DAGs and causal inference

Frank Edwards

2/25/2020

Ways to obtain biased regression estimates (for causal inference)

1. Omitted variable bias (spurious, masked associations)
2. Multicollinearity
3. **Post-treatment bias**

The experiment

In an experimental setting where we want to know the effect of a *treatment* on an outcome, we should not condition on variables whose effect would occur *after* exposure to treatment.

The experiment

In an experimental setting where we want to know the effect of a *treatment* on an outcome, we should not condition on variables whose effect would occur *after* exposure to treatment.

- **Q:** How much do plants grow under different anti-fungal soil treatments?
- **Data:** Initial heights (h0), assignment to treatment (treatment), presence of fungus (fungus), height post-treatment (h1)

Simulation code for this exercise included in the .Rmd file

The data

```
N<-100
d<-data.frame(h0=rnorm(N, 10, 2), # generate initial height
              treatment = rep(c(0,1), each = N/2)) # random assignment to treatment

d<-d %>%
  mutate(fungus = rbinom(N, size = 1, prob = 0.5 - treatment * 0.4), # p(fungus) is lower for treated
         h1 = h0 + rnorm(N, 5-3 * fungus, 1))

summary(d)
```

```
##        h0           treatment       fungus           h1
##  Min.   : 5.571   Min.   :0.0   Min.   :0.00   Min.   : 7.944
##  1st Qu.: 9.012   1st Qu.:0.0   1st Qu.:0.00   1st Qu.:12.673
##  Median :10.228   Median :0.5   Median :0.00   Median :14.498
##  Mean   :10.218   Mean   :0.5   Mean   :0.26   Mean   :14.461
##  3rd Qu.:11.383   3rd Qu.:1.0   3rd Qu.:1.00   3rd Qu.:15.983
##  Max.   :14.803   Max.   :1.0   Max.   :1.00   Max.   :21.458
```

The model

We can model the effect of treatment on the height of a plant from time 0 to time 1 as

$$h_{1i} \sim \text{Normal}(\mu_i, \sigma)$$

$$\mu_i = h_{0i} \times p$$

$$p = \alpha + \beta_T T_i$$

$$\alpha \sim \text{Log-Normal}(0, 0.25)$$

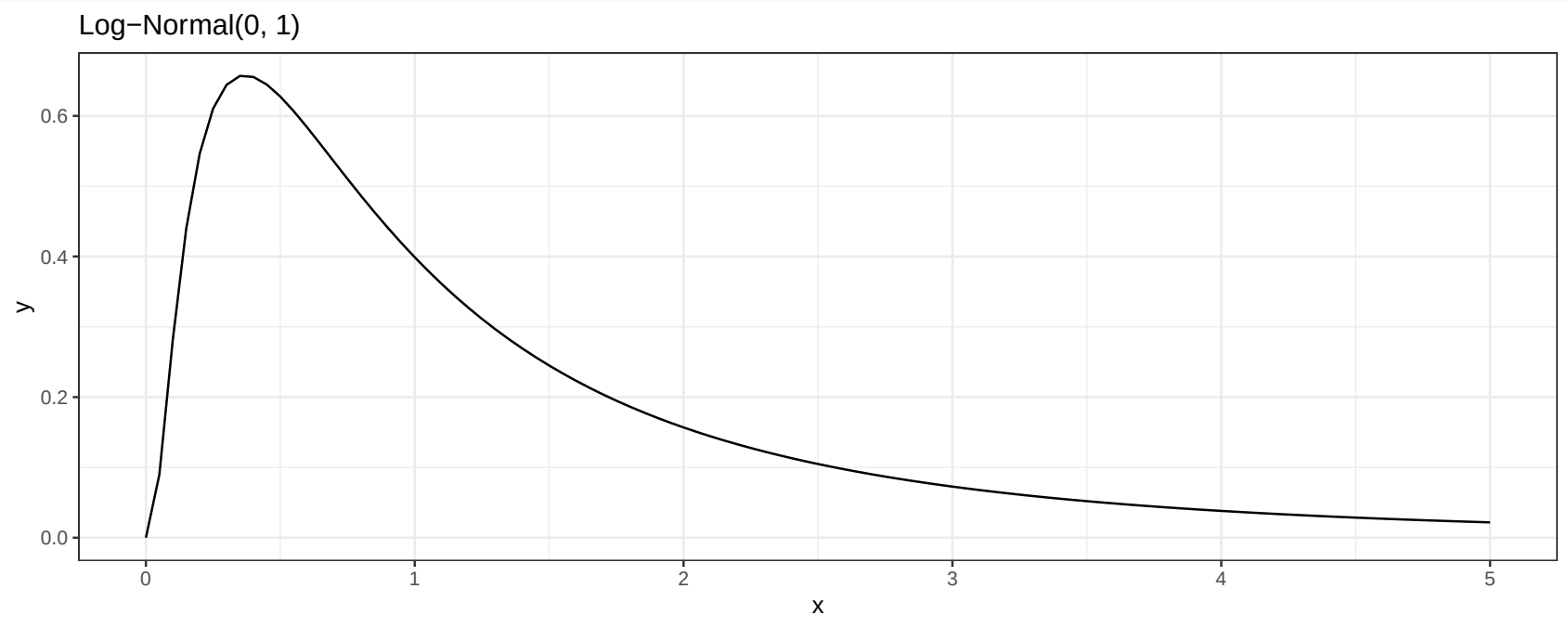
$$\beta_T \sim \text{Normal}(0, 0.5)$$

$$\sigma \sim \text{Exponential}(1)$$

Note that we are modeling height by modeling proportional growth relative to initial height h_0 with the variable p

