

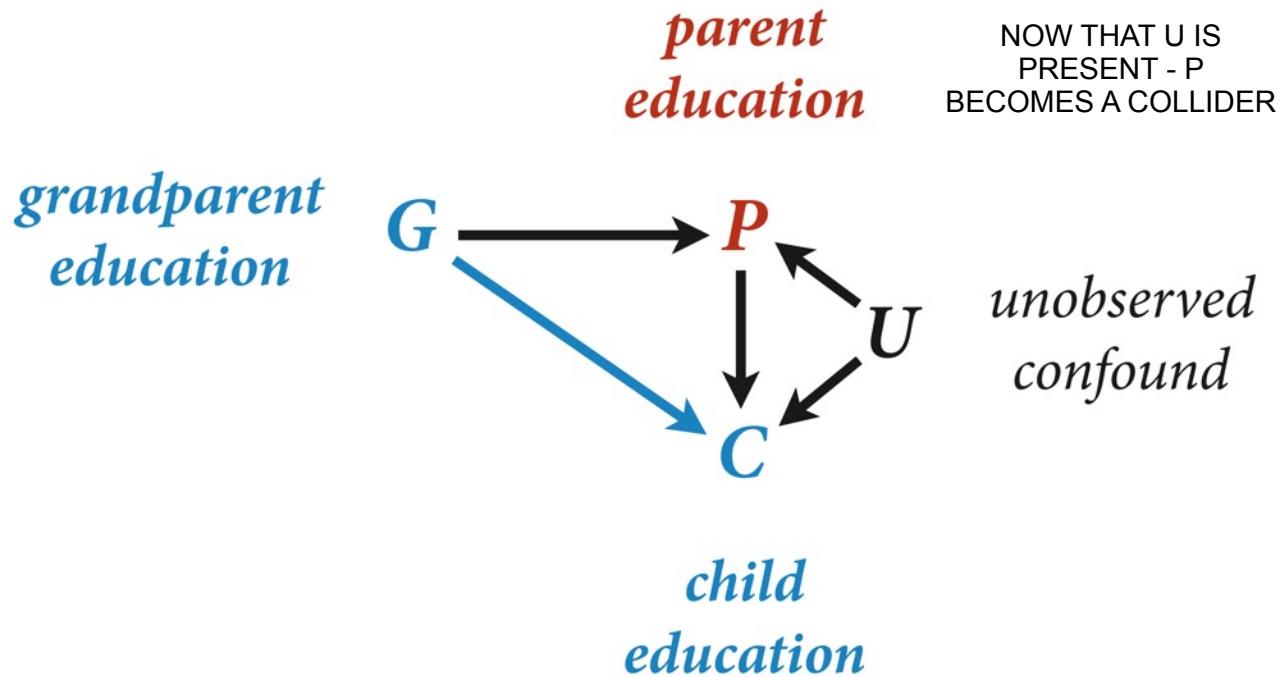
Methods 4 - 6

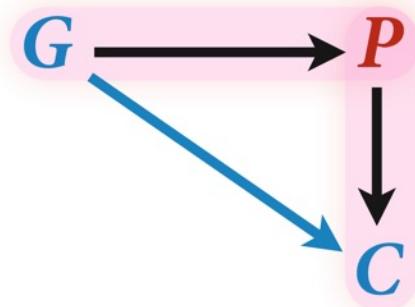
Chris Mathys



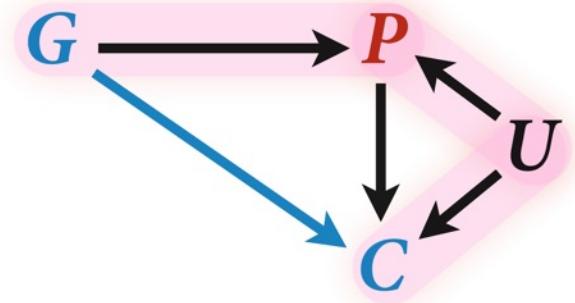
BSc Programme in Cognitive Science

Spring 2022



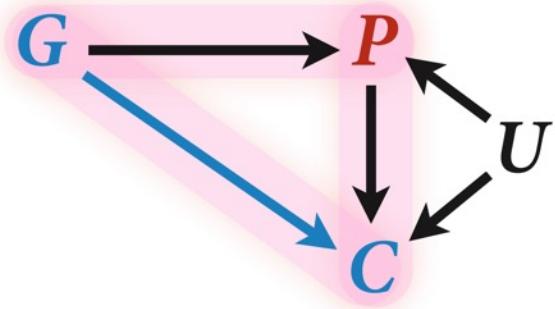


P is a mediator



P is a collider

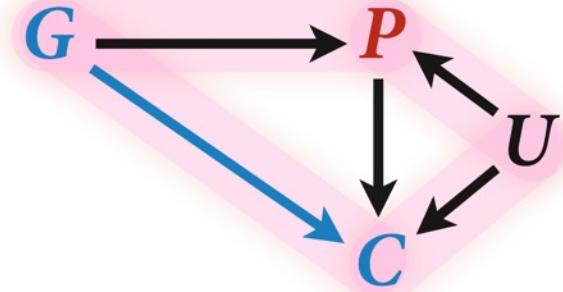
Can estimate **total**
effect of G on C



$$C_i \sim \text{Normal}(\mu_i, \sigma)$$

$$\mu_i = \alpha + \beta_G G_i$$

Cannot estimate
direct effect



$$C_i \sim \text{Normal}(\mu_i, \sigma)$$

$$\mu_i = \alpha + \beta_G G_i + \beta_P P_i$$

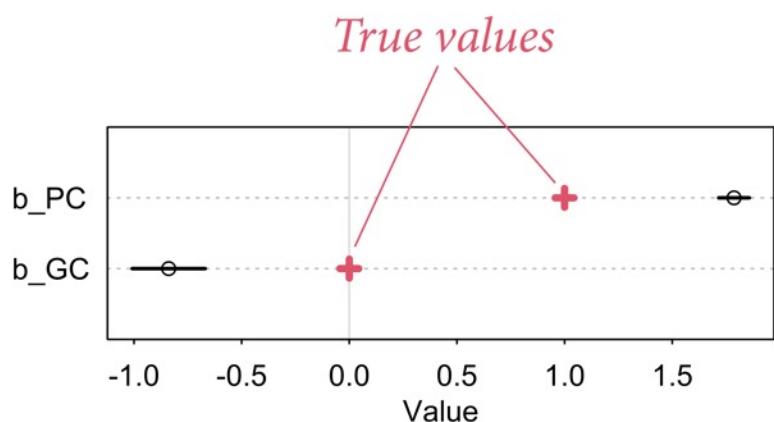
```

N <- 200 # num grandparent-parent-child triads
b_GP <- 1 # direct effect of G on P
b_GC <- 0 # direct effect of G on C
b_PC <- 1 # direct effect of P on C
b_U <- 2 #direct effect of U on P and C

set.seed(1)
U <- 2*rbern( N , 0.5 ) - 1
G <- rnorm( N )
P <- rnorm( N , b_GP*G + b_U*U )
C <- rnorm( N , b_PC*P + b_GC*G + b_U*U )
d <- data.frame( C=C , P=P , G=G , U=U )

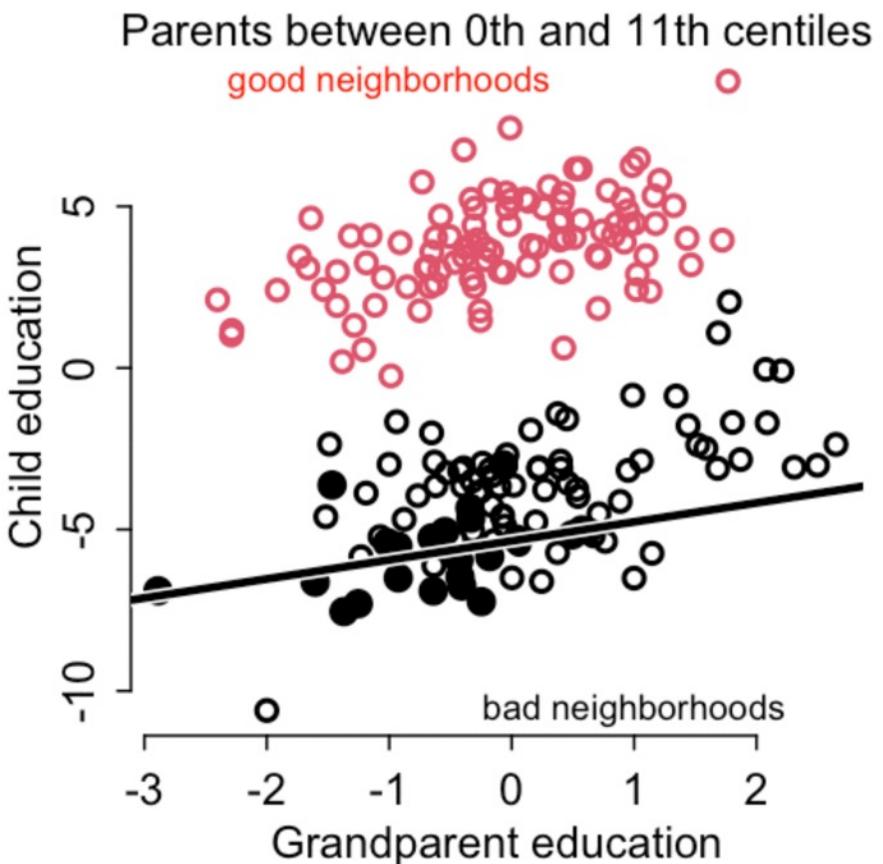
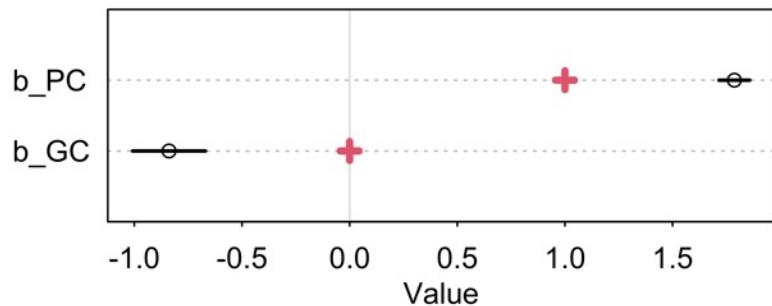
m6.11 <- quap(
  alist(
    C ~ dnorm( mu , sigma ),
    mu <- a + b_PC*P + b_GC*G,
    a ~ dnorm( 0 , 1 ),
    c(b_PC,b_GC) ~ dnorm( 0 , 1 ),
    sigma ~ dexp( 1 )
  ), data=d )

```



Stratify by parent centile
(collider)

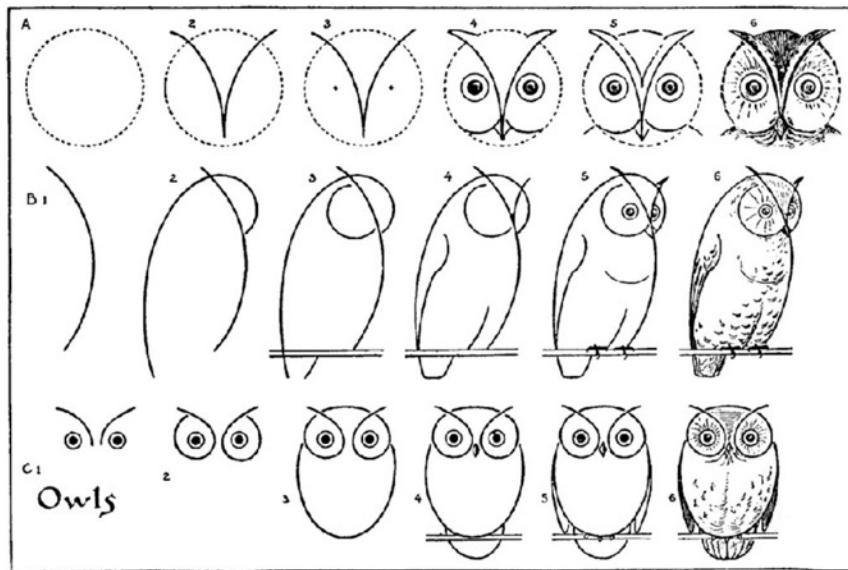
Two ways for parents to
attain their education: from
 G or from U



From Theory to Estimate

Our job is to

- (1) Clearly state assumptions
- (2) Deduce implications
- (3) Test implications



Avoid Being Clever At All Costs

Being clever is neither reliable nor transparent

Now what?

Given a causal model, can use logic to derive implications

Others can use same logic to verify/challenge your work



The Fork

$$X \leftarrow Z \rightarrow Y$$

X and Y associated
unless stratify by Z

The Pipe

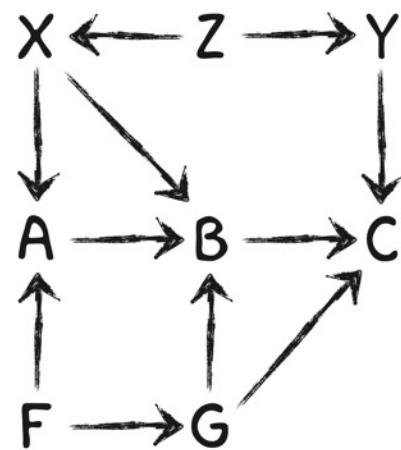
$$X \rightarrow Z \rightarrow Y$$

X and Y associated
unless stratify by Z

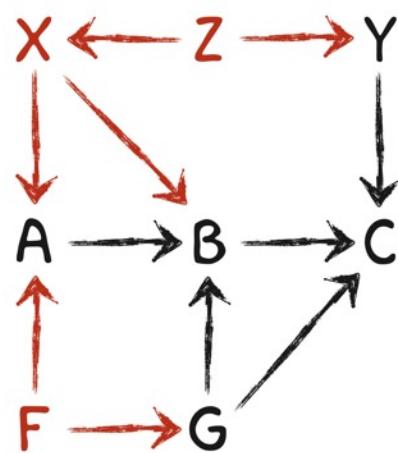
The Collider

$$X \rightarrow Z \leftarrow Y$$

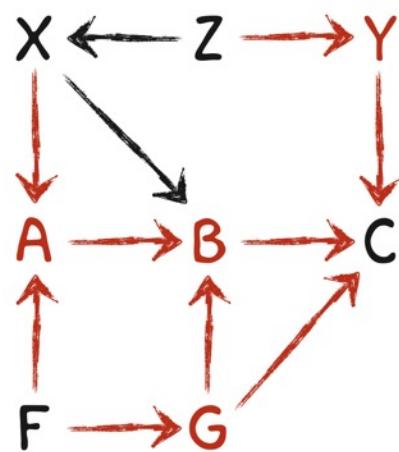
X and Y not associated
unless stratify by Z



Forks



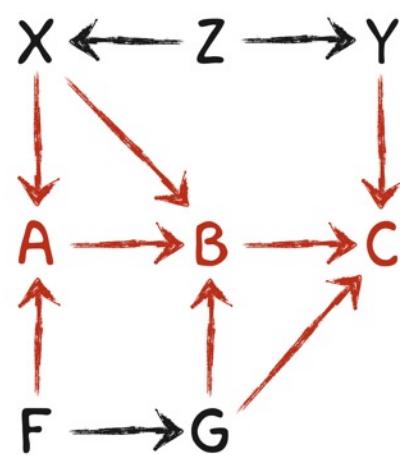
Forks
Pipes

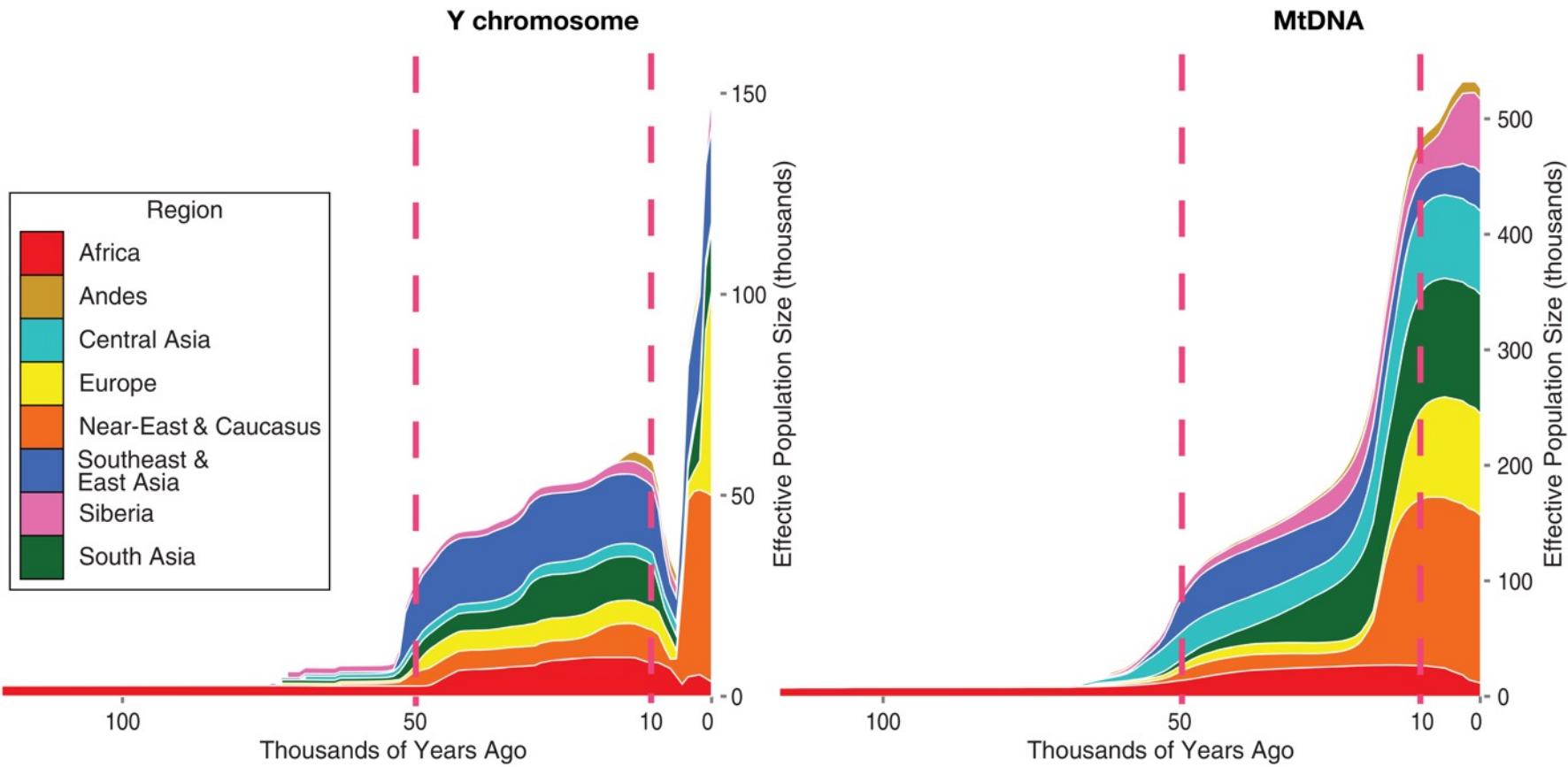


Forks

Pipes

Colliders





Karmin, M., (+100) et al., A recent bottleneck of Y chromosome diversity coincides with a global change in culture, *Genome research* 2015, DOI:10.1101/gr.186684.114

DAG Thinking

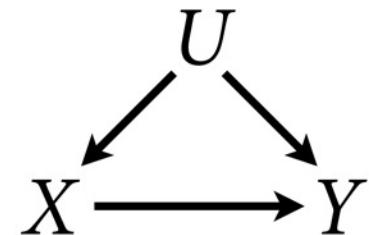
In an experiment, we cut causes of the treatment

We **randomize** (hopefully)

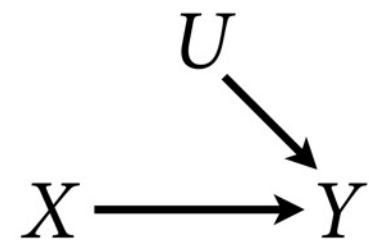
So how does causal inference without randomization ever work?

