Assignment 2 - Methods 4

2025-04-05

Authors: Aneta Simunkova, Emma-Margaret Vaserik, Julianne Silke Swärd, Leona Carina Chi, Marton Ferenc Peterdi

Second assignment

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

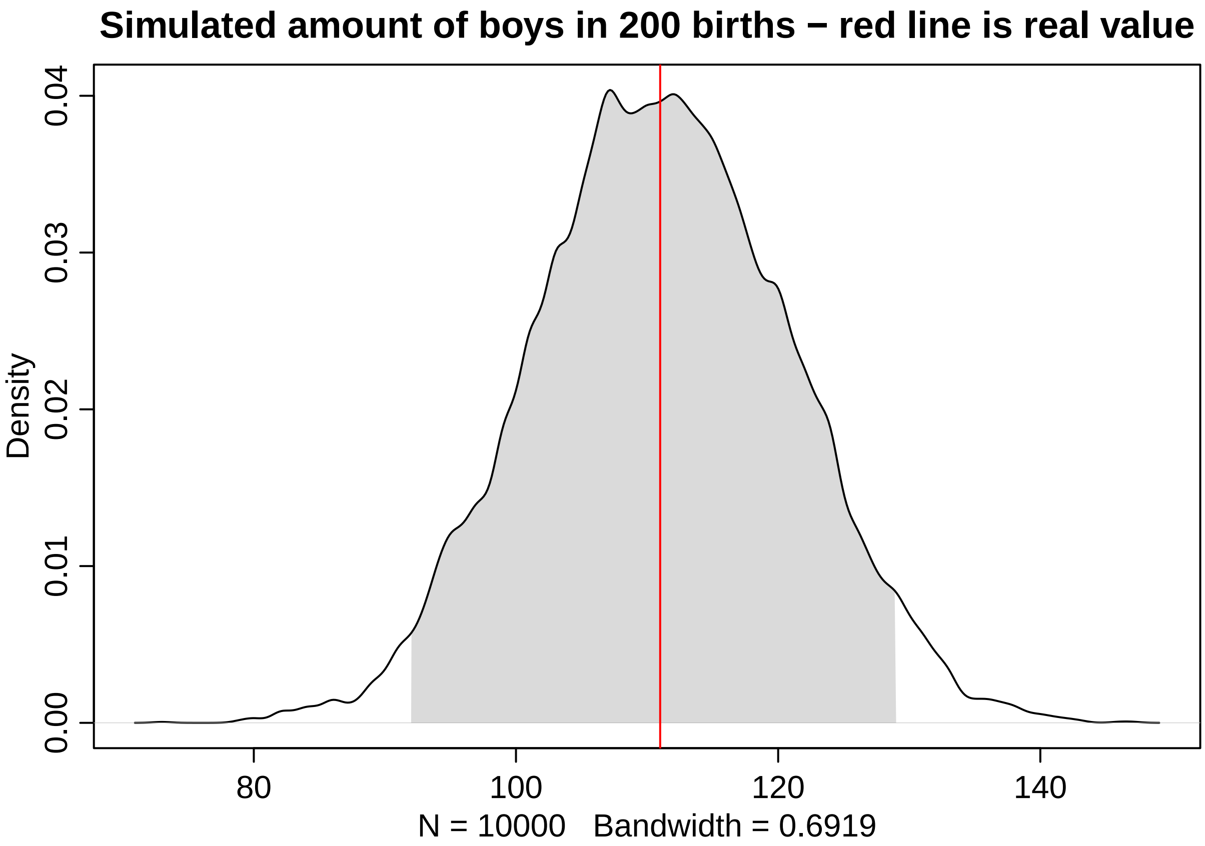
For Chris/Laurits: The different sections are annotated with names of the people responsible for them.

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:)

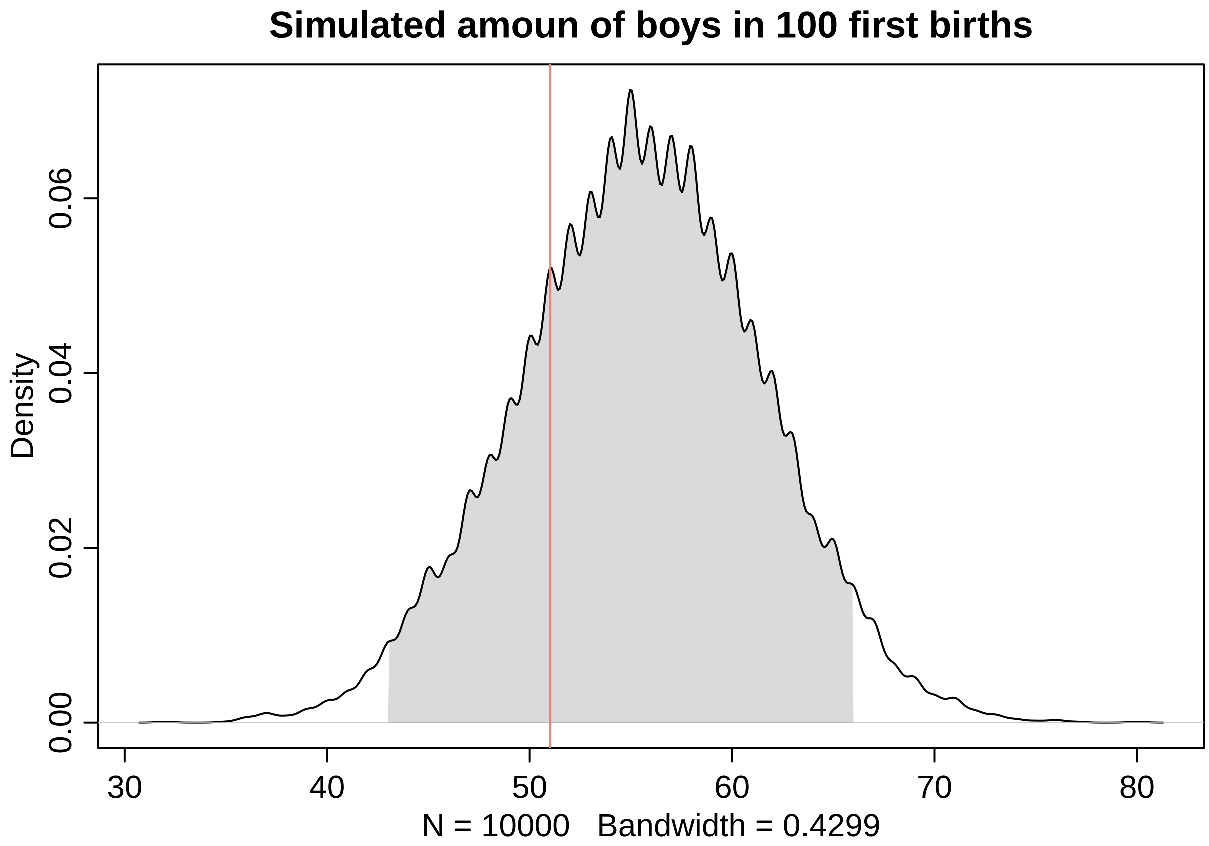
3H3 (Marton) Use rbinom to simulate 10,000 replicates of 200 births. You should end up with 10,000 numbers, each one a count of boys out of 200 births. Compare the distribution of predicted numbers of boys to the actual count in the data (111 boys out of 200 births).

```r
# 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
# Check if they align 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 (Marton) 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?

```
set.seed(1)
n_firstborn <- length(birth1)

sim_firstborn <- rbinom(n = 10000, size = n_firstborn, prob = samples)

rethinking::dens(sim_firstborn, show.HPDI = 0.95)
abline(v = sum(birth1), col = "salmon")
title("Simulated amoun of boys in 100 first births")
```

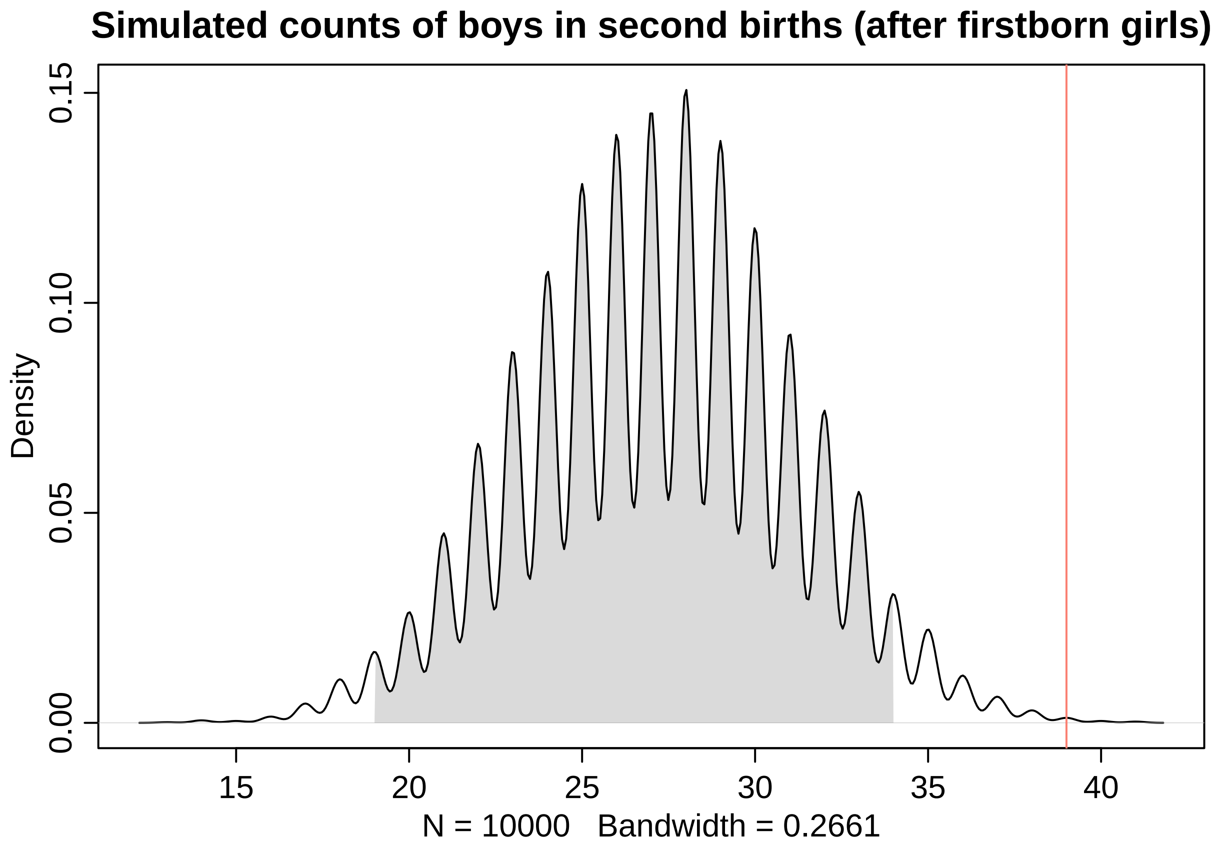


Answer:

The red vertical line represents the actual number of boys in the first births. As one can see, the actual number of boys in first births falls within the HPDI but is left-skewed, which suggests that the model does not capture some aspects of the data but is still broadly reasonable. In comparison to our first model it seems that the second model is not as good of a fit but still consistent with the data.

3H5 (Marton) 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)
n_f_firstborn <- sum(birth1 == 0)

sim_second_births <- rbinom(n = 10000, size = n_f_firstborn, prob = samples)

m_after_f <- sum(birth2[birth1 ==0])

rethinking::dens(sim_second_births, show.HPDI = 0.95)
abline(v = m_after_f, col = "salmon")
title("Simulated counts of boys in second births (after firstborn girls)")
```



Answer:

As can be seen in the plot, the actual number of boy births after a firstborn girl is extremely skewed to the right. This suggests that our model does not cepture the data very well. Additionally, this suggests that the sex of the second child depends on the sex of the first child. Possibly this can be explained by biological or genetic effects influencing the probability of second births being boys after a firstborn girl.

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. (Anet) 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: In an example of a spurious correlation, the outcome variable *number of visitors in a park* is first predicted separately by *weather* and then by *number of earthworms in the park.* They both show to be negatively correlated with *number of visitors in a park.* However, when they are both included as predictors in a bivariate regression, the effect of *number of earthworms in the park* largely disappears, given that it is also driven by *weather.* and thus the correlation of *number of earthworms in the park* and *number of visitors in a park* was spurious. In other words, although there seems to be an association the an increasing number of earthworms, fewer people visit a park, this relationship is spurious since earthworms primarily appear when it rains, and rain is here the main predictor of park visitors.

5M2. (Anet) 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: In an example of a masked relationship, the outcome variable *number of tasks finished at work* shows a weak relationship with *hours of sleep the previous night* and *stress level* when using the predictors separately. However, when both predictors are used in the same model, the relationship of each of them with the outcome variable appears much stronger. The reason for that is possibly that *hours of sleep the previous night* and *stress level* are negatively correlated with each other, and we can only see the full effect of each of them when we get to control for the other one. To describe this more concretely, since we assume that we are more stressed during the day when we don't get enough sleep which in turn affects our work productivity, it is important to include both of these predictors in the model in order to be able to control for their negative correlation and reveal the full effect of each of them.

