Portfolio 2, Methods 4, 2025

Study group 9

Reka Forgo (RF), Clara Holst (CH), Frederikke Lykke Korreborg (FLK), Zuzanna Zyla (ZZ)

2024-09-26

Second assignment

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

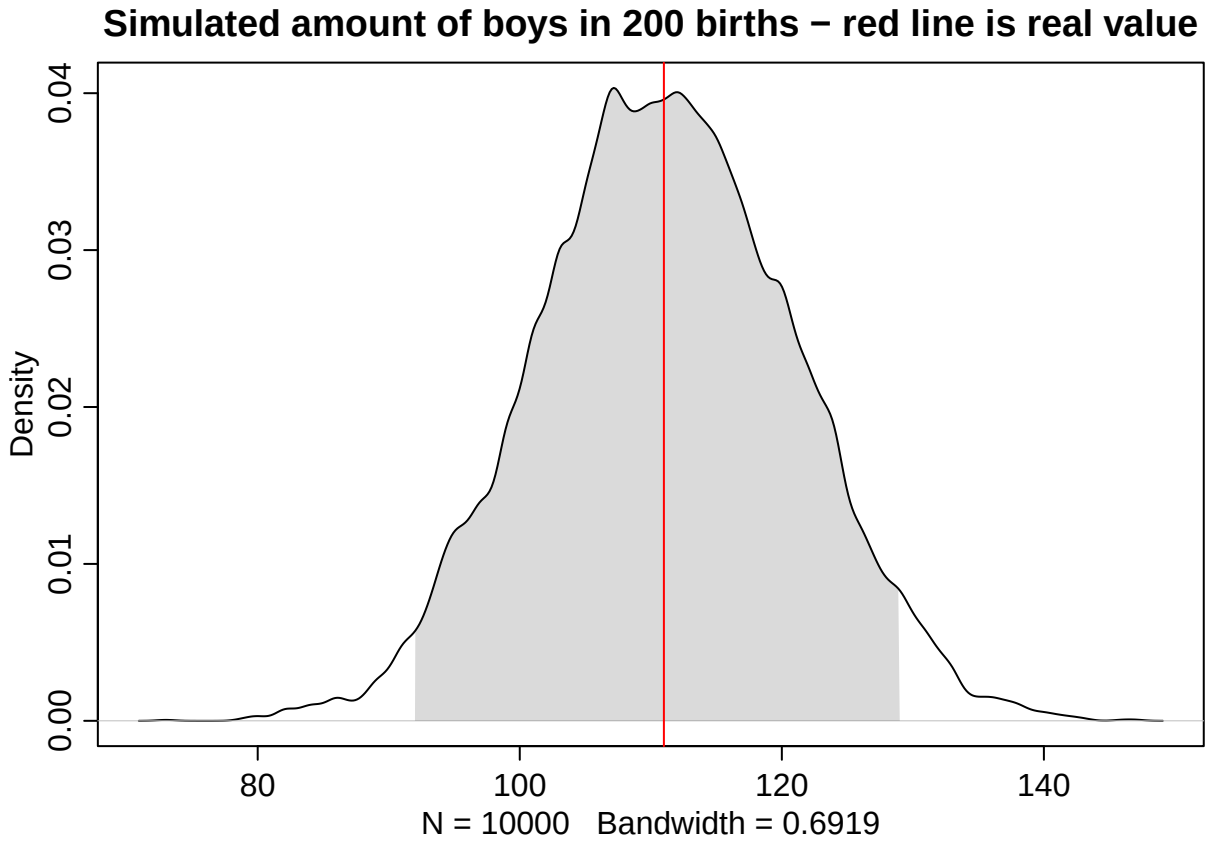
Chapter 3: Causal Confussion

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).

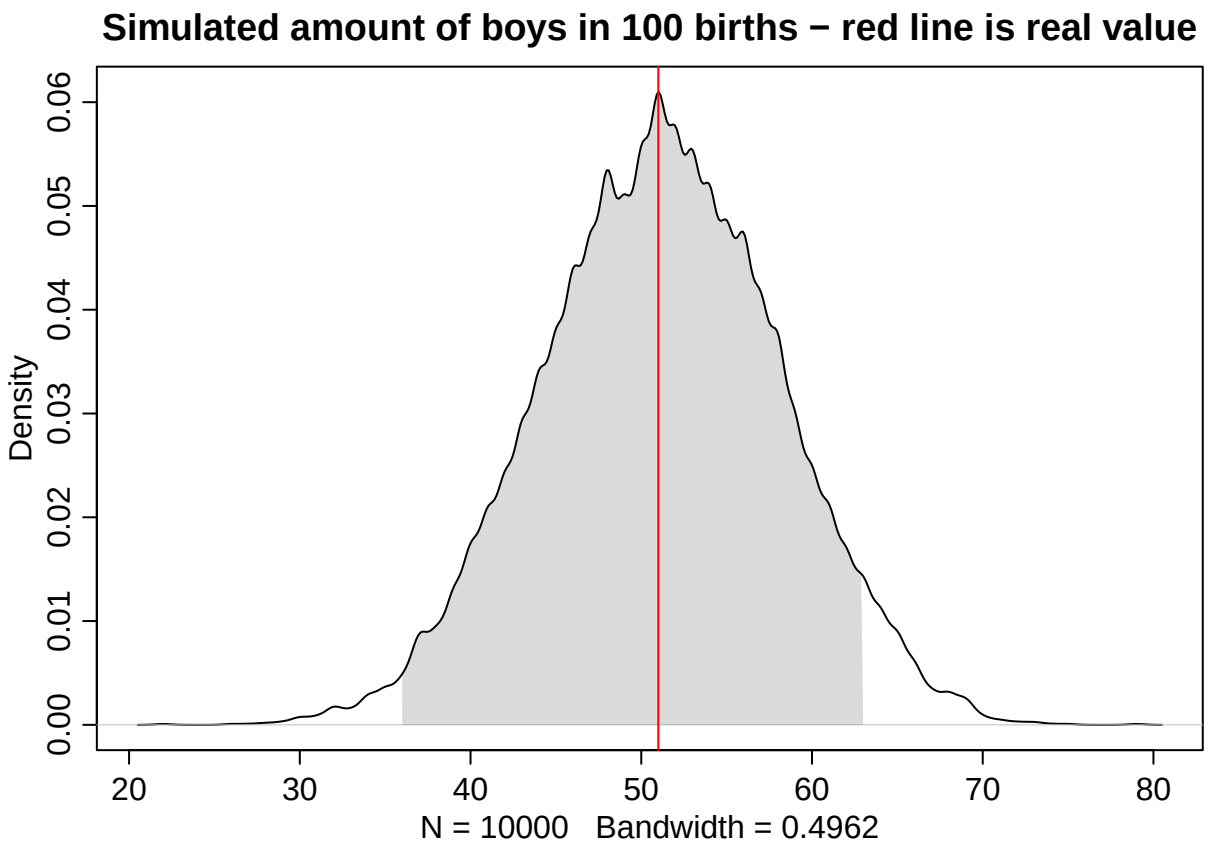
```
# 3H1
# Find the posterior probability of giving birth to a boy:
pacman::p_load(rethinking,
               dplyr,
               tidyverse)
data(homeworkch3)
set.seed(1)
W <- sum(birth1) + sum(birth2)
N <- length(birth1) + length(birth2)
p_grid <-seq(from =0, to = 1, len =1000)
prob_p <- rep(1,1000)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

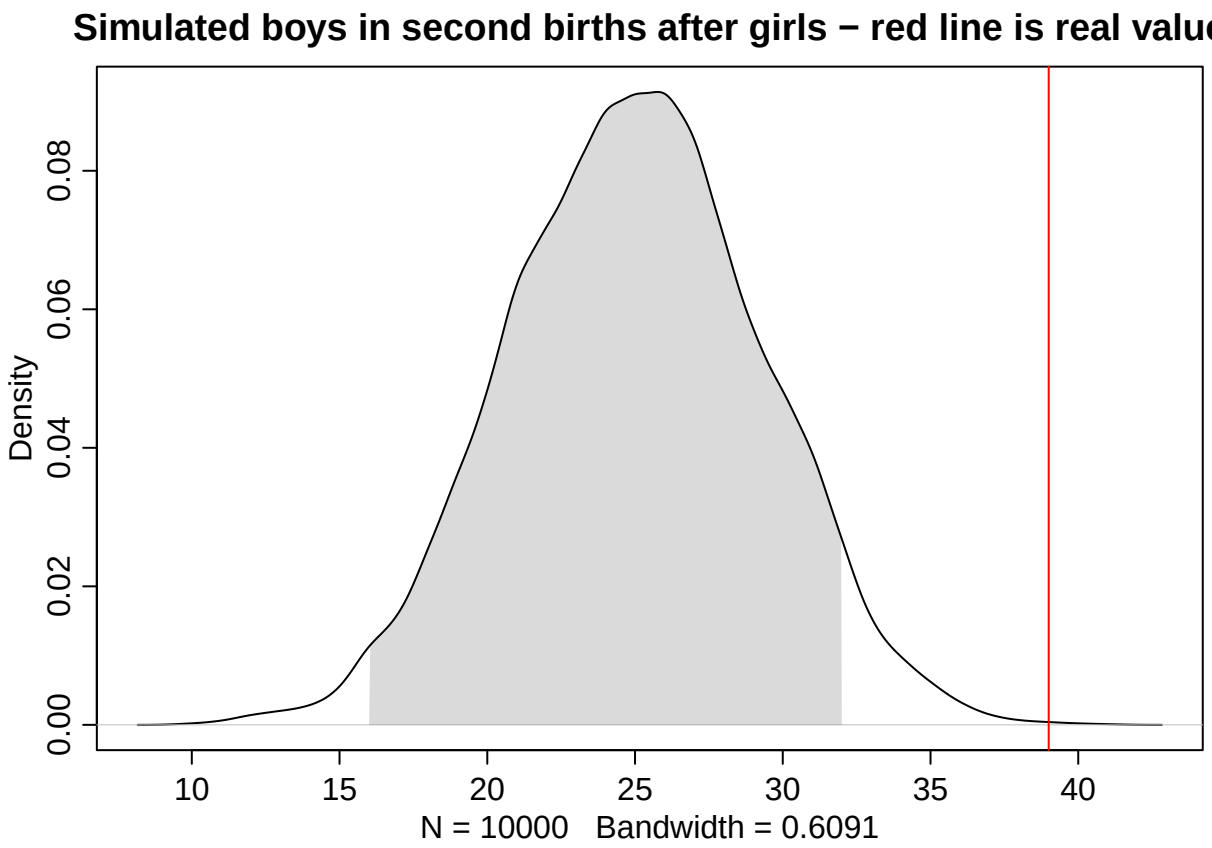
```
#compare
set.seed(1)
W <- sum(birth1)
N <- length(birth1)

