all units (treated or control) using a flexible model.
  - Take the average difference between these predicted values (uncertainty via bootstrap)
- More mathematically, look like this:

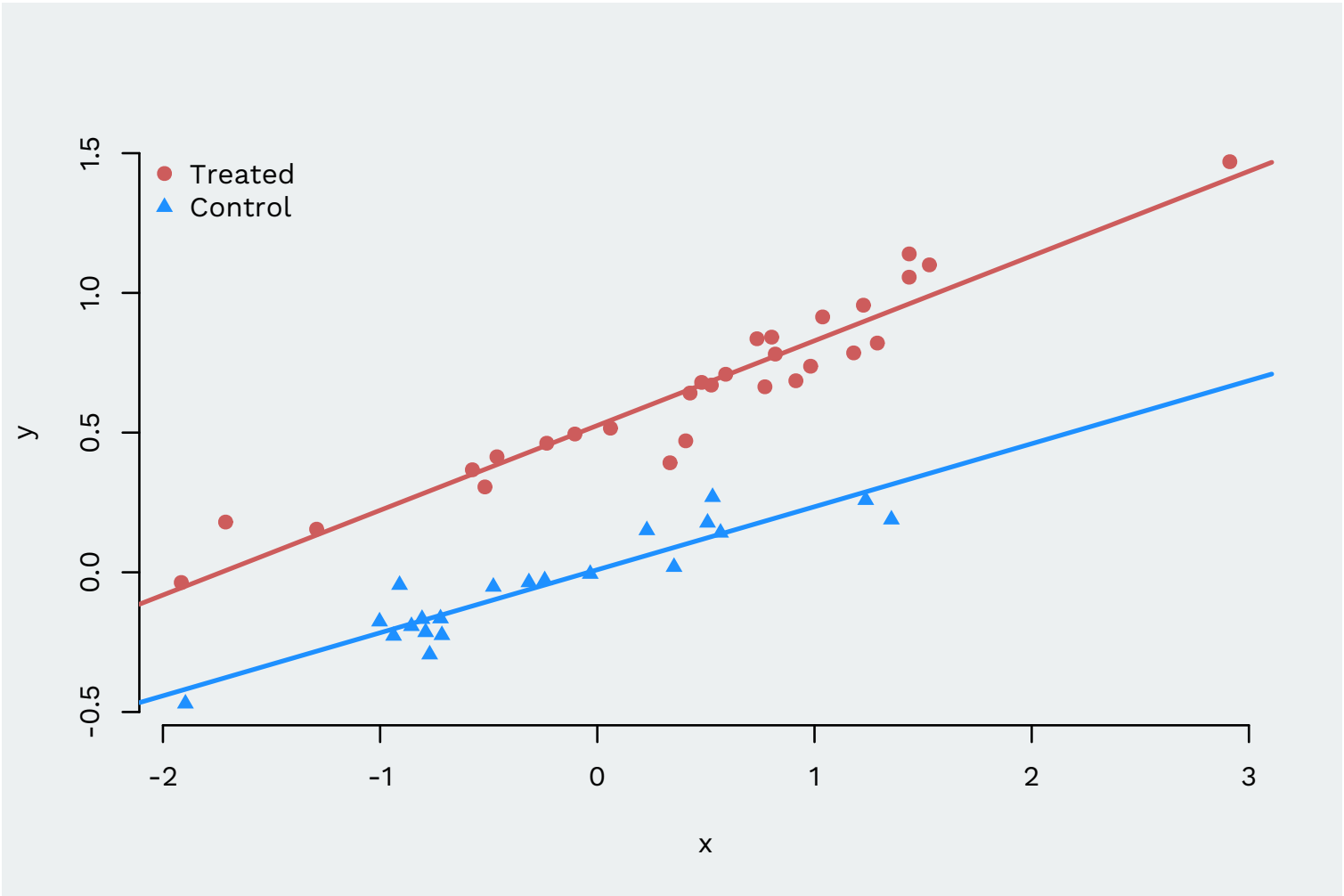
$$\tau_{imp} = \frac{1}{N} \sum_i \hat{\mu}_1(X_i) - \hat{\mu}_0(X_i)$$

- If the models are consistent for the conditional expectation (and identification assumptions hold) will be consistent for the treatment effect!

