Assignment 2 - Methods 4

Mikkel

2025-04-06

Second assignment

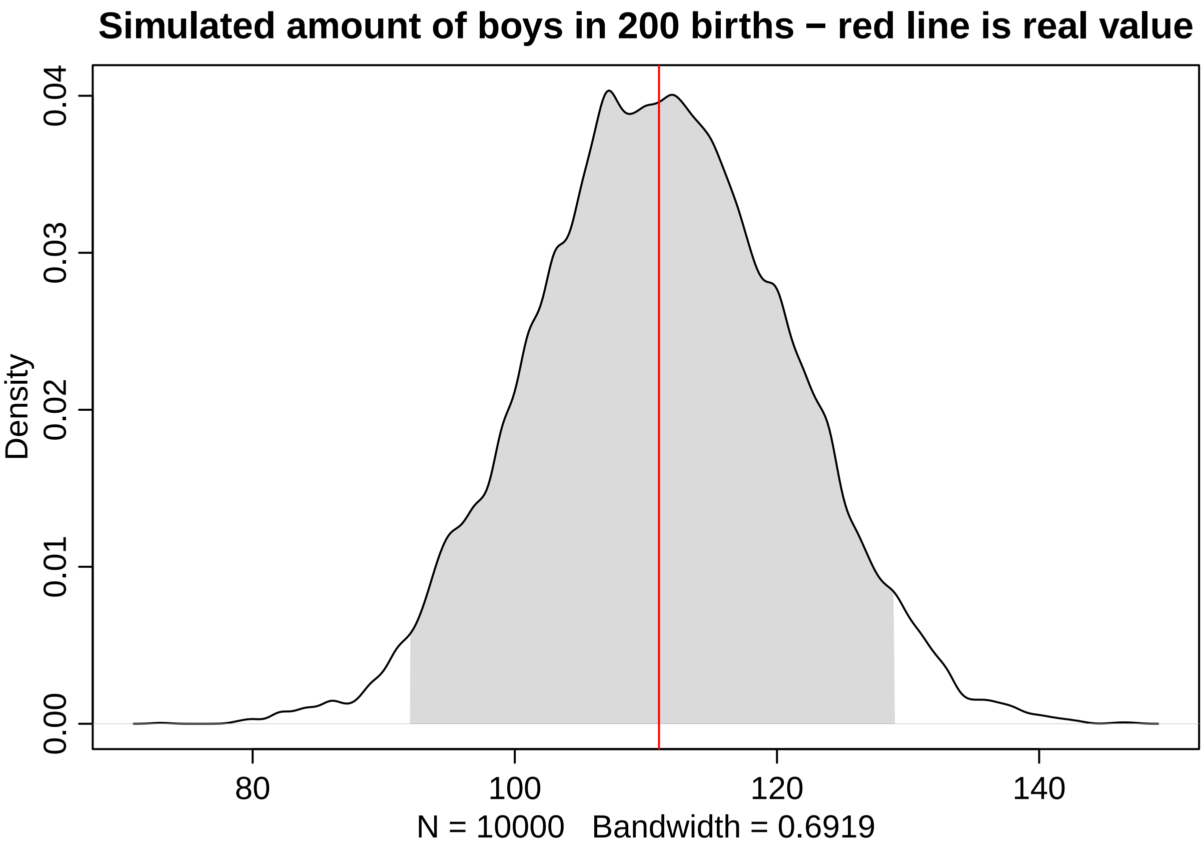
The second assignment uses chapter 3, 5 and 6. The focus of the assignment is getting an understanding of causality.

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).

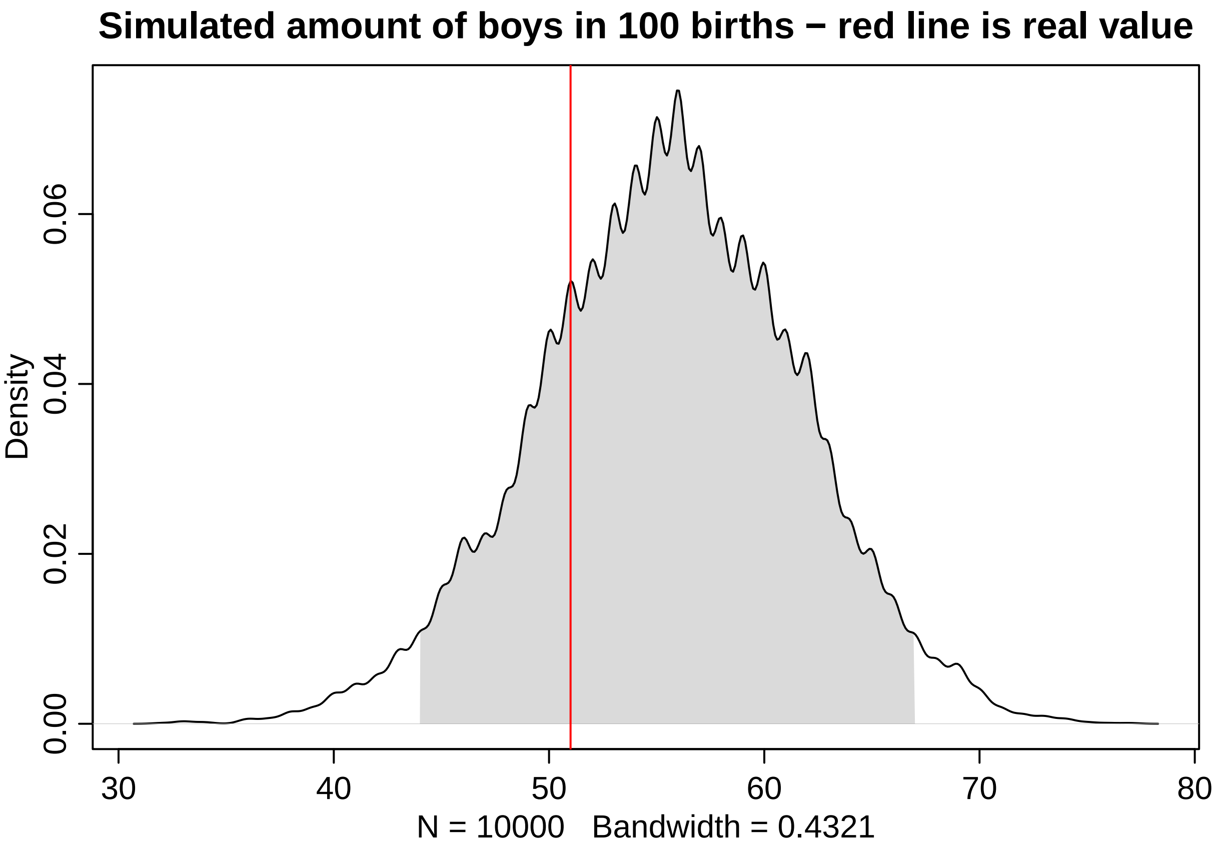
```
# 3H1
# Find the posterior probability of giving birth to a boy:
pacman::p_load(rethinking)
data(homeworkch3)
set.seed(1)
W <- sum(birth1) + sum(birth2)
N <- length(birth1) + length(birth2)
p_grid <-seq(from =0, to = 1, len =1000)
prob_p <- rep(1,1000)
prob_data <- dbinom(W,N,prob=p_grid)
posterior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



The distribution of predicted number of boys match closely with the actual count.

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?

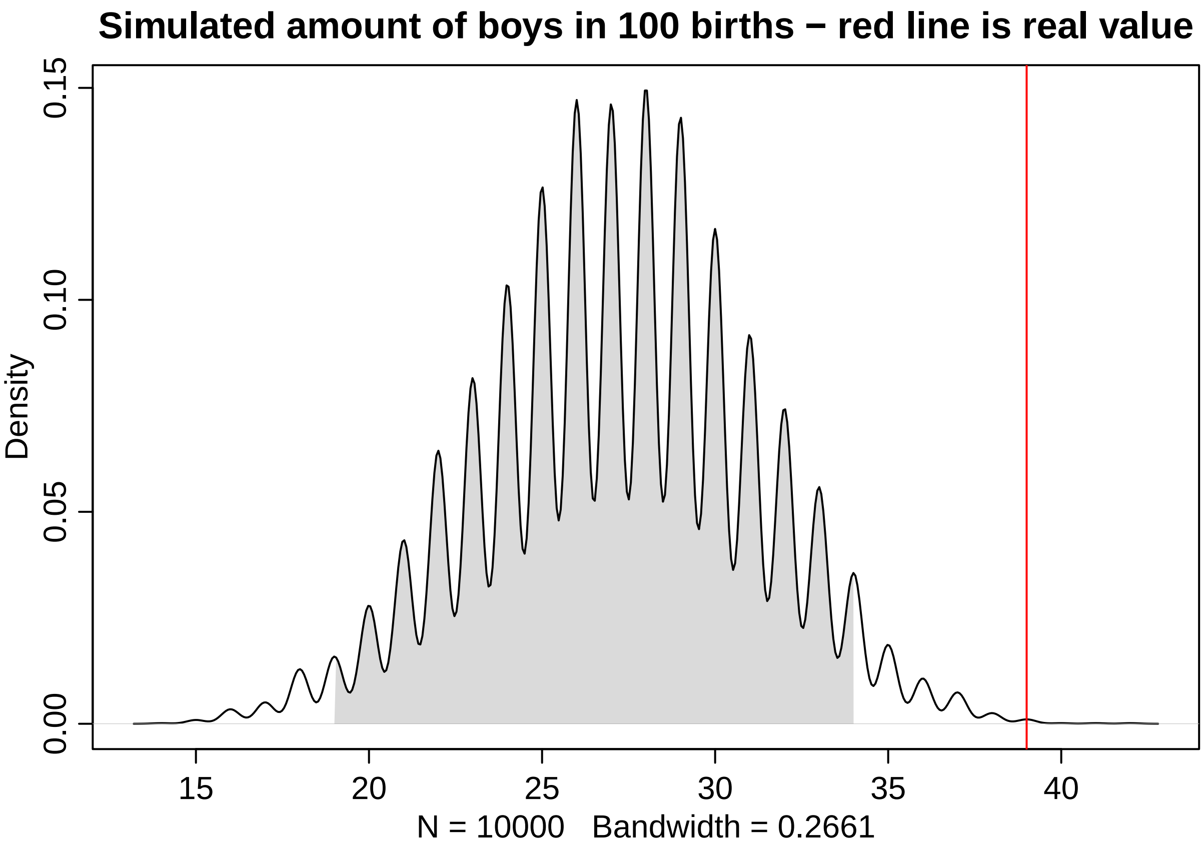
```
# simulate first births
simulated_births <- rbinom(n = 1e4, size = length(birth1), prob = samples)