Refresher on Log-Normal PDFs

- Log-Normal variables are always positive. If x is a Normal random variable, then e^X is a log-normal random variable.



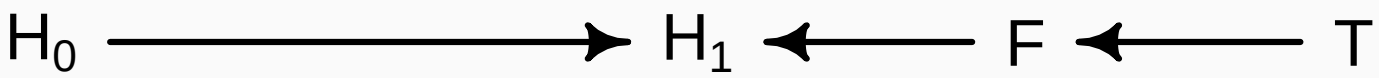
The causal model

Plant height at time 1 is influenced by:

1. Height at time 0
2. Fungus

Fungus is caused by:

1. Treatment



Building the model

```
fungus_0<-quap(alist(
  h1 ~ dnorm(mu, sigma),
  mu<- h0 * p,
  p<- a + bt * treatment,
  a ~ dlnorm(0, 0.25),
  bt ~ dnorm(0, 0.5),
  sigma ~ dexp(1)
), data = d)

summary(fungus_0)
```

```
##              mean         sd       5.5%     94.5%
## a     1.35570664 0.02312421 1.31874969 1.3926636
## bt    0.09179727 0.03256511 0.03975194 0.1438426
## sigma 1.69372441 0.11828100 1.50468852 1.8827603
```

What happens if we condition on fungus?

$$h_{1i} \sim \text{Normal}(\mu_i, \sigma)$$

$$\mu_i = h_{0i} \times p$$

$$p = \alpha + \beta_T T_i + \beta_F F_i$$

$$\alpha \sim \text{Log-Normal}(0, 0.25)$$

$$\beta_T \sim \text{Normal}(0, 0.5)$$

$$\beta_F \sim \text{Normal}(0, 0.5)$$

$$\sigma \sim \text{Exponential}(1)$$

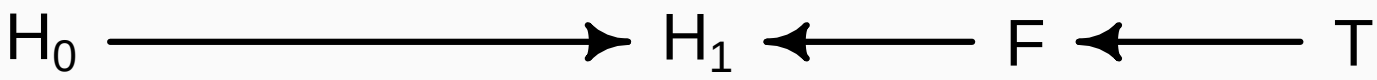
What happens if we condition on fungus?

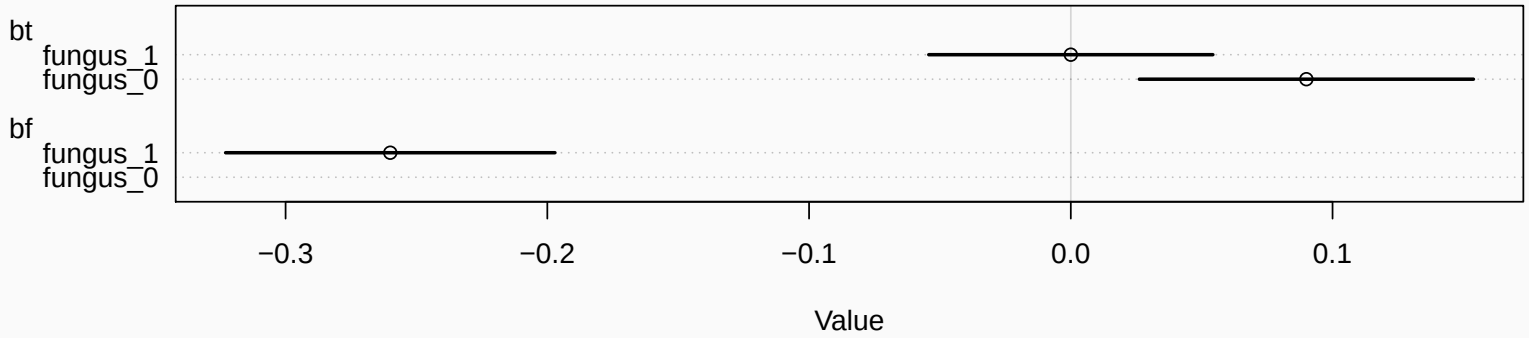
```
fungus_1<-quap(alist(
  h1 ~ dnorm(mu, sigma),
  mu<- h0 * p,
  p<- a + bt * treatment + bf * fungus,
  a ~ dlnorm(0, 0.25),
  bt ~ dnorm(0, 0.5),
  bf ~ dnorm(0, 0.5),
  sigma ~ dexp(1)
), data = d)

summary(fungus_1)
```

```
##                mean         sd        5.5%       94.5%
## a      1.465726648 0.02251993  1.42973545  1.50171784
## bt     0.001377849 0.02769289 -0.04288074  0.04563644
## bf    -0.259499795 0.03209977 -0.31080143 -0.20819816
## sigma  1.317988054 0.09230347  1.17046927  1.46550683
```

What happens if we condition on fungus?





