Homework 3

Emily Maloney

February 4, 2019

Chapter 5 Homework

```
library(rethinking)
library(tidyverse)
library(haven)
library(corrr)
library(brms)
```

Easy Problems

5E1

The linear models that specify a multiple regression are:
2) $\mu\ i = \beta\ x * x\ i + \beta\ z * z\ i$
4) $\mu\ i = \alpha + \beta\ x * x\ i + \beta\ z * z\ i$

5E2

The multiple regression to evaluate the claim that *animal diversity is linearly related to altitude, but only after controlling for plant diversity* is:
Animal Diversity = $\alpha + \beta\ l$(latitude) + $\beta\ p$(plant diversity)

5E3

The multiple regression to evaluate the claim that *neither amount of funding nor size of laboratory is by itself a good predictor of time to PhD degree, but together those variables are both positvely associated with time to degree is*:
Time to PhD = $\alpha + \beta\ f$(amount of funding) + $\beta\ l$(size of lab)
Both $\beta\ f$ and $\beta\ l$ are positive.

5E4

The models that are inferentially equivalent are:
1) $\mu\ i = \alpha + \beta\ A * A\ i + \beta\ B * B\ i + \beta\ D * D\ i$
3) $\mu\ i = \alpha + \beta\ B * B\ i + \beta\ C * C\ i + \beta\ D * D\ i$
4) $\mu\ i = \alpha\ A * A\ i + \alpha\ B * B\ i + \alpha\ C * C\ i + \alpha\ D * D\ i$
5) $\mu\ i = \alpha\ A(1 - B\ i - C\ i - D\ i) + \alpha\ B * B\ i + \alpha\ C * C\ i + \alpha\ D * D\ i$

Medium Problems

5M1

My own example of a spurious relationship is participation in extracurricular activities (eca) and math test scores (math). When SES (ses) is included in the model, the effect of extracurricular activities disappears, showing how SES predicts both involvement in extracurriculars and higher test scores.

```r
#set number of cases
n <- 1e4

#SES
ses <- rnorm(n, 2, 1)
eca <- rnorm(n, ses)
math <- rnorm(n, ses/4)
d <- data.frame(math, ses, eca)

#with only eca predictor
hw.spur.1 <- rethinking::map(
        alist(
                math ~ dnorm(mu, sigma),
                mu <- a + b*eca,
                a ~ dnorm(0, 10),
                b ~ dnorm(0, 1),
                sigma ~ dunif(0, 10)
        ),
        data = d)

#with only ses predictor
hw.spur.2 <- rethinking::map(
        alist(
                math ~ dnorm(mu, sigma),
                mu <- a + b*ses,
                a ~ dnorm(0, 10),
                b ~ dnorm(0, 1),
                sigma ~ dunif(0, 10)
        ),
        data = d)

#with both ses and math
hw.spur.3 <- rethinking::map(
        alist(
                math ~ dnorm(mu, sigma),
                mu <- a + be*eca + bs*ses,
                a ~ dnorm(0, 10),
                be ~ dnorm(0, 1),
                bs ~ dnorm(0, 1),
                sigma ~ dunif(0, 10)
        ),
        data = d)

precis(hw.spur.1)
```

```
##       Mean StdDev 5.5% 94.5%
## a     0.25   0.02 0.22  0.28
## b     0.12   0.01 0.11  0.13
## sigma 1.02   0.01 1.01  1.03
```

```r
precis(hw.spur.2)
```

```
##        Mean StdDev  5.5% 94.5%
## a     -0.02   0.02 -0.06  0.01
```

```
## b      0.26   0.01  0.24  0.28
## sigma  1.00   0.01  0.99  1.01
```

```
precis(hw.spur.3)
```

```
##        Mean StdDev  5.5% 94.5%
## a     -0.02   0.02 -0.06  0.01
## be    -0.01   0.01 -0.03  0.00
## bs     0.27   0.01  0.25  0.30
## sigma  1.00   0.01  0.99  1.01
```

Accordingly, model 1 results show that the beta attached to extracurricular involvement is 0.12, with a 89% credible interval of 0.11-0.13, meaning that for each one unit increase in extracurricular activity, the model predicts a 0.12 point increase in test score. This is equivalent to a 12 percentage point increase in test score as the outcome variable of test score has been scaled from 0 to 1.

The second model similarly shows that the beta attached to socioeconomic status is 0.24, with an 89% credbile interval of 0.23-0.26, meaning that for each one unit increase in socioeconomic status, the model predicts a 0.24 (24 percentage) point increase in test score. This is equivalent to a 24 percentage point increase in test score.

But when socioeconomic status and extracurricular activity are included in the same model, now the beta for extracurricular involvement is 0.00, with an 89% credible interval of -0.01-0.02, and the beta for socioeconomic status is still 0.24, with an 89% credible interval of 0.22-0.26. This shows that socioeconomic status predicts both extracurricular involvement and test score, so there is no causal relationship between participating in extracurricular activities and performing better on math tests.

5M2

An example of a masked relationship would be the relationship between cheating, academic achievement, and anxiety. While there is a negative association between academic achievement and cheating, there is a positive association between anxiety and cheating behavior. Similarly, academic achievement and anxiety are themselves positively correlated with each other. This means that when the variables are not included in the same model, then their effect is masked, but once they are both included in the same model, then the model will predict a greater effect for each.

```
n <- 1e4
rho <- 0.7 #correlation between xpos and xneg
anxiety <- rnorm(n)
achievement <- rnorm(n, rho*anxiety, sqrt(1-rho^2))
cheat <- rnorm(n, anxiety - achievement)
d <- data.frame(cheat, anxiety, achievement)

#with only xpos predictor
hw.mask.1 <- rethinking::map(
        alist(
                cheat ~ dnorm(mu, sigma),
                mu <- a + bax*anxiety,
                a ~ dnorm(0, 10),
                bax ~ dnorm(0, 1),
                sigma ~ dunif(0, 10)
        ),
        data = d)

#with only xneg predictor
hw.mask.2 <- rethinking::map(
```

```
        alist(
                cheat ~ dnorm(mu, sigma),
                mu <- a + bach*achievement,
                a ~ dnorm(0, 10),
                bach ~ dnorm(0, 1),
                sigma ~ dunif(0, 10)
        ),
        data = d)

#with both xpos and xneg
hw.mask.3 <- rethinking::map(
        alist(
                cheat ~ dnorm(mu, sigma),
                mu <- a + bax*anxiety + bach*achievement,
                a ~ dnorm(0, 10),
                bax ~ dnorm(0, 1),
                bach ~ dnorm(0, 1),
                sigma ~ dunif(0, 10)
        ),
        data = d)

precis(hw.mask.1)
```

```
##       Mean StdDev  5.5% 94.5%
## a     0.00   0.01 -0.02  0.02
## bax   0.32   0.01  0.30  0.34
## sigma 1.24   0.01  1.23  1.26
```

```
precis(hw.mask.2)
```

```
##        Mean StdDev  5.5% 94.5%
## a      0.00   0.01 -0.02  0.02
## bach  -0.30   0.01 -0.32 -0.28
## sigma  1.24   0.01  1.23  1.26
```

```
precis(hw.mask.3)
```

```
##        Mean StdDev  5.5% 94.5%
## a      0.00   0.01 -0.01  0.02
## bax    1.02   0.01  1.00  1.05
## bach  -1.02   0.01 -1.04 -0.99
## sigma  1.01   0.01  1.00  1.02
```

In the first model with only anxiety predicting cheating behavior, the model specifies that for every one unit increase in anxiety, there is a 0.3 increase in cheating behavior, with an 89% credible interval of 0.28-0.32.