# plot probability density distribution
dens(simulated_births,show.HPDI = 0.95)
abline(v=sum(birth1), col="red")
title("Simulated amount of boys in 100 births - red line is real value")
```



Looking at only the firstborns the model prediction is not as accurate as before but the actual count is still within high density of the simulated data.

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 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?

```r
# create new data frame of births following female first births
postFemaleBirths = birth2[birth1 == 0]

# simulate births following female first births
simulated_births <- rbinom(n = 1e4, size = length(postFemaleBirths), prob = samples)

# plot probability density distribution
dens(simulated_births,show.HPDI = 0.95)
abline(v=sum(postFemaleBirths), col="red")
title("Simulated amount of boys in 100 births - red line is real value")
```



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 wonderfull 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. 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).

Outcome variable: Healthy BMI Predictor variables: Vegetarianism, Health Consciousness Scale score

Healthy BMI correlates with vegetarianism, but the correlation between vegetarianism and healthy BMI diminishes when Health Consciousness Scale score is entered into the model as vegetarianism is a case of healthy user bias - people who just considers what they eat are healthier than regular people, because they are generally more conscious about their health and make lots of other health beneficial choices such as more regular exercise besides for example becoming vegetarian.

5M2. 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.

Outcome: Miles run per year Predictor variables: Miles run per week, injury rate

Miles run in a week is postively correlated with miles run per year, but running more miles per week could be positively correlated with injury rate, with increased injury rate decreasing the miles run per year.

5M3. 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

More people getting divorced might result in more marriages as the divorced remarry. This hypothesis can be tested by creating a new variable called remarriage rate and using multiple regression to investigate whether divorce or marriage rate predicts remarriage the most.

5M5. 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.

Obesity rate = a + b_price_gasoline*price_gasoline + b_exercise_per_week*exercise_per_week + b_eating_out_per_week*eating_out_per_week

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. 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 data
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
```

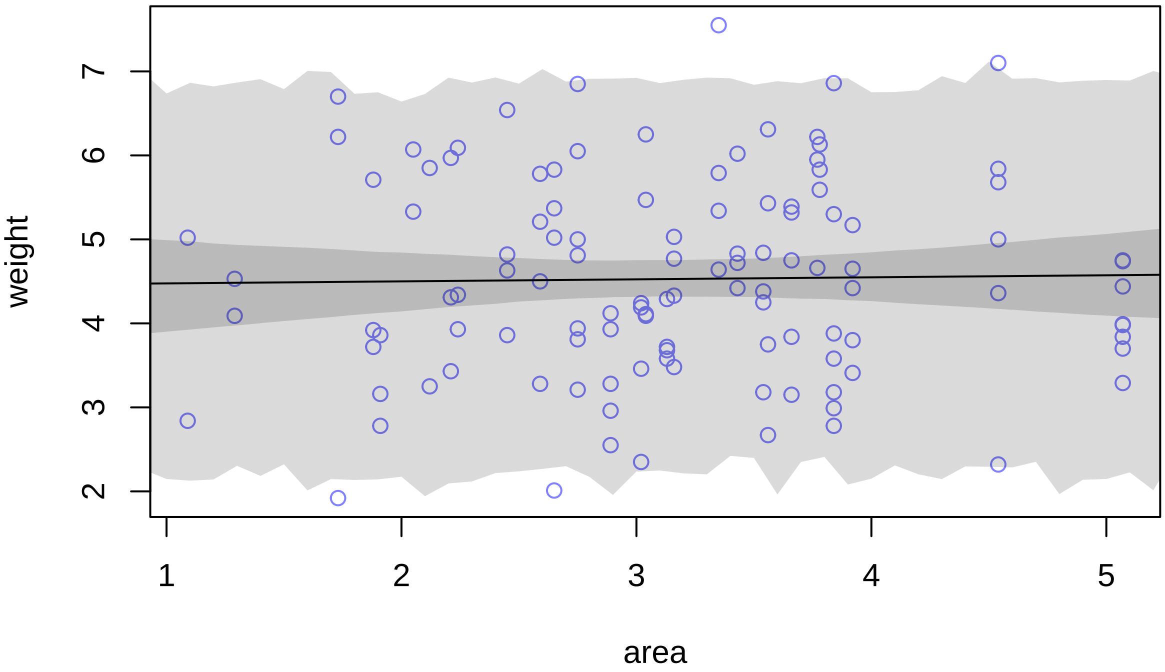
```
# Fit model 1
model1 <- quap(
  alist(
    weight ~ dnorm( mean=mu , sd=sigma ) ,
    mu <- a + b*area ,
    a ~ dnorm(5,5) ,
    b ~ dnorm(0,10) ,
    sigma ~ dunif(0,50)
  ) ,
  data=foxes)

precis(model1)
```

```
##              mean         sd       5.5%     94.5%
## a     4.45426368 0.38963250  3.8315557 5.0769717
## b     0.02387202 0.11805556 -0.1648036 0.2125476
## sigma 1.17868501 0.07738429  1.0550100 1.3023600
```

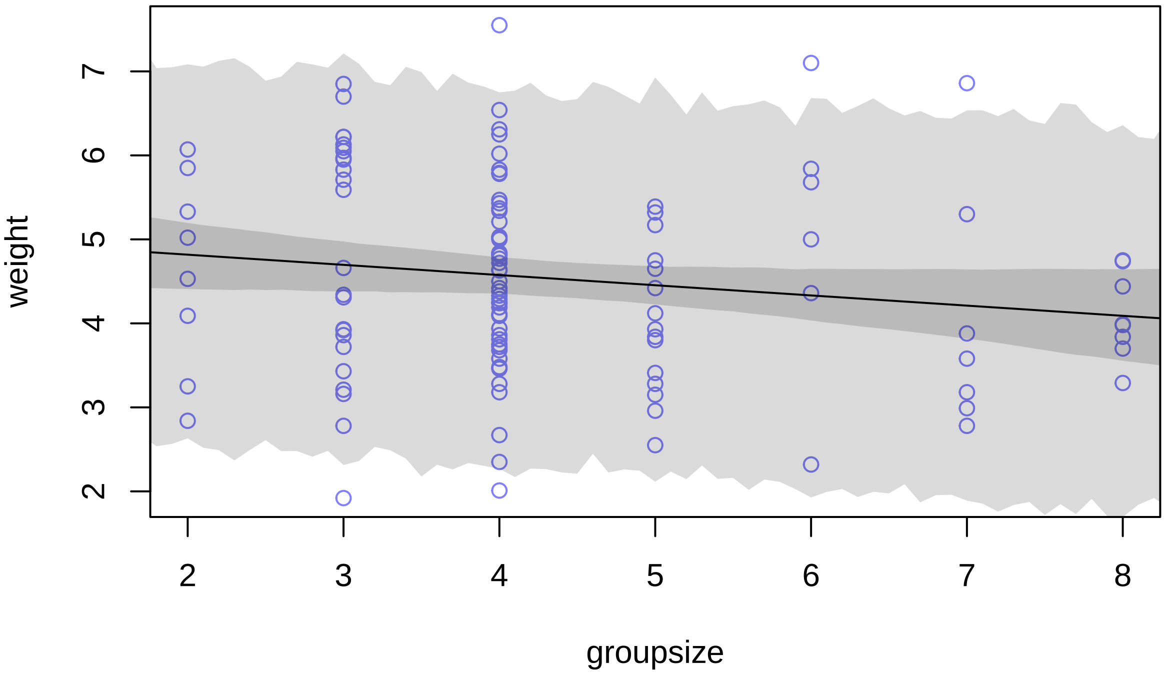
```r
# Fit model 2
model2 <- quap(
  alist(
    weight ~ dnorm( mean=mu , sd=sigma ) ,
    mu <- a + b*groupsize ,
    a ~ dnorm(5,5) ,
    b ~ dnorm(0,10) ,
    sigma ~ dunif(0,50)
  ) ,
  data=foxes)

precis(model2)
```

```
##             mean         sd       5.5%       94.5%
## a      5.0676655 0.32420582  4.5495220  5.58580906
## b     -0.1238372 0.07038918 -0.2363327 -0.01134169
## sigma  1.1635360 0.07639026  1.0414496  1.28562242
```

```r
# Plotting model 1
plot( weight ~ area , data=foxes ,
      col=col.alpha(rangi2,1) )

# Define area sequence
area_seq <- seq(from=0.1,to=10,length.out=100)

# Compute mean and 95% interval of the mean
mu <- link(model1, data=data.frame(area=area_seq))
mu_mean <- apply(mu,2,mean)
mu_PI <- apply(mu,2,PI,prob=0.95)

# Simulated predicted weights
sim_weight <- sim(model1, data=list(area=area_seq))

# Calculate 95% interval for predicted weights
weight_PI <- apply(sim_weight, 2, PI, prob=0.95)


# Plotting
lines(area_seq, mu_mean)      # MAP regression line
shade(mu_PI, area_seq)        # 95% interval of the mean
shade(weight_PI, area_seq)    # 95% interval
```



```
# Plotting model 2
plot( weight ~ groupsize , data=foxes ,
      col=col.alpha(rangi2,1) )

# Define groupsize sequence
groupsize_seq <- seq(from=0.1,to=10,length.out=100)

# Compute mean and 95% interval of the mean
mu <- link(model2, data=data.frame(groupsize=groupsize_seq))
mu_mean <- apply(mu,2,mean)
mu_PI <- apply(mu,2,PI,prob=0.95)

# Simulated predicted weights
sim_weight <- sim(model2, data=list(groupsize=groupsize_seq))

# Calculate 95% interval for predicted weights
weight_PI <- apply(sim_weight, 2, PI, prob=0.95)


# Plotting
lines(groupsize_seq, mu_mean)       # MAP regression line
shade(mu_PI, groupsize_seq)         # 95% interval of the mean
shade(weight_PI, groupsize_seq)     # 95% interval
```