Post-treatment variables

- Treatment is independent of growth conditional on fungus.
- Once fungus is in the model, treatment provides no additional information on growth, because T affects H by suppressing F.
- It is (generally) a bad idea to condition on measures that occurred *after* a focal treatment. Treatment could effect both the outcome and the post-treatment variable

Colliders

Age, marriage and happiness

- Assume happiness is fixed at birth
- Happy people are more likely than sad people to marry
- Living longer makes you more likely to marry

Our DAG

Despite there being no causal relationship, conditioning on M opens a pathway between H and A. This is collider bias.

Simulate some data

1. 20 people are born each year with uniformly distributed happiness values
2. Each person ages one year per year, happiness is unchanged
3. At 16, people can marry with probability proportional to happiness
4. No one divorces
5. At 65 they are removed

```
d <- sim_happiness( seed=1977 , N_years=1000 )
```

How does age relate to happiness?

$$H \sim N(\mu, \sigma)$$

$$\mu_i = \alpha_{Mi} + \beta_A A_i$$

$$\alpha_M \sim N(0, 1)$$

$$\beta_A \sim N(0, 2)$$

$$\sigma \sim Exp(1)$$

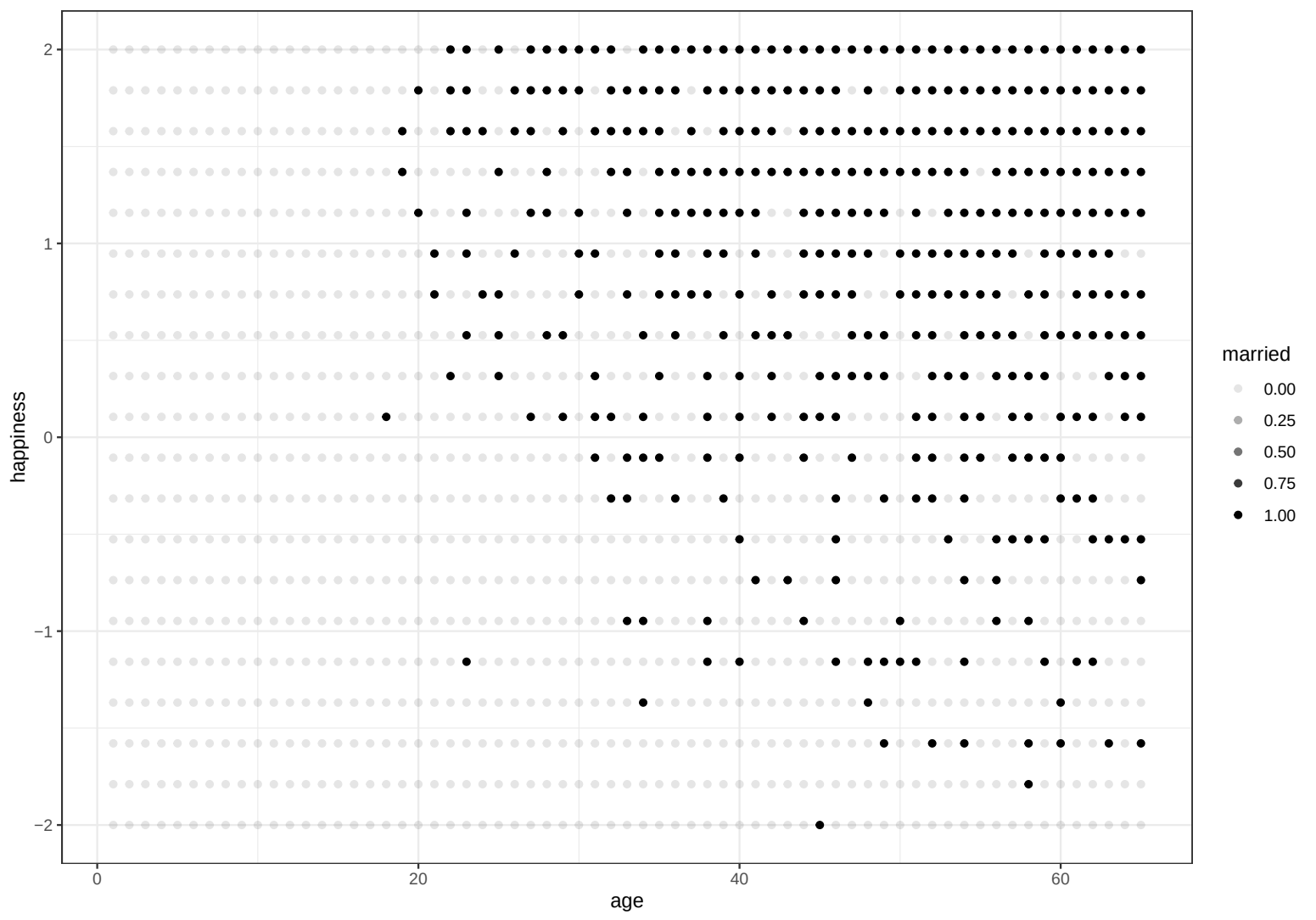
We'll consider two models - one with marriage as a predictor, and one without