Similarly, in the second model with only achievement predicting cheating behavior, the model specifies that for every one unit increase in academic achievement, there will be a 0.28 *decrease* in cheating behavior, with an 89% credible interval of -0.30 to -0.26.

And when they are both included in the same model, both of the beta specifications increase in magnitude, with a one unit increase in anxiety now corresponding to a 0.98 increase in cheating behavior and an 89% credible interval from 0.96-1.00, and a one unit increase in academic achievement corresponding to a -0.98 decrease in cheating behavior, with an 89% credible interval of -1 to -0.96. This is an example of a masked relationship, because when both are included in the same model, the specified effects for both increase in magnitude.

5M3

If there is a high divorce rate then there are more older adults who can get remarried, increasing the marriage rate. To evaulate this relationship, one could first run a regression with divorce rate predicting marriage rate, and then a second regression including both divorce rate and a variable containing the rate of marriages that only includes post-first marriages. If the beta for divorce rate decreases to zero, this would suggest that divorce rates *cause* marriage rates through the amount of people who get remarried.

5M4

```
data("WaffleDivorce")
d <- WaffleDivorce
lds <- read.csv("lds.csv")
d <- left_join(d, lds)
```

```
## Joining, by = "Location"
```

```
## Warning: Column `Location` joining factors with different levels, coercing
## to character vector
```

```
d <- d %>% mutate(medagestd =
                    (MedianAgeMarriage - mean(MedianAgeMarriage, na.rm = T))/sd(MedianAgeMarriage),
                  marriagestd =
                    (Marriage - mean(Marriage, na.rm = T))/sd(Marriage),
                  prop_lds = Raw.1/Raw.2,
                  prop_lds_std = (prop_lds- mean(prop_lds))/sd(prop_lds))

hw.lds.1 <- rethinking::map(
          alist(
            Divorce ~ dnorm(mu, sigma),
            mu <- a + bA*medagestd + bR*marriagestd + bL*prop_lds_std,
            a ~ dnorm(10, 10),
            bA ~ dnorm(0, 1),
            bR ~ dnorm(0, 1),
            bL ~ dnorm(0, 1),
            sigma ~ dunif(0, 10)
          ),
          data = d)

precis(hw.lds.1)
```

```
##        Mean StdDev  5.5% 94.5%
## a      9.69   0.19  9.38  9.99
## bA    -1.27   0.27 -1.70 -0.85
## bR     0.06   0.27 -0.38  0.49
## bL    -0.58   0.22 -0.93 -0.23
## sigma  1.34   0.13  1.13  1.56
```

The model's prediction for the beta associated with the standardized percent LDS population within a state is -0.58, with an 89% credible interval of -0.93 to 0.23, meaning that as a state's percent LDS population increases by one standard deviation, the divorce rate decreases by 0.58. Now the marriage rate variable's 89% credible interval includes 0, spanning from -0.38 to 0.48, with an estimate of 0.06. However, the beta specified for the standardized median age at marriage is still negative, with an estimate of -1.27 and an 89% credible interval of -1.70 to -0.85. This means that with an increase of one standard deviation in median age at marriage, the mdoel predicts that a state should have a 1.27 decrease in divorce rate.

5M5

The two possible mechanisms suggested by which price of gasoline may be negatively associated with obesity rates are: 1. less driving leads to more exercise and 2. less driving leads to less eating out leads to less consumption of large restaurant purchases. To evalute these mechanisms, I would specify a series of regression models.

First, I would run a model predicting obesity rate by gasoline price, to see what the initial relationship looks like, to check that the data show that there is a relationship. Next, I would include a variable with information about how much time one spends driving to the model along with the gasoline price to predict obesity rates. If the beta for gasoline prices decreases in magnitude closer to 0, the first part of the two mechanisms is supported by the model.

Following that, I would specify two more models: one in which obesity rate is predicted by gasoline prices, amount of time spent driving per day, and amount of time spent exercising per day, and second, obesity rate predicted by gasoline prices, amount of time spent driving per day, and number of days a week the respondent eats out. If the beta for driving in both of these models decreases from the previous, then the first hypothesis is supported and the second is partially supported. Next, I would add average calorie count eaten per day to the last model specified, such that it is now: obesity rates predicted by gas prices, amount of time spent per day driving, number of days a week the respondent eats out, and the average calorie count per day. If the beta for number of days a week the respondent eats out decreases, then the second hypothesis is also supported.

Finally, I would include all of the predictors into one large model, to see if either explanation holds more weight than the other in predicting obesity rates. However, I may have to choose carefully between predictors for each mechanism, because multicollinearity may become an issue in a model with so many variables that are related to each other.

Hard Problems

5H1

```
data(foxes)
f <- foxes

hw.fox.1 <- rethinking::map(
        alist(
                weight ~ dnorm(mu, sigma),
                mu <- a + b*area,
                a ~ dnorm(0, 10),
                b ~ dnorm(0, 1),
                sigma ~ dunif(0, 10)
        ),
        data = f)

hw.fox.2 <- rethinking::map(
        alist(
                weight ~ dnorm(mu, sigma),
                mu <- a + b*groupsize,
                a ~ dnorm(0, 10),
                b ~ dnorm(0, 1),
                sigma ~ dunif(0, 10)
        ),
        data = f)

print("Body weight predicted by territory size results:")
```

```
## [1] "Body weight predicted by territory size results:"
```

```
precis(hw.fox.1)
```

```
##       Mean StdDev  5.5% 94.5%
## a     4.45   0.39  3.83  5.07
## b     0.03   0.12 -0.16  0.21
## sigma 1.18   0.08  1.06  1.30
```

```
print("Body weight predicted by group size results:")
```

```
## [1] "Body weight predicted by group size results:"
```