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)

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

# 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 100 births - red line is real value")
```



Answer 3H4. In this light the model looks like it is a good fit and the simulation values align with the real one.

3H5. (CH) 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 cound 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)

# 3H5
# Count observed boys in these second births
W <- sum(birth2[birth1 == 0])  # Count of boys in these second births
N <- length(birth2[birth1 == 0])  # Total number of second births after girls


# Simulate second births following girls
simulated_births <- rbinom(n = 1e4, size = N, prob = samples)

# Plot results
rethinking::dens(simulated_births, show.HPDI = 0.95, adj = 1)
abline(v = W, col = "red")
title("Simulated boys in second births after girls - red line is real value")
```



Answer 3H5. What is happening in this simulation: The plot shows that the real value is way off from the confidence interval of what our model predicts. So, it seems to underestimate the probability of having a boy after a girl. The model might overestimate some components, like a genetic predisposition in the mother that favors consecutive girls. Our model is assuming that the sex of first and second borns are independent, but from the data it seems like they are not.

Chapter 5: Spurrious Correlations

Start of by checking out all the spurious 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.(FLK) 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 5M1. Outcome: number of drownings. Predictors: ice cream sales and temperature. Ice cream sales correlate with drownings. But both are caused by high temperature. In a multiple regression including both predictors, the effect of ice cream should disappear.

5M2.(ZZ) 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 5M2. Outcome: GPA Predictors: study hours (positive effect) and party hours (negative effect) They two predictors are negatively correlated. Without adjusting for both, effects can cancel out. A multiple regression shows their independent effects clearly.

5M3.(RF) It is sometimes observed that the best predictor of fire risk is the presence of firefighters—States and localities with many firefighters also have more fires. Presumably firefighters do not cause fires. Nevertheless, this is not a spurious correlation. Instead fires cause firefighters. Consider the same reversal of causal inference in the context of the divorce and marriage data. How might a high divorce rate cause a higher marriage rate? Can you think of a way to evaluate this relationship, using multiple regression.

Answer 5M3 High divorce rates can cause high marriage rate for ex. with remarriages. Multiple regression: marriage_rate ~ divorce_rate + percent_remarriages. This tests if divorce rate predicts marriage rate, accounting for remarriages.

5M5.CH) 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 5M5. We can use the following models to test the two mechanisms.

Baseline model:

$$ObesityRate_i \sim \alpha + \beta_G GasPrice_i$$

Exercise model:

$$ObesityRate_i \sim \alpha + \beta_G GasPrice_i + \beta_E ExerciseRate_i$$

Restaurant model:

$$ObesityRate_i \sim \alpha + \beta_G GasPrice_i + \beta_R RestaurantSpending_i$$

Full model:

$$ObesityRate_i \sim \alpha + \beta_G GasPrice_i + \beta_E ExerciseRate_i + \beta_R RestaurantSpending_i$$

Chapter 5: Foxes and Pack Sizes

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. (FLK) 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?

```
#load the data
data("foxes")

#standardise the data
foxes <- foxes %>%
  mutate(across(-group, standardize))
```

```r
#fitting first gaussian regression using quap.
#body weight as linear function of territory size(area)
# 1. Regression: Weight ~ Territory Size (Area)
set.seed(1)

m1 <- quap(
  alist(
    weight ~ dnorm(mu, sigma),
    mu <- a + b_area * area,
    a ~ dnorm(0, 0.2),
    b_area ~ dnorm(0, 0.5),
    sigma ~ dexp(1)
  ),
  data = foxes
)

#2
m2 <- quap(
  alist(
    weight ~ dnorm(mu, sigma),
    mu <- a + b_groupsize * groupsize,
    a ~ dnorm(0, 0.2),
    b_groupsize ~ dnorm(0, 0.5),
    sigma ~ dexp(1)
  ),
  data = foxes
)


#Plot the results of these regressions, displaying the MAP
#regression line and the 95% interval of the mean

precis(m1)
```

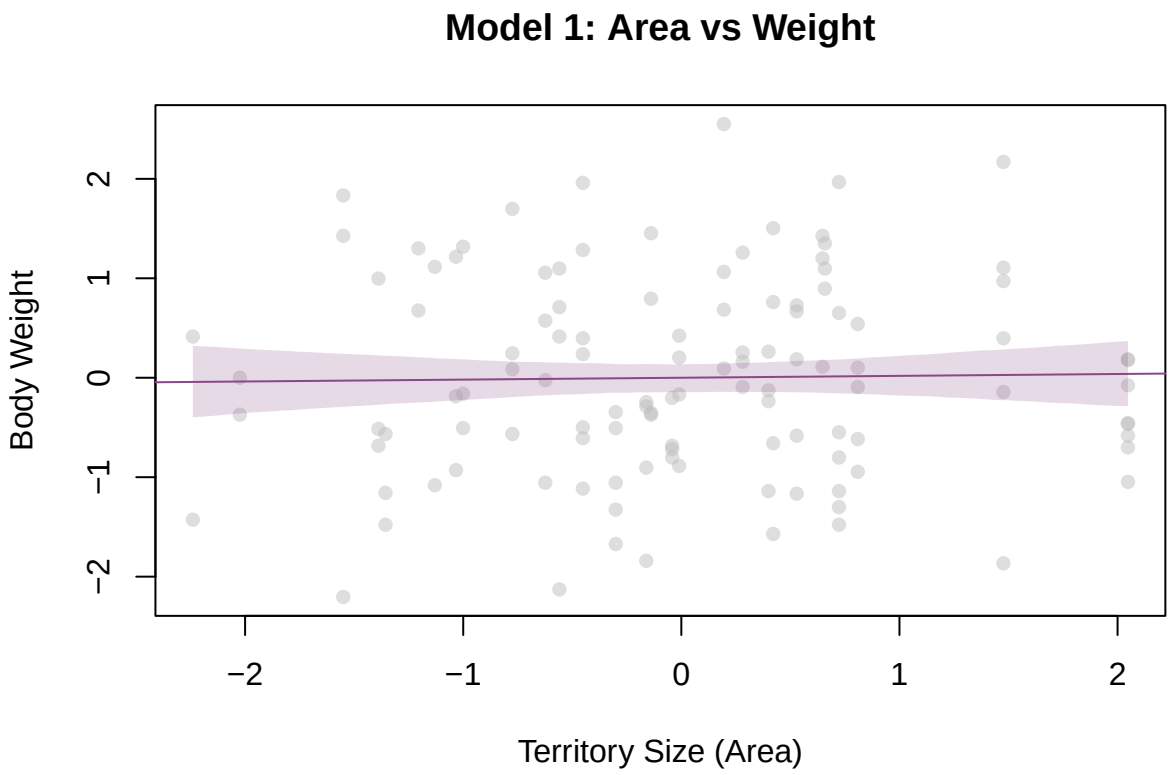
```
##                 mean         sd       5.5%     94.5%
## a      -4.833170e-08 0.08360861 -0.1336228 0.1336227
## b_area  1.883334e-02 0.09089575 -0.1264356 0.1641023
## sigma   9.912653e-01 0.06466635  0.8879159 1.0946146
```

```r
precis(m2)
```