Fit the models and compare β_A

```
##         mean   sd  5.5% 94.5%
## a[1]  -0.24 0.06 -0.34 -0.13
## a[2]   1.26 0.08  1.12  1.39
## bA    -0.75 0.11 -0.93 -0.57
## sigma  0.99 0.02  0.95  1.03
```

```
##       mean   sd  5.5% 94.5%
## a     0.00 0.08 -0.12  0.12
## bA    0.00 0.13 -0.21  0.21
## sigma 1.21 0.03  1.17  1.26
```

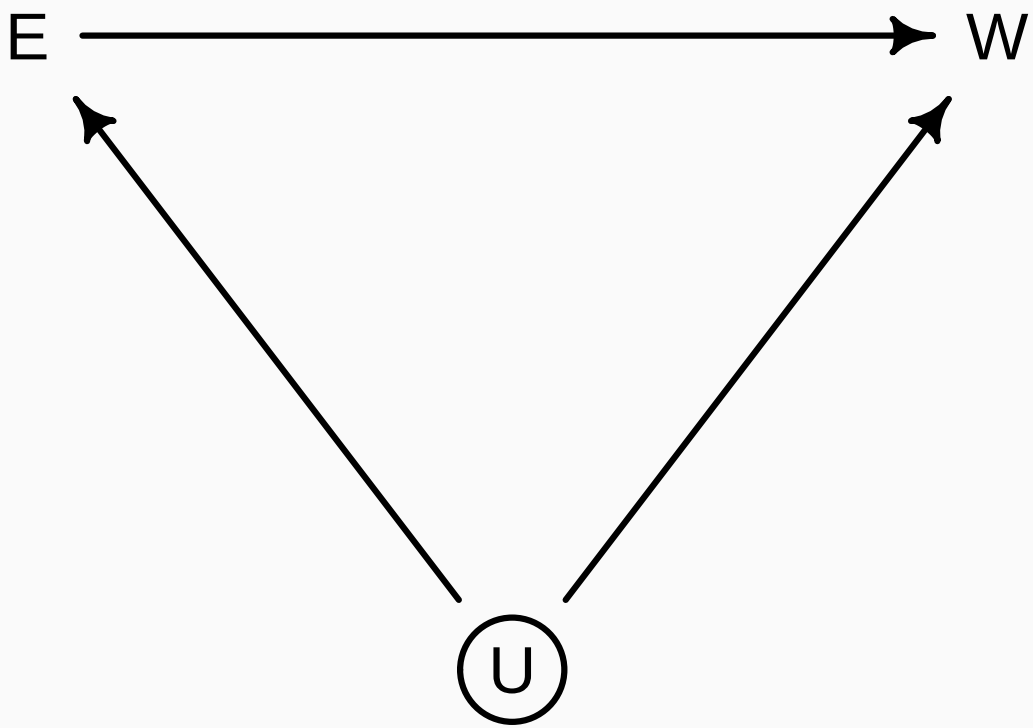
Marriage induces a correlation between age and happiness