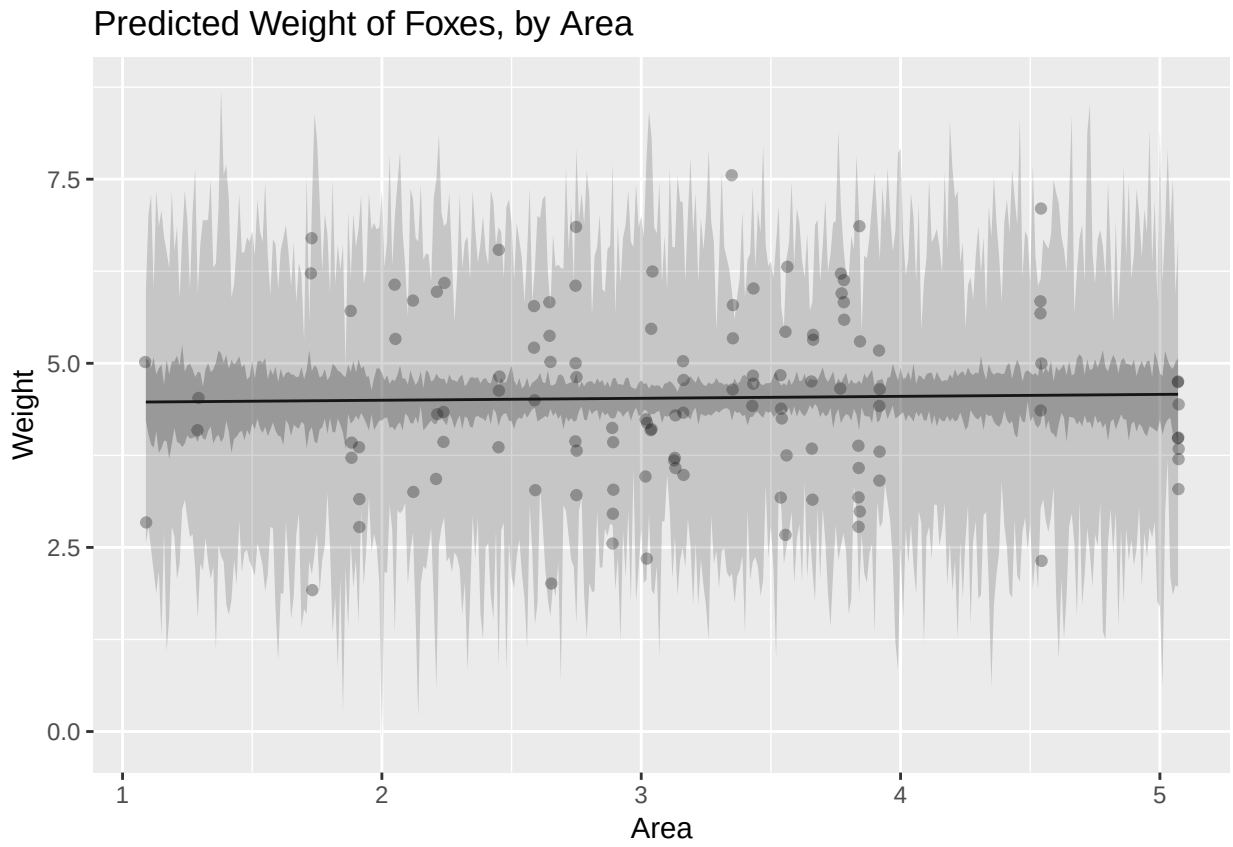
```
precis(hw.fox.2)
```

```
##        Mean StdDev  5.5% 94.5%
## a      5.06   0.32  4.54  5.58
## b     -0.12   0.07 -0.23 -0.01
## sigma  1.16   0.08  1.04  1.29
```

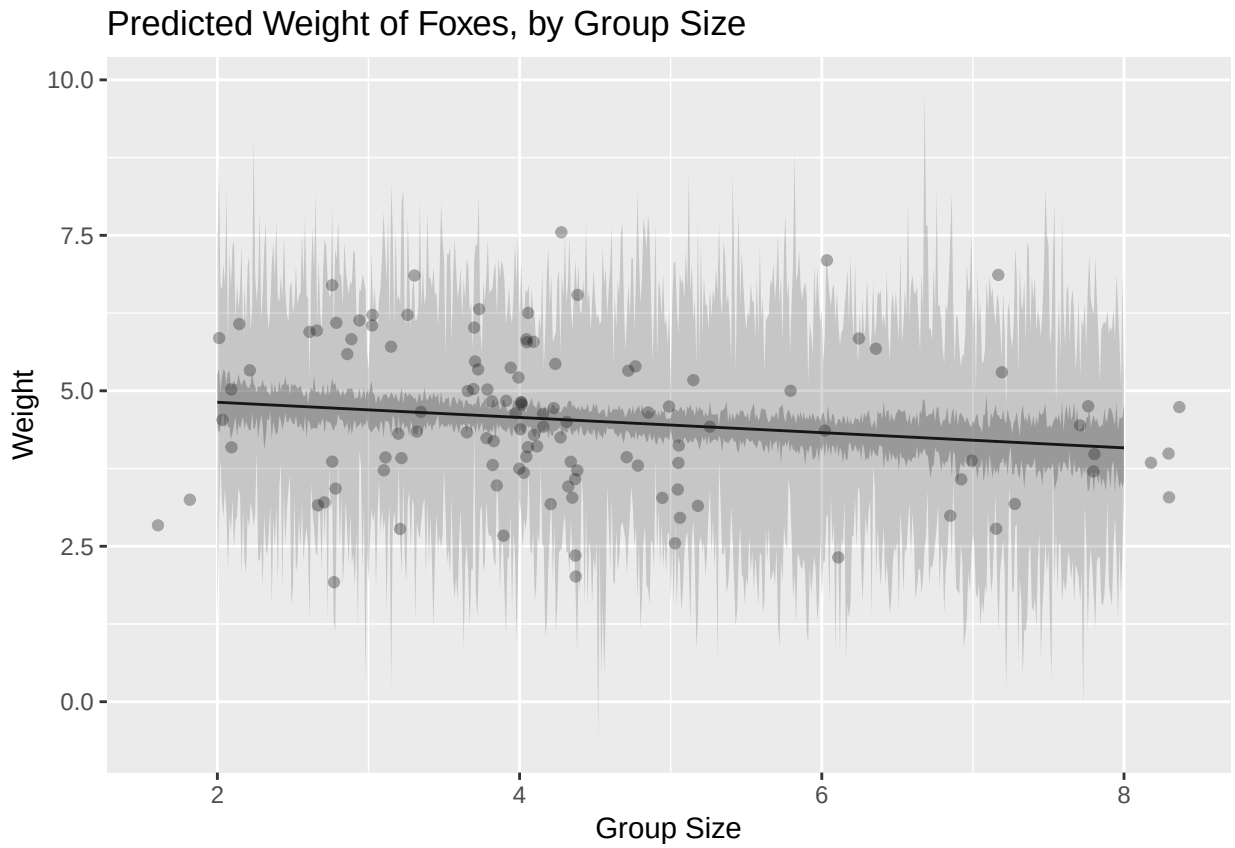
```
N <- 1e4

# Get predictive means and data
preds <-
  as.tibble(MASS::mvrnorm(mu = hw.fox.1@coef,
                          Sigma = hw.fox.1@vcov , n = N )) %>%      # rather than extract.samples
  mutate(area = sample(seq(from = 1.09, to = 5.07, by = 0.01), N, replace = T),
         predmean = a + b * area ,                                  # line uncertainty
         predverb = rnorm(N, a + b*area, sigma )) %>%               # data uncertainty
  group_by(area) %>%
  mutate(lb_mu = rethinking::HPDI(predmean, prob = .95)[1],
         ub_mu = rethinking::HPDI(predmean, prob = .95)[2],
         lb_wt = rethinking::HPDI(predverb, prob = .95)[1],
         ub_wt = rethinking::HPDI(predverb, prob = .95)[2]) %>%
  slice(1) %>%
  mutate(yhat = hw.fox.1@coef["a"] + hw.fox.1@coef["b"] * area) %>%       # yhat for reg line
  select(area, yhat, lb_mu, ub_mu, lb_wt, ub_wt)
```

```
## Warning: `as.tibble()` is deprecated, use `as_tibble()` (but mind the new semantics).
## This warning is displayed once per session.
```

```
#plot
ggplot(f, aes(x = area)) +
  geom_jitter(aes(y = weight), alpha = .3) +
  geom_line(data = preds, aes(y = yhat)) +
  geom_ribbon(data = preds, aes(ymin = lb_mu, ymax = ub_mu), alpha = .3) +
  geom_ribbon(data = preds, aes(ymin = lb_wt, ymax = ub_wt), alpha = .2) +
  labs(x = "Area",
       y = "Weight",
       title = "Predicted Weight of Foxes, by Area")
```




```
# Get predictive means and data
preds <-
  as.tibble(MASS::mvrnorm(mu = hw.fox.2@coef,
                          Sigma = hw.fox.2@vcov , n = N )) %>%      # rather than extract.samples
  mutate(groupsize = sample(seq(from = 2, to = 8, by = 0.01), N, replace = T),
         predmean = a + b * groupsize ,                                  # line uncertainty
         predverb = rnorm(N, a + b*groupsize, sigma )) %>%         # data uncertainty
  group_by(groupsize) %>%
  mutate(lb_mu = rethinking::HPDI(predmean, prob = .95)[1],
         ub_mu = rethinking::HPDI(predmean, prob = .95)[2],
         lb_wt = rethinking::HPDI(predverb, prob = .95)[1],
         ub_wt = rethinking::HPDI(predverb, prob = .95)[2]) %>%
  slice(1) %>%
  mutate(yhat = hw.fox.2@coef["a"] + hw.fox.2@coef["b"] * groupsize) %>%       # yhat for reg line
  select(groupsize, yhat, lb_mu, ub_mu, lb_wt, ub_wt)

#plot
ggplot(f, aes(x = groupsize)) +
  geom_jitter(aes(y = weight), alpha = .3) +
  geom_line(data = preds, aes(y = yhat)) +
  geom_ribbon(data = preds, aes(ymin = lb_mu, ymax = ub_mu), alpha = .3) +
  geom_ribbon(data = preds, aes(ymin = lb_wt, ymax = ub_wt), alpha = .2) +
  labs(x = "Group Size",
       y = "Weight",
       title = "Predicted Weight of Foxes, by Group Size")
```



