Lectures on causal inference and experimental methods

Macartan Humphreys

Section 1

Inquiries

Subsection 1

Estimands and inquiries

Estimands and inquiries

- Your inquiry is your question and the estimand is the true (generally unknown) answer to the inquiry
- The estimand is the thing you want to estimate
- If you are estimating something you should be able to say what your estimand is
- You are responsible for your estimand. Your estimator will not tell you what your estimand is
- Just because you can calculate something does not mean that you have an estimand
- You can test a hypothesis without having an estimand

Read: II ch 4, DD, ch 7

Estimands: ATE, ATT, ATC, S-, P-, C-, ITT, LATE

Say that units are randomly assigned to treatment in different strata (maybe just one); with fixed, though possibly different, shares assigned in each stratum. Then the key estimands and estimators are:

Estimand	Estimator
$\tau_{ATE} \equiv \mathbb{E}[\tau_i]$	$\hat{\tau}_{ATE} = \sum_x \frac{w_x}{\sum_j w_j}\hat{\tau}_x$
$\tau_{ATT} \equiv \mathbb{E}[\tau_i \| Z_i = 1]$	$\hat{\tau}_{ATT} = \sum_x \frac{p_x w_x}{\sum_j p_j w_j}\hat{\tau}_x$
$\tau_{ATC} \equiv \mathbb{E}[\tau_i \| Z_i = 0]$	$\hat{\tau}_{ATC} = \sum_x \frac{(1-p_x) w_x}{\sum_j (1-p_j) w_j}\hat{\tau}_x$

where x indexes strata, p_x is the share of units in each stratum that is treated, and w_x is the size of a stratum.

Here:

- ATE is Average Treatment Effect (all units)
- ATT is Average Treatment Effect on the Treated

Estimands: ATE, ATT, ATC, S-, P-, C-

In addition, each of these can be targets of interest:

- for the **population**, in which case we refer to PATE, PATT, PATC and $\widehat{PATE}, \widehat{PATT}, \widehat{PATC}$
- for a **sample**, in which case we refer to SATE, SATT, SATC, and $\widehat{SATE}, \widehat{SATT}, \widehat{SATC}$

And for different subgroups,

- given some value on a covariate, in which case we refer to CATE (conditional average treatment effect)

Broader classes of estimands: LATE/CATE

The CATEs are **conditional** average treatment effects, for example the effect for men or for women. These are straightfoward.

However we might also imagine conditioning on unobservable or counterfactual features.

- The LATE (or CACE: complier average causal effect) asks about the effect of a treatment (X) on an outcome (Y) *for people that are responsive to an encouragement* (Z)

$$LATE = \frac{1}{|C|}\sum_{j \in C}(Y_j(X=1) - Y_j(X=0))$$

$$C := \{j : X_j(Z=1) > X_j(Z=0)\}$$

Quantile estimands

Other ways to condition on potential outcomes:

Model estimands

Many inquiries are averages of individual effects, even if the groups are not known, but they do not have to be:

- The RDD estimand is a statement about what effects *would be* at a threshold; it can be defined under a model even if no actual individuals are at the threshold. We imagine average potential outcomes as a function of treatment Z and running variable X, $f(z, x)$ and define:

$$\tau_{RDD} := f(1, x*) - f(1, x*)$$

Distribution estimands

Many inquiries are averages of individual effects, even if the groups are not known, but they do not have to be:

- Inquiries might relate to distributional quantities such as:
  - The effect of treatment on the variance in outcomes: $var(Y(1)) - var(Y(0))$
  - The variance of treatment effects: $var(Y(1) - Y(0))$
  - Other inequality measures (e.g. Ginis; (@imbens2015causal 20.3.2))

Spillover estimands

There are lots of interesting "spillover" estimands.

Imagine there are three individuals and each person's outcomes depends on the assignments of all others. For instance $Y_1(Z_1, Z_2, Z_3$, or more generally, $Y_i(Z_i, Z_{i+1 \pmod 3}, Z_{i+2 \pmod 3})$.

