Assignment 2 - Methods 4

Luna Frausing

2025-02-02

Second assignment

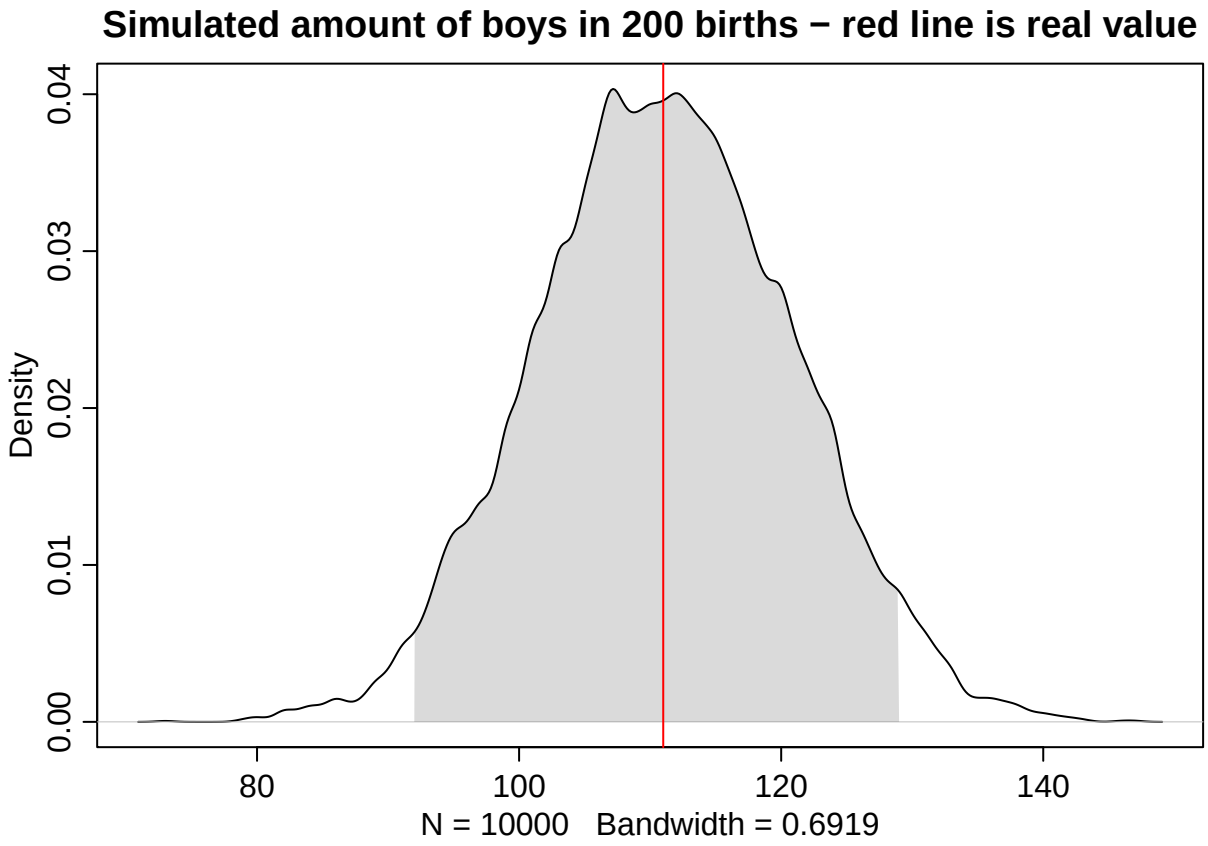
The second assignment uses chapter 3, 5 and 6. The focus of the assignment is getting an understanding of causality.

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:)

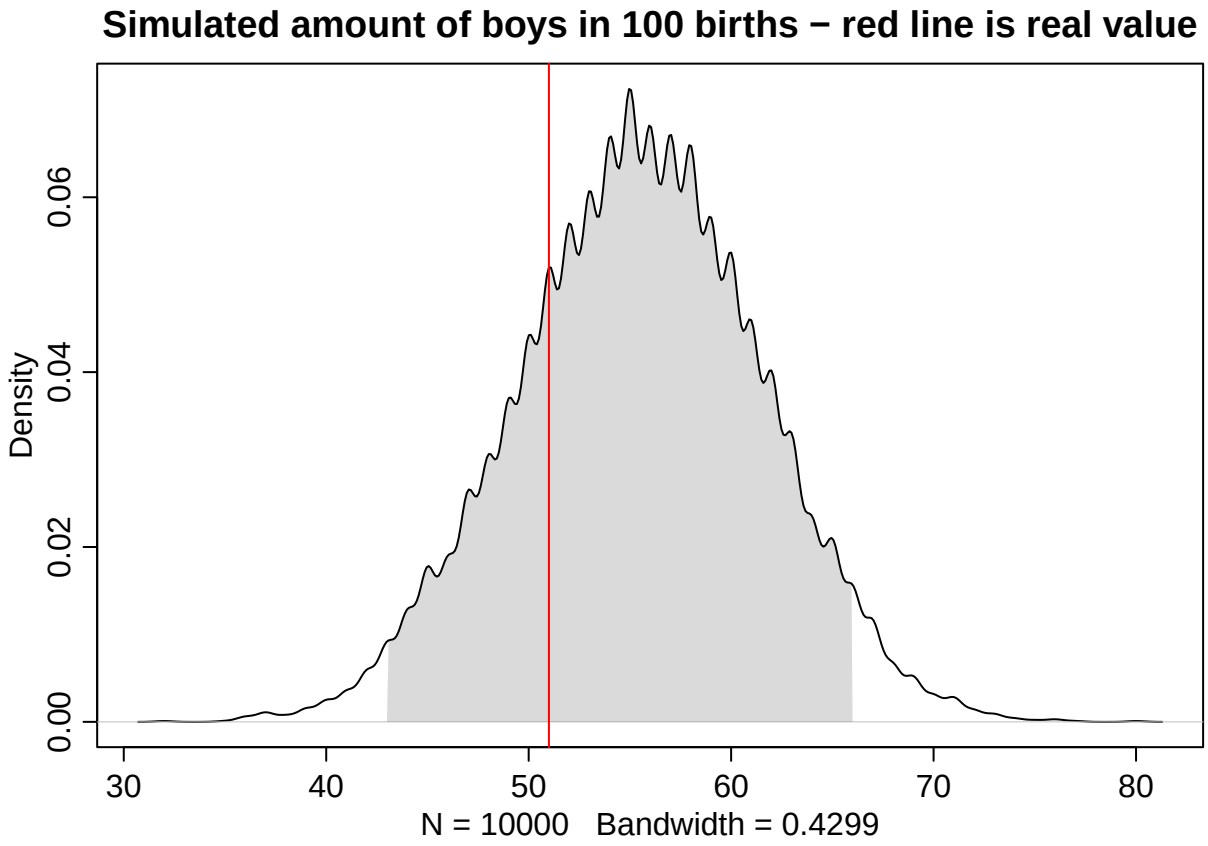
3H3 Use rbinom to simulate 10,000 replicates of 200 births. You should end up with 10,000 numbers, each one a count of boys out of 200 births. Compare the distribution of predicted numbers of boys to the actual count in the data (111 boys out of 200 births).

```
# 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")
```



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?

