Statistical Rethinking: Chapter 5 - Multivariate Linear Models

Chris Hayduk

January 27, 2019

1 Easy

Problem 5E1.
Which of the linear models below are multiple linear regressions?

1. $\mu_i = \alpha + \beta x_i$
2. $\mu_i = \beta_x x_i + \beta_z z_i$
3. $\mu_i = \alpha + \beta(x_i - z_i)$
4. $\mu_i = \alpha + \beta_x x_i + \beta_z z_i$

Linear models 2 and 4 are multiple linear regressions.

Problem 5E2.
Write down a multiple regression to evaluate the claim: *Animal diversity is linearly related to latitude, but only after controlling for plant diversity.* You just need to write down the model definition.

$$\begin{gathered}
\text{animal diversity}_i \sim \text{Normal}(\mu_i, \sigma) \\
\mu_i = \beta_{\text{latitude}}\text{latitude}_i + \beta_{\text{diversity}}\text{diversity}_i \\
\beta_{\text{latitude}} \sim \text{Normal}(0, 10) \\
\beta_{\text{diversity}} \sim \text{Normal}(0, 10) \\
\sigma \sim \text{Uniform}(0, 10)
\end{gathered}$$

Problem 5E3.
Write down a multiple regression to evaluate the claim: *Neither amount of funding nor size of laboratory is by itself a good predictor of time to PhD degree; but together these variables are both positively associated with time to degree.* Write down the model definition and indicate which side of zero each slope parameter should be on.

$$
\begin{aligned}
\text{time}_i &\sim \text{Normal}(\mu_i, \sigma) \\
\mu_i &= \beta_{\text{lab size}}\text{lab size}_i + \beta_{\text{funding}}\text{funding}_i \\
\beta_{\text{lab size}} &\sim \text{Normal}(0, 10) \\
\beta_{\text{funding}} &\sim \text{Normal}(0, 10) \\
\sigma &\sim \text{Uniform}(0, 10)
\end{aligned}
$$

Both parameters should have slopes greater than zero since the problem specifies that "together the variables are both positively associated with time to degree".

Problem 5E4.
Suppose you have a single categorical predictor with 4 levels (unique values), labeled A, B, C, and D. Let A_i be an indicator variable that is 1 where case i is in category A. Also suppose B_i, C_i, and D_i for the other categories. Now which of the following linear models are inferentially equivalent ways to include the categorical variable in a regression? Models are inferentially equivalent when it's possible to compute one posterior distribution from the posterior distribution of another model.

