SCI 2025: Homework 4

Setup

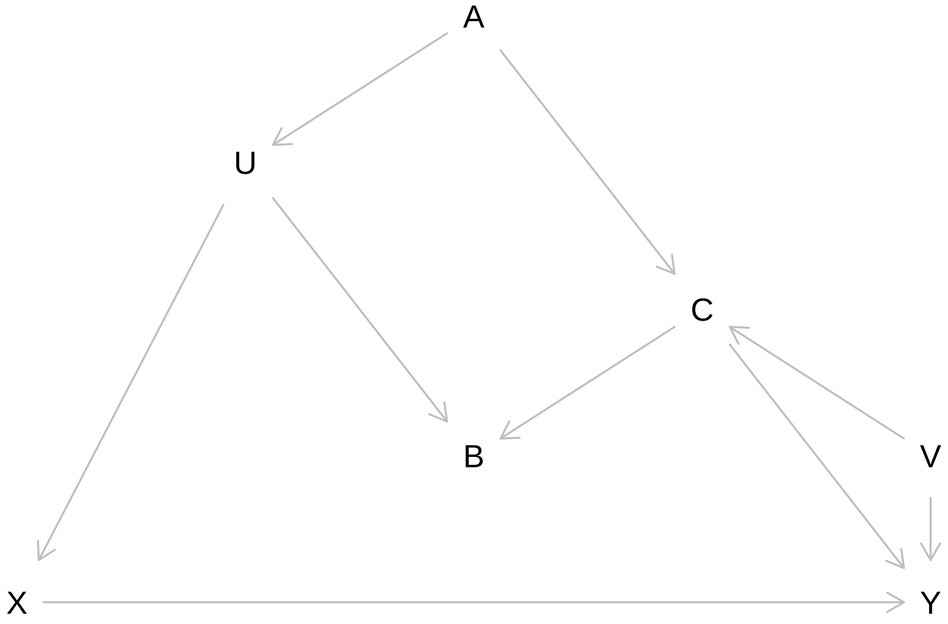
The three homework problems this week are from Chapter 6 of the textbook.

Question 1: 6M1

Modify the DAG on page 186 to include the variable V, an unobserved cause of C and Y: $C \leftarrow V \rightarrow Y$. Reanalyze the DAG. How many paths connect X to Y? Which must be closed? Which variables should you condition on now?

```r
library(dagitty)
dag <- dagitty("dag {
  A -> U
  A -> C
  U -> X
  U -> B
  C -> B
  C -> Y
  V -> C
  V -> Y
  X -> Y
}")
```

```r
# Set coordinates to better match the layout in the image
coordinates(dag) <- list(
  x = c(A = 0, U = -2, C = 2, B = 0, V = 4, X = -4, Y = 4),
  y = c(A = -4, U = -2, C = 0, B = 2, V = 2, X = 4, Y = 4)
)


plot(dag)
```



There are now 5 (instead of just 3) paths connecting X to Y:

1. $X \to Y$
2. $X \leftarrow U \to B \leftarrow C \to Y$
3. $X \leftarrow U \to B \leftarrow C \leftarrow V \to Y$
4. $X \leftarrow U \leftarrow A \to C \to Y$
5. $X \leftarrow U \leftarrow A \to C \leftarrow V \to Y$

2 and 3 are closed already (colliders); 4 and 5 need to be closed. The only valid adjustment set is A alone.

```r
adjustmentSets(dag, exposure = "X", outcome = "Y") # note that A is the only valid adjus
```

```
{ C, V }
{ A }
{ U }
```

```r
# because it doesn't include unobserved variables
```

Question 2: 6M2

Sometimes, in order to avoid multicollinearity, people inspect pairwise correlations among predictors before including them in a model. This is a bad procedure, because what matters is the conditional association, not the association before the variables are included in the model. To highlight this, consider the DAG $X \to Z \to Y$. Simulate data from this DAG so that the correlation between X and Z is very large. Then include both in a model prediction Y. Do you observe any multicollinearity? Why or why not? What is different from the legs example in the chapter?

```r
N <- 500
X <- rnorm(N, 0, 1) # going to the gym
Z <- rnorm(N, X*2.5, 1) # exercising
Y <- rnorm(N, Z, 1) # calories burned

cor(X, Z)
```

```
[1] 0.9313127
```

