homework_6

Christopher Huong

2024-03-02

```
library(rethinking)
library(knitr)
library(dagitty)
```

6E1. List three mechanisms by which multiple regression can produce false inferences about causal effects.

Multicollinearity) Including variables that are highly associated conditional on other model variables (and thus redundant) in a multiple regression may suppress true causal effects

Post-treatment bias) Including variables that are caused by the treatment will suppress the estimated effect of the treatment, because the information provided by the mechanism of the treatment is partialled out.

Collider bias) Including a collider in the model as a predictor can induce a spurious association between that predictor and the dependent variable, when no causal relationship exists.

6E2. For one of the mechanisms in the previous problem, provide an example of your choice, perhaps from your own research.

```
set.seed(1)
n = 1000
diet <- rlnorm(n, 0, 0.5) #kcal deficit per day
coca <- rexp(n, 1) #snorts of cocaine per day
weightloss <- rnorm(n, (diet+coca+5), .5) #weight loss (lbs) per week
```

```
d <- data.frame(diet=standardize(diet), coca=standardize(coca), weightloss=standardize(weightlos
s))

cor(d)
```

```
##                  diet       coca weightloss
## diet       1.00000000 0.01929596  0.4986723
## coca       0.01929596 1.00000000  0.7890876
## weightloss 0.49867230 0.78908755  1.0000000
```

```
summary(d)
```

```
##       diet              coca           weightloss
##  Min.   :-1.4537   Min.   :-0.9929   Min.   :-2.0692
##  1st Qu.:-0.6847   1st Qu.:-0.7270   1st Qu.:-0.6779
##  Median :-0.2443   Median :-0.3205   Median :-0.1859
##  Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.0000
##  3rd Qu.: 0.4372   3rd Qu.: 0.3647   3rd Qu.: 0.4470
##  Max.   : 8.8834   Max.   : 6.1039   Max.   : 4.6511
```

```
m1 <- quap(
  alist(
    coca ~ dnorm(mu, sigma),
    mu <- a + b_diet*diet + b_weightloss*weightloss,
    a ~ dnorm(0, 0.2),
    b_diet ~ dnorm(0, 0.5),
    b_weightloss ~ dnorm(0, 0.5),
    sigma ~ dexp(1)
  ),data=d
)
```

```
precis(m1)
```

```
##                       mean          sd        5.5%      94.5%
## a              1.361542e-07 0.013776785 -0.02201783  0.0220181
## b_diet        -4.970324e-01 0.015929795 -0.52249126 -0.4715735
## b_weightloss   1.036152e+00 0.015929836  1.01069330  1.0616112
## sigma          4.366975e-01 0.009761813  0.42109623  0.4522988
```

The model estimates a negative effect of diet on cocaine use, when no such causal effect actually exists. This demonstrates collider bias, as weightloss is a common cause of diet and cocaine use and was conditioned on in this model.

6E3. List the four elemental confounds. Can you explain the conditional dependencies of each?

The fork: X <- Z -> Y
Z is a common cause of X and Y, thus X is conditionally independent of Y given Z
The pipe: X -> Z -> Y
X effects Y through Z, thus X and Z are conditionally independent given Z
Collider: X -> Z <- Y
Z is a common cause of X and Y, thus X and Y are conditionally dependent given Z

Descendent: X -> Z <- Y, Z -> D
Z is a common cause of X and Y, and D is a cause of Z. Thus conditioning on D weakly conditions on Z, inducing a spurious association between X and Y

6E4. How is a biased sample like conditioning on a collider? Think of the example at the open of the chapter.

If Z is influenced by 2 independent varriables (X -> Z <- Y)
selecting from certain values of Z (i.e., a biased sample) will result in an association between the 2 variables in the sample.

6M1. Modify the DAG on page 186 to include the variable V, an unobserved cause of C and Y: C ← V → Y. Reanalyze the DAG. How many paths connect X to Y? Which must be closed? Which variables should you condition on now?

1. X <- U <- A -> C -> Y (open)
2. X <- U <- A -> C <- V -> Y (C = collider -> closed)
3. X <- U -> B <- C <- V -> Y (B and C = colliders -> closed)
4. X <- U -> B <- C -> Y (B = collider -> closed)

Therefore just need to condition on A, and should not condition on B or C

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 → Z → 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?

