Feature selection for learning causal effects: a finite-sample, stratification estimator perspective

Andrew Herren (joint work with P. Richard Hahn)

8/12/2022

Central research question

Assume that we have data

- Y: observed outcome
- Z: binary treatment
- X: vector of discrete covariates

and we know that we can estimate the average treatment effect [ATE] by conditioning on all of X (we will make this more precise in a moment!) . . .

. . . can we "do better" than adjusting for all of X? In a principled manner? Without knowing or estimating a causal DAG?

Why do we work with discrete X

- We can investigate these phenomena (and many classic asymptotic results in causal inference) in finite samples
- Many common nonparametric methods (i.e. trees, tree ensembles) condition on adaptive discretizations of continuous variables

In this case, feature selection becomes "adaptive stratification"

Define $s(X)$ as a "stratification function" which maps unique level sets of X to strata indices which will be used as adjustment sets

For example, the trivial stratification below assigns every unique value of X to its own stratum

X_1	X_2	X_3	$s(X)$
0	0	0	1
1	0	0	2
0	1	0	3
1	1	0	4
0	0	1	5
1	0	1	6
0	1	1	7
1	1	1	8

We can perform variable selection with $s(X)$ (1/2)

For example, dropping $X_1 \ldots$

X_1	X_2	X_3	$s(X)$
0	0	0	1
1	0	0	1
0	1	0	2
1	1	0	2
0	0	1	3
1	0	1	3
0	1	1	4
1	1	1	4

We can perform variable selection with $s(X)$ (2/2)

. . . or dropping X_2. . .

X_1	X_2	X_3	$s(X)$
0	0	0	1
1	0	0	2
0	1	0	1
1	1	0	2
0	0	1	3
1	0	1	4
0	1	1	3
1	1	1	4

Feature selection vs variable selection

$s(X)$ allows us to define "features" formed out of nonlinear combinations of variables. For example, we can define

$$s(X) = 2 + 2X_3 - \mathcal{I}(X_1 = X_2)$$

X_1	X_2	X_3	$s(X)$
0	0	0	1
1	0	0	2
0	1	0	2
1	1	0	1
0	0	1	3
1	0	1	4
0	1	1	4
1	1	1	3

ATE estimator based on $s(X)$

We can use the levels of $s(X)$ as strata in a classical stratification estimator

$$\bar{\tau}_s = \sum_{s \in s(\mathcal{X})} \frac{n_s}{n} \left(\bar{Y}_{s,Z=1} - \bar{Y}_{s,Z=0} \right)$$

where

$$n_s = \sum_{i=1}^{n} \mathcal{I}\left(s(X_i) = s\right)$$

$$\bar{Y}_{s,Z=1} = \sum_{i=1}^{n} Y_i \mathcal{I}\left(s(X_i) = s\right) \mathcal{I}\left(Z_i = 1\right)$$

$$\bar{Y}_{s,Z=0} = \sum_{i=1}^{n} Y_i \mathcal{I}\left(s(X_i) = s\right) \mathcal{I}\left(Z_i = 0\right)$$

Before we talk about identification, we need some notation

Potential outcomes (Y^0, Y^1): counterfactual random outcome variables in which the treatment is set to 0 or 1 (regardless of the individual's covariates)

Observed outcome, $Y = ZY^1 + (1 - Z)Y^0$: product of potential outcomes and observed treatment assignment

Structural model: rewriting Y^0 and Y^1 in terms of mean and error components

$$\mu(X) = \mathsf{E}\left(Y^0 \mid X\right) \qquad \tau(X) = \mathsf{E}\left(Y^1 \mid X\right) - \mu(X)$$
$$\nu(X, \epsilon_y) = Y^0 - \mu(X) \qquad \delta(X, \epsilon_y) = Y^1 - Y^0 - \tau(X)$$

where $\mathsf{E}\left(\nu(X, \epsilon_y) \mid X = x\right) = 0$ and $\mathsf{E}\left(\delta(X, \epsilon_y) \mid X = x\right) = 0$ for all x

So when does $s(X)$ identify the ATE

We can use the structural model decomposition to write

$$Y = \underbrace{\mu(X) + \tau(X)Z}_{\text{Mean term}} + \underbrace{[\upsilon(X, \epsilon_y) + \delta(X, \epsilon_y)Z]}_{\text{Error term}}$$

And through this lens, we identify the ATE ($\mathrm{E}\left[Y^1\right] - \mathrm{E}\left[Y^0\right]$) when $\mu(X), \tau(X) \perp\!\!\!\perp Z \mid s(X)$

Note that this assumption, **mean** conditional unconfoundedness, is weaker than the typically-invoked conditional unconfoundedness assumption $(Y^0, Y^1) \perp\!\!\!\perp Z \mid s(X)$.

With mean conditional unconfoundedness, estimands such as the quantile treatment effect are not identified.

Recap: where are we so far?

We have

1. A way of representing feature selection mathematically ($s(X)$)
2. ATE estimator that uses $s(X)$
3. Identification criterion for any $s(X)$

. . . but $(\mu(X), \tau(X))$ are unobservable. . .

A Bayesian approach to feature selection for statistical control in ATE estimation

We choose $s(X)$ by MAP estimation in the following model

$$
\begin{aligned}
Y \mid s(X) = s, Z = z &\sim \mathcal{N}\left(\mu(s) + \tau(s)z, \sigma^2\right) \\
s(X) &\sim p_s \\
\mu(s) \mid s(X) &\propto 1 \\
\tau(s) \mid s(X) &\propto 1 \\
\sigma^2 &\propto 1
\end{aligned}
$$

where p_s assigns prior probabilities to each $s(X)$ function.

Setting a flat prior p_s, we can see by a standard result in linear regression that we will prefer $s(X) = X$ (i.e. the likelihood alone will not perform feature selection)

We can "favor" certain $s(X)$ functions with an informative prior

$$
\begin{aligned}
Y \mid s(X) = s, Z = z &\sim \mathcal{N}\left(\mu(s) + \tau(s)z, \sigma^2\right) \\
s(X) &\sim p_s \\
\mu(s) \mid s(X) &\propto 1 \\
\tau(s) \mid s(X) &\propto 1 \\
\sigma^2 &\propto 1
\end{aligned}
$$

Note: not every $s(X)$ identifies the average treatment effect

A problem with the naive approach

Define a naive prior that penalizes the size of $s(X)$

$$p_s \propto \text{Beta}\left(\frac{|s(X)|}{|X|}; 1, \alpha\right)$$

Increasing α expresses a preference for smaller $|s(X)|$, so this prior penalizes large stratification functions...

...but this can bias the ATE severely by dropping confounding features in favor of non-confounders that are strongly associated with the outcome (discussed in Hahn et al. (2018) as "regularization-induced confounding")

The propensity score offers one way to coarsen X without confounding the ATE

Rosenbaum and Rubin (1983) demonstrated that rather than controlling for all of X, we can condition on the "propensity score" (which we'll call $\pi(X) = \mathrm{P}(Z = 1|X)$)

So setting $s(X)$ to have the same unique levels as $\pi(X)$ is one viable option for "doing better" than $s(X) = X$.

But we need not think of a single propensity score, we can project Z onto any $s(X)$:

$$\pi(s(X)) = \mathrm{P}(Z = 1|s(X))$$

We can penalize "excess control" using the relationship between $s(X)$ and Z

With $\pi(s(X)) = \mathrm{E}\left(Z \mid s(X)\right)$, we define the $s(X)$ prior

$$p_s \propto \text{Beta}\left(\frac{|\pi(s(X))|}{|s(X)|}; \alpha, 1\right)$$

Increasing α expresses a preference for $|\pi(s(X))| = |s(X)|$, so this prior penalizes

- Noise features (associated with neither treatment nor outcome), and
- Features weakly associated with outcome, but not associated with treatment

However, it still leaves us with features that are associated with treatment but not the outcome (typically referred to as "instruments")

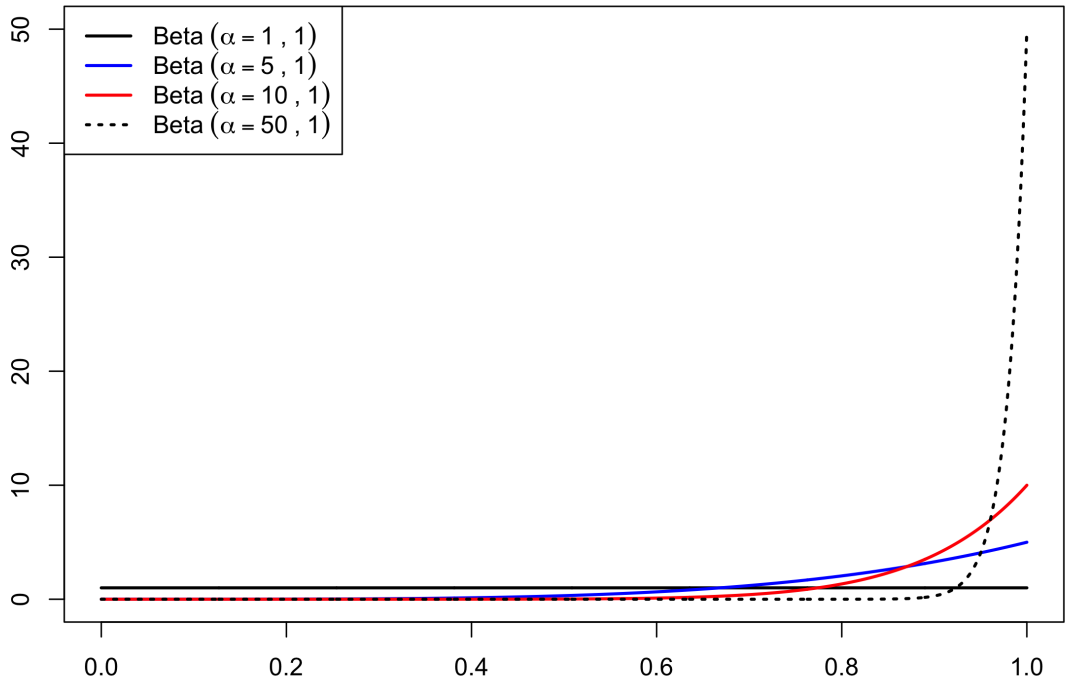
Our proposal: a strong, unbiased excess control prior and a much weaker size prior