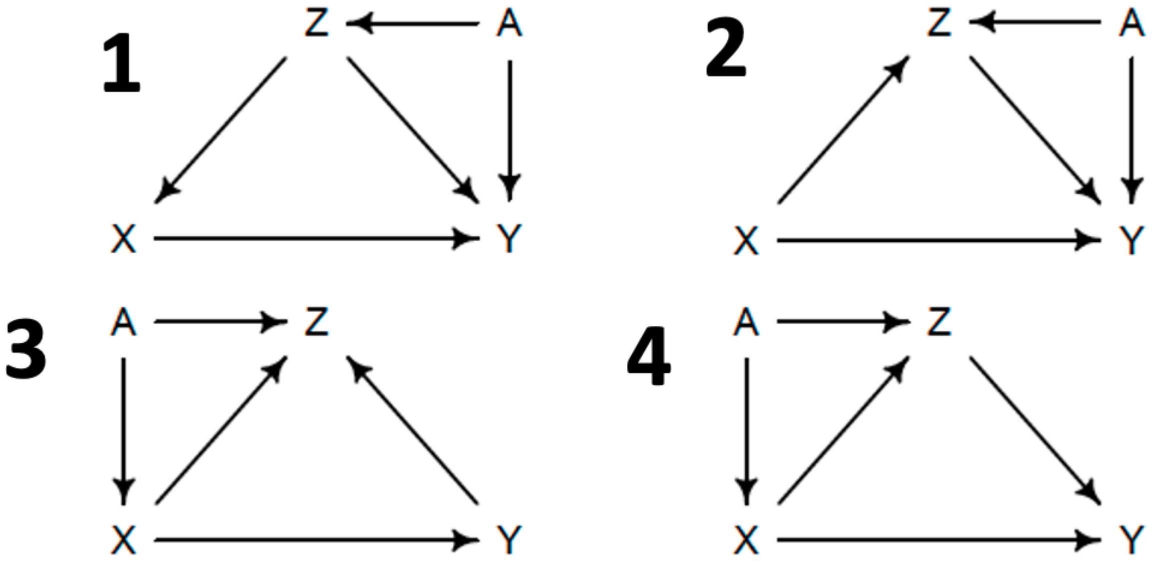
```r
d <- data.frame(X, Z, Y)

library(rethinking)

model <- quap(
  alist(
    Y ~ dnorm(mu, sigma),
    mu <- a + bX*X + bZ*Z,
    a ~ dnorm(0, 1),
    bX ~ dnorm(0, 1),
    bZ ~ dnorm(0, 1),
    sigma ~ dexp(1)
  ),
  data = d
)

precis(model)
```

```
             mean         sd        5.5%     94.5%
a     0.055589077 0.04379295 -0.01440051 0.1255787
bX    0.004807632 0.11754073 -0.18304516 0.1926604
bZ    0.983474789 0.04399607  0.91316057 1.0537890
sigma 0.979251250 0.03092124  0.92983314 1.0286694
```

The parameter estimates look fine (lack of multicollinearity), despite high correlation between X and Z. The difference here is: (1) mediation rather than common causes of Y and (2) there is not a perfect correlation between X and Z.

Question 3: 6M3

Learning to analyze DAGs requires practice. For each of the four DAGs below, state which variables, if any, you must adjust for (condition on) to estimate the total causal influence of X on Y.



1. Adjust for Z (Z is a confounder)
2. No adjustment (Z is a mediator/collider)
3. No adjustment (Z is a collider)
4. Adjust for A (Z is a mediator, A creater a backdoor path between X and Y)

Key to this question: the direction of the arrow matters a lot!

Question 4

Can you think of potential examples of collider bias from your own field or a related literature? Do you think the bias is likely to be positive or negative (with respect to the causal effect of interest)? Positive bias would mean that the estimated relationship between the focal variable and the outcome is too strong, while negative bias would mean that the estimated relationship is too weak.

Example from evolutionary biology: survival bias in comparative studies. If two traits both reduce the likelihood that a species will go extinct, this can induce a spurious positive association between the traits if we only look at extant species.