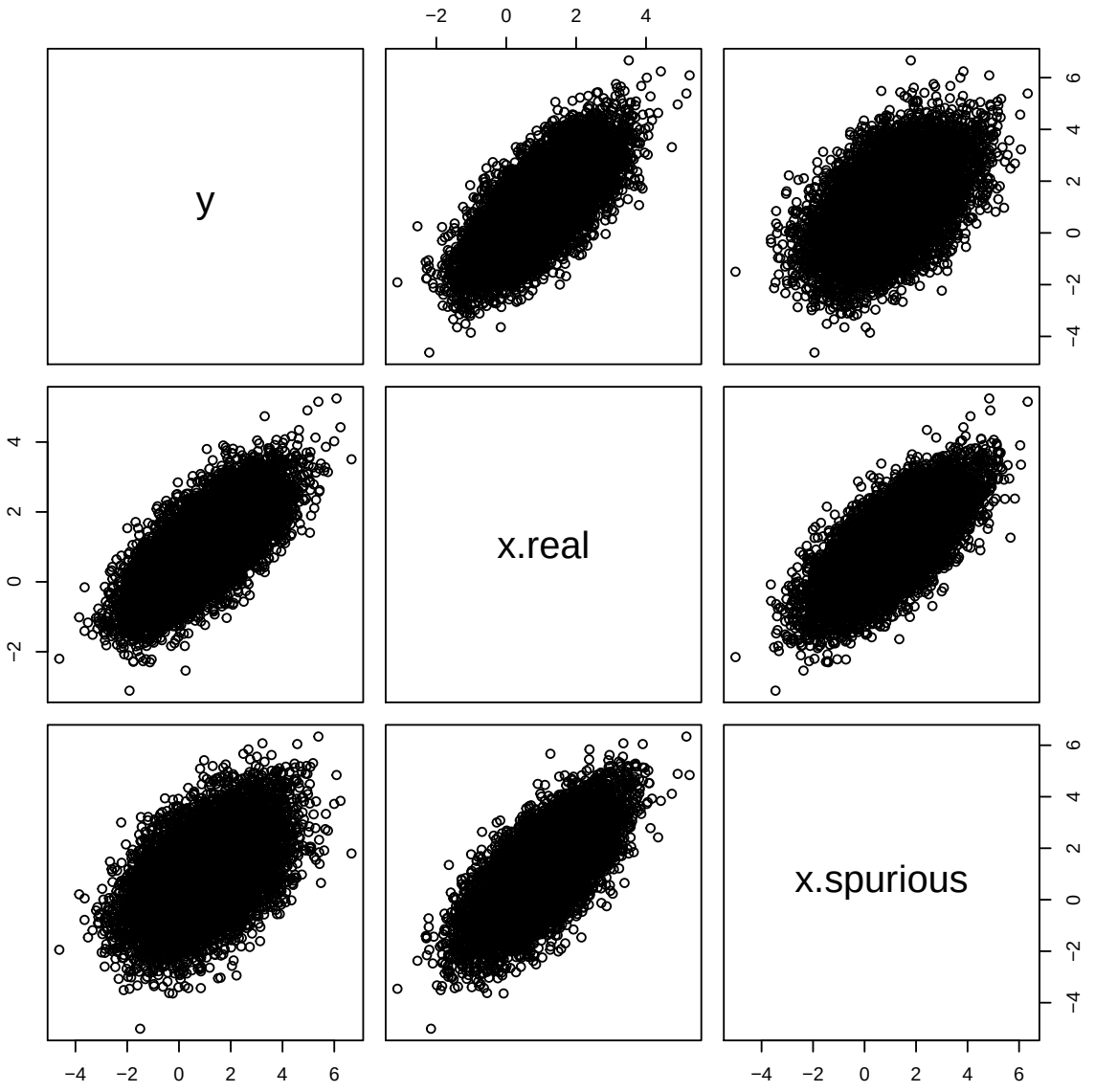
1. $\mu_i = \alpha + \beta_A A_i + \beta_B B_i + \beta_D D_i$
2. $\mu_i = \alpha + \beta_A A_i + \beta_B B_i + \beta_C C_i + \beta_D D_i$
3. $\mu_i = \alpha + \beta_B B_i + \beta_C C_i + \beta_D D_i$
4. $\mu_i = \alpha_A A_i + \alpha_B B_i + \alpha_C C_i + \alpha_D D_i$
5. $\mu_i = \alpha_i(1 - B_i - C_i - D_i) + \alpha_B B_i + \alpha_C C_i + \alpha_D D_i$

Models 1, 3, 4, and 5 are all inferentially equivalent.

2 Medium

Problem 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).

```
N <- 1e4
x.real <- rnorm(n = N, mean = 1, sd = 1)
x.spurious <- rnorm(n = N, mean = x.real)
y <- rnorm(n = N, mean = x.real)

df <- data.frame(y = y, x.real = x.real, x.spurious = x.spurious)
pairs(df)
```