We define

$$p_s \propto \underbrace{\text{Beta}\left(\frac{|\pi(s(X))|}{|s(X)|}; \alpha, 1\right)}_{\text{Excess control penalty}} \underbrace{\text{Beta}\left(\frac{|\pi(s(X))|}{\max_s |\pi(s(X))|}; 1, \eta\right)}_{\text{Deconfounding strata size penalty}}$$

Recap: three informative priors p_s

Name	**Density**
Size prior	$\text{Beta}\left(\frac{\lvert s(X)\rvert}{\lvert X\rvert}; 1, \alpha\right)$
Excess control prior	$\text{Beta}\left(\frac{\lvert \pi(s(X))\rvert}{\lvert s(X)\rvert}; \alpha, 1\right)$
Combined prior	$\text{Beta}\left(\frac{\lvert \pi(s(X))\rvert}{\lvert s(X)\rvert}; \alpha, 1\right) \text{Beta}\left(\frac{\lvert \pi(s(X))\rvert}{\max_s \lvert \pi(s(X))\rvert}; 1, \eta\right)$



Future Direction

Methods:

- Develop a performant nonparametric implementation of this regularization approach (i.e. neural networks, trees)

Analytical:

- Study the bias incurred by the combined penalty, develop practical recommendations on setting η

References I

Hahn, P Richard, Carlos M Carvalho, David Puelz, and Jingyu He. 2018. "Regularization and Confounding in Linear Regression for Treatment Effect Estimation." *Bayesian Analysis* 13 (1): 163–82.

Hahn, P Richard, Jared S Murray, and Carlos M Carvalho. 2020. "Bayesian Regression Tree Models for Causal Inference: Regularization, Confounding, and Heterogeneous Effects." *Bayesian Analysis*.

Hansen, Ben B. 2008. "The Prognostic Analogue of the Propensity Score." *Biometrika* 95 (2): 481–88.

Heckman, James J, and Edward Vytlacil. 2005. "Structural Equations, Treatment Effects, and Econometric Policy Evaluation." *Econometrica* 73 (3): 669–738.

Imbens, Guido W, and Donald B Rubin. 2015. *Causal Inference in Statistics, Social, and Biomedical Sciences*. Cambridge University Press.

Pearl, Judea. 2009. *Causality*. Cambridge University Press.

References II

Rosenbaum, Paul R, and Donald B Rubin. 1983. "The Central Role of the Propensity Score in Observational Studies for Causal Effects." *Biometrika* 70 (1): 41–55.

Shalizi, Cosma. 2021. *Advanced Data Analysis from an Elementary Point of View*. Cambridge University Press. https://www.stat.cmu.edu/~cshalizi/ADAfaEPoV/.

Appendix A: Review of Common Causal Frameworks

Potential Outcomes

Most commonly associated with Donald Rubin and Jerzy Neyman (see Imbens and Rubin (2015) for a standard reference).

Defines Y^z as the counterfactual outcome when $Z = z$, so that for binary Z, the observed outcome is $Y = ZY^1 + (1 - Z)Y^0$.

In this context, the ATE is defined as $\mathrm{E}\left[Y^1 - Y^0\right]$ and identified conditional on X assuming:

- X satisfies strong ignorability:
  - $Y^1, Y^0 \perp Z \mid X$
  - $0 < \mathrm{P}(Z = 1 \mid X) < 1$
- SUTVA: no between-subject interference of potential outcomes

Causal DAGs

Most commonly associated with Judea Pearl (see Pearl (2009)).

Represents (X, Z, Y) in a directed graph with edges representing causal relationships and provides algorithms for identifying and estimating causal effects given a graph $\mathcal{G}$.

ATE estimand can be represented in this framework as $\mathrm{E}[Y \mid \mathrm{do}(Z = 1)] - \mathrm{E}[Y \mid \mathrm{do}(Z = 0)]$ where $\mathrm{E}[Y \mid \mathrm{do}(Z = z)]$ involves an intervention on a graph that sets $Z = z$ while leaving other variables unchanged.

The ATE can be identified conditional on X if X satisfies the "backdoor criterion." In words (visuals to follow!), this happens when X blocks any open paths from Z to Y and does not open any new paths from Z to Y.

Structural Equation Models (1/2)

Most commonly associated with James Heckman (see Heckman and Vytlacil (2005) for a review).

Represents Y in terms of conditional mean functions and exogenous error terms, defining

$$\begin{aligned}
\mu(x) &\equiv \mathsf{E}(F(x, 0, \epsilon_y)), \\
\tau(x) &\equiv \mathsf{E}(F(x, 1, \epsilon_y)) - \mu(x), \\
\upsilon(x, \epsilon_y) &\equiv F(x, 0, \epsilon_y) - \mu(x), \\
\delta(x, \epsilon_y) &\equiv F(x, 1, \epsilon_y) - F(x, 0, \epsilon_y) - \tau(x)
\end{aligned}$$

where $F(X, Z, \epsilon_y)$ is a deterministic causal function that generates Y in terms of X, Z, and a random error term ϵ_y.

Structural Equation Models (2/2)

This gives a "structural model"

$$\begin{aligned} Y &= \mu(x) + \upsilon(x, \epsilon_y) + (\tau(x) + \delta(x, \epsilon_y))z \\ &= \underbrace{\mu(x) + \tau(x)z}_{\text{Mean term}} + \underbrace{[\upsilon(x, \epsilon_y) + \delta(x, \epsilon_y)z]}_{\text{Error term}} \end{aligned}$$

The ATE is defined as $\mathsf{E}[\tau(X)]$ and is identified if $(\upsilon(X, \epsilon_y), \delta(X, \epsilon_y)) \perp\!\!\!\perp Z \mid X$

Equivalence between the frameworks

Each framework states a different version of the requirement that X can be used to deconfound the effect of Z on Y

1. X satisfies the "backdoor criterion."
2. $Y^1, Y^0 \perp Z \mid X$
3. $(\upsilon(X, \epsilon_y), \delta(X, \epsilon_y)) \perp\!\!\!\perp Z \mid X$

Without further assumptions, we have that $1 \Rightarrow 2 \Leftrightarrow 3$, and assuming "faithfulness" (see for example Shalizi (2021)) gives $1 \Leftrightarrow 2 \Leftrightarrow 3$

All three frameworks require (implicitly or explicitly) a version of SUTVA and positivity

Verifying identification of the ATE by $s(X)$

Assuming mean conditional unconfoundedness $(\mu(X), \tau(X) \perp\!\!\!\perp Z \mid s(X))$, we see that the ATE is identified:

$$
\begin{aligned}
\mathsf{E}\left(Y \mid s(X), Z=1\right) &= \mathsf{E}\left(\mu(X) \mid s(X), Z=1\right) \\
&\quad + \mathsf{E}\left(\tau(X) \mid s(X), Z=1\right) \\
&\quad + \mathsf{E}\left(\upsilon(X, \epsilon_y) \mid s(X), Z=1\right) \\
&\quad + \mathsf{E}\left(\delta(X, \epsilon_y) \mid s(X), Z=1\right) \\
&= \mathsf{E}\left(\mu(X) \mid s(X)\right) + \mathsf{E}\left(\tau(X) \mid s(X)\right) + 0 + 0 \\
\mathsf{E}\left(Y \mid s(X), Z=0\right) &= \mathsf{E}\left(\mu(X) \mid s(X), Z=1\right) \\
&\quad + \mathsf{E}\left(\upsilon(X, \epsilon_y) \mid s(X), Z=1\right) \\
&= \mathsf{E}\left(\mu(X) \mid s(X)\right)
\end{aligned}
$$