Area does not appear to be a meaningful predictor of weight, while group size might be slightly negatively correlated with weight.

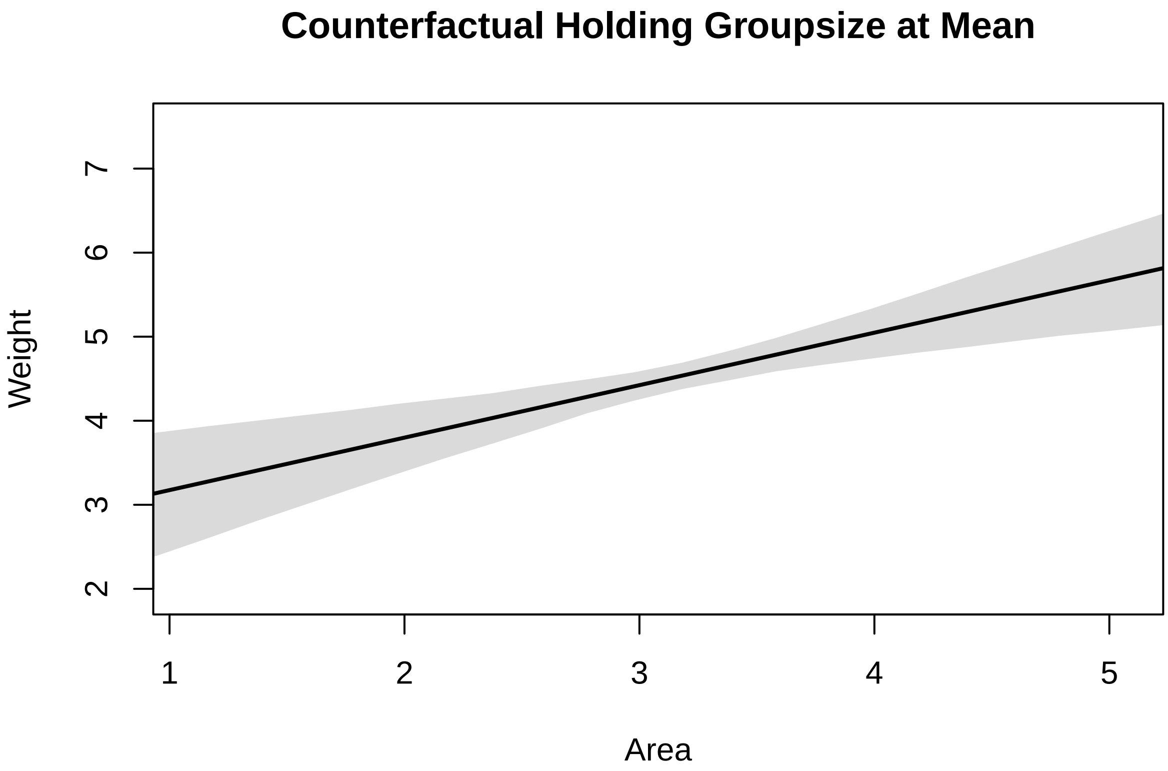
5H2. Now fit a multiple linear regression with weight as the outcome and both area and groupsize as predictor variables. Plot the predictions of the model for each predictor, holding the other predictor constant at its mean. What does this model say about the importance of each variable? Why do you get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

```
# Fit model 3
model3 <- quap(
  alist(
    weight ~ dnorm( mean=mu , sd=sigma ) ,
    mu <- a + b_area*area + b_groupsize*groupsize ,
    a ~ dnorm(5,5) ,
    b_area ~ dnorm(5,5) ,
    b_groupsize ~ dnorm(5,5) ,
    sigma ~ dexp(1)
  ) ,
  data=foxes)

precis(model3)
```

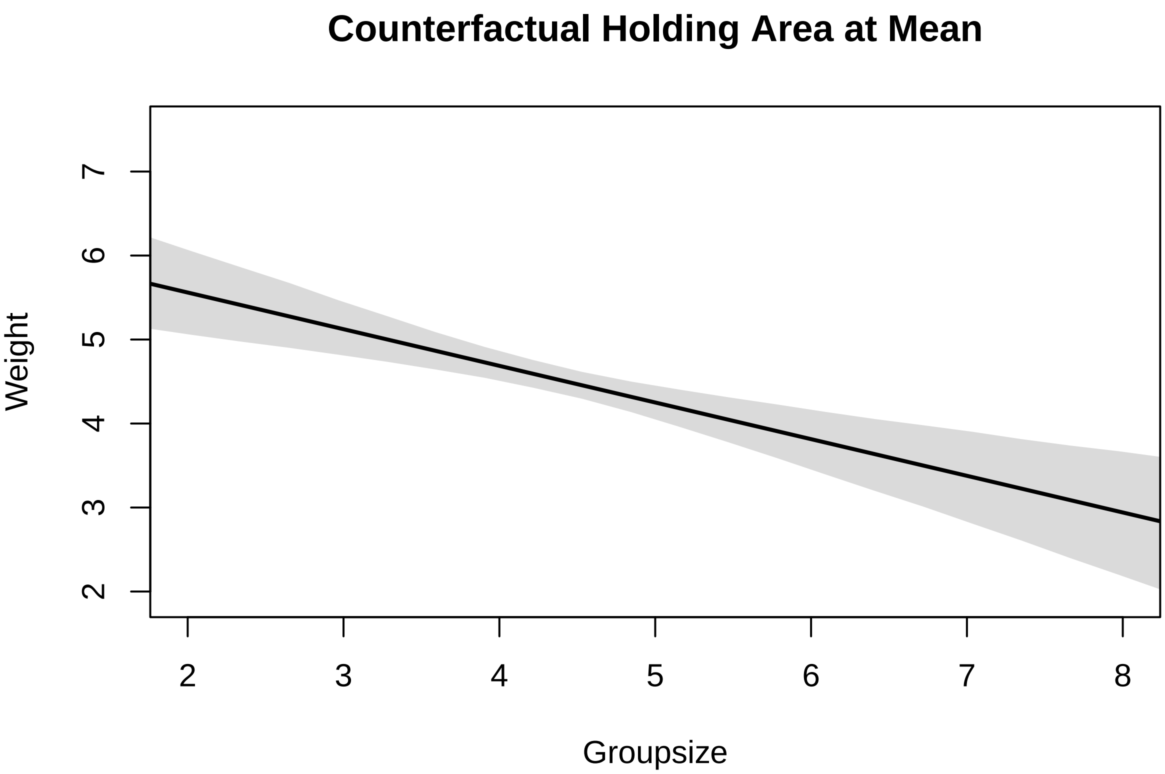
```
##                   mean         sd       5.5%      94.5%
## a            4.4462527 0.36802080  3.8580844  5.0344210
## b_area       0.6199218 0.19883539  0.3021445  0.9376992
## b_groupsize -0.4329236 0.12009298 -0.6248554 -0.2409919
## sigma        1.1131245 0.07255985  0.9971599  1.2290892
```

```
# Plotting counterfactual holding groupsize constant at mean
area_seq <- seq( from=min(foxes$area) - sd(foxes$area),
                 to=max(foxes$area) + sd(foxes$area),
                 length.out=30 )

mu <- link(model3, data = data.frame(area=area_seq,
                                     groupsize = mean(foxes$groupsize)))
```

```
mu_mean <- apply(mu,2,mean)
mu_PI <- apply(mu,2,PI)
plot(NULL, xlim=range(foxes$area), ylim=range(foxes$weight),
     xlab="Area", ylab="Weight", main="Counterfactual Holding Groupsize at Mean")
lines(area_seq, mu_mean, lwd=2)
shade(mu_PI, area_seq)
```



```
# Plotting counterfactual holding area constant at mean
groupsize_seq <- seq( from=min(foxes$groupsize) - sd(foxes$groupsize) ,
                  to=max(foxes$groupsize) + sd(foxes$groupsize) ,
                  length.out=30)

mu <- link(model3, data = data.frame(groupsize=groupsize_seq,
                                     area = mean(foxes$area)))
mu_mean <- apply(mu,2,mean)
mu_PI <- apply(mu,2,PI)
plot(NULL, xlim=range(foxes$groupsize), ylim=range(foxes$weight),
     xlab="Groupsize", ylab="Weight",
     main="Counterfactual Holding Area at Mean")
lines(groupsize_seq, mu_mean, lwd=2)
shade(mu_PI, groupsize_seq)
```



Area is positively correlated with weight, but groupsize is negatively correlated with weight. This is the case of a masked relationship as groupsize and area washes out the the effect of each other when combined.