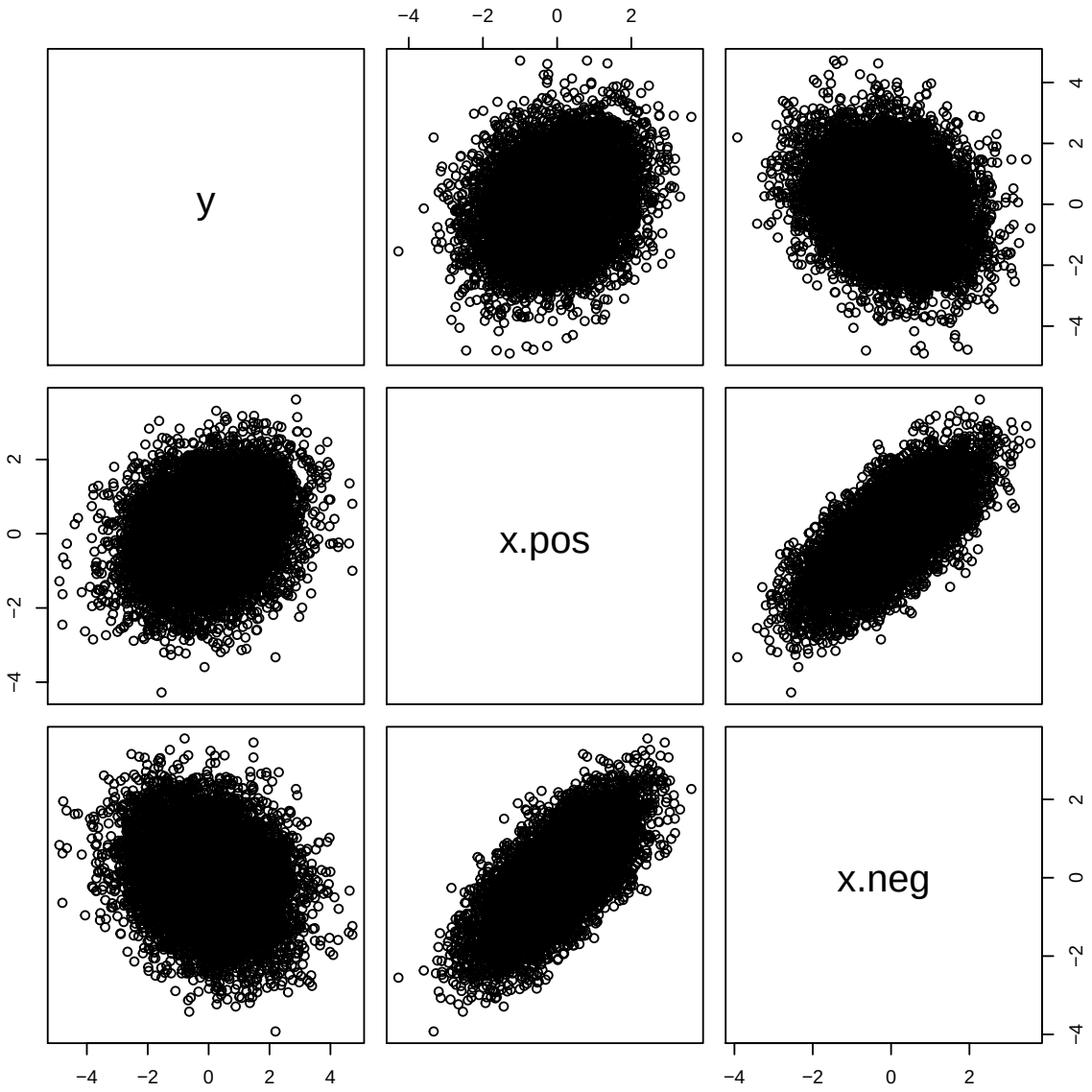



```
model <- lm(y ~ x.real + x.spurious)
precis(model)

##             Mean StdDev  5.5% 94.5%
## (Intercept) 0.01   0.01 -0.01  0.03
## x.real      1.00   0.01  0.97  1.02
## x.spurious  0.00   0.01 -0.02  0.01
```

Problem 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.

```
N <- 1e4
rho <- 0.7

x.pos <- rnorm(N)
x.neg <- rnorm(N, rho*x.pos, sqrt(1-rho^2))
y <- rnorm(N, x.pos - x.neg)

df <- data.frame(y = y, x.pos = x.pos, x.neg = x.neg)
pairs(df)
```




```
model <- lm(y ~ x.pos + x.neg)
precis(model)

##              Mean StdDev  5.5% 94.5%
## (Intercept)  0.01   0.01 -0.01  0.02
## x.pos        1.00   0.01  0.97  1.02
## x.neg       -0.99   0.01 -1.01 -0.97
```