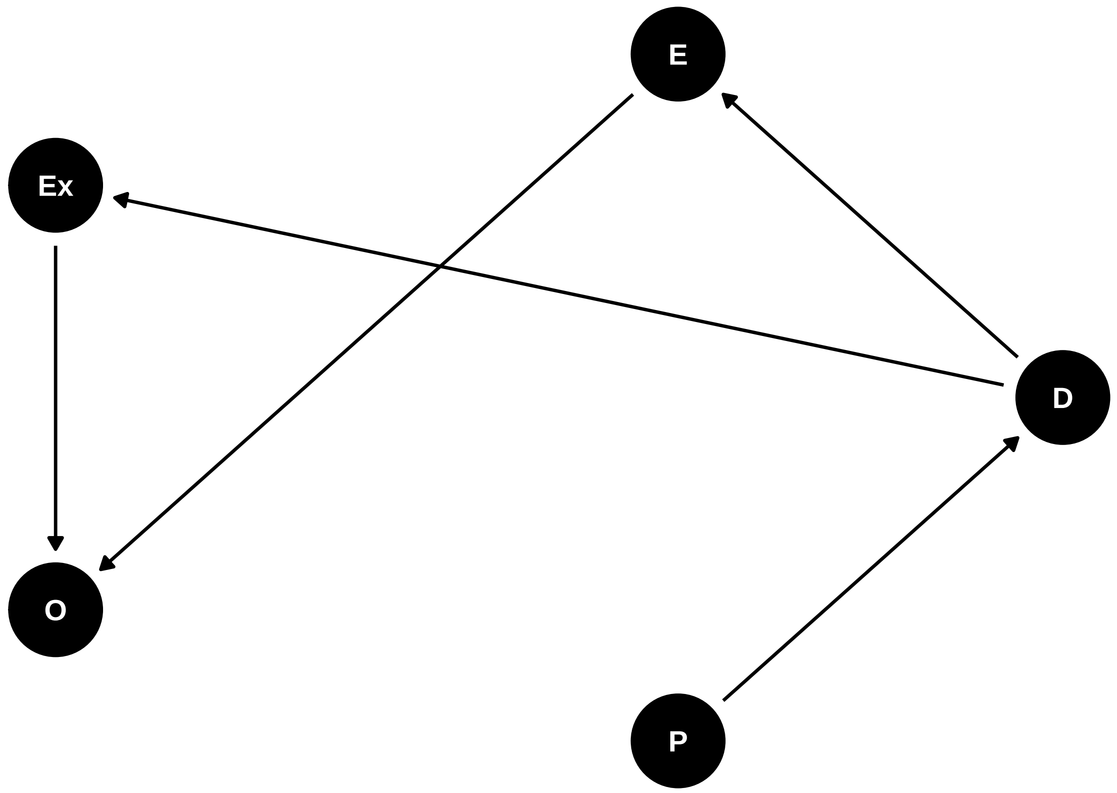
5M3. (Anet) It is sometimes observed that the best predictor of fire risk is the presence of firefighters—States and localities with many firefighters also have more fires. Presumably firefighters do not cause fires. Nevertheless, this is not a spurious correlation. Instead fires cause firefighters. 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: One way to think about it could be that a divorce sets up the ground for remarrying, i.e. ‘’every time’ ’ a person gets divorced, they might - and many of them do - marry again. A way to test this could be to model a multiple regression where *marriage rate* is predicted by both *divorce rate* as well as *remarrying rate*. The size of the beta coefficients would provide more information about whether higher marriage rates are driven by the will to remarry.

5M5. (Anet) 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: The following DAG illustrates the causal relationships among the variables: *P* represents *Price of gas* *O* represents *Obesity rates* *D* represents *Amount of driving* *E* represents *Amount of eating out in restaurants* *Ex* represents *Amount of exercise*

```r
pacman::p_load(dagitty,
               ggdag)
dag <- dagitty("dag{
P -> D -> E -> O
P -> D -> Ex -> O}")
ggdag(dag, layout = "circle")+
  theme_dag()
```



Mechanism 1

Main effect Gas prices are argued to be negatively correlated with obesity rates.

Mediators (indirect effect) Gas prices are negatively correlated with the amount of driving; The amount of driving is negatively correlated with the amount of exercise; The amount of exercise is negatively correlated with obesity rates.

Multiple regression for Mechanism 1 This mechanism could be addressed by modelling a multiple regression which includes three predictors, gas prices, amount of driving and amount of exercise, and obesity rates as the outcome variable. By accounting for the mediators, it would be possible to see the direct effect of gas prices on obesity rates, while also seeing the individual effects of amount of driving and amount of exercise.

It might also be helpful to model a simple regression where obesity rates are predicted only by gas prices. This would show only the direct effect of gas prices and would be informative for comparing with the full model with the mediators.

Mechanism 2

Main effect Gas prices are argued to be negatively correlated with obesity rates.

Mediators (indirect effect) Gas prices are negatively correlated with the amount of driving; The amount of driving is negatively correlated with the amount of eating out in restaurants; The amount of eating out in restaurant is negatively correlated with obesity rates.

Multiple regression for Mechanism 2 This mechanism could be addressed by modelling a multiple regression which includes three predictors, gas prices, amount of driving and amount of eating out in restaurants, and obesity rates as the outcome variable. By accounting for the mediators, it would be possible to see the direct effect of gas prices on obesity rates, while also seeing the individual effects of amount of driving and amount of eating out in restaurants.

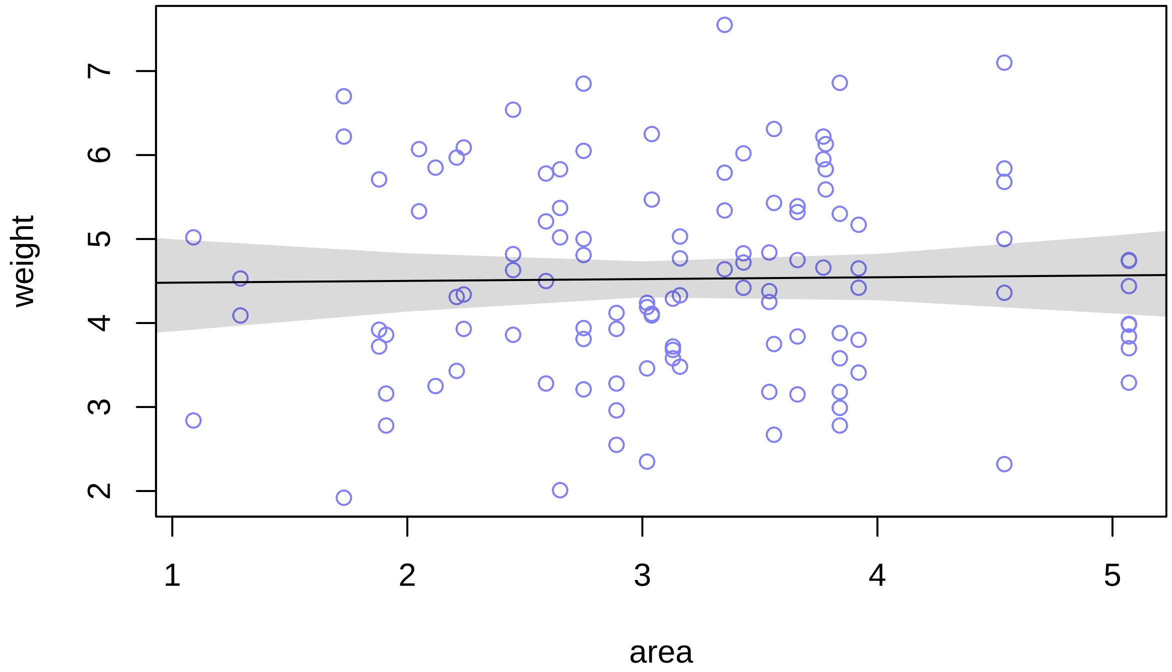
It might also be helpful to model a simple regression where obesity rates are predicted only by gas prices. This would show only the direct effect of gas prices and would be informative for comparing with the full model with the mediators.

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. (Emma) 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?

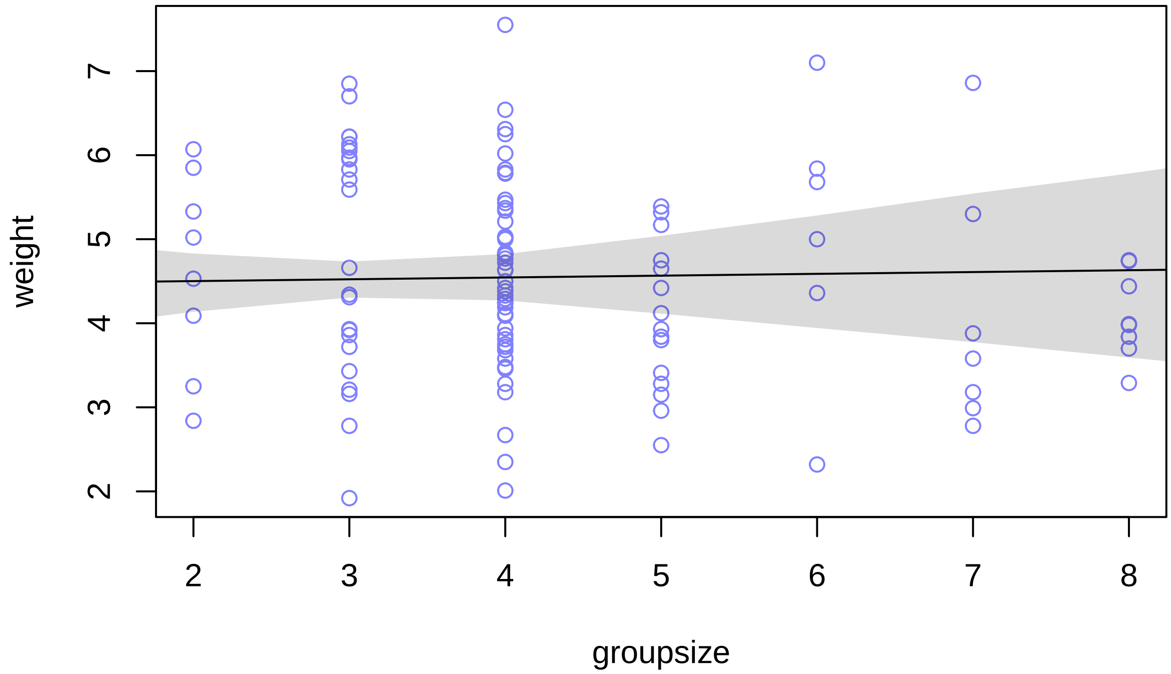
```r
data(foxes)

# bivariate gaussian regressions
m1 <- quap(
  alist(
    weight ~ dnorm(mu, sigma),
    mu ~ a + b_area * area,
    a ~ dnorm(4.5, 2), # intercept
    b_area ~ dnorm(0, 15), # slope
    sigma ~ dexp(1) # sd
  ), data = foxes
)

#precis(foxes)
plot(weight ~ area, foxes, col=rangi2)

area.seq <- seq( from=0 , to=70 , by=1 )
weight <- link(m1, data=data.frame(area=area.seq))
weight.mean <- apply(weight , 2 , mean)
weight.PI <- apply(weight , 2 , PI , prob=0.95) # get 95% PI

lines(area.seq, weight.mean) # plot regression line
shade(weight.PI, area.seq) # 89% interval for the mean
```



```
m2 <- quap(
  alist(
    weight ~ dnorm(mu, sigma),
    mu ~ a + b_groupsize * groupsize,
    a ~ dnorm(4.5, 2), # intercept
    b_groupsize ~ dnorm(0, 15), # slope
    sigma ~ dexp(1) # sd
  ), data = foxes
)

plot(weight ~ groupsize, foxes, col=rangi2)

groupsize.seq <- seq( from=0 , to=70 , by=1 )
weight <- link(m2, data=data.frame(groupsize=groupsize.seq))
```

```
## Error in mvrnorm(n = n, mu = mu, Sigma = vcov(object)): 'Sigma' is not positive definite
```

```
weight.mean <- apply(weight , 2 , mean)
weight.PI <- apply(weight , 2 , PI , prob=0.95) # get 95% PI

lines(groupsize.seq, weight.mean) # plot regression line
shade(weight.PI, groupsize.seq)
```



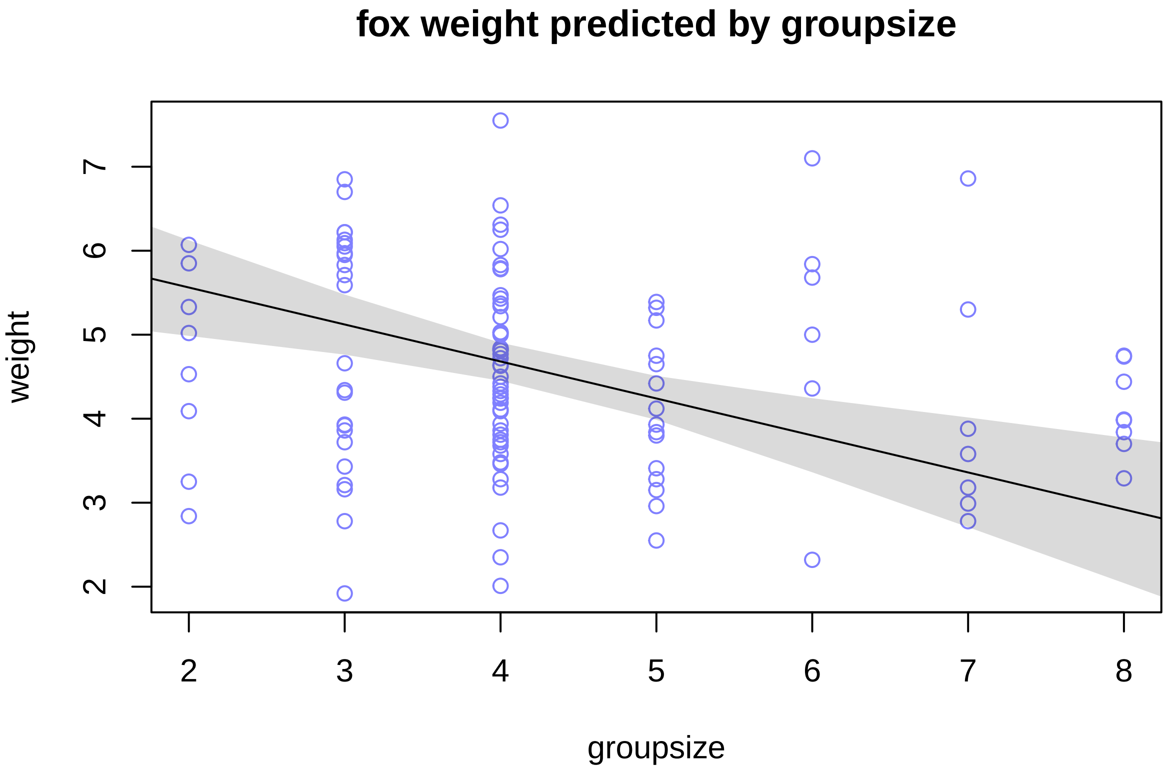
Answer:

Neither of the variables seem to predict weight well, although model 2 (m2) with groupsize as a predictor suggests a slight negative relationship.

