Assignment 2 - Methods 4

study group 18

2025-03-18

Second assignment

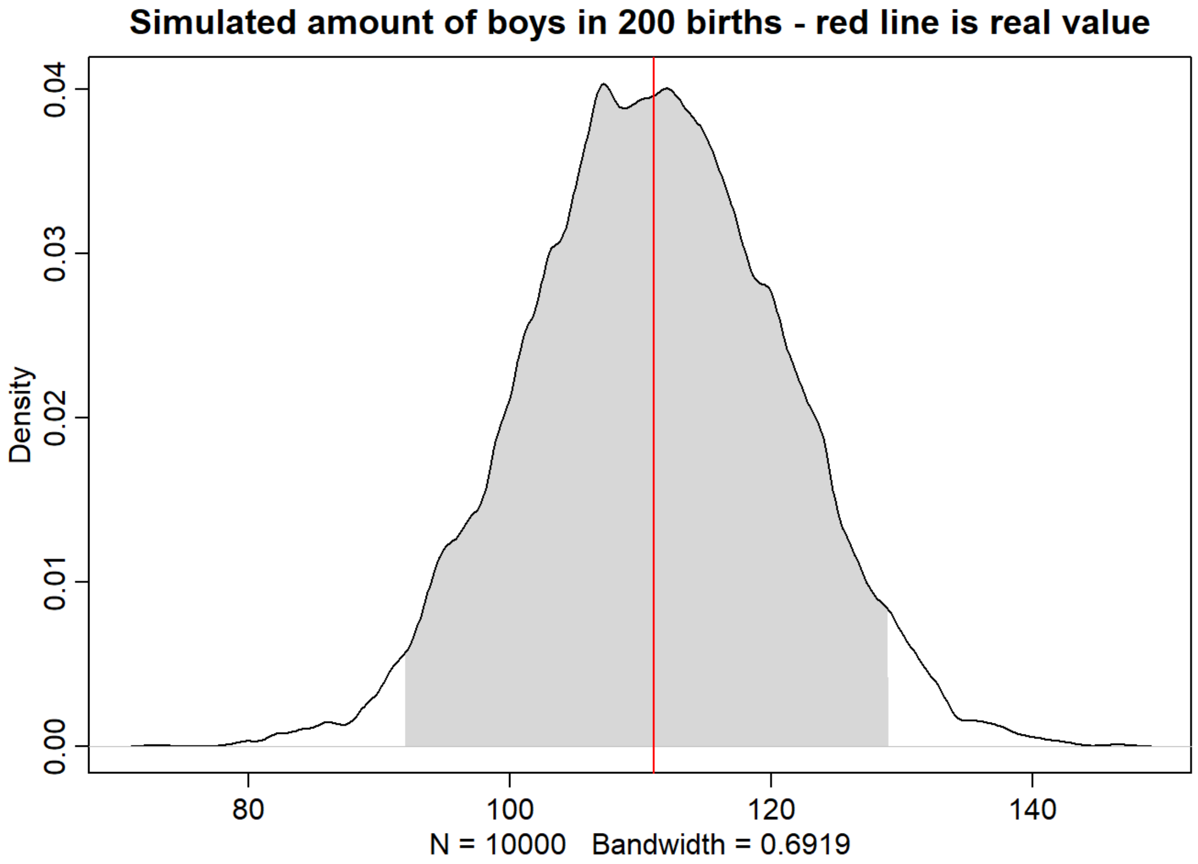
The second assignment uses chapter 3, 5 and 6. The focus of the assignment is getting an understanding of causality.

Chapter 3: Causal Confussion - by Regitze Hammer Holt

Reminder: We are tying to estimate the probability of giving birth to a boy I have pasted a working solution to questions 6.1-6.3 so you can continue from here:)

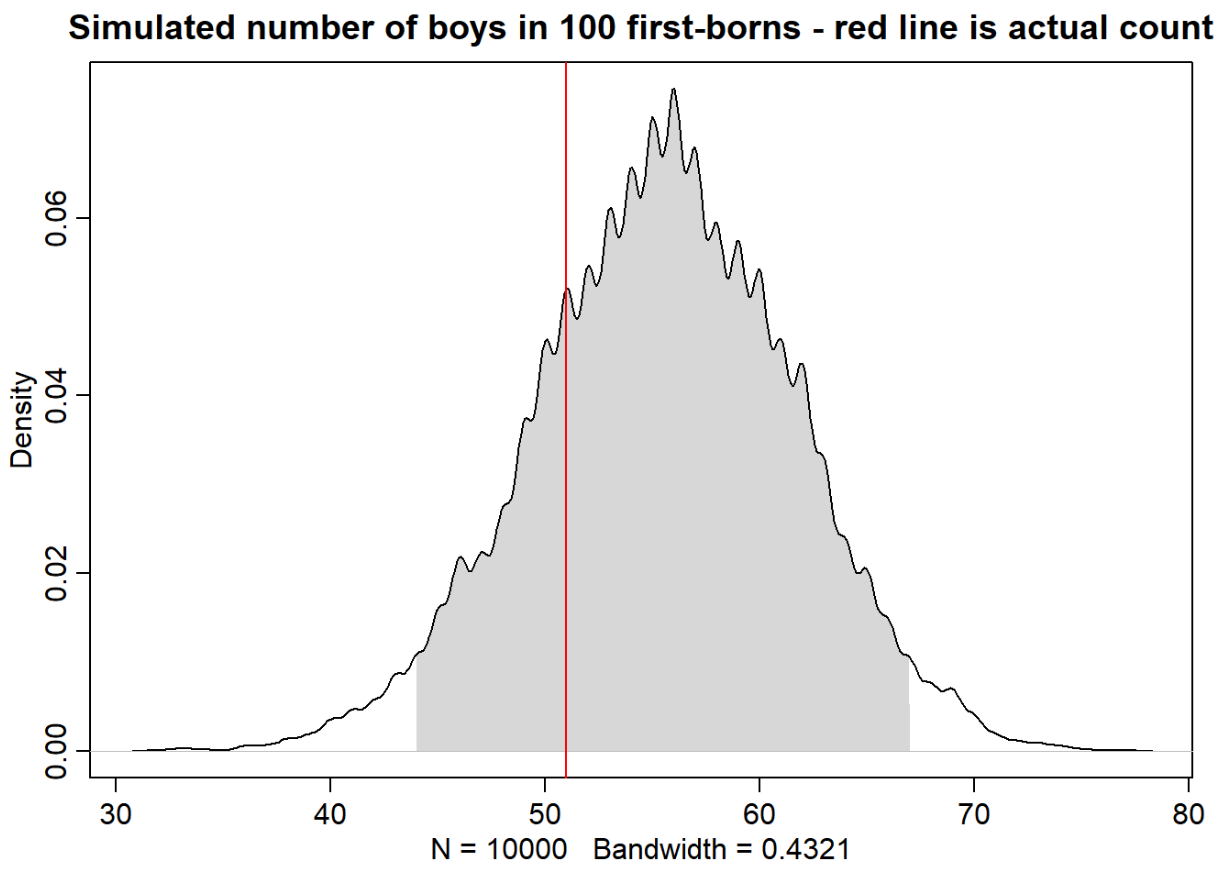
3H3 Use rbinom to simulate 10,000 replicates of 200 births. You should end up with 10,000 numbers, each one a count of boys out of 200 births. Compare the distribution of predicted numbers of boys to the actual count in the data (111 boys out of 200 births).

```
# 3H1
# Find the posterior probability of giving birth to a boy:
pacman::p_load(rethinking)
data(homeworkch3)
set.seed(1)
W <- sum(birth1) + sum(birth2)
N <- length(birth1) + length(birth2)
p_grid <-seq(from =0, to = 1, len =1000)
prob_p <- rep(1,1000)
prob_data <- dbinom(W,N,prob=p_grid)
posterior <-prob_data * prob_p
posterior <- posterior / sum(posterior)

# 3H2
# Sample probabilities from posterior distribution:
samples <- sample (p_grid, prob = posterior, size =1e4, replace =TRUE)


# 3H3
# Simulate births using sampled probabilities as simulation input, and check
if they allign with real value.
simulated_births <- rbinom(n = 1e4, size = N, prob = samples)
rethinking::dens(simulated_births,show.HPDI = 0.95)
abline(v=W, col="red")
title("Simulated amount of boys in 200 births - red line is real value")
```



3H4. Now compare 10,000 counts of boys from 100 simulated first borns only to the number of boys in the first births, birth1. How does the model look in this light?

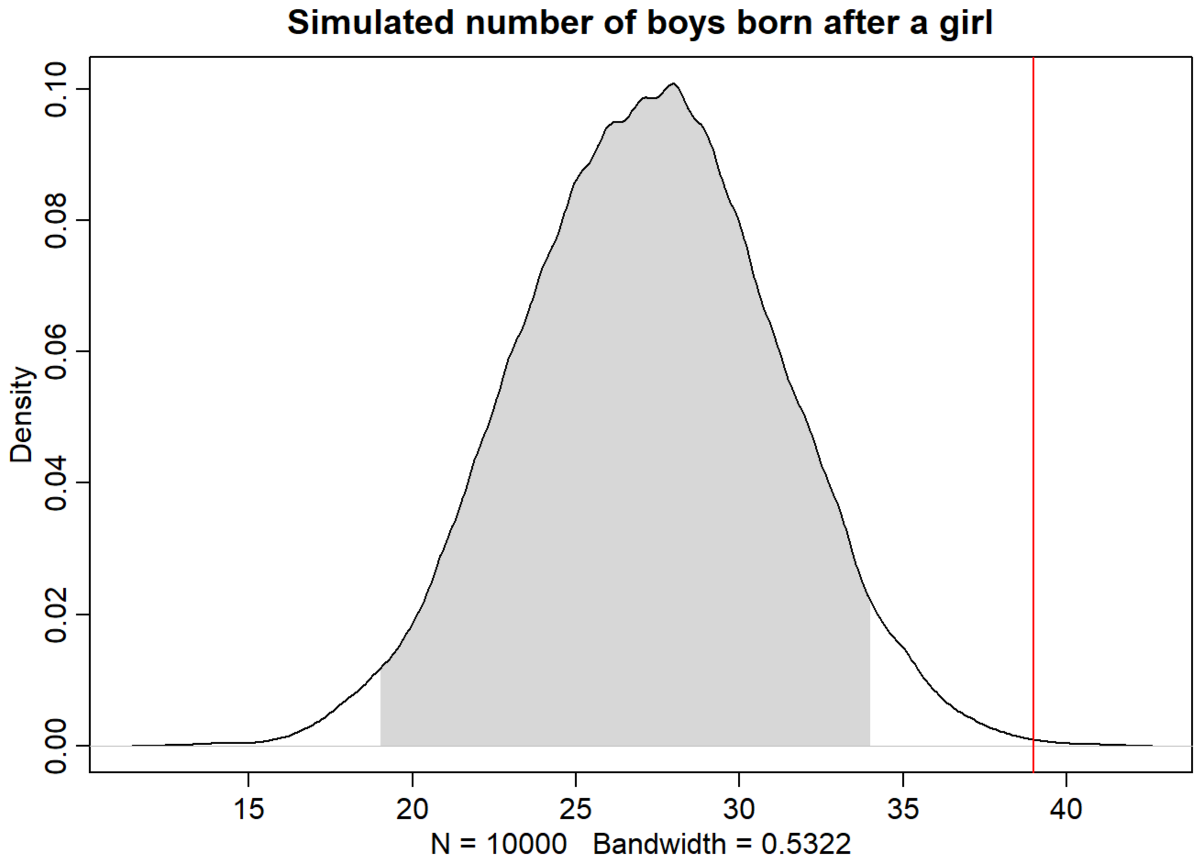
```
simulated_first_borns <- rbinom(n = 1e4, size = 100, prob = samples)
W_first_born <- sum(birth1)
rethinking::dens(simulated_first_borns, show.HPDI = 0.95)
abline(v = W_first_born, col = "red")
title("Simulated number of boys in 100 first-borns - red line is actual coun
t")
```