Is there a statistical procedure that mimics randomization?

Without randomization



With randomization



DAG Thinking

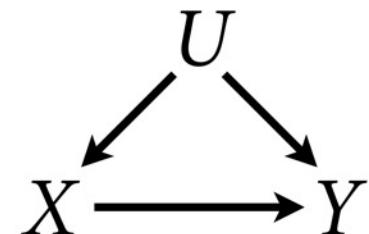
Is there a statistical procedure
that mimics randomization?

$$P(Y | \text{do}(X)) = P(Y | ?)$$

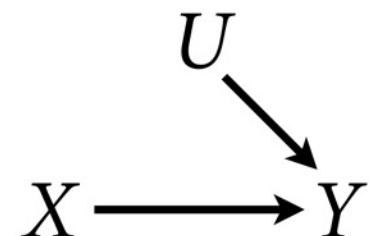
$\text{do}(X)$ means intervene on X

Can analyze causal model to
find answer (if it exists)

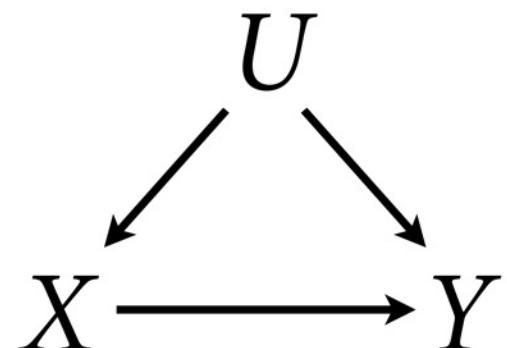
Without randomization



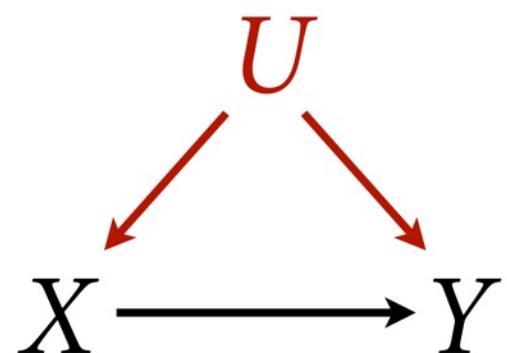
With randomization



Example: Simple Confound



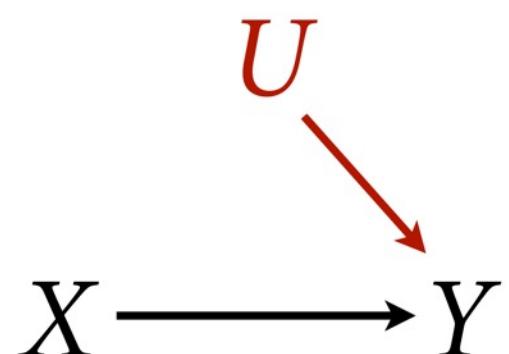
Example: Simple Confound



*Non-causal path
 $X <- U -> Y$*

*Close the fork!
Condition on U*

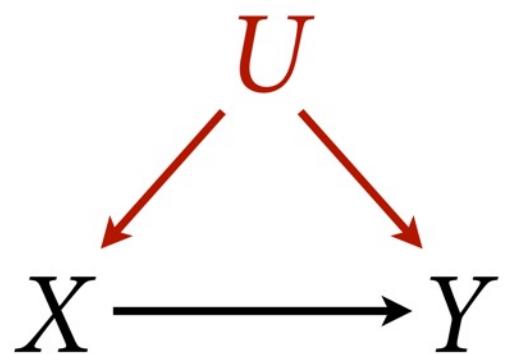
Example: Simple Confound



*Non-causal path
 $X <- U -> Y$*

*Close the fork!
Condition on U*

Example: Simple Confound



*Non-causal path
X <- U -> Y*

*Close the fork!
Condition on U*

$$P(Y | \text{do}(X)) = \sum_U P(Y | X, U)P(U) = \text{E}_U P(Y | X, U)$$

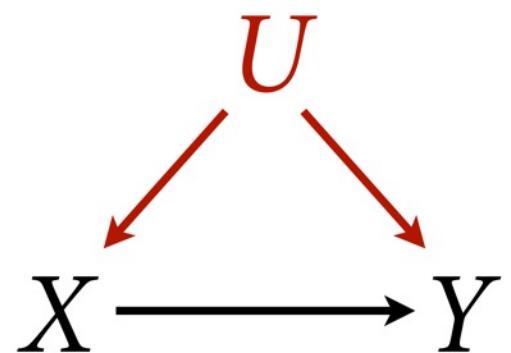
*“The distribution of Y, stratified by X and U,
averaged over the distribution of U.”*

$$P(Y | \text{do}(X)) = \sum_U P(Y | X, U)P(U) = \mathbb{E}_U P(Y | X, U)$$

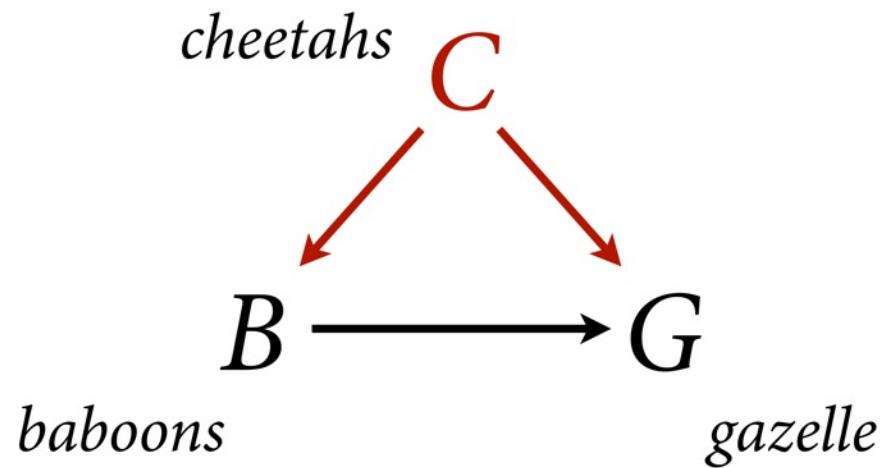
“The distribution of Y, stratified by X and U, averaged over the distribution of U.”

The causal effect of X on Y is **not** (in general) the **coefficient** relating X to Y

It is the distribution of Y when we change X , **averaged** over the distributions of the control variables (here U)

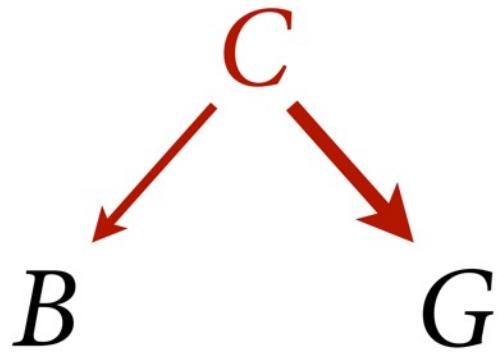


Marginal Effects Example

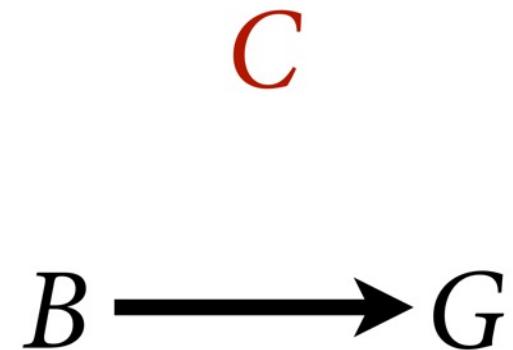


Marginal Effects Example

cheetahs present



cheetahs absent



Causal effect of baboons depends upon distribution of cheetahs

do-calculus

For DAGs, rules for finding
 $P(Y|do(X))$ known as **do-calculus**

do-calculus says what is possible
to say **before** picking functions

Additional assumptions yield
additional implications

236

THE BOOK OF WHY

DO-CALCULUS AT WORK

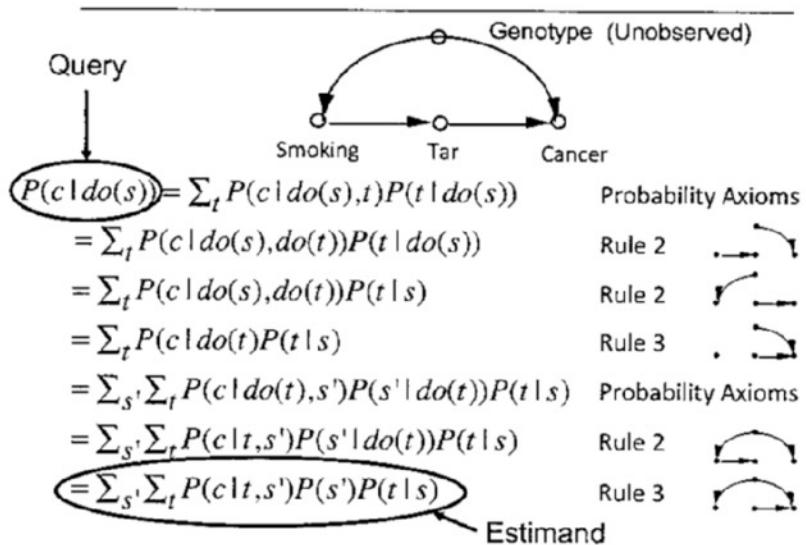


FIGURE 7.4. Derivation of the front-door adjustment formula from the rules of *do*-calculus.

do-calculus

do-calculus is **worst case**:
additional assumptions often
allow stronger inference

do-calculus is **best case**:
if inference possible by do-
calculus, does not depend on
special assumptions



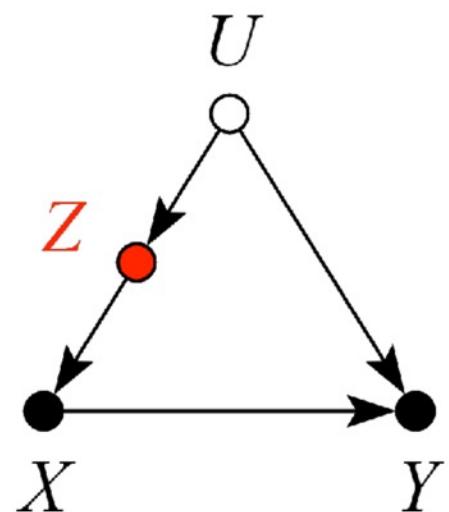
*Judea Pearl, father of
do-calculus, in 1966*

Backdoor Criterion

Very useful implication of do-calculus is the **Backdoor Criterion**

Backdoor Criterion is a shortcut to applying rules of do-calculus

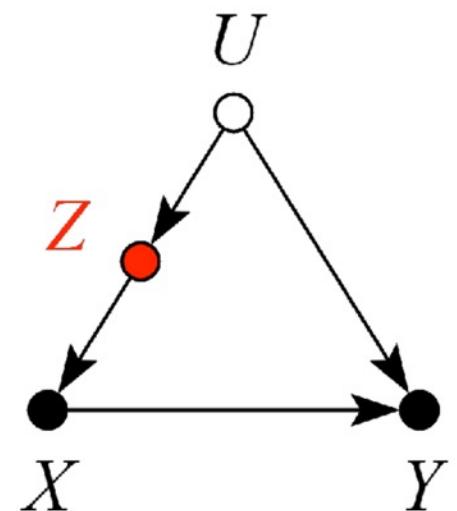
Also inspires **strategies** for research design that yield valid estimates



paper here = Cinelli et al. 2020: A crash course in good and bad control

Backdoor Criterion

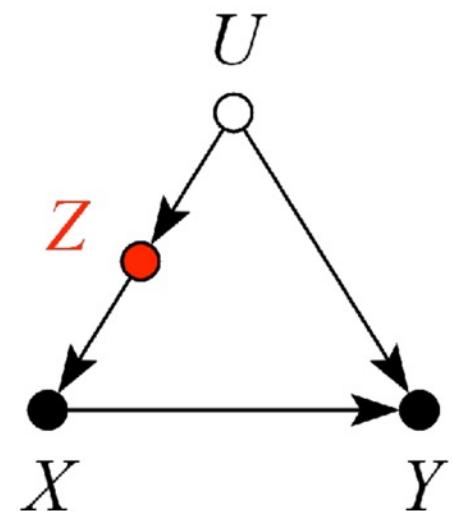
Backdoor Criterion: Rule to find a set of variables to stratify (condition) by to yield $P(Y|\text{do}(X))$



Backdoor Criterion

Backdoor Criterion: Rule to find a set of variables to stratify (condition) by to yield $P(Y|\text{do}(X))$

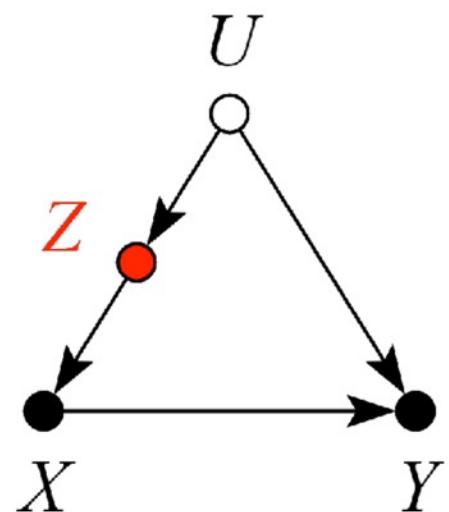
- (1) Identify all **paths** connection the treatment (X) to the outcome (Y)



Backdoor Criterion

Backdoor Criterion: Rule to find a set of variables to stratify (condition) by to yield $P(Y|\text{do}(X))$

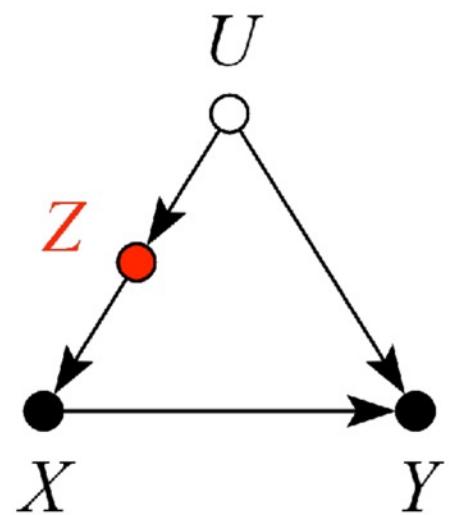
- (1) Identify all **paths** connection the treatment (X) to the outcome (Y)
- (2) Paths with arrows **entering** X are backdoor paths (non-causal paths)



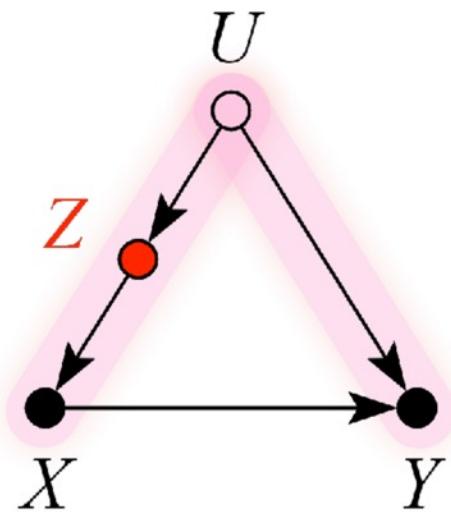
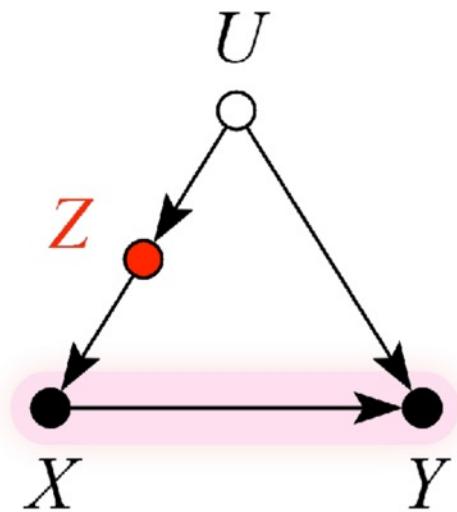
Backdoor Criterion

Backdoor Criterion: Rule to find a set of variables to stratify (condition) by to yield $P(Y|\text{do}(X))$

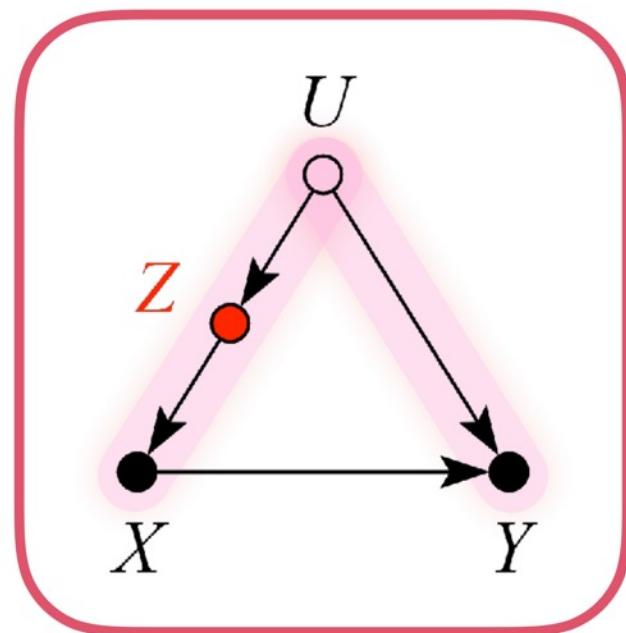
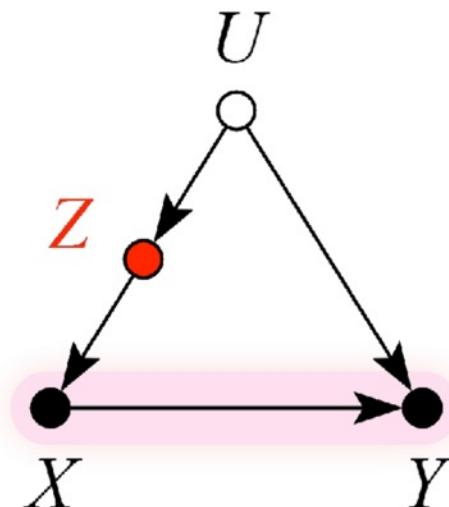
- (1) Identify all **paths** connection the treatment (X) to the outcome (Y)
- (2) Paths with arrows **entering** X are backdoor paths (non-causal paths)
- (3) Find **adjustment set** that closes/blocks all backdoor paths