Then three estimands might be:

- $\frac{1}{3}\left(\sum_i Y_i(1,0,0) - Y_i(0,0,0)\right)$
- $\frac{1}{3}\left(\sum_i Y_i(1,1,1) - Y_i(0,0,0)\right)$
- $\frac{1}{3}\left(\sum_i Y_i(0,1,1) - Y_i(0,0,0)\right)$

Interpret these. What others might be of interest?

Differnces in CATEs and interaction estimands

A difference in CATEs is a well defined estimand that might involve interventions on one node only:

- $\mathbb{E}_{\{W=1\}}[Y(X=1)-Y(X=0)]-\mathbb{E}_{\{W=0\}}[Y(X=1)-Y(X=0)]$

It captures differences in effects.

An *interaction* is an effect on an effect:

- $\mathbb{E}[Y(X=1,W=1)-Y(X=0,W=1)]-\mathbb{E}[Y(X=1,W=0)-Y(X=0,W=0)]$

Note in the latter the expectation is taken over the whole population.

Mediation estimands and complex counterfactuals

Say X can affect Y directly, or indirectly through M. then we can write potential outcomes as:

- $Y(X = x, M = m)$
- $M(X = x)$

We can then imagine inquiries of the form:

- $Y(X = 1, M = M(X = 1)) - Y(X = 0, M = M(X = 0))$
- $Y(X = 1, M = 1) - Y(X = 0, M = 1)$
- $Y(X = 1, M = M(X = 1)) - Y(X = 1, M = M(X = 0))$

Interpret these. What others might be of interest?

Mediation estimands and complex counterfactuals

Again we might imagine that these are defined with respect to some group:

- $A = \{i|Y_i(X = 1, M = M(X = 1)) > Y_i(X = 0, M = M(X = 0))\}$
- $\frac{1}{|A|}\sum_{i \in A}(Y(X = 1, M = 1) > Y(X = 0, M = 1))$

here, among those for whom X has a positive effect on Y, for what share would there be a positive effect if M were fixed at 1.

Causes of effects and effects of causes

In qualitative research a particularly common inquiry is “did $X = 1$ cause $Y = 1$?

This is often given as a probability, the “probability of causation” (though at the case level we might better think of this probabiliy as an estimate rather than an estimand):

$$\Pr(Y_i(0) = 0 | Y_i(1) = 1, X = 1)$$

Causes of effects and effects of causes

Intuition: What's the probability $X = 1$ caused $Y = 1$ in an $X = 1, Y = 1$ case drawn from a large population with the following experimental distribution:

	Y=0	Y=1	All
X=0	1	0	1
X=1	0.25	0.75	1

Causes of effects and effects of causes

Intuition: What's the probability $X = 1$ caused $Y = 1$ in an $X = 1, Y = 1$ case drawn from a large population with the following experimental distribution:

	Y=0	Y=1	All
X=0	0.75	0.25	1
X=1	0.25	0.75	1

Subsection 2

Actual causation

Actual causation

Other inquiries focus on distinguishing between causes.

For the Billy Suzy problem [@hall2004two], @halpern2016actual focuses on "actual causation" as a way to distinguish between Suzy and Billy:

> *Imagine Suzy and Billy, simultaneously throwing stones at a bottle. Both are excellent shots and hit whatever they aim at. Suzy's stone hits first, knocks over the bottle, and the bottle breaks. However, Billy's stone would have hit had Suzy's not hit, and again the bottle would have broken. Did Suzy's throw cause the bottle to break? Did Billy's?*

Subsection 3

Actual causation

Actual causation

Actual Causation:

1. $X = x$ and $Y = y$ both happened;
2. there is some set of variables, $\mathcal{W}$, such that if they were fixed at the levels that they *actually took* on in the case, and if X were to be changed, then Y would change (where $\mathcal{W}$ can also be an empty set);
3. no strict subset of X satisfies 1 and 2 (there is no redundant part of the condition, $X = x$).

Subsection 4

Actual causation

Actual causation

- Suzy: Condition 2 is met if Suzy's throw made a difference, counterfactually speaking—with the important caveat that, in determining this, we are permitted to condition on Billy' stone not hitting the bottle.
- Billy: Condition 2 is not met.

An inquiry: for what share in a population is a possible cause an actual cause?

Subsection 5

Pearl's ladder

Subsection 6

Inquiries as statements about principal strata

Subsection 7

Identification

Identification

What it is. When you have it. What it's worth.