Problem 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 firefightes do not

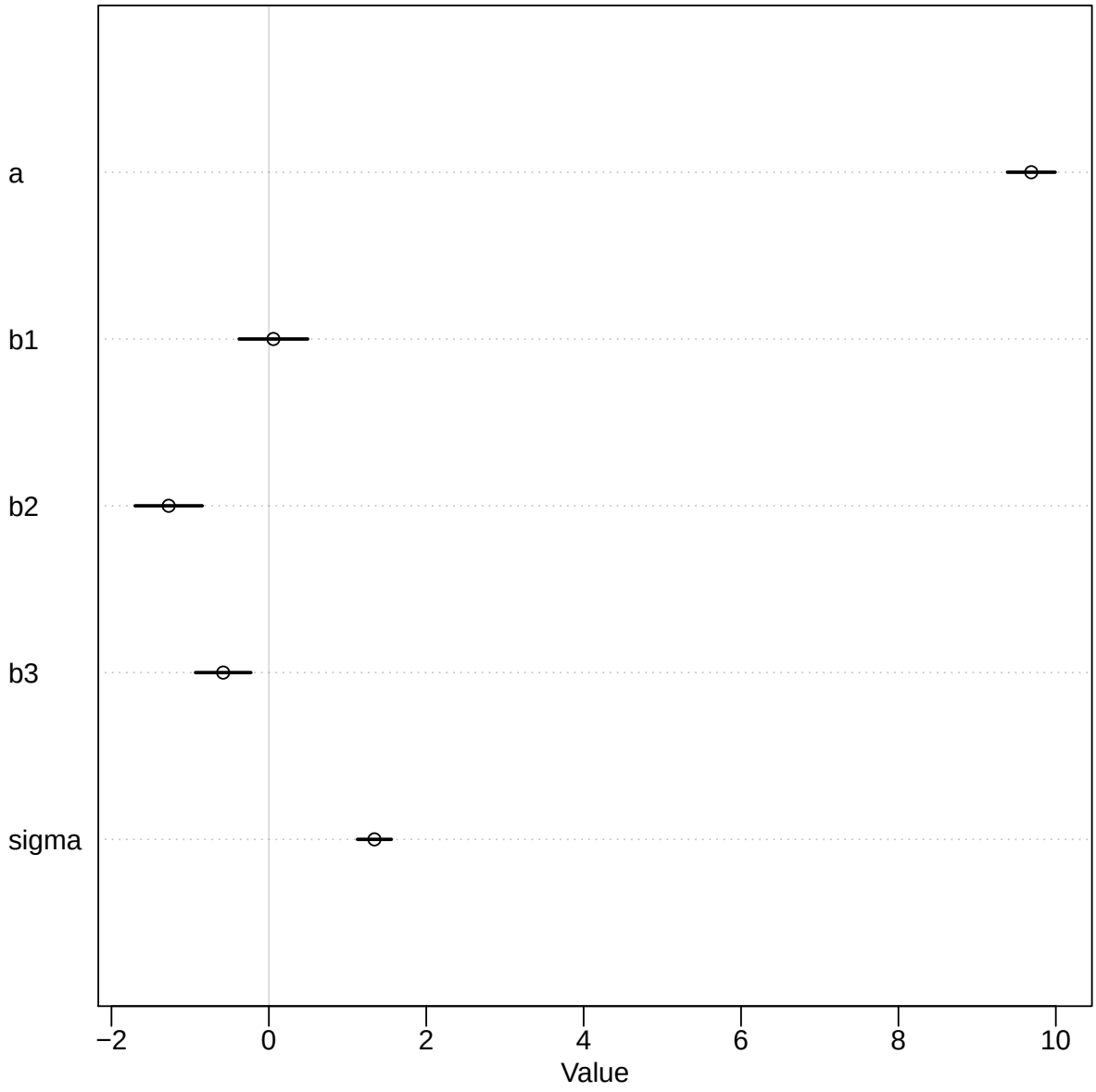
cause fires. Nevertheless, this is not a spurious correlation. Instead fires cause firefighters. Consider the same reversal of causal inference in the context of the divorice and marriage data. How might a high divorce rate cause a higher marriage rate? Can you think of a way to evaluate this relationship, using multiple regression?