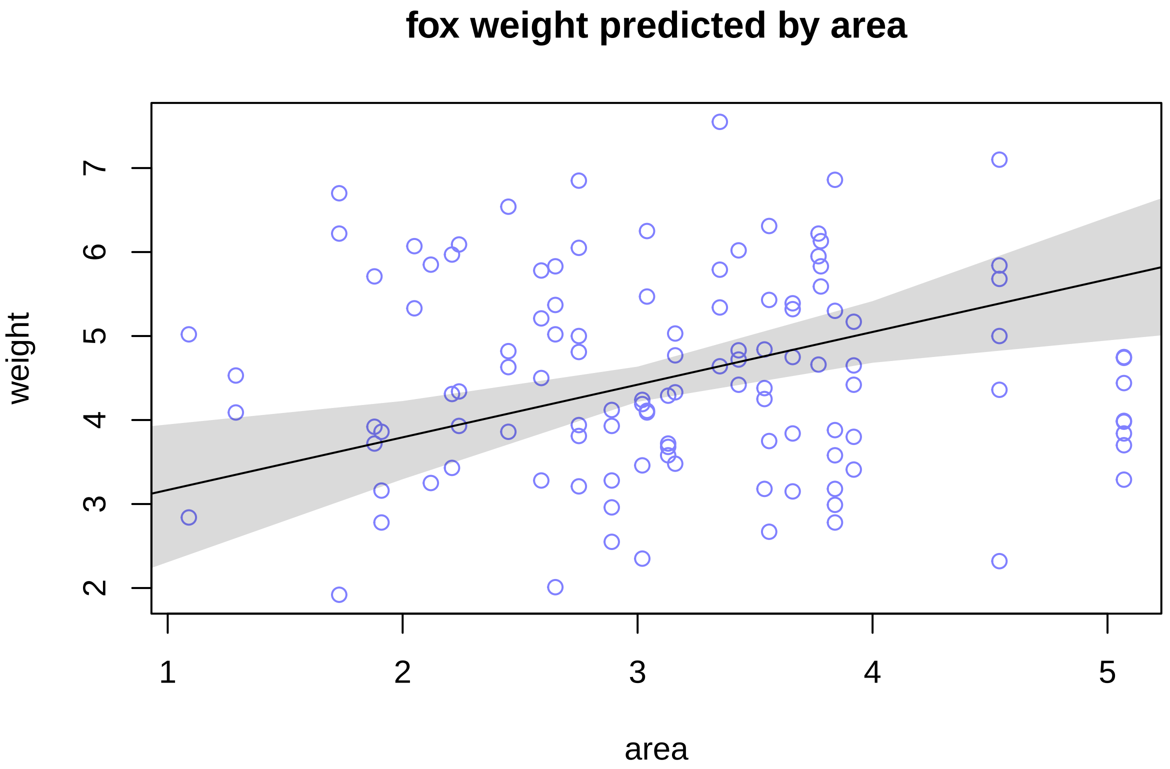
5H2. (Emma) Now fit a multiple linear regression with weight as the outcome and both area and groupsize as predictor variables. Plot the predictions of the model for each predictor, holding the other predictor constant at its mean. What does this model say about the importance of each variable? Why do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

```
m3 <- quap(
  alist(
    weight ~ dnorm(mu, sigma),
    mu <- a + b_groupsize * groupsize + b_area * area,
    a ~ dnorm(4.5, 2), # intercept
    b_groupsize ~ dnorm(0, 15), # slope
    b_area ~ dnorm(0, 15),
    sigma ~ dexp(1) # sd
  ), data = foxes
)

plot(weight ~ groupsize, foxes, col=rangi2)

groupsize.seq <- seq( from=0 , to=70 , by=1 )
weight <- link(m3, data=data.frame(groupsize=groupsize.seq, area=mean(foxes$area)))
```

```
weight.mean <- apply(weight , 2 , mean)
weight.PI <- apply(weight , 2 , PI , prob=0.95) # get 95% PI

lines(groupsize.seq, weight.mean) # plot regression line
shade(weight.PI, groupsize.seq)
title("fox weight predicted by groupsize")
```



```
plot(weight ~ area, foxes, col=rangi2)

area.seq <- seq( from=0 , to=70 , by=1 )
weight <- link(m3, data=data.frame(area=area.seq, groupsize=mean(foxes$groupsize)))
weight.mean <- apply(weight , 2 , mean)
weight.PI <- apply(weight , 2 , PI , prob=0.95) # get 95% PI

lines(area.seq, weight.mean) # plot regression line
shade(weight.PI, area.seq)
title("fox weight predicted by area")
```



Answer:

In model 3 (m3), we include both predictor variables, area and groupsize, to predict fox weight. Previously, we used separate models with only one predictor at a time. The difference in results between these models suggests that in the simpler models, the full effect of groupsize and area was not fully accounted for. By including both predictors in Model 3, we can see their independent contributions, which were previously masked due to potential negative correlations between the predictor variables.

5H3. (Emma) 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.

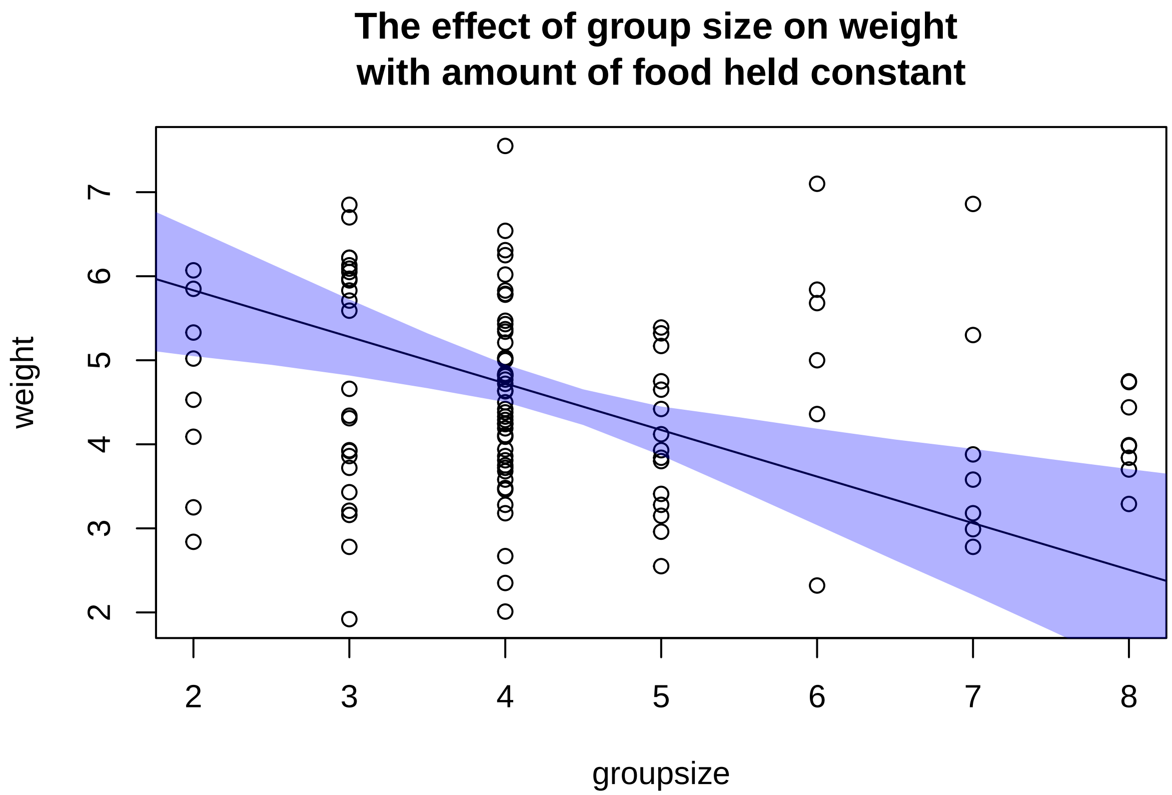
ANSWER It seems that avgfood is a better predictor based on the plots from model 4 as well as model 3 results, especially in the extreme values where avgfood predicts extreme weight values that area does not. Therefore, I would choose avgfood to include in a model.

(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?

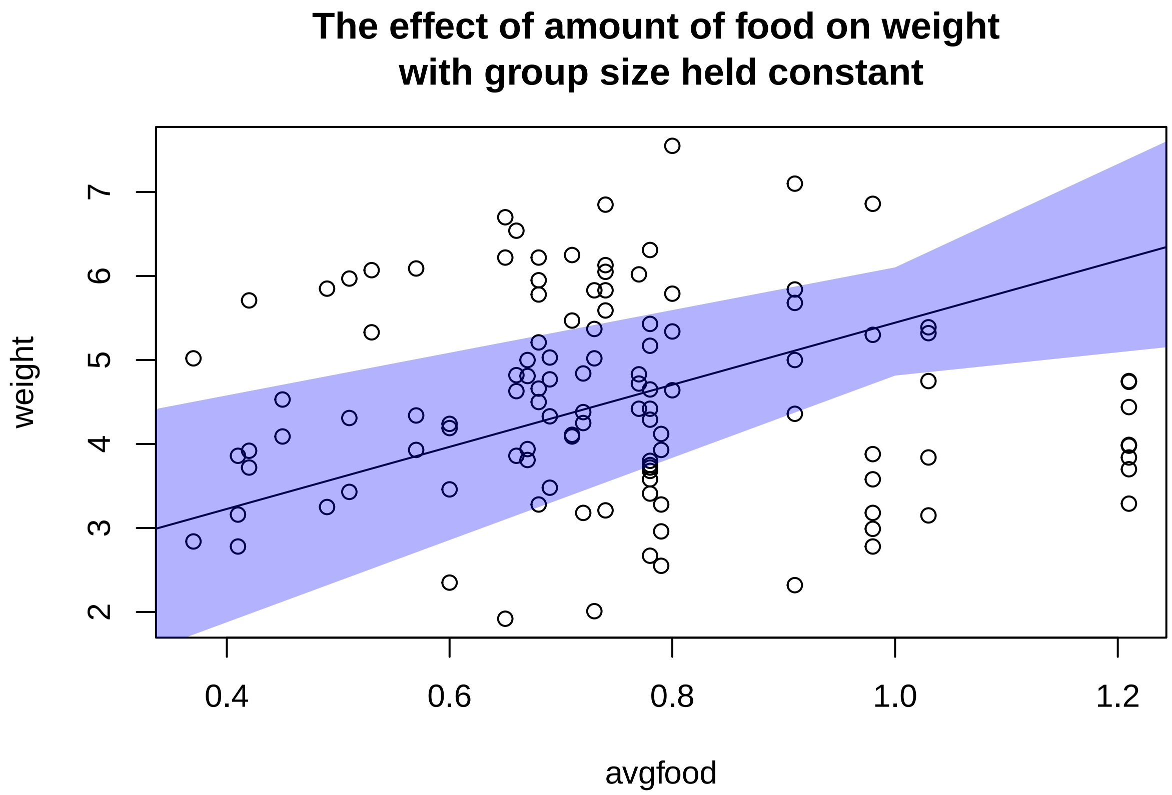
ANSWER Since both of the variables have a positive effect on the outcome variable, there are pretty much an infinite amount of different combinations of positive parameter estimates that create a given overall positive effect (product of the two), which explains the larger standard errors. The effects come out smaller as the effect is being 'spread out' over two predictors that both have a similar effect.

```r
# creating  formula for model 4
model_4_list <- alist(
  weight ~ dnorm(mu, sd),
  mu <- a + bavgfood*avgfood + bg*groupsize,
  a ~ dnorm(0,100),
  bavgfood ~ dnorm(0,10),
  bg ~ dnorm(0,10),
  sd ~ dunif(0,50)
)
# using quadratic approximation for model 4
model_4 <- quap(model_4_list, foxes)
```

```r
# making dataframes with one variable having only the mean value
group_seq <- seq(1,9, by = 0.5)
avgfood_seq <- seq(from = -2, to = 2, by = 1)
data_1 <- data.frame(groupsize = group_seq , avgfood = mean(foxes$avgfood))
data_2 <- data.frame(avgfood = avgfood_seq, groupsize= mean(foxes$groupsize))

# getting predicted values of model 4
model_pred_1 <- link(model_4 , data=data_1)
model_pred_2 <- link(model_4 , data=data_2)

# getting the mean values of model 4 predictions
model_pred_1_mean <- apply(model_pred_1, 2, mean)
model_pred_2_mean <- apply(model_pred_2, 2, mean)

# getting the prediction intervals for the mean values of model 4 predictions
pi_mu_1 <- apply(model_pred_1, 2, function(x){PI(x,0.95)})
pi_mu_2 <- apply(model_pred_2, 2, function(x){PI(x,0.95)})

# plotting the results
plot( weight ~ groupsize , data=foxes)
lines( group_seq, model_pred_1_mean )
shade(pi_mu_1, group_seq, col=adjustcolor("blue", alpha.f = 0.3))
title(main = 'The effect of group size on weight \nwith amount of food held constant')
```