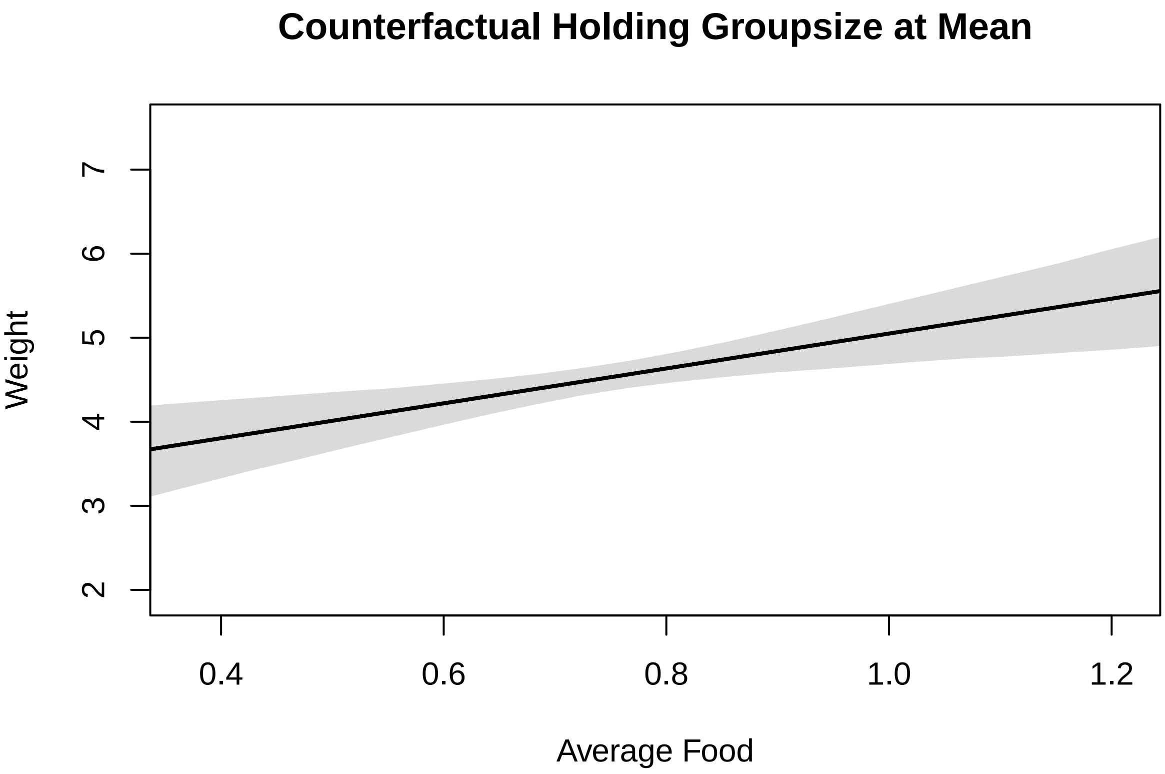
5H3. Finally, consider the avgfood variable. Fit two more multiple regressions: (1) body weight as an additive function of avgfood and groupsize, and (2) body weight as an additive function of all three variables, avgfood and groupsize and area. Compare the results of these models to the previous models 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)

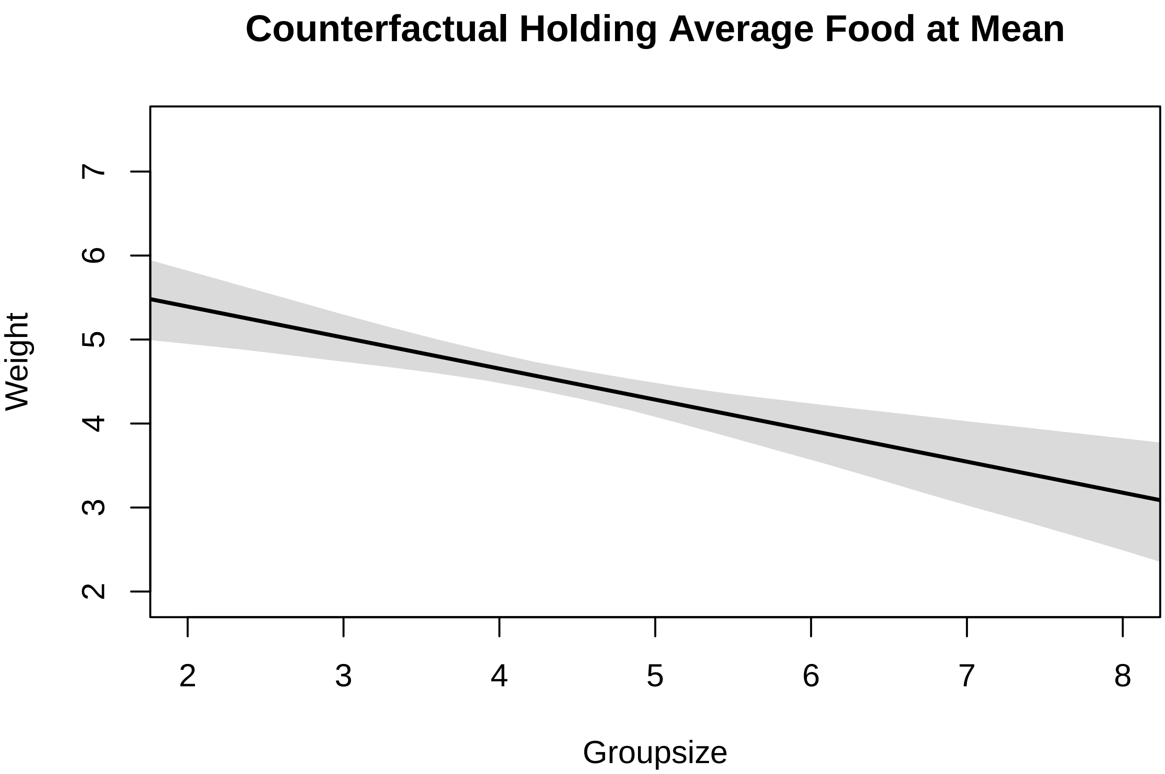
```
model4 <- quap(
  alist(
    weight ~ dnorm(mu, sigma),
    mu <- a + b_avgfood*avgfood + b_groupsize*groupsize,
    a ~ dnorm(5,5),
    b_avgfood ~ dnorm(1,1),
    b_groupsize ~ dnorm(5,5),
    sigma ~ dexp(1)
  ), data = foxes)

precis(model4)
```

```
##                   mean         sd       5.5%     94.5%
## a            4.5417905 0.36615531  3.9566036  5.126977
## b_avgfood    2.1214599 0.77730666  0.8791737  3.363746
## b_groupsize -0.3697936 0.11305743 -0.5504812 -0.189106
## sigma        1.1206812 0.07354396  1.0031437  1.238219
```

```r
# Plotting counterfactual holding groupsize constant at mean
avgfood_seq <- seq( from=min(foxes$avgfood) - sd(foxes$avgfood) ,
                    to=max(foxes$avgfood) + sd(foxes$avgfood) ,
                    length.out=30)

mu <- link(model4, data = data.frame(avgfood = avgfood_seq,
                                     groupsize = mean(foxes$groupsize)))
mu_mean <- apply(mu,2,mean)
mu_PI <- apply(mu,2,PI)
plot(NULL, xlim=range(foxes$avgfood), ylim=range(foxes$weight),
     xlab="Average Food", ylab="Weight",
     main="Counterfactual Holding Groupsize at Mean")
lines(avgfood_seq, mu_mean, lwd=2)
shade(mu_PI, avgfood_seq)
```



```r
# Plotting counterfactual holding avgfood constant at mean
groupsize_seq <- seq( from=min(foxes$groupsize) - sd(foxes$groupsize) ,
                    to=max(foxes$groupsize) + sd(foxes$groupsize) ,
                    length.out=30)

mu <- link(model4, data = data.frame(groupsize = groupsize_seq,
                                     avgfood = mean(foxes$avgfood)))
mu_mean <- apply(mu,2,mean)
mu_PI <- apply(mu,2,PI)
plot(NULL, xlim=range(foxes$groupsize), ylim=range(foxes$weight),
     xlab="Groupsize", ylab="Weight",
     main="Counterfactual Holding Average Food at Mean")
lines(groupsize_seq, mu_mean, lwd=2)
shade(mu_PI, groupsize_seq)
```



(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.