DAG: X -> Z -> Y

```
set.seed(1)
X <- rnorm(n, 0, 1)
Z <- rnorm(n, X, .5)
Y <- rnorm(n, Z, 1)
```

```
d2 <- data.frame(X=X, Z=Z, Y=Y)
cor(d2)
```

```
##           X         Z         Y
## X 1.0000000 0.8941315 0.6832559
## Z 0.8941315 1.0000000 0.7634258
## Y 0.6832559 0.7634258 1.0000000
```

```
m2 <- quap(
  alist(
    Y ~ dnorm(mu, sigma),
    mu <- a + bZ*Z + bX*X,
    a ~ dnorm(0, 0.2),
    c(bZ, bX) ~ dnorm(0, .5),
    sigma ~ dexp(1)
  ),
  data=d2
)
precis(m2)
```

```
##             mean         sd        5.5%      94.5%
## a     0.01569440 0.03211061 -0.03562456 0.06701336
## bZ    1.02823818 0.06162944  0.92974244 1.12673393
## bX    0.02077481 0.06911950 -0.08969149 0.13124112
## sigma 1.02858528 0.02298351  0.99185319 1.06531737
```

No multicollinearity was observed as Y was not causally related to both X and Z. The correlation between X and Y was through Z, thus including X and Z in the model partials out the influence of Z on Y, leaving no association between X on Y. In the leg example, each leg were a function of height, and thus both causally related to height.

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.

Left to right, top down

1. Condition on A and Z (common influences of X and Y)
2. None. No open door from A to X, and Z is a collider
3. None. No open door from A to Y, and Z is a collider
4. Condition on A (common influence of X and Y). No open door from Z to X

6H1. Use the Waffle House data, data(WaffleDivorce), to find the total causal influence of number of Waffle Houses on divorce rate. Justify your model or models with a causal graph

```
data("WaffleDivorce"); d3 <- WaffleDivorce
```

```
d3 <- d3[, c("MedianAgeMarriage", "Marriage", "Divorce", "WaffleHouses", "South")]

colnames(d3) <- c("A", "M", "D", "W", "S")
for (i in c("A", "M", "D", "W", "S")) d3[[i]] <- standardize(d3[[i]])
```

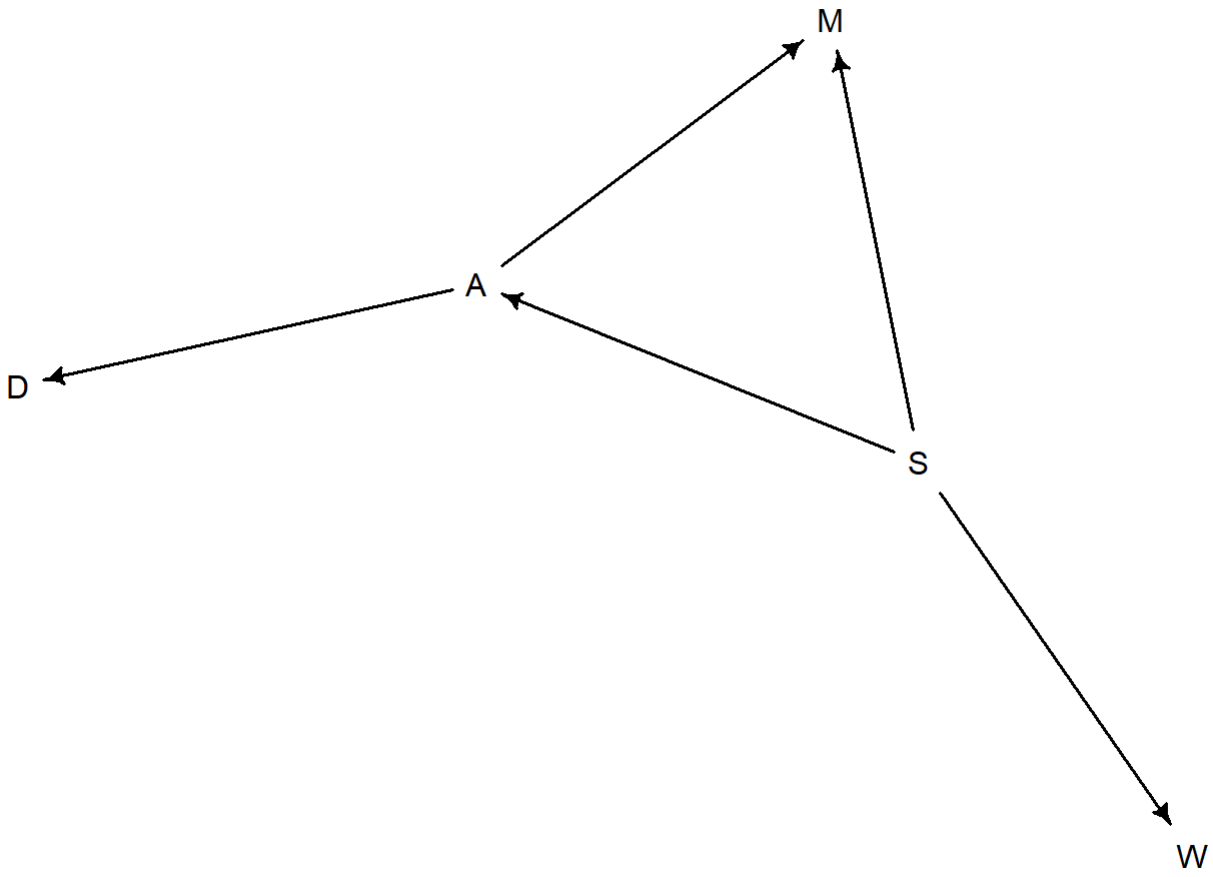
Proposed DAG:

W <- S -> A
D <- A -> M <- S

Waffle houses are more common in the south.
Lower age of marriage, and higher marriage rates associated with the South due to culture
Lower age of marriage causes higher marriage rates because there are more opportunities to marry (and remarry) if people tend to start younger. Lower age of marriage is also associated with divorce rate, because there is less probability the marriage was carefully thought out.

```
dag <- dagitty("dag{
               W <- S -> A;
               D <- A -> M <- S}")

drawdag(dag)
```



6H2. Build a series of models to test the implied conditional independencies of the causal graph you used in the previous problem. If any of the tests fail, how do you think the graph needs to be amended? Does the graph need more or fewer arrows? Feel free to nominate variables that aren't in the data

```
impliedConditionalIndependencies(dag)
```

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

1. S is a common cause of W and A
2. A is a common cause of D and M
3. A is a mediator of S on D
4. S is a common cause of W and D (via a pipe)
5. You can also condition on the pipe (A) to block the path from D to W
6. S is a common cause of M and W

Test all these implied conditional independencies

```
# A ~ W + S
ici1 <- quap(alist(
  A ~ dnorm(mu, sigma),
  mu <- a + bW*W + bS*S,
  a ~ dnorm(0,0.2),
  c(bW, bS) ~ dnorm(0, 0.5),
  sigma ~ dexp(1)
), data=d3)

# D ~ M + A
ici2 <- quap(alist(
  D ~ dnorm(mu, sigma),
  mu <- a + bM*M + bA*A,
  a ~ dnorm(0, 0.2),
  c(bM, bA) ~ dnorm(0, 0.5),
  sigma ~ dexp(1)
),data=d3)

# D ~ S + A
ici3 <- quap(alist(
  D ~ dnorm(mu, sigma),
  mu <- a + bS*S + bA*A,
  a ~ dnorm(0, 0.2),
  c(bS, bA) ~ dnorm(0, 0.5),
  sigma ~ dexp(1)
),data=d3)

# D ~ W + S
ici4 <- quap(alist(
  D ~ dnorm(mu, sigma),
  mu <- a + bW*W + bS*S,
  a ~ dnorm(0, 0.2),
  c(bW, bS) ~ dnorm(0, 0.5),
  sigma ~ dexp(1)
),data=d3)

# D ~ W + A
ici5 <- quap(alist(
  D ~ dnorm(mu, sigma),
  mu <- a + bW*W + bA*A,
  a ~ dnorm(0, 0.2),
  c(bW, bA) ~ dnorm(0, 0.5),
  sigma ~ dexp(1)
),data=d3)

# M ~ W + S
ici6 <- quap(alist(
  M ~ dnorm(mu, sigma),
  mu <- a + bW*W + bS*S,
  a ~ dnorm(0, 0.2),
  c(bW, bS) ~ dnorm(0, 0.5),
```

```
  sigma ~ dexp(1)
),data=d3)
```

```
impliedConditionalIndependencies(dag)
```