Answer: From the plot, we see that the observed number of first-born boys was approximately 51 out of 100 births. However, the simulated density seems to be skewed to the right of this value, indicating that the model tends to predict more boys than observed in the actual data. This differs from the previous plot where the distribution was more centered around the observed value. This suggests that the model performs less well when we subset the data to only the first-borns. This indicates that proportionally more boys were born as second-borns than first-borns. Nevertheless, the model's predictions are not drastically off, as the observed count is still within the range of plausible outcomes in the simulated distribution.

3H5. The model assumes that sex of first and second births are independent. To check this assumption, focus now on second births that followed female first borns. Compare 10,000 simulated counts of boys to only those second births that followed girls. To do this correctly, you need to count the number of first borns who were girls and simulate that many births, 10,000 times. Compare the counts of boys in your simulations to the actual observed count of boys following girls. How does the model look in this light? Any guesses what is going on in these data?

```
set.seed(1)
count_g_then_b <- sum(birth1 == 0 & birth2 == 1)
n_first_born_girls <- 100-sum(birth1)
simulated_girl_then_boy <- rbinom(n = 1e4, size = n_first_born_girls , prob
= samples)
rethinking::dens(simulated_girl_then_boy, show.HPDI = 0.95, adj = 1)
abline(v = count_g_then_b, col = "red")
title("Simulated number of boys born after a girl")
```



Answer: The actual count of boys born after a girl is outside the range of a plausible counts from the simulated model. Therefore, the model doesn't perform well when sub-setting the data. The plot indicates that for the actual observed values, it is more likely getting a boy after a girl than for the simulated data, hence the variables are not independent from each other.

Chapter 5: Spurrious Correlations

Start of by checking out all the spurrious correlations that exists in the world. Some of these can be seen on this wonderful website: https://www.tylervigen.com/spurious/random All the medium questions are only asking you to explain a solution with words, but feel free to simulate the data and prove the concepts.

5M1. - by Katrine Tikjøb Wøldike Invent your own example of a spurious correlation. An outcome variable should be correlated with both predictor variables. But when both predictors are entered in the same model, the correlation between the outcome and one of the predictors should mostly vanish (or at least be greatly reduced).

Answer: A spurious correlation could be something like a correlation between the number of students that are tired before a school day and the number of students who have bad breath. This seems at first sight unrelated and spurious, since tiredness doesn't lead to bad breath. However, there is also a correlation between having bad breath and the predictor of drinking coffee - when the students are more tired, they will probably drink more coffee, leading to bad breath. When coffee drinking is then included in the model, level of sleep might lose its influence as a predictor.

5M2. - by Rahel Meyer-Stromfeldt Invent your own example of a masked relationship. An outcome variable should be correlated with both predictor variables, but in opposite directions. And the two predictor variables should be correlated with one another.

Answer: An example for a masked relationship is predicting mood, by sleep deprivation and partying. Sleep derivation is negatively affecting mood.On the other side (we assume that) partying is positively affecting mood. Since partying is can cause sleep derivation, these variables will correlate. Therefore, the effect of the predictors will even each other out and be masked when only one of them is included in the model.

5M3. - by Regitze Hammer Holt An example for a masked relationship is predicting mood, by sleep deprivation and partying. Sleep derivation is negatively affecting mood.On the other side (we assume that) partying is positively affecting mood. Since partying is can cause sleep derivation, these variables will correlate. Therefore, the effect of the predictors will even each other out and be masked when only one of them is included in the model.

Answer: When people divorce they can marry again. Also divorces requires for a marriage to have occurred in the first place, hence having more marriages enlarge the space for potential divorces, just like how more fires requires more firefighters. Worded differently: fires are prerequisite for firefighters, and marriages are prerequisite for divorces. To model it with multiple regression, we can change the outcome to be marriage rate an the predictor to be divorce rate. We can add in predictors like religion (assuming religious beliefs may effect divorce), or age of getting married which may say something about how liberal a society is.

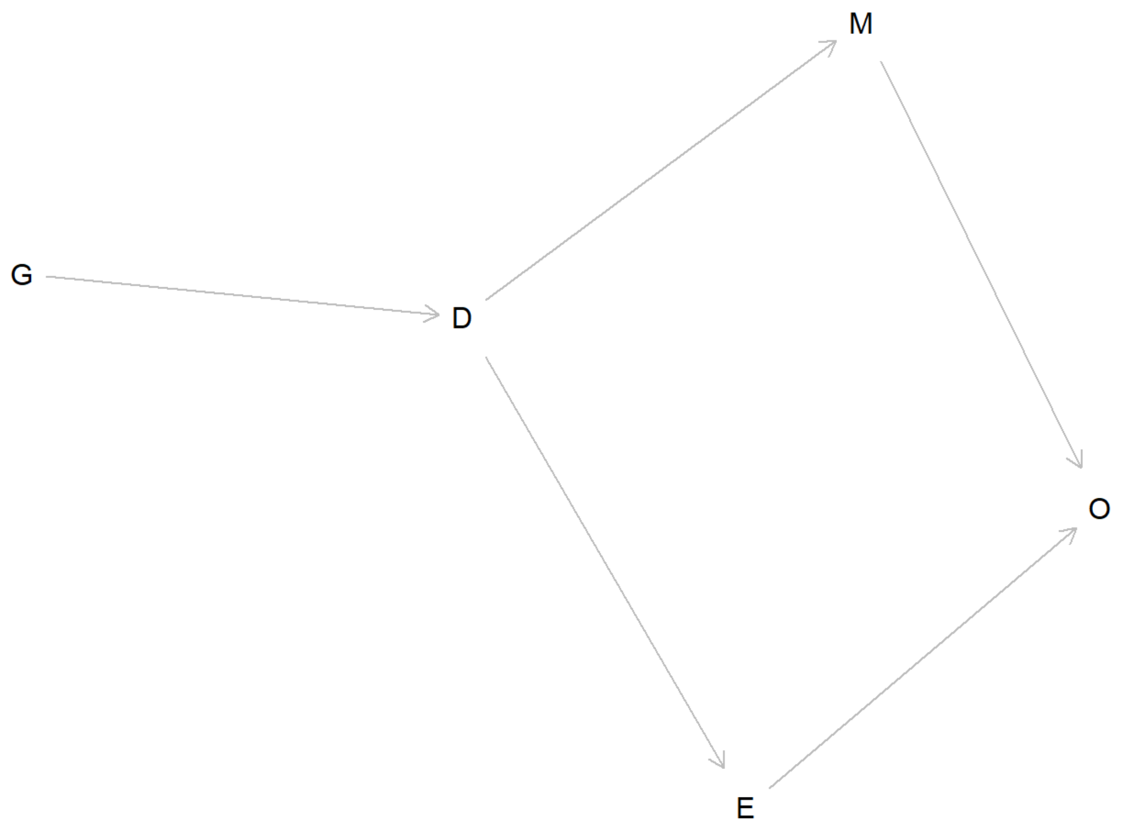
5M5. - Sára Fernezelyi One way to reason through multiple causation hypotheses is to imagine detailed mechanisms through which predictor variables may influence outcomes. For example, it is sometimes argued that the price of gasoline (predictor variable) is positively associated with lower obesity rates (outcome variable). However, there are at least two important mechanisms by which the price of gas could reduce obesity. First, it could lead to less driving and therefore more exercise. Second, it could lead to less driving, which leads to less eating out, which leads to less consumption of huge restaurant meals. Can you outline one or more multiple regressions that address these two mechanisms? Assume you can have any predictor data you need.

Answer: First we create notations for the predictors and outcomes: - Price of Gasoline (predictor): G - Amount of Driving (confound): D - Amount of Excercise (confound): E - Amount of Restaurant Meals (confound): M - Obesity rate (outcome): O

We have two ways that the predictors influence each other and thus the outcome, which can be our two multiple regressions: - Path 1: G -> D -> E -> O - Path 2: G -> D -> M -> O

```
library(dagitty)

dag_5m5 <- dagitty("dag {
                     G -> D
                     D -> E
                     D -> M
                     E -> O
                     M -> O
                     }")

plot(dag_5m5)
```

```
## Plot coordinates for graph not supplied! Generating coordinates, see ?coo
rdinates for how to set your own.
```



From this we can conclude:

- If gasoline is cheaper, people drive more, so they exercise less, so they are more obese.
- If gasoline is cheaper, people drive more, so they eat out more, so they are more obese.

Looking at the two paths we can see that driving has the same explanatory value in both, so we can just create a multiple regression with the outcome being obesity rate, an intercept, a beta estimate multiplied by a measure of how much people exercise by walking, a beta estimate multiplied by a measure of how much people eat out, and a beta estimate multiplied by the price of gasoline. By looking at the mean estimates of the beta parameters we can check which predictor describes the correlation better. We can also create a full-model, combining beta estimates of gas prices, eating out and exercise, so we can see if it is really through these paths that obesity is mediated.