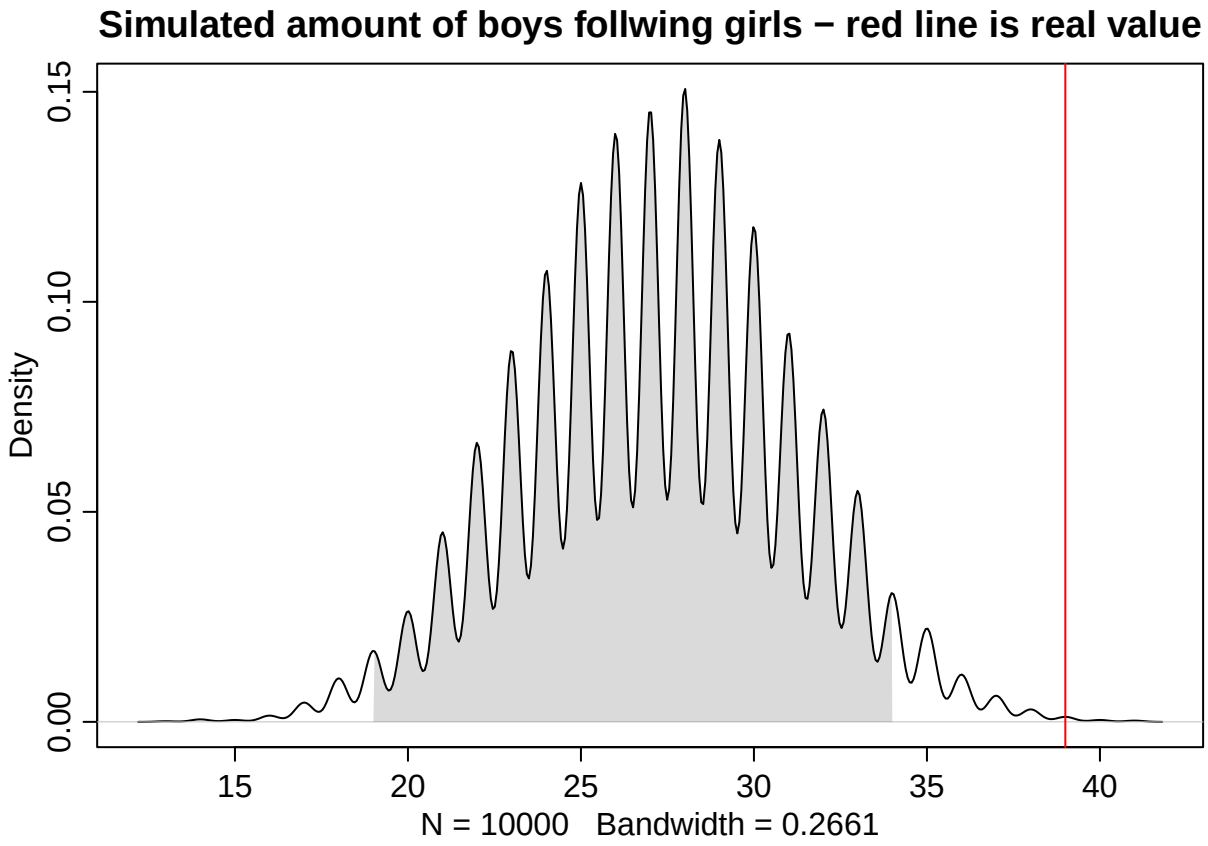
```
# Find the posterior probability of giving birth to a boy:
set.seed(1)
W <- sum(birth1)
N <- length(birth1)

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



Answer - The real value still lies withibn the moedls confodence interval, but is no longer close to the median of the distribution.

3H5. The model assumes that sex of first and second births are independent. To check this assumption, focus now on second births that followed female first borns. Compare 10,000 simulated counts of boys to only those second births that followed girls. To do this correctly, you need to 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)
W <- sum(birth2[birth1==0]) # counting number of boys following girls
N <- length(birth2[birth1==0])

# 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 follwing girls - red line is real value")
```



Answer The distribution of predicted amount if births gets smaller, as the we input fewer births as our value. We now see that the model is a really poor fit of the real value. The real value lies outside of the confidence interval.

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). *Answer* - Number of times people have watched the harry potter movies with leading office positions on income.

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. *Answer* - going to the gym and amount of food eaten on weight.

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 *Answer* - If the divorce rate is higher, more people can get married multiple times. We could therefor add a variable that measures the amount of re-marriages to evaluate this relationship.

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. *Answer* - Job type (predictor) may influence vacation (outcome), first that of you have a higher salary you are more able to go on vacaytion, ans second by some jobs being more flexible and therefore allowing more possibilities to go on vacation.

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?

```
# inspecting data to choose priors
data(foxes)

table(foxes$groupsize)
```

```
##
##  2  3  4  5  6  7  8
##  8 24 48 15  6  7  8
```

```
table(foxes$area)
```

```
##
## 1.09 1.29 1.73 1.88 1.91 2.05 2.12 2.21 2.24 2.45 2.59 2.65 2.75 2.89 3.02 3.04
##    2    2    3    3    3    2    2    3    3    4    4    4    7    5    4    4
## 3.13 3.16 3.35 3.43 3.54 3.56 3.66 3.77 3.78 3.84 3.92 4.54 5.07
##    4    4    4    4    4    4    5    3    3    7    5    6    8
```

```
# Making models

# Weight predicted by area
model_area <- quap(
  alist(
    weight ~ dnorm(mu,sigma),
    mu <- a + b*area,
    a~dnorm(2.5, 1),
    b~dnorm(0, 20),
    sigma~dexp(1)),
  data = foxes
)

# Weight predicted by groupsize
model_gz <- quap(
  alist(
    weight ~ dnorm(mu,sigma),
    mu <- a + b*groupsize,
```

```
    a~dnorm(4, 1.5),
    b~dnorm(0, 15),
    sigma~dexp(1)),
  data = foxes
)
```

```
## Caution, model may not have converged.
```

```
## Code 1: Maximum iterations reached.
```

```
precis(model_area)
```

```
##             mean         sd        5.5%     94.5%
## a     4.19384033 0.36419411  3.61178781 4.7758929
## b     0.09960696 0.11094472 -0.07770414 0.2769181
## sigma 1.17495151 0.07680728  1.05219864 1.2977044
```