(1) Identify all **paths** connection the treatment (X) to the outcome (Y)

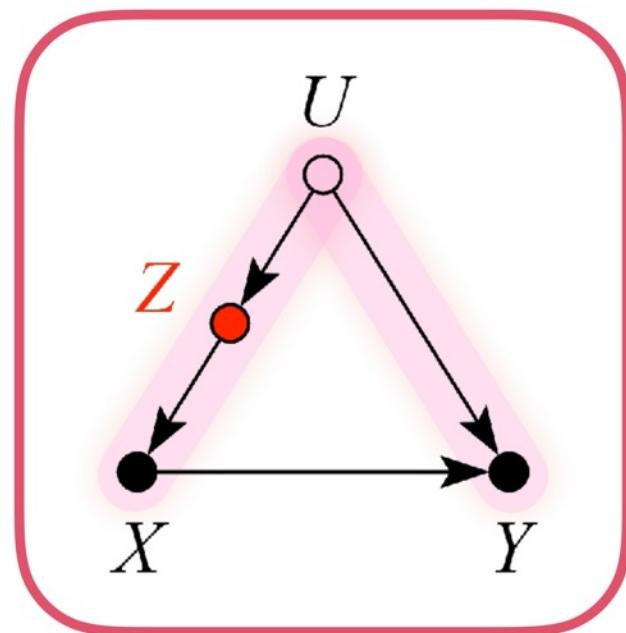


(2) Paths with arrows **entering** X are backdoor paths (non-causal paths)



(3) Find a set of control variables that close/block all backdoor paths

Block the pipe: $X \perp\!\!\!\perp U | Z$



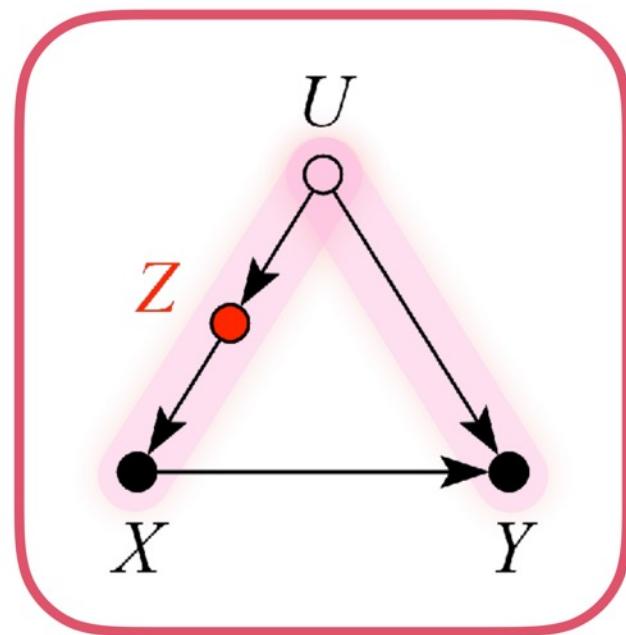
(3) Find a set of control variables that close/block all backdoor paths

Block the pipe: $X \perp\!\!\!\perp U | Z$

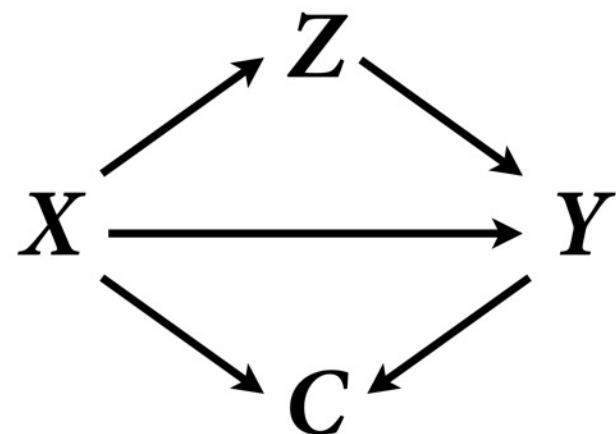
$$P(Y | \text{do}(X)) = \sum_U P(Y | X, Z)P(Z)$$

$$Y_i \sim \text{Normal}(\mu_i, \sigma)$$

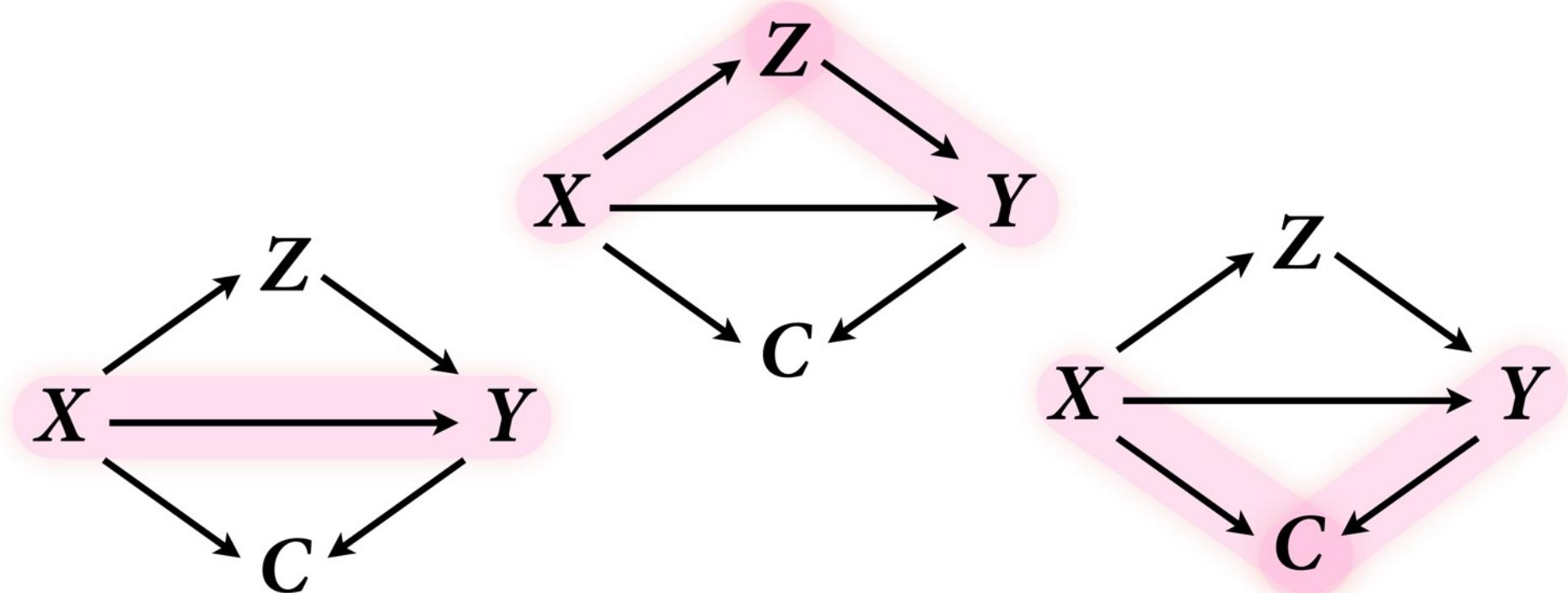
$$\mu_i = \alpha + \beta_X X_i + \beta_Z Z_i$$

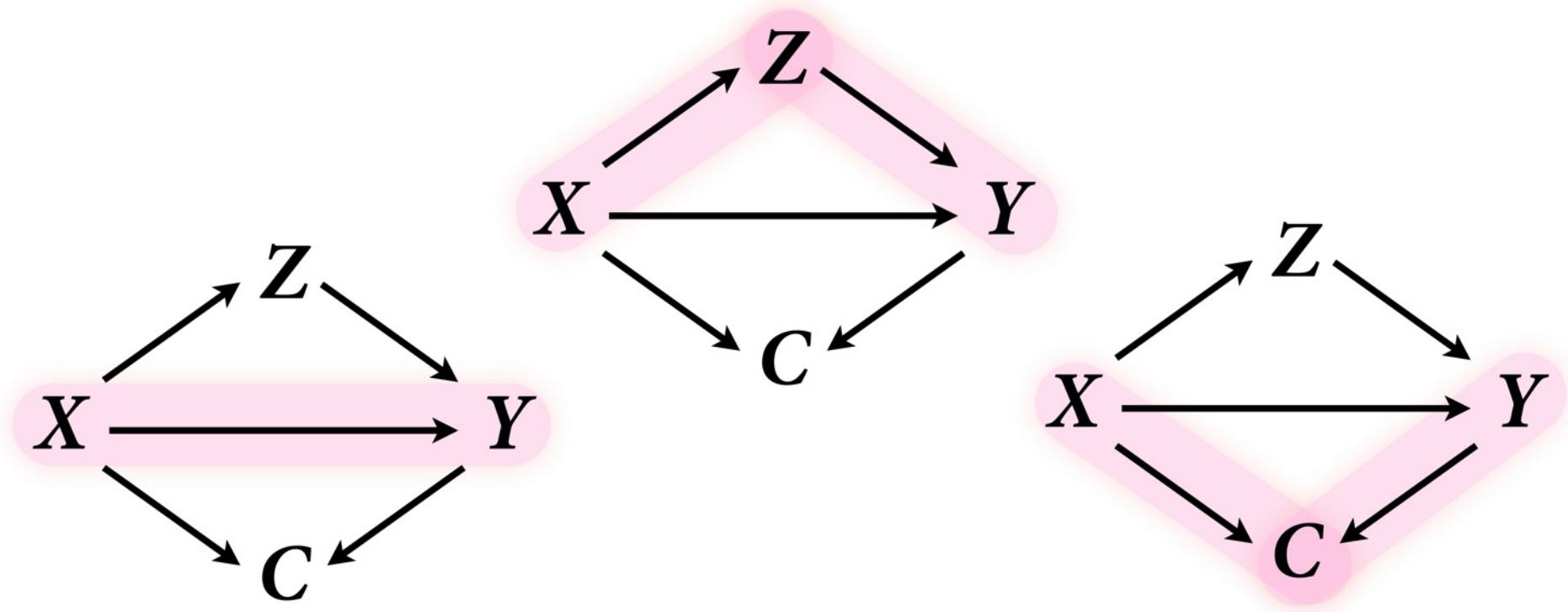


List all the paths connecting **X** and **Y**.
Which need to be closed to estimate
effect of **X** on **Y**?



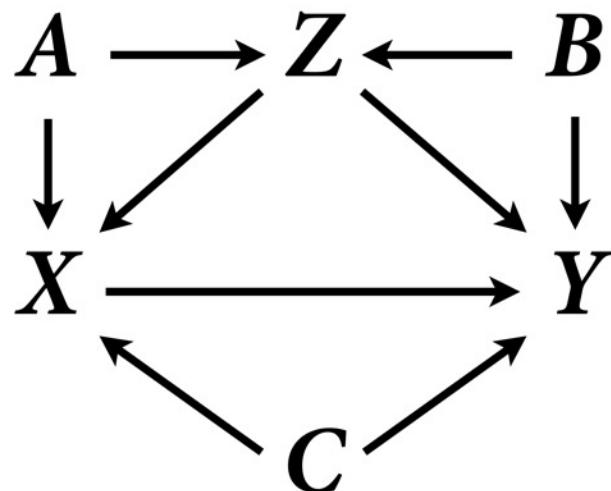
List all the paths connecting **X** and **Y**.
Which need to be closed to estimate
effect of **X** on **Y**?



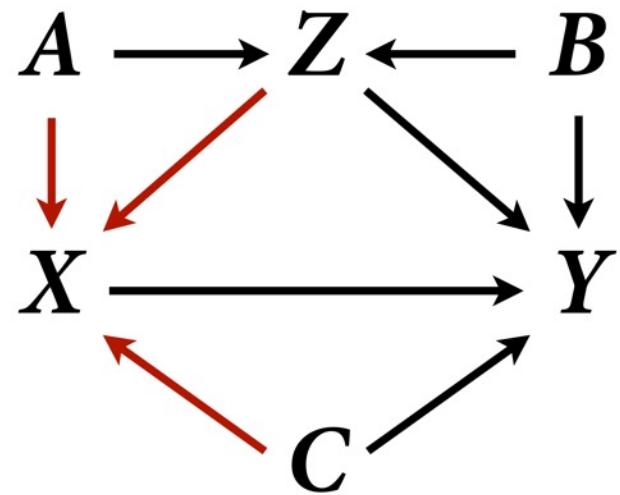
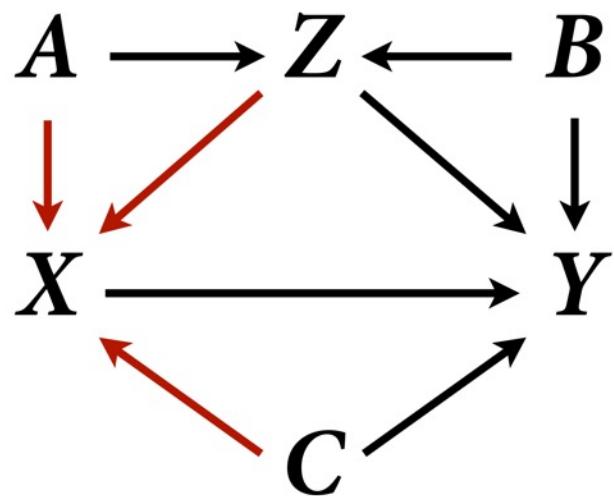


Adjustment set: *nothing!*

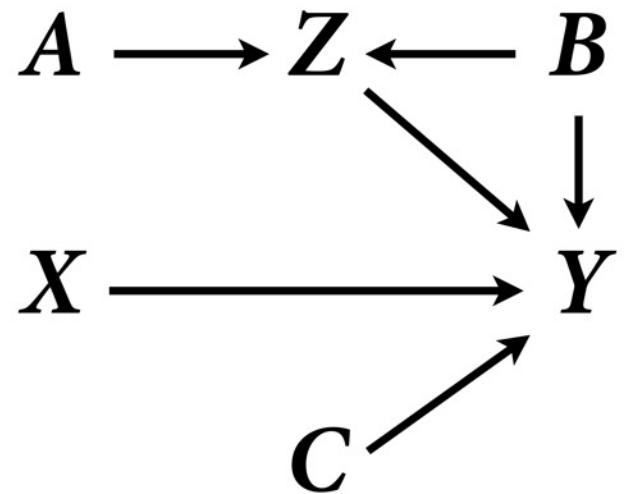
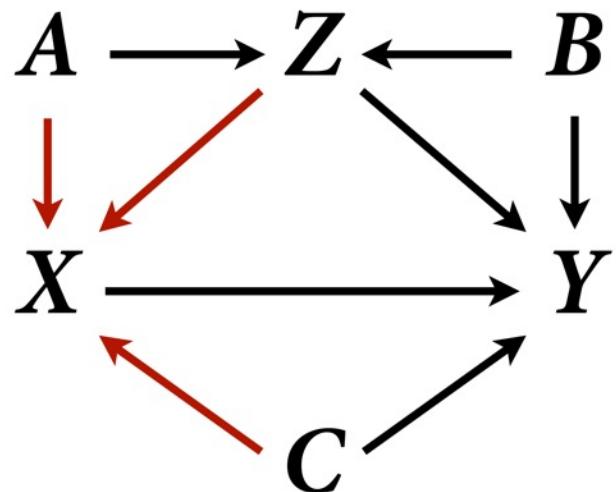
List all the paths connecting **X** and **Y**.
Which need to be closed to estimate
effect of **X** on **Y**?

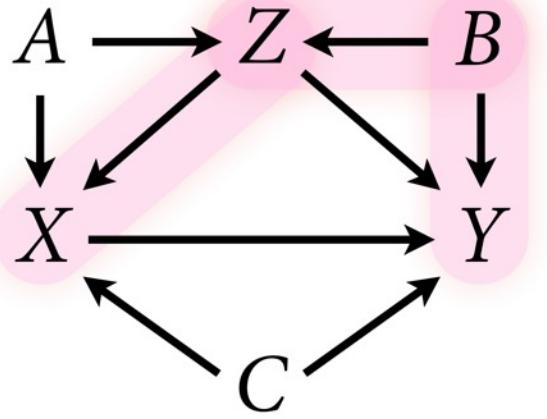
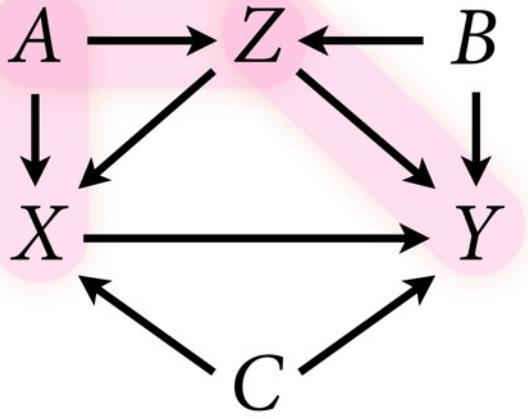
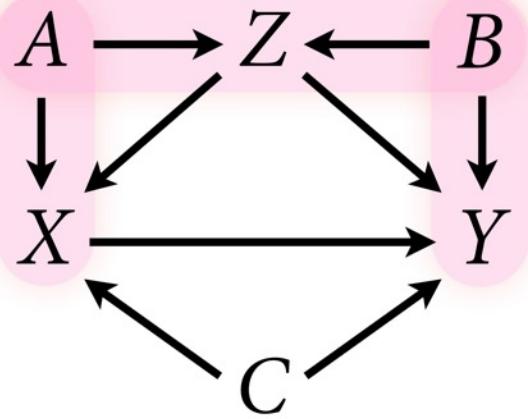
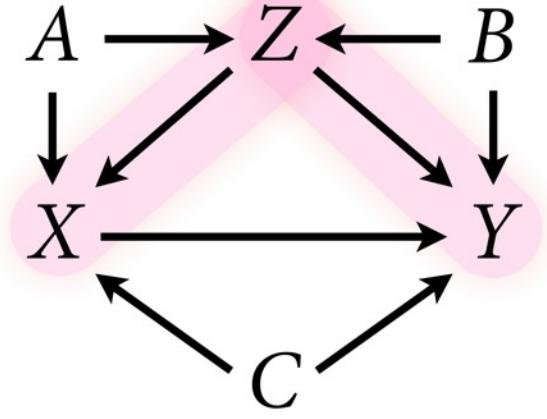
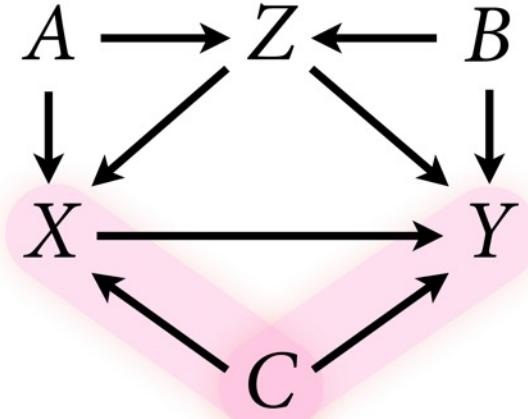
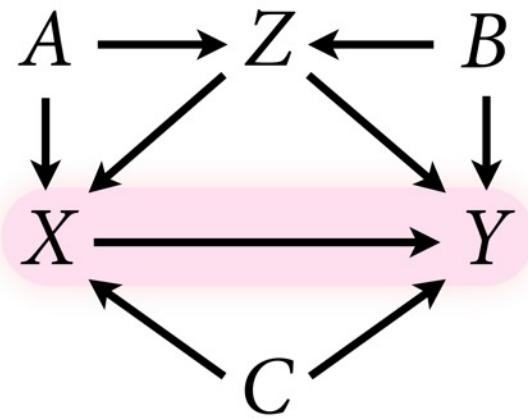