Looking at the results from the models and the plots, it appears that group size is important for predicting body weight, but territory size (area) is not. The second model which predicts body weight based on group size has a beta of -0.12 and an 89% credible interval of -0.23 to -0.01, meaning that for every one unit increase in group size, the model predicts a 0.12 decrease in body weight. The second plot similarly shows a slightly negative slope.

On the other hand, the first model and first plot indicate that there is largely no relationship between the territory size (area) and body weight. In particular, the model predicts a beta of 0.03, but an 89% credible interval that includes 0, going from -0.16 to 0.21. Along these lines, the first plot shows a basically horizontal line, meaning that as area increases, there doesn't seem to be much of a corresponding increase in weight.

5H2

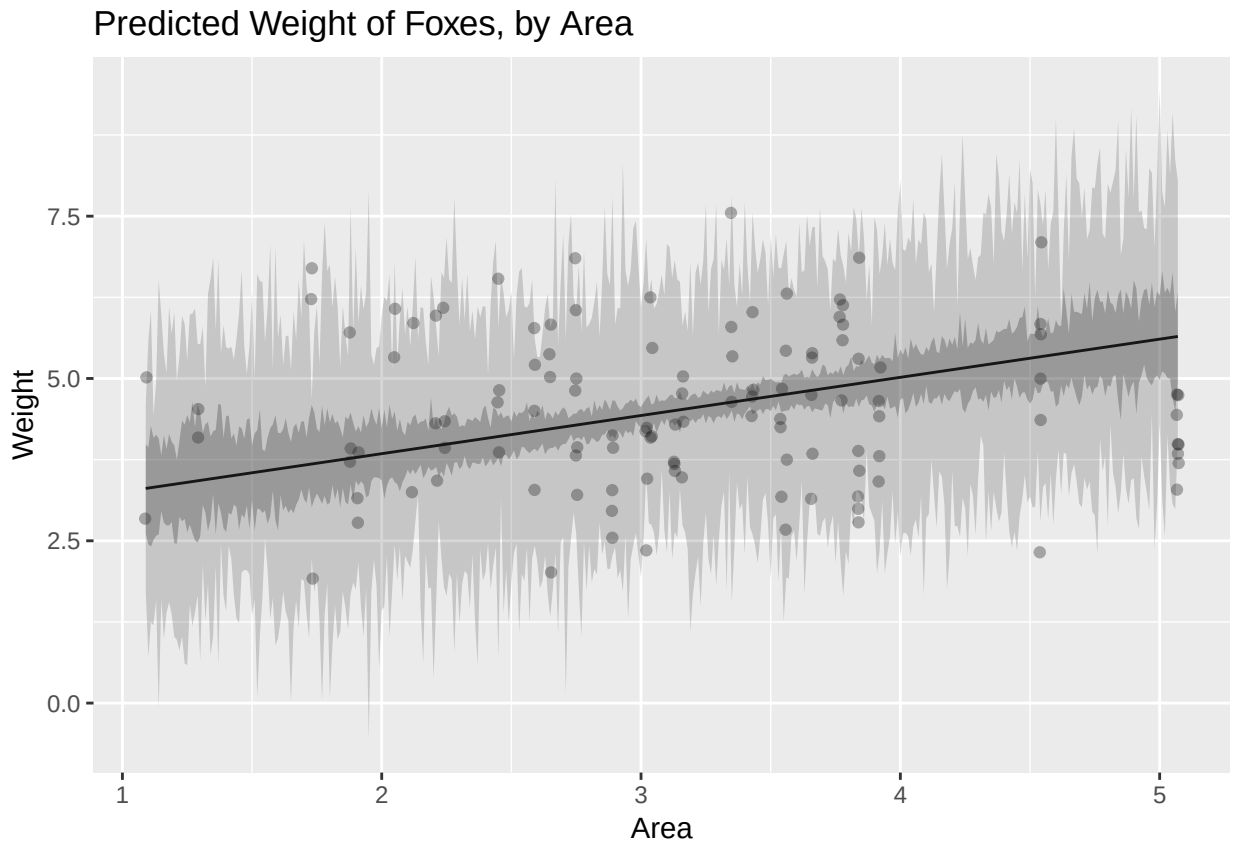
```
hw.fox.3 <- rethinking::map(
        alist(
                weight ~ dnorm(mu, sigma),
                mu <- a + bg*groupsize + ba*area,
                a ~ dnorm(0, 10),
                bg ~ dnorm(0, 1),
                ba ~ dnorm(0, 1),
                sigma ~ dunif(0, 10)
        ),
        data = f)

print("Predicting Body Weight by both Territory Size and Group Size")
```

```
## [1] "Predicting Body Weight by both Territory Size and Group Size"
```