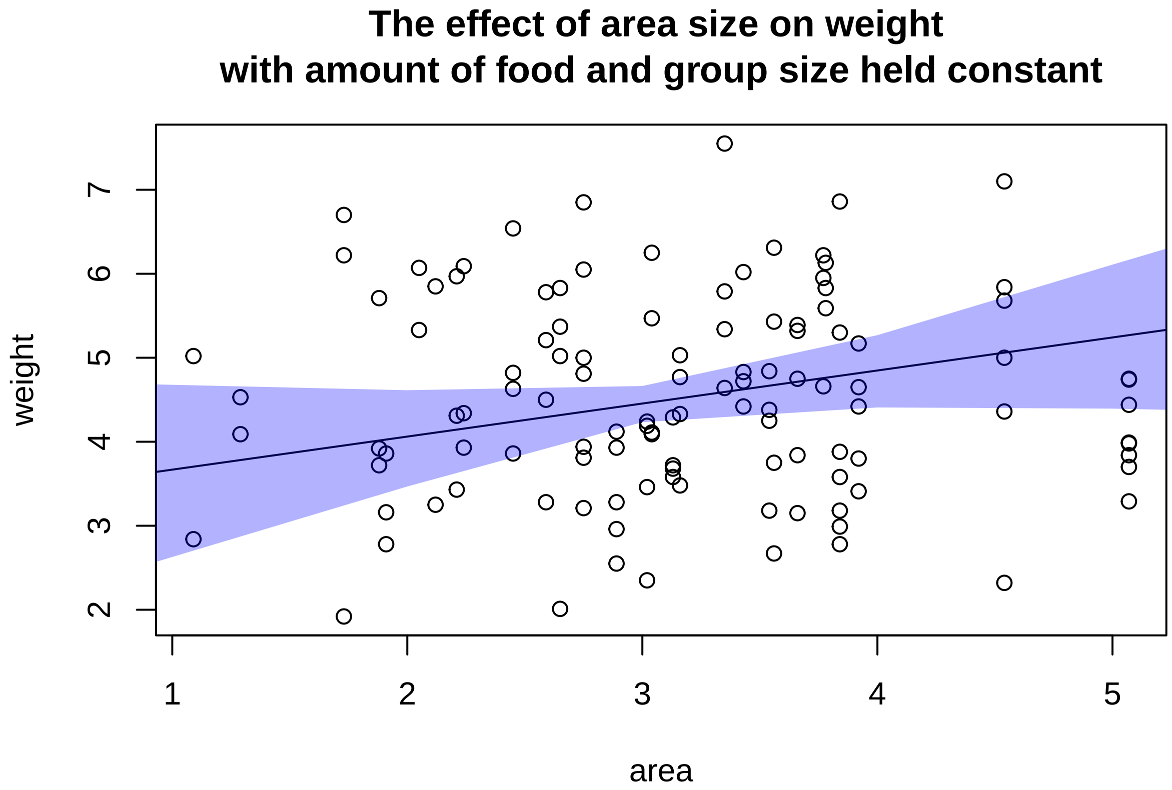
```
plot( weight ~ avgfood , data= foxes)
lines( avgfood_seq , model_pred_2_mean )
shade(pi_mu_2, avgfood_seq, col=adjustcolor("blue", alpha.f = 0.3))
title(main = 'The effect of amount of food on weight \nwith group size held constant')
```



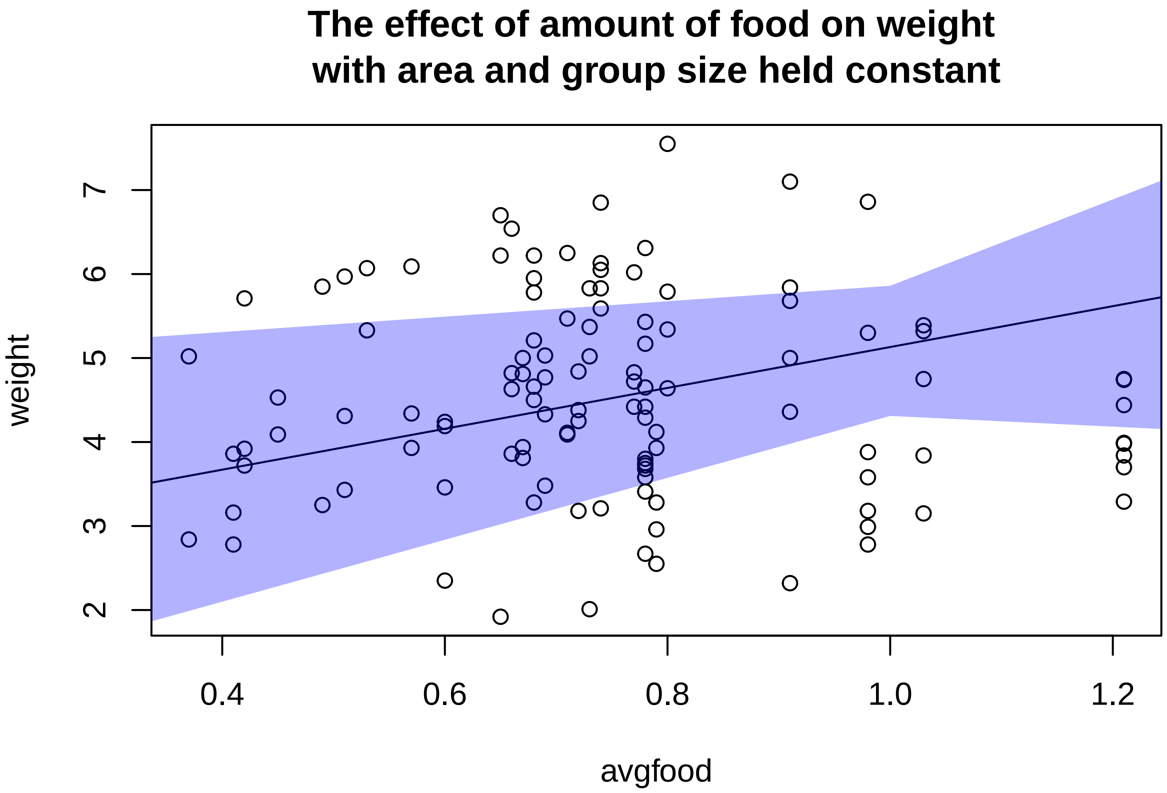
```r
# creating  formula for model 5
model_5_list <- alist(
  weight ~ dnorm(mu, sd),
  mu <- a + ba*area + bg*groupsize + bavgf*avgfood,
  a ~ dnorm(0,100),
  ba ~ dnorm(0,10),
  bg ~ dnorm(0,10),
  bavgf ~ dnorm(0,10),
  sd ~ dunif(0,50)
)
# using quadratic approximation for model 5
model_5 <- quap(model_5_list, foxes)
```

```r
#Emma Margaret
# making dataframes with one variable having only the mean value
area_seq <- seq(from = -2, to = 6, by = 1)
data_1 <- data.frame(area = area_seq, groupsize = mean(foxes$groupsize), avgfood = mean(foxes$avgfood))
data_2 <- data.frame(avgfood = avgfood_seq, area= mean(foxes$area), groupsize = mean(foxes$groupsize))
data_3 <- data.frame(groupsize = group_seq, area= mean(foxes$area), avgfood= mean(foxes$avgfood))

# getting predicted values of model 5
model_pred_1 <- link(model_5 , data=data_1)
model_pred_2 <- link(model_5 , data=data_2)
model_pred_3 <- link(model_5 , data=data_3)

# getting the mean values of model 5 predictions
```

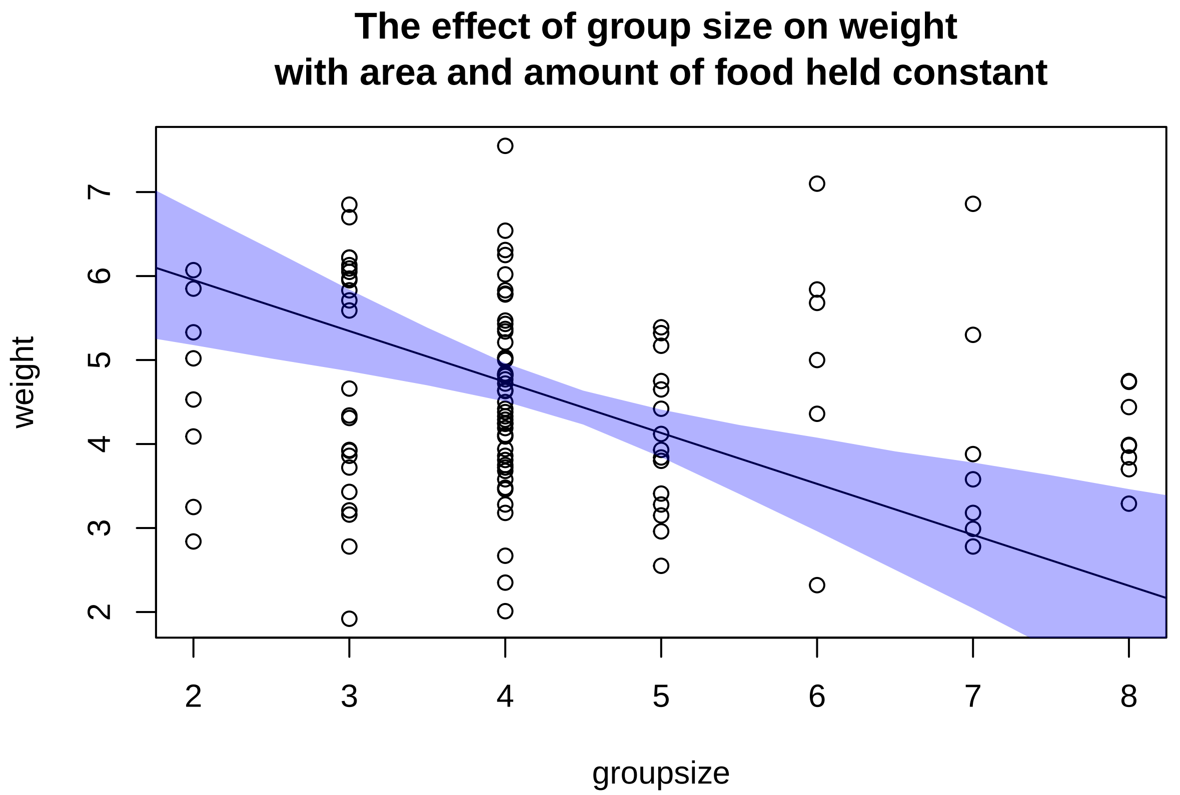
```r
model_pred_1_mean <- apply(model_pred_1, 2, mean)
model_pred_2_mean <- apply(model_pred_2, 2, mean)
model_pred_3_mean <- apply(model_pred_3, 2, mean)

# getting the prediction intervals for the mean values of model 5 predictions
pi_mu_1 <- apply(model_pred_1, 2, function(x){PI(x,0.95)})
pi_mu_2 <- apply(model_pred_2, 2, function(x){PI(x,0.95)})
pi_mu_3 <- apply(model_pred_3, 2, function(x){PI(x,0.95)})

# plotting the results
plot( weight ~ area , data=foxes )
lines( area_seq, model_pred_1_mean )
shade(pi_mu_1, area_seq, col=adjustcolor("blue", alpha.f = 0.3))
title(main = 'The effect of area size on weight \nwith amount of food and group size held constant')
```



```r
plot( weight ~ avgfood , data= foxes)
lines( avgfood_seq , model_pred_2_mean )
shade(pi_mu_2, avgfood_seq, col=adjustcolor("blue", alpha.f = 0.3))
title(main = 'The effect of amount of food on weight \nwith area and group size held constant')
```



```r
plot( weight ~ groupsize , data= foxes)
lines( group_seq , model_pred_3_mean )
shade(pi_mu_3, group_seq, col=adjustcolor("blue", alpha.f = 0.3))
title(main = 'The effect of group size on weight \nwith area and amount of food held constant')
```



```r
# getting the summaries of model 4 and model 5
precis(model_4)
```

```
##                 mean         sd       5.5%      94.5%
## a          4.1369139 0.43078867  3.4484304  4.8253974
## bavgfood   3.7712281 1.20416427  1.8467410  5.6957152
## bg        -0.5620904 0.15542842 -0.8104951 -0.3136858
## sd         1.1166134 0.07331056  0.9994490  1.2337778
```