Chapter 5: Foxes and Pack Sizes - by Rahel Meyer-Stromfeldt

All five exercises below use the same data, data(foxes) (part of rethinking).84 The urban fox (Vulpes vulpes) is a successful exploiter of human habitat. Since urban foxes move in packs and defend territories, data on habitat quality and population density is also included. The data frame has five columns: (1) group: Number of the social group the individual fox belongs to (2) avgfood: The average amount of food available in the territory (3) groupsize: The number of foxes in the social group (4) area: Size of the territory (5) weight: Body weight of the individual fox

5H1. Fit two bivariate Gaussian regressions, using quap: (1) body weight as a linear function of territory size (area), and (2) body weight as a linear function of groupsize. Plot the results of these regressions, displaying the MAP regression line and the 95% interval of the mean. Is either variable important for predicting fox body weight?

```
data(foxes)
head(foxes)
```

```
##   group avgfood groupsize area weight
## 1     1    0.37         2 1.09   5.02
## 2     1    0.37         2 1.09   2.84
## 3     2    0.53         2 2.05   5.33
## 4     2    0.53         2 2.05   6.07
## 5     3    0.49         2 2.12   5.85
## 6     3    0.49         2 2.12   3.25
```

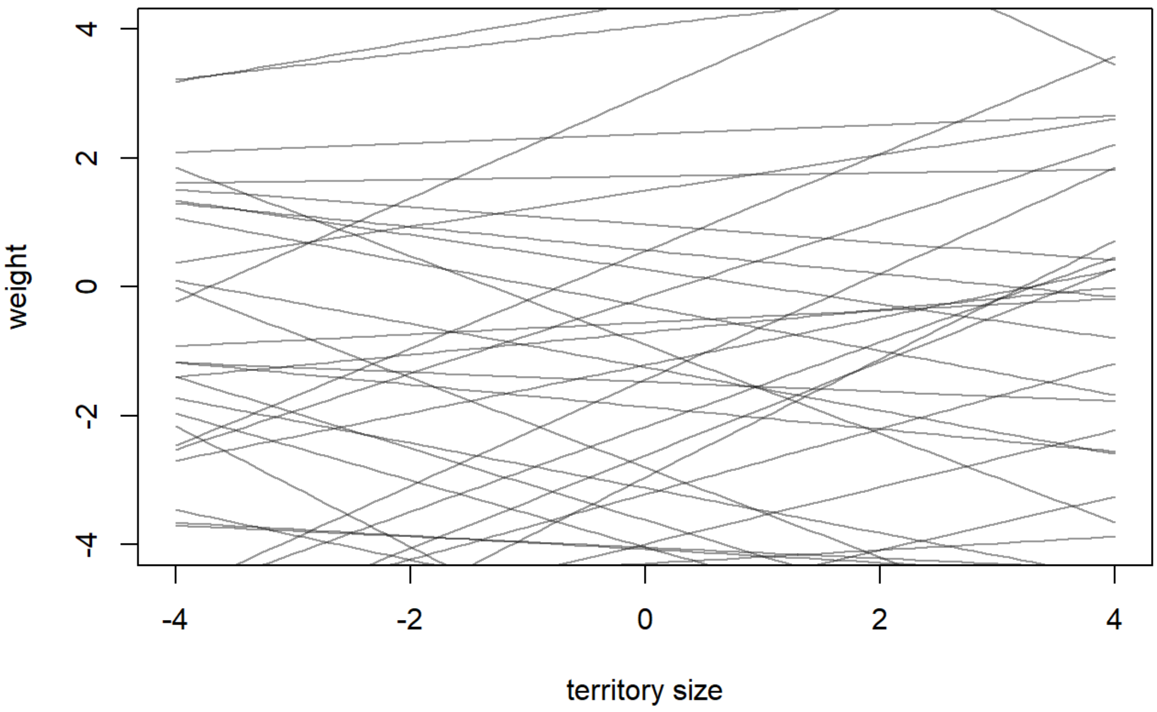
```
d <- list()

# since we want to compare effect sizes of different variables, it makes sen
se to scale them fisrt


d$area <- scale(foxes$area)
d$groupsize <- scale(foxes$groupsize)
d$avgfood <-  scale(foxes$avgfood)
d$weight <- scale(foxes$weight)

# linear model of body weight predicted by territory size

m_area <- quap(
  alist(
    weight ~ dnorm(mu, sigma),
    mu <- a + b_area *area,
    a ~ dnorm(0, 5),
    b_area ~ dnorm (0, 0.5),
    sigma ~ dexp(1)
  ), data = d )
precis(m_area)
```

```
##                  mean         sd       5.5%     94.5%
## a        -2.649397e-07 0.09202151 -0.1470684 0.1470679
## b_area    1.883313e-02 0.09089616 -0.1264365 0.1641027
## sigma     9.912699e-01 0.06466710  0.8879194 1.0946204
```

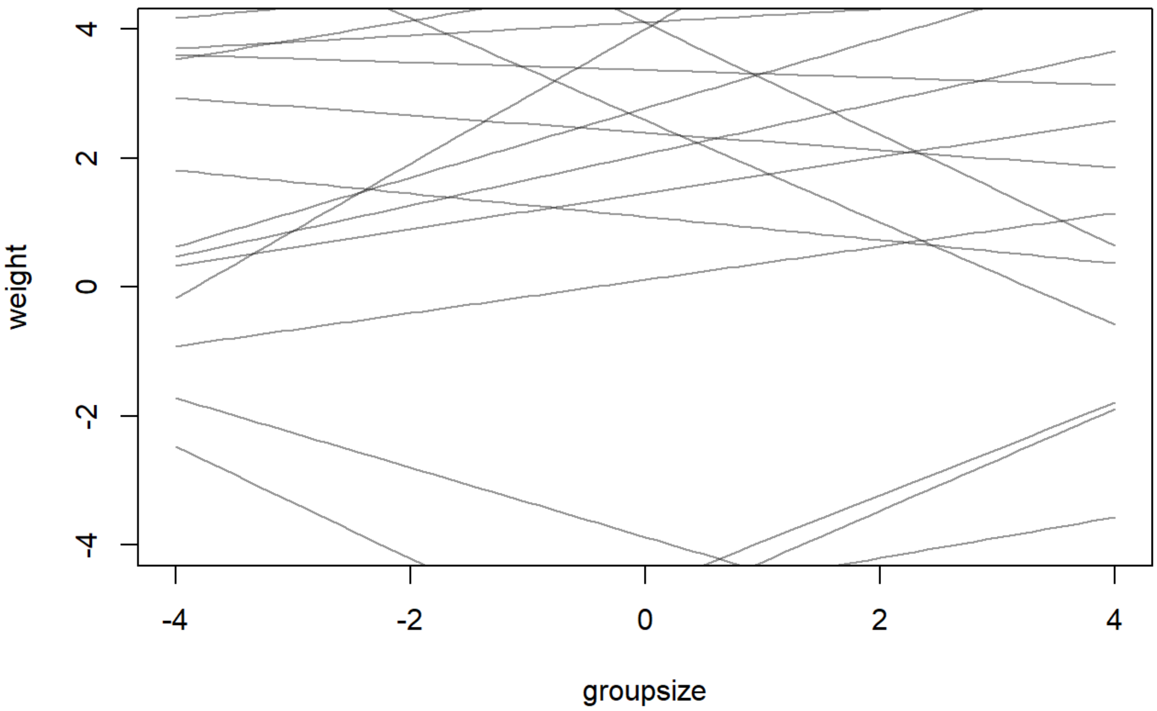
```
# model for body weight predicted by groupsize !! probably wrong because gro
upsize is categorical!!

m_groupsize <- quap(
  alist(
    weight ~ dnorm(mu, sigma),
    mu <- a + b_groupsize *groupsize,
    a ~ dnorm(0, 10),    # plausible priors after plotting different options
    b_groupsize~ dnorm (0, 0.5),    # random prior
    sigma ~ dexp(1)
  ), data = d )
precis(m_groupsize)
```

```
##                      mean         sd       5.5%       94.5%
## a           -1.669890e-08 0.09085578 -0.1452051  0.14520507
## b_groupsize -1.558040e-01 0.08977327 -0.2992791 -0.01232901
## sigma        9.785871e-01 0.06384616  0.8765486  1.08062562
```

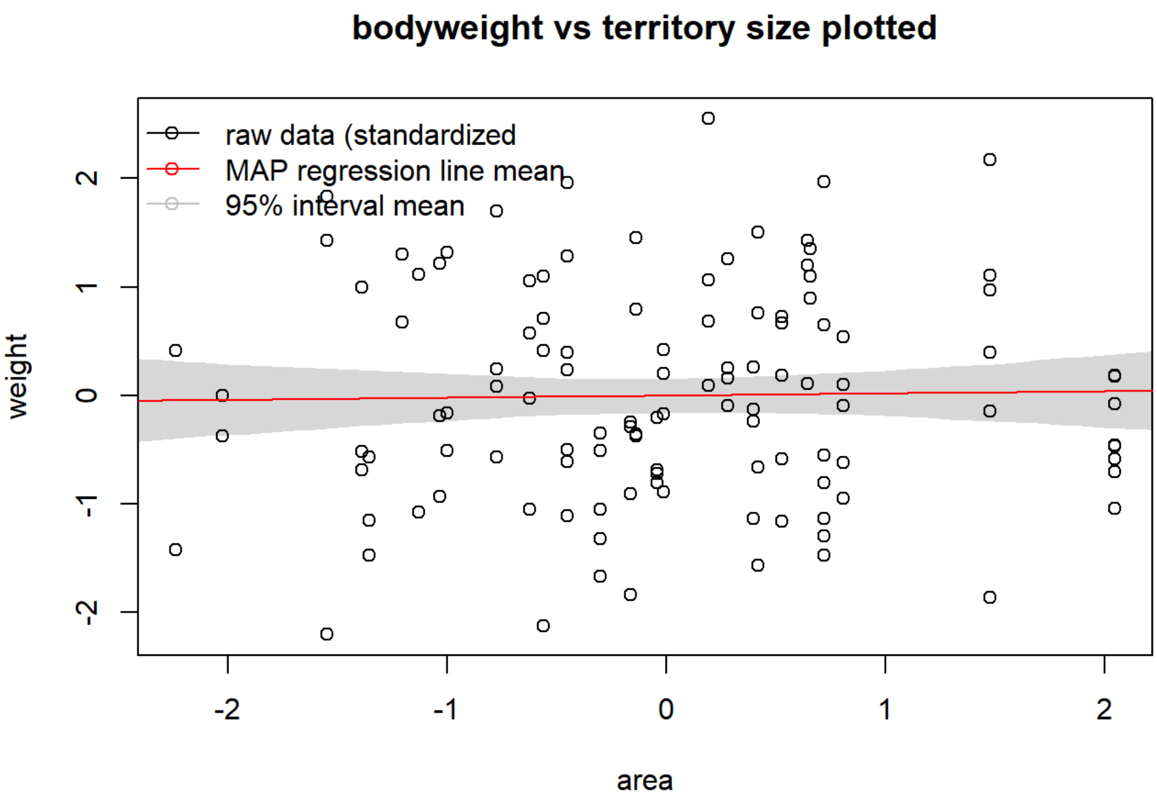
```
area_seq <- seq(-4,4, by = 0.1)
groupsize_seq <- seq(-4,4, by = 0.1)
# # simulate and plot prior regression lines to find reasonable priors
prior <- extract.prior(m_area)
mu <- link(m_area , post = prior, data = list(area = area_seq))
plot( NULL , xlim= range(area_seq) , ylim= range(area_seq), xlab = "territor
y size", ylab = "weight")
for ( i in 1:50 ) lines( area_seq , mu[i,] , col=col.alpha("black",0.4) )
```