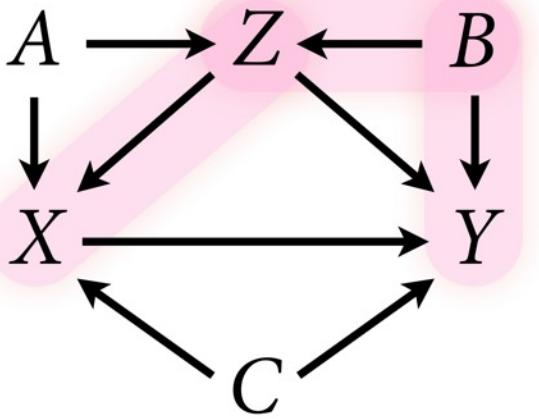
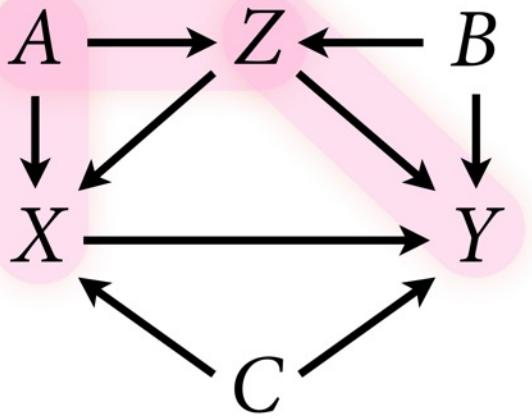
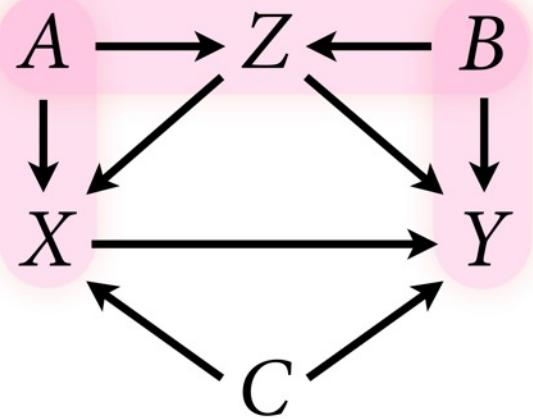
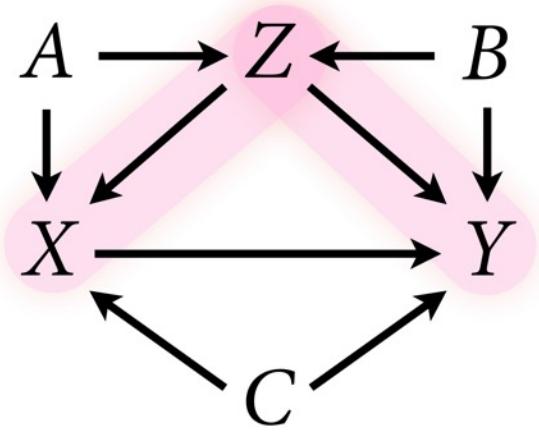
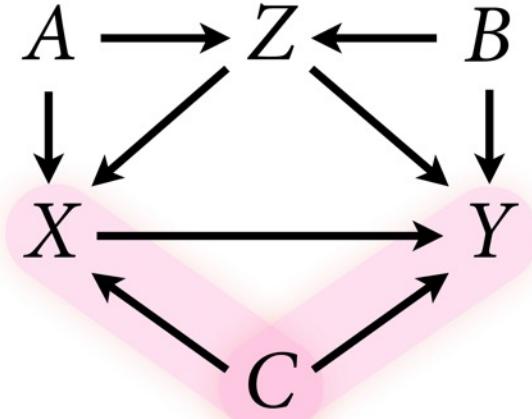
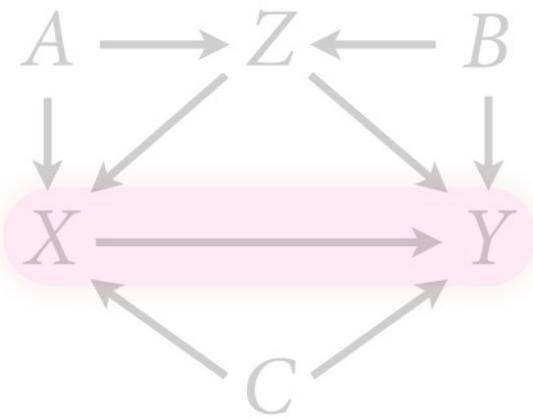
$$P(Y|\text{do}(X))$$

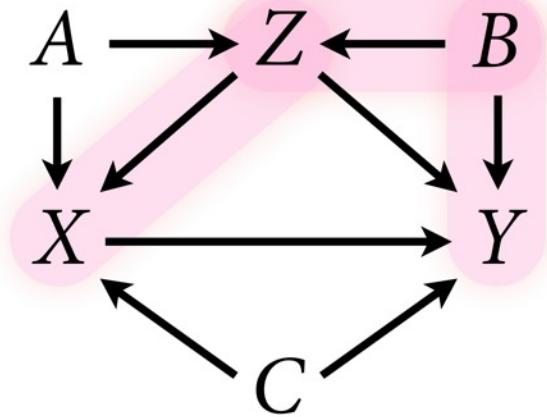
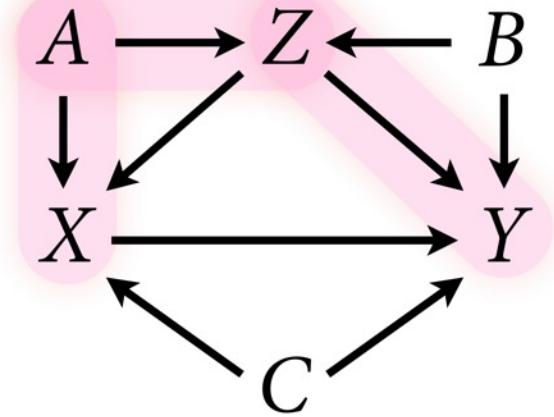
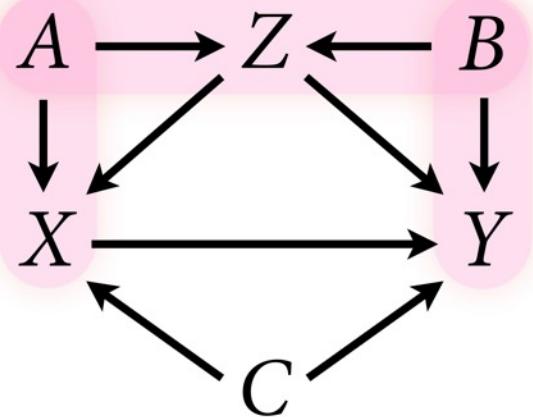
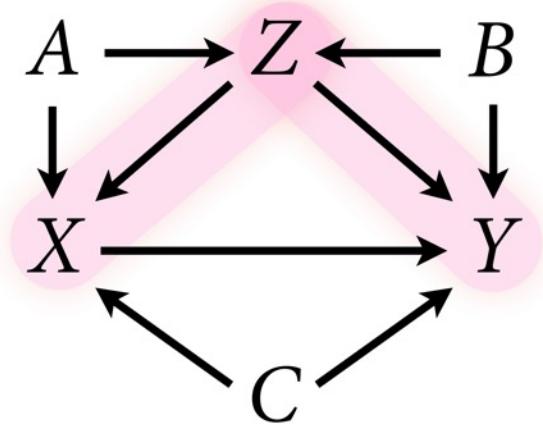
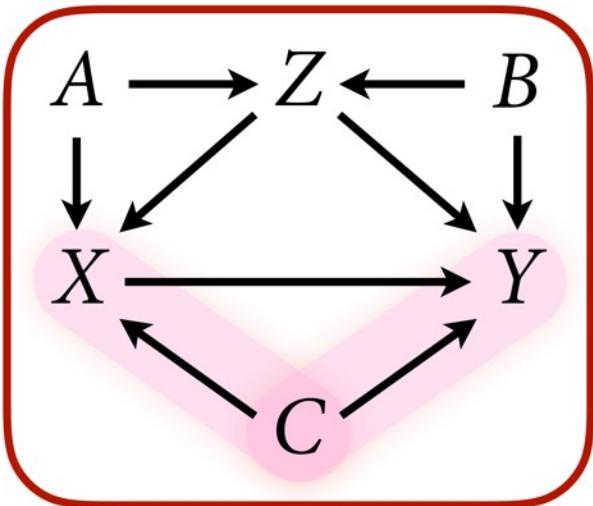
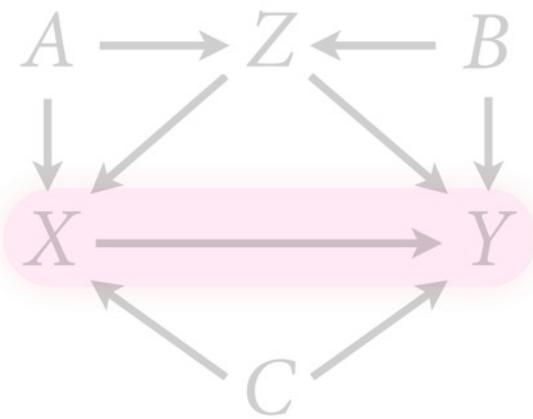


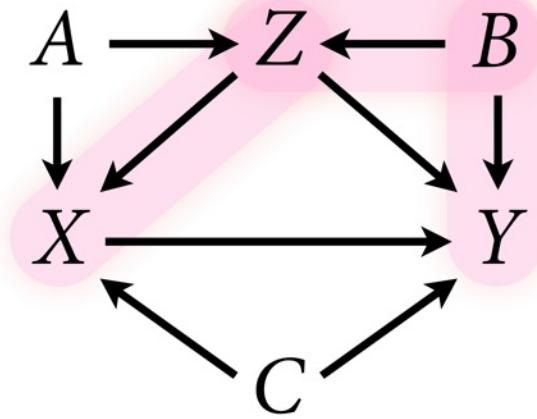
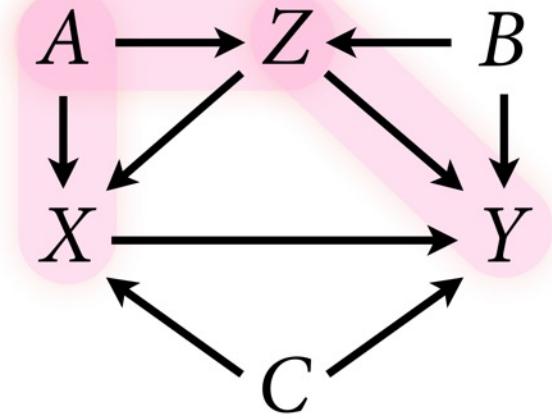
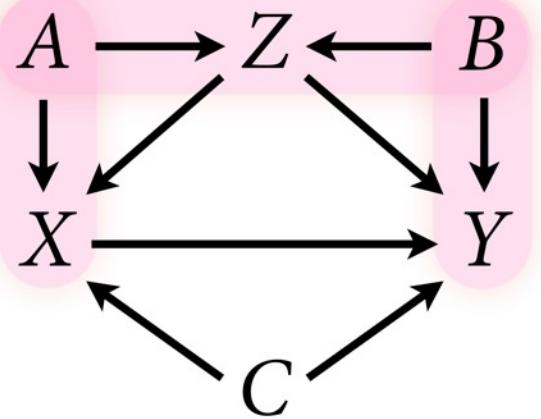
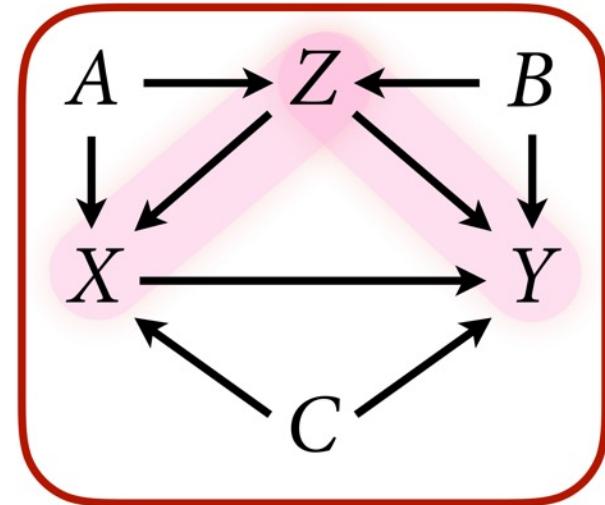
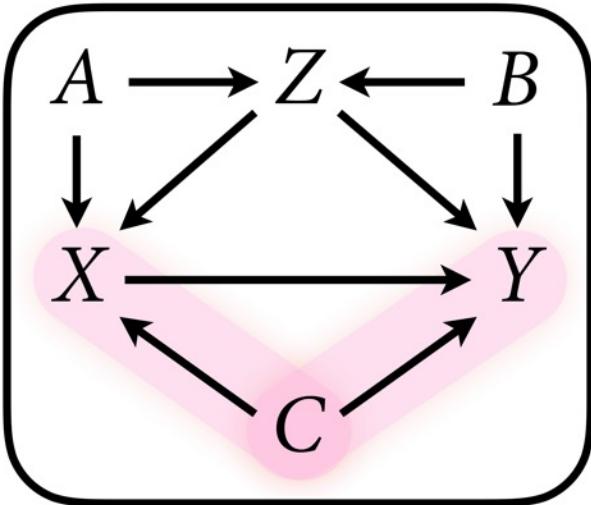
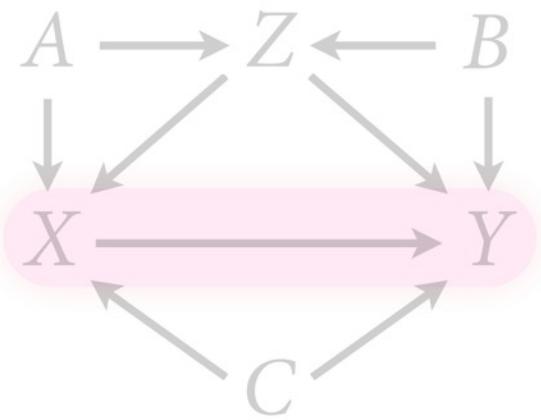
$$P(Y|\text{do}(X))$$

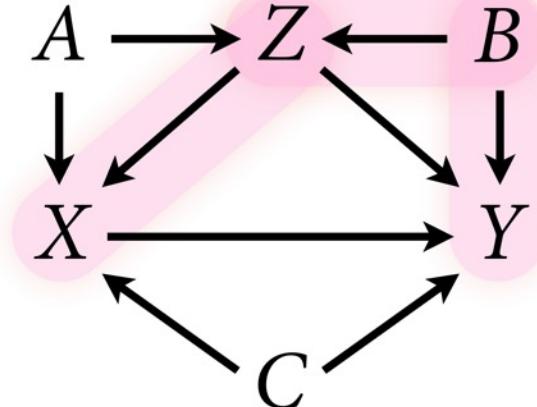
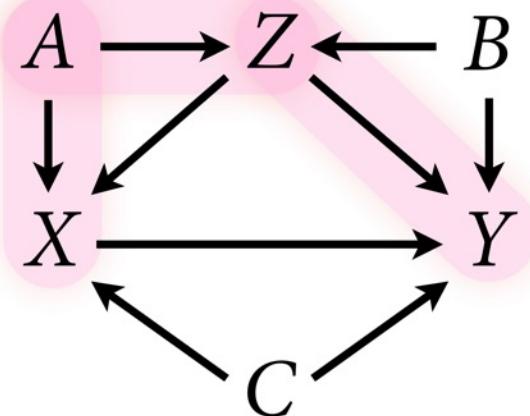
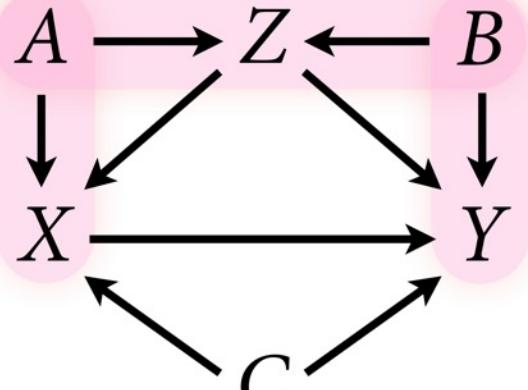
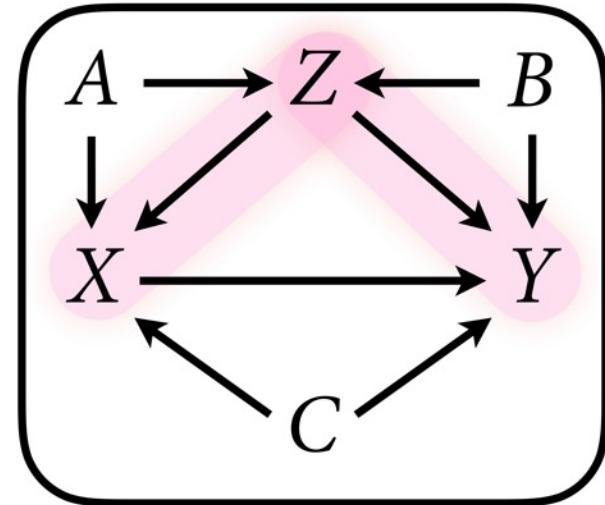
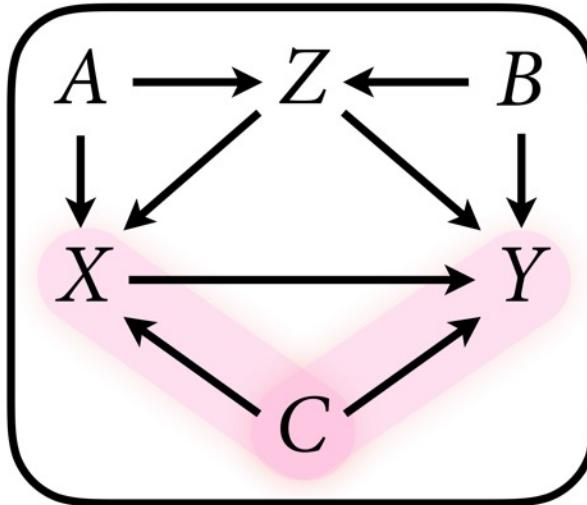
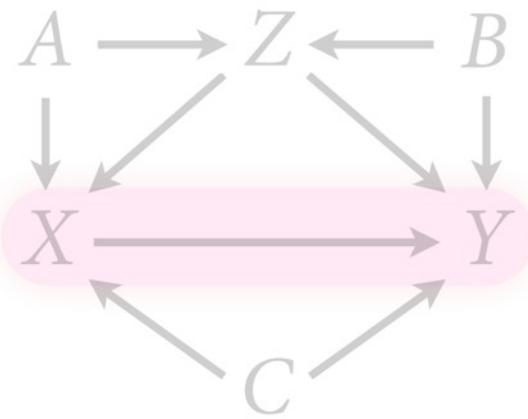