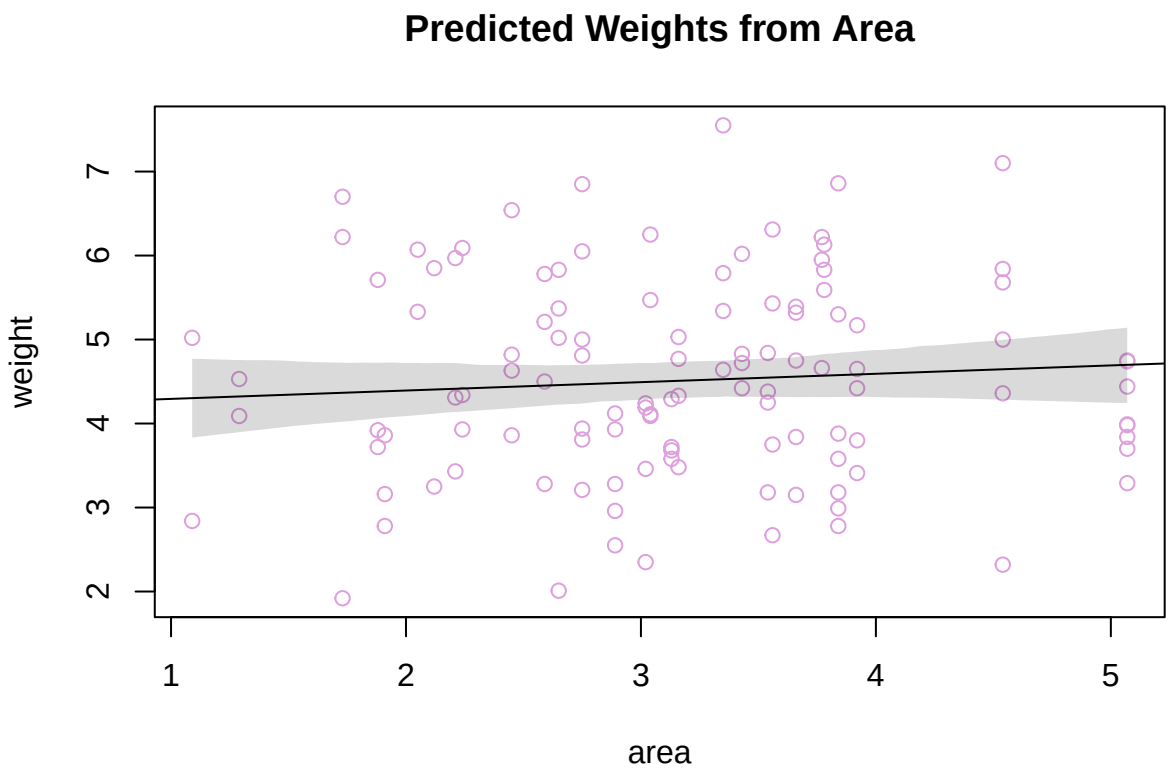
```
precis(model_gz)
```

```
##             mean         sd       5.5%     94.5%
## a      5.6187123 0.32851838  5.0936765  6.143748
## b     -0.2259301 0.07103025 -0.3394502 -0.112410
## sigma  1.1762981 0.07895032  1.0501202  1.302476
```

```
#Plotting
### AREA
area.seq <- seq(from = min(foxes$area), to = max(foxes$area), length.out = 1e4)
mu <- link(model_area, data = data.frame(area = area.seq))
mu.PI <-  apply(mu, 2, PI, prob = 0.95)

plot(weight ~ area, data = foxes, col = "plum", main = "Predicted Weights from Area")
abline(model_area)
```

```
## Warning in abline(model_area): only using the first two of 3 regression
## coefficients
```

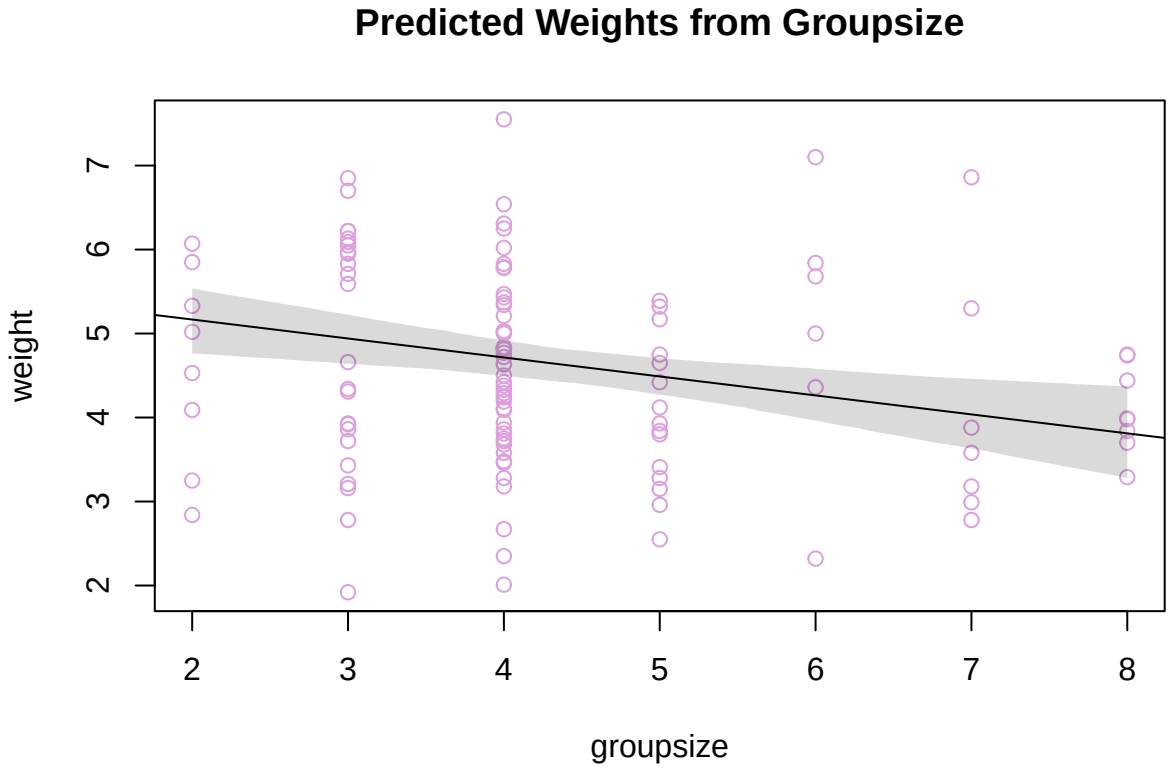
```
shade(mu.PI, area.seq)
```




```
### Groupsize
groupsize.seq <- seq(from = min(foxes$groupsize), to = max(foxes$groupsize), length.out = 1e4)
mu <- link(model_gz, data = data.frame(groupsize = groupsize.seq))
mu.PI <-  apply(mu, 2, PI, prob = 0.95)

plot(weight ~ groupsize, data = foxes, col = "plum", main = "Predicted Weights from Groupsize")
abline(model_gz)
```

```
## Warning in abline(model_gz): only using the first two of 3 regression
## coefficients
```

```
shade(mu.PI, groupsize.seq)
```



answer

- For area size the effect seem to be very small, and the plot also highlight that there does not seem to be a ralationship betwwen area and weight. The groupsize seem to have a negative relationship with weight, however taking the error in to considertion this relationship does not seem to be super robust.