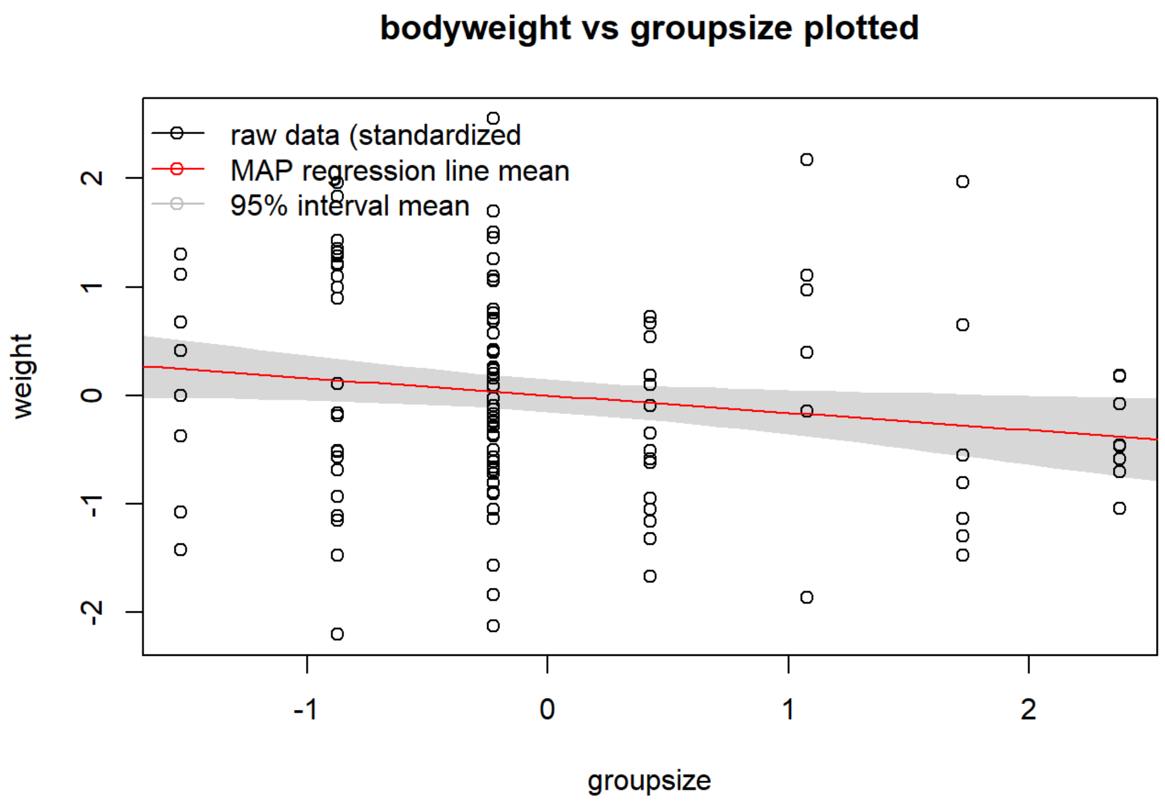
```
# same for groupsize:
prior <- extract.prior(m_groupsize)
mu <- link(m_groupsize , post = prior, data = list(groupsize = groupsize_se
q))
plot( NULL , xlim= range(groupsize_seq) , ylim= range(groupsize_seq), xlab =
"groupsize", ylab = "weight")
for ( i in 1:50 ) lines( area_seq , mu[i,] , col=col.alpha("black",0.4) )
```



```
## link to compute mean
mu <- link(m_area, data = data.frame(area = area_seq))
# summarize distribution of mu
mu.mean <-  apply(mu,2,mean)
mu.PI <- apply(mu,2,PI, prob = 0.89)
# plot raw  data
plot(weight ~ area, data = d,
     main = "bodyweight vs territory size plotted")

# add shaded area for 89% interval
shade(mu.PI, area_seq )

# add MAP line
lines(area_seq, mu.mean, col = "red")
# add a legend to not get confused
legend("topleft",legend = c("raw data (standardized", "MAP regression line m
ean", "95% interval mean"), col = c("black","red", "grey"), lty = 1, pch =
1, bty = "n")
```



```
# and same for groupsize


## link to compute mean
mu <- link(m_groupsize, data = data.frame(groupsize = groupsize_seq))
# summarize distribution of mu
mu.mean <-  apply(mu,2,mean)
mu.PI <- apply(mu,2,PI, prob = 0.89)
# plot raw  data
plot(weight ~ groupsize, data = d,
     main = "bodyweight vs groupsize plotted")

# add shaded area for 89% interval
shade(mu.PI, groupsize_seq )

# add MAP line
lines(groupsize_seq, mu.mean, col = "red")
# add a legend to not get confused
legend("topleft",legend = c("raw data (standardized", "MAP regression line m
ean", "95% interval mean"), col = c("black","red", "grey"), lty = 1, pch =
1, bty = "n")
```



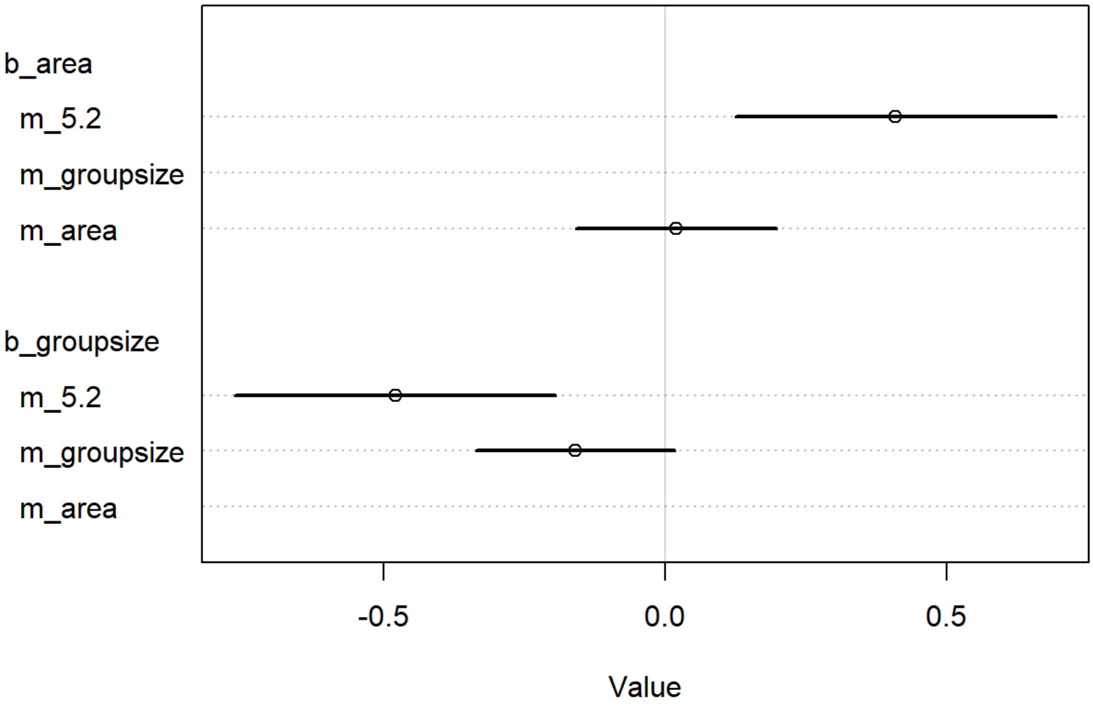
Answer: Both variables show rather flat regression lines, with fairly narrow 95% intervals. Looking at the data and the spread of it, both variables seem to be unimportant to predict weight.

5H2. Now fit a multiple linear regression with weight as the outcome and both area and groupsize as predictor variables. Plot the predictions of the model for each predictor, holding the other predictor constant at its mean. What does this model say about the importance of each variable? Why do you get different results than you got in the exercise just above?

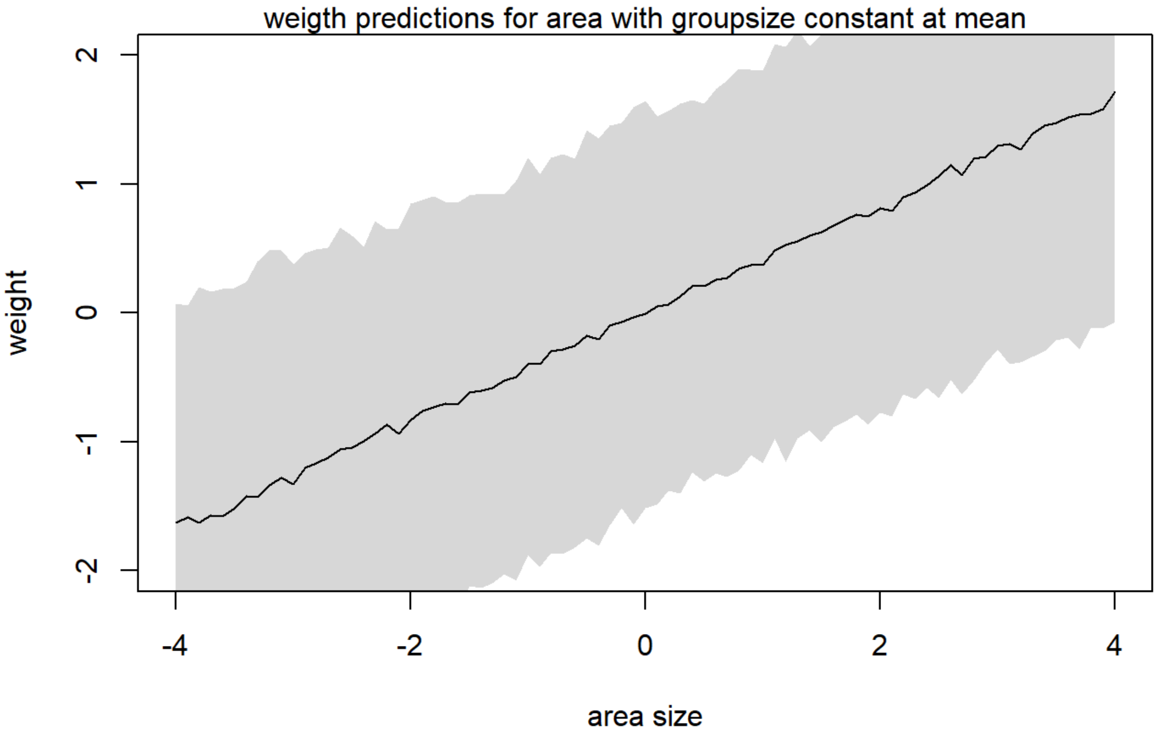
```
# multivariate regression
m_5.2 <- quap(
  alist(
    weight ~ dnorm(mu, sigma),
    mu <- a + b_area* area + b_groupsize*groupsize,
    a ~ dnorm(0,10),
    b_area ~ dnorm (0, 0.5),
    b_groupsize ~ dnorm(0,0.5),
    sigma ~ dexp(1)
  ), data = d)
precis(m_5.2)
```

```
##                      mean         sd       5.5%      94.5%
## a            3.461725e-06 0.08743526 -0.1397350  0.1397419
## b_area       4.055240e-01 0.14533773  0.1732462  0.6378017
## b_groupsize -4.816176e-01 0.14534822 -0.7139121 -0.2493231
## sigma        9.417426e-01 0.06156111  0.8433561  1.0401291
```