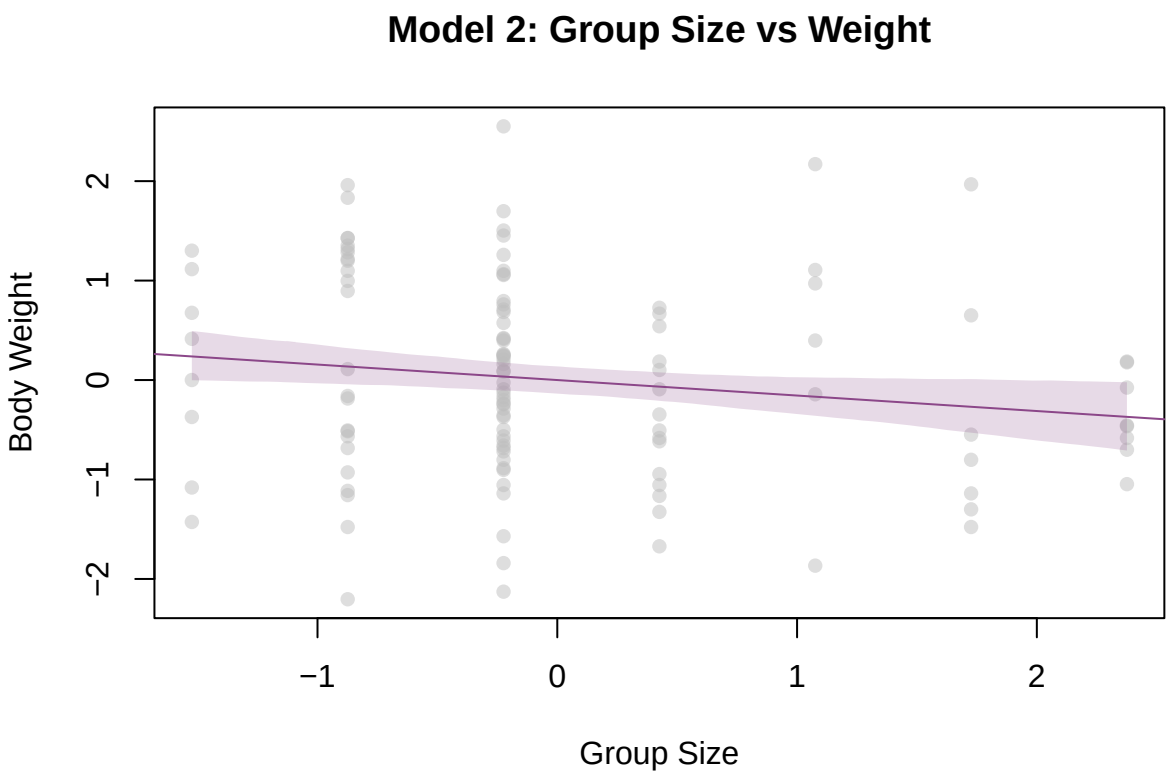
```
##                      mean         sd       5.5%       94.5%
## a           -5.758899e-07 0.08272193 -0.1322062  0.13220505
## b_groupsize -1.558037e-01 0.08977167 -0.2992762 -0.01233124
## sigma        9.785691e-01 0.06384323  0.8765353  1.08060293
```

```r
# Superimpose the MAP regression line and it's 89% interval
plot(foxes$area, foxes$weight, col = col.alpha("grey", 0.5), pch = 16,
     xlab = "Territory Size (Area)", ylab = "Body Weight")
abline( a=coef(m1)["a"], b=coef(m1)["b_area"], col = "orchid4" ) #regression line
weight_seq <- seq(min(foxes$area), max(foxes$area), length.out = 100)
mu_samples <- link(m1, data = data.frame(area = weight_seq)) # Posterior samples
```

```r
interval_mean_w <- apply(mu_samples, 2, PI, prob = 0.89)
shade(interval_mean_w, weight_seq, col = col.alpha("orchid4", 0.2))
title("Model 1: Area vs Weight")
```




```r
#now for model 2
plot(foxes$groupsize, foxes$weight, col = col.alpha("grey", 0.5), pch = 16,
     xlab = "Group Size", ylab = "Body Weight")
abline( a=coef(m2)["a"], b=coef(m2)["b_groupsize"], col = "orchid4" )
groupsize_seq <- seq(min(foxes$groupsize), max(foxes$groupsize), length.out = 100)
mu_samples <- link(m2, data = data.frame(groupsize = groupsize_seq))
interval_mean_g <- apply(mu_samples, 2, PI, prob = 0.89)
shade(interval_mean_g, groupsize_seq, col = col.alpha("orchid4", 0.2))
title("Model 2: Group Size vs Weight")
```



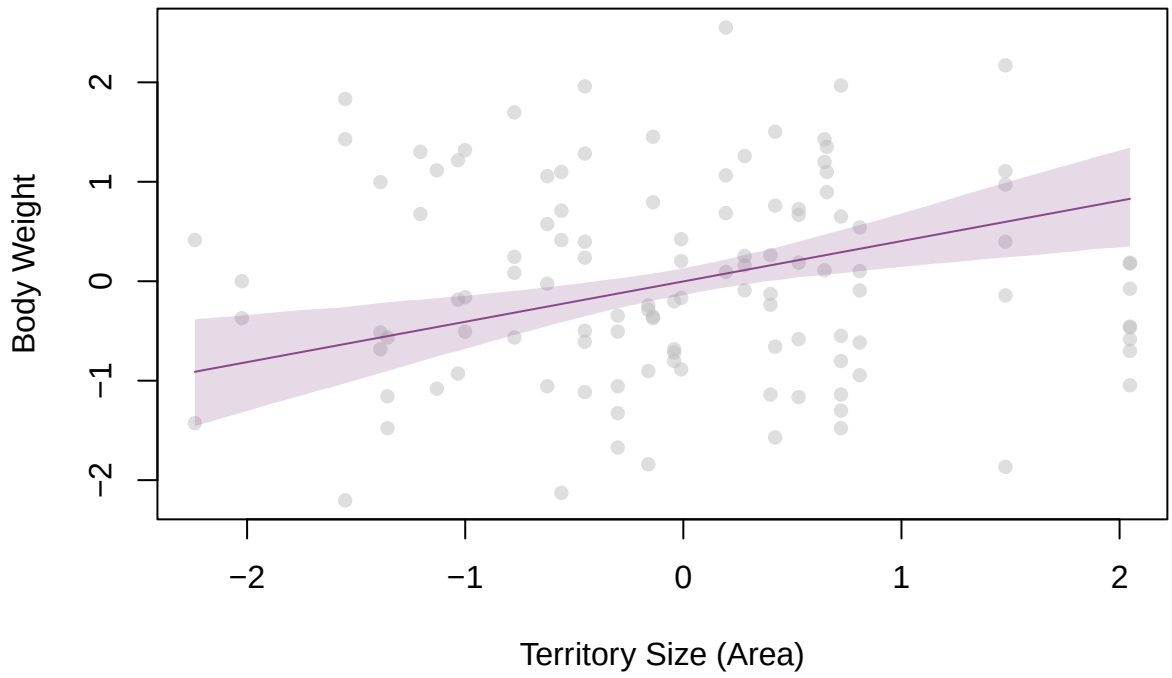
Answer 5H1. While area shows no consistent trend, Groupsize seems to be negatively correlated to weight, but not a strong association.