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

```
ici_list <- list(ici1, ici2, ici3, ici4, ici5, ici6)

lapply(ici_list, function(a) precis(a))
```

```
## [[1]]
##        mean   sd  5.5% 94.5%
## a      0.00 0.11 -0.18  0.18
## bW     0.06 0.17 -0.21  0.34
## bS    -0.27 0.17 -0.54  0.00
## sigma  0.95 0.09  0.80  1.10
##
## [[2]]
##        mean   sd  5.5% 94.5%
## a      0.00 0.10 -0.16  0.16
## bM    -0.07 0.15 -0.31  0.18
## bA    -0.61 0.15 -0.85 -0.37
## sigma  0.79 0.08  0.66  0.91
##
## [[3]]
##        mean   sd  5.5% 94.5%
## a      0.00 0.09 -0.15  0.15
## bS     0.21 0.11  0.03  0.38
## bA    -0.52 0.11 -0.70 -0.35
## sigma  0.76 0.08  0.64  0.88
##
## [[4]]
##       mean   sd  5.5% 94.5%
## a     0.00 0.11 -0.17  0.17
## bW    0.05 0.17 -0.21  0.32
## bS    0.29 0.17  0.02  0.55
## sigma 0.92 0.09  0.78  1.07
##
## [[5]]
##        mean   sd  5.5% 94.5%
## a      0.00 0.10 -0.15  0.15
## bW     0.18 0.11  0.01  0.35
## bA    -0.55 0.11 -0.72 -0.38
## sigma  0.77 0.08  0.65  0.89
##
## [[6]]
##        mean   sd  5.5% 94.5%
## a      0.00 0.11 -0.18  0.18
## bW    -0.04 0.17 -0.32  0.24
## bS     0.10 0.17 -0.18  0.38
## sigma  0.98 0.10  0.82  1.13
```

1. True; no effect of W on A given S
2. True; no effect of M on D given A
3. Maybe; seems to be some small effect of S on D given A
4. True; no effect of W on D given S
5. Maybe; seems to be some small effect of W on D given A
6. True; no effect of W on M given S

Overall, the conditional dependencies implied by the causal model (DAG) seem to mostly supported by the data

```
adjustmentSets(dag, exposure="W", outcome="D")
```

```
## { A }
## { S }
```

Condition on A or S to estimate the causal effect of Wafflehouses on divorce

```
m3 <- quap(
  alist(
    D ~ dnorm(mu, sigma),
    mu <- a + bW*W + bA*A,
    a ~ dnorm(0, 0.2),
    c(bW, bA) ~ dnorm(0, 0.5),
    sigma ~ dexp(1)
  ),
  data=d3)
precis(m3)
```

```
##                mean         sd         5.5%      94.5%
## a      2.131324e-06 0.09536449 -0.152408739  0.1524130
## bW     1.808148e-01 0.10774579  0.008616198  0.3530134
## bA    -5.494698e-01 0.10784987 -0.721834721 -0.3771049
## sigma  7.671548e-01 0.07593281  0.645799473  0.8885101
```

Conditioning on A reveals no causal effect of W on D

```
m4 <- quap(
  alist(
    D ~ dnorm(mu, sigma),
    mu <- a + bW*W + bS*S,
    a ~ dnorm(0, 0.2),
    c(bW, bS) ~ dnorm(0, 0.5),
    sigma ~ dexp(1)
  ),
  data=d3)
precis(m4)
```

```
##               mean         sd        5.5%     94.5%
## a     3.830832e-07 0.10911700 -0.17438966 0.1743904
## bW    5.241293e-02 0.16588049 -0.21269613 0.3175220
## bS    2.892012e-01 0.16594814  0.02398406 0.5544184
## sigma 9.206714e-01 0.09087318  0.77543851 1.0659043
```

Conditioning on S reveals a neglible causal effect of W on D

6H3. Use a model to infer the total causal influence of area on weight. Would increasing the area available to each fox make it heavier (healthier)? You might want to standardize the variables. Regardless, use prior predictive simulation to show that your model's prior predictions stay within the possible outcome range

```
rm(list=ls())
data("foxes"); d <- foxes
```