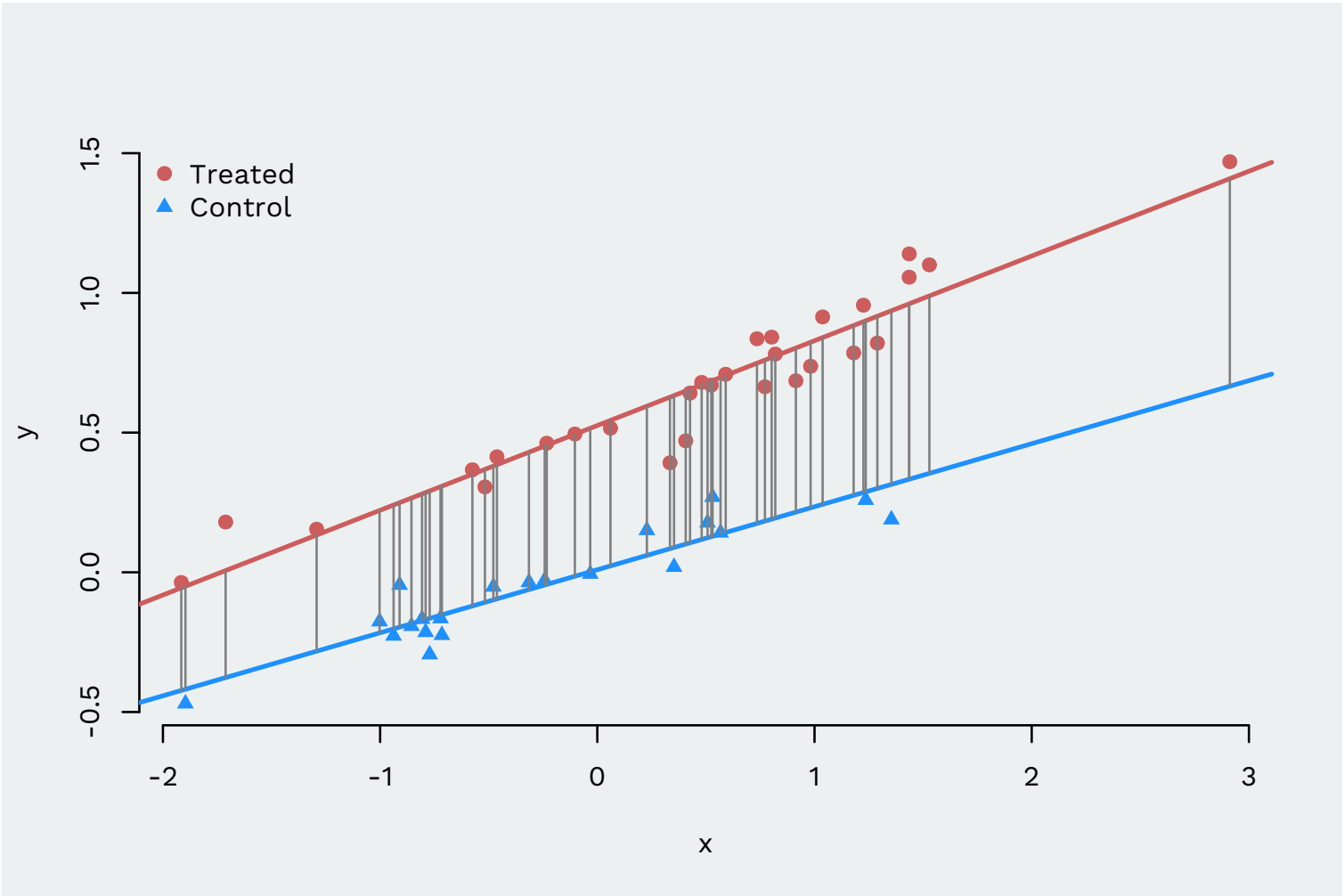
# Imputation estimator visualization



# Imputation estimator visualization

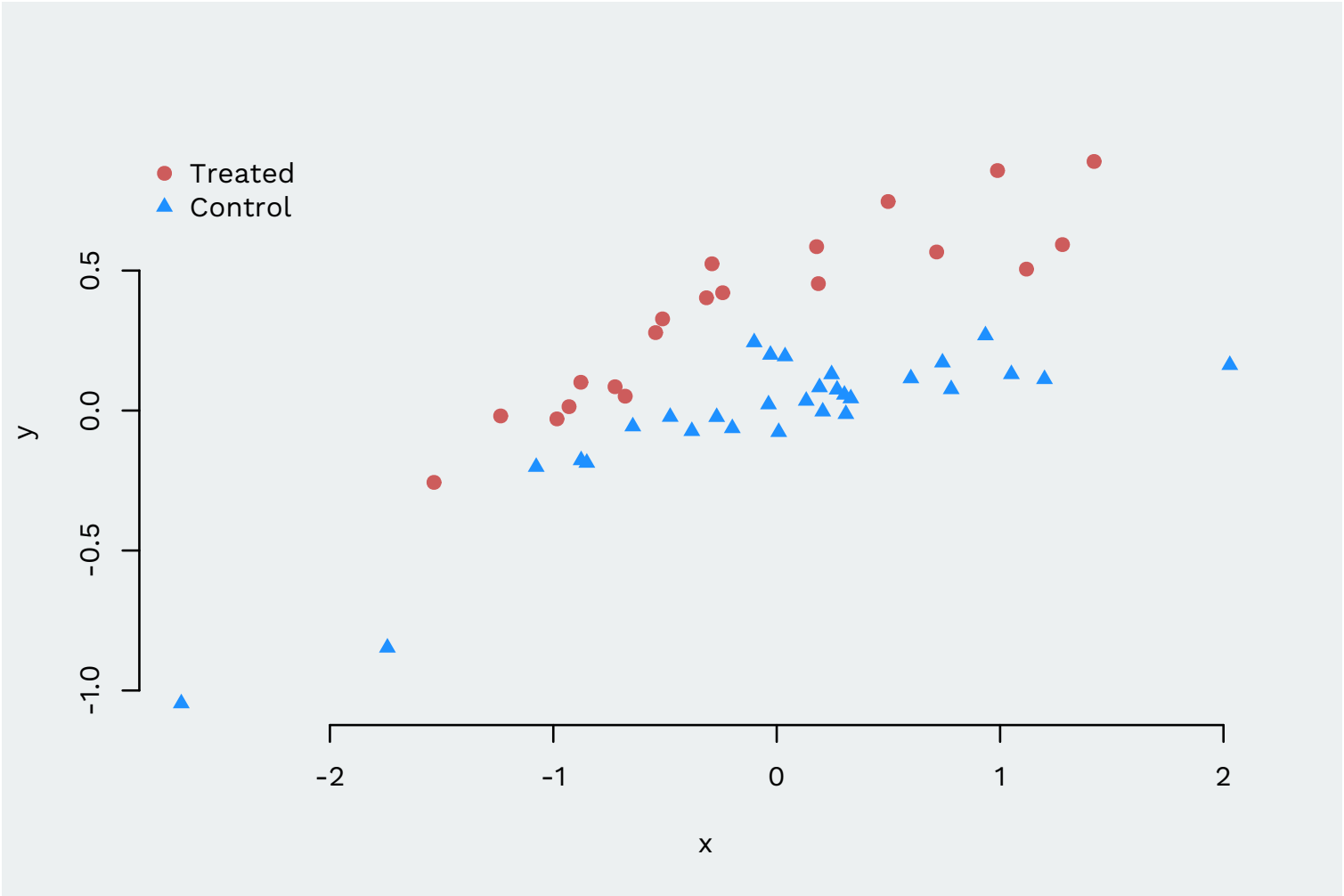