5H2. (ZZ) 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?

```
set.seed(1)

#fitting multiple linear regression with weight as outcome and both area and
#groupsize as predictor variables

m3 <- quap(
  alist(
    weight ~ dnorm(mu, sigma),
    mu <- a + b_area * area + b_groupsize * groupsize,
    a ~ dnorm(0, 0.2),
    b_area ~ dnorm(0, 0.5),
    b_groupsize ~ dnorm(0, 0.5),
    sigma ~ dexp(1)
  ),
  data = foxes
)

precis(m3)
```

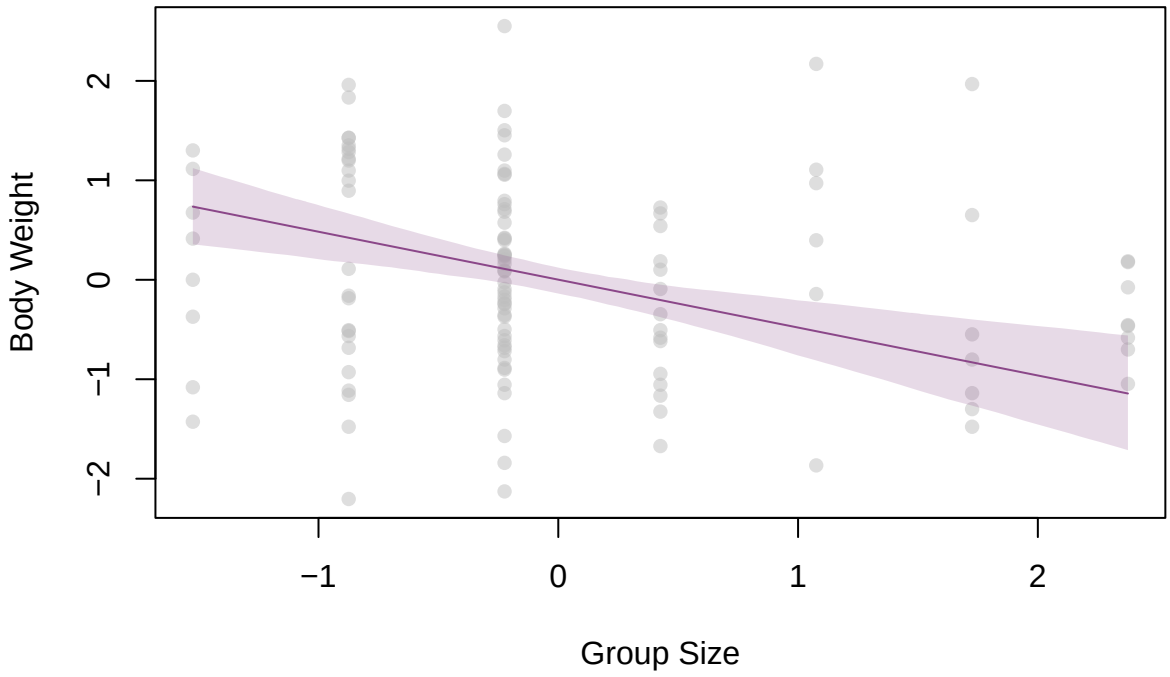
```
##                      mean         sd       5.5%      94.5%
## a           -3.213848e-08 0.08013106 -0.1280649  0.1280649
## b_area       4.058526e-01 0.14536262  0.1735351  0.6381702
## b_groupsize -4.820000e-01 0.14537265 -0.7143336 -0.2496664
## sigma        9.419458e-01 0.06159413  0.8435065  1.0403851
```

```
#extracting posterior samples while holding the other predictor constant at its mean
weight_seq <- seq(min(foxes$area), max(foxes$area), length.out = 100)
samples_1 <- link(m3, data = data.frame(area = weight_seq,
                                        groupsize = mean(foxes$groupsize)))

weight_seq2 <- seq(min(foxes$groupsize), max(foxes$groupsize), length.out = 100)
samples_2 <- link(m3, data = data.frame(groupsize = weight_seq2,
                                        area = mean(foxes$area)))

#plotting the predictions of the model for holding area constant at its mean
plot(foxes$area, foxes$weight, col = col.alpha("grey", 0.5), pch = 16,
     xlab = "Territory Size (Area)", ylab = "Body Weight")
interval_mean <- apply(samples_1, 2, PI, prob = 0.89)
shade(interval_mean, weight_seq, col = col.alpha("orchid4", 0.2))
lines(weight_seq, apply(samples_1, 2, mean), col = "orchid4")
```




```
#plotting the predictions of the model for holding groupsize constant at its mean
plot(foxes$groupsize, foxes$weight, col = col.alpha("grey", 0.5), pch = 16,
     xlab = "Group Size", ylab = "Body Weight")
```

```
interval_mean2 <- apply(samples_2, 2, PI, prob = 0.89)
shade(interval_mean2, weight_seq2, col = col.alpha("orchid4", 0.2))
lines(weight_seq2, apply(samples_2, 2, mean), col = "orchid4")
```



Answer 5H2: It seems there is a masked relationship between the variables. Area is positively related to weight, while group size is negatively related, hence canceling each other out. Using multiple regression, we can see the real effects that was not apparent in the bivariate models.

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?

```
#fitting 1: body weight + avgfood + groupsize

m4 <- quap(
  alist(
    weight ~ dnorm(mu, sigma),
    mu <- a + b_avgfood * avgfood + b_groupsize * groupsize,
    a ~ dnorm(0, 0.2),
    b_avgfood ~ dnorm(0, 0.5),
    b_groupsize ~ dnorm(0, 0.5),
    sigma ~ dexp(1)
```

```
  ),
  data = foxes
)


#fitting 2: body weight + avgfood + groupsize + area

m5 <-  quap(
  alist(
    weight ~ dnorm(mu, sigma),
    mu <- a + b_avgfood * avgfood + b_groupsize * groupsize + b_area * area,
    a ~ dnorm(0, 0.2),
    b_avgfood ~ dnorm(0, 0.5),
    b_groupsize ~ dnorm(0, 0.5),
    b_area ~ dnorm(0, 0.5),
    sigma ~ dexp(1)
  ),
  data = foxes
)
```

Answer 5H3. a. (RF)

In m4 (without area), b_avgfood is strong (0.48) with a CI not overlapping with 0. In m5 (including area as a predictor), b_avgfood shrinks to 0.30, and b_area is 0.28 In m5 two of the coefficients have CIs that include zero, meaning that they are not as strong predictors of body weight. Choosing the model 4 with avgfood as a predictor is better than model 5 which also includes area as a predictor.

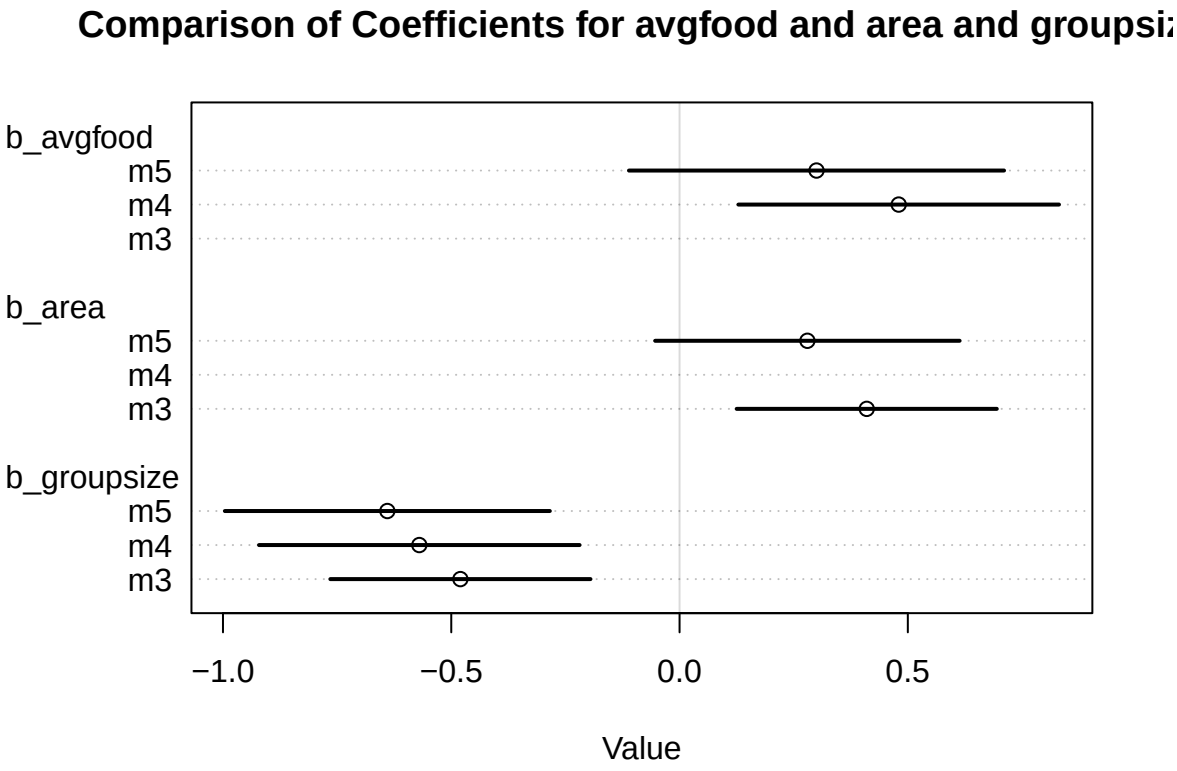
```
precis(m4)
```

```
##                       mean         sd       5.5%      94.5%
## a           -1.868764e-05 0.08014104 -0.1280996  0.1280622
## b_avgfood    4.773359e-01 0.17912849  0.1910540  0.7636178
## b_groupsize -5.735090e-01 0.17914816 -0.8598224 -0.2871956
## sigma        9.420856e-01 0.06175937  0.8433822  1.0407890
```