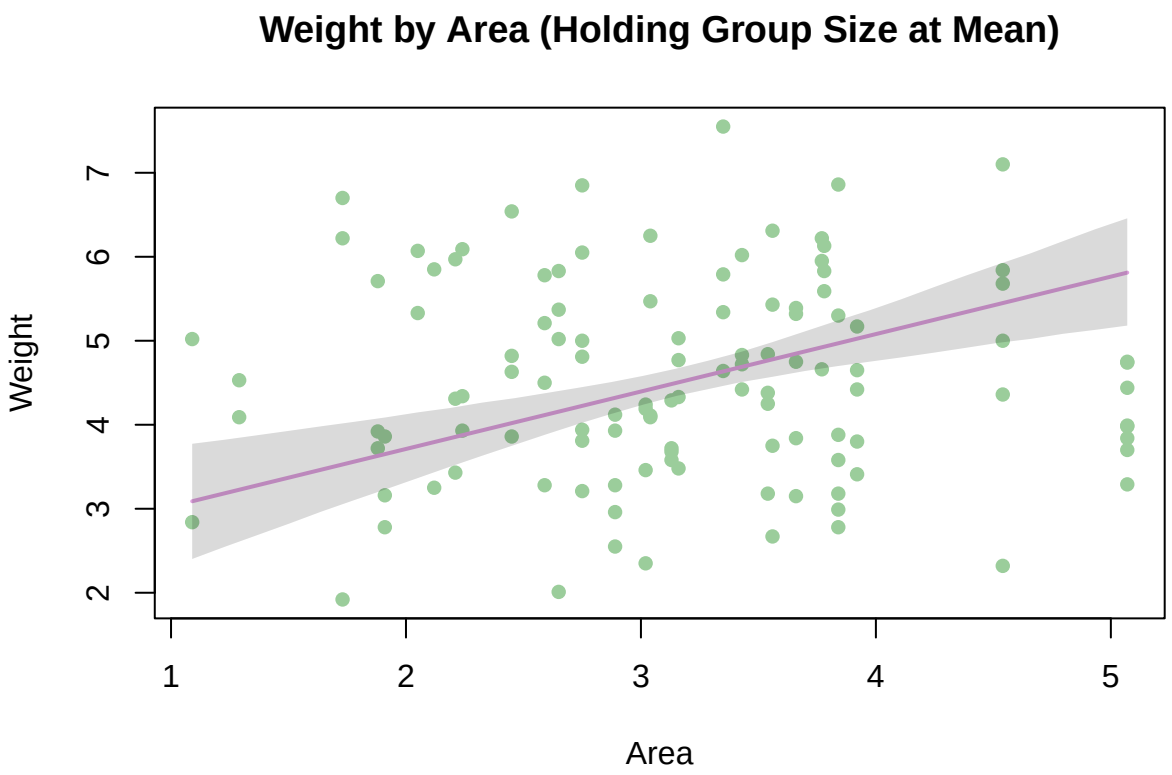
5H2. Now fit a multiple linear regression with weight as the outcome and both area and groupsize as predictor variables. Plot the predictions of the model for each predictor, holding the other predictor constant at its mean. What does this model say about the importance of each variable? Why do you get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

```
model_area_gz <- quap(
  alist(
    weight ~ dnorm(mu,sigma),
    mu <- a + b*area + b2*groupsize,
    a~dnorm(2.5, 1),
    b~dnorm(0, 20),
    b2~dnorm(0, 15),
    sigma~dexp(1)),
  data = foxes
)

precis(model_area_gz)
```

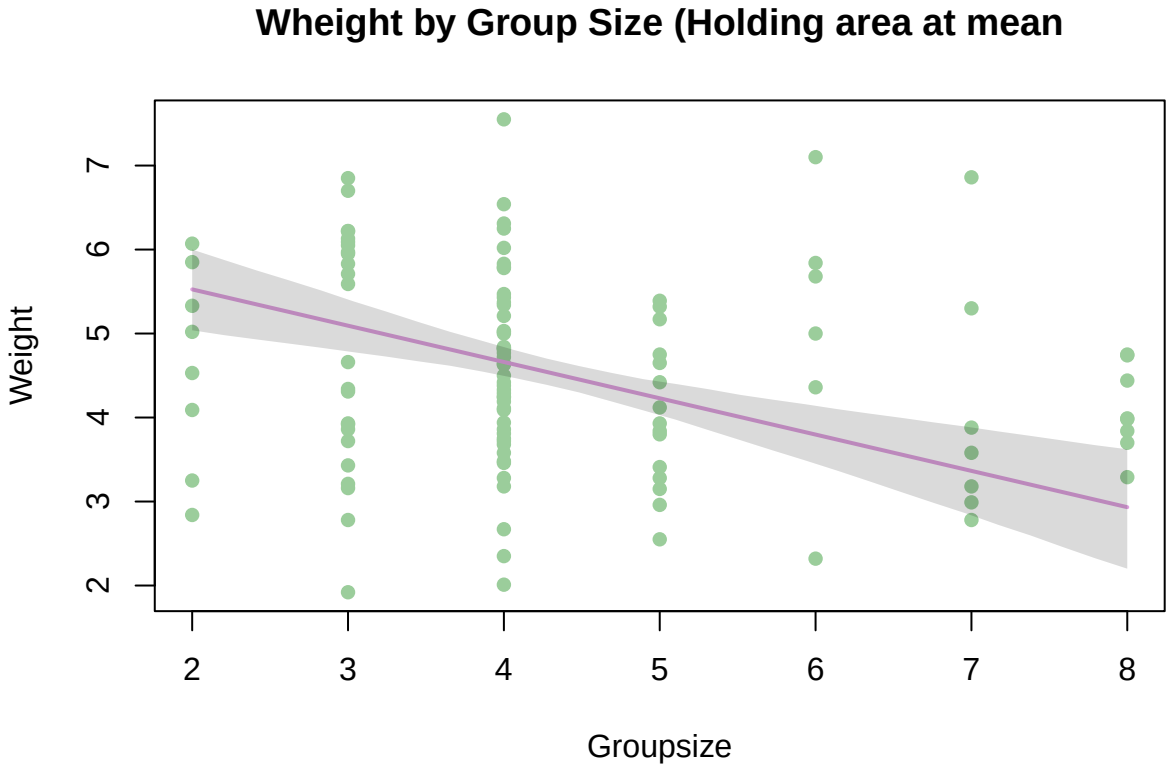
```
##             mean         sd       5.5%      94.5%
## a      4.2157070 0.34783600  3.6597979  4.7716161
## b      0.6863116 0.19615178  0.3728231  0.9998000
## b2    -0.4325732 0.12039066 -0.6249807 -0.2401657
## sigma  1.1150303 0.07290257  0.9985179  1.2315427
```

```
# holding group size at mean
xseq <- seq(from = min(foxes$area), to=max(foxes$area), length.out=30)
mu <- link(model_area_gz, data = data.frame( area = xseq, groupsize = mean(foxes$groupsize)))
```

```r
#summarize samples across cases
mu_mean <- apply(mu, 2, mean)
mu_PI <- apply(mu, 2, PI)

#plot
plot(foxes$area, foxes$weight, col = "darkseagreen3", pch = 16,
     xlab = "Area", ylab = "Weight",
     main = "Weight by Area (Holding Group Size at Mean)")
lines(xseq, mu_mean, lwd=2, col = "plum")
shade(mu_PI, xseq)
```




```r
# holding area at mean
xseq <- seq(from = min(foxes$groupsize), to=max(foxes$groupsize), length.out=30)
mu <- link(model_area_gz, data = data.frame( groupsize = xseq, area = mean(foxes$area)))

#summarize samples across cases
mu_mean <- apply(mu, 2, mean)
mu_PI <- apply(mu, 2, PI)

#plot
plot(foxes$groupsize, foxes$weight, col = "darkseagreen3", pch = 16,
     xlab = "Groupsize", ylab = "Weight",
     main = "Wheight by Group Size (Holding area at mean")
lines(xseq, mu_mean, lwd=2, col = "plum")
shade(mu_PI, xseq)
```