# Imputation estimator visualization



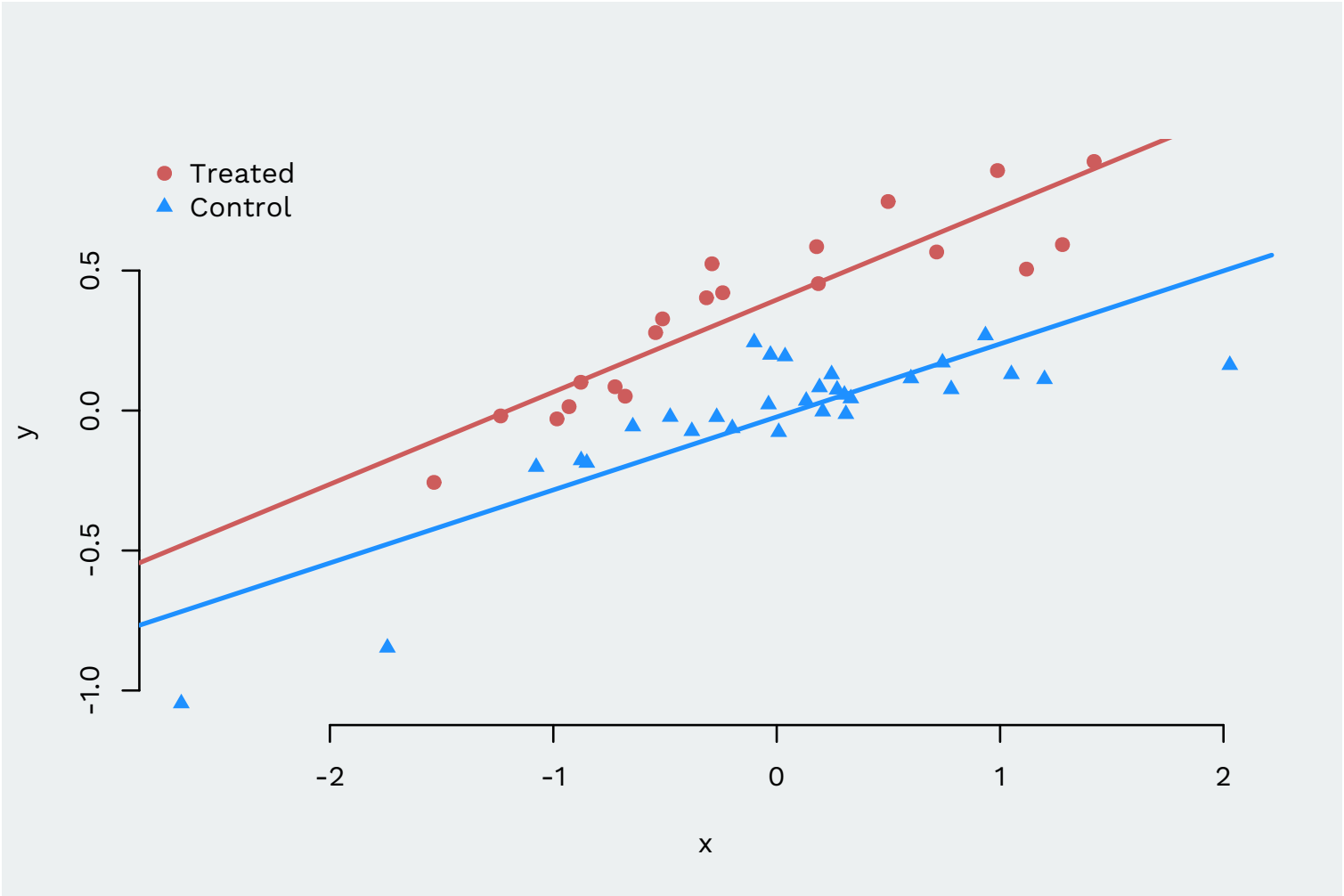
# Nonlinear relationships