Standardize continuous variables

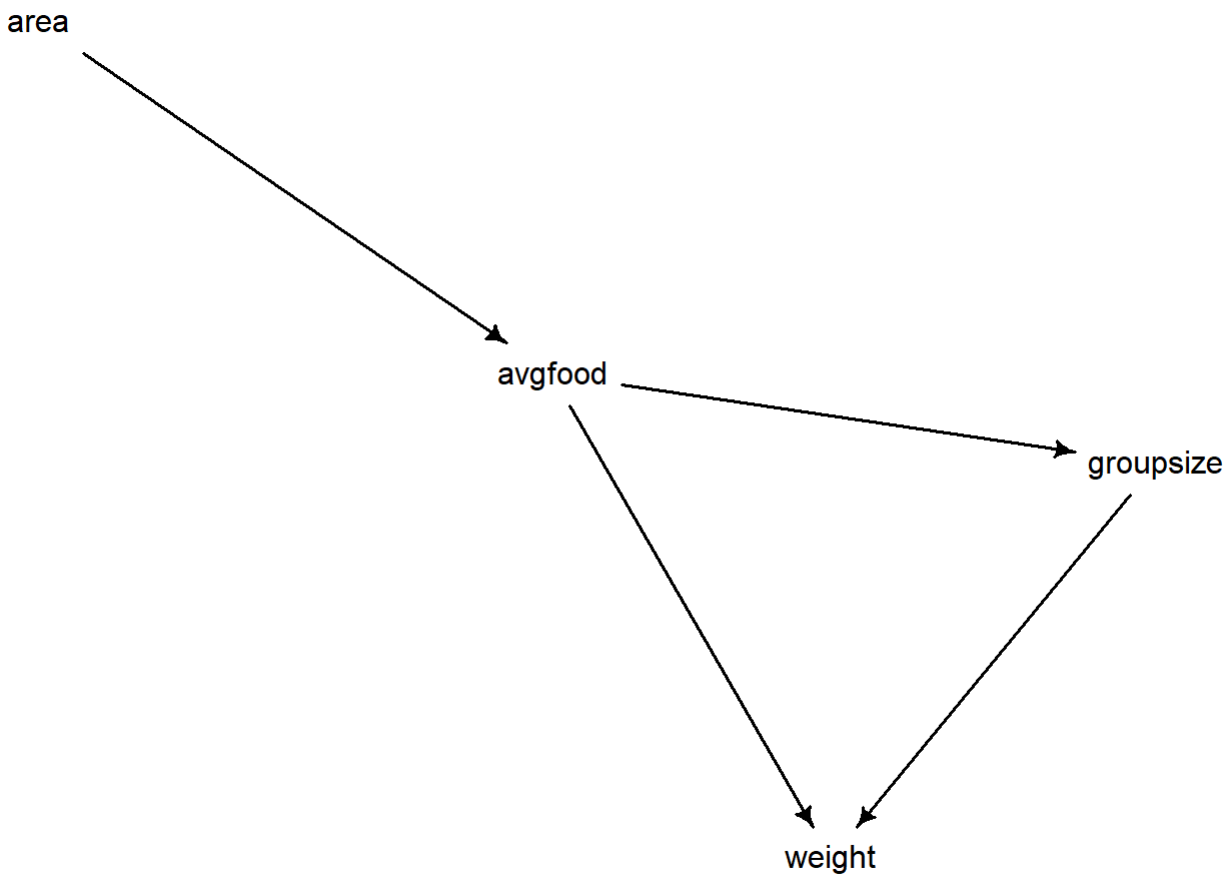
```
for (i in c("avgfood", "groupsize", "area", "weight")){d[[i]] <- standardize(d[[i]])}
summary(d)
```

```
##      group          avgfood           groupsize            area         
##  Min.   : 1.00   Min.   :-1.92483   Min.   :-1.5241   Min.   :-2.23960  
##  1st Qu.:11.75   1st Qu.:-0.46252   1st Qu.:-0.8741   1st Qu.:-0.62383  
##  Median :18.00   Median :-0.08433   Median :-0.2241   Median :-0.04216  
##  Mean   :17.21   Mean   : 0.00000   Mean   : 0.0000   Mean   : 0.00000  
##  3rd Qu.:24.00   3rd Qu.: 0.24343   3rd Qu.: 0.4258   3rd Qu.: 0.64993  
##  Max.   :30.00   Max.   : 2.31084   Max.   : 2.3758   Max.   : 2.04756  
##      weight       
##  Min.   :-2.20406  
##  1st Qu.:-0.68382  
##  Median :-0.09261  
##  Mean   : 0.00000  
##  3rd Qu.: 0.71396  
##  Max.   : 2.55092
```