answer

- we now observe much stronger relationships for both variables. We observe stronger effects because area has a positive relationship with weight and group size has a negative relationship, leading them to cancel each other out when constructing separate models. Including both variables in the same model allows us to model this, and keep both effects.

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?

```
table(foxes$avgfood)
```

```
##
## 0.37 0.41 0.42 0.45 0.49 0.51 0.53 0.57  0.6 0.65 0.66 0.67 0.68 0.69 0.71 0.72
##    2    3    3    2    2    3    2    3    4    3    4    4    7    4    4    4
## 0.73 0.74 0.77 0.78 0.79  0.8 0.91 0.98 1.03 1.21
##    4    6    4   13    5    4    6    7    5    8
```

```
# AvgFood and Groups size
model_food_gz <- quap(
  alist(
    weight ~ dnorm(mu,sigma),
    mu <- a + b*avgfood + b2*groupsize,
    a~dnorm(0.72, 0.5),
    b~dnorm(0, 20),
    b2~dnorm(0, 15),
    sigma~dexp(1)),
  data = foxes
```

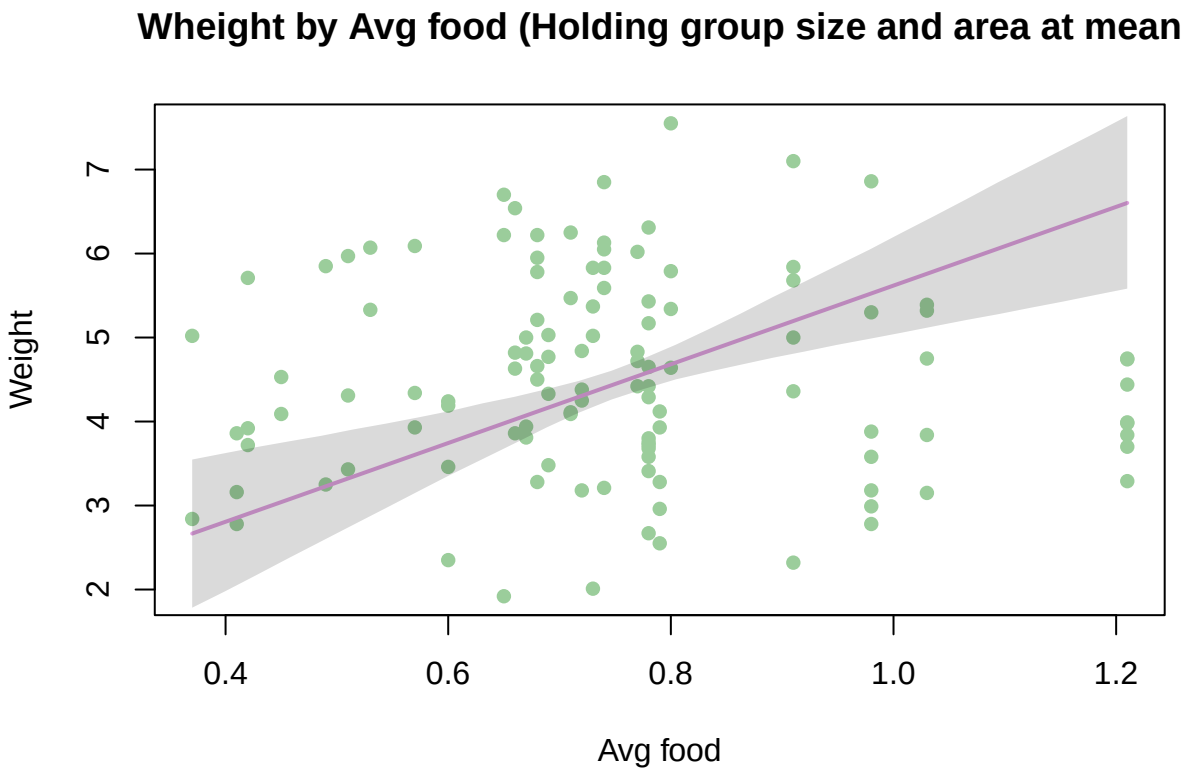
```
)

# AvgFood, Groups size and Area
model_food_gz_area <- quap(
  alist(
    weight ~ dnorm(mu,sigma),
    mu <- a + b*avgfood + b2*groupsize + b3*area,
    a~dnorm(0.72, 0.5),
    b~dnorm(0, 20),
    b2~dnorm(0, 15),
    b3~dnorm(2.5, 1),
    sigma~dexp(1)),
  data = foxes
)

precis(model_food_gz)
```

```
##             mean         sd      5.5%      94.5%
## a      2.5937105 0.35483304  2.026619  3.1608023
## b      6.7857179 1.14479630  4.956112  8.6153235
## b2    -0.7487989 0.16003344 -1.004563 -0.4930345
## sigma  1.1694247 0.08061123  1.040592  1.2982570
```