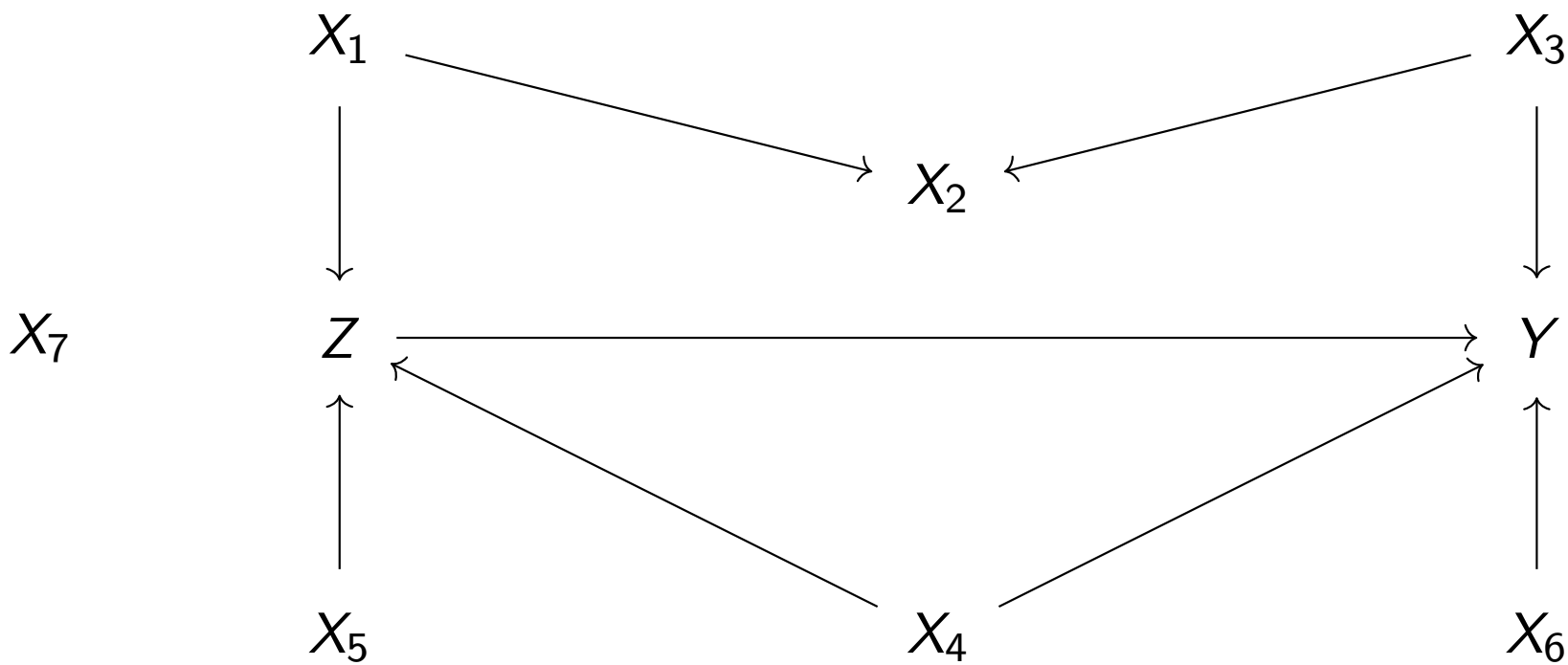
$$
\begin{aligned}
&\mathsf{E}\left(\mathsf{E}\left(Y \mid s(X), Z=1\right) - \mathsf{E}\left(Y \mid s(X), Z=0\right)\right) = \mathsf{E}\left(\mathsf{E}\left(\tau(X) \mid s(X)\right)\right) \\
&\quad = \mathsf{E}\left(\mathsf{E}\left(\mathsf{E}\left(Y^1 \mid X\right) - \mathsf{E}\left(Y^0 \mid X\right) \mid s(X)\right)\right) = \mathsf{E}\left(Y^1\right) - \mathsf{E}\left(Y^0\right)
\end{aligned}
$$

Appendix B: Feature selection for graphs

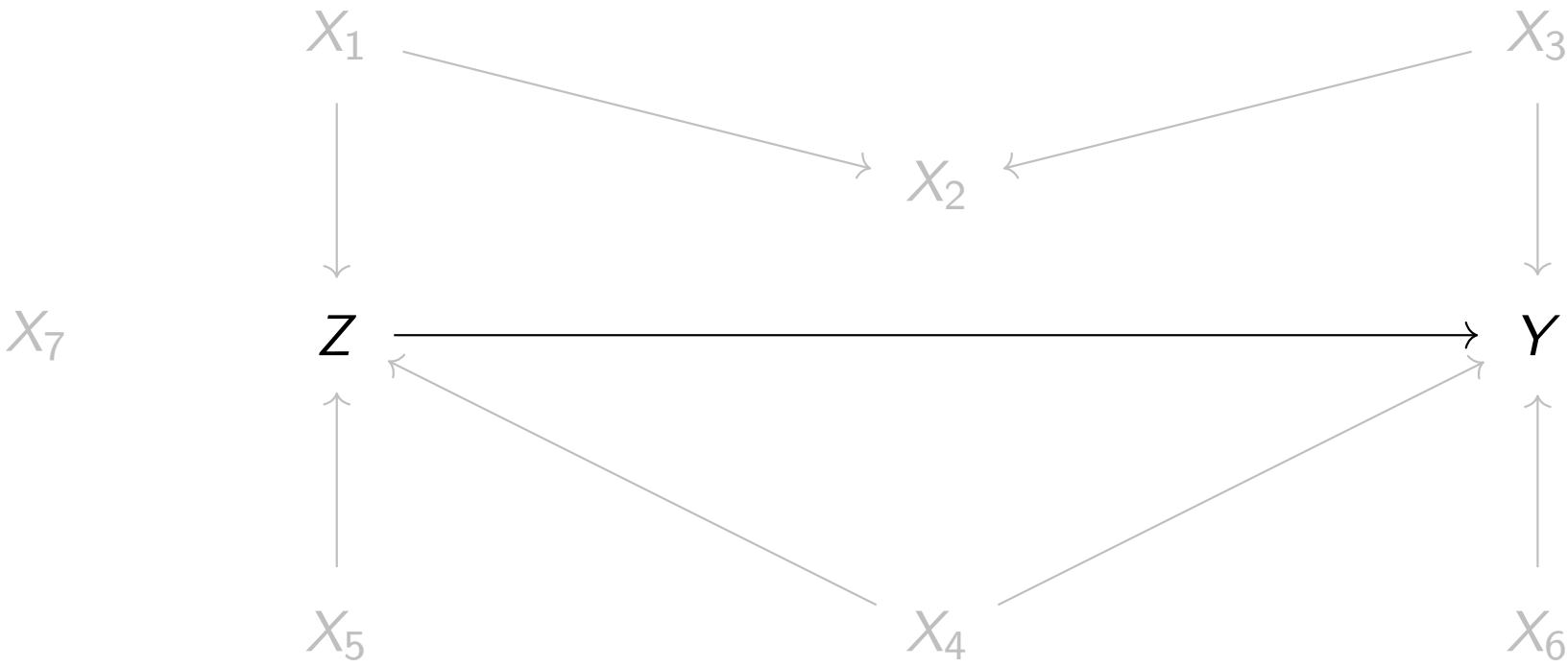
Let's briefly review graph notation and terminology

We have to define / categorize the types of variables that we might encounter in a causal inference problem.