The effect of average food seems to have a larger effect on the weight than area, which makes sense causally as food consummation is more directly connected to weight than area. The predictive value of the two variables can also by comparing the fit of model 3 and 4, because these are identical besides the two predictor variables we are interested in assessing. The fit can be compared by calculating the log-likelihood of each model:

```
-logLik(model3)
```

```
## 'log Lik.' 177.5851 (df=4)
```

```
-logLik(model4)
```

```
## 'log Lik.' 178.3738 (df=4)
```

This comparison suggests that average food is the best predictor of the two in consideration.

(2)

```
model5 <- quap(
  alist(
    weight ~ dnorm(mu, sigma),
    mu <- a + b_avgfood*avgfood + b_area*area +b_groupsize*groupsize ,
    a ~ dnorm(5,5),
    b_avgfood ~ dnorm(1,1),
    b_area ~ dnorm(5,5),
    b_groupsize ~ dnorm(5,5),
    sigma ~ dexp(1)
  ), data = foxes)

precis(model5)
```

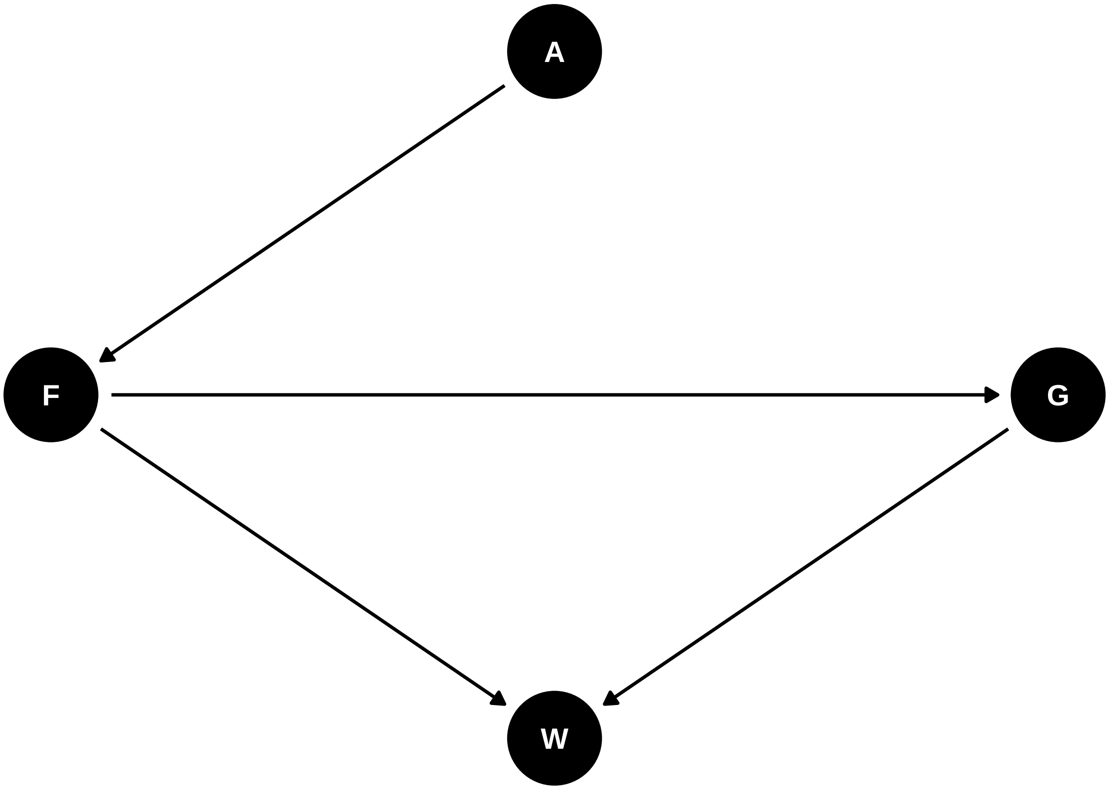
```
##                     mean          sd        5.5%       94.5%
## a             4.2221685 0.38551489  3.6060412  4.8382957
## b_avgfood     1.4611288 0.82387110  0.1444237  2.7778339
## b_area        0.4839418 0.21120040  0.1464027  0.8214808
## b_groupsize -0.5349448 0.13204796 -0.7459829 -0.3239067
## sigma         1.1016293 0.07192128  0.9866852  1.2165734
```

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?

The effect of each predictor is reduced as a result of multicollinearity - food and area is correlated because larger areas have more food. This makes each coefficient harder to separate and estimate therefore smaller and more uncertain leading to larger standard errors.

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

```r
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?

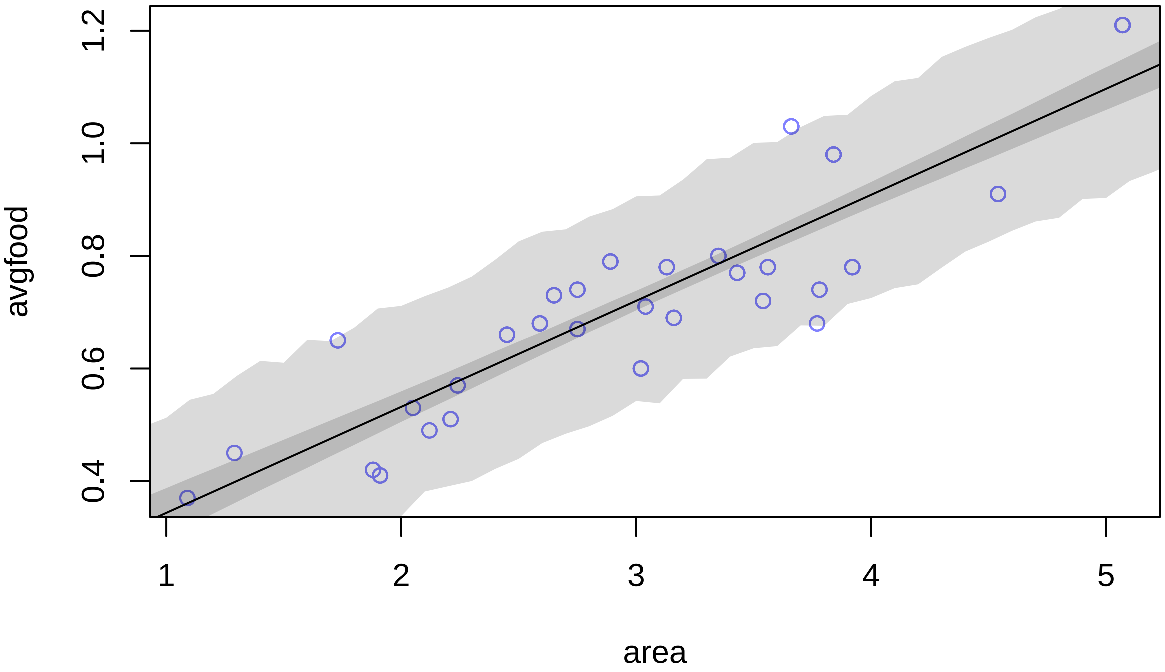
```
model_total_A_on_F <- quap(
  alist(
    avgfood ~ dnorm(mu, sigma),
    mu <- a + b_area*area,
    a ~ dnorm(1.5,2),
    b_area ~ dnorm(5,5),
    sigma ~ dexp(1)
  ), data = foxes
)

precis(model_total_A_on_F)
```

```
##              mean          sd       5.5%     94.5%
## a      0.15413477 0.030707077 0.10505893 0.2032106
## b_area 0.18857308 0.009301908 0.17370684 0.2034393
## sigma  0.09261519 0.006074518 0.08290694 0.1023234
```

Increasing the area results in a increase in the amount of food in the area.

```
# Plotting
plot( avgfood ~ area , data=foxes ,
      col=col.alpha(rangi2,1) )

# Define area sequence
area_seq <- seq(from=0.1,to=10,length.out=100)
```

```
# Compute mean and 95% interval of the mean
mu <- link(model_total_A_on_F, data=data.frame(area=area_seq))
mu_mean <- apply(mu,2,mean)
mu_PI <- apply(mu,2,PI,prob=0.95)