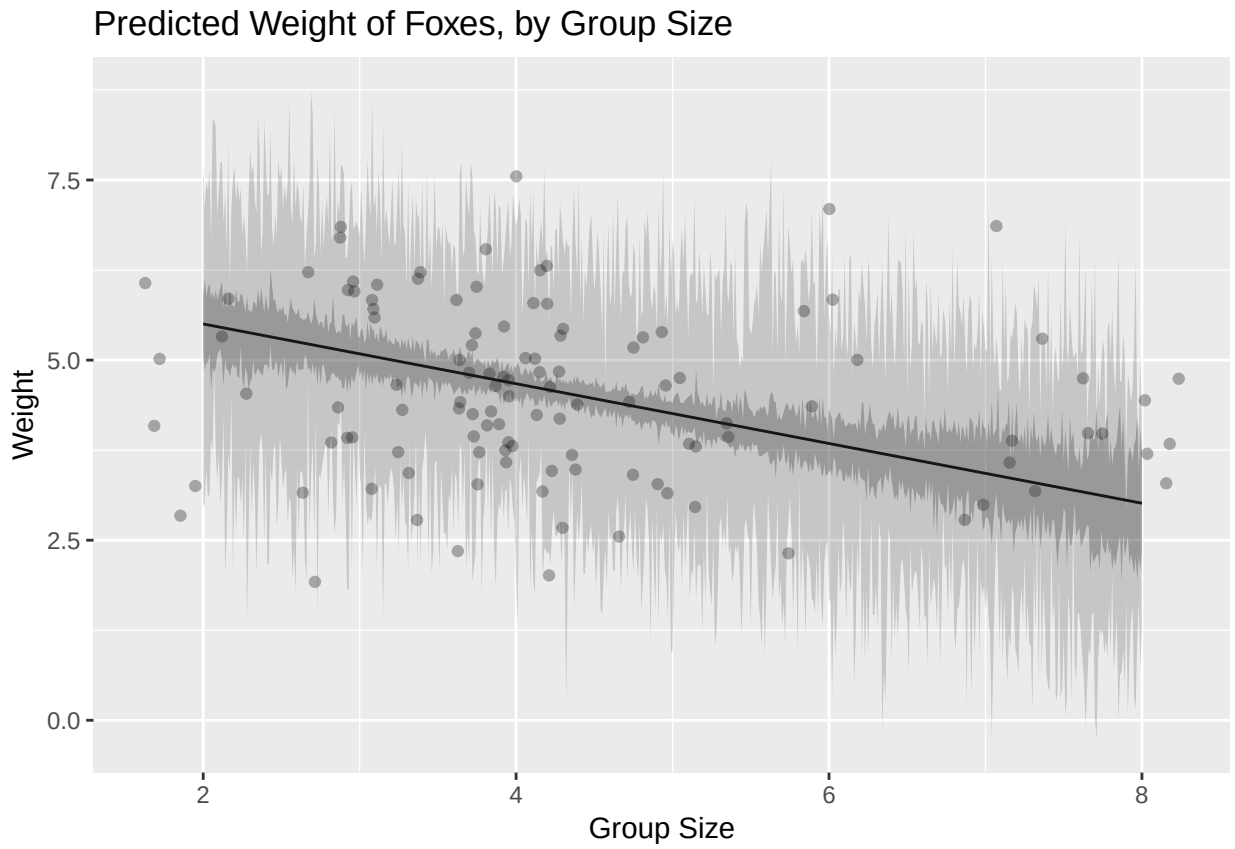
```
precis(hw.fox.3)
```

```
##         Mean StdDev  5.5% 94.5%
## a       4.47   0.37  3.88  5.06
## bg     -0.41   0.12 -0.60 -0.23
## ba      0.59   0.20  0.28  0.90
## sigma   1.12   0.07  1.00  1.24
```

```
preds <-
  as.tibble(MASS::mvrnorm(mu = hw.fox.3@coef,
                          Sigma = hw.fox.3@vcov , n = N )) %>%      # rather than extract.samples
  mutate(area = sample(seq(from = 1.09, to = 5.07, by = 0.01), N, replace = T),
         groupsize = mean(f$groupsize, na.rm = T),
         predmean = a + ba*area + bg*groupsize,                              # line uncertainty
         predverb = rnorm(N, a + ba*area + bg*groupsize, sigma )) %>%         # data uncertainty
  group_by(area) %>%
  mutate(lb_mu = rethinking::HPDI(predmean, prob = .95)[1],
         ub_mu = rethinking::HPDI(predmean, prob = .95)[2],
         lb_wt = rethinking::HPDI(predverb, prob = .95)[1],
         ub_wt = rethinking::HPDI(predverb, prob = .95)[2]) %>%
  slice(1) %>%
  mutate(yhat = hw.fox.3@coef["a"] + hw.fox.3@coef["ba"] * area +
                hw.fox.3@coef["bg"]*groupsize) %>%       # yhat for reg line
  select(area, groupsize, yhat, lb_mu, ub_mu, lb_wt, ub_wt)

#plot
ggplot(f, aes(x = area)) +
  geom_jitter(aes(y = weight), alpha = .3) +
  geom_line(data = preds, aes(y = yhat)) +
  geom_ribbon(data = preds, aes(ymin = lb_mu, ymax = ub_mu), alpha = .3) +
  geom_ribbon(data = preds, aes(ymin = lb_wt, ymax = ub_wt), alpha = .2) +
  labs(x = "Area",
       y = "Weight",
       title = "Predicted Weight of Foxes, by Area")
```




```
preds <-
  as.tibble(MASS::mvrnorm(mu = hw.fox.3@coef,
                          Sigma = hw.fox.3@vcov , n = N )) %>%      # rather than extract.samples
  mutate(area = mean(f$area, na.rm = T),
         groupsize = sample(seq(from = 2, to = 8, by = 0.01), N, replace = T),
         predmean = a + ba*area + bg*groupsize,                              # line uncertainty
         predverb = rnorm(N, a + ba*area + bg*groupsize, sigma )) %>%         # data uncertainty
  group_by(groupsize) %>%
  mutate(lb_mu = rethinking::HPDI(predmean, prob = .95)[1],
         ub_mu = rethinking::HPDI(predmean, prob = .95)[2],
         lb_wt = rethinking::HPDI(predverb, prob = .95)[1],
         ub_wt = rethinking::HPDI(predverb, prob = .95)[2]) %>%
  slice(1) %>%
  mutate(yhat = hw.fox.3@coef["a"] + hw.fox.3@coef["ba"] * area +
                hw.fox.3@coef["bg"]*groupsize) %>%       # yhat for reg line
  select(area, groupsize, yhat, lb_mu, ub_mu, lb_wt, ub_wt)

ggplot(f, aes(x = groupsize)) +
  geom_jitter(aes(y = weight), alpha = .3) +
  geom_line(data = preds, aes(y = yhat)) +
  geom_ribbon(data = preds, aes(ymin = lb_mu, ymax = ub_mu), alpha = .3) +
  geom_ribbon(data = preds, aes(ymin = lb_wt, ymax = ub_wt), alpha = .2) +
  labs(x = "Group Size",
       y = "Weight",
       title = "Predicted Weight of Foxes, by Group Size")
```