Identification

Informally a quantity is "identified" if it can be "recovered" once you have enough data.

Say for example average wage is x in some very large population. If I gather lots and lots of data on the wages of individuals and take the average then then my estimate will ultimately let be figure out x. If x is 1 then by estimate will end up centered on \$1. If it is \$2 it will end up centered on \$2.

Essentially: Each underlying value produces a unique data distribution. When you see that distribution you recover the parameter.

Identification (Example without identification)

Informally a quantity is "identified" if it can be "recovered" once you have enough data.

- Say for example average wage is x^m for men and x^w for women (in some very large population).
- If I gather lots and lots of data on the wages of (male and female) couples, e.g. $x_i^c = x_i^m + x_i^w$ then, although this will be informative, it will never be sufficient to recover x^m for men and x^w.
- I can recover x^c, but there are too many combinations of possible values of x^m and x^w consistent with the observed data.

Identification : Goal

Our goal in causal inference is to estimate quantities such as:

$$\Pr(Y|\hat{x})$$

where $\hat{x}$ is interpreted as X set to x by "external" control. Equivalently: $do(X = x)$ or sometimes $X \leftarrow x$.

If this quantity is **identifiable** then we can recover it with infinite data.

If it is not identifiable, then, even in the best case, we are not guaranteed to get the right answer.

Are there general rules for determining whether this quantitiy can be identified? Yes.

Identification : Goal

Note first, identifying

$$\Pr(Y|x)$$

is easy.

But we are not interested in identifying the distribution of Y given observed values of x, but rather, the distribution of Y if X is *set* to x.

Subsection 8

Levels and effects

Levels and effects

If we can identify the controlled disteribution we can calculate other causal quantities of interest.

For example for a binary X, Y the causal effect of X on the probability that $Y = 1$ is:

$$\Pr(Y = 1|\hat{x} = 1) - \Pr(Y = 1|\hat{x} = 0)$$

Again, **this is not the same as**:

$$\Pr(Y = 1|x = 1) - \Pr(Y = 1|x = 0)$$

It's the difference between seeing and doing.

When to condition? What to condition on?

The key idea is that you want to find a set of variables such that when you condition on these you get what you would get if you used a do operation.

Intuition:

- You could imagine creating a “mutilated” graph by removing all the arrows leading *out* of X
- Then select a set of variables, Z, such that X and Y are d-separated by Z on the the mutilated graph
- When you condition on these you are making sure that any covariation between X and Y is covariation that is due to the effects of X