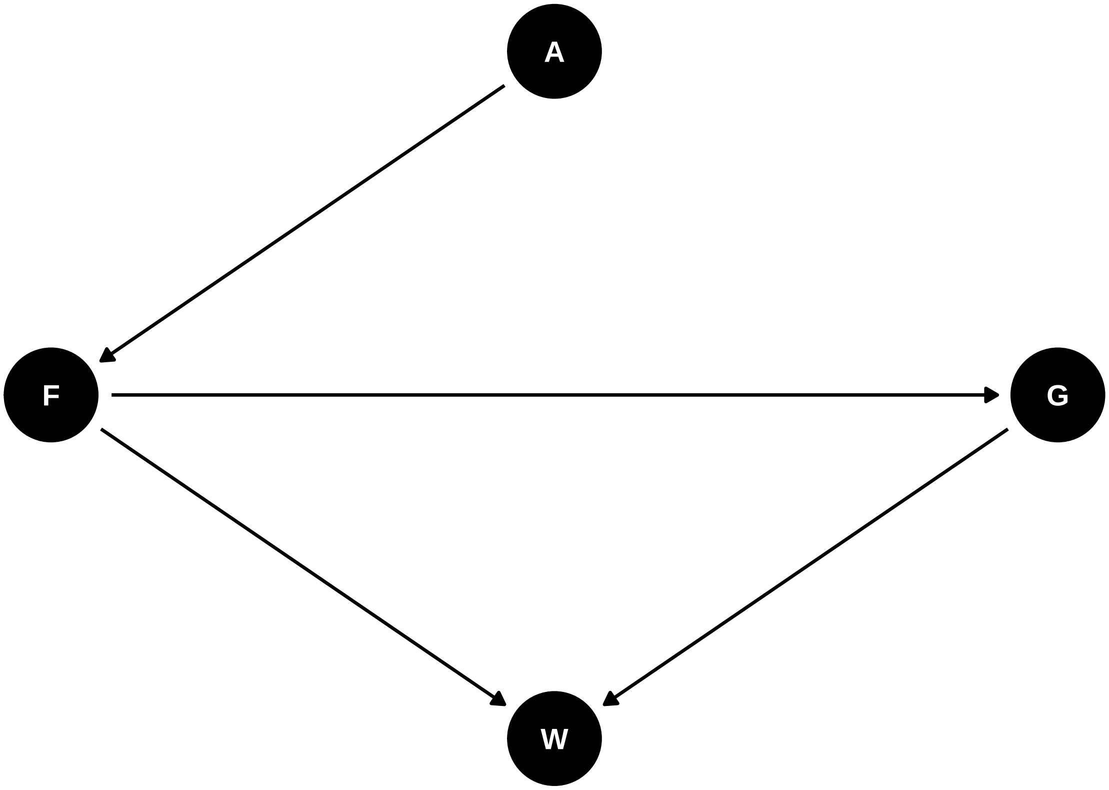
```r
precis(model_5)
```

```
##              mean         sd         5.5%      94.5%
## a       4.0718815 0.42792022  3.387982397  4.7557807
## ba      0.3895610 0.23846944  0.008440807  0.7706813
## bg     -0.6040034 0.15584420 -0.853072526 -0.3549343
## bavgf   2.4576710 1.43769097  0.159963180  4.7553789
## sd      1.1043413 0.07250289  0.988467706  1.2202149
```

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:

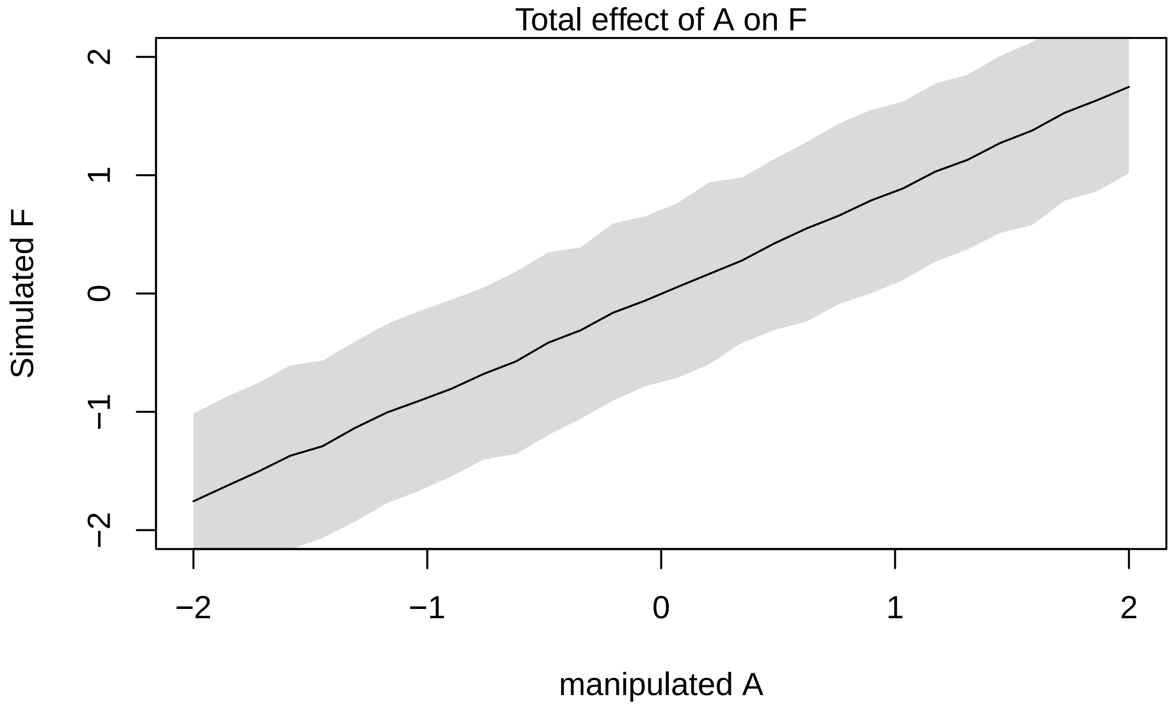
1) Estimate the total causal influence of A on F. What effect would increasing the area of a territory have on the amount of food inside of it? (Julianne)

```
#standardise data
data <- data(foxes)
fox = foxes
data_fox <- fox
```

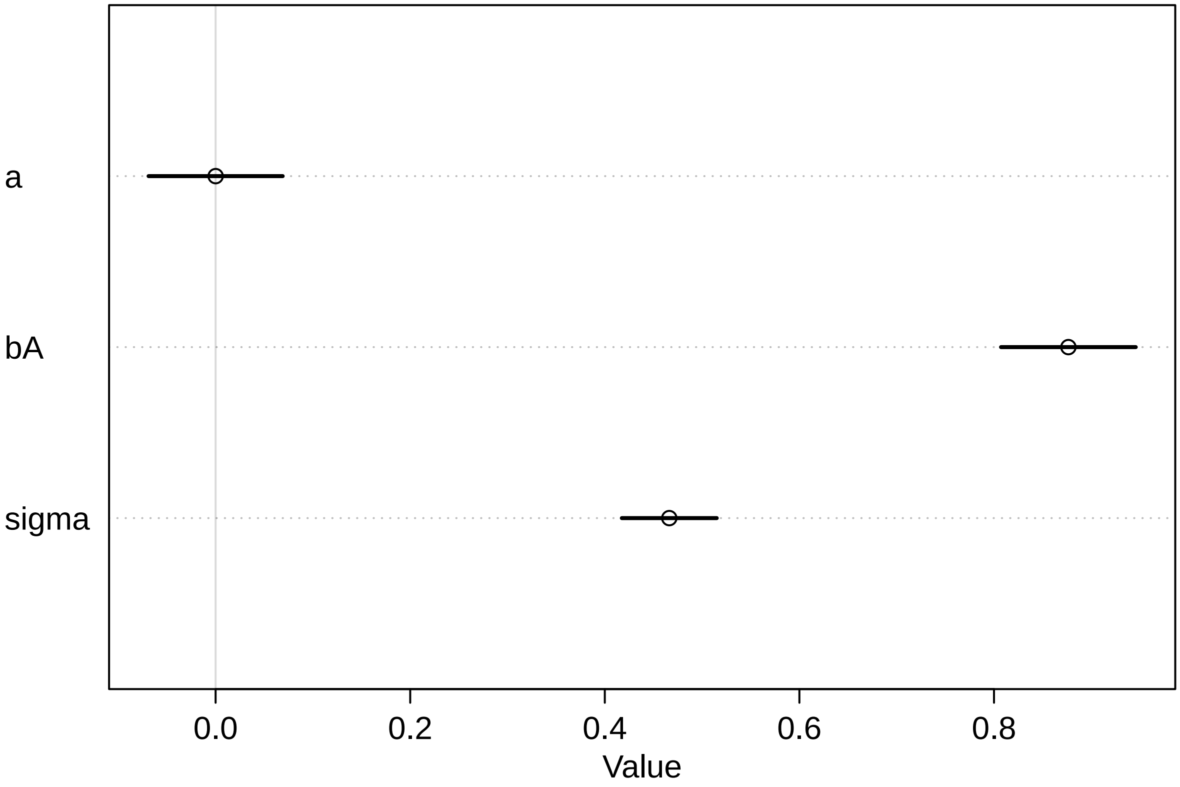
```
d = list()
d$F <- standardize(data_fox$avgfood)
d$A <- standardize(data_fox$area)

model7 <- quap(
  alist(
    ##A -> F
    F ~ dnorm(mu, sigma),
    mu <- a + bA*A,
    a ~ dnorm(0, 0.5),
    bA ~ dnorm(0, 0.5),
    sigma ~ dexp(1)
  ), data = d
)

precis(model7)
```

```
##               mean         sd        5.5%      94.5%
## a     3.265069e-05 0.04312977 -0.06889705 0.06896235
## bA    8.764615e-01 0.04332408  0.80722125 0.94570175
## sigma 4.662597e-01 0.03052482  0.41747515 0.51504425
```

```
#simulating and plotting
A_seq <- seq(from = -2, to = 2, length.out = 30)

sim_data <- data.frame(A=A_seq)
s <- sim(model7, data = sim_data, vars = "F")

plot(sim_data$A, colMeans(s), ylim=c(-2,2), type="l", xlab="manipulated A", ylab="Simulated F")

shade(apply(s,2,PI), sim_data$A)
mtext("Total effect of A on F")
```



```
plot(precis(model7))
```



ANSWER

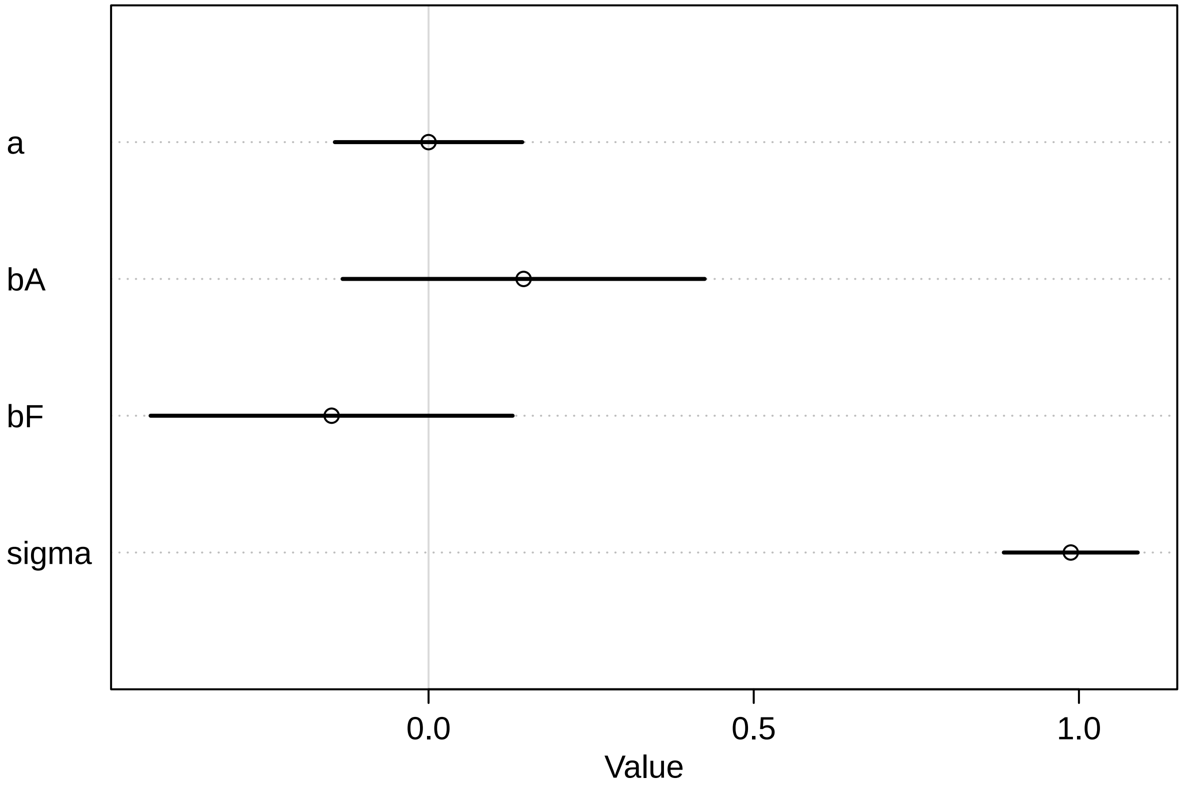
The amount of food increases as the area increases, with a coefficient of 0.88, meaning that as the area expands by one unit, the quantity of food increases by 0.88 units.