# Simulated predicted weights
sim_avgfood <- sim(model_total_A_on_F, data=list(area=area_seq))

# Calculate 95% interval for predicted weights
avgfood_PI <- apply(sim_avgfood, 2, PI, prob=0.95)


# Plotting
lines(area_seq, mu_mean)      # MAP regression line
shade(mu_PI, area_seq)        # 95% interval of the mean
shade(avgfood_PI, area_seq)    # 95% interval
```



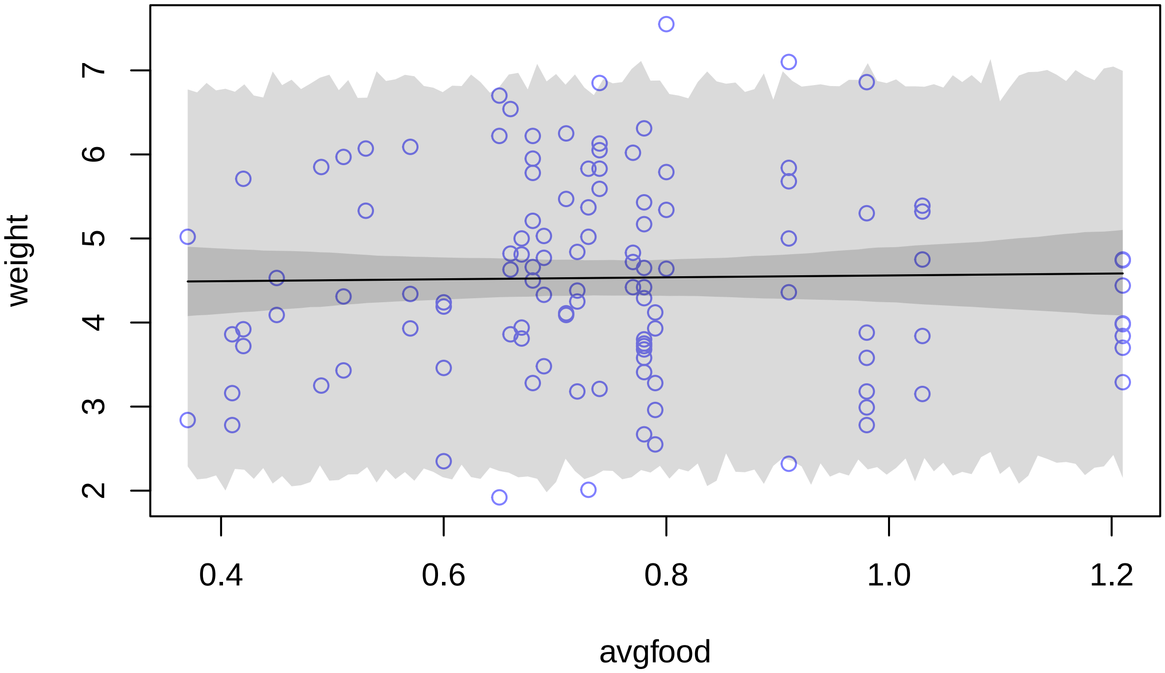
2) Infer the **total** causal effect of adding food F to a territory on the weight W of foxes. Can you calculate the causal effect by simulating an intervention on food?

```
# Create model
model_avgfood_on_weight <- quap(
  alist(
    weight ~ dnorm(mu, sigma),
    mu <- a + b_avgfood*avgfood,
    a ~ dnorm(5,5),
    b_avgfood ~ dnorm(1,1),
    sigma ~ dexp(1)
  ), data=foxes
)

# Check previews
```

```r
precis(model_avgfood_on_weight)
```

```
##                  mean         sd      5.5%     94.5%
## a           4.4434759 0.37866580  3.838295 5.0486570
## b_avgfood 0.1149944 0.48265846 -0.656387 0.8863759
## sigma     1.1737716 0.07659738  1.051354 1.2961890
```

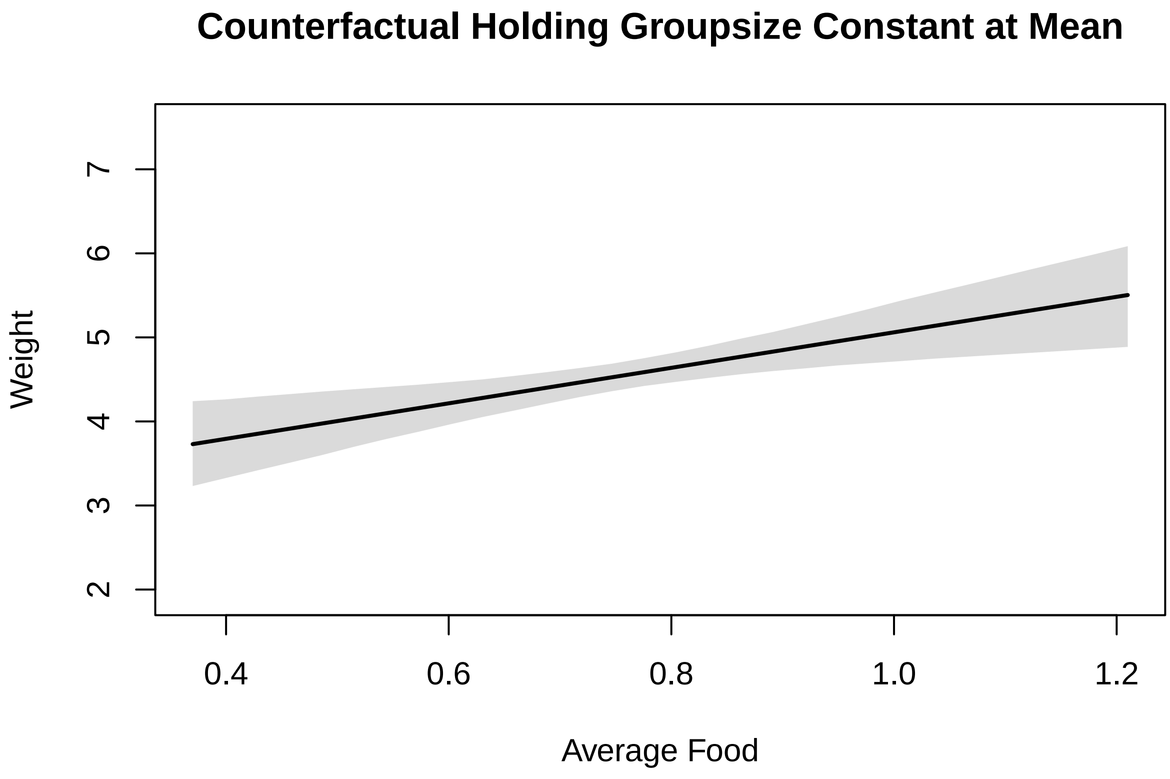
```r
# Plotting
plot( weight ~ avgfood , data=foxes ,
      col=col.alpha(rangi2,1) )

# Define average food sequence
avgfood_seq <- seq(from=min(foxes$avgfood),to=max(foxes$avgfood),
                   length.out=100)

# Compute mean and 95% interval of the mean
mu <- link(model_avgfood_on_weight, data=data.frame(avgfood=avgfood_seq))
mu_mean <- apply(mu,2,mean)
mu_PI <- apply(mu,2,PI,prob=0.95)

# Simulated predicted weights
sim_weight <- sim(model_avgfood_on_weight, data=list(avgfood=avgfood_seq))

# Calculate 95% interval for predicted weights
weight_PI <- apply(sim_weight, 2, PI, prob=0.95)


# Plotting
lines(avgfood_seq, mu_mean)      # MAP regression line
shade(mu_PI, avgfood_seq)        # 95% interval of the mean
shade(weight_PI, avgfood_seq)    # 95% interval
```



The total effect is close to negligible.

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?

```
# Creating model
model_avgfood_and_groupsize_on_weight <- quap(
  alist(
    weight ~ dnorm(mu, sigma),
    mu <- a + b_avgfood*avgfood + b_groupsize*groupsize,
    a ~ dnorm(5,5),
    b_avgfood ~ dnorm(1,1),
    b_groupsize ~ dnorm(5,5),
    sigma ~ dexp(1)
  ), data=foxes
)

# Plotting counterfactual groupsize constant at mean
avgfood_seq <- seq( from=min(foxes$avgfood),
                 to=max(foxes$avgfood),
                 length.out=30 )

mu <- link(model_avgfood_and_groupsize_on_weight,
           data = data.frame(avgfood=avgfood_seq,
                                        groupsize = mean(foxes$groupsize)))
mu_mean <- apply(mu,2,mean)
mu_PI <- apply(mu,2,PI)
plot(NULL, xlim=range(foxes$avgfood), ylim=range(foxes$weight),
     xlab="Average Food", ylab="Weight",
     main="Counterfactual Holding Groupsize Constant at Mean")
lines(avgfood_seq, mu_mean, lwd=2)
shade(mu_PI, avgfood_seq)
```



Increasing food increases groupsize, which in turn leads to less food per fox, but if groupsize is hold constant increasing food leads to bigger foxes.

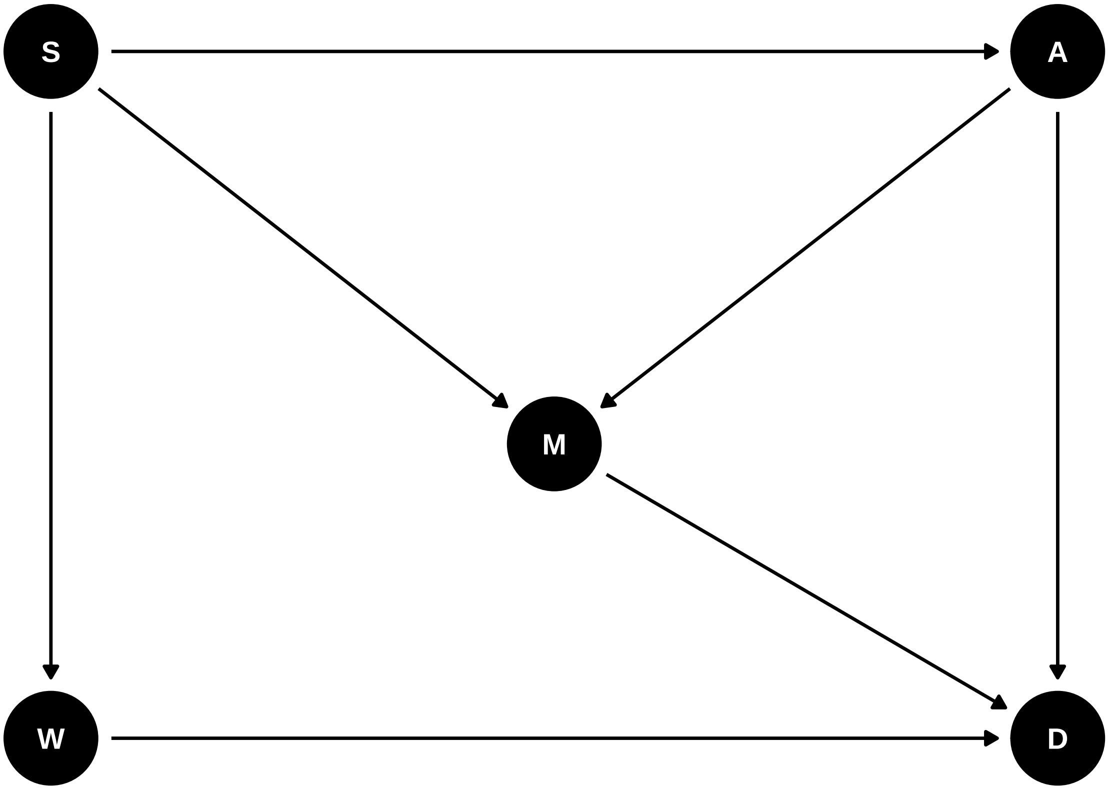
Chapter 6: Investigating the Waffles and Divorces

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.