Let's break down this graph into its component variables

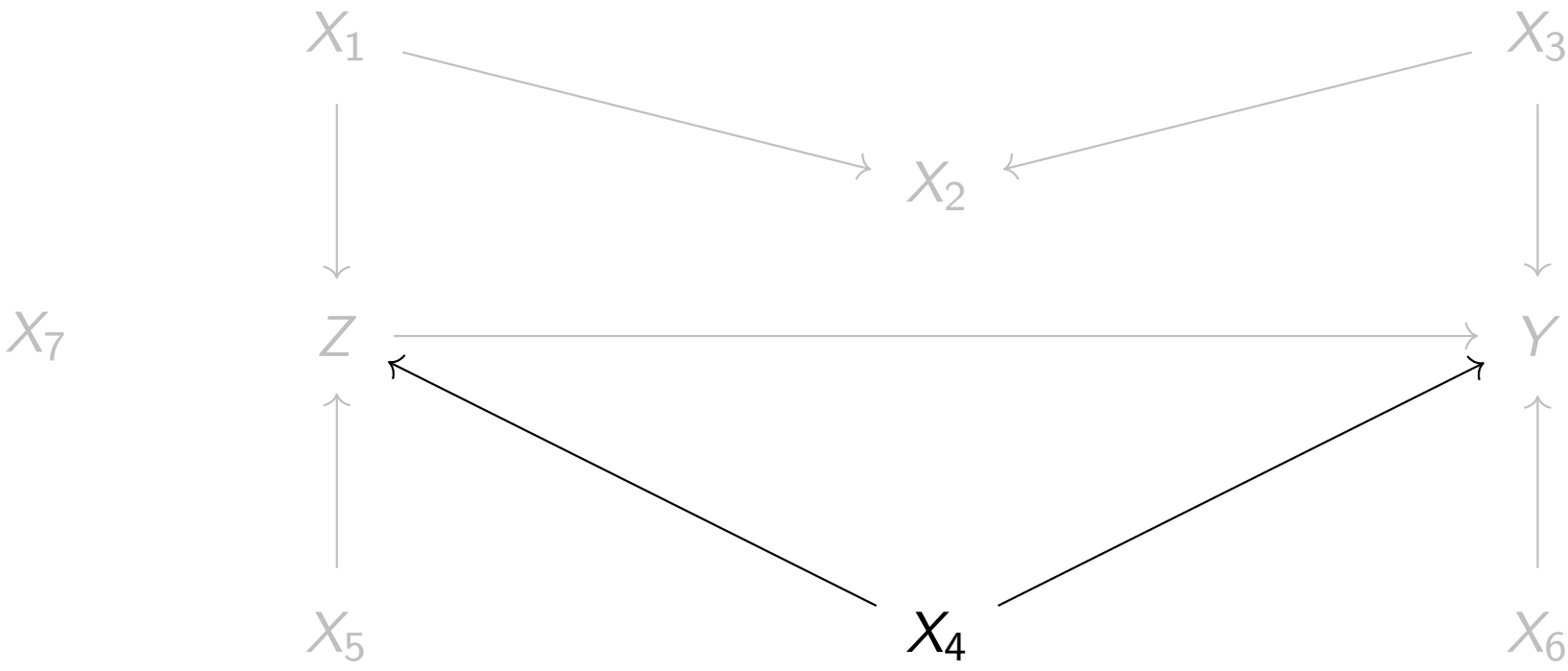


Outcome and Treatment



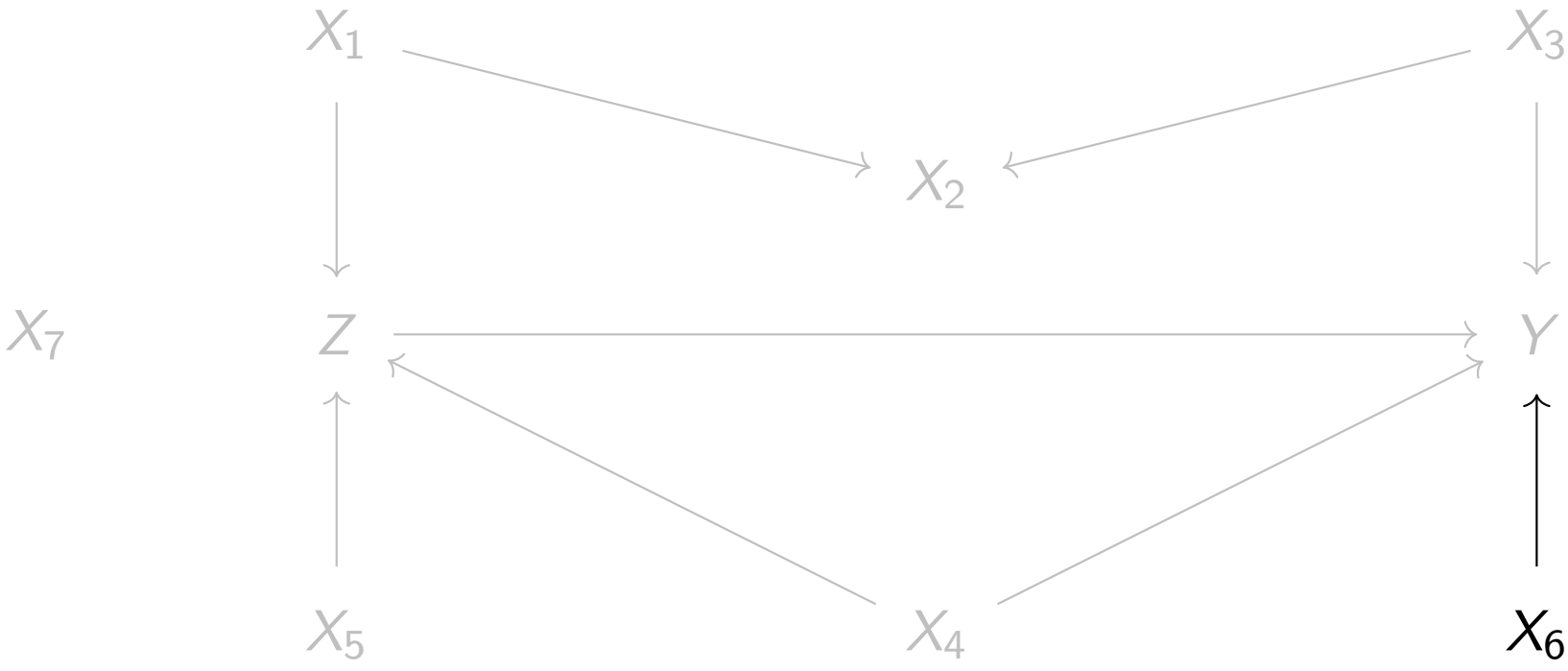
This the "relationship of interest," which we hope to deconfound by defining an appropriate adjustment set.

Confounders



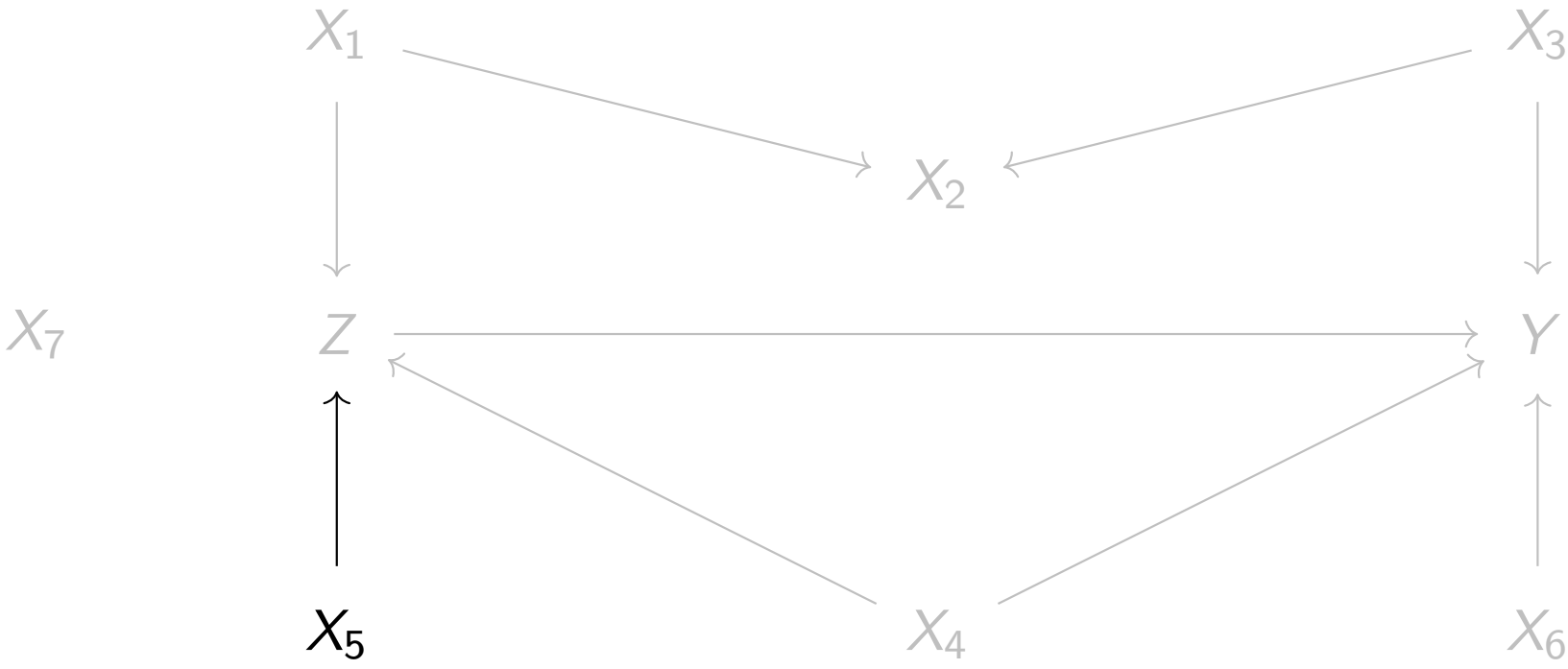
These are variables that impact both treatment and outcome. We definitely need to adjust for them or our estimate of the ATE will be biased.