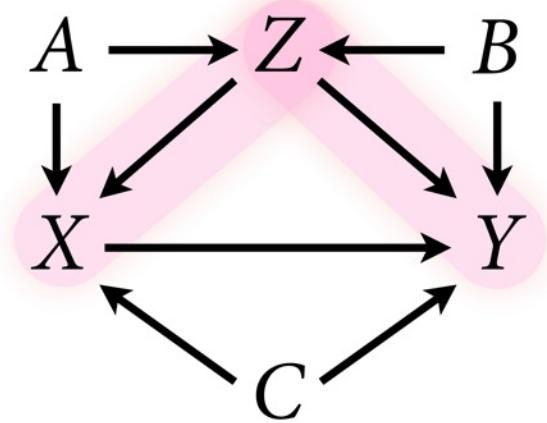
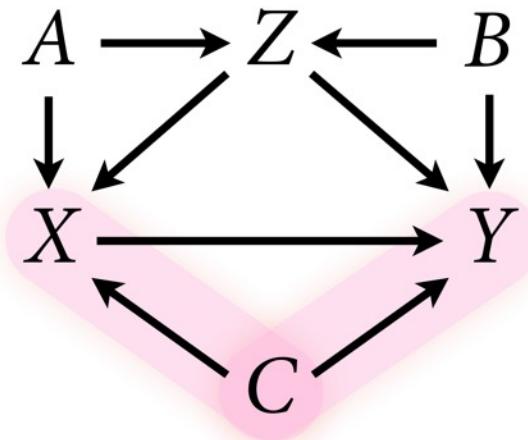
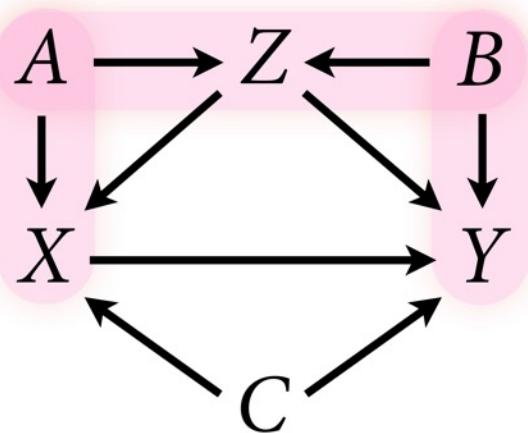








We need to stratify by A or B, (plus Z and C, since confounds)
close the path between X and Y. Since B has a direct effect of Y
we want to control on that on, so here B.



Tekst

Adjustment set: *C, Z, and either A or B*

(*B* is better choice)

www.dagitty.net

Model | Examples | How to ... | Layout | Help

Causal effect identification

Adjustment (total effect) ▾

Minimal sufficient adjustment sets for estimating the total effect of X on Y:

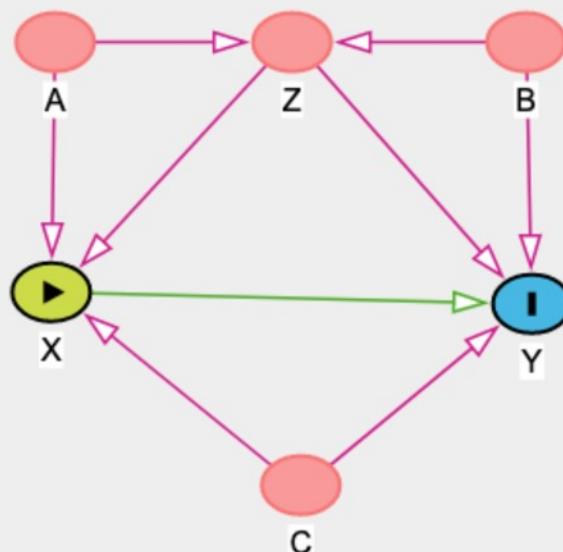
- A, C, Z
- B, C, Z

Testable implications

The model implies the following conditional independences:

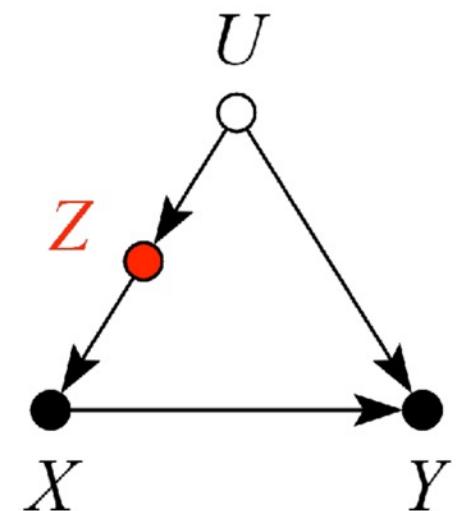
- $X \perp\!\!\!\perp B \mid A, Z$
- $Y \perp\!\!\!\perp A \mid B, C, X, Z$
- $A \perp\!\!\!\perp B$
- $A \perp\!\!\!\perp C$
- $B \perp\!\!\!\perp C$
- $Z \perp\!\!\!\perp C$

[Export R code](#)



Backdoor Criterion

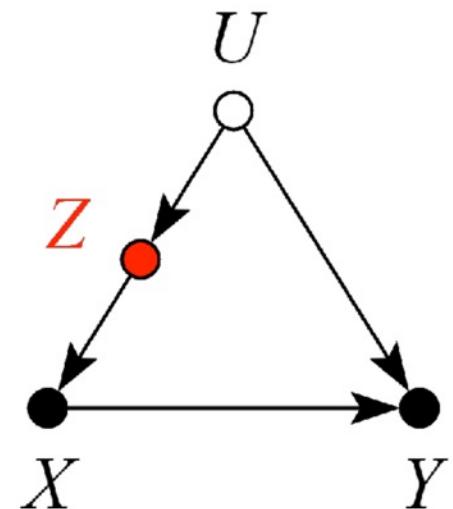
Backdoor Criterion: Rule to find adjustment set to yield $P(Y|\text{do}(X))$



Backdoor Criterion

Backdoor Criterion: Rule to find adjustment set to yield $P(Y|\text{do}(X))$

Beware non-causal paths that you open while closing other paths!

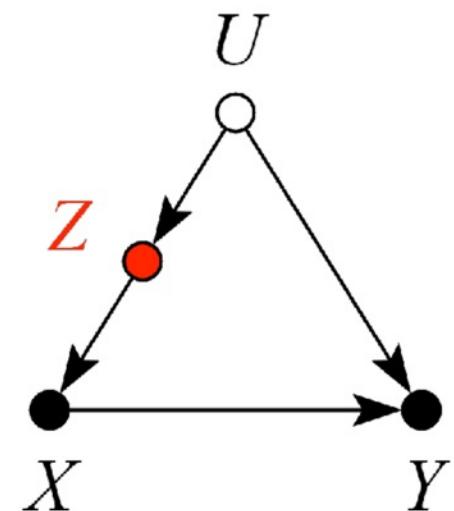


Backdoor Criterion

Backdoor Criterion: Rule to find adjustment set to yield $P(Y|\text{do}(X))$

Beware non-causal paths that you open while closing other paths!

More than backdoors:



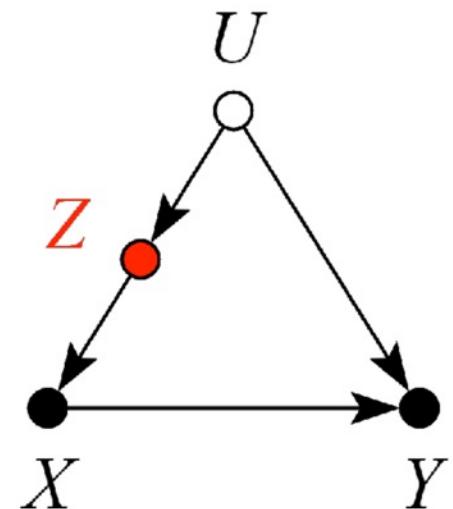
Backdoor Criterion

Backdoor Criterion: Rule to find adjustment set to yield $P(Y|\text{do}(X))$

Beware non-causal paths that you open while closing other paths!

More than backdoors:

Also solutions with simultaneous equations
(instrumental variables e.g.)



Backdoor Criterion

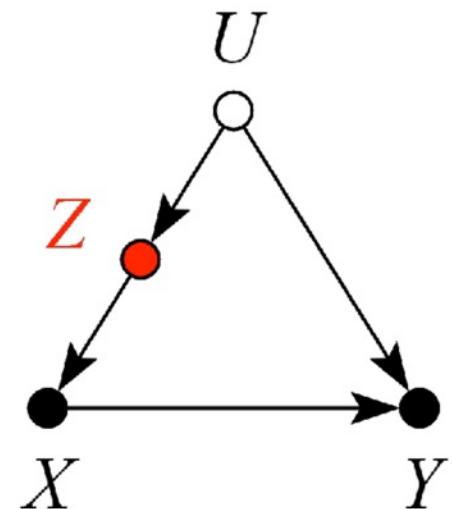
Backdoor Criterion: Rule to find adjustment set to yield $P(Y|\text{do}(X))$

Beware non-causal paths that you open while closing other paths!

More than backdoors:

Also solutions with simultaneous equations
(instrumental variables e.g.)

Full Luxury Bayes: use all variables, but in separate sub-models instead of single regression



Good & Bad Controls

“Control” variable: Variable introduced to an analysis so that a causal estimate is possible

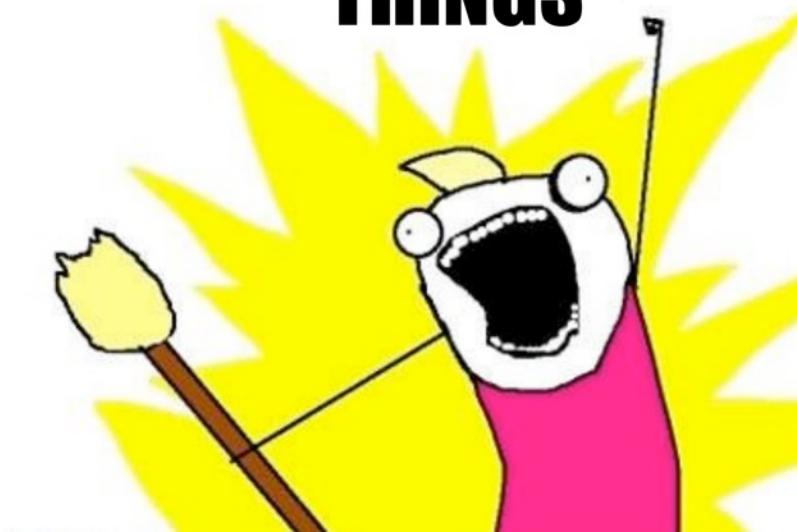
Common **wrong** heuristics for choosing control variables

Anything in the spreadsheet **YOLO!**

Any variables not highly **collinear**

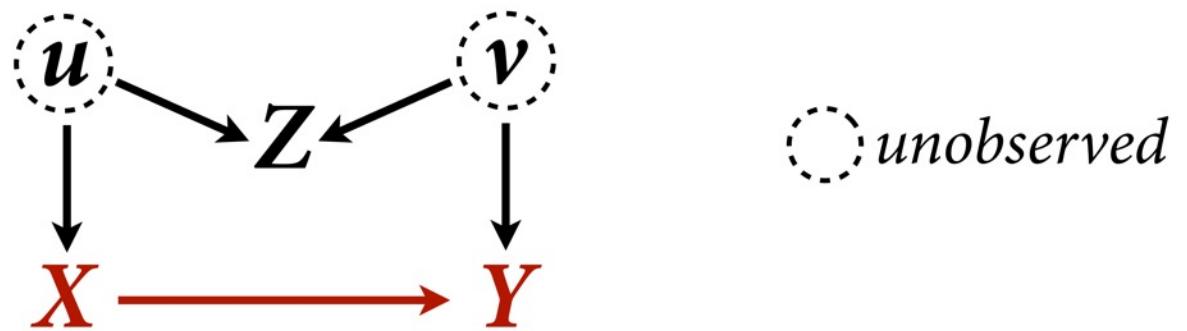
Any **pre-treatment** measurement (baseline)

**CONTROL
ALL THE
THINGS**

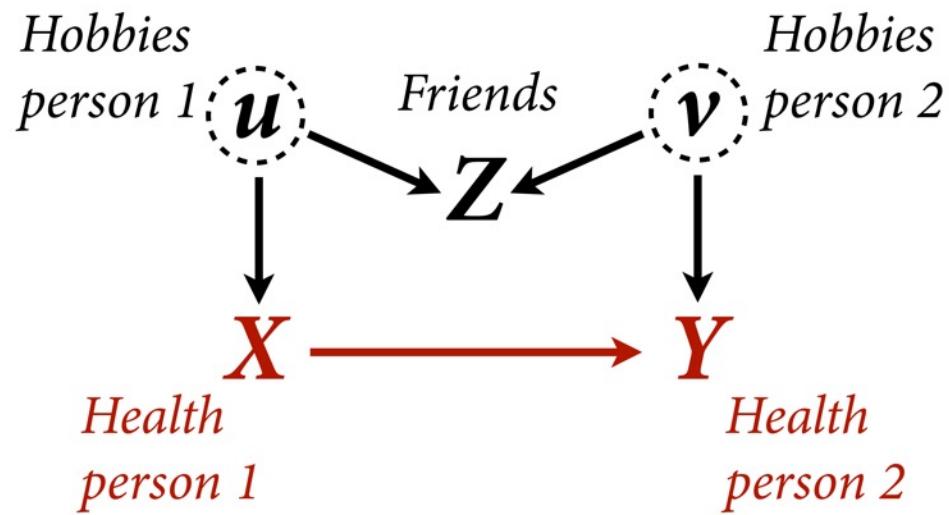


$$X \longrightarrow Y$$

Cinelli, Forney, Pearl 2021 A Crash Course in Good and Bad Controls

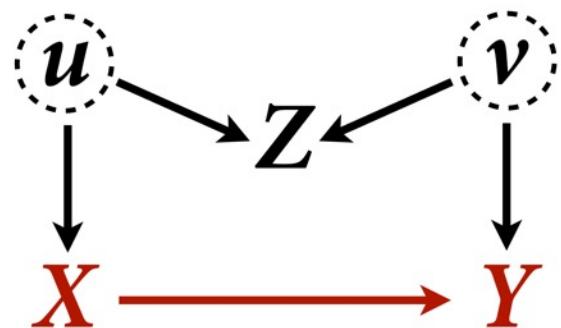


Cinelli, Forney, Pearl 2021 A Crash Course in Good and Bad Controls



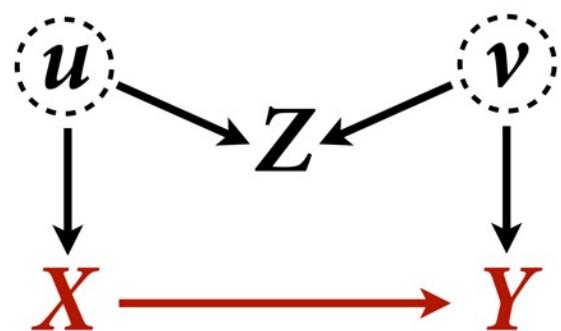
Cinelli, Forney, Pearl 2021 A Crash Course in Good and Bad Controls

(1) List the paths



(1) List the paths

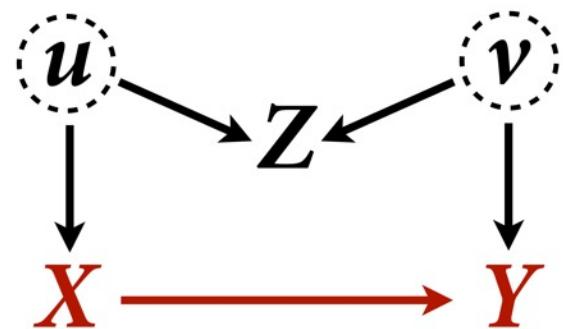
$$X \rightarrow Y$$



(1) List the paths

$$X \rightarrow Y$$

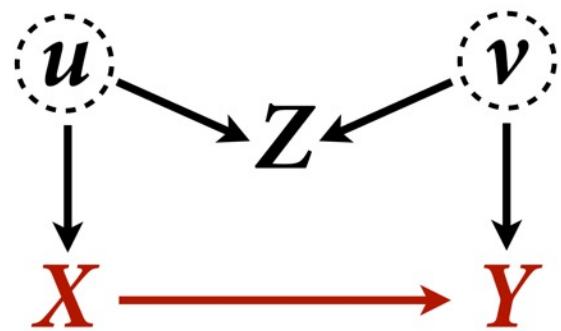
$$X \leftarrow u \rightarrow Z \leftarrow v \rightarrow Y$$



(1) List the paths (2) Find backdoors

$X \rightarrow Y$
frontdoor & open

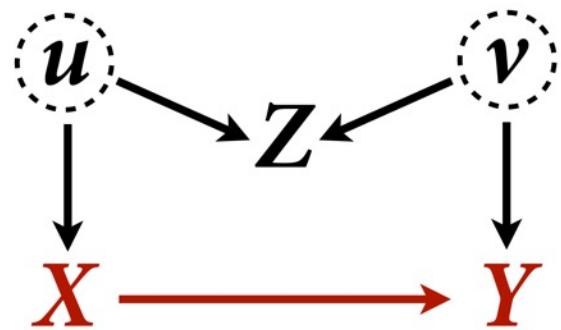
$X \leftarrow u \rightarrow Z \leftarrow v \rightarrow Y$
backdoor & closed



(1) List the paths (2) Find backdoors

$X \rightarrow Y$
frontdoor & open

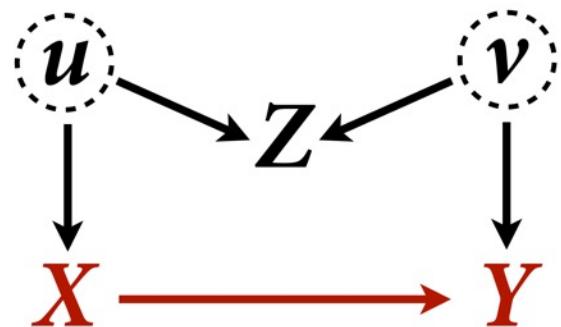
$X \leftarrow u \rightarrow Z \leftarrow v \rightarrow Y$
backdoor & closed



- (1) List the paths (2) Find backdoors (3) Close backdoors

$X \rightarrow Y$
frontdoor & open

$X \leftarrow u \rightarrow Z \leftarrow v \rightarrow Y$
backdoor & closed

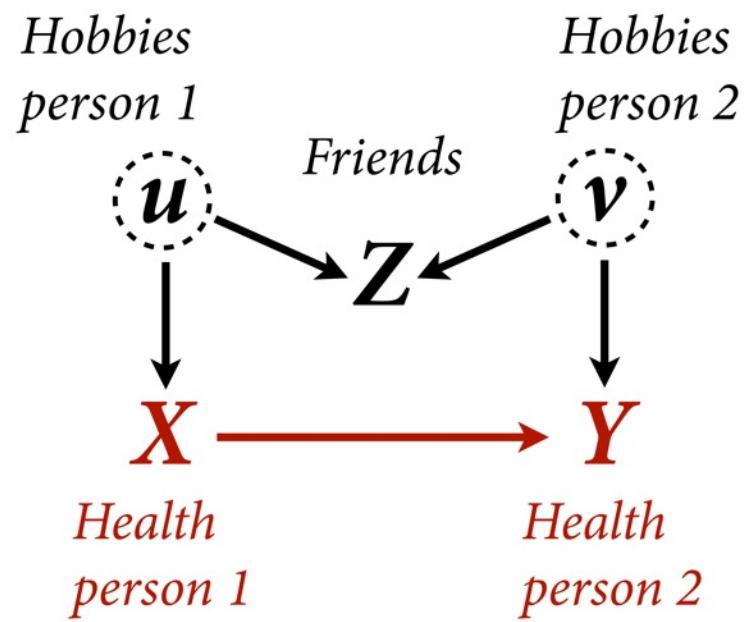


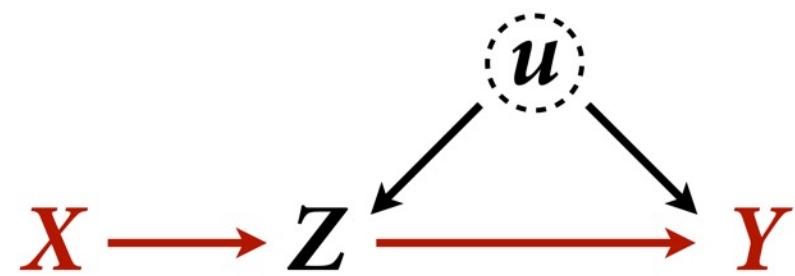
What happens if you stratify by Z?