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? (Julianne)

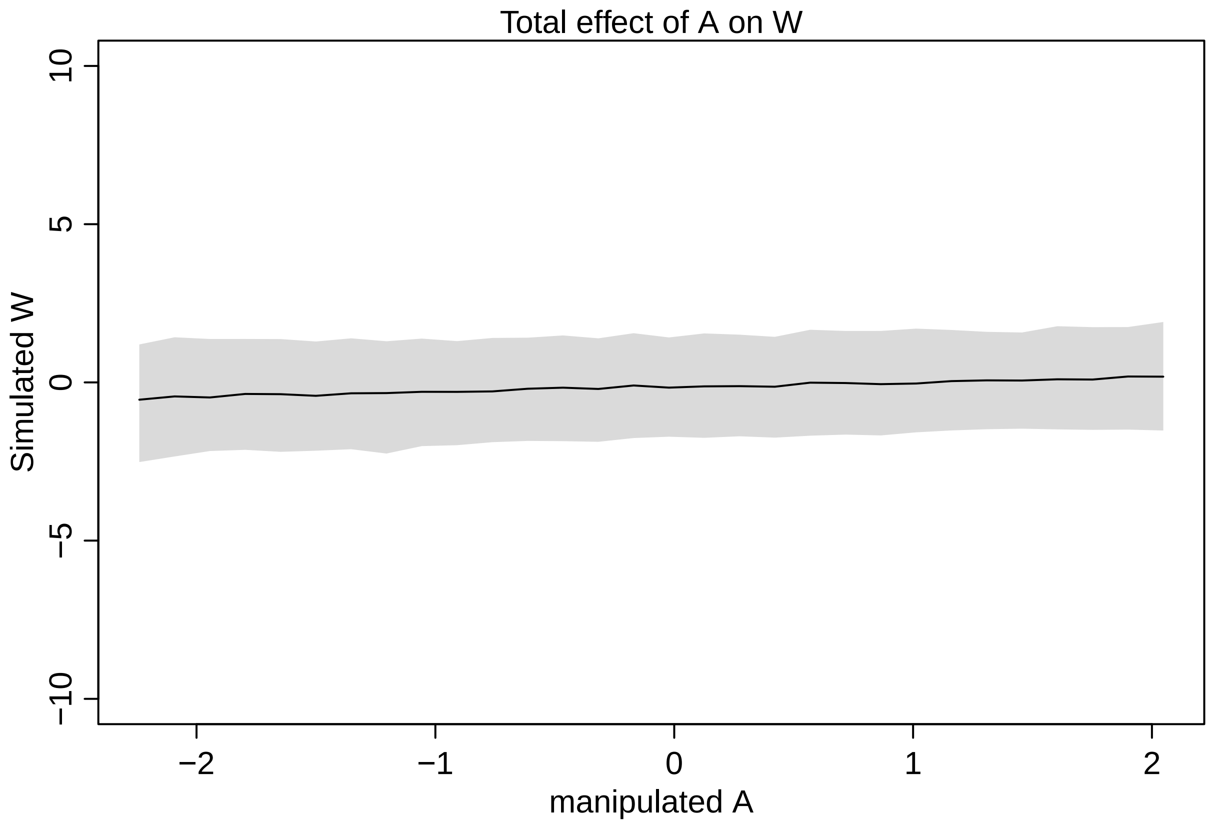
```
d$W <- standardize(data_fox$weight)

model8 <- quap(
  alist(
    ##A -> F -> W
    W ~ dnorm(mu, sigma),
    mu <- a + bF*F + bA*A,
    a ~ dnorm(0, 0.5),
    bA ~ dnorm(0, 0.5),
    bF ~ dnorm(0, 0.5),
    sigma ~ dexp(1)
  ), data = d
)

precis(model8)
```

```
##                  mean          sd        5.5%     94.5%
## a      -6.459286e-08 0.09018056 -0.1441260 0.1441259
## bA      1.461373e-01 0.17418834 -0.1322493 0.4245239
## bF     -1.490382e-01 0.17418851 -0.4274251 0.1293487
## sigma   9.874684e-01 0.06444177  0.8844780 1.0904588
```

```
plot(precis(model8))
```



```
A_seq2 <- seq(from = min(d$A), to = max(d$A), length.out = 30)
sim_data2 <- data.frame(A = A_seq2)
s2 <- sim(model8, data = sim_data2, vars=c("F", "W"))

plot(sim_data2$A, colMeans(s2$W), ylim=c(-10,10), type="l" , xlab="manipulated A" ,ylab="Simulated W" )
shade(apply(s2$W,2,PI), sim_data2$A)
mtext("Total effect of A on W")
```



ANSWER

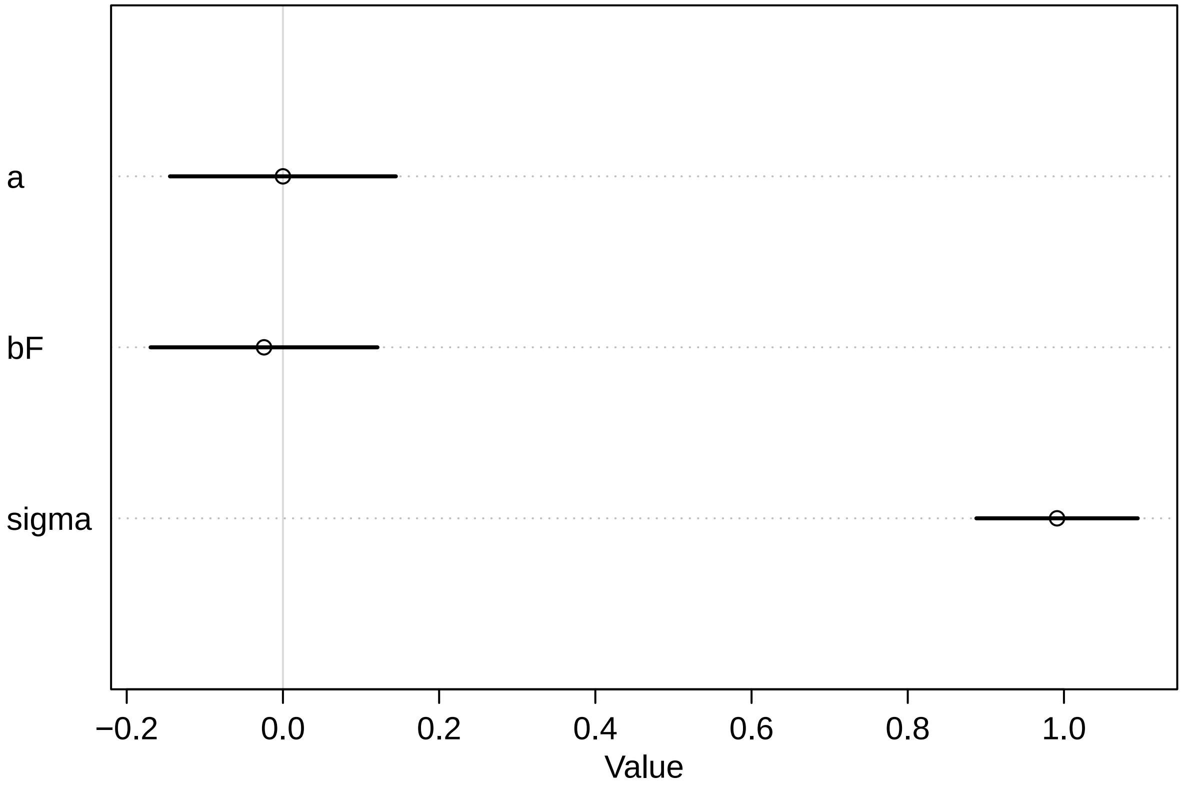
The total causal effect of adding food to a territory on the weight of foxes in the plot appears rather small, because there is an equally big positive influence of area on food (0.15) as there is a negative one on food on fox weight (-0.15), so the line becomes flat. This is likely due to a strong correlation (as seen in question 1 above) between area and food availability. The sigma of 0.99 indicates a very large uncertainty in the data, so the predictors we are using are not ideal for explaining the relationship between the foxes' living conditions and their weight.

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? (Julianne)

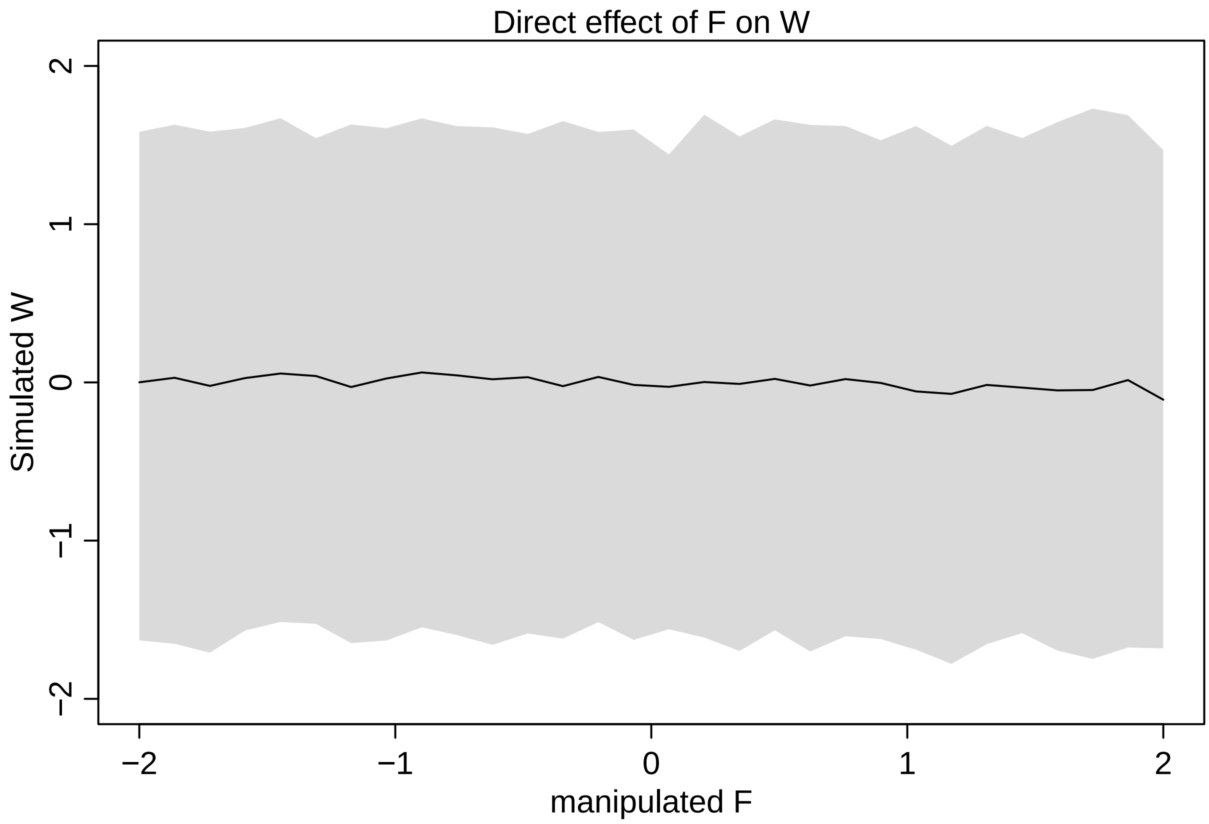
```
model9 <- quap(
  alist(
    ##F -> W
    W ~ dnorm(mu, sigma),
    mu <- a + bF*F,
    a ~ dnorm(0, 0.5),
    bF ~ dnorm(0, 0.5),
    sigma ~ dexp(1)
  ), data = d
)

precis(model9)
```

```
##                  mean         sd       5.5%     94.5%
## a       2.999301e-07 0.09050663 -0.1446468 0.1446474
## bF     -2.421022e-02 0.09088640 -0.1694642 0.1210438
## sigma   9.911596e-01 0.06466112  0.8878186 1.0945005
```

```
plot(precis(model9))
```



```
#simulating and plotting
F_seq3 <- seq(from = -2, to = 2, length.out = 30)

sim_data3 <- data.frame(F=F_seq3)
s3 <- sim(model9, data = sim_data3, vars = "W")

plot(sim_data3$F, colMeans(s3), ylim=c(-2,2), type="l", xlab="manipulated F", ylab="Simulated W")

shade(apply(s3,2,PI), sim_data3$F)
mtext("Direct effect of F on W")
```