Illustration

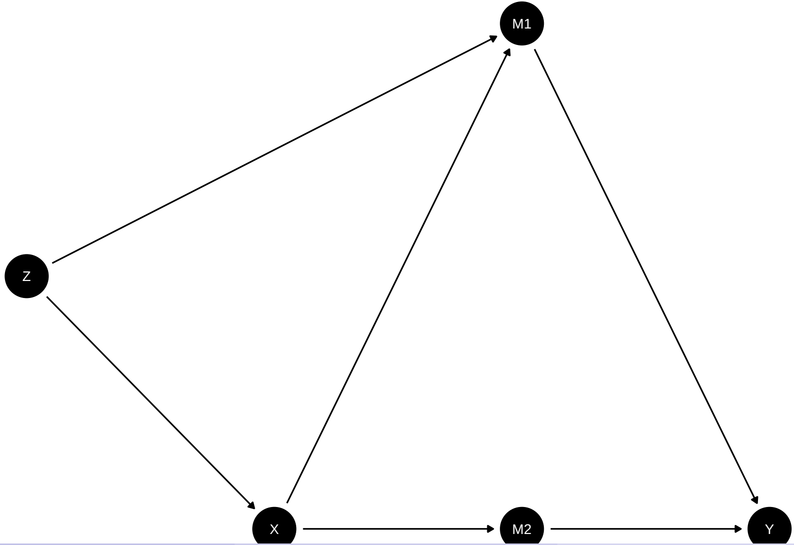


Illustration: Remove paths out

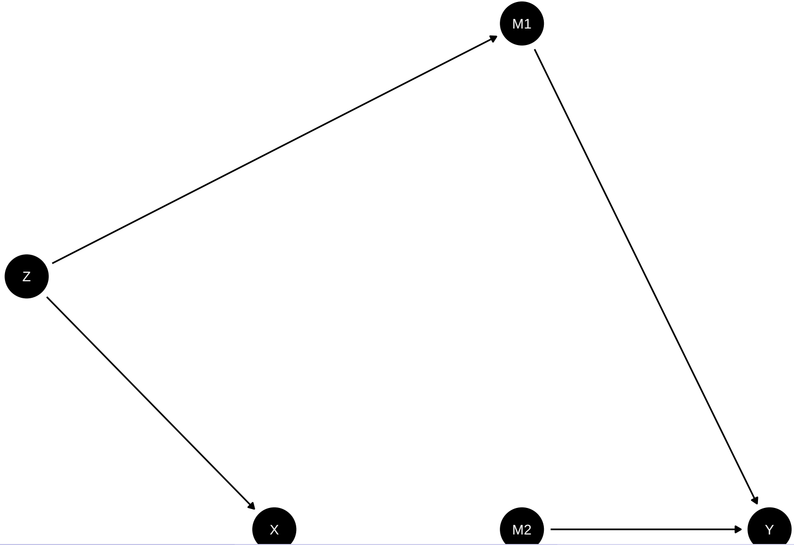


Illustration: Block backdoor path

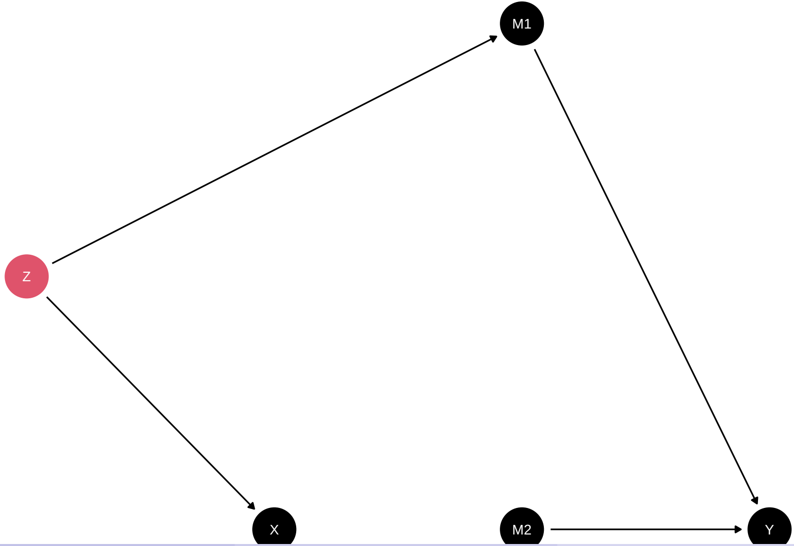
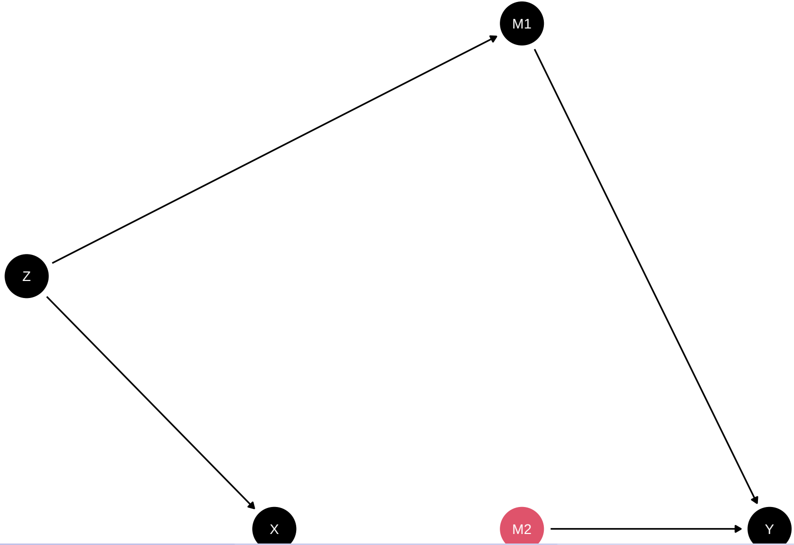


Illustration: Why not like this?