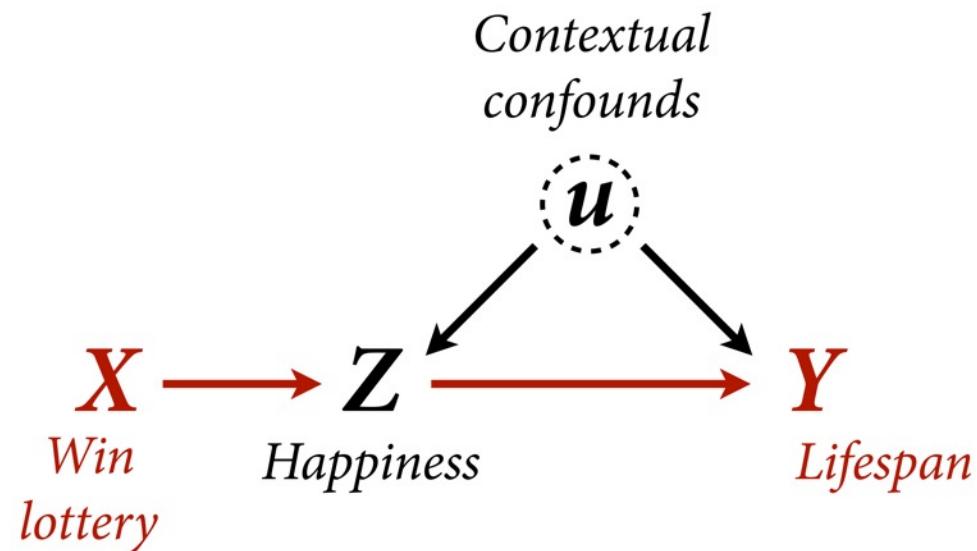
Opens the backdoor path

Z could be a **pre-treatment** variable

Not safe to always control pre-treatment measurements



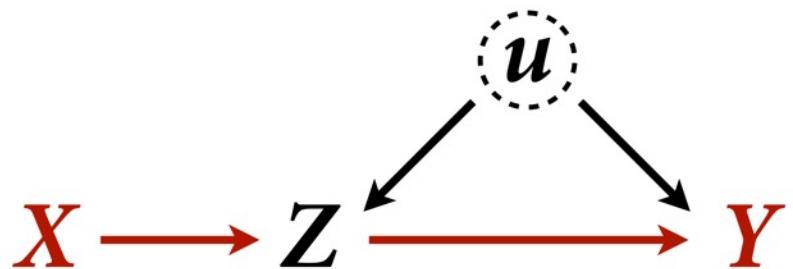




$$X \rightarrow Z \rightarrow Y$$

$$X \rightarrow Z \leftarrow u \rightarrow Y$$

No backdoor, no need
to control for Z



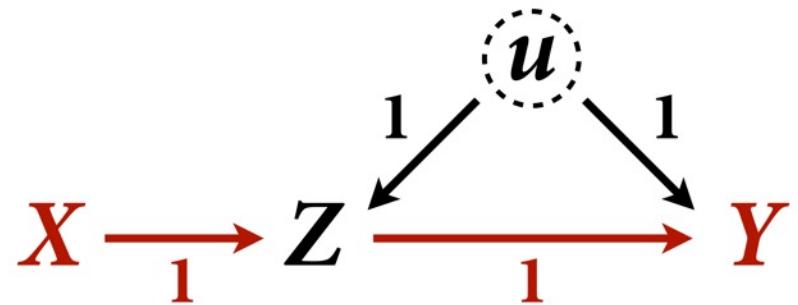
```

f <- function(n=100,bXZ=1,bZY=1) {
  X <- rnorm(n)
  u <- rnorm(n)
  Z <- rnorm(n, bXZ*X + u)
  Y <- rnorm(n, bZY*Z + u )
  bX <- coef( lm(Y ~ X) )['X']
  bXZ <- coef( lm(Y ~ X + Z) )['X']
  return( c(bX,bXZ) )
}

sim <- mcreplicate( 1e4 , f() , mc.cores=8 )

dens( sim[1,] , lwd=3 , xlab="posterior mean" )
dens( sim[2,] , lwd=3 , col=2 , add=TRUE )

```



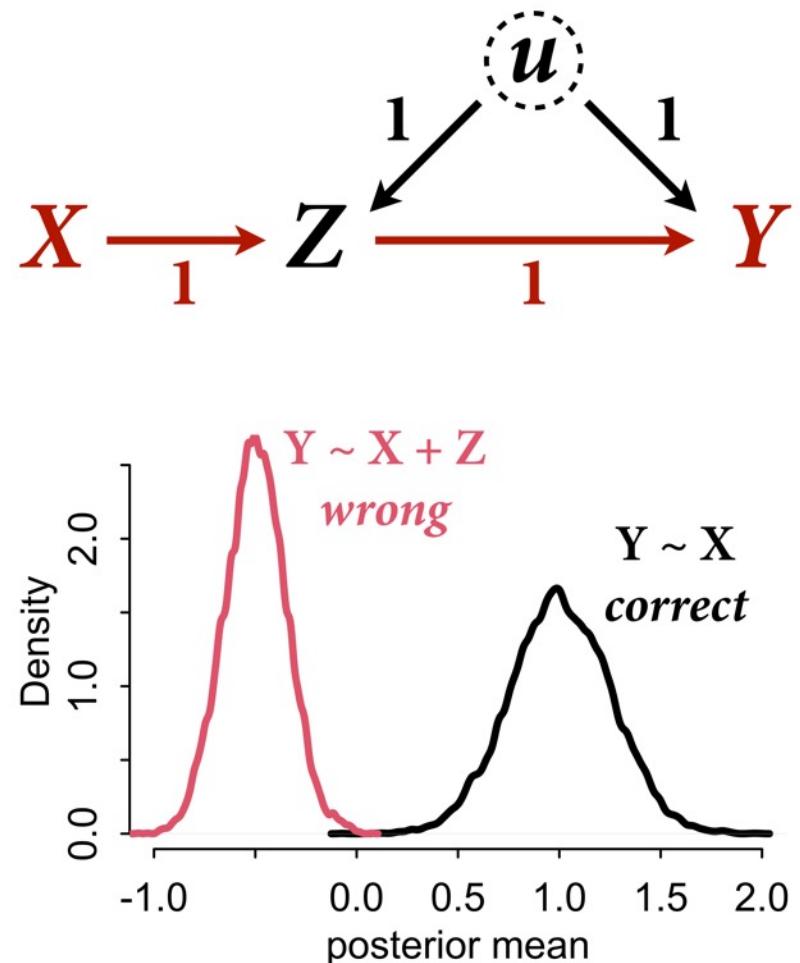
```

f <- function(n=100,bXZ=1,bZY=1) {
  X <- rnorm(n)
  u <- rnorm(n)
  Z <- rnorm(n, bXZ*X + u)
  Y <- rnorm(n, bZY*Z + u )
  bX <- coef( lm(Y ~ X) )['X']
  bXZ <- coef( lm(Y ~ X + Z) )['X']
  return( c(bX,bXZ) )
}

sim <- mcreplicate( 1e4 , f() , mc.cores=8 )

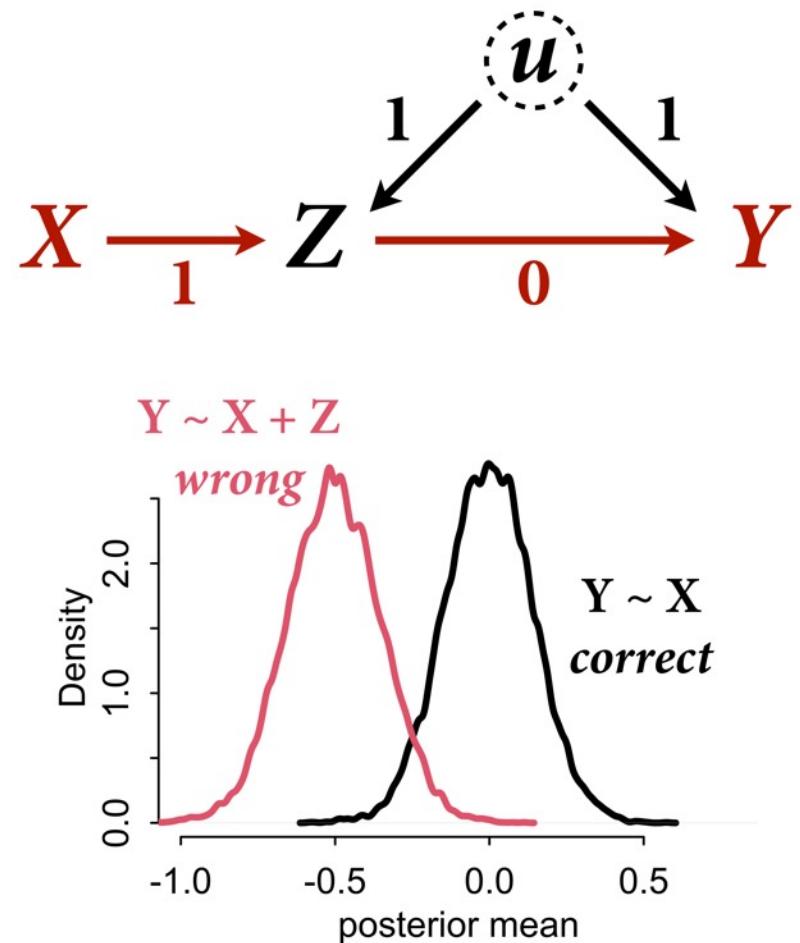
dens( sim[1,] , lwd=3 , xlab="posterior mean" )
dens( sim[2,] , lwd=3 , col=2 , add=TRUE )

```



Change bZY to zero

```
f <- function(n=100,bXZ=1,bZY=1) {  
  X <- rnorm(n)  
  u <- rnorm(n)  
  Z <- rnorm(n, bXZ*X + u)  
  Y <- rnorm(n, bZY*Z + u )  
  bX <- coef( lm(Y ~ X) )['X']  
  bXZ <- coef( lm(Y ~ X + Z) )['X']  
  return( c(bX,bXZ) )  
}  
  
sim <- mcreplicate( 1e4 , f(bZY=0) , mc.cores=8 )  
  
dens( sim[1,] , lwd=3 , xlab="posterior mean" )  
dens( sim[2,] , lwd=3 , col=2 , add=TRUE )
```



$$X \rightarrow Z \rightarrow Y$$

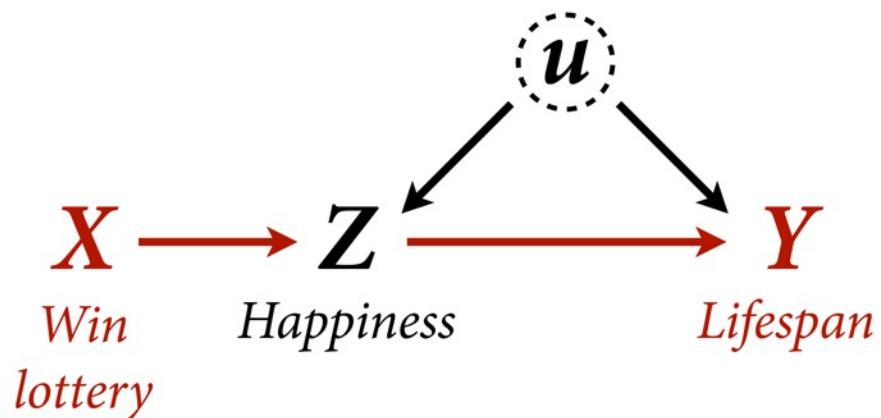
$$X \rightarrow Z \leftarrow u \rightarrow Y$$

Controlling for Z biases treatment estimate X

Controlling for Z opens biasing path through u

Can estimate effect of X ; Cannot estimate mediation effect Z

No backdoor, no need to control for Z

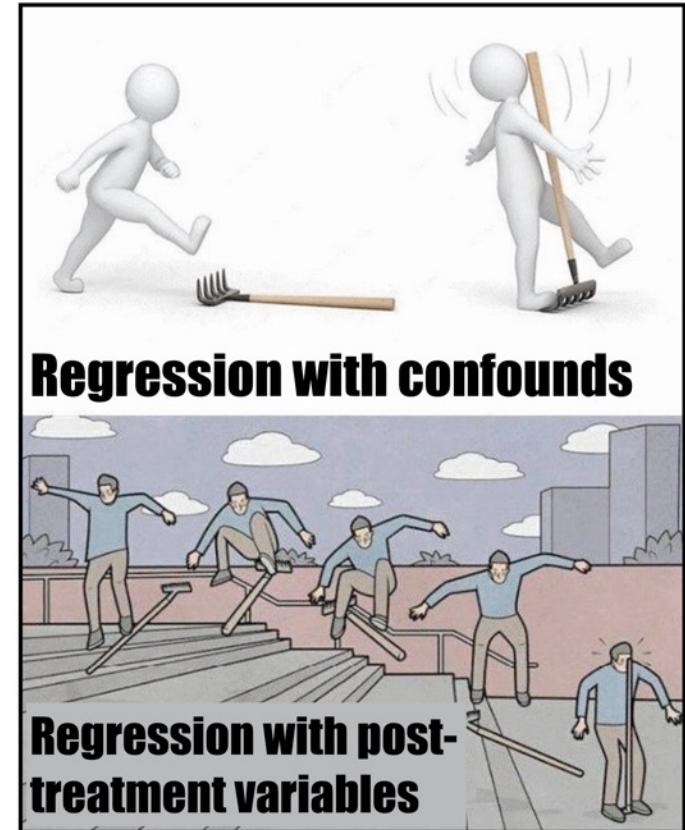


Post-treatment bias is common

TABLE 1 Posttreatment Conditioning in Experimental Studies

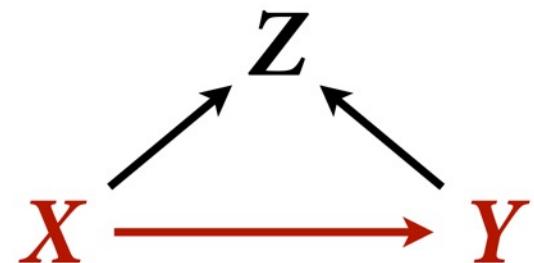
Category	Prevalence
Engages in posttreatment conditioning	46.7%
Controls for/interacts with a posttreatment variable	21.3%
Drops cases based on posttreatment criteria	14.7%
Both types of posttreatment conditioning present	10.7%
No conditioning on posttreatment variables	52.0%
Insufficient information to code	1.3%

Note: The sample consists of 2012–14 articles in the *American Political Science Review*, the *American Journal of Political Science*, and the *Journal of Politics* including a survey, field, laboratory, or lab-in-the-field experiment (n = 75).

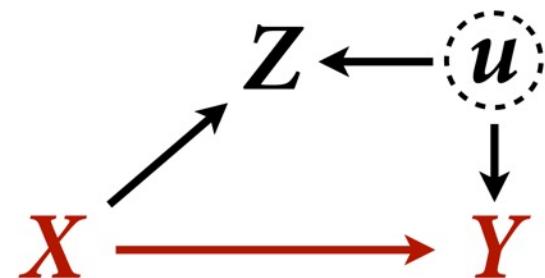


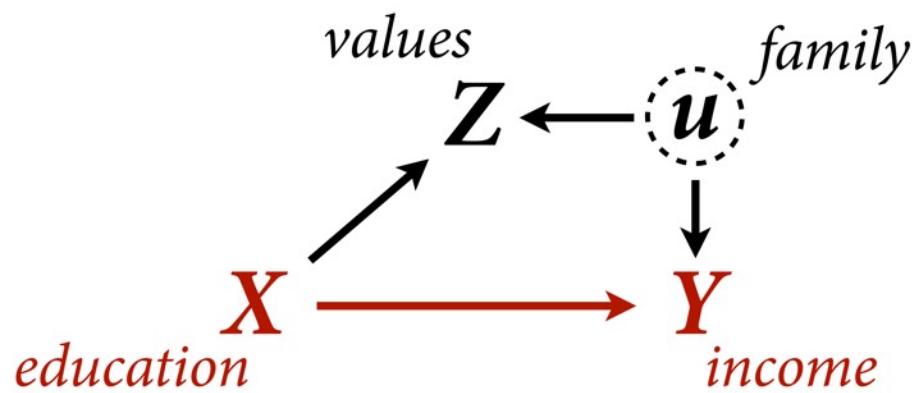
Montgomery et al 2018 How Conditioning on Posttreatment Variables Can Ruin Your Experiment

Do not touch the collider!