ANSWER

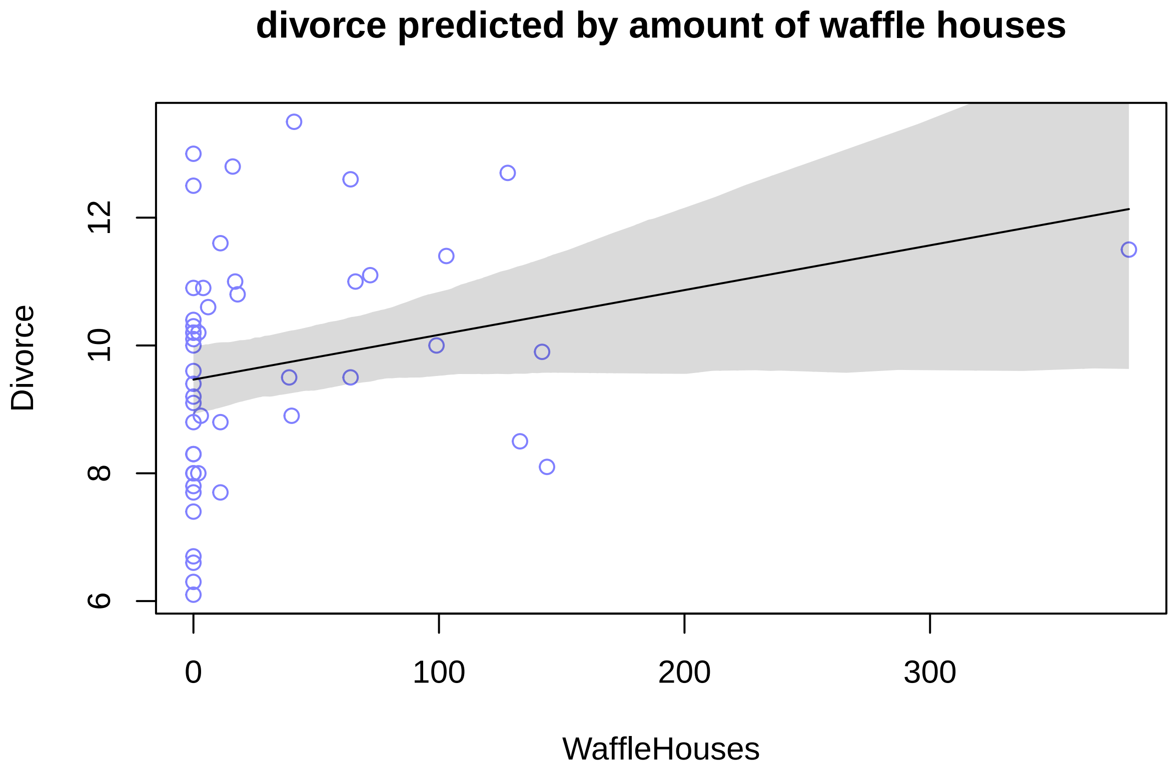
Here we can see that the direct effect of food on weight is negative -0.2, which means that increasing the amount of food by one unit, decreases the weight by -0.2 units. Again the variance is very large (0.99), so most likely we would need to include groupsize´or control for other predictors in the model to get a better explanation for how the amount of food affects the foxes' weights.

Chapter 6: Investigating the Waffles and Divorces

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

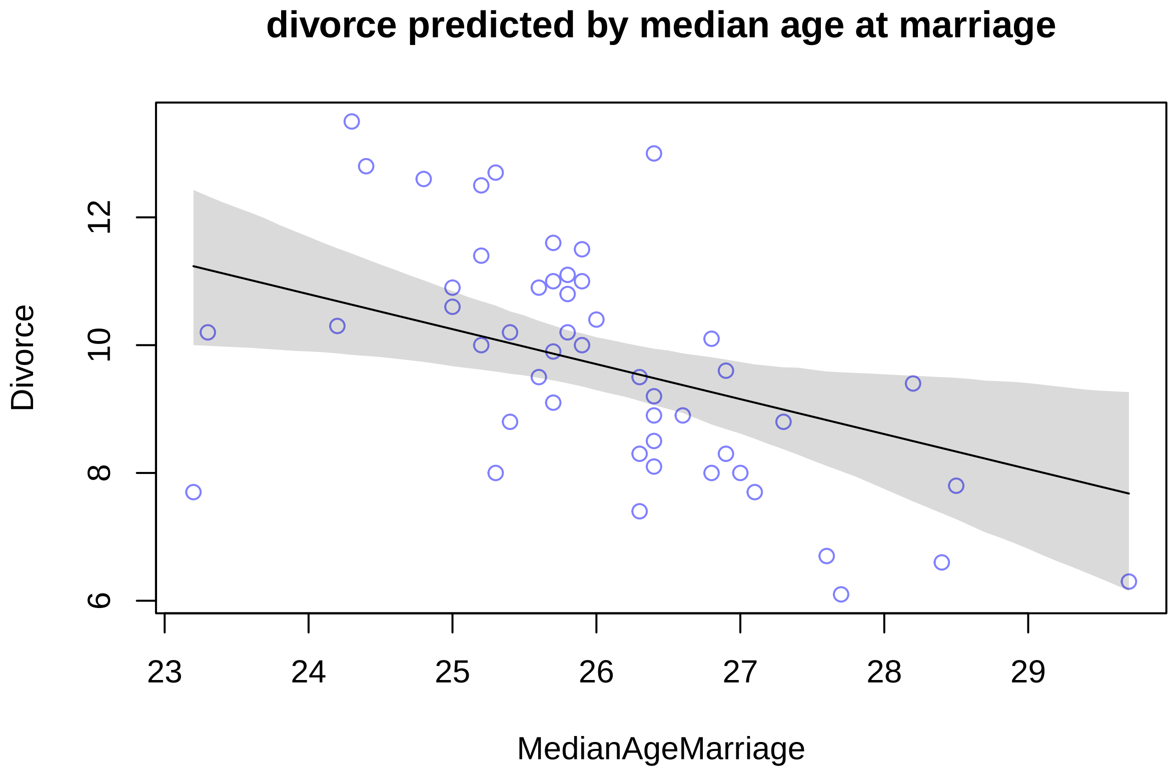
```
data(WaffleDivorce)

# onle waffle houses and divorce
m6.1 <- quap(
  alist(
    Divorce ~ dnorm(mu, sigma),
    mu <- a + b_waffle * WaffleHouses,
    a ~ dnorm(0, 10), # intercept
    b_waffle ~ dnorm(0, 5), # slope
    sigma ~ dexp(1) # sd
  ), data = WaffleDivorce
)
```

```
plot(Divorce ~ WaffleHouses, WaffleDivorce, col=rangi2)

waffle.seq <- seq( from=0 , to=381 , by=1 )
divorce <- link(m6.1, data=data.frame(WaffleHouses=waffle.seq))
divorce.mean <- apply(divorce , 2 , mean)
divorce.PI <- apply(divorce , 2 , PI , prob=0.95) # get 95% PI

lines(waffle.seq, divorce.mean) # plot regression line
shade(divorce.PI, waffle.seq)
title("divorce predicted by amount of waffle houses")
```

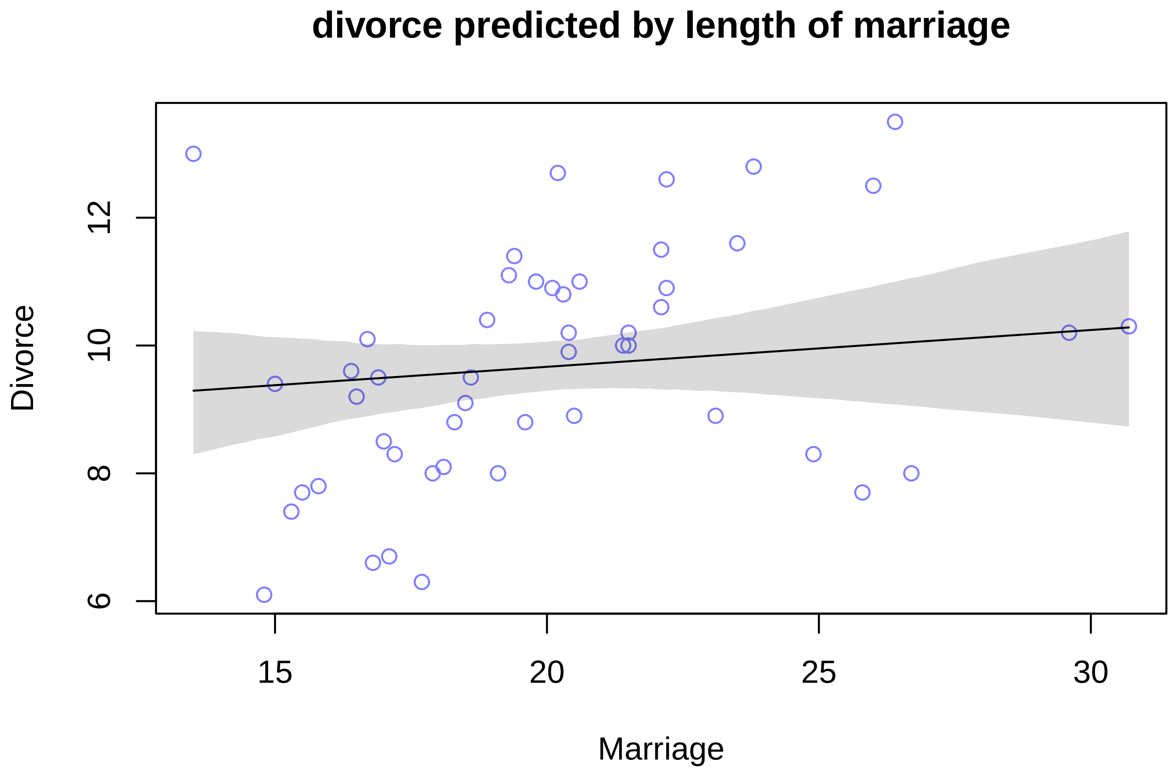


```
# key confounders
m6.2 <- quap(
  alist(
    Divorce ~ dnorm(mu, sigma),
    mu <- a + b_waffle * WaffleHouses + b_age * MedianAgeMarriage + b_marriage * Marriage,
    a ~ dnorm(0, 10), # intercept
    b_waffle ~ dnorm(0, 5), # slope
    b_age ~ dnorm(0, 5),
    b_marriage ~ dnorm(0, 5),
    sigma ~ dexp(1) # sd
  ), data = WaffleDivorce
)

plot(Divorce ~ MedianAgeMarriage, WaffleDivorce, col=rangi2)
```

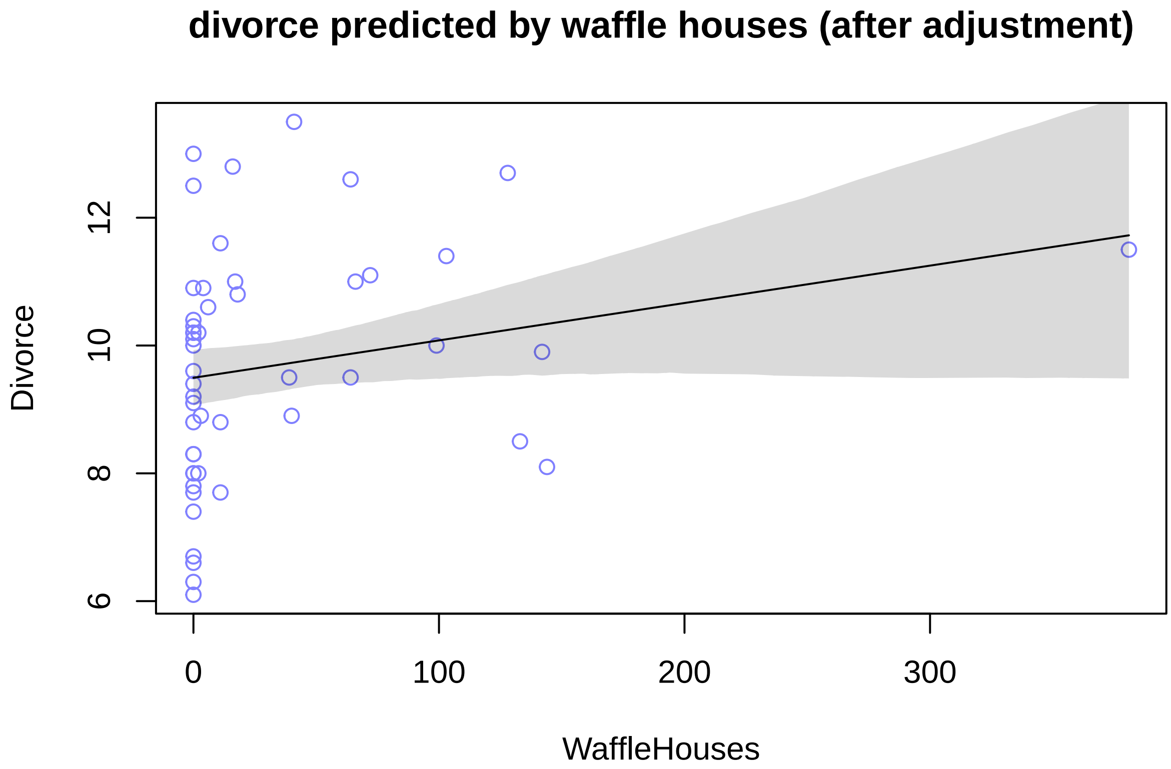
```r
age.seq <- seq( from=23.2 , to=29.7 , by=0.1 )
divorce <- link(m6.2, data=data.frame(MedianAgeMarriage=age.seq,
                                      WaffleHouses=mean(WaffleDivorce$WaffleHouses),
                                      Marriage=mean(WaffleDivorce$Marriage)))
divorce.mean <- apply(divorce , 2 , mean)
divorce.PI <- apply(divorce , 2 , PI , prob=0.95) # get 95% PI

lines(age.seq, divorce.mean) # plot regression line
shade(divorce.PI, age.seq)
title("divorce predicted by median age at marriage")
```