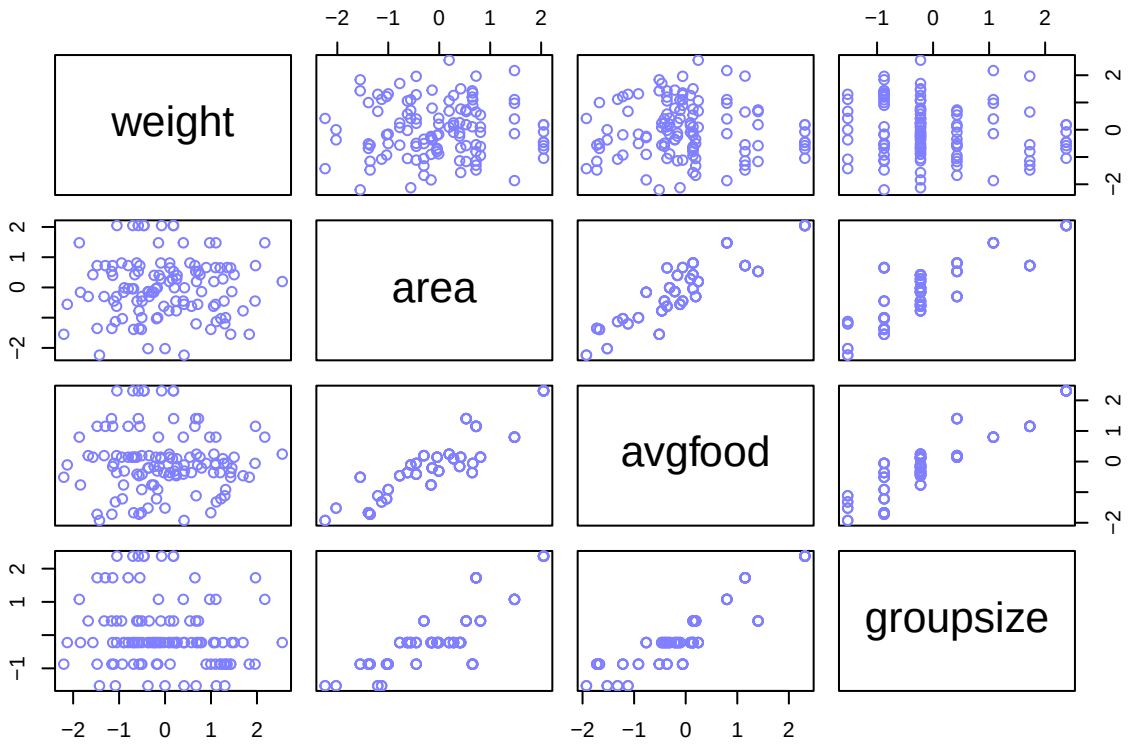
```
precis(m5)
```

```
##                       mean         sd        5.5%      94.5%
## a            7.075997e-08 0.07936194 -0.12683563  0.1268358
## b_avgfood    2.969009e-01 0.20960006 -0.03808052  0.6318822
## b_groupsize -6.396185e-01 0.18161469 -0.92987388 -0.3493632
## b_area       2.782369e-01 0.17011212  0.00636487  0.5501089
## sigma        9.312053e-01 0.06099992  0.83371569  1.0286950
```

```
# Compare coefficients for avgfood and area
plot(coeftab(m3, m4, m5), pars = c("b_avgfood", "b_area", "b_groupsize"),
     main = "Comparison of Coefficients for avgfood and area and groupsize")
```




```
cor <- cor(foxes$avgfood, foxes$area)

pairs(~ weight+area + avgfood+groupsize, data=foxes, col=rangi2)
```



Answer 5H3. b. (CH)

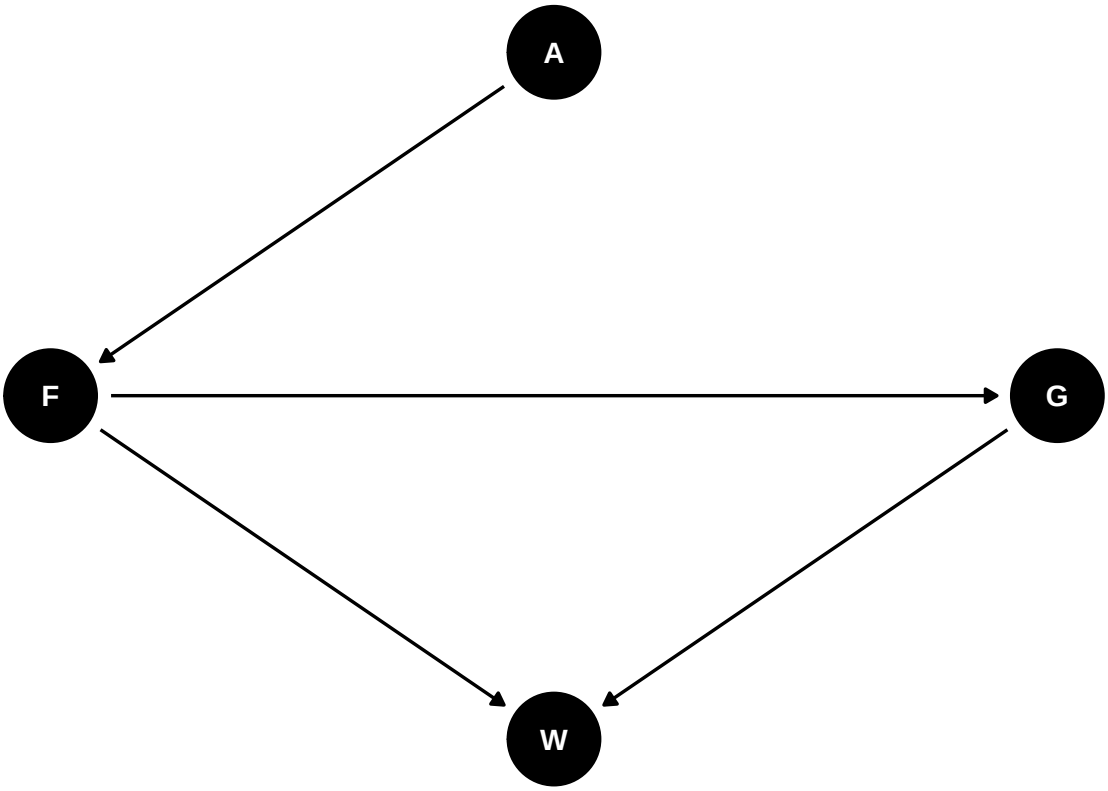
Adding area weakens avgfood's effect, suggesting multicollinearity. The correlation between avgfood and area is `cor`, which suggests they are highly correlated. This means that the model is trying to estimate the effect of each variable while controlling for the other, which can lead to smaller effect sizes and larger standard errors.

—- Here we have different solutions —-

Defining our theory with explicit DAGs (FLK) Assume this DAG as an causal explanation of fox weight:

```
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()
```




```
impliedConditionalIndependencies(dag)
```

```
## A _||_ G | F
## A _||_ W | F
```

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? (ZZ)

```
#The only direct path from A (Area) to F (Avgfood) is A → F.
#Since this is a direct causal link, the total causal effect of A on F
#is simply the coefficient of A in this relationship.

# Estimate the total causal influence of A on F
set.seed(1)
m6 <- quap(
  alist(
    avgfood ~ dnorm(mu, sigma),
    mu <- a + b_area * area,
```

```
    a ~ dnorm(0, 0.2),
    b_area ~ dnorm(0, 0.5),
    sigma ~ dexp(1)
  ),
  data = foxes
)

precis(m6)
```

```
##                mean         sd        5.5%      94.5%
## a      0.0000134981 0.04231668 -0.06761673 0.06764372
## b_area 0.8765112263 0.04332971  0.80726198 0.94576047
## sigma  0.4663221418 0.03053500  0.41752131 0.51512298
```

Answer 1: The total causal influence of A on F is 0.88, with a 95% credible interval of [0.81, 0.95]. This means that increasing the area of a territory would increase the amount of food. This makes sense: larger areas contain more food sources.