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



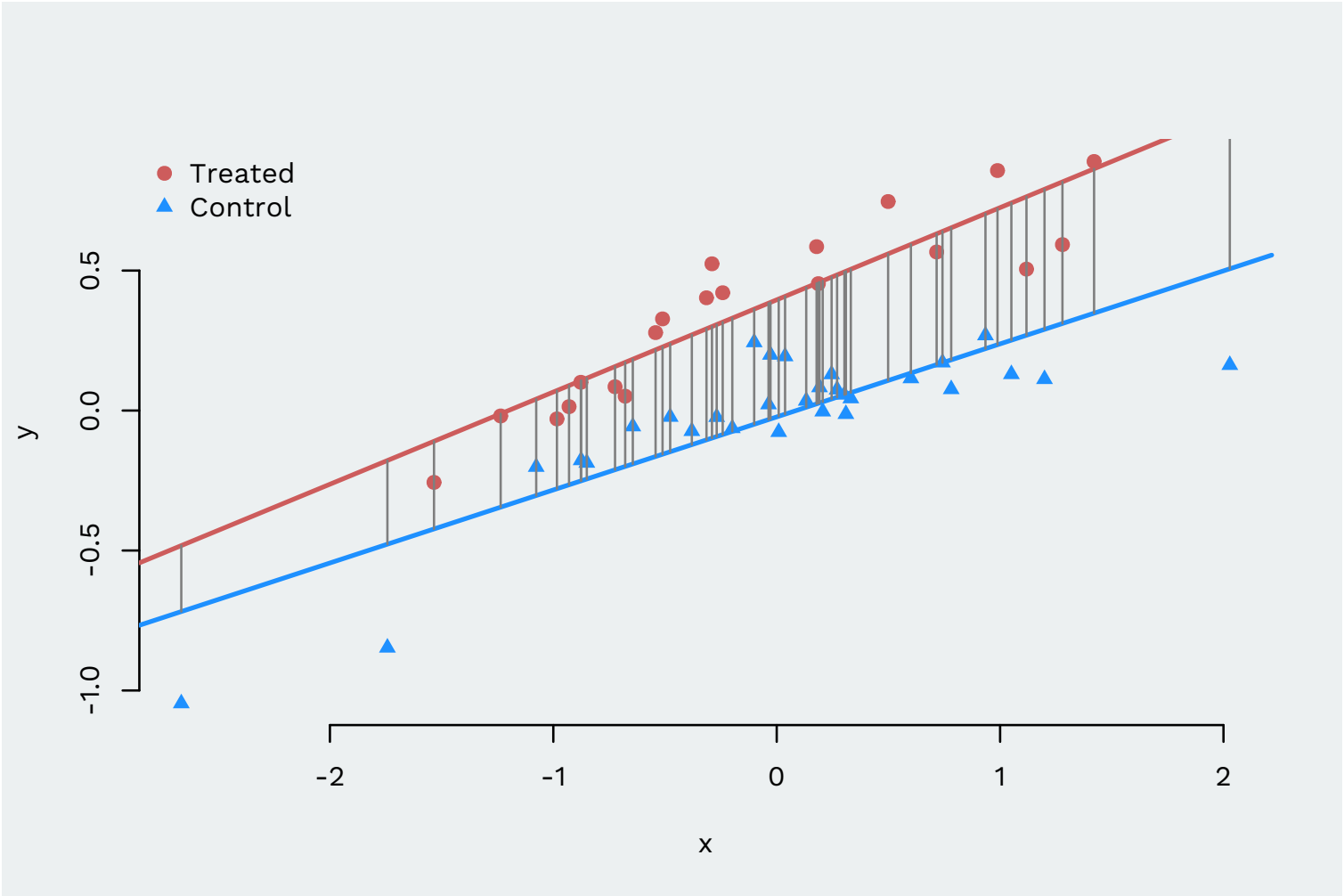
# Nonlinear relationships

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