```r
# loading and previewing data
data(WaffleDivorce)
head(WaffleDivorce)
```

```
##     Location Loc Population MedianAgeMarriage Marriage Marriage.SE Divorce
## 1    Alabama  AL       4.78              25.3     20.2        1.27    12.7
## 2     Alaska  AK       0.71              25.2     26.0        2.93    12.5
## 3    Arizona  AZ       6.33              25.8     20.3        0.98    10.8
## 4   Arkansas  AR       2.92              24.3     26.4        1.70    13.5
## 5 California  CA      37.25              26.8     19.1        0.39     8.0
## 6   Colorado  CO       5.03              25.7     23.5        1.24    11.6
##   Divorce.SE WaffleHouses South Slaves1860 Population1860 PropSlaves1860
## 1       0.79          128     1     435080         964201           0.45
## 2       2.05            0     0          0              0           0.00
## 3       0.74           18     0          0              0           0.00
## 4       1.22           41     1     111115         435450           0.26
## 5       0.24            0     0          0         379994           0.00
## 6       0.94           11     0          0          34277           0.00
```

```r
# Draw causal graph
dag <- dagitty('dag {
bb="0,0,1,1"
A [pos="0.600,0.700"]
D [pos="0.600,0.350"]
M [pos="0.450,0.500"]
S [pos="0.300,0.700"]
W [pos="0.300,0.350"]
A -> D
A -> M
M -> D
S -> A
S -> M
S -> W
W -> D
}
')

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



```r
# Scale data
WaffleDivorce$WaffleHouses <- scale( WaffleDivorce$WaffleHouses )
WaffleDivorce$Divorce <- scale( WaffleDivorce$Divorce )
WaffleDivorce$MedianAgeMarriage <- scale( WaffleDivorce$MedianAgeMarriage )
WaffleDivorce$Marriage <- scale( WaffleDivorce$Marriage )

# Create categorical variable for south or not
WaffleDivorce$SouthIndicator <- ifelse( WaffleDivorce$South==1 , 2 , 1 )

# Create and fit model
model_total_W_on_D <- quap(
  alist(
    Divorce ~ dnorm(mu, sigma),
    mu <- a + b_WaffleHouses*WaffleHouses,
    a ~ dnorm(0,1),
    b_WaffleHouses ~ dnorm(0,1),
    sigma ~ dexp(1)
  ), data=WaffleDivorce,
)

# See result
precis(model_total_W_on_D, depth=2)
```

```
##                        mean         sd        5.5%     94.5%
## a              1.810343e-07 0.13293612 -0.21245741 0.2124578
## b_WaffleHouses 2.498881e-01 0.13426449  0.03530752 0.4644687
## sigma          9.484179e-01 0.09352228  0.79895121 1.0978845
```

The total effect of a increasing WaffleHouses by 1 standard deviation is a 0.25 standard deviation increase in Divorce rate.

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 int he data.

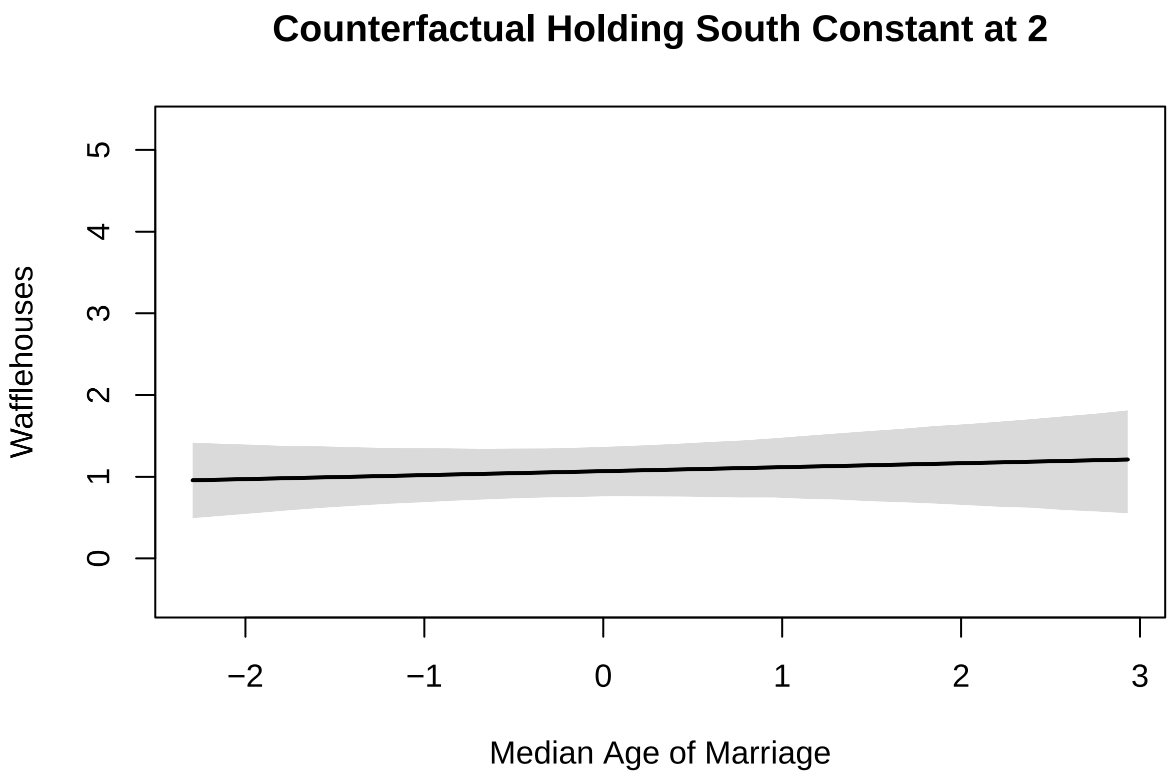
```
# Finding the implied conditinoal independencies:
impliedConditionalIndependencies(dag)
```

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

```
## Testing the implied conditional independencies

# Testing: A _||_ W | S
# Create and fit model
A_on_W <- quap(
  alist(
    WaffleHouses ~ dnorm(mu, sigma),
    mu <- a[SouthIndicator] + b_MedianAgeMarriage*MedianAgeMarriage,
    a[SouthIndicator] ~ dnorm(0,1),
    b_MedianAgeMarriage ~ dnorm(0,1),
    sigma ~ dexp(1)
  ), data=WaffleDivorce,
)
precis(A_on_W, depth = 2)
```

```
##                           mean         sd       5.5%      94.5%
## a[1]                -0.4251449 0.11913459 -0.6155450 -0.2347448
## a[2]                 1.0697971 0.19162945  0.7635362  1.3760580
## b_MedianAgeMarriage  0.0494185 0.10453957 -0.1176559  0.2164929
## sigma                0.7134507 0.07065581  0.6005290  0.8263723
```

```
# Plotting counterfactual South Constant at 2 (South)
MedianAgeMarriage_seq <- seq( from=min(WaffleDivorce$MedianAgeMarriage),
                 to=max(WaffleDivorce$MedianAgeMarriage),
                 length.out=30 )

mu <- link(A_on_W,
           data = data.frame(MedianAgeMarriage=MedianAgeMarriage_seq,
                             SouthIndicator=2))
mu_mean <- apply(mu,2,mean)
mu_PI <- apply(mu,2,PI)
plot(NULL, xlim=range(WaffleDivorce$MedianAgeMarriage),
     ylim=range(WaffleDivorce$WaffleHouses),
     xlab="Median Age of Marriage", ylab="Wafflehouses",
     main="Counterfactual Holding South Constant at 2")
lines(MedianAgeMarriage_seq, mu_mean, lwd=2)
shade(mu_PI, MedianAgeMarriage_seq)
```