—- Solution 1, leads to same conclusion but different implementation —-

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? (RF)

```
#The total causal effect of adding food F to a territory on the weight W of foxes
#can be calculated using the coefficients from the models we have already fitted.

#       F → W  (direct)
#       F → G → W (indirect via groupsize)

#we can use the estimate from a previous model m4: avgfood ~ weight + groupsize
set.seed(1)

#precis(m4)
coef_avgfood <- coef(m4)["b_avgfood"]
print(coef_avgfood)
```

```
## b_avgfood
## 0.4773359
```

```
#Since food also affects groupsize, which then affects weight,
# we need a model for groupsize ~ food
m7 <- quap(
  alist(
    groupsize ~ dnorm(mu, sigma),
    mu <- a + b_avgfood * avgfood,
    a ~ dnorm(0, 0.2),
    b_avgfood ~ dnorm(0, 0.5),
    sigma ~ dexp(1)
  ),
  data = foxes
)
```

```r
#precis(m7)
coef_avgfood_groupsize <- coef(m7)["b_avgfood"]

# now estimate how groupsize affects weight (W ~ G) - this is our model m2
#precis(m2)
coef_groupsize <- coef(m2)["b_groupsize"]


# Compute Total Effect
total_effect <- coef_avgfood + (coef_avgfood_groupsize * coef_groupsize)
print(total_effect)
```

```
## b_avgfood
## 0.3377798
```

```r
#simulation
sim_data <- foxes
sim_data$avgfood <- sim_data$avgfood + 1 #Simulating a 1-unit increase in avgfood

# Adjust groupsize using F → G effect
sim_data$groupsize <- sim_data$groupsize + coef_avgfood_groupsize

# Predict new weights using both direct (m4) and indirect effects
mu_new <- link(m4, data = sim_data) + (coef_groupsize * coef_avgfood_groupsize)

# Get mean predicted weight
mean_weight <- apply(mu_new, 2, mean)

# Compare new vs old weights
mean(mean_weight - foxes$weight)  # Total causal effect of increasing food
```

```
## [1] -0.1735432
```

A negative total effect suggests that adding food actually reduces fox weight overall. This is surprising at first because we expect more food → higher weight. However, this result makes sense given the DAG structure.

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? (CH)

```r
# Direct Effect
direct_effect <- coef_avgfood
print(direct_effect)
```

```
## b_avgfood
## 0.4773359
```

The direct effect of food on weight is positive, but the indirect effect through groupsize is negative. The negative indirect effect is stronger than the positive direct effect, leading to an overall negative total effect. What is happening to the foxes: More food leads to larger groups, which in turn leads to lower weight. This suggests that larger groups may have to share food resources, leading to lower individual weights.

—- Solution 2, leads to same conclusion but different implementation —-

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? (FLK)

```
set.seed(1)

# There is both a path from F -> W and F -> G -> W, so we need to account
# for both paths and therefore do not want to include groupsize as a variable

m_w <- quap(
  alist(
    weight ~ dnorm(mu, sigma),
    mu <- a + baf * avgfood,
    a ~ dnorm(0, 0.2),
    baf ~ dnorm(0, 0.5),
    sigma ~ dexp(1)
  ),
  data = foxes
)
# Model output
precis(m_w)
```

```
##                mean         sd       5.5%     94.5%
## a     -7.828488e-10 0.08360017 -0.1336092 0.1336092
## baf   -2.421163e-02 0.09088502 -0.1694634 0.1210402
## sigma  9.911440e-01 0.06465859  0.8878071 1.0944809
```

Answer 2): We cannot calculate the causal effect of an intervention on food as area is not included in the model and therefore the model is not aware of the relationship between the two. If we included more predictors in the model then it would no longer be estimating the total causal effect.

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? (ZZ)

```
set.seed(1)

#accounting for groupsize to get the direct effect
m_w_direct <- quap(
  alist(
    weight ~ dnorm(mu, sigma),
    mu <- a + baf * avgfood + bg * groupsize,
    a ~ dnorm(0, 0.2),
    baf ~ dnorm(0, 0.5),
    bg ~ dnorm(0, 0.5),
    sigma ~ dexp(1)
  ),
  data = foxes
)
# Model output
precis(m_w_direct)
```

```
##                mean         sd       5.5%      94.5%
## a     -6.179144e-07 0.08013805 -0.1280767  0.1280755
## baf    4.772541e-01 0.17912317  0.1909806  0.7635275
## bg    -5.735266e-01 0.17914167 -0.8598296 -0.2872236
## sigma  9.420437e-01 0.06175251  0.8433512  1.0407361
```

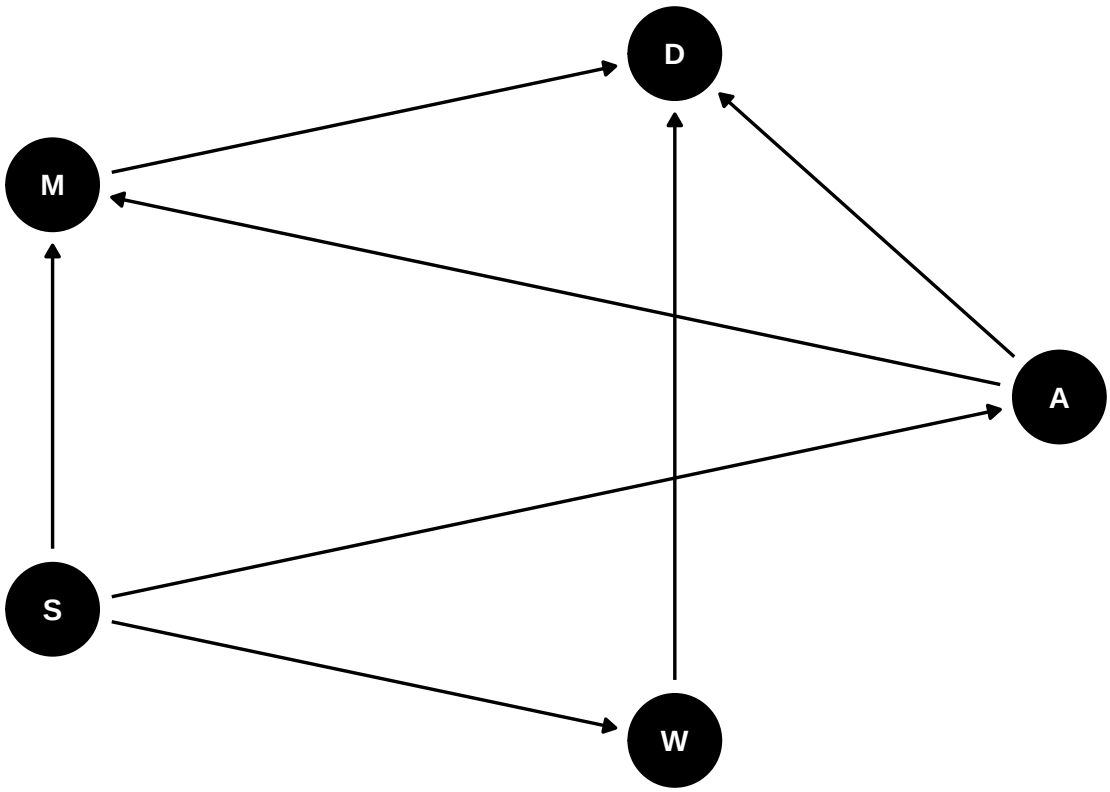
Answer 3): The last model shows that weight increases when average food available increases, but at the same time an increase in groupsize decreases weight. This explains the estimate from the first model, where avgfood had almost no effect on weight, as these two variables create a masked relationship.

—- Continue with Assignment —-

Chapter 6: Investigating the Waffles and Divorces

6H1. (RF) 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.

```
# Load data
data(WaffleDivorce)

# Standardize the variables
WaffleDivorce$WaffleHouses <- standardize(WaffleDivorce$WaffleHouses)
WaffleDivorce$Divorce <- standardize(WaffleDivorce$Divorce)
WaffleDivorce$MedianAgeMarriage <- standardize(WaffleDivorce$MedianAgeMarriage)
WaffleDivorce$Marriage <- standardize(WaffleDivorce$Marriage)
WaffleDivorce$South <- standardize(WaffleDivorce$South)

# Define DAG paths (from Statistical Rethinking page 187)
dag <- dagitty("dag {
    S [Southern State]
    A [Median Age Marriage]
    M [Marriage Rate]
    W [Waffle House Count]
    D [Divorce Rate]
    A -> D
    A -> M -> D
    A <- S -> M
    S -> W -> D
}")

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




```
# Model divorce rate as a function of number of Waffle Houses alone
m_w <- quap(
  alist(
    Divorce ~ dnorm(mu, sigma),
    mu <- a + bw * WaffleHouses,
    a ~ dnorm(0, 0.2),
    bw ~ dnorm(0, 0.5),
    sigma ~ dexp(1)
  ),
  data = WaffleDivorce
)
precis(m_w)
```