Prognostic variables



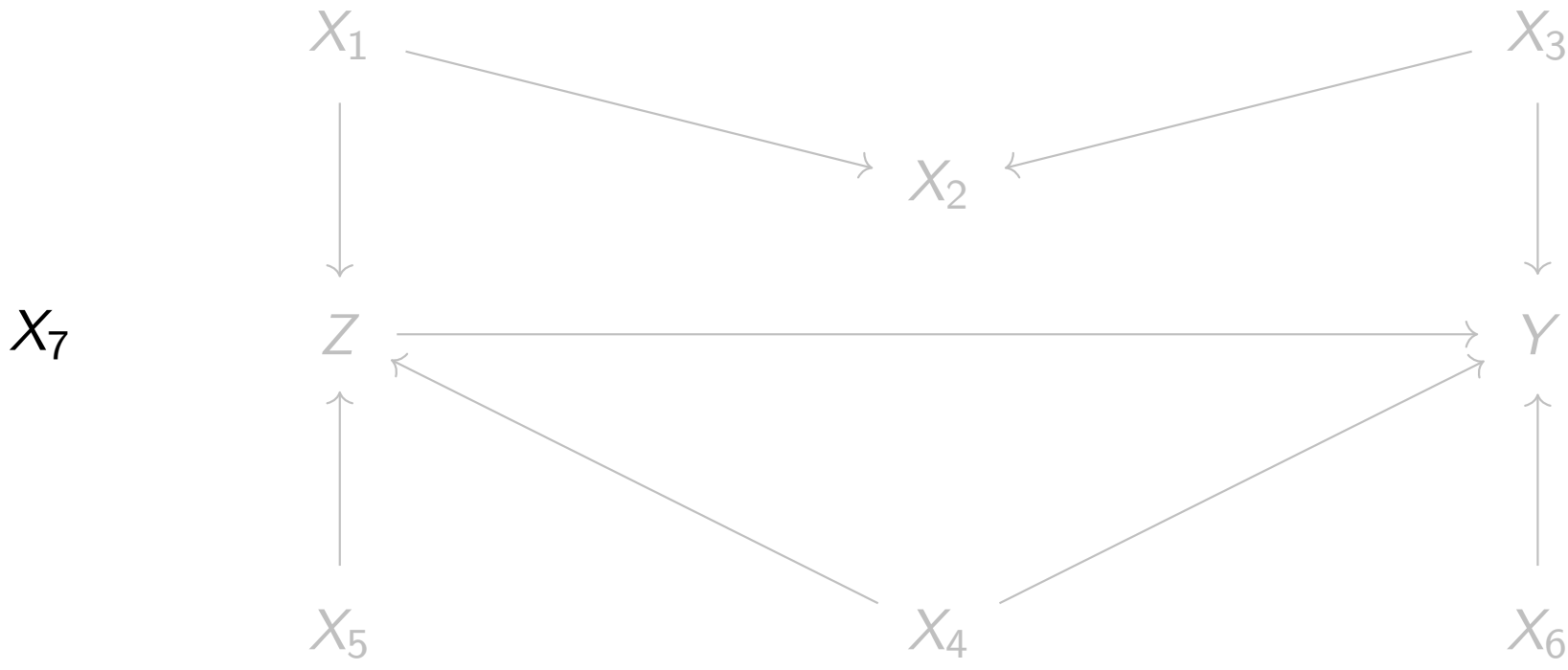
These are variables that only impact the outcome. Even though they don't confound our estimate, they may still be worth controlling for, since they control variability in the outcome.

Instruments



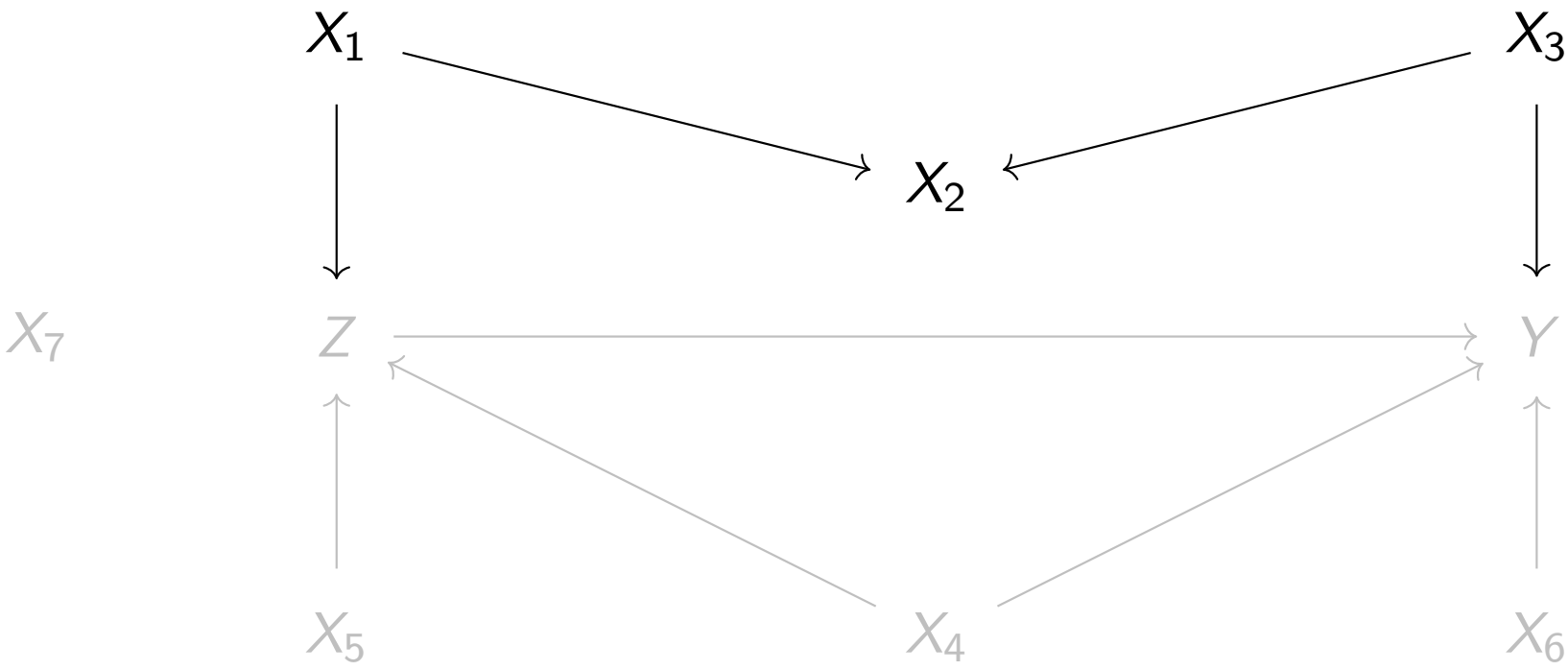
These variables only impact the treatment. We **do not** want to control for these variables if we can help it (unnecessary stratification).

Noise variables



These variables ended up in the covariate set but are just noise. We definitely want to avoid conditioning on these variables.

Colliders



This graph structure has the strange property where:

- If we control for none of (X_1, X_2, X_3), the ATE is identified
- If we **only control for** X_2, the ATE is confounded
- If we control for X_2 and X_1 or X_3, the ATE is identified

Feature selection goals

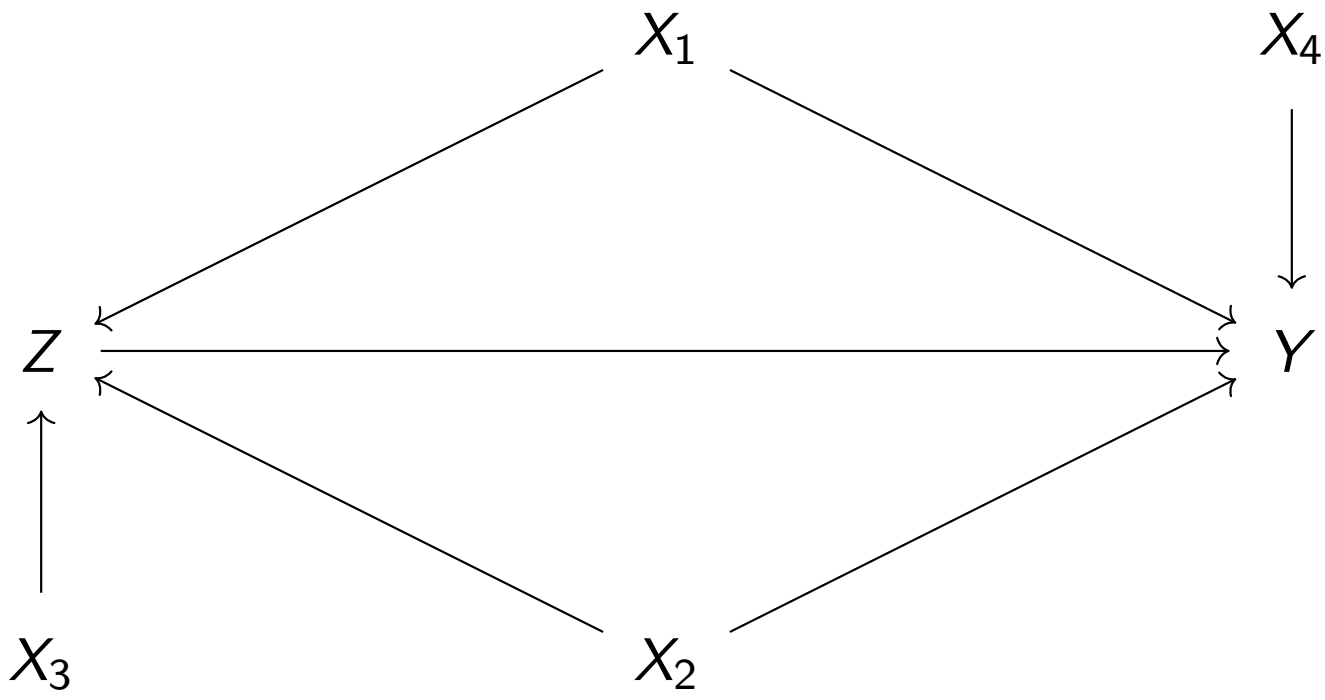
1. Remove noise variables
2. Remove instruments
3. (Maybe) remove any weak prognostic variables
4. Avoid unblocking colliders

This is where the "in a principled manner" comes into play

Appendix C: “Variable selection” vs “feature selection” in more detail

Variable selection is easy when we have the causal graph

Take this graph as an example



By Pearl (2009), (X_1, X_2) satisfy the "back door criterion" which means we must adjust for those variables to identify the ATE

We **might** also want to adjust for X_4, but with this graph, we can confidently remove X_3 from the adjustment set.