Confounding

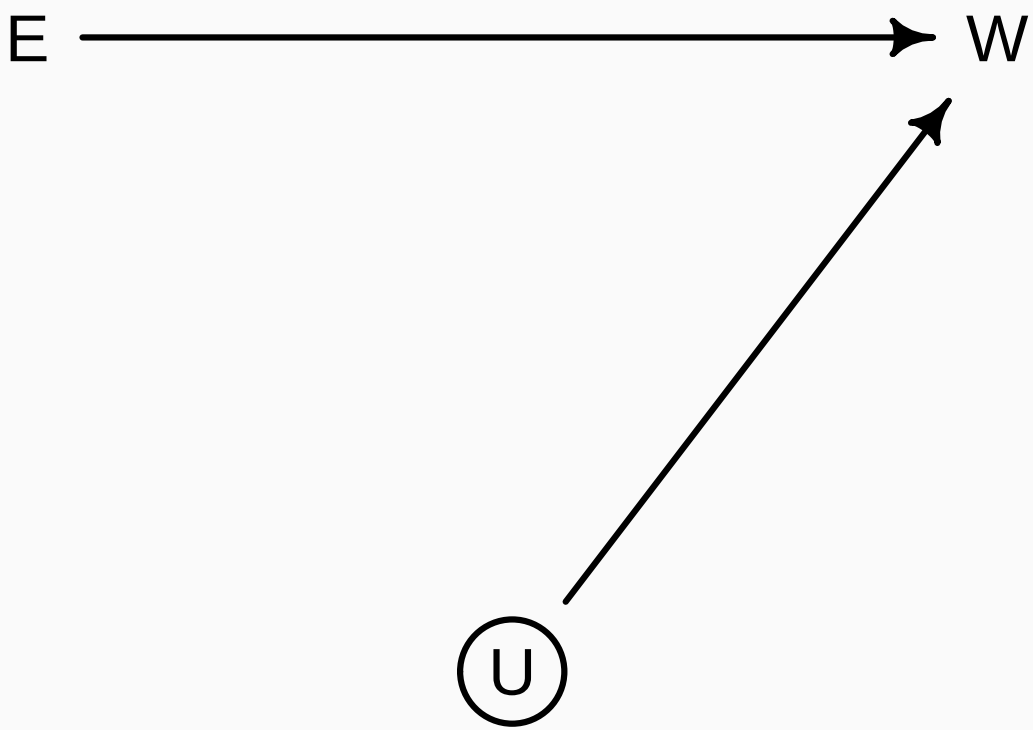
The classic example

An unobserved variable U confounds the relationship between education and wages



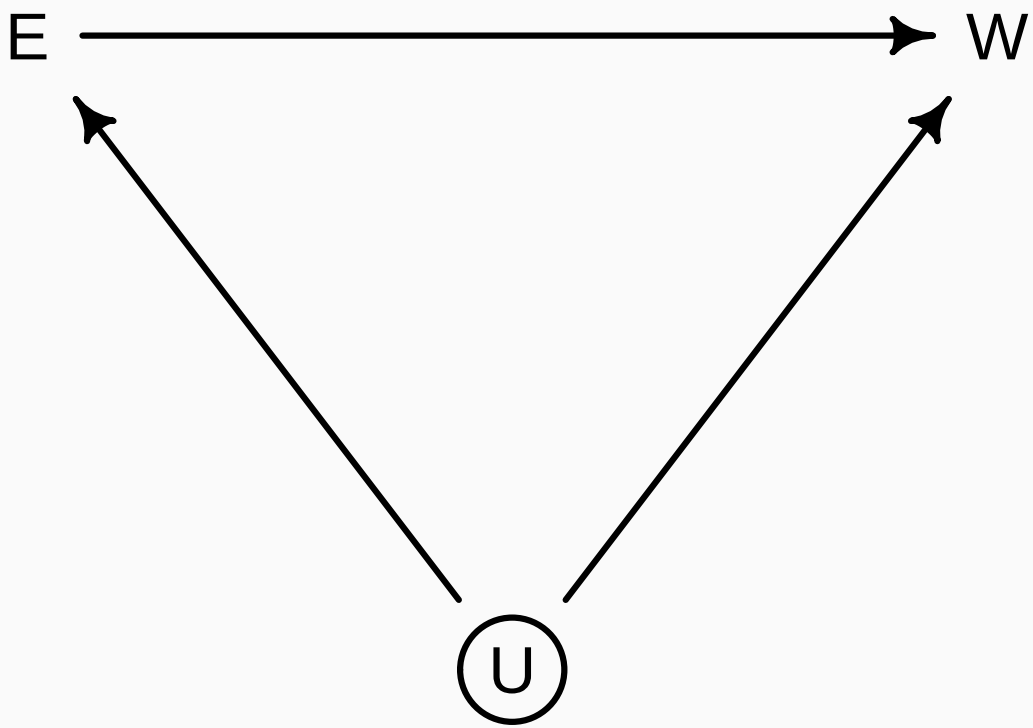
The classic solution: randomization

- Randomizing education (experimentally or otherwise) breaks the relationship between E and U.
- U still influences W, but because it no longer influences E, we can estimate the effect of E on W without bias.



The statistical solution: conditioning

By adding U to the model, we block the flow of information on $E \leftarrow U \rightarrow W$, leaving only the path $E \rightarrow W$