Identification

- Three results (“Graphical Identification Criteria”)
  - Backdoor criterion
  - Adjustment criterion
  - Frontdoor criterion
- There are more

Backdoor Criterion: (Pearl 1995)

The **backdoor criterion** is satisfied by Z (relative to X, Y) if:

1. No node in Z is a descendant of X
2. Z blocks every **backdoor** path from X to Y (i.e. every path that contains an arrow into X)

In that case you can identify the effect of X on Y by conditioning on Z:

$$P(Y = y|\hat{x}) = \sum_z P(Y = y|X = x, Z = z)P(z)$$

(This is eqn 3.19 in Pearl (2000))

Backdoor Criterion: (Pearl 1995)

$$P(Y = y|\hat{x}) = \sum_z P(Y = y|X = x, Z = z)P(z)$$

- Note notion of a linear control of anything like that; idea really is like blocking: think lots of discrete data and no missing patterns
- Note this is a formula for a (possibly counterfactual) *level*; a counterfactual difference would be given in the obvious way by:

$$P(Y = y|\hat{x}) - P(Y = y|\hat{x}') = \sum_z P(Y = y|X = x, Z = z)P(z) - \sum_z P(Y =$$

Backdoor Proof

Following Pearl (2009), Chapter 11. Let T denote the set of parents of X: $T := pa(X)$, with (possibly vector valued) realizations t.

If the backdoor criterion is satisfied, we have:

1. Y is independent of T, given X and observed data, Z (since Z blocks backdoor paths)
2. X is independent of Z given T. (Since Z includes only nondescendents)

- From the DAG we have:

$$p(y|\hat{x}) = \sum_{t \in T} p(t)p(y|\hat{x}, t)$$

Backdoor Proof

- But we do not observe T, rather we observe Z. OK, but we can write:

$$p(y|\hat{x}) = \sum_{t \in T} p(t) \sum_{z} p(y|\hat{x}, t, z) p(z|\hat{x}, pa(X))$$

- Then using the two conditions above:
  1. replace $p(y|\hat{x}, pa(X), z)$ with $p(y, \hat{x}, z)$
  2. replace $p(z|\hat{x}, pa(X))$ with $p(z|\hat{x})$

This gives:

$$p(y|\hat{x}) = \sum_{pa(X)} p(pa(X)) \sum_{z} p(y|\hat{x}, z) p(z|pa(X))$$

Now Clean up:

$$p(y|\hat{x}) = \sum_{pa(X)} p(pa(X)) \sum_{z} p(y|\hat{x}, z) p(z|pa(X))$$

$$\leftrightarrow$$

$$p(y|\hat{x}) = \sum_{z} p(y|\hat{x}, z) \sum_{pa(X)} p(pa(X)) p(z|pa(X)) = \sum_{z} p(y|\hat{x}) p(z)$$

Adjustment criterion

See @shpitser2012validity

The adjustment criterion is satisfied by Z (relative to X, Y) if:

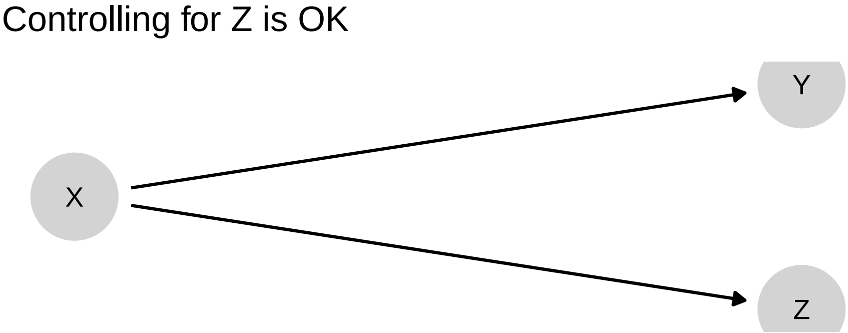
1. no element of Z is a descendant (in the mutilated graph[1]) of any variable $W \notin X$ which lies on a proper causal path from X to Y[2]
2. Z blocks all **noncausal paths** from X to Y

[1] remove arrows pointing into X

[2] A *proper* causal pathway nodes in X to nodes in Y only intersects X at the endpoint

These are different. Simple illustration.