# Nonlinear relationships

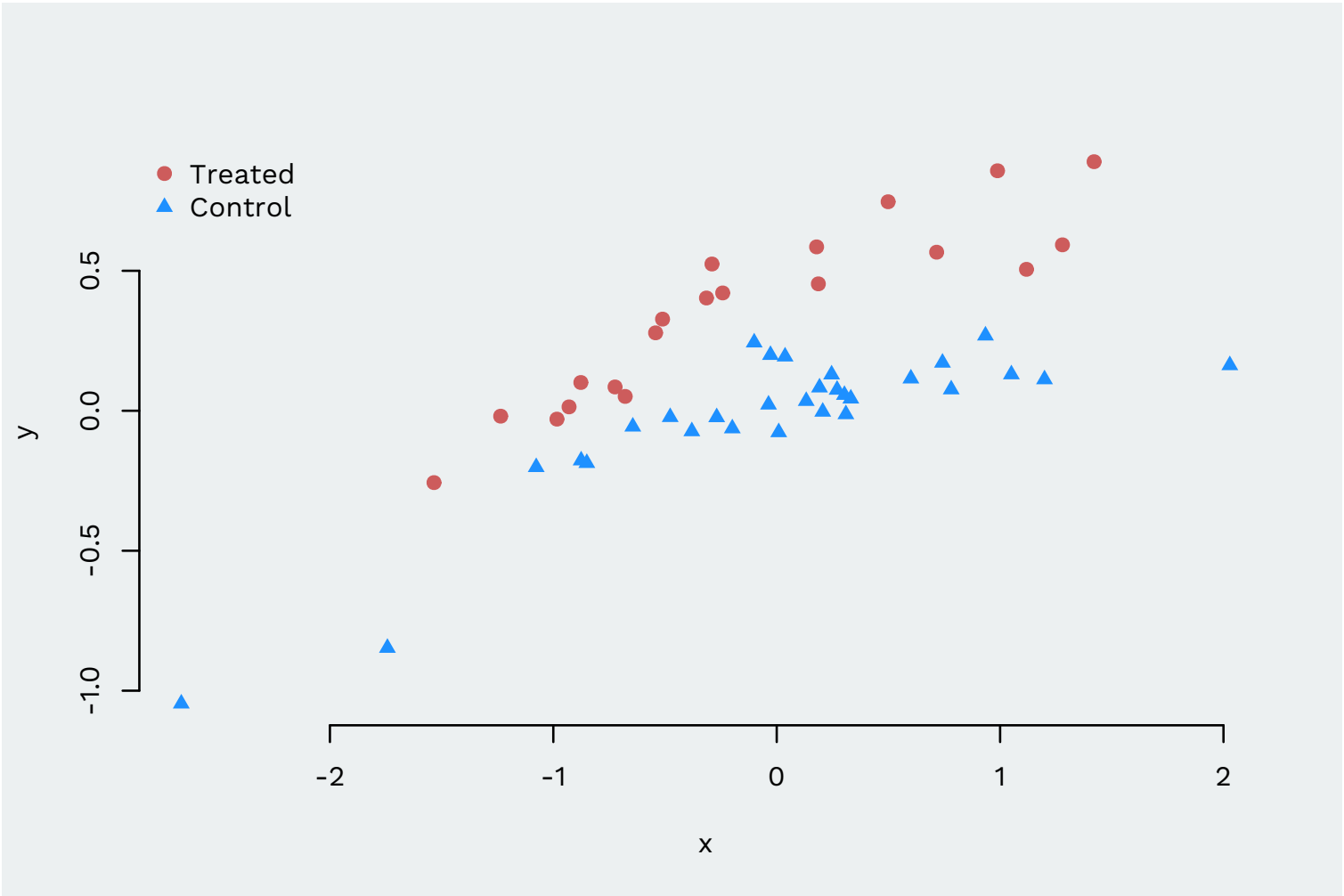
- Same idea but with nonlinear relationship between $Y_i$ and $X_i$:



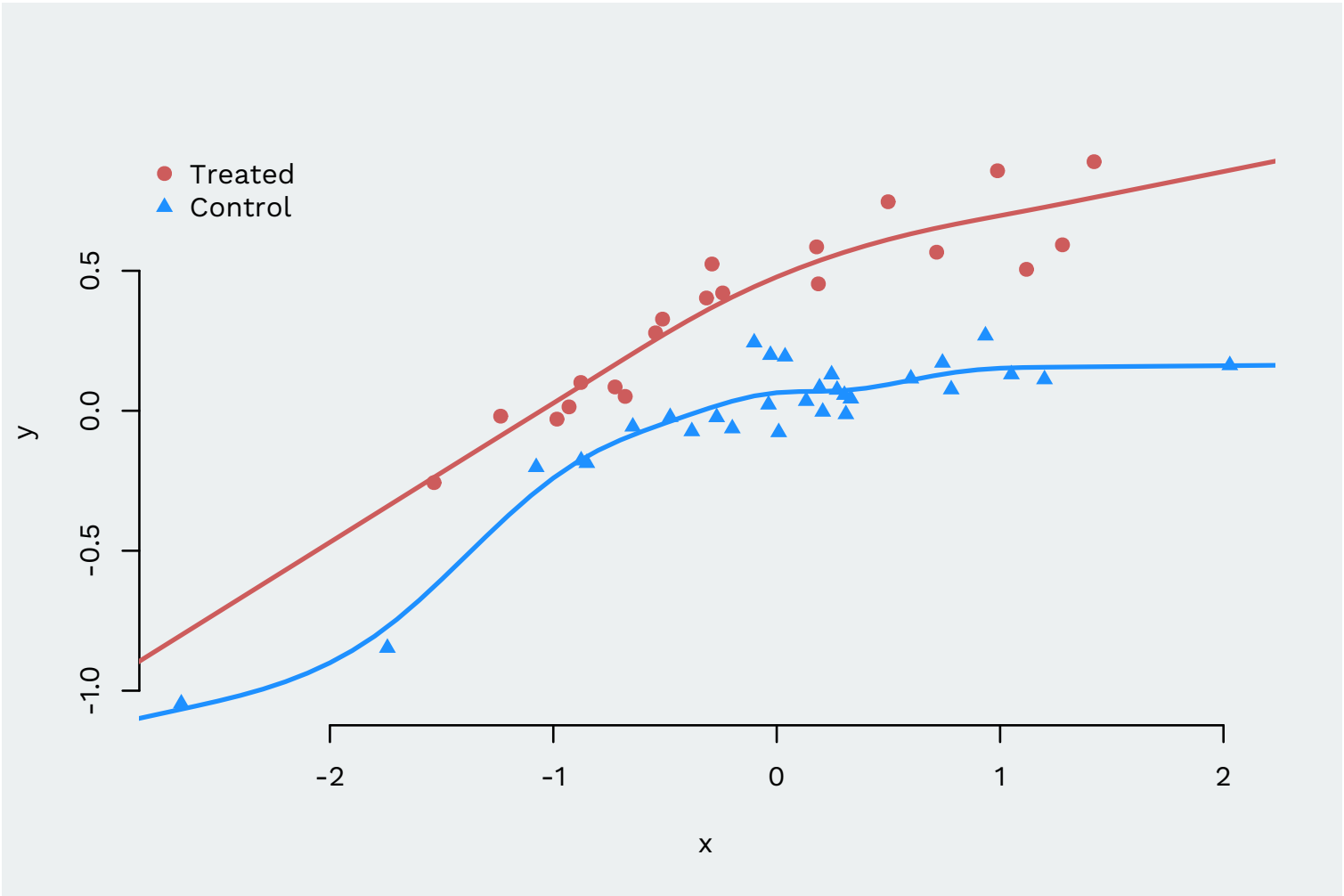
# Using semiparametric regression

- Here, CEFs are nonlinear, but we don't know their form.
- We can use Generalized Additive Models (GAMs) from the `mgcv` package for flexible estimate (a kind of machine learning).
- More generally we can use general machine learning estimators for this kind of adjustment (see e.g. Double Machine Learning from Chernozhukov et. al.)