Searching for confounds, closing backdoors

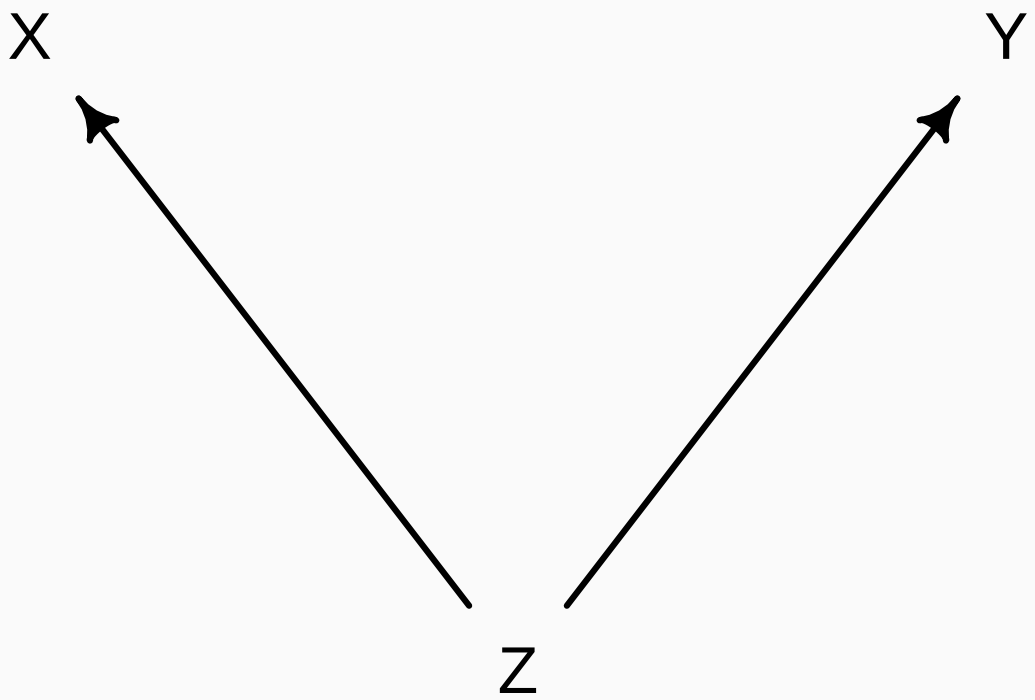
- Confounds are spurious correlations between some outcome Y and some predictor X
- In a DAG, we should pay careful attention to causal paths that enter the “back” of predictor X and connect to outcome Y

With DAGs, we have two general goals:

1. Close all backdoor paths between X and Y
2. Leave focal causal paths between X and Y open

The four basic DAGs: fork

- There is a backdoor path from X to Y, through Z
- X and Y are independent, conditional on Z

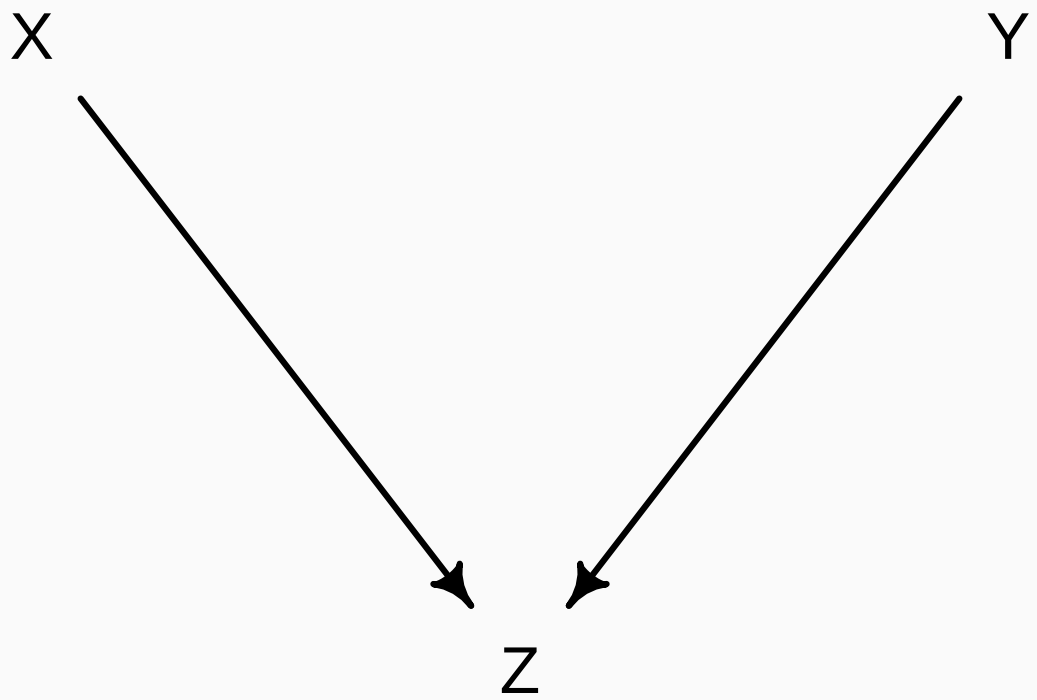


The four basic DAGs: pipe

- There is a causal path between X and Y through Z
- X and Y are independent, conditional on Z