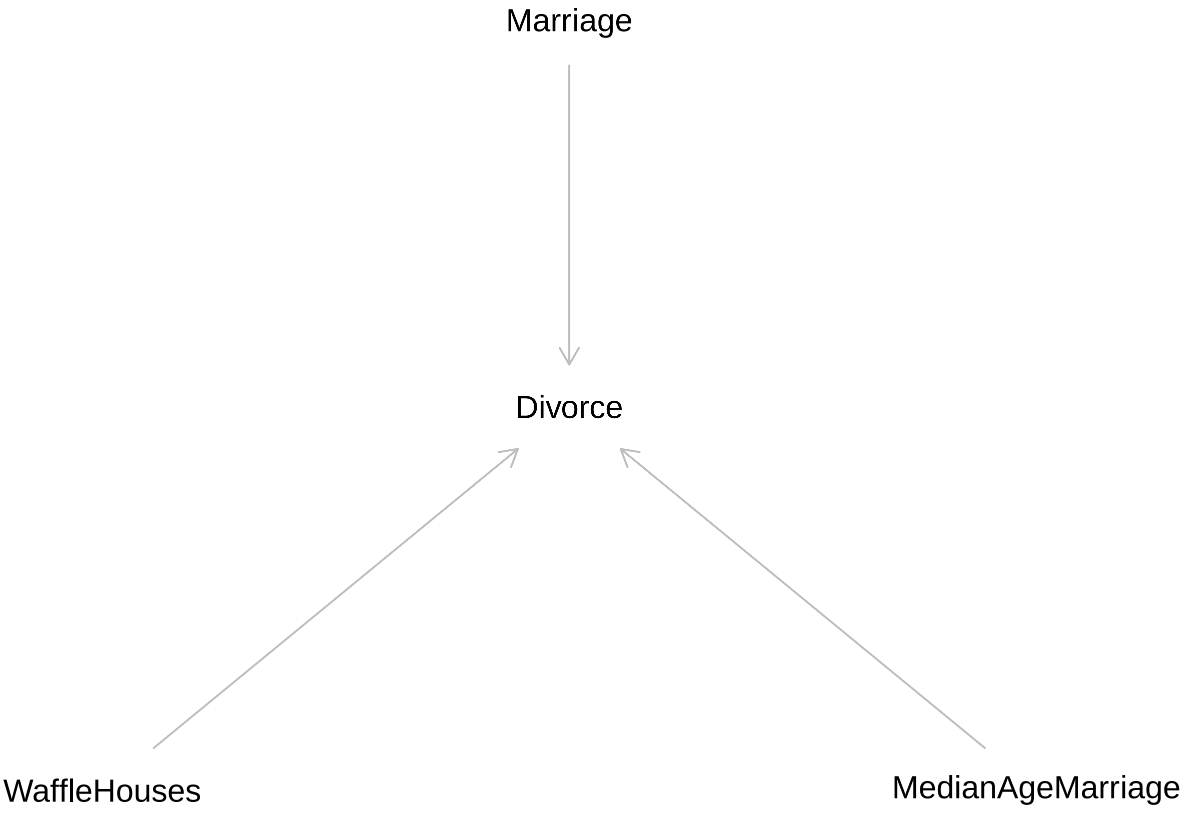
```r
plot(Divorce ~ Marriage, WaffleDivorce, col=rangi2)

marriage.seq <- seq( from=13.5 , to=30.7 , by=0.1 )
divorce <- link(m6.2, data=data.frame(Marriage=marriage.seq,
                                      WaffleHouses=mean(WaffleDivorce$WaffleHouses),
                                      MedianAgeMarriage=mean(WaffleDivorce$MedianAgeMarriage)))
divorce.mean <- apply(divorce , 2 , mean)
divorce.PI <- apply(divorce , 2 , PI , prob=0.95) # get 95% PI

lines(marriage.seq, divorce.mean) # plot regression line
shade(divorce.PI, marriage.seq)
title("divorce predicted by length of marriage")
```



```
plot(Divorce ~ WaffleHouses, data = WaffleDivorce, col = rangi2)

waffle.seq <- seq(from = 0, to = 381, by=1)
divorce <- link(m6.2, data = data.frame(WaffleHouses = waffle.seq,
                                        MedianAgeMarriage = mean(WaffleDivorce$MedianAgeMarriage),
                                        Marriage = mean(WaffleDivorce$Marriage)))

divorce.mean <- apply(divorce, 2, mean)
divorce.PI <- apply(divorce, 2, PI, prob = 0.95)

lines(waffle.seq, divorce.mean)
shade(divorce.PI, waffle.seq)
title("divorce predicted by waffle houses (after adjustment)")
```



```
dag <- dagitty('dag {
  Marriage[pos="1,-0.5"]
  Divorce[pos="1,0"]
  WaffleHouses[pos="0.5,0.5"]
  MedianAgeMarriage[pos="1.5,0.5"]
  MedianAgeMarriage -> Divorce
  Marriage -> Divorce
  WaffleHouses -> Divorce
}')

plot(dag)
```



Answer:

The initial model m6.1 suggests a potential relationship between number of waffle houses and divorce rate. However, looking at the shaded region, the model appears less confident at higher numbers of waffle houses, while confident in the lower ranges, where high fluctuations of divorce rates are present, even when no waffle houses are present at all. This suggests a correlation, but not enough evidence for a causality. Additionally, adding key confounders during the second iteration m6.2, MedianAgeMarriage and Marriage, gives the following: - The regression line for WaffleHouses and divorce rates became closer to y=constant, suggesting that the effect number of waffle houses on divorce suggested by m6.1 lessened after adding these factors. - The plot for MedianAgeMarriage and its effect on divorce rate suggests that there is a correlation that for example a higher age at marriage leads to less divorces. The model does however remain most confident at the midpoint of the age range, where there are still fluctuations in divorce rates. As such, the above DAG has been added to explain the findings. Median age at marriage seems to affect divorce rate, and additionally when accounting for it appears to lessen the effect number of waffle houses have on divorce rate, but a possible relationship may still exist.

6H2 (Leona) 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.

```
precis(m6.1)
```

```
##                 mean          sd        5.5%      94.5%
## a        9.453306531 0.270547845 9.020918821 9.88569424
## b_waffle 0.007085293 0.003721677 0.001137333 0.01303325
## sigma    1.714053697 0.167163385 1.446894321 1.98121307
```

```
precis(m6.2)
```

```
##                     mean          sd          5.5%       94.5%
## a           22.587792018 6.177887046 12.7143353216 32.46124871
## b_waffle     0.005791885 0.003109248  0.0008227058  0.01076106
## b_age       -0.548856938 0.199040080 -0.8669614283 -0.23075245
## b_marriage   0.059843194 0.069082665 -0.0505642477  0.17025063
## sigma        1.421715376 0.144370303  1.1909837483  1.65244700
```

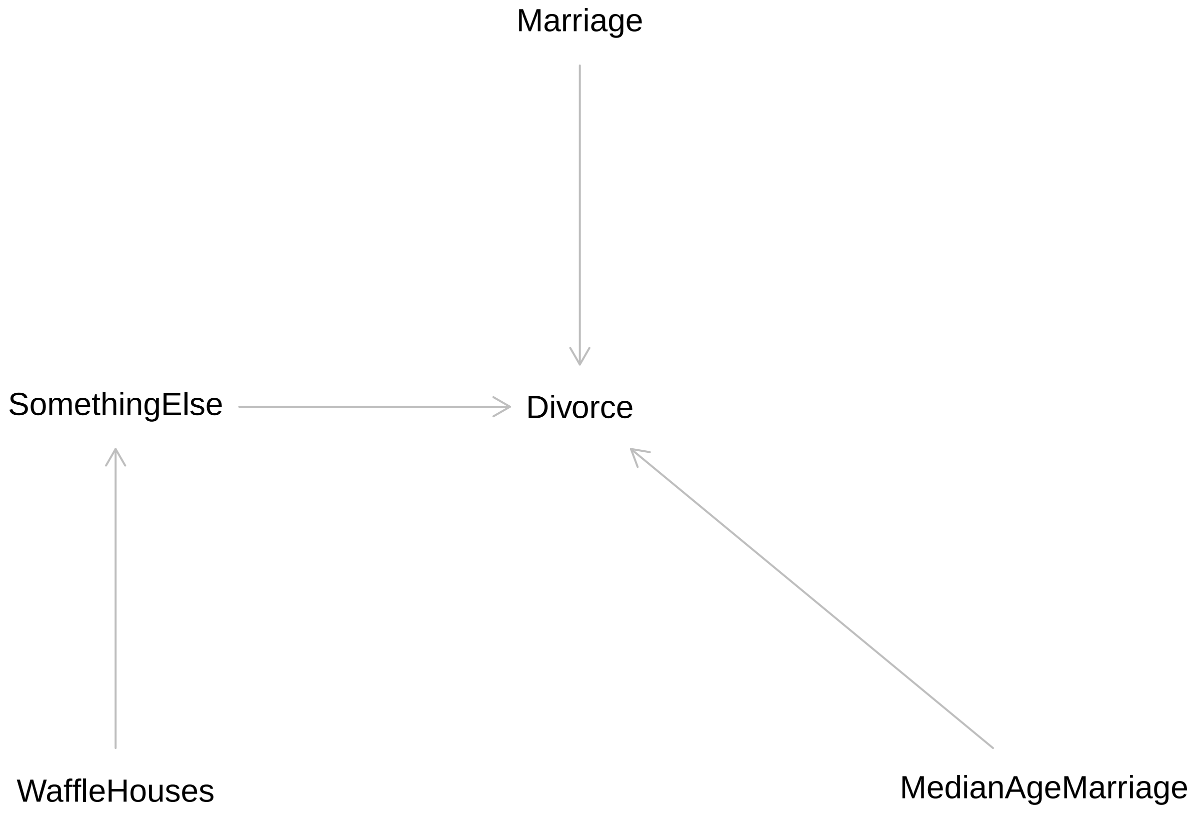
```
cor(WaffleDivorce$Marriage, WaffleDivorce$WaffleHouses)
```

```
## [1] 0.027122
```

```
m_test <- quap(
  alist(
    Marriage ~ dnorm(mu, sigma),
    mu <- a + b_waffle * WaffleHouses + b_age * MedianAgeMarriage,
    a ~ dnorm(0, 10),
    b_waffle ~ dnorm(0, 5),
    b_age ~ dnorm(0, 5),
    sigma ~ dexp(1)
  ), data = WaffleDivorce
)

precis(m_test)
```

```
##                   mean          sd        5.5%       94.5%
## a        42.9767179658 8.035781665 30.13398684 55.81944910
## b_waffle  0.0001855089 0.006514447 -0.01022584  0.01059685
## b_age    -0.8806765707 0.307263349 -1.37174275 -0.38961039
## sigma     2.9851922844 0.343379667  2.43640526  3.53397931
```

```
dag <- dagitty('dag {
  Marriage[pos="1,-0.5"]
  Divorce[pos="1,0"]
  WaffleHouses[pos="0.5,0.5"]
  MedianAgeMarriage[pos="1.5,0.5"]
  SomethingElse[pos="0.5, 0"]
  MedianAgeMarriage -> Divorce
  Marriage -> Divorce
  WaffleHouses -> SomethingElse
  SomethingElse -> Divorce
}')

plot(dag)
```



Answer:

4 tests were conducted in total: - Using the model m6.1, using precis to check the slope and intercept. This was to check the arrow from WaffleHouses to Divorce - Using the model m6.2, the same as above. To check the arrow from MedianAgeMarriage to Divorce and the one to WaffleHouses - Testing the correlation between marriage and WaffleHouses with cor. To check that there indeed should not be an arrow between Marriage and WaffleHouses. - Finally, defining a test model that checks marriage and WaffleHouses remain independent after controlling for MedianAgeMarriage.

The results showcase that b_waffle is near zero both before and after accounting for MedianAgeMarriage. This suggests that, in connection with the plots in 6H1, MedianAgeMarriage does not significantly change the effect WaffleHouses has on divorce, which in itself does not have a strong causality, although a relationship may be present. As such, these two arrows can be removed for now. Notably, there may be a missing pathway due to not considered factors, such as economic or even regional factors, such as number of waffle houses in regions that already have high divorce rates, or the general financial differences between regions with and without waffle houses. Additionally, the correlation between marriage and waffle houses is near zero, with and without MedianAgeMarriage as a confounder, meaning no arrow needs to be added. Finally, all tests show a large negative slope for b_age, suggesting that indeed the median age at marriage has a causal effect on divorce rates, with a higher age leading to lower divorce rates and vice versa. Additional variables may be needed for further causal investigation, however the DAG has for now been adjusted as above.