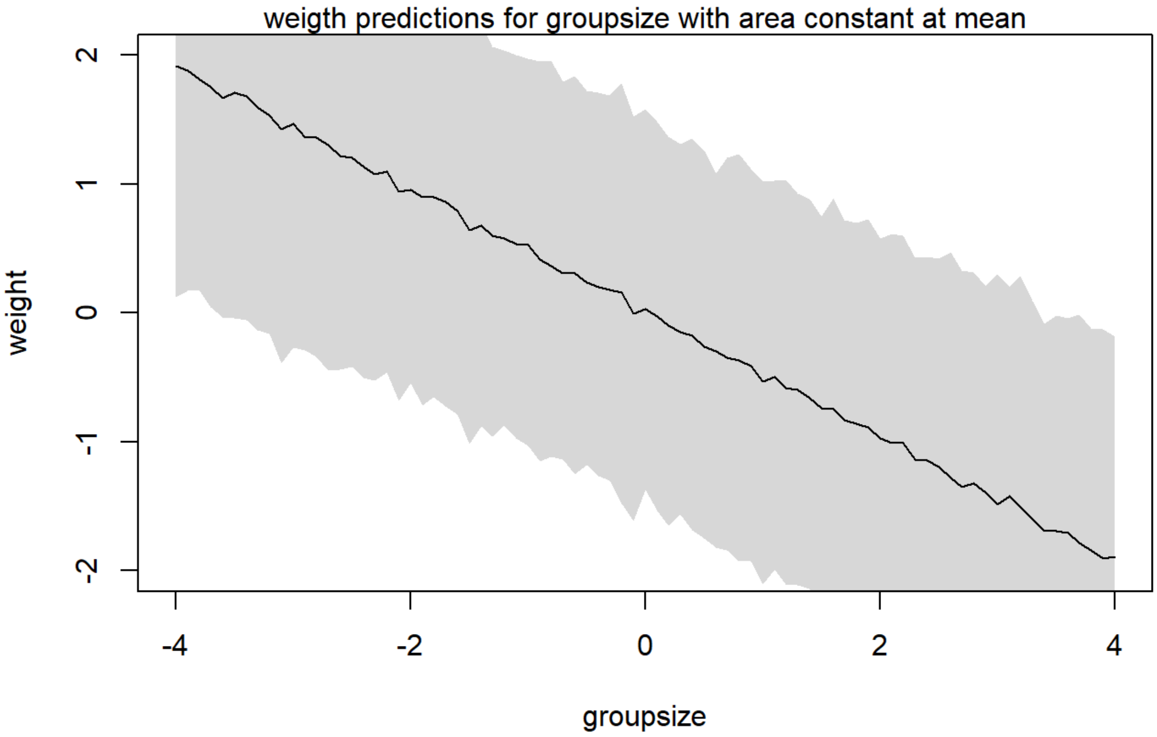
```
# plot coefficients against each other: ??????
plot(coeftab(m_area, m_groupsize, m_5.2), pars = c("b_area", "b_groupsize"))
```



```
# create means:
groupsize_mean = mean(d$groupsize)
area_mean = mean(d$area)
# sim data for area holding groupsize 0
sim_dat_area<- data.frame(area = area_seq, groupsize = groupsize_mean)
s_area <- sim(m_5.2, data = sim_dat_area, vars = c("weight"))
# sim data for groupsize holding area 0
sim_dat_groupsize <- data.frame(groupsize = groupsize_seq, area = area_mean)
s_groupsize <- sim(m_5.2, data = sim_dat_groupsize, vars = c("weight"))

plot(sim_dat_area$area, colMeans(s_area), ylim = c(-2,2), type = "l",
     xlab = "area size", ylab = "weight")
shade( apply(s_area,2,PI), sim_dat_area$area)
mtext("weigth predictions for area with groupsize constant at mean")
```



```
plot(sim_dat_groupsize$groupsize, colMeans(s_groupsize), ylim = c(-2,2), typ
e = "l",
     xlab = "groupsize", ylab = "weight")
shade( apply(s_groupsize,2,PI), sim_dat_groupsize$groupsize)
mtext("weigth predictions for groupsize with area constant at mean")
```



Answer: This model shows that the variables are both predictors for bodyweight but with effects in opposite directions. Therefore they are confounders when only one of them is included in the model. Groupsize seems to have a negative effect (mean = -0.48) on weight, whereas area has a positive effect (mean = 0.41). Judging by its now bigger effect and compared to that smaller standard deviation both predictors are important. The reason this was not seen in the model above is, that we did not include both predictors, creating a masked relationship.

5H3. Finally, consider the avgfood variable. Fit two more multiple regressions: (1) body weight as an additive function of avgfood and groupsize, and (2) body weight as an additive function of all three variables, avgfood and groupsize and area. Compare the results of these models to the previous models you've fit, in the first two exercises. (a) Is avgfood or area a better predictor of body weight? If you had to choose one or the other to include in a model, which would it be? Support your assessment with any tables or plots you choose. (b) When both avgfood or area are in the same model, their effects are reduced (closer to zero) and their standard errors are larger than when they are included in separate models. Can you explain this result?

```
# 1. multiple regression of body weight predicted by avg food and groupsize


m_5.3.1 <- quap(
  alist(
    weight ~ dnorm(mu, sigma),
    mu <- a + b_avgfood*avgfood + b_groupsize*groupsize,
    a ~ dnorm(0,5),
    b_avgfood ~ dnorm (0, 0.5),
    b_groupsize ~ dnorm(0,0.5),
    sigma ~ dexp(1)
  ), data = d)
precis(m_5.3.1)
```

```
##                     mean         sd       5.5%      94.5%
## a           2.824776e-06 0.08745801 -0.1397720  0.1397776
## b_avgfood   4.770707e-01 0.17913251  0.1907823  0.7633590
## b_groupsize -5.733189e-01 0.17915132 -0.8596373 -0.2870005
## sigma        9.420957e-01 0.06176104  0.8433897  1.0408018
```

```
# 2. multiple regression of body weight predicted by avg food and groupsize
and area

m_5.3.2 <- quap(
  alist(
    weight ~ dnorm(mu, sigma),
    mu <- a +b_area*area + b_avgfood*avgfood + b_groupsize*groupsize,
    a ~ dnorm(0,5),
    b_avgfood ~ dnorm (0, 0.5),
    b_area ~ dnorm (0, 0.5),
    b_groupsize ~ dnorm(0,0.5),
    sigma ~ dexp(1)
  ), data = d)
precis(m_5.3.2)
```

```
##                     mean         sd         5.5%      94.5%
## a           1.169929e-08 0.08644744 -0.138159691  0.1381597
## b_avgfood   2.968997e-01 0.20960027 -0.038082038  0.6318814
## b_area      2.782382e-01 0.17011231  0.006365882  0.5501105
## b_groupsize -6.396203e-01 0.18161487 -0.929875948 -0.3493647
## sigma        9.312066e-01 0.06100013  0.833716608  1.0286966
```

```
precis(m_5.2)
```

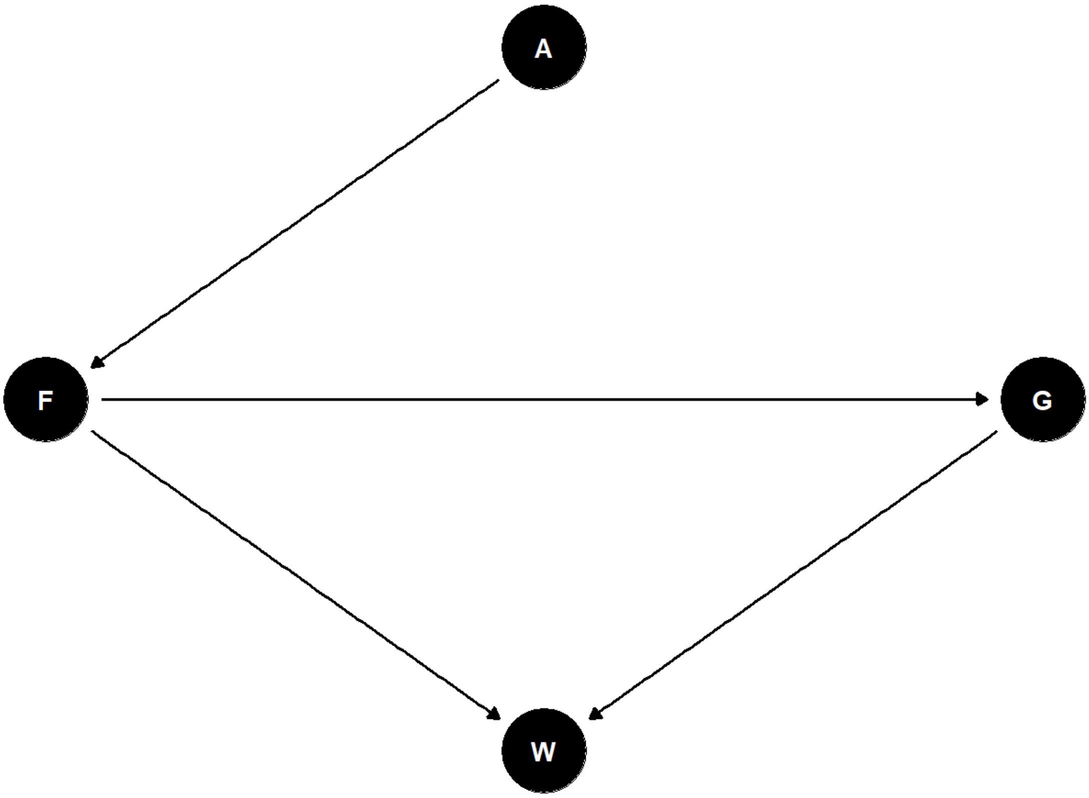
```
##                     mean         sd       5.5%      94.5%
## a           3.461725e-06 0.08743526 -0.1397350  0.1397419
## b_area      4.055240e-01 0.14533773  0.1732462  0.6378017
## b_groupsize -4.816176e-01 0.14534822 -0.7139121 -0.2493231
## sigma        9.417426e-01 0.06156111  0.8433561  1.0401291
```

Answer: a) According to the model, both average food and area size have a similar predictive effect on weight. Whereas average food has a slightly bigger effect, area size has a slightly smaller standard deviation. Therefore I think both of them could be argued to be used. If I would need to choose I would choose food, because it has the bigger effect.

b. When both of the predictors are included, their effects are smaller with a larger standard deviation, because they explain a lot of the same variance. The variables are correlated with each other, so the model cannot estimate the effect size as precisely.