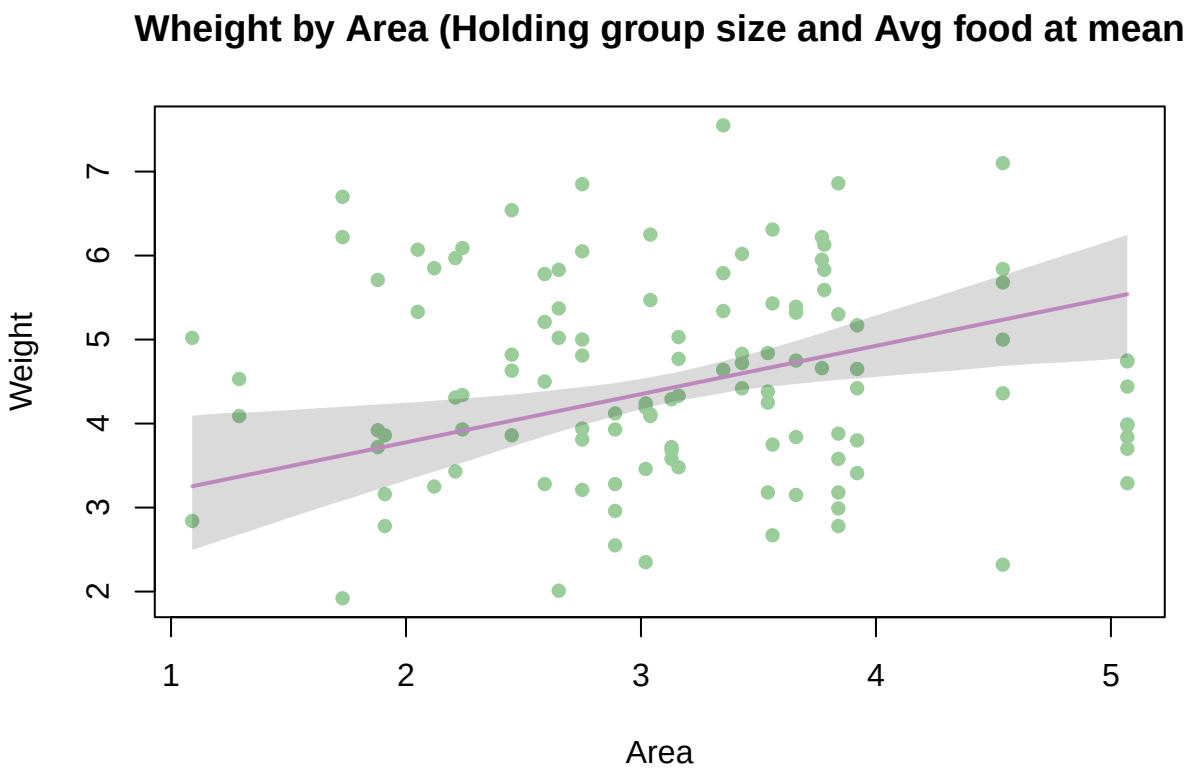
```
precis(model_food_gz_area)
```

```
##             mean         sd       5.5%      94.5%
## a      2.5614491 0.35316052  1.9970304  3.1258678
## b      4.6947730 1.41355334  2.4356417  6.9539042
## b2    -0.8015416 0.16016528 -1.0575166 -0.5455665
## b3     0.5795411 0.24346116  0.1904431  0.9686391
## sigma  1.1566680 0.07983648  1.0290738  1.2842621
```

```
#plot
# Define sequences for area and avgfood
xseq_area <- seq(from = min(foxes$area), to = max(foxes$area), length.out = 30)
xseq_food <- seq(from = min(foxes$avgfood), to = max(foxes$avgfood), length.out = 30)

# Predictions for area (holding avgfood at mean)
mu_area <- link(model_food_gz_area, data = data.frame(area = xseq_area, avgfood = mean(foxes$avgfood), g
mu_mean_area <- apply(mu_area, 2, mean)
mu_PI_area <- apply(mu_area, 2, PI)

# Predictions for avgfood (holding area at mean)
mu_food <- link(model_food_gz_area, data = data.frame(avgfood = xseq_food, area = mean(foxes$area), grou
mu_mean_food <- apply(mu_food, 2, mean)
mu_PI_food <- apply(mu_food, 2, PI)

# plotting
plot(foxes$avgfood, foxes$weight, col = "darkseagreen3", pch = 16,
     xlab = "Avg food", ylab = "Weight",
     main = "Wheight by Avg food (Holding group size and area at mean")
lines(xseq_food, mu_mean_food, lwd=2, col = "plum")
shade(mu_PI_food, xseq_food)
```

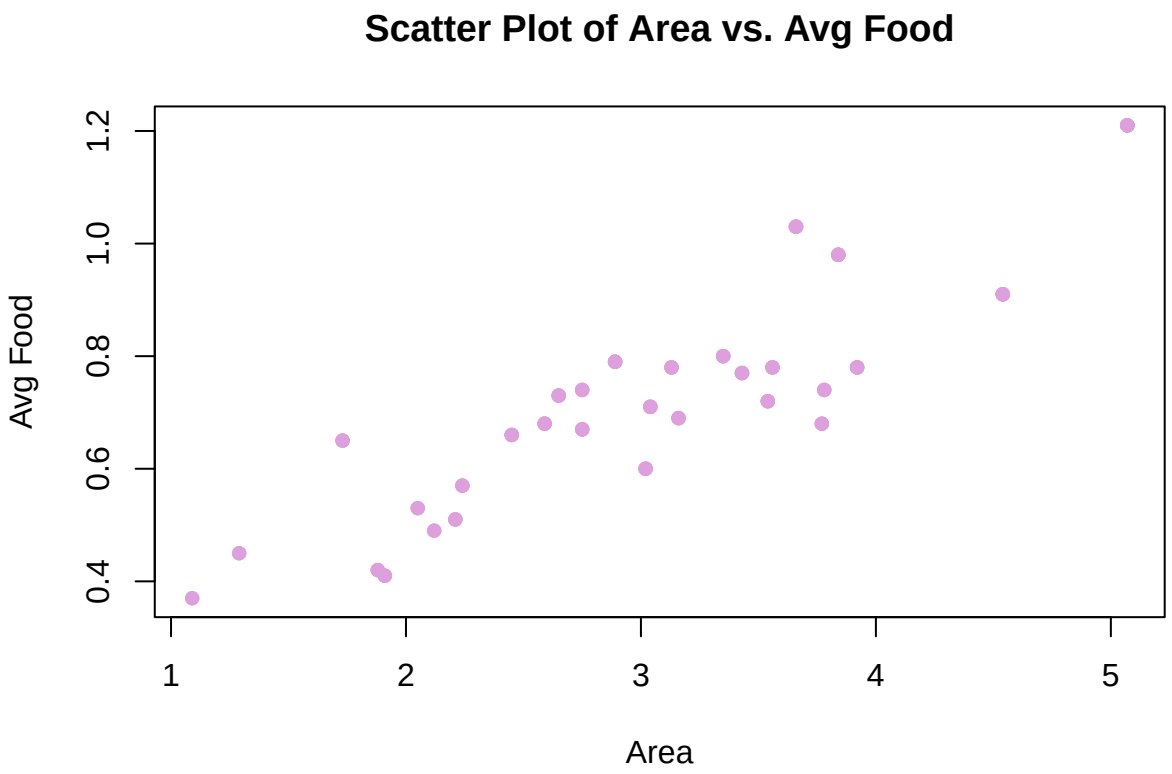



```
# plotting
plot(foxes$area, foxes$weight, col = "darkseagreen3", pch = 16,
     xlab = "Area", ylab = "Weight",
     main = "Wheight by Area (Holding group size and Avg food at mean")
lines(xseq_area, mu_mean_area, lwd=2, col = "plum")
shade(mu_PI_area, xseq_area)
```