Including both predictors in the same model increases the magnitude of both of the beta specifications. Most drastically, the beta for area size now is 0.59 and the 89% credible interval does not include 0, spanning from 0.28 to 0.90. This suggests that for every one unit increase in territory area size, there is a corresponding 0.59 unit increase in body weight. Similarly, the beta for group size changed from -0.12 to -0.41, with an 89% credible interval of -0.60 to -0.23. This model now predicts that for every one unit increase in group size, there will be a 0.41 decrease in body weight. Again, one can see the difference in slopes visually in the two plots, with the first revealing a positive relationship between area and body weight and the second a more sharply negative slope for the regression line relating group size and body weight than in the previous model that had only included group size.

The reason why the model with both of the predictors gives such different outcomes for beta for these two variables than when they are on their own is because this is an example of a masked relationship. Because area is positively correlated with weight (0.0194773), groupsize is negative correlated with weight (0.8275945), and area and groupsize are positively correlated with each other you need to put them in the same model to account for the relationships between the variables.

```
hw.fox.4 <- rethinking::map(
        alist(
                weight ~ dnorm(mu, sigma),
                mu <- a + bg*groupsize + bf*avgfood,
                a ~ dnorm(0, 10),
                bg ~ dnorm(0, 1),
                bf ~ dnorm(0, 1),
                sigma ~ dunif(0, 10)
        ),
        data = f)
```

```
hw.fox.5 <- rethinking::map(
        alist(
                weight ~ dnorm(mu, sigma),
                mu <- a + bg*groupsize + ba*area + bf*avgfood,
                a ~ dnorm(0, 10),
                bg ~ dnorm(0, 1),
                ba ~ dnorm(0, 1),
                bf ~ dnorm(0, 1),
                sigma ~ dunif(0, 10)
        ),
        data = f)

print("Body weight predicted by avgfood and groupsize results:")
```

```
## [1] "Body weight predicted by avgfood and groupsize results:"
```

```
precis(hw.fox.4)
```

```
##         Mean StdDev  5.5% 94.5%
## a       4.69   0.37  4.09  5.28
## bg     -0.30   0.11 -0.48 -0.12
## bf      1.51   0.78  0.26  2.76
## sigma   1.13   0.08  1.01  1.25
```

```
print("Body weight predicted by avgfood, groupsize, and area results:")
```

```
## [1] "Body weight predicted by avgfood, groupsize, and area results:"
```

```
precis(hw.fox.5)
```

```
##         Mean StdDev  5.5% 94.5%
## a       4.34   0.39  3.72  4.96
## bg     -0.47   0.13 -0.68 -0.26
## ba      0.51   0.21  0.18  0.85
## bf      0.81   0.82 -0.50  2.13
## sigma   1.11   0.07  0.99  1.23
```

Predicting body weight by average food and group size indicates that there is a positive relationship between average food, with a beta of 1.51 and an 89% credible interval of 0.26 to 2.76, indicating that for every one unit increase in average food consumption, there should be a 1.51 increase in body weight. The relationship between group size and body weight is still negative as it has been in the previous models, with a beta of -0.30 and 89% credible interval of -0.47 to -0.11, indicating that a one unit increase in group size should correspond to a 0.30 unit decrease in body weight. Including all 3 variables in one model (avgfood, area, and groupsize), leads to an increase in magnitude of the beta for groupsize, which is now -0.47 and has an 89% credible interval from -0.68 to -0.26. This is similar, but slightly larger in magnitude to the beta specified by the model with only groupsize and area, which was -0.41.

However, the beta specifications for both area and average food are smaller than when only one or the other is included in a model with groupsize. For instance, the beta for average food is 0.81 with a credible interval of -0.51 to 2.13, while it had been 1.51 and had a credible interval that did not include 0 in the previous model. Similarly, but not as drastically, the beta for area is 0.51, with an 89% credible interval of 0.18 to 0.85, while in the previous exercise the model with only area and group size had a beta specification of 0.59 and an 89% credible interval of 0.28 to 0.90.

b) One consideration we might take into account is including either average food or territory area size in the final model, but not both. To do this, we need to evaluate which variable adds more to the model's ability to predict the data we already have, in other words, we should do a posterior predictive check for