A high divorce rate may cause a higher marriage rate because people who are divorced may get married multiple times throughout their lives.

Problem 5M4.
In the divorce data, States with high numbers of Mormons have much lower divorce rates than the regression models expected. Find a list of LDS population by State and use those numbers as a predictor variable, predicting divorce rate using marriage rate, median age at marriage, and percent LDS population (possibly standardized). You may want to consider transformations of the raw percent LDS variable.

```
setwd(dirname(rstudioapi::getActiveDocumentContext()$path))

## Error:  RStudio not running

mormon <- read_excel("Mormon.xlsx")

data(WaffleDivorce)

d <- WaffleDivorce

names(mormon) <- c("Location", "MormonPop", "StatePop", "PercMormon")

mormon <- mormon[,c("Location", "MormonPop", "PercMormon")]

d <- left_join(d, mormon)

d$MedianAgeMarriage.s <- (d$MedianAgeMarriage - mean(d$MedianAgeMarriage))/sd(d$MedianAg

d$Marriage.s <- (d$Marriage - mean(d$Marriage))/sd(d$Marriage)

d$PercMormon.s <- (d$PercMormon - mean(d$PercMormon))/sd(d$PercMormon)

model <- map(
  alist(
    Divorce ~ dnorm(mu, sigma),
    mu <- a + b1 * Marriage.s + + b2 * MedianAgeMarriage.s + b3 * PercMormon.s,
    a ~ dnorm(10, 10),
    b1 ~ dnorm(0, 1),
    b2 ~ dnorm(0, 1),
    b3 ~ dnorm(0, 1),
```

```
    sigma ~ dunif(0, 10)
  ), data = d)

precis(model)

##         Mean StdDev  5.5% 94.5%
## a       9.69   0.19  9.38  9.99
## b1      0.06   0.27 -0.38  0.49
## b2     -1.27   0.27 -1.70 -0.85
## b3     -0.58   0.22 -0.93 -0.23
## sigma   1.34   0.13  1.13  1.56

plot(precis(model))
```



Problem 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 positiviely 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.

Model 1:

$$\text{obesity rate}_i \sim \text{Normal}(\mu_i, \sigma)$$
$$\mu_i = \beta_{\text{gas}}\text{gas price}_i + \beta_{\text{driving}}\text{hours driving per year per capita}_i$$
$$\beta_{\text{gas}} \sim \text{Normal}(0, 10)$$
$$\beta_{\text{driving}} \sim \text{Normal}(0, 10)$$
$$\sigma \sim \text{Uniform}(0, 10)$$

Model 2:

$$\text{obesity rate}_i \sim \text{Normal}(\mu_i, \sigma)$$
$$\mu_i = \beta_{\text{gas}}\text{gas price}_i + \beta_{\text{restaurant}}\text{restaurant trips per year per capita}_i$$
$$\beta_{\text{gas}} \sim \text{Normal}(0, 10)$$
$$\beta_{\text{restaurant}} \sim \text{Normal}(0, 10)$$
$$\sigma \sim \text{Uniform}(0, 10)$$

3 Hard

All three exercises below use the same data, data(foxes) (part of rethinking). The urban fox (*Vulpes vulpes*) is a successful exploiter of human habitat. Since urbran foxes move in packs and defend territories, data on haitat quality and population density is also included. The data frame has fix 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