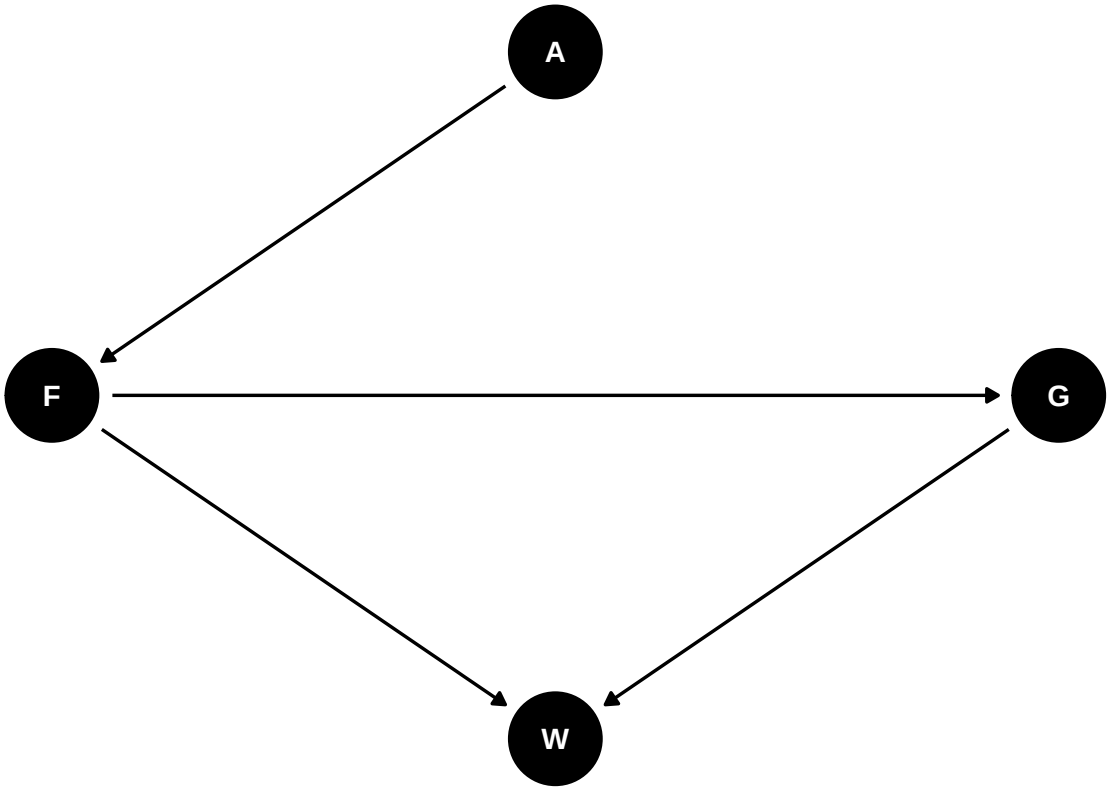
answer - avg food has a much stronger positive relationship with body weight(b=4.60 in together model and 6.79 in alone) than area (b = 0.58 in together model, and 0.69 in alone), and is therefore a better predictor, and the one i would choose. This is also illustrated visually above. - The total effect being reduced might indicate that area and average food might be correlated - this is indeed the case illustrated in the plot below; there is a positive linear relationship between area size and Avg Food. This is a case of multicollinearity, which usually leads to high standard errors and causes the estimation of effects to be difficult.

```
plot(foxes$area, foxes$avgfood,
     col = "plum", pch = 16,
     xlab = "Area", ylab = "Avg Food",
     main = "Scatter Plot of Area vs. Avg Food")
```



Defining our theory with explicit DAGs 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()
```



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_area_food <- quap(
  alist(
    avgfood ~ dnorm(mu,sigma),
    mu <- a + b*area,
    a~dnorm(0.75, 0.2),
    b~dnorm(0, 5),
    sigma~dexp(1)),
  data = foxes
)

precis(model_area_food)
```

```
##             mean          sd       5.5%     94.5%
## a     0.16762529 0.030431813 0.11898938 0.2162612
## b     0.18464897 0.009225725 0.16990448 0.1993935
## sigma 0.09269516 0.006090030 0.08296212 0.1024282
```

answer increasing the area by 1 would result in an icrease of 0.19 on food

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?

```
model_food_total <- quap(
  alist(
    weight ~ dnorm(mu,sigma),
```

```
    mu <- a + b_f*avgfood,
    a~dnorm(4.5, 1.2),
    b_f~dnorm(0, 5),
    sigma~dexp(1)),
  data = foxes
)

precis(model_food_total)
```

```
##             mean        sd       5.5%     94.5%
## a      4.6249327 0.4015730  3.9831415 5.2667238
## b_f   -0.1281109 0.5185934 -0.9569233 0.7007016
## sigma  1.1726346 0.0764108  1.0505154 1.2947539
```

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?

```
#Stratify by G
model_food_direct <- quap(
  alist(
    weight ~ dnorm(mu,sigma),
    mu <- a + b_f*avgfood + b_gz*groupsize,
    a~dnorm(4.5, 1.2),
    b_f~dnorm(0, 5),
    b_gz~dnorm(0, 5),
    sigma~dexp(1)),
  data = foxes
)

precis(model_food_direct)
```

```
##             mean         sd       5.5%      94.5%
## a      4.2116798 0.40039922  3.5717645  4.8515950
## b_f    3.5452298 1.14306801  1.7183863  5.3720732
## b_gz  -0.5397098 0.15130659 -0.7815270 -0.2978927
## sigma  1.1115579 0.07248876  0.9957068  1.2274089
```