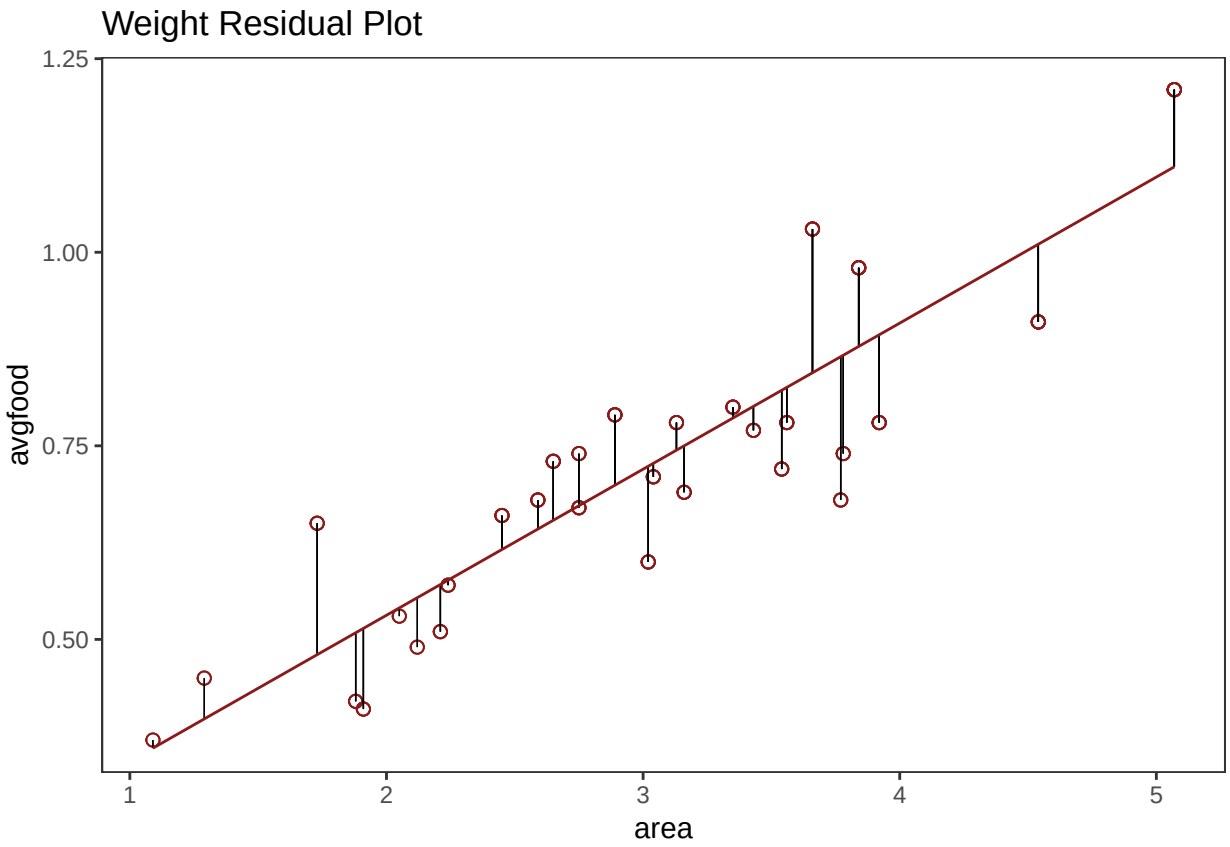
the model with groupsize and area and the model with groupsize and avgfood and see which appears to fit the data better. Below, a residual plot and predictor residual plots for the weight data are produced to help determine which variable contributes more to explaining the variance in weight.

```
bhw.c <-
  brm(data = f, family = gaussian,
      avgfood ~ 1 + area,
      prior = c(prior(normal(0, 10), class = Intercept),
                prior(normal(0, 1), class = b),
                prior(uniform(0, 10), class = sigma)),
      iter = 2000, warmup = 500, chains = 4, cores = 4)
```

```
## Compiling the C++ model
```

```
## Start sampling
```

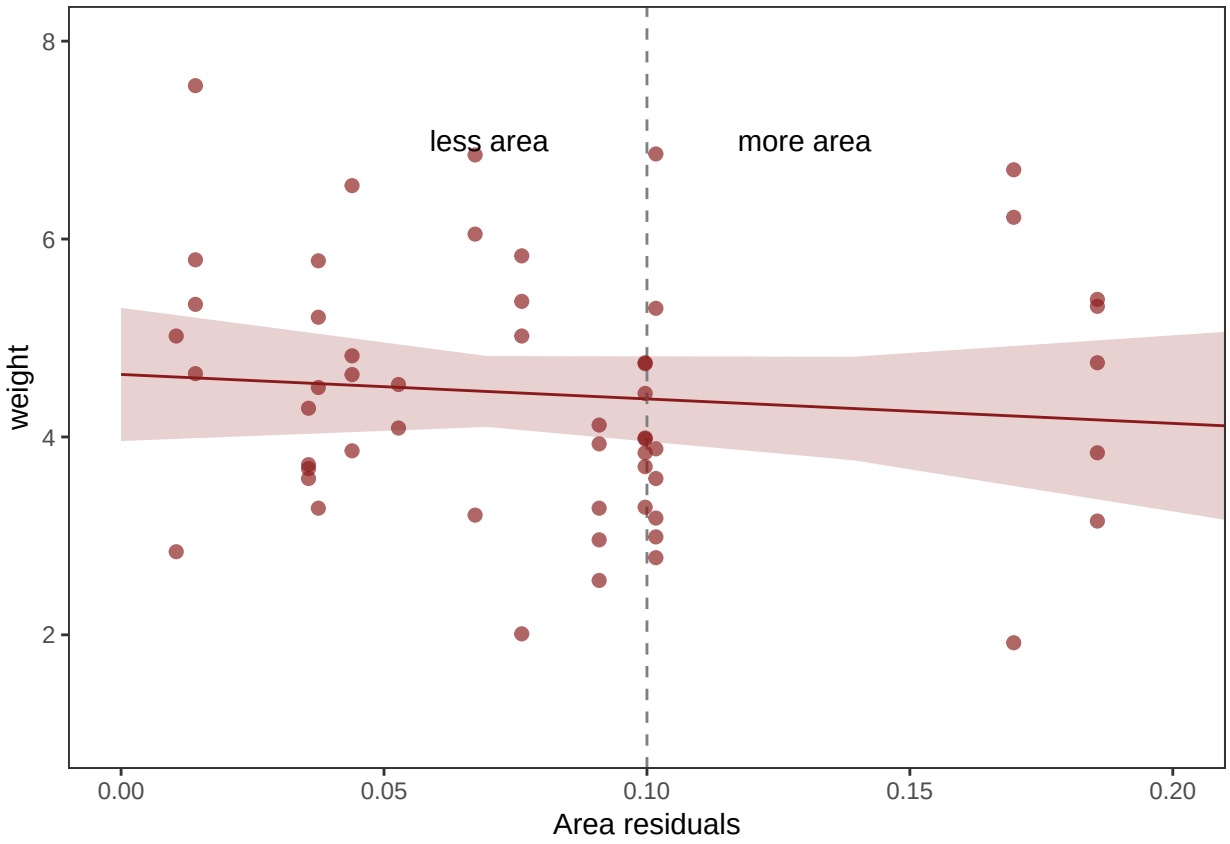
```
fitd_bhw.c <-
  fitted(bhw.c) %>%
  as_tibble() %>%
  bind_cols(f)

fitd_bhw.c %>%

  ggplot(aes(x = area, y = avgfood)) +
  geom_point(size = 2, shape = 1, color = "firebrick4") +
  geom_segment(aes(xend = area, yend = Estimate),
               size = 1/4) +
  geom_line(aes(y = Estimate),
            color = "firebrick4") +
  coord_cartesian(ylim = range(f$avgfood)) +
  theme_bw() +
  theme(panel.grid = element_blank()) + ggtitle("Weight Residual Plot")
```




```
text <-
  tibble(Estimate = c(0.07, 0.13),
         weight  = 7,
         label = c("less area", "more area"))

res_bhw.c <-
  residuals(bhw.c) %>%
  as_tibble() %>%
  bind_cols(f)

res_bhw.c %>%
  ggplot(aes(x = Estimate, y = weight)) +
  stat_smooth(method = "lm", fullrange = T,
              color = "firebrick4", fill = "firebrick4",
              alpha = 1/5, size = 1/2) +
  geom_vline(xintercept = 0.1, linetype = 2, color = "grey50") +
  geom_point(size = 2, color = "firebrick4", alpha = 2/3) +
      geom_text(data = text,
            aes(label = label)) +
  scale_x_continuous(limits = c(0, 5.5)) +
  coord_cartesian(xlim = c(0,0.2),
                  ylim = c(1, 8)) +
  labs(x = "Area residuals") +
  theme_bw() +
  theme(panel.grid = element_blank())
```

```
## Warning: Removed 61 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 61 rows containing missing values (geom_point).
```




```
bhw.d <-
  brm(data = f, family = gaussian,
      area ~ 1 + avgfood,
      prior = c(prior(normal(0, 10), class = Intercept),
                prior(normal(0, 1), class = b),
                prior(uniform(0, 10), class = sigma)),
      iter = 2000, warmup = 500, chains = 4, cores = 4)
```

```
## Compiling the C++ model
```

```
## recompiling to avoid crashing R session
```