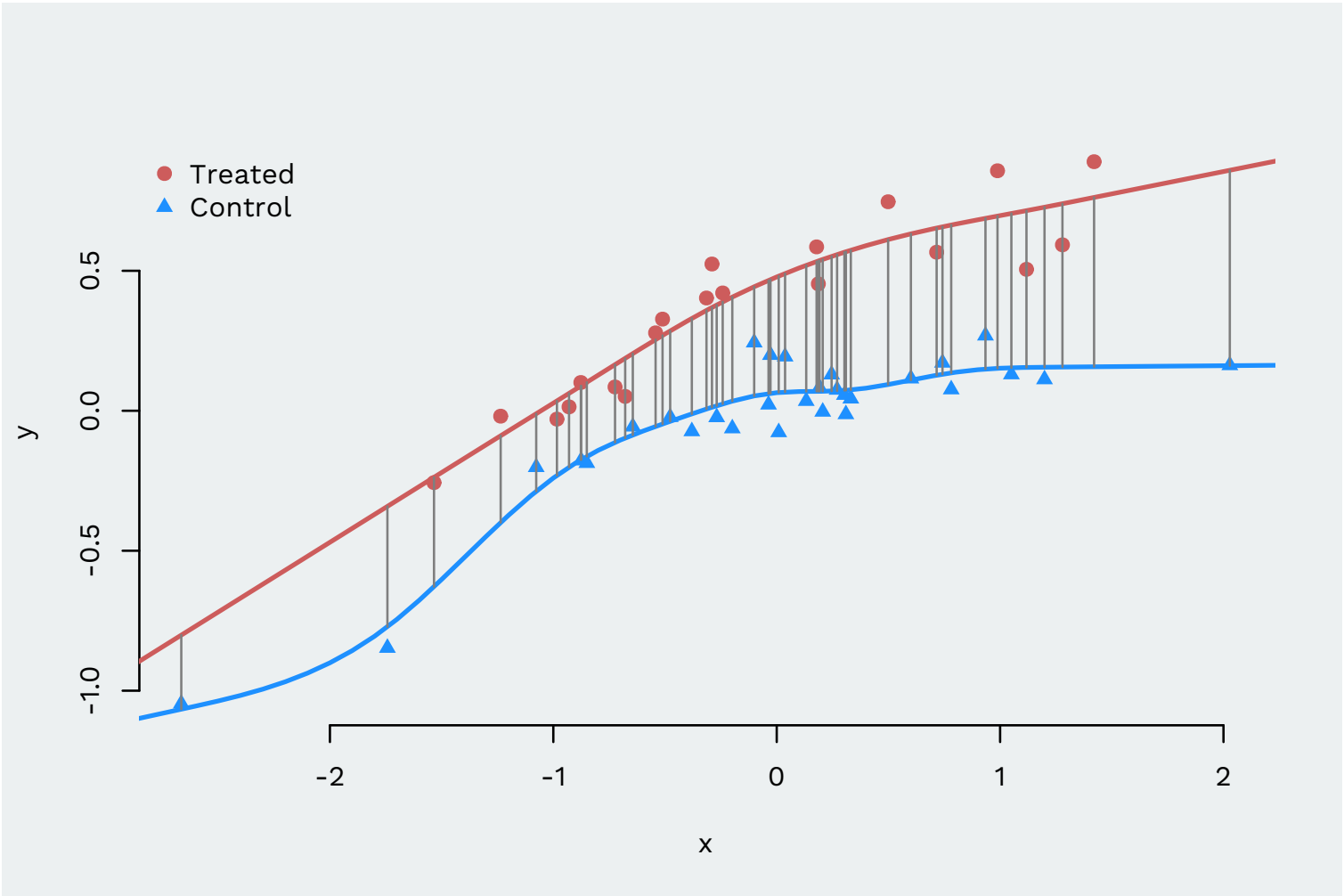
# Using GAMs



# Using GAMs



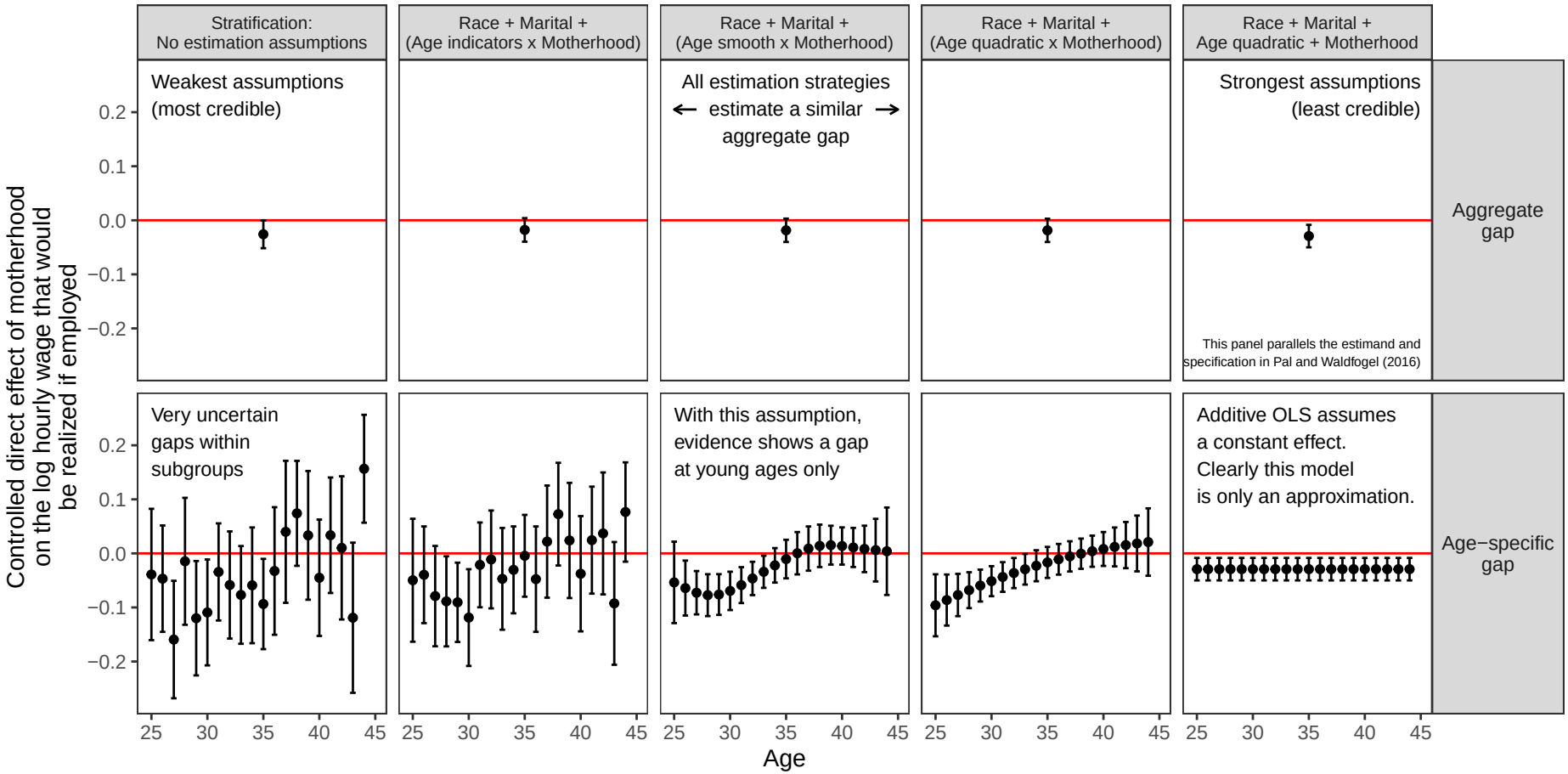
# Using GAMs



# Parametric $g$-Formula Estimators

- Imputation estimators (via 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 $\tau_{imp}$ 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



# Advantages of the Regression Framework

- Learning how regression behaves as an estimator is helpful for many reasons:
  - people use regression all the time and it is useful to know what the coefficient represents
  - the variance-weighting is a key ingredient in understanding how many additively separable models behave.
  - it lends itself nicely to various kinds of sensitivity analysis (e.g. Cinelli and Hazlett 2019) and diagnostics
- Basic linear regression is an efficient estimator for relatively small datasets that in practice is probably at least going to get the sign correct.
- Remember a fancier regression only makes the estimation assumptions more credible, not the identification assumption!

# Other Approaches for Measured Confounding

- There are many non-regression ways to adjust for confounding: matching and weighting being the two most common examples.
- Next semester you will learn about adjustment using the propensity score which is the probability of treatment given the covariates.
- What you choose to adjust for is much more important than how you do the conditioning.

# We Covered

- The experimental ideal
- Identification with measured confounding
- Stratification
- 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.

Next Time: A Discussion about Estimands (Thanksgiving Monday), then What to Do About Unmeasured Confounding (Week 11)