Here Z satisfies the adjustment criterion but not the backdoor criterion:



Z is descendant of X but it does not a descendant of a node on a path from X to Y. No harm adjusting for Z here, but not necessary either.

Frontdoor criterion

In code: Dagitty

There is a package for this

```
library(dagitty)
```

Then define a dag using dagitty syntax:

```
g <- dagitty("dag{X -> M -> Y ; Z -> X ; Z -> R -> Y}")
```

There is then a simple command to check whether two sets are d-separated by a third set:

```
dseparated(g, "X", "Y", "M")
```

```
[1] FALSE
```

```
dseparated(g, "X", "Y", c("Z","M"))
```

```
[1] TRUE
```

Dagitty: Find adjustment sets

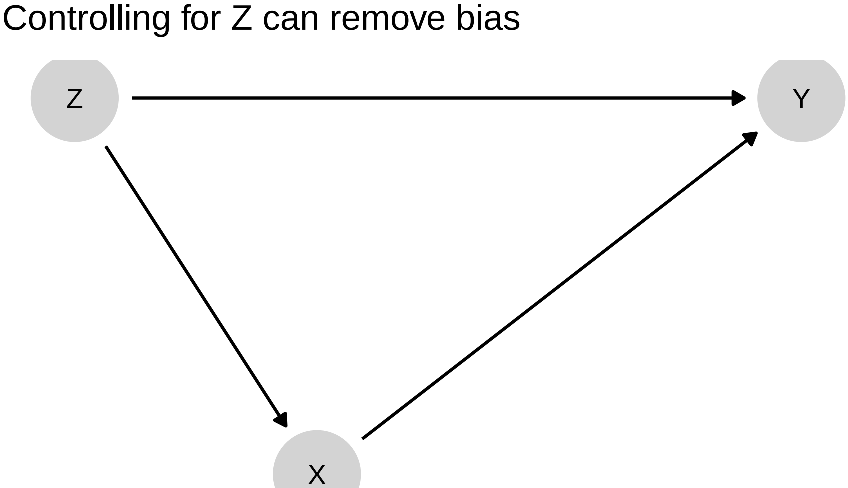
And a simple command to identify the adjustments needed to identify the effect of one variable on another:

```
adjustmentSets(g, exposure = "X", outcome = "Y")
```

```
{ R }
{ Z }
```

Important Examples : Confounding

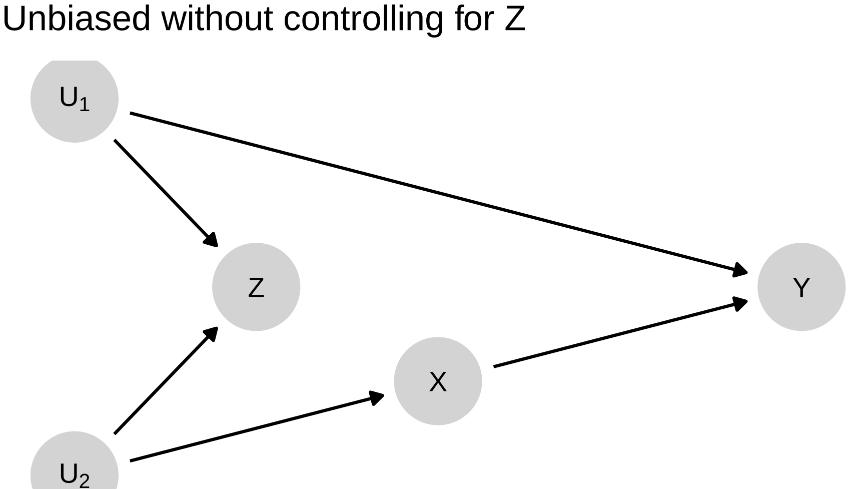
Example where Z is correlated with X and Y and is a confounder