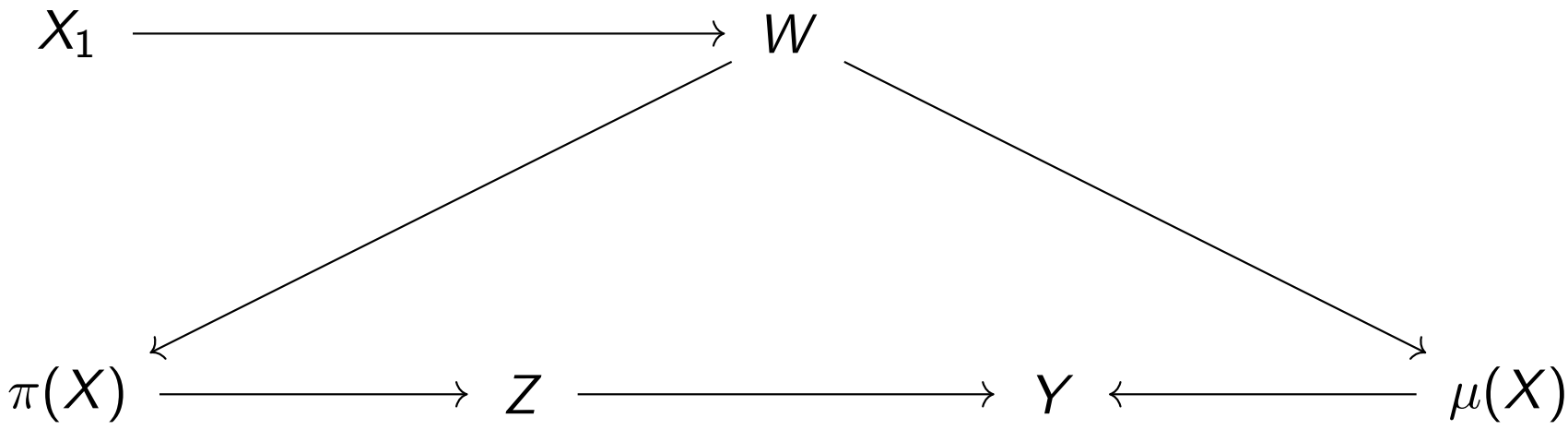
Now, let's consider a specific data generating process that matches this graph

$$
\begin{aligned}
X_1 &\sim \text{Bernoulli}\left(p_1\right) \\
X_2 &\sim \text{Bernoulli}\left(p_2\right) \\
X_3 &\sim \text{Bernoulli}\left(p_3\right) \\
X_4 &\sim \text{Bernoulli}\left(p_4\right) \\
\mu(X) &= \alpha_0 + \alpha_1(2X_1X_2 - X_1 - X_2 + 1) + \alpha_2 X_4 \\
\pi(X) &= \beta_0 + \beta_1(2X_1X_2 - X_1 - X_2 + 1) + \beta_2 X_3 \\
Z &\sim \text{Bernoulli}\left(\pi(X)\right) \\
\varepsilon &\sim \mathcal{N}\left(0, \sigma_\varepsilon^2\right) \\
\tau(X) &= \gamma_0 \\
Y &= \mu(X) + \tau(X)Z + \epsilon
\end{aligned}
$$

We can rewrite the random variables as follows

- $W \mid X_1$ takes X_1, generates a new random variable with distribution Bernoulli (p_2) and combines the two to yield a variable identical to $(2X_1X_2 - X_1 - X_2 + 1)$
- $\mu(X) \mid W$ takes W, generates a new random variable with distribution Bernoulli (p_4) and combines the two to yield a variable identical to $\alpha_0 + \alpha_1 W + \alpha_2 X_4$
- $\pi(X) \mid W$ takes W, generates a new random variable with distribution Bernoulli (p_3) and combines the two to yield a variable identical to $\beta_0 + \beta_1 W + \beta_2 X_3$

This gives a new causal graph



Now, we see that we don't actually have to control for X_1 and X_2, just a synthetic "feature" (W) created by interacting them.

Checking back in: what have we learned?

1. Variable selection for causal inference is dangerous
2. We can do variable selection if we have a causal graph. . .
3. . . . but even then, we could do better!
4. Ultimately, we want to select "features" that deconfound or help with prognostic stratification

Appendix D: Past and present of deconfounding functions in causal inference

From X to $s(X)$

Let X be a discrete (possibly multivariate) random variable with $|X| = k$ unique levels, and define $s(X)$ as a discrete random variable with $|s(X)| \leq |X|$

1. $s(X) = X$ is a trivial stratification function which performs no feature selection
2. $s(X) = 1$ is the coarsest possible stratification function which does not sub-divide any observations

Item 1 is what we are trying to improve upon; 2 is only a valid conditioning set in limited settings (i.e. completely randomized experiments)

There is already an $s(X)$ commonly used in causal inference: the propensity score

Rosenbaum and Rubin (1983) demonstrated that rather than controlling for all of X, we can condition on the "propensity score" (which we'll call $\pi(X) = \mathrm{P}(Z = 1|X)$)

So setting $s(X) = \pi(X)$ is one viable option for "doing better" than $s(X) = X$, however the propensity score still leaves us with instruments.

Another $s(X)$ used in causal inference: the prognostic score

Hansen (2008) defined a different projection: "prognostic score" (which we'll call $\mu(X) = E\left[Y^0|X\right]$)

When the treatment effect is constant (i.e. $\tau(x) = \tau$ for all x), $s(X) = \mu(X)$ provides another viable conditioning set. More generally, $(\mu(X), \tau(X))$ together identify the ATE when there is heterogeneity.

However, defining $s(X)$ by the level sets of $\mu(X)$ or $(\mu(X), \tau(X))$ leaves us with prognostic variables.

Furthermore, even if $\mu(X)$ can be estimated from the data, learning (a function of) $\tau(X)$ is the goal of most of causal inference. If we already had $\tau(X)$, we could stop right there!

Relaxing assumptions for the ATE: **mean** conditional unconfoundedness vs conditional unconfoundedness

The conditional unconfoundedness assumption reviewed in the context of all three frameworks is stronger than necessary for ATE estimation.

The averages of Y^0 and Y^1 are driven by their mean functions $\mu(X)$ and $\mu(X) + \tau(X)$. Compare

- Conditional unconfoundedness: $(Y^0, Y^1) \perp\!\!\!\perp Z \mid s(X)$
- Mean conditional unconfoundedness: $(\mu(X), \tau(X)) \perp\!\!\!\perp Z \mid s(X)$

Mean conditional unconfoundedness identifies the ATE, but not other estimands such as the quantile treatment effect (QTE).

Principal deconfounding function: smallest possible conditioning set

Define

$$\lambda(X) = \mathsf{E}\left(\pi(X) \mid \mu(X), \tau(X)\right)$$

where $\pi(X) = \mathsf{P}\left(Z = 1 \mid X\right)$

The unique level sets of $\lambda(X)$ define the coarsest possible stratification function that satisfies mean conditional unconfoundedness.

Great! So let's just compute this and condition on it for our empirical work?

. . . Unfortunately, no. There are several problems:

- If we could reliably estimate $\tau(X)$ from the data (such that we were comfortable conditioning on it), we wouldn't need to do feature selection at all
- Conditioning on $\lambda(X)$ removes prognostic variables, but we'd prefer a procedure that may optionally include prognostic variables if they have strong effects (for variance reduction purposes)

Appendix E: Regularization-induced confounding in more detail

Regularization-induced confounding (RIC) in linear models

Hahn et al. (2018) introduce the idea of "regularization-induced confounding"

Suppose we are fitting a linear model

$$Y = \alpha + \tau Z + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon$$

and we put a steep regularization penalty on the weights $\beta_1, \ldots, \beta_p$ because:

(a) there are many covariates, and
(b) we are not sure they're all "necessary" andwant to control the variance of our estimator $\hat{\tau}$.

If some of the X_i variables are strongly associated with Z, the estimate of τ will incorporate some of the effects $\beta_1 \ldots X_p$ on Y since we placed constraints on how big $\hat{\beta}_1, \ldots, \hat{\beta}_p$ can be.