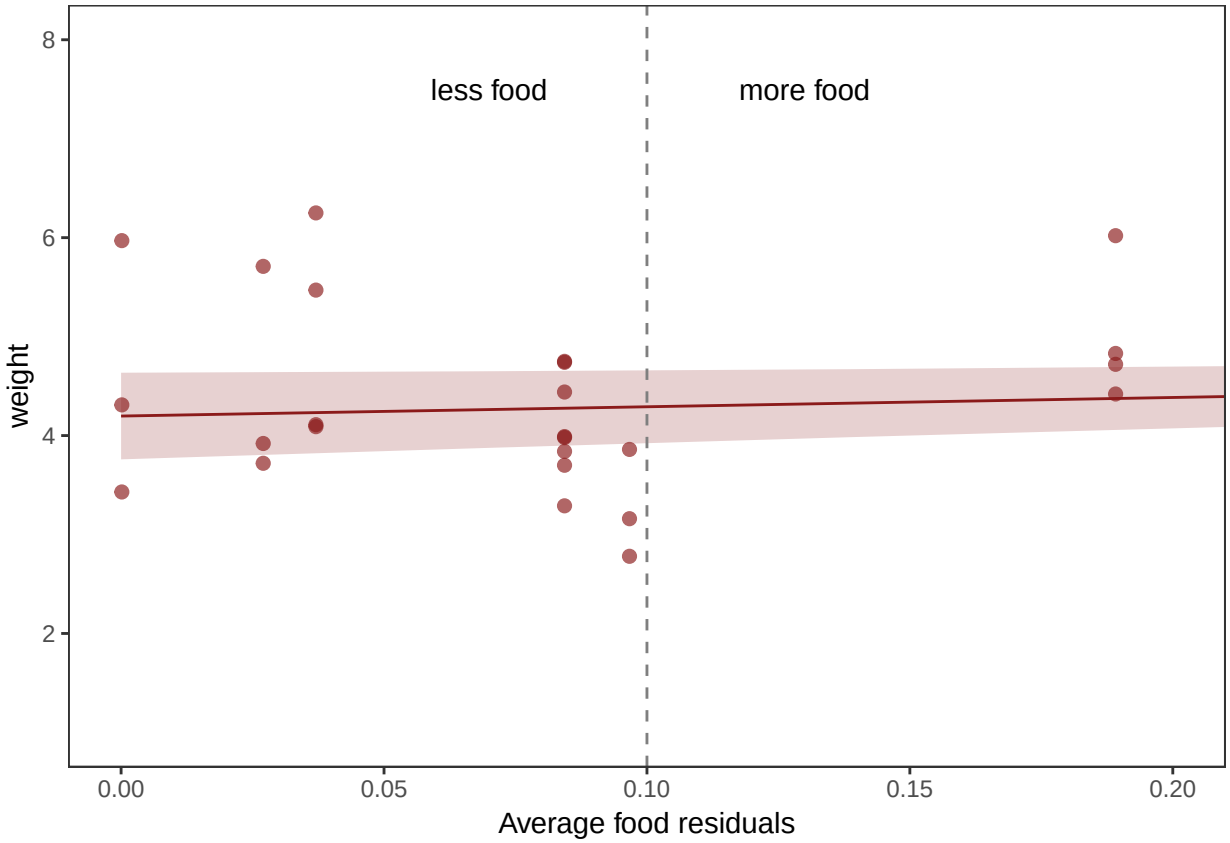
```
## Start sampling
```

```
res_bhw.d <-
  residuals(bhw.d) %>%
  as_tibble() %>%
  bind_cols(f)

text <-
  tibble(Estimate = c(0.07, 0.13),
         weight  = 7.5,
         label = c("less food", "more food"))

res_bhw.d %>%
```

```
  ggplot(aes(x = Estimate, y = weight)) +
  stat_smooth(method = "lm", fullrange = T,
              color = "firebrick4", fill = "firebrick4",
              alpha = 1/5, size = 1/2) +
  geom_vline(xintercept = 0.1, linetype = 2, color = "grey50") +
  geom_point(size = 2, color = "firebrick4", alpha = 2/3) +
    geom_text(data = text,
            aes(label = label)) +
  scale_x_continuous(limits = c(0, 5.5)) +
  coord_cartesian(xlim = c(0,0.2),
                  ylim = c(1, 8)) +
  labs(x = "Average food residuals") +
  theme_bw() +
  theme(panel.grid = element_blank())
```

```
## Warning: Removed 58 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 58 rows containing missing values (geom_point).
```



Based on the plots produced above, I think the more important variable to keep in the model is area, rather than average food size. As can particularly be observed from looking at the second two plots, groups with more average food than their area size have about the same weight as those with less average food than their area size do, as can be seen in the relatively flat shaded area of the last plot. However, in the second plot, groups that have more territorial area for their average food intake have lower body weight, while groups with less territorial area for their average food intake have higher body weight. This indicates to me that knowing the territorial area contributes more to the model than knowing the average food intake.

c) As you can see in the model results from above, the standard errors for the estimations of the betas for average food consumption and territory area size increased by quite a bit when both were included in the same model. This is likely because the two predictors are highly correlated to each other, meaning the model is weakened by *multicollinearity*. As McElreath explains, when two variables are highly correlated with each other, this means that the additional knowledge gained from adding one variable or the other when you already know the other predictor, is quite low. The correlation between avgfood and area is 0.8831038, which is quite high, suggesting that it is indeed multicollinearity at play in influencing the larger standard errors and more imprecise estimates when both are in the same model.