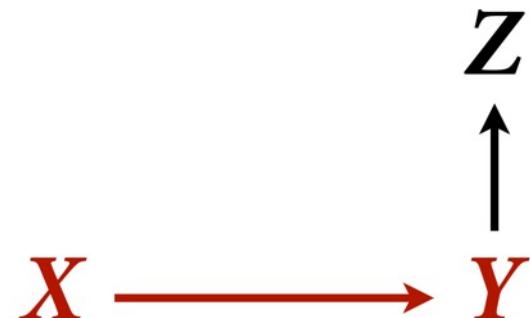


Colliders not always so obvious

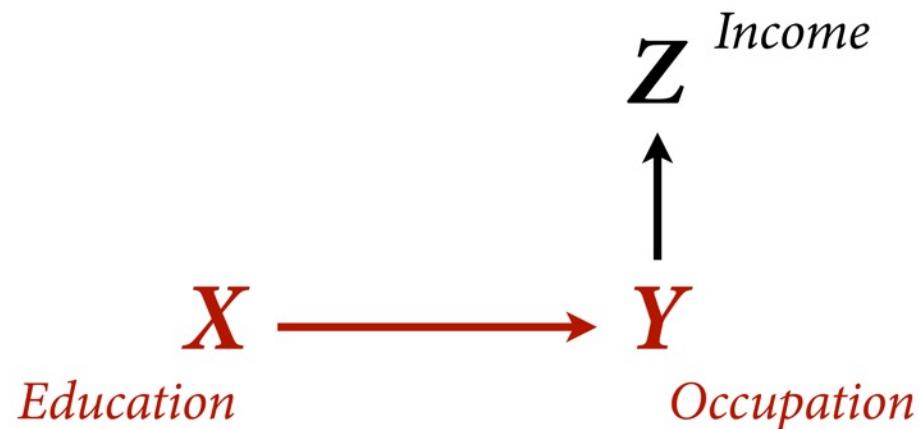




“Case-control bias”

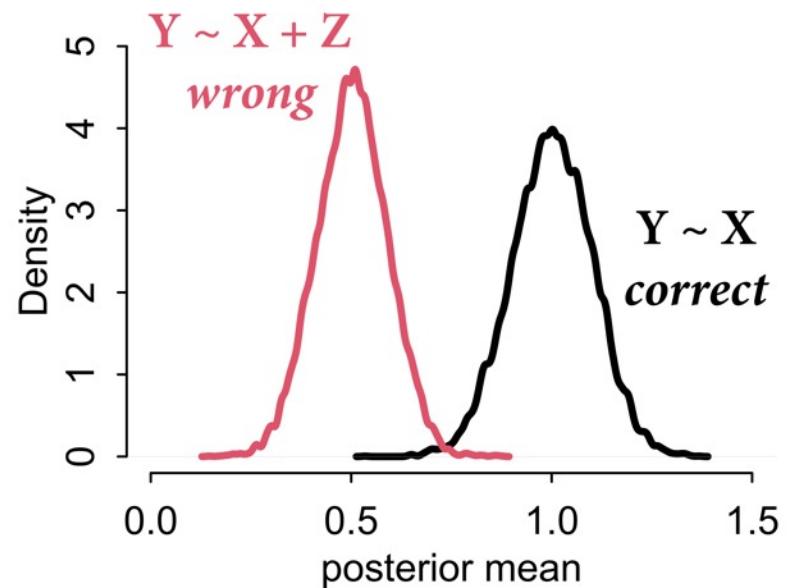


“Case-control bias”



“Case-control bias”

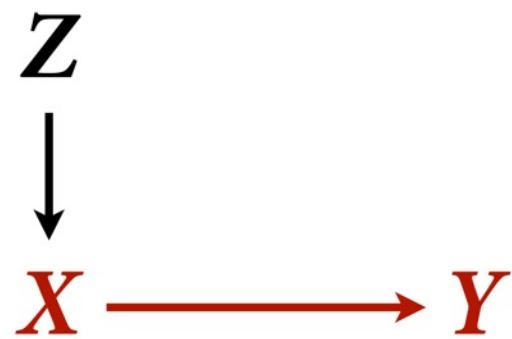
```
f <- function(n=100,bXY=1,bYZ=1) {  
  X <- rnorm(n)  
  Y <- rnorm(n, bXY*X )  
  Z <- rnorm(n, bYZ*Y )  
  bX <- coef( lm(Y ~ X) )['X']  
  bXZ <- coef( lm(Y ~ X + Z) )['X']  
  return( c(bX,bXZ) )  
}  
  
sim <- mcreplicate( 1e4 , f() , mc.cores=8 )  
  
dens( sim[1,] , lwd=3 , xlab="posterior mean" )  
dens( sim[2,] , lwd=3 , col=2 , add=TRUE )
```



“Precision parasite”

No backdoors

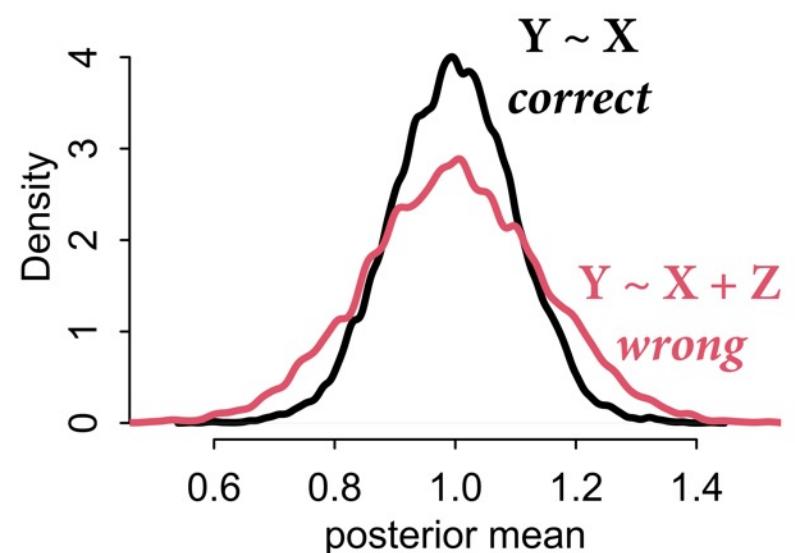
But still not good to
condition on Z



“Precision parasite”

```
f <- function(n=100,bZX=1,bXY=1) {  
  Z <- rnorm(n)  
  X <- rnorm(n, bZX*Z )  
  Y <- rnorm(n, bXY*X )  
  bX <- coef( lm(Y ~ X) )['X']  
  bXZ <- coef( lm(Y ~ X + Z) )['X']  
  return( c(bX,bXZ) )  
}  
  
sim <- mcreplicate( 1e4 , f(n=50) , mc.cores=8 )  
  
dens( sim[1,] , lwd=3 , xlab="posterior mean" )  
dens( sim[2,] , lwd=3 , col=2 , add=TRUE )
```

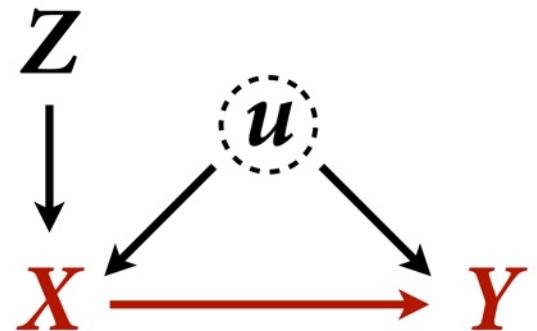
$Z \rightarrow X \rightarrow Y$



“Bias amplification”

X and Y confounded by u

Something **truly awful** happens
when we add Z



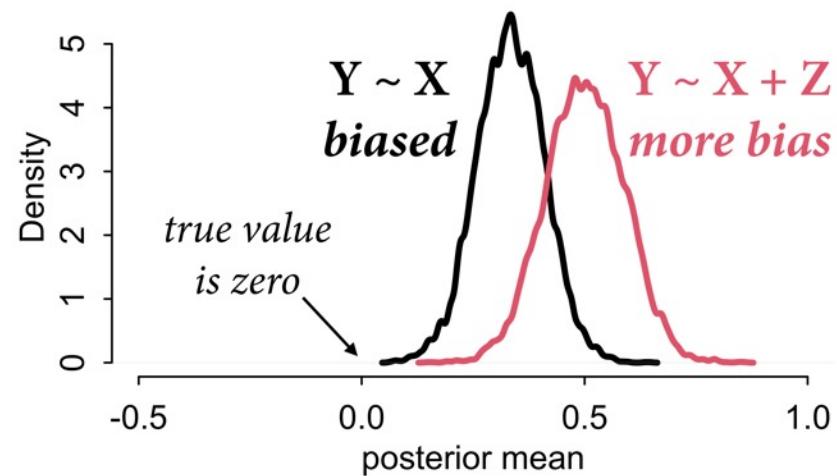
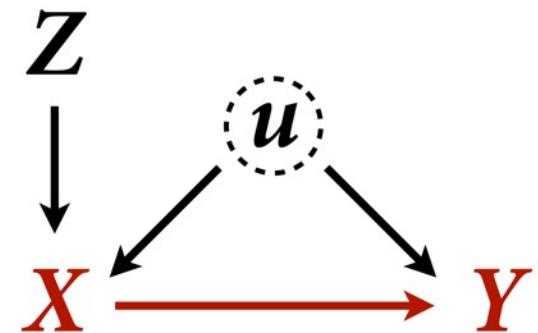
```

f <- function(n=100,bZX=1,bXY=1) {
  Z <- rnorm(n)
  u <- rnorm(n)
  X <- rnorm(n, bZX*Z + u )
  Y <- rnorm(n, bXY*X + u )
  bX <- coef( lm(Y ~ X) )['X']
  bXZ <- coef( lm(Y ~ X + Z) )['X']
  return( c(bX,bXZ) )
}

sim <- mcreplicate( 1e4 , f(bXY=0) , mc.cores=8 )

dens( sim[1,] , lwd=3 , xlab="posterior mean" )
dens( sim[2,] , lwd=3 , col=2 , add=TRUE )

```

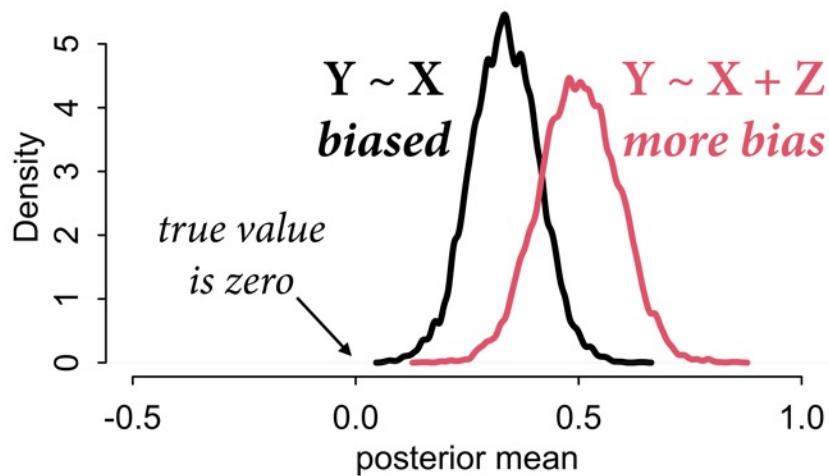
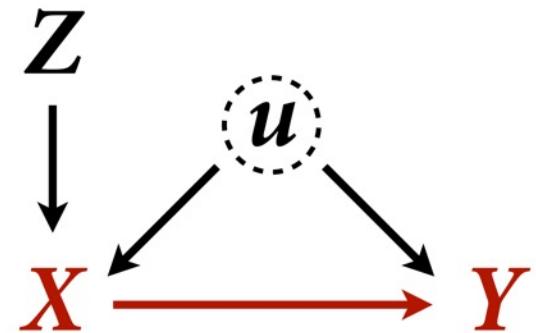


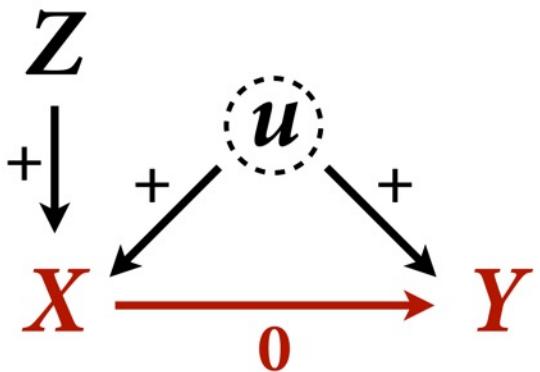
WHY?

Covariation X & Y requires variation in their causes

Within each level of Z , less variation in X

Confound u relatively more important within each Z

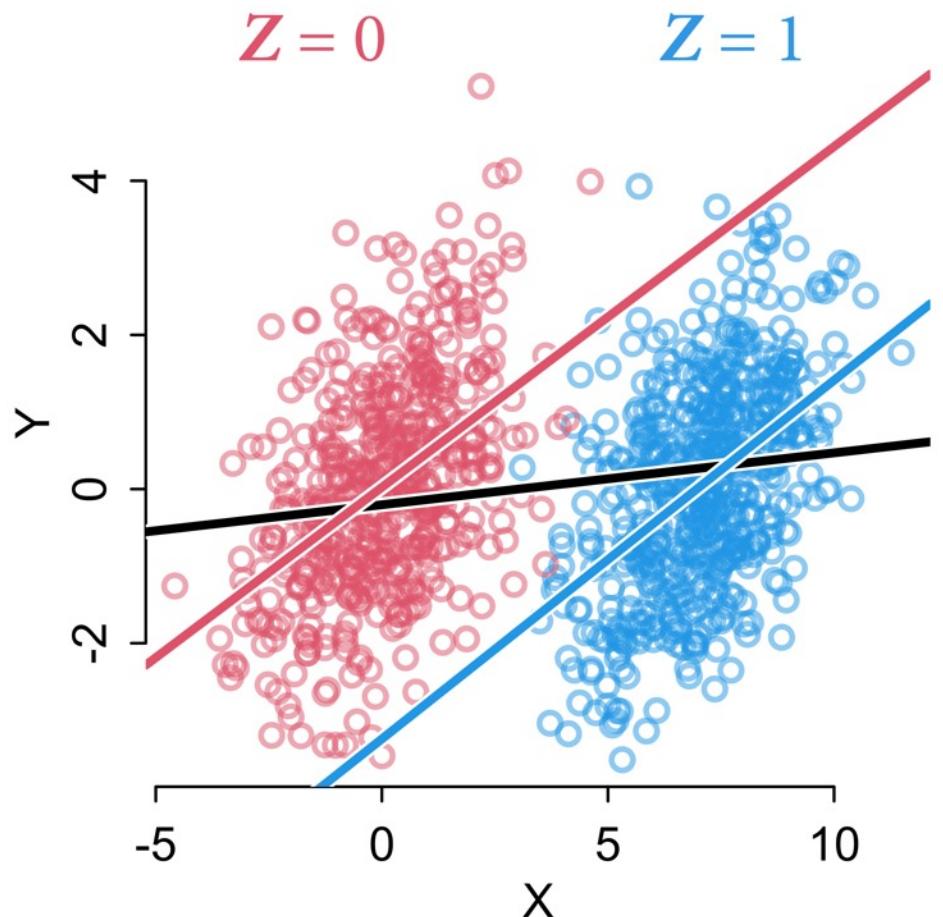




```

n <- 1000
Z <- rbern(n)
u <- rnorm(n)
X <- rnorm(n, 7*Z + u )
Y <- rnorm(n, 0*X + u )

```



Good & Bad Controls

“Control” variable: Variable introduced to an analysis so that a causal estimate is possible

Heuristics fail — adding control variables can be worse than omitting

Make assumptions explicit

**MODEL
ALL THE
THINGS**

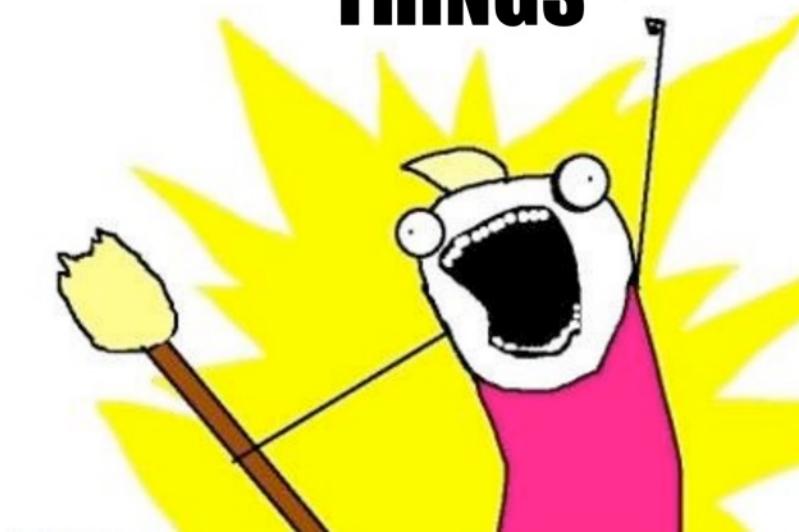


Table 2 Fallacy

Not all coefficients are causal effects

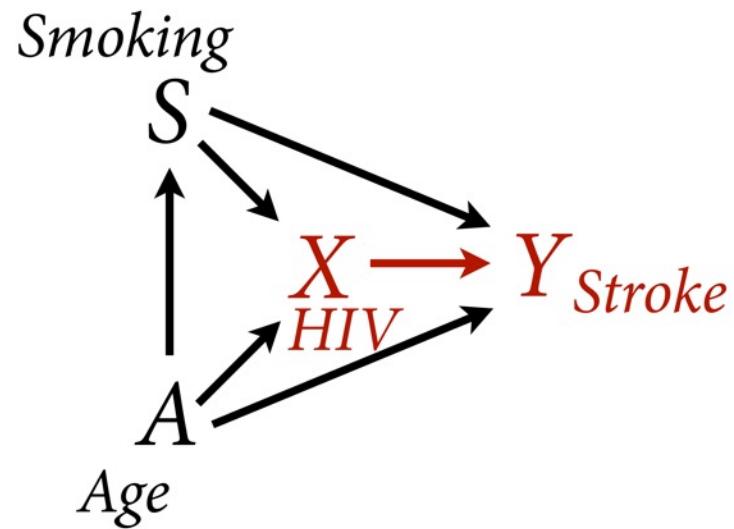
Statistical model designed to identify $X \rightarrow Y$ will not also identify effects of control variables

Table 2 is dangerous

TABLE 2—ESTIMATED PROBIT MODELS
FOR THE USE OF A SCREEN

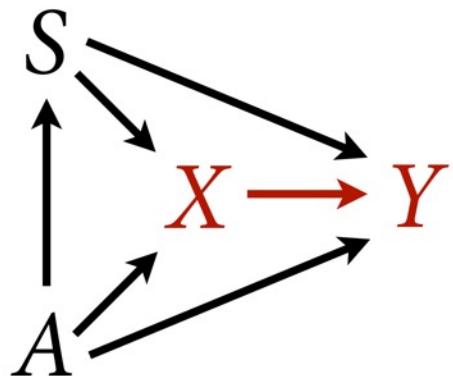
	Preliminaries	Finals	blind
	(1)	(2)	(3)
(Proportion female) _{t-1}	2.744 (3.265) [0.006]	3.120 (3.271) [0.004]	0.490 (1.163) [0.011]
(Proportion of orchestra personnel with <6 years tenure) _{t-1}	-26.46 (7.314) [-0.058]	-28.13 (8.459) [-0.039]	-9.467 (2.787) [-0.207]
“Big Five” orchestra		0.367 (0.452) [0.001]	
pseudo R^2	0.178	0.193	0.050
Number of observations	294	294	434