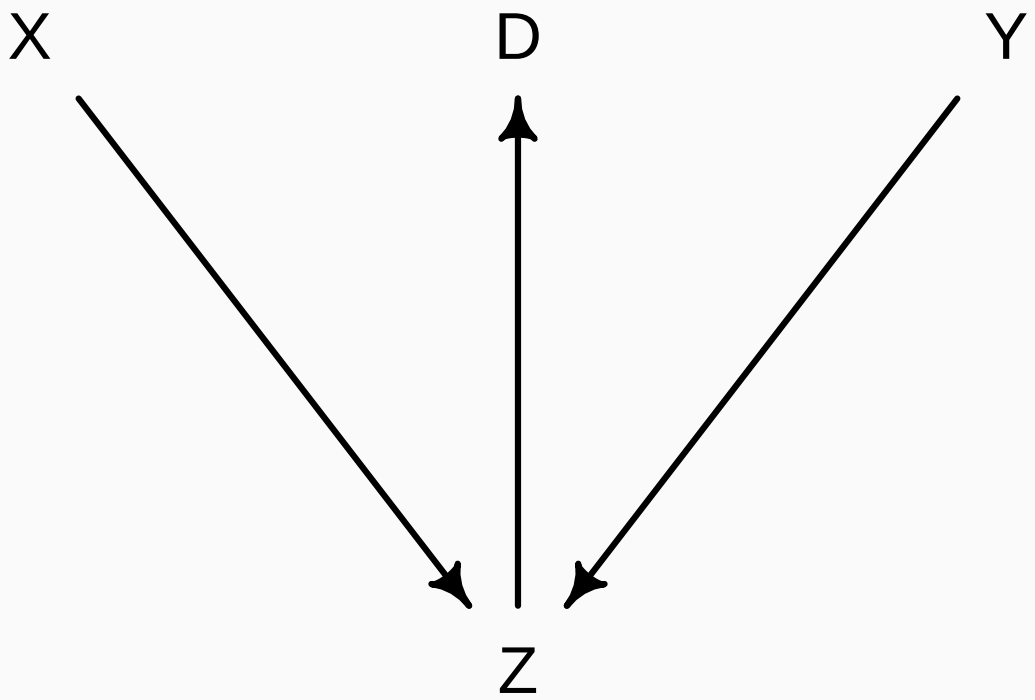
The four basic DAGs: collider

- No causal path between X and Y
- Conditioning on Z opens a path between X and Y



The four basic DAGs: descendant

- No causal path between X and Y
- Conditioning on D opens a weaker path between X and Y than conditioning on Z (a collider)



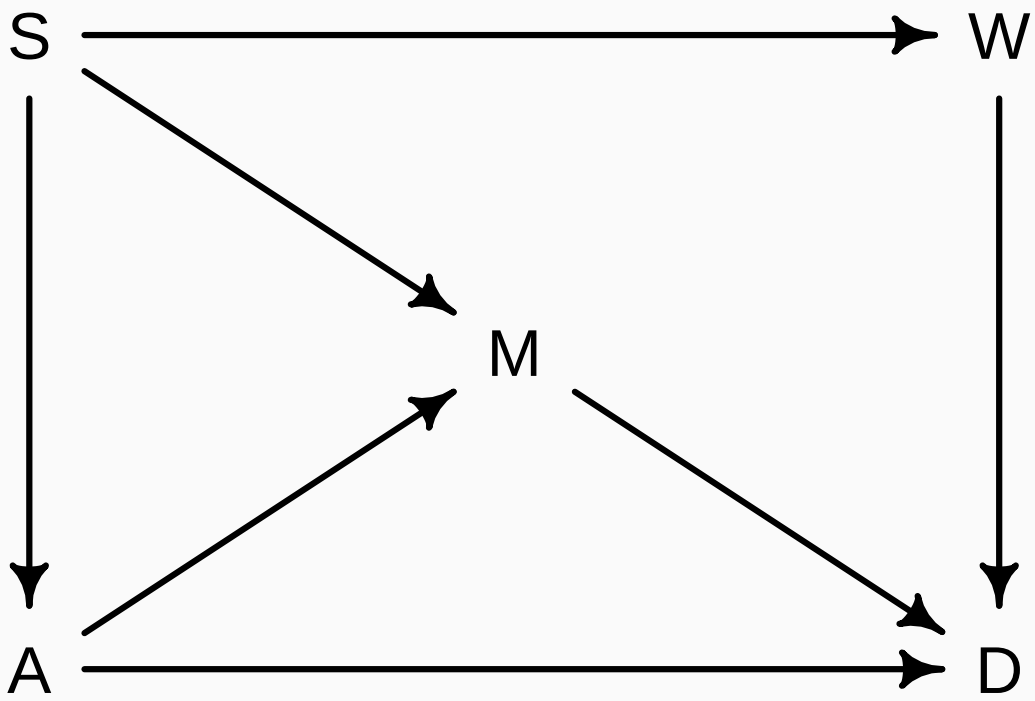
General method for causal inference using DAGs

1. List all paths connecting X (focal cause) and Y (focal effect)
2. Classify each path is open or closed. A path is open unless it contains a collider
3. Identify backdoor paths (arrow entering X)
4. For any open backdoor paths, decide which variables to condition on to close it