Draw given DAG

```
dag2 <- dagitty("dag{
               area -> avgfood -> weight <- groupsize
               avgfood -> groupsize
               }")

drawdag(dag2)
```



The total causal effect of area on weight will include avgfood and groupsize. There are no observed confounders that must be conditioned on.

```
m5 <- quap(alist(
  weight ~ dnorm(mu, sigma),
  mu <- a + bA*area,
  a ~ dnorm(0, 0.2),
  bA ~ dnorm(0, 0.5),
  sigma ~ dexp(1)
),data=d)
precis(m5)
```

```
##                mean         sd       5.5%     94.5%
## a     -6.277041e-08 0.08360866 -0.1336228 0.1336227
## bA     1.883357e-02 0.09089580 -0.1264355 0.1641026
## sigma  9.912659e-01 0.06466645  0.8879164 1.0946153
```

No total causal effect of area on weight.

6H4. Now infer the causal impact of adding food to a territory. Would this make foxes

heavier? Which covariates do you need to adjust for to estimate the total causal influence of food?

To estimate the direct causal effect of avgfood to weight, we need to control for the indirect effect through groupsize. Since we are estimating the total causal effect, there is no need to control for groupsize.

```
m6 <- quap(alist(
  weight ~ dnorm(mu, sigma),
  mu <- a + bF*avgfood,
  a ~ dnorm(0, 0.2),
  bF ~ dnorm(0, 0.5),
  sigma ~ dexp(1)
),data=d)

precis(m6)
```

```
##                mean         sd       5.5%     94.5%
## a      2.150470e-08 0.08360017 -0.1336092 0.1336092
## bF    -2.421158e-02 0.09088502 -0.1694634 0.1210402
## sigma  9.911440e-01 0.06465859  0.8878071 1.0944809
```

There seems to be no total causal effect of avgfood on weight, though this may be masked through the indirect effect of groupsize.

6H5. Now infer the causal impact of group size. Which covariates do you need to adjust for? Looking at the posterior distribution of the resulting model, what do you think explains these data? That is, can you explain the estimates for all three problems? How do they go together?

To infer the causal effect of groupsize on weight, we need to control for avgfood, which is a common cause of groupsize and avgfood, and thus a confounder.

```
m7 <- quap(alist(
  weight ~ dnorm(mu, sigma),
  mu <- a + bG*groupsize + bF*avgfood,
  a ~ dnorm(0, 0.2),
  c(bG, bF) ~ dnorm(0, 0.5),
  sigma ~ dexp(1)
),data=d)

precis(m7)
```

```
##                mean         sd       5.5%      94.5%
## a     -4.568920e-07 0.08013217 -0.1280671  0.1280662
## bG    -5.735885e-01 0.17912888 -0.8598711 -0.2873060
## bF     4.772819e-01 0.17911079  0.1910282  0.7635355
## sigma  9.419613e-01 0.06173904  0.8432904  1.0406322
```

We estimate a negative direct causal effect of groupsize on weight. To explain these results, lets look at the direct effect of avgfood on groupsize

Lets look at the direct effect of

```
m8 <- quap(alist(
  groupsize ~ dnorm(mu, sigma),
  mu <- a + bF*avgfood,
  a ~ dnorm(0, 0.2),
  bF ~ dnorm(0, 0.5),
  sigma ~ dexp(1)
),
data=d)

precis(m8)
```

```
##               mean         sd        5.5%     94.5%
## a     4.205710e-08 0.03916796 -0.06259792 0.0625980
## bF    8.957171e-01 0.03999322  0.83180016 0.9596339
## sigma 4.301819e-01 0.02816843  0.38516331 0.4752005
```

There is a large positive effect of avgfood on groupsize, and a medium positive effect of avg food on groupsize, which in turn has a medium negative effect on weight. Thus, the indirect path from avgfood to weight through groupsize is a masking effect.