Westreich & Greenland 2013 The Table 2 Fallacy

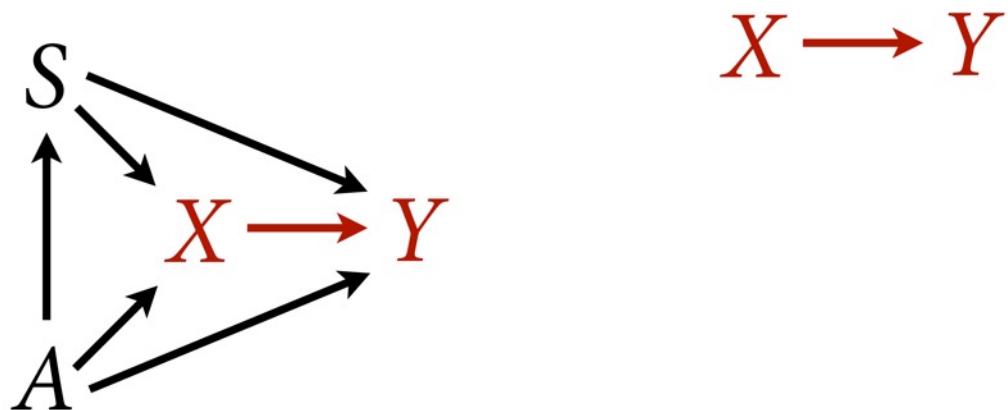


Westreich & Greenland 2013 The Table 2 Fallacy

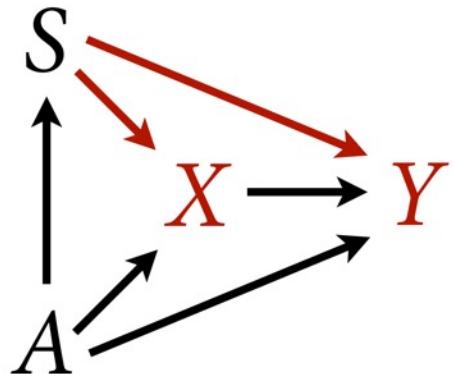
Use Backdoor Criterion



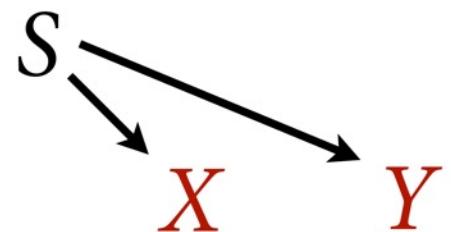
Use Backdoor Criterion



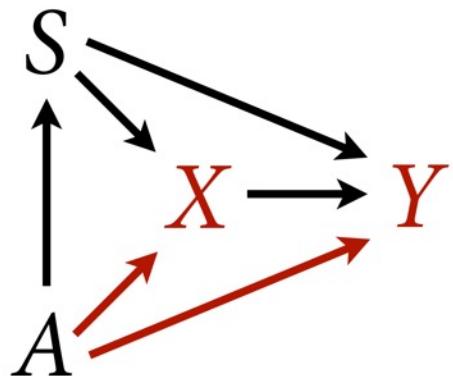
Use Backdoor Criterion



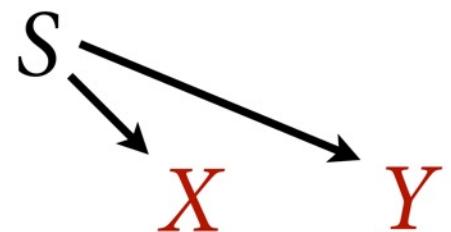
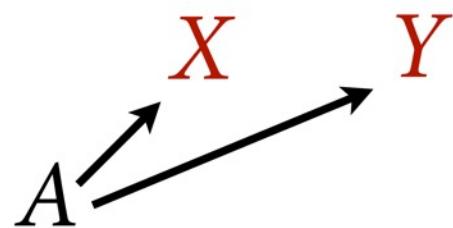
$X \rightarrow Y$



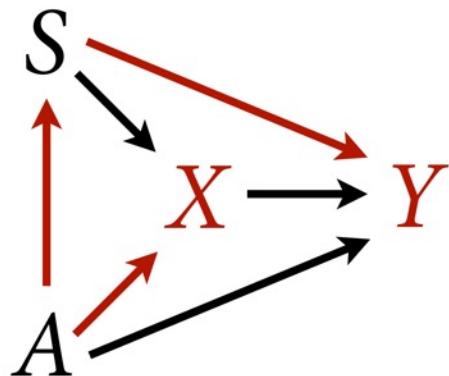
Use Backdoor Criterion



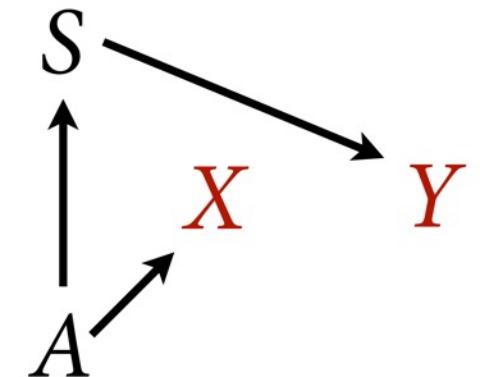
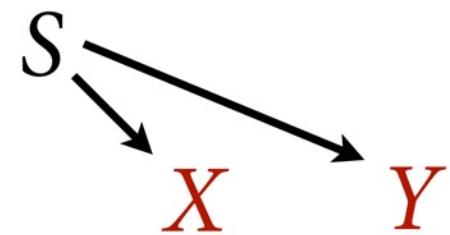
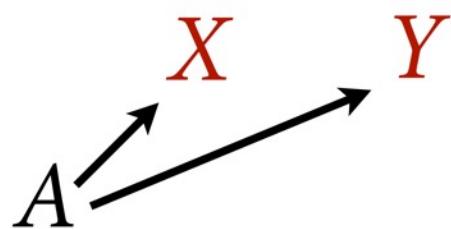
$$X \rightarrow Y$$



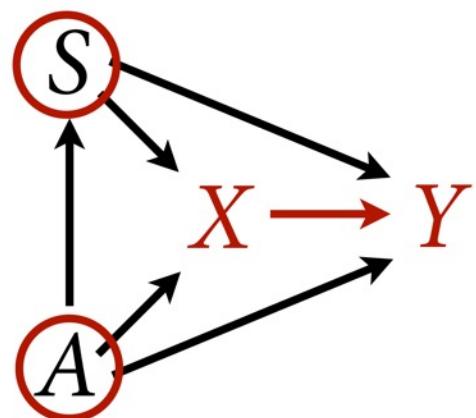
Use Backdoor Criterion



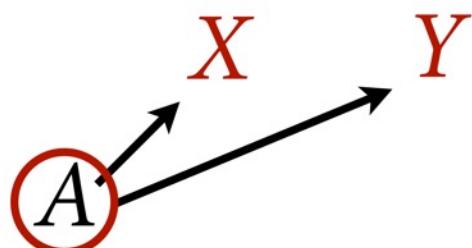
$X \rightarrow Y$



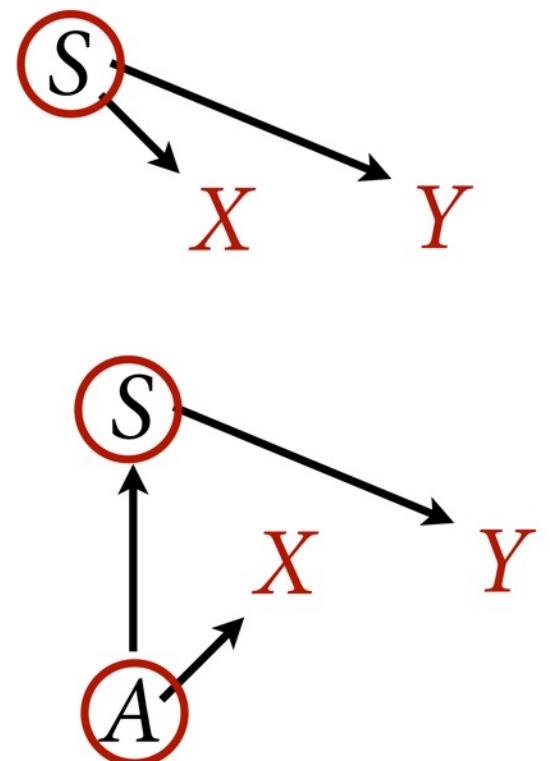
Use Backdoor Criterion

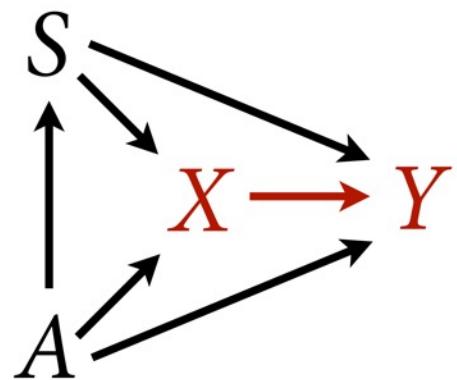


$$X \rightarrow Y$$



$$X \rightarrow Y$$



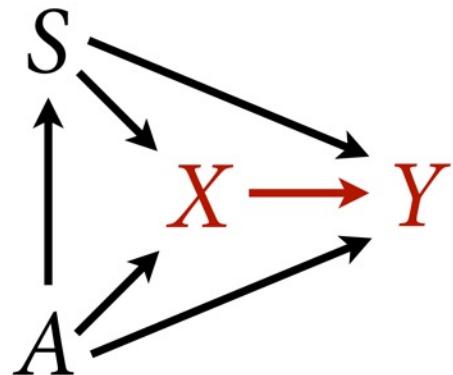


$$Y_i \sim \text{Normal}(\mu_i, \sigma)$$

$$\mu_i = \alpha + \beta_X X_i + \beta_S S_i + \beta_A A_i$$

X

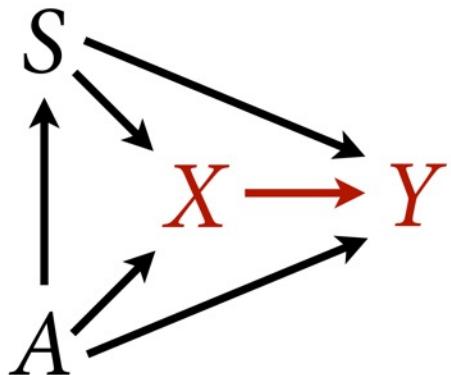
Unconditional



Confounded by A
and S

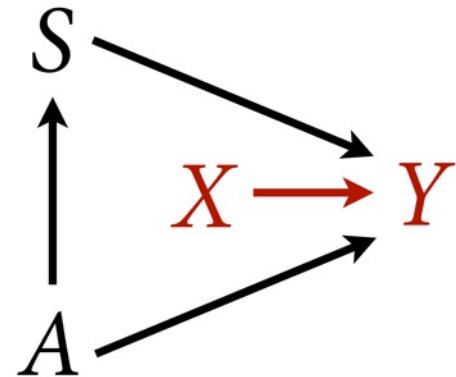
X

Unconditional



Confounded by A
and S

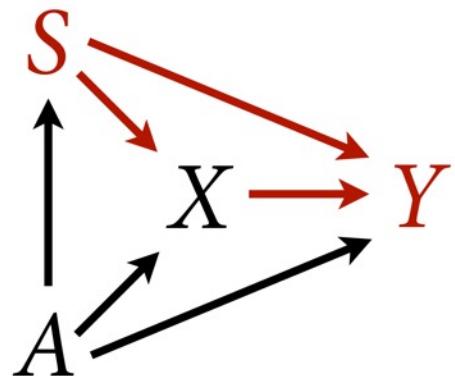
Conditional on A and S



Coefficient for \mathbf{X} :
Effect of X on Y
(still must
marginalize!)

S

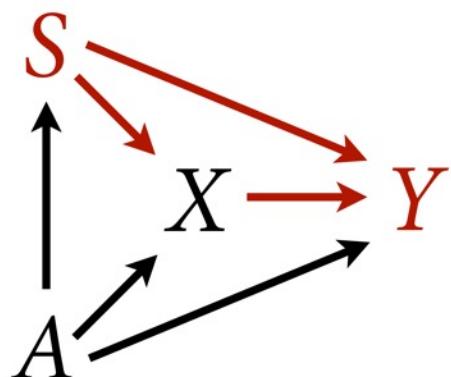
Unconditional



Effect of S
confounded by A

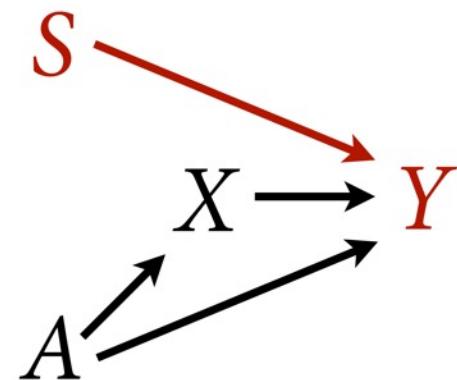
S

Unconditional



Effect of S
confounded by A

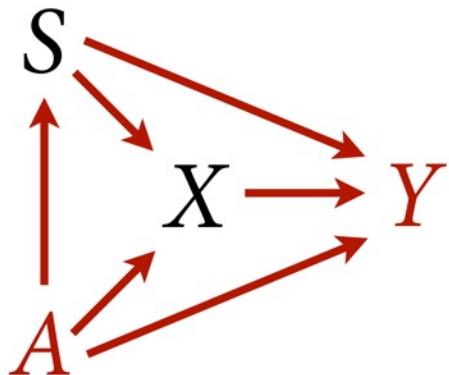
Conditional on A and X



Coefficient for S:
Direct effect of S on Y

A

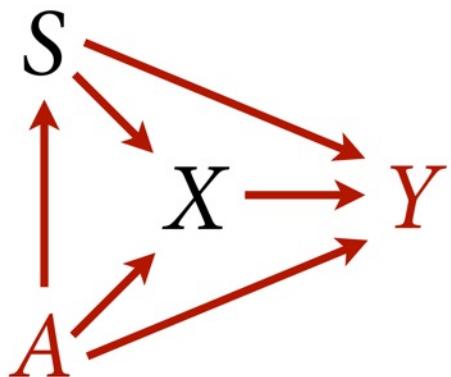
Unconditional



Total causal effect
of A on Y flows
through all paths

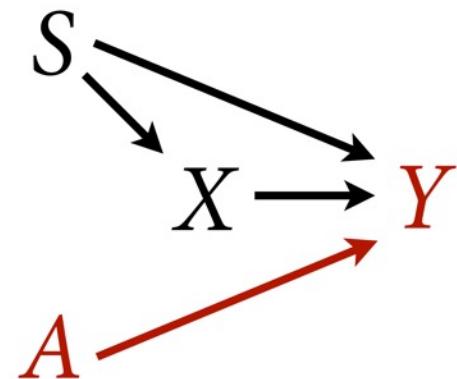
A

Unconditional

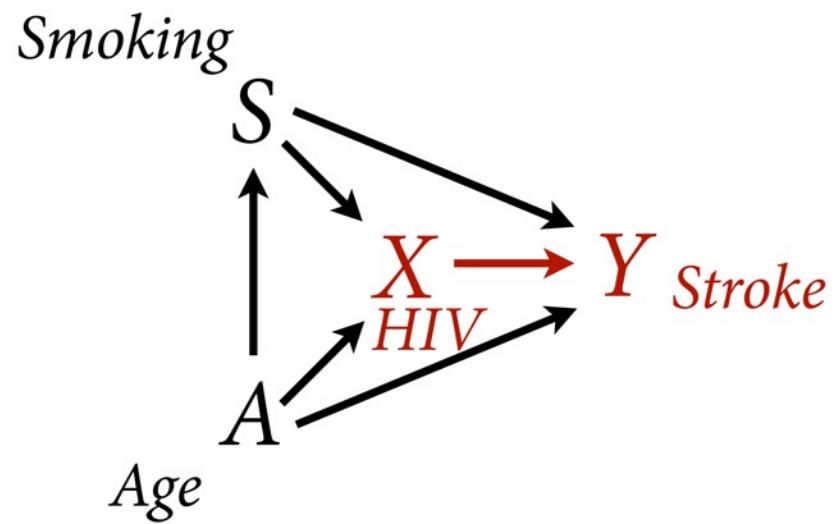


Total causal effect
of A on Y flows
through all paths

Conditional on X and S



Coefficient for A :
Direct effect of A on Y



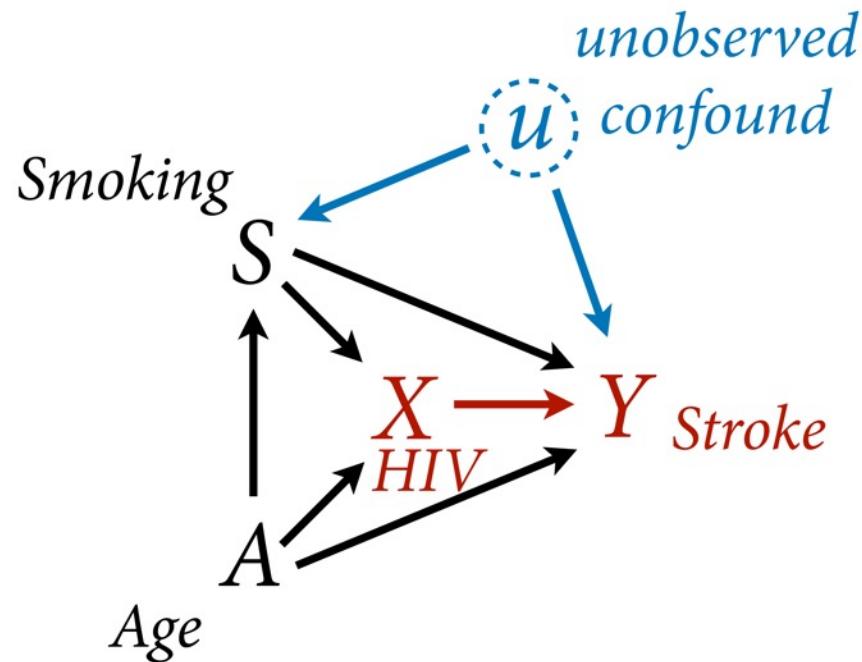


Table 2 Fallacy

Not all coefficients created equal

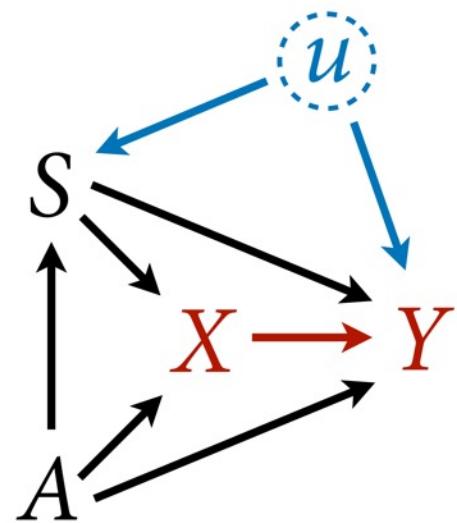
So do not present them as equal

Options:

Do not present control coefficients

Give explicit interpretation of each

No causal model, no interpretation



Imagine Confounding

Often we cannot credibly adjust for all confounding

Do not give up!

Biased estimate can be better than no estimate

Sensitivity analysis: draw the implications of what you don't know

Find natural experiment or design one