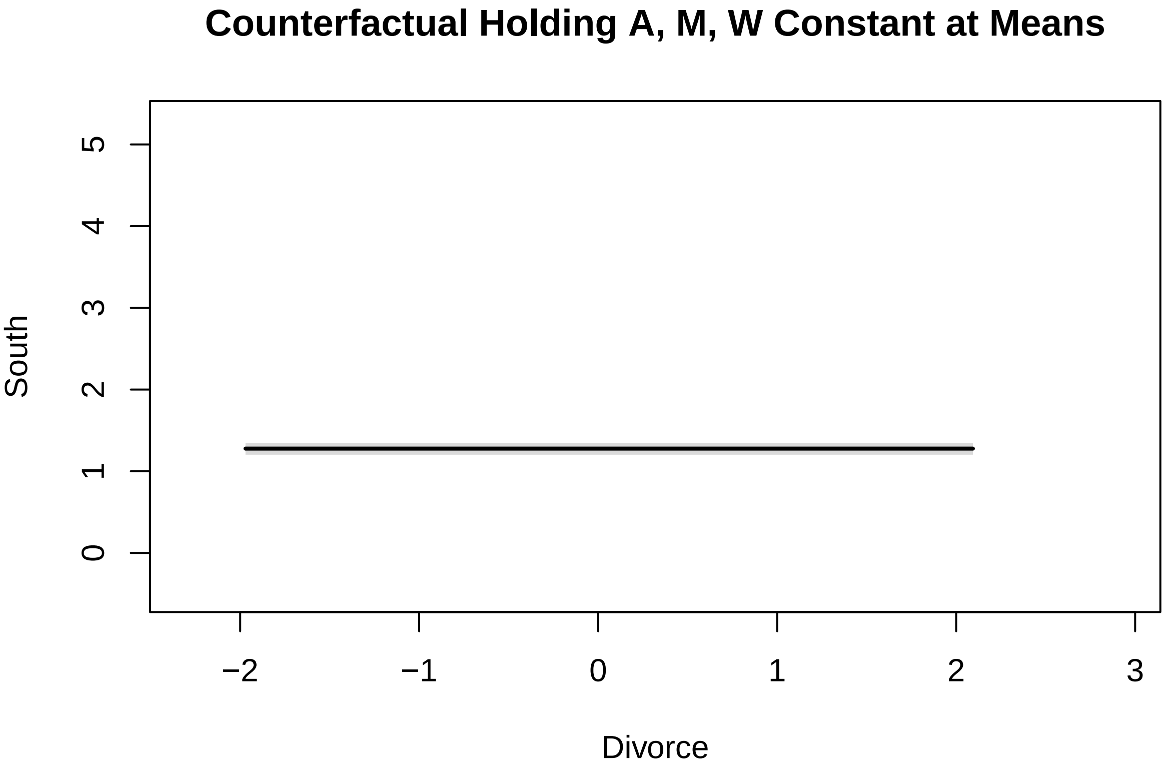
There seem to be no real effect suggesting that this implied conditional independency holds.

```
# Testing: D _||_ S | A, M, W
# Create and fit model
D_on_S <- quap(
  alist(
    SouthIndicator ~ dnorm(mu, sigma),
    mu <- a +
      b_MedianAgeMarriage*MedianAgeMarriage +
      b_Marriage*Marriage +
      b_WaffleHouses*WaffleHouses,
    a ~ dnorm(0,1),
    b_MedianAgeMarriage ~ dnorm(0,1),
    b_Marriage ~ dnorm(0,1),
    b_WaffleHouses ~ dnorm(0,1),
    sigma ~ dexp(1)
  ), data=WaffleDivorce,
)
precis(D_on_S, depth = 2)
```

```
##                           mean         sd       5.5%       94.5%
## a                    1.2774720 0.04444188  1.2064453  1.34849868
## b_MedianAgeMarriage -0.1166744 0.06531159 -0.2210549 -0.01229385
## b_Marriage          -0.0549967 0.06487889 -0.1586857  0.04869230
## b_WaffleHouses       0.2983030 0.04536543  0.2258002  0.37080567
## sigma                0.3145422 0.03130803  0.2645059  0.36457845
```

```
# Plotting counterfactual holding A, M, W constant at means
Divorce_seq <- seq( from=min(WaffleDivorce$Divorce),
                 to=max(WaffleDivorce$Divorce),
                 length.out=30 )
```

```
mu <- link(D_on_S,
          data = data.frame(Divorce=Divorce_seq,
                            MedianAgeMarriage = mean(WaffleDivorce$MedianAgeMarriage),
                            Marriage = mean(WaffleDivorce$Marriage),
                            WaffleHouses = mean(WaffleDivorce$WaffleHouses)))
mu_mean <- apply(mu,2,mean)
mu_PI <- apply(mu,2,PI)
plot(NULL, xlim=range(WaffleDivorce$MedianAgeMarriage),
     ylim=range(WaffleDivorce$WaffleHouses),
     xlab="Divorce", ylab="South",
     main="Counterfactual Holding A, M, W Constant at Means")
lines(Divorce_seq, mu_mean, lwd=2)
shade(mu_PI, Divorce_seq)
```

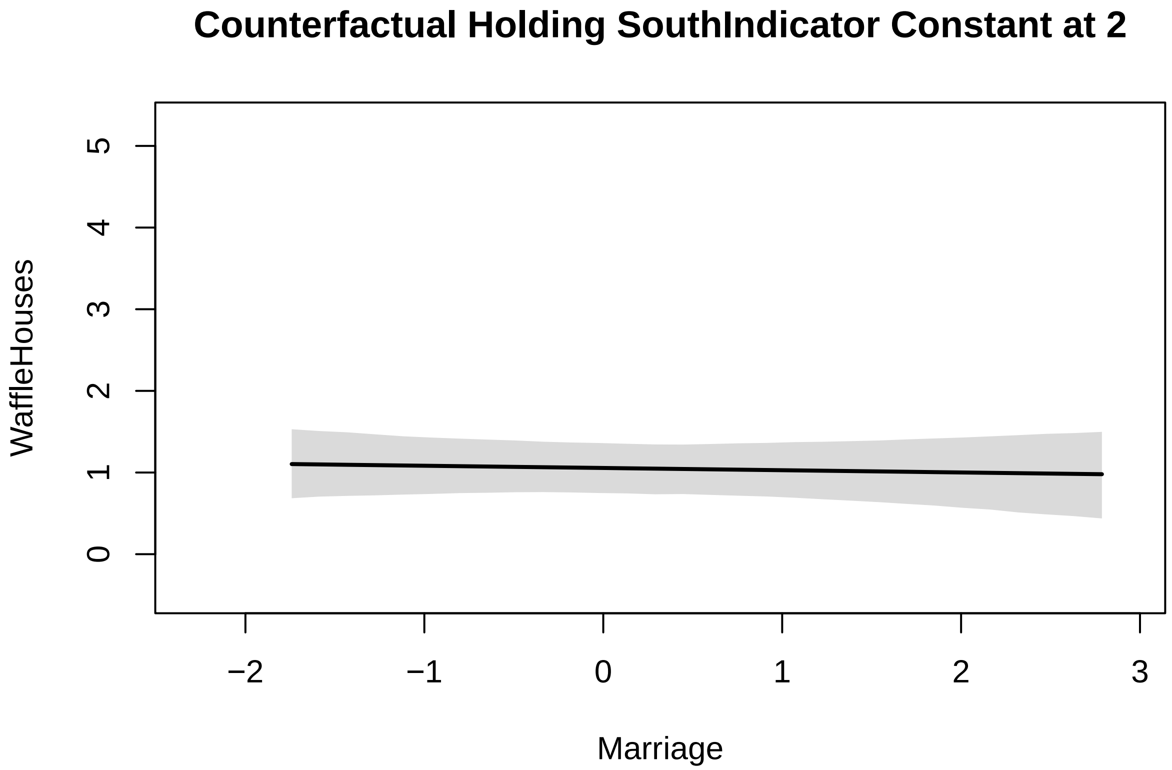


There seem to be no effect suggesting that this implied conditional independency holds.

```
# Testing: M _||_ W | S
# Create and fit model
M_on_W <- quap(
  alist(
    WaffleHouses ~ dnorm(mu, sigma),
    mu <- a[SouthIndicator] + b_Marriage*Marriage,
    a[SouthIndicator] ~ dnorm(0,1),
    b_Marriage ~ dnorm(0,1),
    sigma ~ dexp(1)
  ), data=WaffleDivorce,
)
precis(M_on_W, depth = 2)
```

```
##                     mean          sd         5.5%        94.5%
## a[1]       -0.41901844 0.11840968 -0.6082600 -0.2297769
## a[2]        1.05429318 0.18822142  0.7534790  1.3551074
## b_Marriage -0.02725168 0.10191322 -0.1901287  0.1356253
## sigma       0.71478907 0.07078216  0.6016655  0.8279126
```

```
# Plotting counterfactual holding S constant at 2
Marriage_seq <- seq( from=min(WaffleDivorce$Marriage),
                 to=max(WaffleDivorce$Marriage),
                 length.out=30 )

mu <- link(M_on_W,
           data = data.frame(Marriage=Marriage_seq,
                             SouthIndicator = 2))

mu_mean <- apply(mu,2,mean)
mu_PI <- apply(mu,2,PI)
plot(NULL, xlim=range(WaffleDivorce$MedianAgeMarriage),
     ylim=range(WaffleDivorce$WaffleHouses),
     xlab="Marriage", ylab="WaffleHouses",
     main="Counterfactual Holding SouthIndicator Constant at 2")
lines(Marriage_seq, mu_mean, lwd=2)
shade(mu_PI, Marriage_seq)
```



Again there seem to be no effect suggesting that the last implied conditional independency holds meaning that the causal model is plausible.