WAFFLE HOUSE

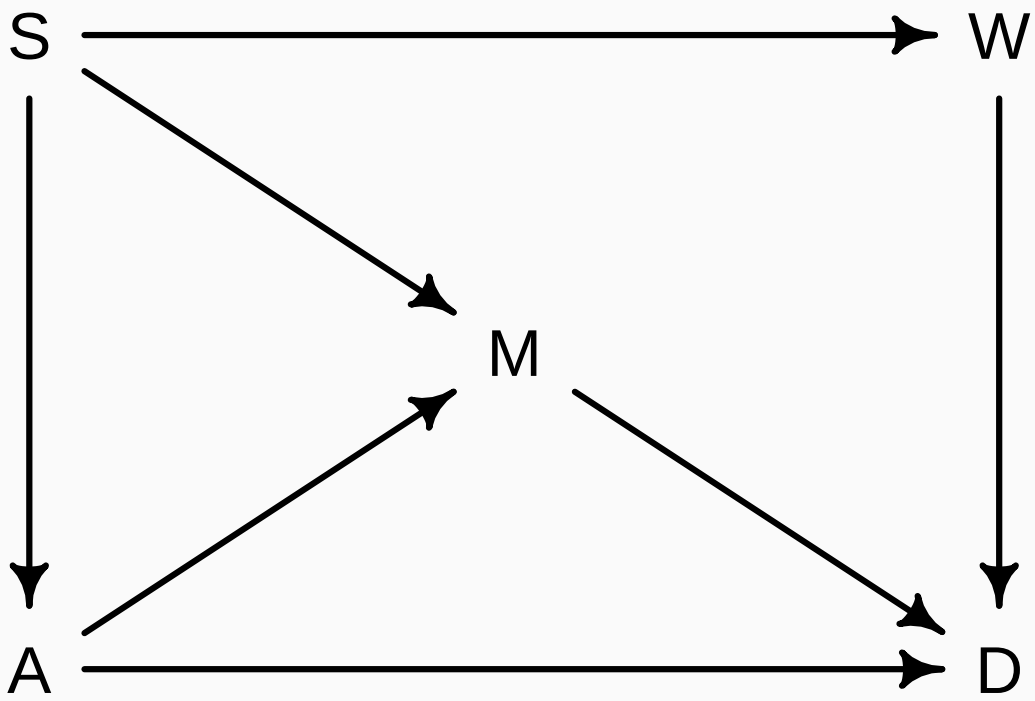
Does Waffle House effect divorce rates

S: State in the South; **W:** Waffle houses per capita; **M:** Marriage rates; **A:** Median age at first marriage; **D:** Divorce rates



Does Waffle House effect divorce rates

S: State in the South; **W:** Waffle houses per capita; **M:** Marriage rates; **A:** Median age at first marriage; **D:** Divorce rates



What causal relationships does this DAG assume?

General method for causal inference using DAGs

1. List all paths connecting W (focal cause) and D (focal effect)
2. Classify each path is open or closed. A path is open unless it contains a collider
3. Identify backdoor paths (arrow entering W)
4. For any open backdoor paths, decide which variables to condition on to close it

Checking our intuition

```
waffledag <- dagitty( "dag {
  A -> D
  A -> M -> D
  A <- S -> M
  S -> W -> D
  }")
adjustmentSets(waffledag, exposure = "W", outcome = "D")
```

```
##  { A, M }
##  { S }
```

Evaluating the causal effect, assuming the prior DAG

```
data("WaffleDivorce")
d<-WaffleDivorce %>%
  mutate(D = scale(Divorce),
         S = South + 1,
         A = scale(MedianAgeMarriage),
         W = scale(WaffleHouses / Population),
         M = scale(Marriage))

mWaffles<-quap(alist(
  D ~ dnorm(mu, sigma),
  mu <- a[S] + bW * W,
  a[S] ~ dnorm(0, 1),
  bW ~ dnorm(0, 0.5),
  sigma ~ dexp(1)
), data = d)

precis(mWaffles)
```

```
##             mean         sd        5.5%     94.5%
## bW    0.2245572 0.18044887 -0.06383495 0.5129494
## sigma 0.9074929 0.08956264  0.76435450 1.0506313
```

Do we believe this DAG?

We can check whether the assumptions of the DAG hold:

```
impliedConditionalIndependencies(waffledag)
```

```
## A _||_ W | S
## D _||_ S | A, M, W
## M _||_ W | S
```

These conditional independencies imply a series of regressions that we could estimate to test the validity of the DAG

Summary

- DAGs are powerful tools that we can use to clarify our thinking and develop statistical models
- DAGs are assumptions with testable implications
- We've just scratched the surface! Check out `http://www.dagitty.net/learn/` for more introductory materials
- Elwert and Winship (2014) provides a great review of DAGs and colliders: `https://doi.org/10.1146/annurev-soc-071913-043455`
- No homework this week, HW5 has been postponed: now due on 3/6
- More ggplot in lab on Friday (regular lecture time)