Defining our theory with explicit DAGs - by Katrine Tikjøb Wøldike Assume this DAG as an causal explanation of fox weight:

```r
data(foxes)

#scaling
foxes$avgfood <- scale(foxes$avgfood)
foxes$area <- scale(foxes$area)
foxes$weight <- scale(foxes$weight)
foxes$groupsize <- scale(foxes$groupsize)



pacman::p_load(dagitty,
               ggdag)
dag <- dagitty('dag {
A[pos="1.000,0.500"]
F[pos="0.000,0.000"]
G[pos="2.000,0.000"]
W[pos="1.000,-0.500"]
A -> F
F -> G
F -> W
G -> W
}')

# Plot the DAG
ggdag(dag, layout = "circle")+
  theme_dag()
```



where A is area, F is avgfood,G is groupsize, and W is weight.

Using what you know about DAGs from chapter 5 and 6, solve the following three questions:

1. Estimate the total causal influence of A on F. What effect would increasing the area of a territory have on the amount of food inside of it?

```
# since there are no other variables influencing F, we'll make a simple model
model_areafood <- quap(
  alist(
    avgfood ~ dnorm(mu, sigma),
    mu <- a + b * area,
    a ~ dnorm(0.5, 0.5),
    b ~ dnorm(0,1),
    sigma ~ dexp(1)
  ),
  data = foxes
)
precis( model_areafood )
```

```
##              mean         sd        5.5%      94.5%
## a     0.003715278 0.04312632 -0.06520891 0.07263947
## b     0.881437556 0.04343222  0.81202448 0.95085063
## sigma 0.466192723 0.03051295  0.41742713 0.51495831
```

Answer: Increasing an area seems to also increase the amount of food inside of it with a small sigma, which makes sense also from a conceptual perspective.

2. Infer the **total** causal effect of adding food F to a territory on the weight W of foxes. Can you calculate the causal effect by simulating an intervention on food?

```
set.seed(123)
# first making the model that we want to simulate from
# Since we want the total causal effect

model_foodweight <- quap(
  alist(
    weight ~ dnorm(mu, sigma),
    mu <- a + bF * avgfood,
    a ~ dnorm(0, 0.5),
    bF ~ dnorm(0,1), # effect of food on weight
    sigma ~ dexp(1)
  ),
  data = foxes
)

precis(model_foodweight)
```

```
##                mean         sd       5.5%     94.5%
## a     -8.120403e-05 0.09050377 -0.1447237 0.1445613
## bF    -2.486612e-02 0.09203085 -0.1719492 0.1222170
## sigma  9.911271e-01 0.06465583  0.8877946 1.0944596
```

Answer: Since it would make a lot of sense if more food lead to an increase in weight of the foxes, this model makes us suspect that there could be a masked relationship hidden somewhere in our structure. *We'll return to the intervention after next question:))*

3. Infer the **direct** causal effect of adding food F to a territory on the weight W of foxes. In light of your estimates from this problem and the previous one, what do you think is going on with these foxes?

```
set.seed(123)
# conditioning on groupsize to take away the indirect effect
model_foodweight2 <- quap(
  alist(
    weight ~ dnorm(mu, sigma),
    mu <- a + bF * avgfood + bG * groupsize,
    a ~ dnorm(0, 10),
    bF ~ dnorm(0,2), # effect of food on weight
    bG ~ dnorm(0,2), # effect of group size on weight
    sigma ~ dexp(1)
  ),
   data = foxes
)
precis(model_foodweight2)
```

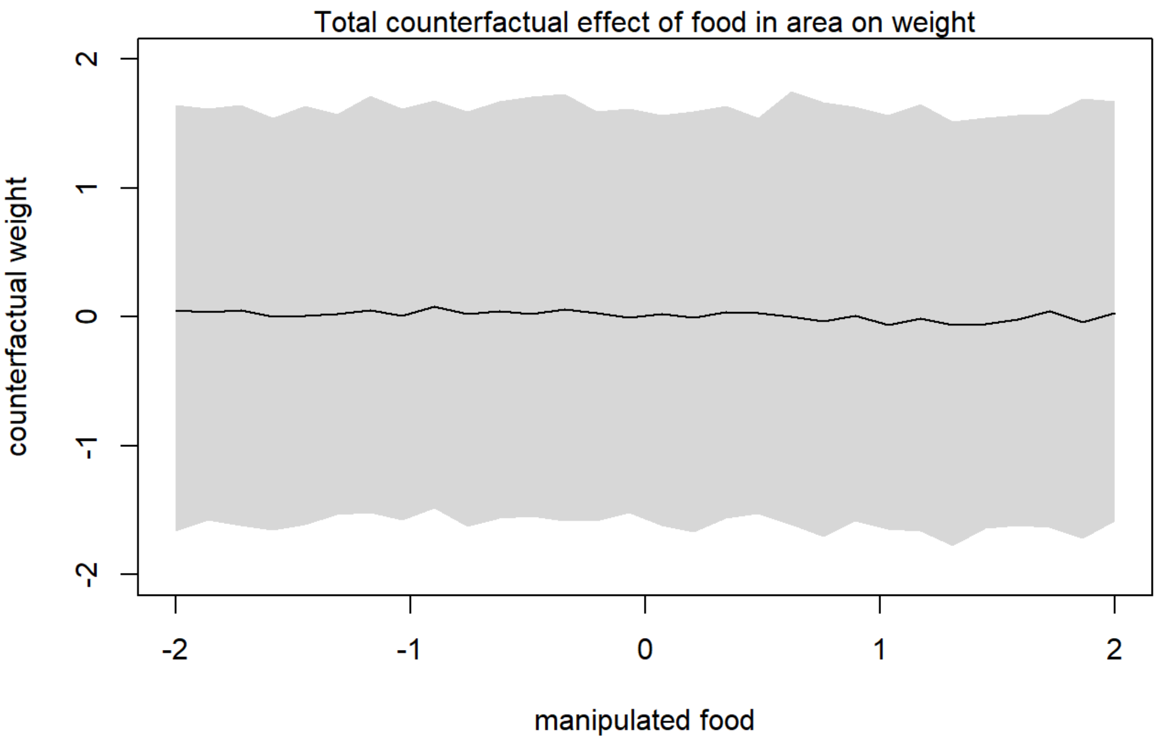
```
##                 mean         sd       5.5%      94.5%
## a      2.557851e-06 0.08720814 -0.1393729  0.1393780
## bF     6.278962e-01 0.20052901  0.3074121  0.9483803
## bG    -7.256325e-01 0.20052912 -1.0461168 -0.4051483
## sigma  9.392961e-01 0.06130025  0.8413265  1.0372657
```

```
set.seed(123)
# prepare a model for the simulation, implying our causal assumptions

model_foodweight3 <- quap(
  alist(
    weight ~ dnorm(mu, sigma),
    mu <- a + bF * avgfood + bG * groupsize,
    a ~ dnorm(0, 10),
    bF ~ dnorm(0,2), # effect of food on weight
    bG ~ dnorm(0,2), # effect of group size on weight
    sigma ~ dexp(1),
    ## F -> G
        groupsize ~ dnorm( mu_groupsize , sigmaG),
        mu_groupsize <- aG + bFG*avgfood,
        aG ~ dnorm( 0 , 0.2 ),
        bFG ~ dnorm( 0 , 0.5 ),
        sigmaG ~ dexp( 1 )
  ),
  data = foxes
)

precis(model_foodweight3)
```

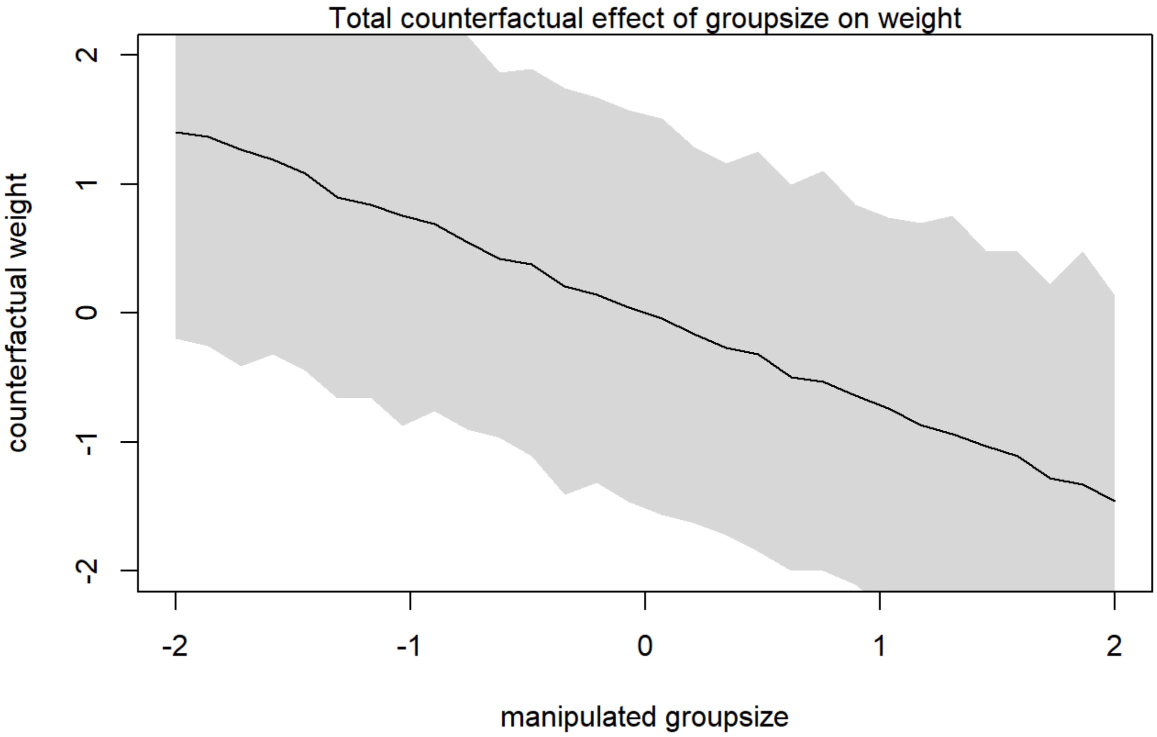
```
##                 mean         sd        5.5%       94.5%
## a       1.435455e-04 0.08717653 -0.13918138  0.13946847
## bF      6.279437e-01 0.20045758  0.30757380  0.94831366
## bG     -7.255652e-01 0.20045782 -1.04593554 -0.40519492
## sigma   9.389556e-01 0.06124484  0.84107449  1.03683667
## aG     -6.073356e-05 0.03916730 -0.06265765  0.06253618
## bFG     8.957107e-01 0.03999255  0.83179489  0.95962653
## sigmaG  4.301744e-01 0.02816721  0.38515779  0.47519108
```

```
# Simulating intervention
# using the model from 3) because it's the only way this makes sense?
f_seq <- seq( from=-2 , to=2 , length.out=30 )

# prep data
sim_dat <- data.frame( avgfood=f_seq )

# simulate groupsize and then weight, using avgfood simulations
s <- sim( model_foodweight3 , data=sim_dat , vars=c("groupsize","weight") )

# display counterfactual predictions
```

```
plot( sim_dat$avgfood , colMeans(s$weight) , ylim=c(-2,2) , type="l" ,
    xlab="manipulated food" , ylab="counterfactual weight"  )
shade( apply(s$weight,2,PI) , sim_dat$avgfood )
mtext( "Total counterfactual effect of food in area on weight" )
```



```
sim_dat <- data.frame( groupsize=seq(from=-2,to=2,length.out=30) , avgfood=0
)
s <- sim( model_foodweight3 , data=sim_dat , vars="weight" )
plot( sim_dat$groupsize , colMeans(s) , ylim=c(-2,2) , type="l" ,
    xlab="manipulated groupsize" , ylab="counterfactual weight"  )
shade( apply(s,2,PI) , sim_dat$groupsize )
mtext( "Total counterfactual effect of groupsize on weight" )
```