This occurs more generally in a stratification setting (1/n)

For a given stratification function $s(X)$, we define

$$\begin{aligned}
\mu(s(X)) &\equiv \mathsf{E}(Y^0 \mid s(X)), \\
\tau(s(X)) &\equiv \mathsf{E}(Y^1 \mid s(X)) - \mu(s(X)), \\
\upsilon(s(X), \epsilon_y) &\equiv Y^0 - \mu(s(X)), \\
\delta(s(X), \epsilon_y) &\equiv (Y^1 - Y^0) - \tau(s(X))
\end{aligned}$$

giving a new "structural model" characterized by $s(X)$ rather than all of X

$$Y = \mu(s(X)) + \upsilon(s(X), \epsilon_y) + (\tau(s(X)) + \delta(s(X), \epsilon_y))Z$$

This occurs more generally in a stratification setting (2/n)

We use the familiar stratification estimator of the ATE:

$$\bar{\tau}_s = \sum_{s \in s(\mathcal{X})} \frac{n_s}{n} \left(\bar{Y}_{s,z=1} - \bar{Y}_{s,z=0} \right)$$

and we can represent the predicted value of Y for any s by $\bar{\tau}_s$ and two other terms, which we define below

$$\begin{aligned} \hat{Y} &= \hat{\mu}(s) + Z \left(\bar{\tau}_s + \hat{t}(s) \right) \\ \hat{\mu}(s) &= \bar{Y}_{s,Z=0} \\ \hat{t}(s) &= \left(\bar{Y}_{s,Z=1} - \bar{Y}_{s,Z=0} \right) - \bar{\tau}_s \end{aligned}$$

This occurs more generally in a stratification setting (3/n)

The true structural component of Y can be similarly decomposed for any x

$$\mu(x) + Z\tau(x) = \mu(x) + Z\left(\tau_x + t(x)\right)$$
$$\tau_x = \mathbb{E}\left[\tau(X)\right]$$
$$t(x) = \tau(x) - \mathbb{E}\left[\tau(X)\right]$$

This occurs more generally in a stratification setting (4/n)

This can also be written in terms of covariate strata $s(X)$

$$\tau_s = \mathbb{E}_{s(X)} \left[\mathbb{E} \left(\Delta_s \mid s(X) \right) \right]$$
$$t(s) = \mathbb{E} \left(\Delta_s \mid s(X) = s \right) - \mathbb{E}_{s(X)} \left[\mathbb{E} \left(\Delta_s \mid s(X) \right) \right]$$

where $\Delta_s = \mathbb{E} \left[Y \mid s(X) = s, Z = 1 \right] - \mathbb{E} \left[Y \mid s(X) = s, Z = 0 \right]$. If $s(X)$ does not satisfy mean conditional unconfoundedness, then τ_s is not necessarily equal to $\mathrm{E} \left(\tau(X) \right)$.

This occurs more generally in a stratification setting (5/n)

For any random vector (Y, X, Z) for which $s(X) = s$, we have that

$$\begin{aligned}\left(\hat{Y} - Y\right)^2 &= \left(\hat{\mu}(s) + Z\bar{\tau}_s + Z\hat{t}(s) - Y\right)^2 \\ &= \left(\hat{\mu}(s) + Z\bar{\tau}_s + Z\hat{t}(s) - \mu(x) - Z\tau_x - Zt(x)\right)^2 \\ &\quad + \left(\mu(x) + Z\tau_x + Zt(x) - Y\right)^2 \\ &\quad + 2\left(\hat{Y} - \mu(x) - Z\tau_x - Zt(x)\right)\left(\mu(x) + Z\tau_x + Zt(x) - Y\right)\end{aligned}$$

1. The first term constitutes the "prediction error" of $\hat{\mu}(s) + Z\bar{\tau}_s + Z\hat{t}(s)$ with respect to the true structural model $\mu(x) + Z\tau_x + Zt(x)$
2. The second term is a stratification-independent measure of the magnitude of the outcome noise, and
3. The third term is the double the covariance of $\hat{Y}$ and the outcome noise term

This occurs more generally in a stratification setting (6/n)

We compare estimators based on different stratification functions $s(x)$ via their MSE $\mathbb{E}\left(\hat{Y} - Y\right)^2$. Since $(\mu(x) + Z\tau_x + Zt(x) - Y)^2$ does not depend on the choice of $s(X)$, we denote its expectation as σ^2. Similarly, $\text{Cov}\left(\hat{Y}, \mu(x) + Z\tau_x + Zt(x) - Y\right)$ is 0 in expectation, so we focus our analysis on the first term.

$$
\begin{aligned}
&\left(\hat{\mu}(s) + Z\bar{\tau}_s + Z\hat{t}(s) - \mu(x) - Z\tau_x - Zt(x)\right)^2 \\
&\quad = \left((\hat{\mu}(s) - \mu(x)) + Z\left(\bar{\tau}_s - \tau_x\right) + Z\left(\hat{t}(s) - t(x)\right)\right)^2 \\
&\quad = \left(\hat{\mu}(s) - \mu(x)\right)^2 + Z\left(\bar{\tau}_s - \tau_x\right)^2 + Z\left(\hat{t}(s) - t(x)\right)^2 \\
&\quad + 2Z\left(\hat{\mu}(s) - \mu(x)\right)\left(\bar{\tau}_s - \tau_x\right) \\
&\quad + 2Z\left(\hat{\mu}(s) - \mu(x)\right)\left(\hat{t}(s) - t(x)\right) \\
&\quad + 2Z\left(\bar{\tau}_s - \tau_x\right)\left(\hat{t}(s) - t(x)\right)
\end{aligned}
$$

This occurs more generally in a stratification setting (7/n)

We evaluate the expectation of this expression in parts. First, note that

$$\begin{aligned}
\mathbb{E}\left(\hat{\mu}(s) - \mu(x)\right)^2 &= \mathbb{E}\left(\left(\hat{\mu}(s) - \mu(s)\right) + \left(\mu(s) - \mu(x)\right)\right)^2 \\
&= \mathbb{E}\left(\hat{\mu}(s) - \mu(s)\right)^2 + \mathbb{E}\left(\mu(s) - \mu(x)\right)^2 \\
&\quad + 2\mathbb{E}\left[\left(\hat{\mu}(s) - \mu(s)\right)\left(\mu(s) - \mu(x)\right)\right] \\
&= \mathbb{E}\left(\mathbb{E}\left(\left(\hat{\mu}(s) - \mu(s)\right)^2 \mid s(X) = s\right)\right) + 0 + 0 \\
&= \mathbb{E}\left(\ \mathrm{Var}\left(\hat{\mu}(s) \mid s(X) = s\right)\right)
\end{aligned}$$

This occurs more generally in a stratification setting (8/n)

Now, note that

$$\begin{aligned}
\mathbb{E}\left[Z\left(\bar{\tau}_s - \tau_x\right)^2\right] &= \mathbb{E}\left[\mathbb{E}\left[Z\left(\left(\bar{\tau}_s - \tau_s\right) + \left(\tau_s - \tau_x\right)\right)^2 \mid Z\right]\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[Z\left(\bar{\tau}_s - \tau_s\right)^2 \mid Z\right]\right] \\
&\quad + \mathbb{E}\left[\mathbb{E}\left[Z\left(\tau_s - \tau_x\right)^2 \mid Z\right]\right] \\
&\quad + 2\mathbb{E}\left[\mathbb{E}\left[Z\left(\bar{\tau}_s - \tau_s\right)\left(\tau_s - \tau_x\right) \mid Z\right]\right] \\
&= \mathbb{E}\left(\pi(X)\right)\left[\ \mathrm{Var}\left(\bar{\tau}_s\right) + \ \mathrm{Bias}\left(\bar{\tau}_s\right)^2\right]
\end{aligned}$$

This occurs more generally in a stratification setting (9/n)

Also consider that

$$\begin{aligned}
\mathbb{E}\left(Z\left(\hat{t}(s)-t(x)\right)^2\right) &= \mathbb{E}Z\left(\left(\hat{t}(s)-t(s)\right)+\left(t(s)-t(x)\right)\right)^2 \\
&= \mathbb{E}\left(Z\hat{t}(s)-t(s)\right)^2 + Z\mathbb{E}\left(t(s)-t(x)\right)^2 \\
&\quad + 2\mathbb{E}Z\left[\left(\hat{t}(s)-t(s)\right)\left(t(s)-t(x)\right)\right] \\
&= \mathbb{E}Z\left(\hat{t}(s)-t(s)\right)^2 + \mathbb{E}Z\left(t(s)-t(x)\right)^2 + 0
\end{aligned}$$

This occurs more generally in a stratification setting (10/n)