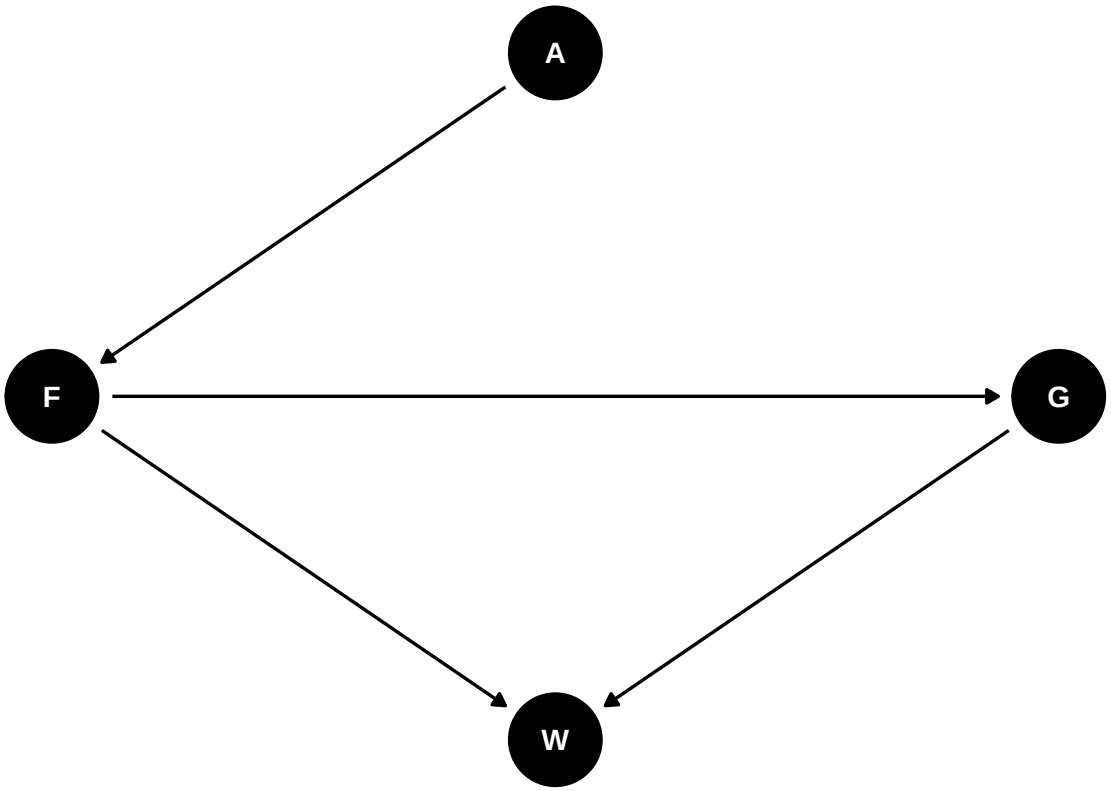
answer - it looks like there is a causal effect of avg food on weight (is sensible), but when taking the totalk effect it is hidden by the negatuve effect of groupsize on wheight.

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.

```
# Creating the causal graph
data(WaffleDivorce)

# dag code
pacman::p_load(dagitty,
               ggdag)
dag_WH <- dagitty('dag {
S[pos="1.0,1"]
A[pos="2,0"]
W[pos="0.0,0.0"]
M[pos="1.0,0.0"]
```

```
D[pos="1.0,-1.0"]
W -> D
A -> D
S -> M
M -> D
A -> M
S -> W
S -> A

}')

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



W = waffle-House, D = Divorce rate, A = Age of marriage, S = south, M = marriage rate

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

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

```
#total effect of Wafflhouse on divorce
model_divorce_totalt <- quap(
  alist(
    Divorce ~ dnorm(mu,sigma),
    mu <- a + b_W*WaffleHouses,
    a~dnorm(9.7, 1.8),
    b_W~dnorm(0, 1),
    sigma~dexp(1)),
```

```
  data = WaffleDivorce
)

# Direct effect of waffflehouse on divorce
model_divorce_direct <- quap(
  alist(
    Divorce ~ dnorm(mu,sigma),
    mu <- a + b_W*WaffleHouses + b_S*South,
    a~dnorm(9.7, 1.8),
    b_W~dnorm(0, 1),
    b_S~dnorm(0, 1),
    sigma~dexp(1)),
  data = WaffleDivorce
)

precis(model_divorce_totalt)
```

```
##              mean          sd       5.5%      94.5%
## a     9.465534851 0.267635974 9.03780087 9.89326883
## b_W   0.007010427 0.003713761 0.00107512 0.01294573
## sigma 1.714048588 0.167161912 1.44689157 1.98120561
```

```
precis(model_divorce_direct)
```

```
##              mean          sd         5.5%      94.5%
## a     9.360177068 0.270984450  8.927091579 9.79326256
## b_W   0.003000311 0.004569984 -0.004303406 0.01030403
## b_S   0.845168140 0.588251108 -0.094970745 1.78530702
## sigma 1.669864770 0.163746405  1.408166388 1.93156315
```

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.

```
# Getting the conditional independencies
impliedConditionalIndependencies(dag_WH)
```

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

```
mean(WaffleDivorce$Marriage)
```

```
## [1] 20.114
```

```
M_con1 <- quap(
  alist(
    MedianAgeMarriage ~ dnorm(mu,sigma),
    mu <- a + b_W*WaffleHouses + b_S*South,
    a~dnorm(26, 1.3),
    b_W~dnorm(0, 5),
    b_S~dnorm(0, 5),
    sigma~dexp(1)),
  data = WaffleDivorce
)

M_con2 <- quap(
```

```
  alist(
    Divorce ~ dnorm(mu,sigma),
    mu <- a + b_S*South + b_A*MedianAgeMarriage + b_M*Marriage + b_W*WaffleHouses,
    a~dnorm(9.7, 1.8),
    b_S~dnorm(0, 5),
    b_A~dnorm(0, 5),
    b_M~dnorm(0, 5),
    b_W~dnorm(0, 5),
    sigma~dexp(1)),
  data = WaffleDivorce
)

precis(M_con1)
```

```
##               mean          sd         5.5%        94.5%
## a     26.229776842 0.194263538 25.919306187 26.540247496
## b_W    0.001821363 0.003500601 -0.003773272  0.007415999
## b_S   -0.851566047 0.505821203 -1.659966023 -0.043166070
## sigma  1.175682961 0.115550874  0.991010348  1.360355575
```

```
precis(M_con2)
```

```
##               mean          sd         5.5%        94.5%
## a     10.840681248 1.769128161  8.013272757 13.668089740
## b_S    0.996991053 0.643951435 -0.032167713  2.026149818
## b_A   -0.173532157 0.067227899 -0.280975324 -0.066088990
## b_M    0.149878686 0.049686819  0.070469552  0.229287819
## b_W    0.001823894 0.004436956 -0.005267219  0.008915007
## sigma  1.490204651 0.146885092  1.255453906  1.724955397
```

answer - The tests for the implied conditional independencies revealed that the median age if marriage and marriage rate is independent of waffelhouse, when conditioning on South. - However the when testing if Independence of south in divorce rate when conditioning in Marriage rate, age and waffelhouse, it revaled a small positive relationship. This would indicate that there are still other factors related to a state being in the south or not that effects Divorce rate. This could for example be religion, state laws, political view and so on.