Controlling for Z can remove bias

Confounding

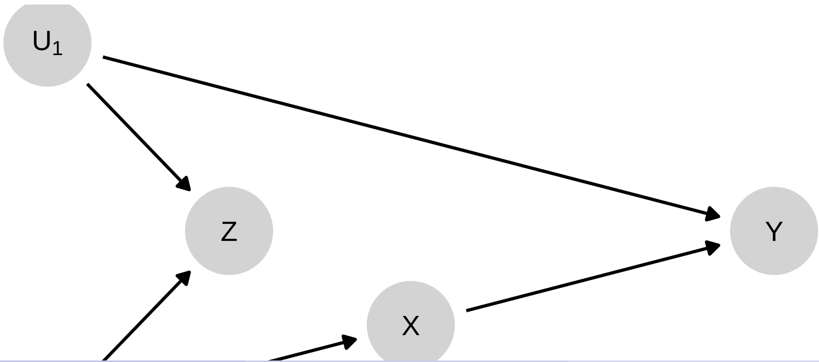
Example where Z is correlated with X and Y but it is *not* a confounder



Unbiased without controlling for Z

Important Examples : Collider

But controlling can also cause problems. In fact conditioning on a temporally pre-treatment variable could cause problems. Who'd have thunk? Here is an example from Pearl (2005):



Controlling for Z can induce bias

Illustration of identification failure from conditioning on a collider

```
U1 <- rnorm(10000);  U2 <- rnorm(10000)
Z  <- U1+U2
X  <- U2 + rnorm(10000)/2
Y  <- U1*2 + X

lm_robust(Y ~ X) |> tidy() |> kable(digits = 2)
```

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	-0.02	0.02	-0.85	0.39	-0.06	0.02
X	0.98	0.02	54.27	0.00	0.94	1.02

```
lm_robust(Y ~ X + Z) |> tidy() |> kable(digits = 2)
```

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	-0.01	0.01	-0.65	0.51	-0.02	0.01
X	-0.33	0.01	-35.01	0.00	-0.35	-0.31

Let's look at that in dagitty

```
g <- dagitty("dag{U1 -> Z  ; U1 -> y ; U2 -> Z ; U2 -> x  -> y
adjustmentSets(g, exposure = "x", outcome = "y")
```

```
 {}
```

```
isAdjustmentSet(g, "Z", exposure = "x", outcome = "y")
```

```
[1] FALSE
```

```
isAdjustmentSet(g, NULL, exposure = "x", outcome = "y")
```

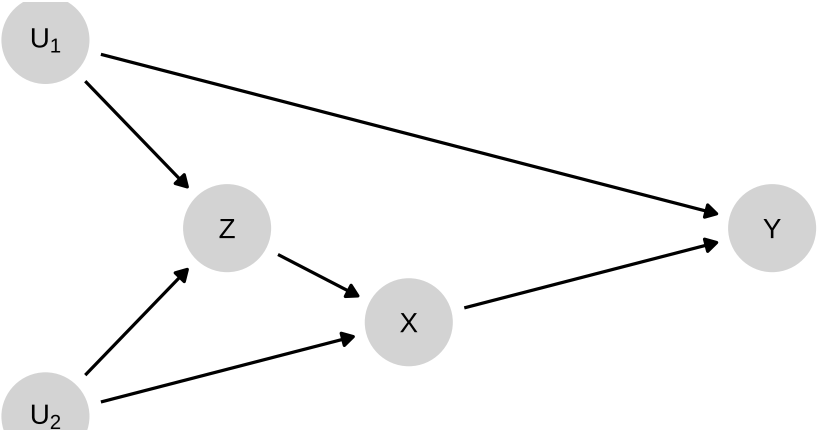
```
[1] TRUE
```

Which means, no need to condition on anything.

Collider & Confounder

A bind: from Pearl 1995.

Z is a confound but controlling for it can induce bias



Let's look at that in dagitty

```
g <- dagitty("dag{U1 -> Z  ; U1 -> y ;
             U2 -> Z ; U2 -> x  -> y;
             Z -> x}")
adjustmentSets(g, exposure = "x", outcome = "y")
```

```
{ U1 }
{ U2, Z }
```

which means you have to adjust on an unobservable. Here we double check that including or not including "Z" is enough:

```
isAdjustmentSet(g, "Z", exposure = "x", outcome = "y")
```

```
[1] FALSE
```

```
isAdjustmentSet(g, NULL, exposure = "x", outcome = "y")
```

```
[1] FALSE
```