Similarly, note that

$$
\begin{aligned}
&\mathbb{E}\left[2Z\left(\hat{\mu}(s)-\mu(x)\right)\left(\bar{\tau}_s-\tau_x\right)\right] \\
&\quad=2\mathbb{E}\left[Z\left(\hat{\mu}(s)-\mu(s)+\mu(s)-\mu(x)\right)\left(\bar{\tau}_s-\tau_s+\tau_s-\tau_x\right)\right] \\
&\quad=2\mathbb{E}\left[Z\left(\hat{\mu}(s)-\mu(s)+\mu(s)-\mu(x)\right)\left(\bar{\tau}_s-\tau_s+\tau_s-\tau_x\right)\right] \\
&\quad=2\mathbb{E}\left[Z\left(\hat{\mu}(s)-\mu(s)\right)\left(\bar{\tau}_s-\tau_s\right)\right] \\
&\qquad+2\mathbb{E}\left[Z\left(\mu(s)-\mu(x)\right)\left(\bar{\tau}_s-\tau_s\right)\right] \\
&\qquad+2\mathbb{E}\left[Z\left(\hat{\mu}(s)-\mu(s)\right)\left(\tau_s-\tau_x\right)\right] \\
&\qquad+2\mathbb{E}\left[Z\left(\mu(s)-\mu(x)\right)\left(\tau_s-\tau_x\right)\right] \\
&\quad=2\mathbb{E}\left[Z\left(\hat{\mu}(s)-\mu(s)\right)\left(\bar{\tau}_s-\tau_s\right)\right]+0+0+0 \\
&\quad=2\mathbb{E}_{s(X)}\left[\mathbb{E}\left[Z\left(\hat{\mu}(s)-\mu(s)\right)\left(\bar{\tau}_s-\tau_s\right)\mid s(X)\right]\right] \\
&\quad=2\mathbb{E}_{s(X)}\left[\mathrm{Cov}\left(\hat{\mu}(s), Z\bar{\tau}_s\mid s(X)\right)\right] \\
&\quad=2\left[\mathrm{Cov}\left(\hat{\mu}(s), Z\bar{\tau}_s\right)-\mathrm{Cov}\left(\mathbb{E}\left[\hat{\mu}(s)\mid s(X)\right], \mathbb{E}\left[Z\bar{\tau}_s\mid s(X)\right]\right)\right] \\
&\quad=2\left[\mathrm{Cov}\left(\hat{\mu}(s), Z\bar{\tau}_s\right)-\mathrm{Cov}\left(\mu(s), \pi(s(X))\bar{\tau}_s\right)\right]
\end{aligned}
$$

This occurs more generally in a stratification setting (11/n)

Thus, our objective function decomposes into a sum of several objectives

$$\begin{aligned}
\mathbb{E}\left(\hat{Y} - Y\right)^2 = & \mathbb{E}\left(\text{Var}\left(\hat{\mu}(s) \mid s(X) = s\right)\right) \\
& + \mathbb{E}\left(\pi(X)\right)\left[\text{Var}\left(\bar{\tau}_s\right) + \text{Bias}\left(\bar{\tau}_s\right)^2\right] \\
& + \mathbb{E}Z\left(\hat{t}(s) - t(s)\right)^2 + \mathbb{E}Z\left(t(s) - t(x)\right)^2 \\
& + 2\left[\text{Cov}\left(\hat{\mu}(s), Z\bar{\tau}_s\right) - \text{Cov}\left(\mu(s), \pi(s(X))\bar{\tau}_s\right)\right] \\
& + 2\mathbb{E}_{s(X)}\left[\text{Cov}\left(\hat{\mu}(s), Z\hat{t}(s) \mid s(X)\right)\right] \\
& + 2\mathbb{E}_{s(X)}\left[\text{Cov}\left(Z\hat{t}(s), Z\bar{\tau}_s \mid s(X)\right)\right]
\end{aligned}$$

In an unconstrained optimization of $\mathbb{E}\left(\hat{Y} - Y\right)^2$, we minimize the MSE by setting $s(X) = X$. But if place a regularization penalty on $|s(X)|$, we may favor smaller subsets of X that do not increase the MSE substantially.

This occurs more generally in a stratification setting (12/n)

Here we see that the degree of bias this can induce in $\bar{\tau}_s$ does not substitute in a bivariate fashion with $\text{Var}(\bar{\tau}_s)$ as in the classic *bias-variance tradeoff*.

Instead, the bias incurred by penalizing $|s(X)|$ trades off with 8 other terms, namely. We can reduce the MSE by

1. Conditioning on sets $s(X)$ that reduce variance of $\hat{\mu}(s)$
2. Conditioning on sets $s(X)$ that increase covariance of $\mu(s)$ and $\pi(s(X))\bar{\tau}_s$

The first scenario is possible when $\mu(X)$ is large in magnitude relative to $\tau(X)$. The second scenario is possible when $\mu(X)$ is strongly correlated to treatment probabilities (a phenomenon referred to as "targeted selection" in Hahn, Murray, and Carvalho (2020)).

Appendix F: Equivalence of the stratification, [saturated] regression, and IPW estimators

When you hear "conditional on" / "adjusting for," think regression

With discrete X, we can obtain the exact same result using a stratification estimator, an IPW estimator, or a saturated linear regression model

$$Y = \beta_0 + \beta_1 Z + \beta_2 X_1 + \cdots + \beta_{p+2} X_1 Z + \cdots + \epsilon$$

Essentially, we estimate a different model $Y_x = \beta_{0,x} + \beta_{1,x} Z_x$ for each unique $x \in \mathcal{X}$ and weight the estimates accordingly.

When we talk about "controlling for" variables, what we typically mean for the purposes of this talk is this style of regression.

Define the three estimators as follows

$$\bar{\tau}_{IPW} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{Y_i Z_i}{\hat{p}(X_i)} - \frac{Y_i(1-Z_i)}{1-\hat{p}(X_i)}\right)$$

$$\bar{\tau}_{strat} = \sum_{x\in\mathcal{X}}\frac{n_x}{n}\left(\bar{Y}_{x,Z=1} - \bar{Y}_{x,Z=1}\right)$$

$$\bar{\tau}_{reg} = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{Y}_{Z=1,X=x_i} - \hat{Y}_{Z=0,X=x_i}\right)$$

where

$$\hat{p}(x) = \frac{N_{x,Z=1}}{n_x}$$

$$N_{x,Z=1} = \sum_{i=1}^{n}\mathbf{1}(X_i = x, Z = 1)$$

$$n_x = \sum_{i=1}^{n}\mathbf{1}(X_i = x)$$

and the regression fit for $\bar{\tau}_{reg}$ is a fully saturated linear model

$$Y = \beta_I + \beta_Z Z + \beta_{X_1} X_1 + \cdots + \beta_{X_1,Z} X_1 Z + \cdots + \epsilon$$

Now, we compare the IPW and stratification estimators

$$
\begin{aligned}
\bar{\tau}_{IPW} &= \frac{1}{n}\sum_{i=1}^{n}\left(\frac{Y_i Z_i}{\hat{p}(X_i)} - \frac{Y_i(1-Z_i)}{1-\hat{p}(X_i)}\right) \\
&= \frac{1}{n}\sum_{x\in\mathcal{X}}\left(\frac{n_x N_{x,Z=1}\bar{Y}_{x,Z=1}}{N_{x,Z=1}} - \frac{n_x N_{x,Z=0}\bar{Y}_{x,Z=0}}{N_{x,Z=0}}\right) \\
&= \frac{1}{n}\sum_{x\in\mathcal{X}}\left(n_x\bar{Y}_{x,Z=1} - n_x\bar{Y}_{x,Z=0}\right) \\
&= \sum_{x\in\mathcal{X}}\frac{n_x}{n}\left(\bar{Y}_{x,Z=1} - \bar{Y}_{x,Z=0}\right) = \bar{\tau}_{strat}
\end{aligned}
$$

Finally, we compare the regression and stratification estimators

$$\begin{aligned}
\bar{\tau}_{reg} &= \frac{1}{n}\sum_{i=1}^{n}\left(\hat{Y}_{Z=1,X=x_i} - \hat{Y}_{Z=0,X=x_i}\right) \\
&= \sum_{x\in\mathcal{X}}\frac{n_x}{n}\left(\hat{Y}_{Z=1,X=x} - \hat{Y}_{Z=0,X=x}\right)
\end{aligned}$$

since X is discrete, we can represent $\hat{Y}_{z,x}$ using cell means

$$= \sum_{x\in\mathcal{X}}\frac{n_x}{n}\left(\bar{Y}_{x,Z=1} - \bar{Y}_{x,Z=1}\right) = \bar{\tau}_{strat}$$

We can see this in R. First, we generate some data.

```r
# Define functions
pi_func <- function(x) .25 + .25*(x > 2) + .25*(x > 4)
mu_func <- function(x) x
tau_func <- function(x) 2 - 1*(x > 2) + 0.5*(x > 4)

# Generate data
n <- 1000
x <- sample(1:5, size = n, replace = T)
pi_x = pi_func(x)
mu_x = mu_func(x)
tau_x = tau_func(x)
z <- rbinom(n, 1, pi_x)
eps <- rnorm(n, 0, 0.5*sd(mu_x))
y <- mu_x + z*tau_x + eps
ATE_true = mean(tau_x)
```

Fit the stratification estimator

```
contrast_func = function(i) {
  mean(y[z==1 & x == i]) - mean(y[z==0 & x == i])
}
strata_contrast = sapply(1:5, contrast_func)
strata_weights = sapply(1:5, function(i) sum(x == i) / n)
(tau_hat_strat = sum(strata_contrast*strata_weights))
```

```
## [1] 1.550431
```

Fit the IPW estimator

```
pi_hat_x = sapply(
  1:5, function(i) mean(z[x == i])
)[x]
ipw_summand = (
  ((y*z)/(pi_hat_x)) - ((y*(1-z))/(1-pi_hat_x))
)
(tau_hat_ipw = sum(ipw_summand))/n
```

```
## [1] 1.550431
```

Fit the regression estimator

```
saturated_model = lm(y ~ as.factor(x)*as.factor(z))
y_hat_1 = predict(
  saturated_model, newdata = data.frame(z=1, x=x)
)
y_hat_0 = predict(
  saturated_model, newdata = data.frame(z=0, x=x)
)
(tau_hat_reg = mean(y_hat_1 - y_hat_0))
```

```
## [1] 1.550431
```