Problem 5H1.
Fit two bivariate Gaussian regressions, using map: (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 95% interval of the mean. Is either variable important for predicting fox body weight?

```
data(foxes)

d <- foxes

model1 <- map(
  alist(
    weight ~ dnorm(mu, sigma),
    mu <- a + b * area,
    a ~ dnorm(10, 10),
```

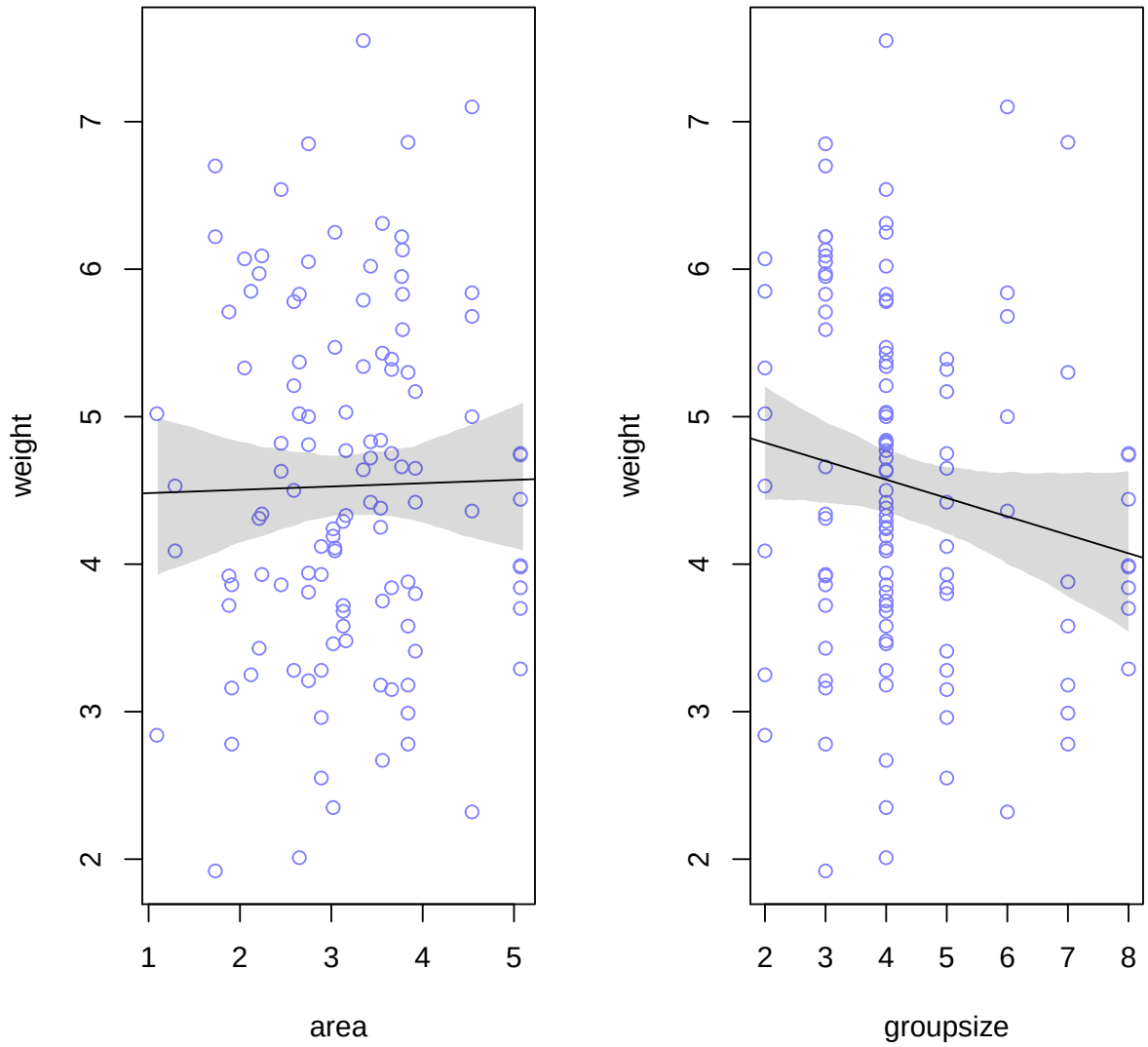
```
    b ~ dnorm(0, 100),
    sigma ~ dunif(0, 10)
  ), data = d)

model2 <- map(
  alist(
    weight ~ dnorm(mu, sigma),
    mu <- a + b * groupsize,
    a ~ dnorm(10, 10),
    b ~ dnorm(0, 100),
    sigma ~ dunif(0, 10)
  ), data = d)

area.seq <- seq(from = round(min(d$area), 1), to = round(max(d$area),1), length.out = 10

groupsize.seq <- seq(from = round(min(d$groupsize), 1), to = round(max(d$groupsize),1),

mu1 <- link(model1, data = data.frame(area=area.seq))
```

```
## [ 100 / 1000 ]
[ 200 / 1000 ]
[ 300 / 1000 ]
[ 400 / 1000 ]
[ 500 / 1000 ]
[ 600 / 1000 ]
[ 700 / 1000 ]
[ 800 / 1000 ]
[ 900 / 1000 ]
[ 1000 / 1000 ]
```

```
mu1.PI <- apply(mu1, 2, PI, 0.95)

mu2 <- link(model2, data = data.frame(groupsize = groupsize.seq))
```

```
## [ 100 / 1000 ]
[ 200 / 1000 ]
[ 300 / 1000 ]
[ 400 / 1000 ]
[ 500 / 1000 ]
[ 600 / 1000 ]
[ 700 / 1000 ]
[ 800 / 1000 ]
[ 900 / 1000 ]
[ 1000 / 1000 ]
```

```r
mu2.PI <- apply(mu2, 2, PI, 0.95)

par(mfrow=c(1,2))

plot(weight ~ area, data = d, col=rangi2)
abline(model1)
shade(mu1.PI, area.seq)

plot(weight ~ groupsize, data = d, col = rangi2)
abline(model2)
shade(mu2.PI, groupsize.seq)
```