Answer: According to our model output, the direct effect of F on W is positive, whereas the indirect effect mediated by G is negative. In the first plot, part of the effect of F on W is mediated by G. It's only when we manually set the avgfood to 0, that the effect of groupsize gets to show itself. This makes sense if more food in the area leads to a growth in the fox groups, leading to more competition for the food, leading to lower weights.

Chapter 6: Investigating the Waffles and Divorces - by Sára Fernezelyi

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.

```
#setting up data
library(rethinking)
data("WaffleDivorce")
data <- WaffleDivorce
data$S <- scale(data$South)
data$A <- scale(data$MedianAgeMarriage)
data$D <- scale(data$Divorce)
data$M <- scale(data$Marriage)
data$W <- scale(data$WaffleHouses)

#causal graph
library(dagitty)
library(ggplot2)
library(dplyr)
```

```
##
## Kapcsolódás csomaghoz: 'dplyr'
```

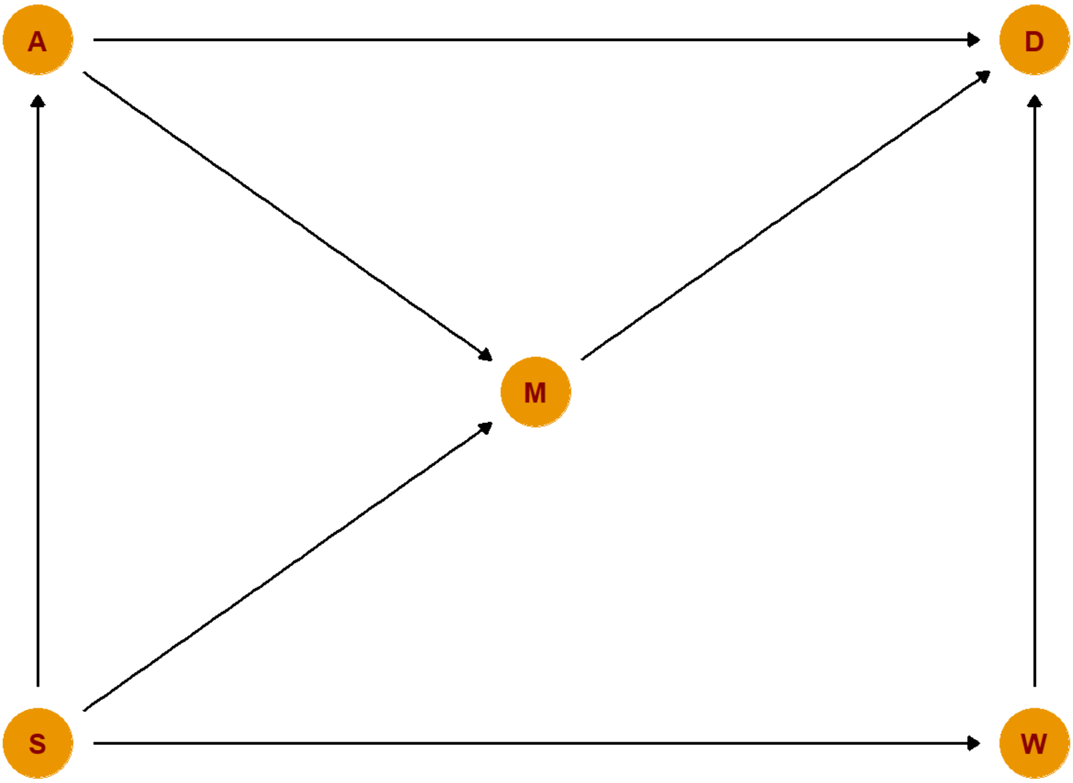
```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
print("S: if a state is southern, A: median age at marriage, D: divorce rat
e, M: marriage rate, W: number of Waffle Houses")
```

```
## [1] "S: if a state is southern, A: median age at marriage, D: divorce rat
e, M: marriage rate, W: number of Waffle Houses"
```

```
dag_6h1 <- dagitty("dag {
                   A -> D
                   A -> M -> D
                   A <- S -> M
                   S -> W -> D
                   }")
coordinates(dag_6h1) <- list(
  x = c(A = 1, S = 1, M = 2, W = 3, D = 3),
  y = c(A = 3, S = 1, M = 2, W = 1, D = 3)
)

library(ggdag)
tidy_dag <- tidy_dagitty(dag_6h1)
ggplot(dag_6h1,
       aes(x = x, y = y, xend = xend, yend = yend)) +
  geom_dag_point(size = 20, shape = 20, color = "orange2") +
  geom_dag_text(color = "red4") +
  geom_dag_edges() +
  theme_dag()
```

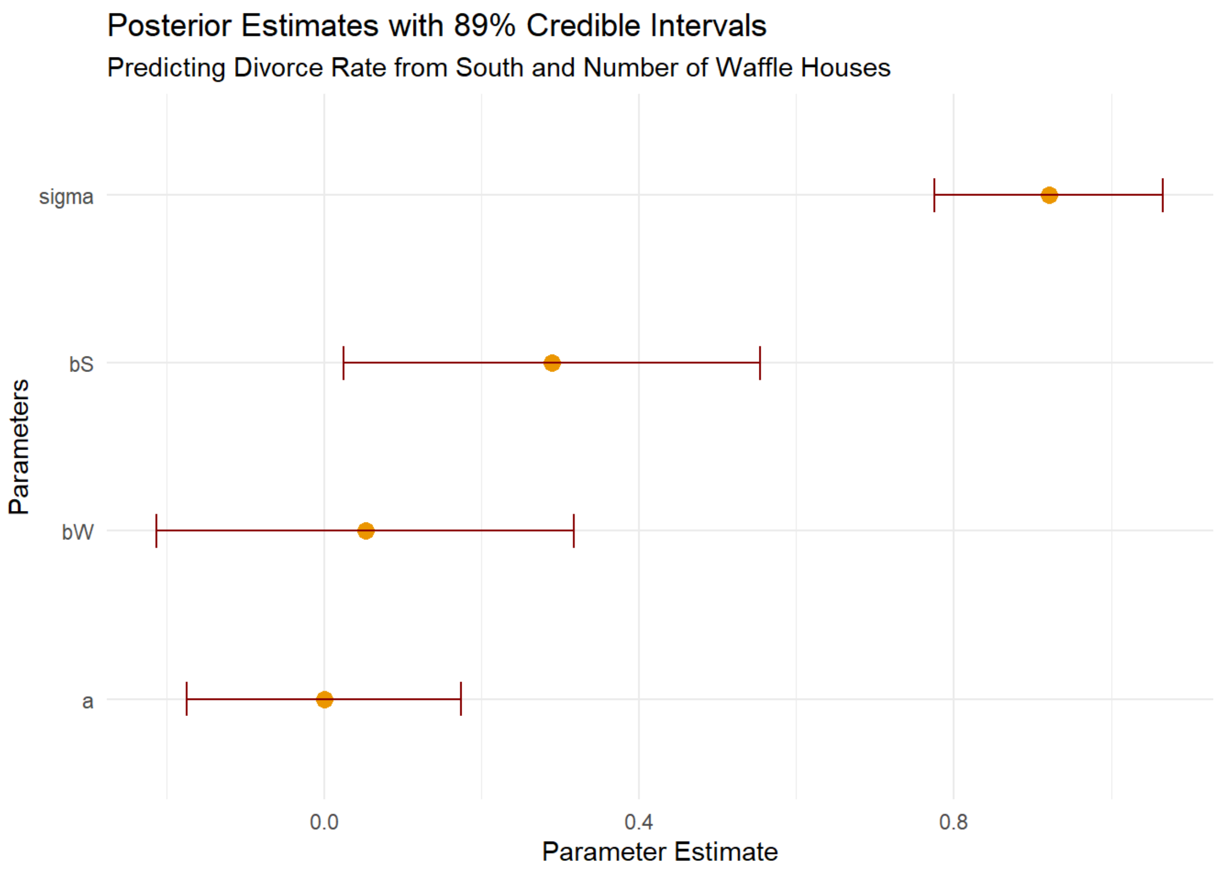


```
#modelling
##conditioning on south as it is the open only back door on divorce reate th
at is not connected to the number of waffle houses
model_6h1 <- quap(
  alist(
    D ~ dnorm(mu, sigma),
    mu <- a + bS*S + bW*W,
    a ~ dnorm(0, 0.2),
    bS ~ dnorm(0, 0.5),
    bW ~ dnorm(0, 0.5),
    sigma ~ dexp(1)
  ), data = data
)

precis(model_6h1)
```

```
##                mean        sd        5.5%     94.5%
## a     -3.207148e-05 0.1091146 -0.17441834 0.1743542
## bS     2.892660e-01 0.1659435  0.02405626 0.5544757
## bW     5.233165e-02 0.1658759 -0.21277013 0.3174334
## sigma  9.206430e-01 0.0908662  0.77542129 1.0658648
```

```
#getting summary of model
summary_6h1 <- precis(model_6h1, depth = 2) %>% as.data.frame()
summary_6h1$parameter <- rownames(summary_6h1)
```

```
#plotting
ggplot(summary_6h1, aes(x = mean, y = reorder(parameter, mean))) +
  geom_point(size = 3, color = "orange2") +
  geom_errorbarh(aes(xmin = `5.5%`, xmax = `94.5%`), height = 0.2, color = "red4") +
  theme_minimal() +
  labs(
    title = "Posterior Estimates with 89% Credible Intervals",
    subtitle = "Predicting Divorce Rate from South and Number of Waffle Houses",
    x = "Parameter Estimate",
    y = "Parameters"
  )
```



Answer: We can see that with a model conditioned on if a state is in the south of the US or not, we can see that the number of Waffle Houses contains no information on the divorce rate of the sate, that the southernity does not already explain. From this we could conclude that it is just a spurious correlation between the number of Waffle Houses in a state and divorce rate, that is mediated by the connection of the location of the state (southern states having both more Waffle Houses and more divorces).