6H6. Consider your own research question. Draw a DAG to represent it. What are the testable implications of your DAG? Are there any variables you could condition on to close

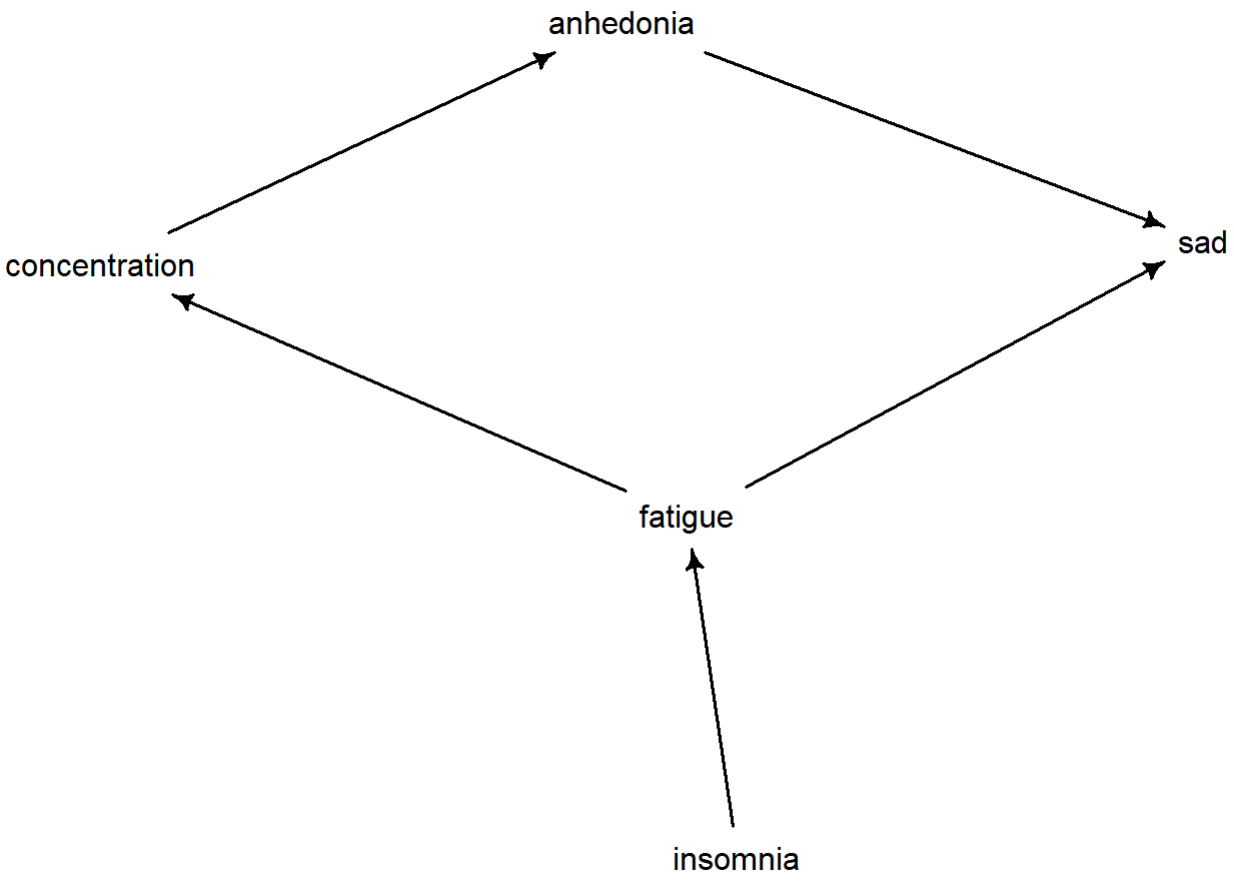
all backdoor paths? Are there unobserved variables that you have omitted? Would a reasonable colleague imagine additional threats to causal inference that you have ignored?

A hypothesized causal graph among depression symptoms.

```
rm(list=ls())
dag <- dagitty("dag{
               insomnia -> fatigue -> sad
               fatigue -> concentration -> anhedonia -> sad

               }")

drawdag(dag)
```




```
impliedConditionalIndependencies(dag)
```

```
## anhd _||_ fatg | cncn
## anhd _||_ insm | fatg
## anhd _||_ insm | cncn
## cncn _||_ insm | fatg
## cncn _||_ sad | anhd, fatg
## insm _||_ sad | fatg
```

1. concentration mediates the effect of fatigue on anhedonia
   2 & 3) fatigue & concentration mediates the effect of insomnia on anhedonia
2. fatigue mediates the effect of insomnia on concentration
3. anhedonia mediates the effect of concentration on sadness. fatigue is a common cause of concentration and sadness
4. fatigue mediates the effect of insomnia on sadness

There are very likely to be unobserved confounders, such as physical illness causing fatigue & concentration issues. A colleague would point out the many potential unobserved confounders and colliders, and that the acyclic assumption is almost certainly violated in reality.