```
##                mean         sd        5.5%     94.5%
## a     -1.398555e-05 0.11140765 -0.17806493 0.1780370
## bw     2.371383e-01 0.13082987  0.02804692 0.4462297
## sigma  9.485654e-01 0.09356338  0.79903305 1.0980977
```

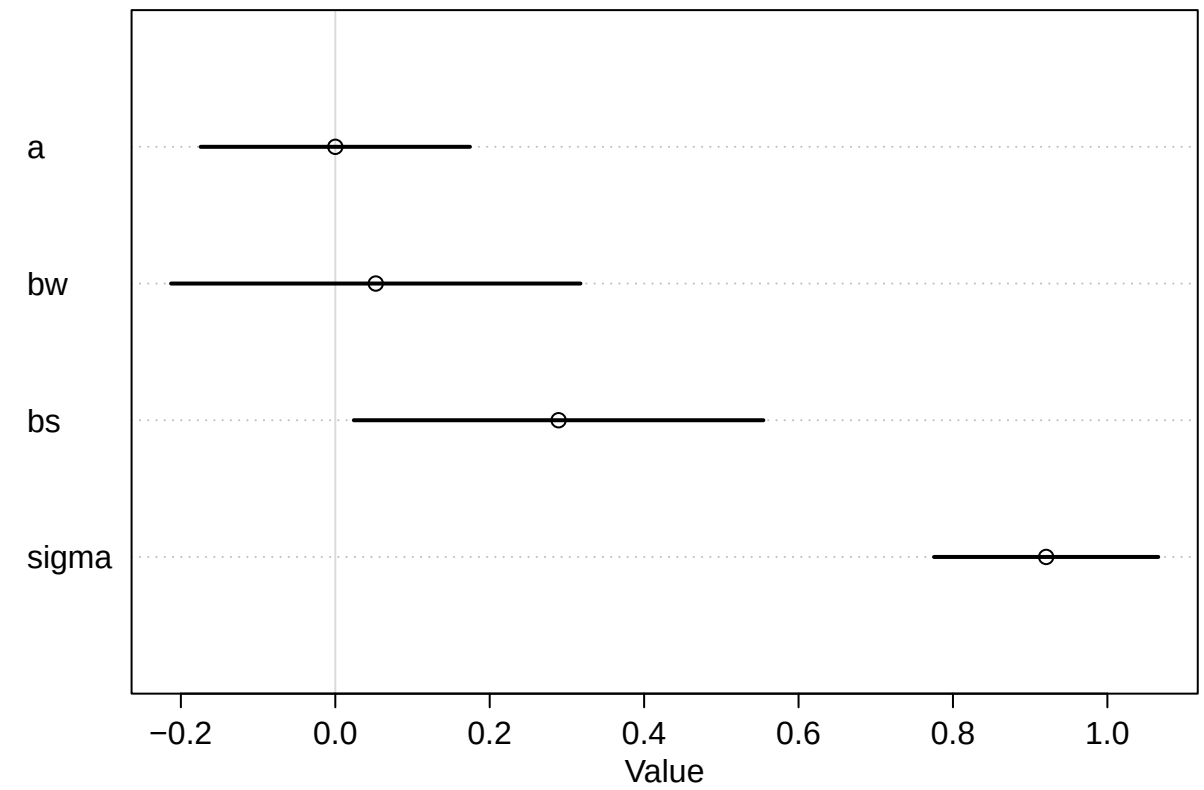
```
# Help decide which variable to condition on
adjustmentSets(dag, exposure="W", outcome="D")
```

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

```
# Model divorce rate as a function of number of Waffle Houses
# conditioning on Southern State
m_ws <- quap(
  alist(
    Divorce ~ dnorm(mu, sigma),
    mu <- a + bw * WaffleHouses + bs * South,
    a ~ dnorm(0, 0.2),
    bw ~ dnorm(0, 0.5),
    bs ~ dnorm(0, 0.5),
    sigma ~ dexp(1)
  ),
  data = WaffleDivorce
)
precis(m_ws)
```

```
##                mean         sd        5.5%     94.5%
## a      3.514652e-08 0.10911702 -0.17439004 0.1743901
## bw     5.241289e-02 0.16588052 -0.21269622 0.3175220
## bs     2.892013e-01 0.16594817  0.02398405 0.5544185
## sigma  9.206716e-01 0.09087323  0.77543865 1.0659046
```