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_6h1)
```

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

```
#testing first one - A _||_ W | S
model_6h2_1 <- quap(
  alist(
    A ~ dnorm(mu, sigma),
    mu <- a + bS*S + bW*W,
    a ~ dnorm(0, 0.2),
    bS ~ dnorm(0, 0.5),
    bW ~ dnorm(0, 0.5),
    sigma ~ dexp(1)
  ), data = data
)

precis(model_6h2_1)
```

```
##                mean         sd       5.5%        94.5%
## a     -6.449371e-07 0.11136061 -0.1779764 0.1779751246
## bS    -2.717004e-01 0.17013265 -0.5436052 0.0002044409
## bW     6.350817e-02 0.17006951 -0.2082958 0.3353120915
## sigma  9.479849e-01 0.09356219  0.7984544 1.0975153177
```

```
#getting summary of model
summary_6h2_1 <- precis(model_6h2_1, depth = 2) %>% as.data.frame()
summary_6h2_1$parameter <- rownames(summary_6h2_1)

#plotting
plot1 <- ggplot(summary_6h2_1, aes(x = mean, y = reorder(parameter, mean)))
+
  geom_point(size = 3, color = "orange2") +
  geom_errorbarh(aes(xmin = `5.5%`, xmax = `94.5%`), height = 0.2, color =
"red4") +
  theme_minimal() +
  labs(
    title = "Posterior Estimates with 89% Credible Intervals",
    subtitle = "Predicting Median Age at Marriage from South and Number of W
affle Houses",
    x = "Parameter Estimate",
    y = "Parameters"
  )

#testing second one - D _||_ S | A, M, W
model_6h2_2 <- quap(
  alist(
    D ~ dnorm(mu, sigma),
    mu <- a + bS*S + bW*W + bA*A + bM*M,
    a ~ dnorm(0, 0.2),
    bS ~ dnorm(0, 0.5),
    bW ~ dnorm(0, 0.5),
    bA ~ dnorm(0, 0.5),
    bM ~ dnorm(0, 0.5),
    sigma ~ dexp(1)
  ), data = data
)
precis(model_6h2_2)
```

```
##                 mean         sd        5.5%      94.5%
## a      1.982964e-06 0.09456359 -0.15112890  0.1511329
## bS     1.452649e-01 0.14437919 -0.08548092  0.3760107
## bW     8.654976e-02 0.14050274 -0.13800075  0.3111003
## bA    -5.511781e-01 0.15155297 -0.79338902 -0.3089672
## bM    -3.621895e-02 0.14738127 -0.27176269  0.1993248
## sigma  7.588468e-01 0.07520434  0.63865577  0.8790379
```

```
#getting summary of model
summary_6h2_2 <- precis(model_6h2_2, depth = 2) %>% as.data.frame()
summary_6h2_2$parameter <- rownames(summary_6h2_2)

#plotting
plot2 <- ggplot(summary_6h2_2, aes(x = mean, y = reorder(parameter, mean)))
+
  geom_point(size = 3, color = "orange2") +
  geom_errorbarh(aes(xmin = `5.5%`, xmax = `94.5%`), height = 0.2, color =
"red4") +
  theme_minimal() +
  labs(
    title = "Posterior Estimates with 89% Credible Intervals",
    subtitle = "Predicting Divorce Rate from South, Medium Age at Marriage,
Marriage Rate and Number of Waffle Houses",
    x = "Parameter Estimate",
    y = "Parameters"
  )

#testing third one - M _||_ W | S
model_6h2_3 <- quap(
  alist(
    M ~ dnorm(mu, sigma),
    mu <- a + bS*S + bW*W,
    a ~ dnorm(0, 0.2),
    bS ~ dnorm(0, 0.5),
    bW ~ dnorm(0, 0.5),
    sigma ~ dexp(1)
  ), data = data
)
precis(model_6h2_3)
```

```
##                  mean         sd       5.5%     94.5%
## a       4.064910e-10 0.11363587 -0.1816121 0.1816121
## bS      1.006653e-01 0.17425231 -0.1778235 0.3791542
## bW     -3.892027e-02 0.17424388 -0.3173956 0.2395551
## sigma   9.764519e-01 0.09626046  0.8226091 1.1302947
```

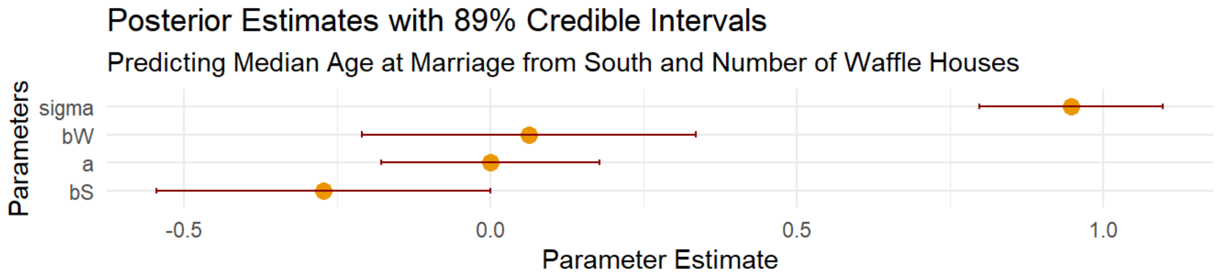
```
#getting summary of model
summary_6h2_3 <- precis(model_6h2_3, depth = 2) %>% as.data.frame()
summary_6h2_3$parameter <- rownames(summary_6h2_3)

#plotting
plot3 <- ggplot(summary_6h2_3, aes(x = mean, y = reorder(parameter, mean)))
+
  geom_point(size = 3, color = "orange2") +
  geom_errorbarh(aes(xmin = `5.5%`, xmax = `94.5%`), height = 0.2, color =
"red4") +
  theme_minimal() +
  labs(
    title = "Posterior Estimates with 89% Credible Intervals",
    subtitle = "Predicting Marriage Rate from South and Number of Waffle Hou
ses",
    x = "Parameter Estimate",
    y = "Parameters"
  )

library(gridExtra)
```

```
## Warning: a(z) 'gridExtra' csomag az R 4.4.2 verziójával lett fordítva
```

```
##
## Kapcsolódás csomaghoz: 'gridExtra'
```

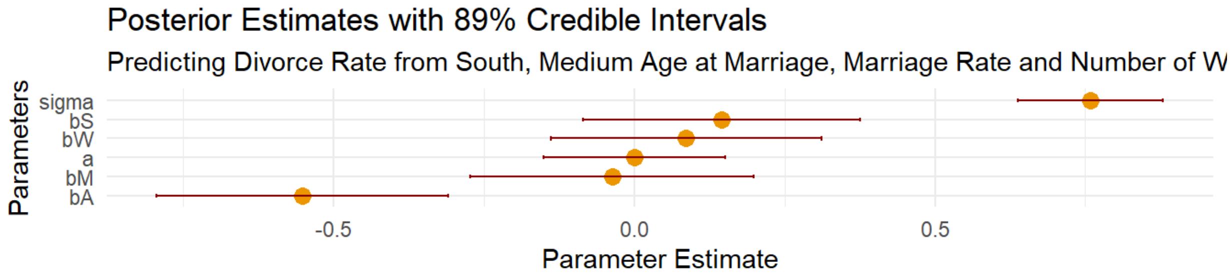
```
## The following object is masked from 'package:dplyr':
##
##     combine
```

```
grid.arrange(plot1, plot2, plot3, nrow = 3)
```



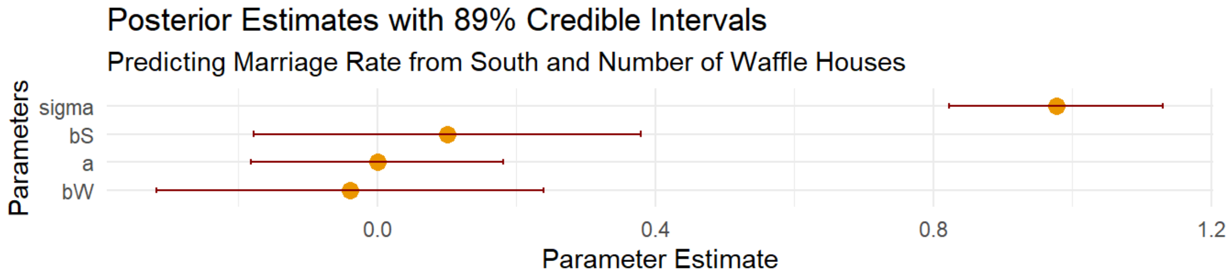
Posterior Estimates with 89% Credible Intervals

Predicting Median Age at Marriage from South and Number of Waffle Houses



Posterior Estimates with 89% Credible Intervals

Predicting Divorce Rate from South, Medium Age at Marriage, Marriage Rate and Number of W



Posterior Estimates with 89% Credible Intervals

Predicting Marriage Rate from South and Number of Waffle Houses

Answer: - From the first test we can see that median age and the number of Waffle Houses are independent form each other when controlling for the southernity of the state. - From the second test we can see that even when controlling for median age, number of Waffle Houses and marriage rate, sourthernity still has a significant effect on divorce rate. This suggest that we should have a direct relationship between S and D in our DAG graph. - from the third test we can see that marriage rate and the number of Waffle Houses are independent from each other when controlling for the southernity of the state.