```
plot(precis(m_ws))
```



Answer 6H1: The models show that when whether the location is a southern state or not is included in the model, then the amount of Waffle Houses in a state does not provide any more new information about

divorce rates. This suggests that the number of Waffle Houses does not have a causal influence on divorce rates, but that the relationship is confounded by the fact that southern states have more Waffle Houses and higher divorce rates.

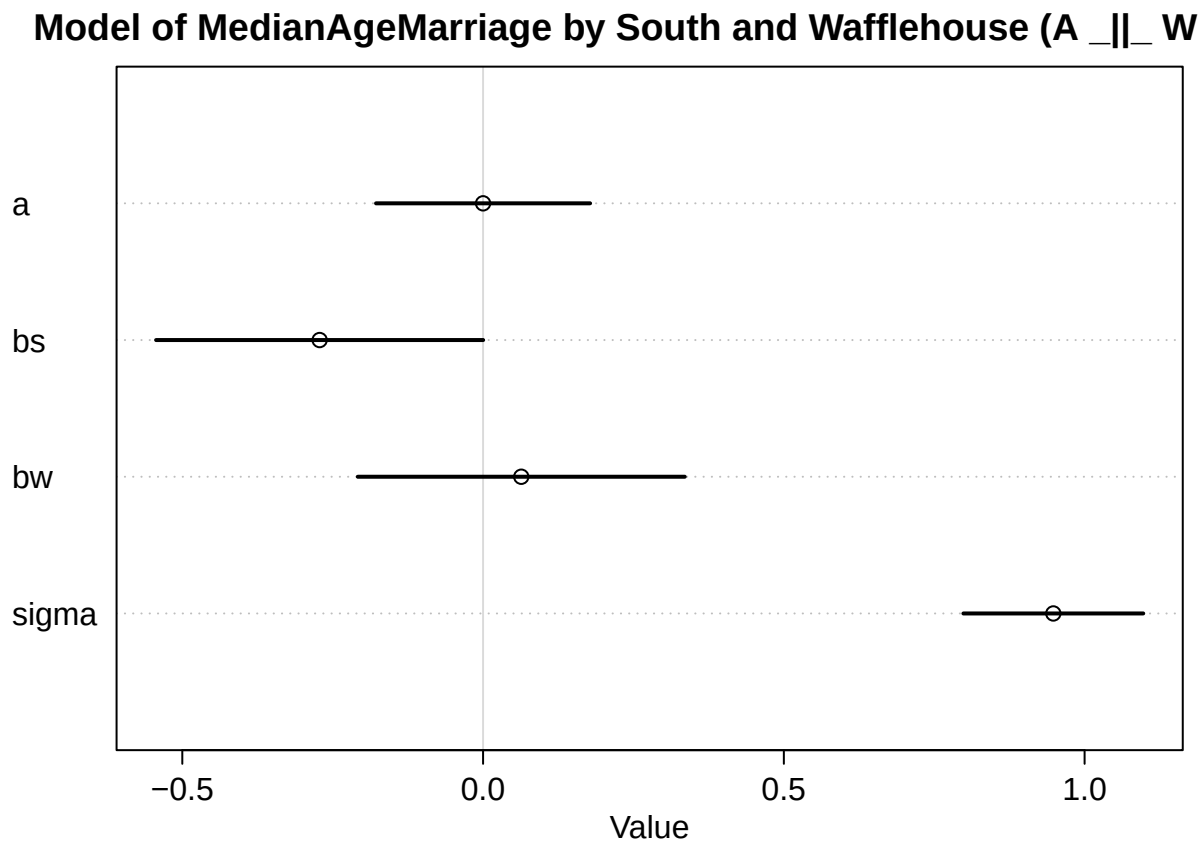
6H2. (CH)

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.

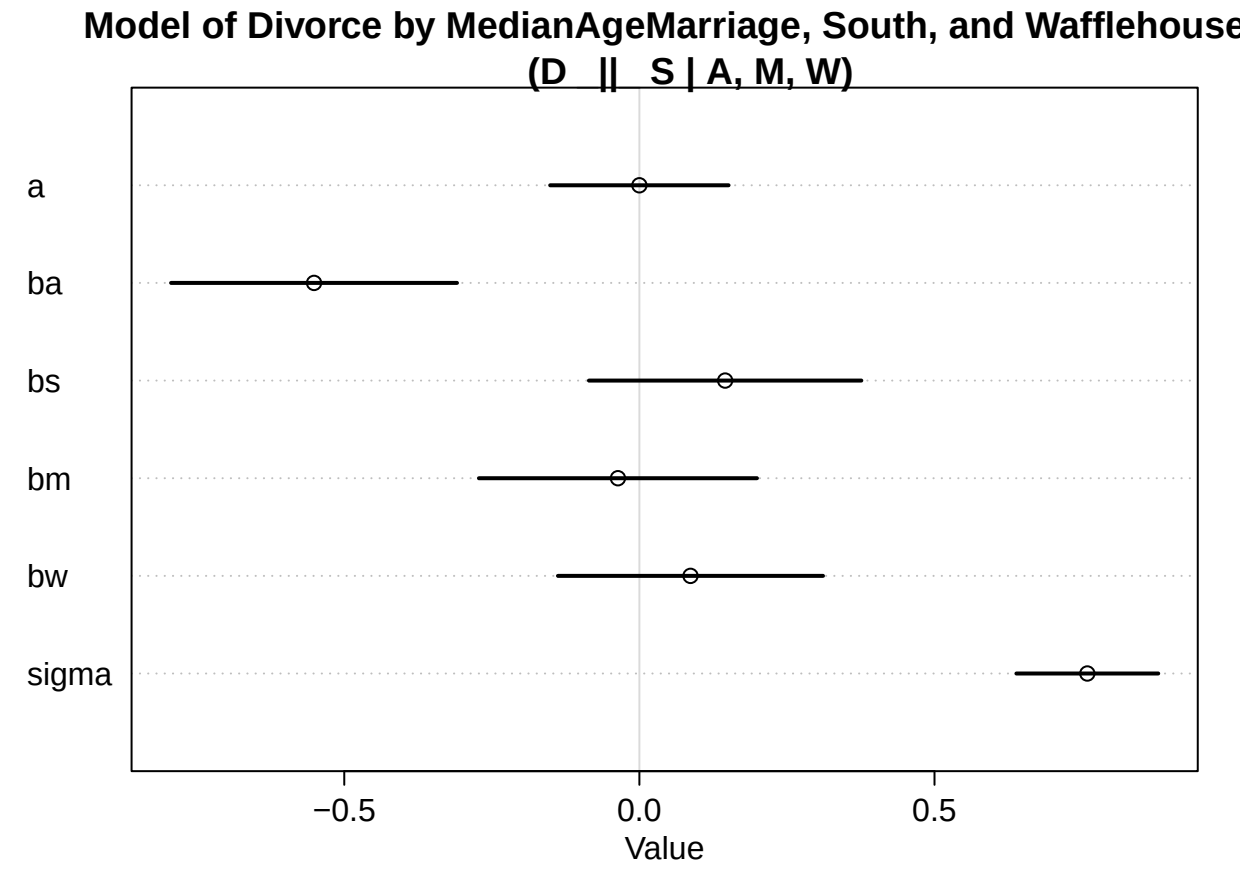
```
# Get implied conditional independencies for the DAG
impliedConditionalIndependencies(dag)
```

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

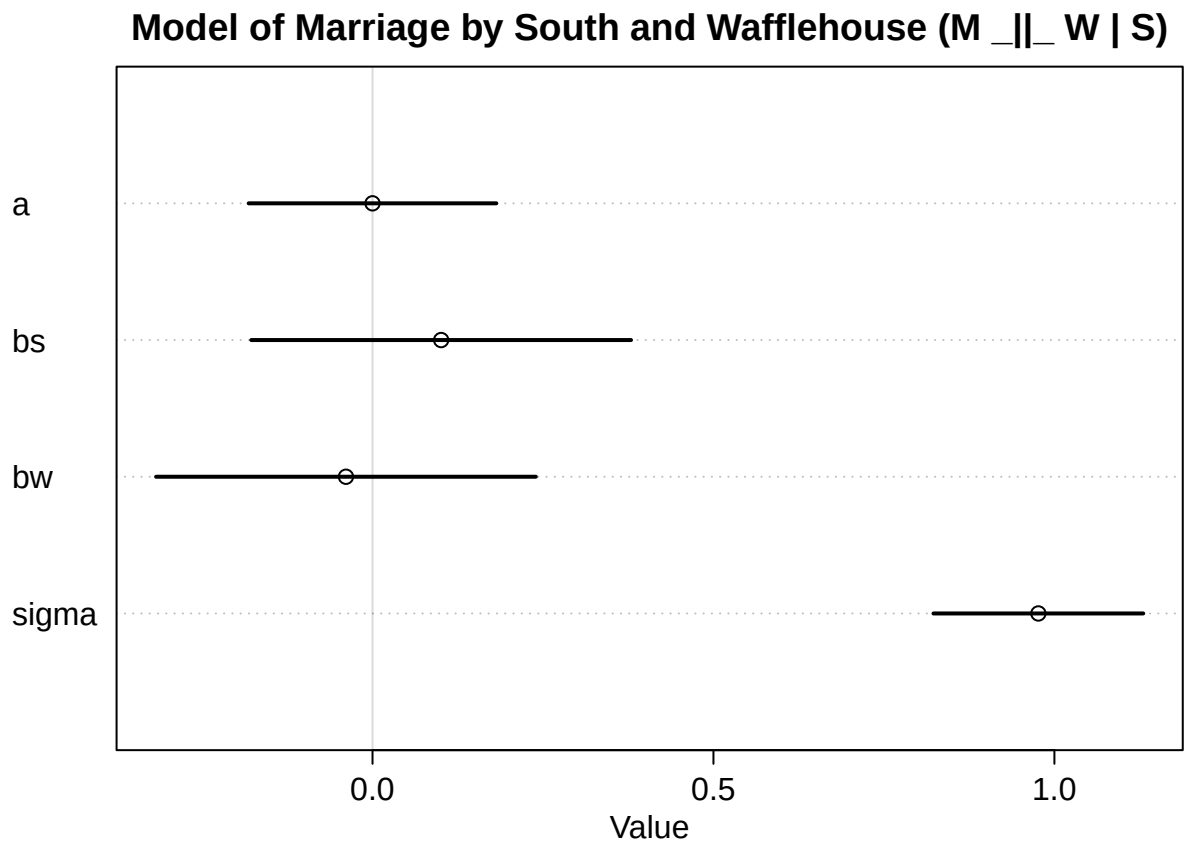
```
# Testing first implied conditional independency (A _||_ W | S)
model1 <- quap(
  alist(
    MedianAgeMarriage ~ dnorm(mu, sigma),
    mu <- a + bs * South + bw * WaffleHouses,
    a ~ dnorm(0, 0.2),
    bs ~ dnorm(0, 0.5),
    bw ~ dnorm(0, 0.5),
    sigma ~ dexp(1)
  ),
  data=WaffleDivorce
)

plot(precis(model1),
     main = "Model of MedianAgeMarriage by South and Wafflehouse (A _||_ W | S)")
```




```
# Testing second implied conditional independency (D _||_ S | A, M, W)
model2 <- quap(
  alist(
    Divorce ~ dnorm(mu, sigma),
    mu <- a + ba * MedianAgeMarriage + bs * South + bm * Marriage + bw * WaffleHouses,
    a ~ dnorm(0, 0.2),
    ba ~ dnorm(0, 0.5),
    bs ~ dnorm(0, 0.5),
    bm ~ dnorm(0, 0.5),
    bw ~ dnorm(0, 0.5),
    sigma ~ dexp(1)
  ),
  data=WaffleDivorce
)

plot(precis(model2),
main = "Model of Divorce by MedianAgeMarriage, South, and Wafflehouse
     (D _||_ S | A, M, W)")
```




```r
# Testing third implied conditional independency (M _||_ W | S)
model3 <- quap(
  alist(
    Marriage ~ dnorm(mu, sigma),
    mu <- a + bs * South + bw * WaffleHouses,
    a ~ dnorm(0, 0.2),
    bs ~ dnorm(0, 0.5),
    bw ~ dnorm(0, 0.5),
    sigma ~ dexp(1)
  ),
  data=WaffleDivorce
)

plot(precis(model3),
     main = "Model of Marriage by South and Wafflehouse (M _||_ W | S)")
```




```
# See if bw (estimate for WaffleHouses) overlaps with 0
round(precis(model1)["bw", ], 2)
```

```
##    mean   sd  5.5% 94.5%
## bw 0.06 0.17 -0.21  0.34
```

```
# See if bs (estimate for South) overlaps with 0
round(precis(model2)["bs", ], 2)
```

```
##    mean   sd  5.5% 94.5%
## bs 0.15 0.14 -0.09  0.38
```

```
# See if bw (estimate for WaffleHouses) overlaps with 0
round(precis(model3)["bw", ], 2)
```

```
##     mean   sd  5.5% 94.5%
## bw -0.04 0.17 -0.32  0.24
```

```
# compare(model1, model2, model3)
#if we wanted to compare the models, we could look at dWAIC scores too.
```

Answer 6H2: All three model estimates have credible intervals that overlap with 0, meaning that the implied conditional independencies are correct. The DAG does not need to be amended, as it is already a good representation of the data.