Neither predictor appears to be heavily associated with weight, although group size does slightly better than area.

Problem 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 above?

```
model3 <- map(
  alist(
    weight ~ dnorm(mu, sigma),
    mu <- a + b1 * area + b2 * groupsize,
    a ~ dnorm(10, 10),
    b1 ~ dnorm(0, 100),
    b2 ~ dnorm(0, 100),
    sigma ~ dunif(0, 10)
  ), data = d)

precis(model3)
```

```
##        Mean StdDev  5.5% 94.5%
## a      4.46   0.37  3.87  5.05
## b1     0.62   0.20  0.30  0.94
## b2    -0.43   0.12 -0.63 -0.24
## sigma  1.12   0.07  1.00  1.24
```

```
A.avg <- mean(d$area)
G.avg <- mean(d$groupsize)

pred.data1 <-data.frame(area = area.seq, groupsize = G.avg)
pred.data2 <- data.frame(area = A.avg, groupsize = groupsize.seq)

mu1 <- link(model3, data = pred.data1)
```

```
## [ 100 / 1000 ]
[ 200 / 1000 ]
[ 300 / 1000 ]
[ 400 / 1000 ]
[ 500 / 1000 ]
[ 600 / 1000 ]
[ 700 / 1000 ]
[ 800 / 1000 ]
[ 900 / 1000 ]
[ 1000 / 1000 ]
```

```r
mu1.mean <- apply(mu1, 2, mean)
mu1.PI <- apply(mu1, 2, PI, 0.95)

mu2 <- link(model3, data = pred.data2)
```

```
## [ 100 / 1000 ]
[ 200 / 1000 ]
[ 300 / 1000 ]
[ 400 / 1000 ]
[ 500 / 1000 ]
[ 600 / 1000 ]
[ 700 / 1000 ]
[ 800 / 1000 ]
[ 900 / 1000 ]
[ 1000 / 1000 ]
```

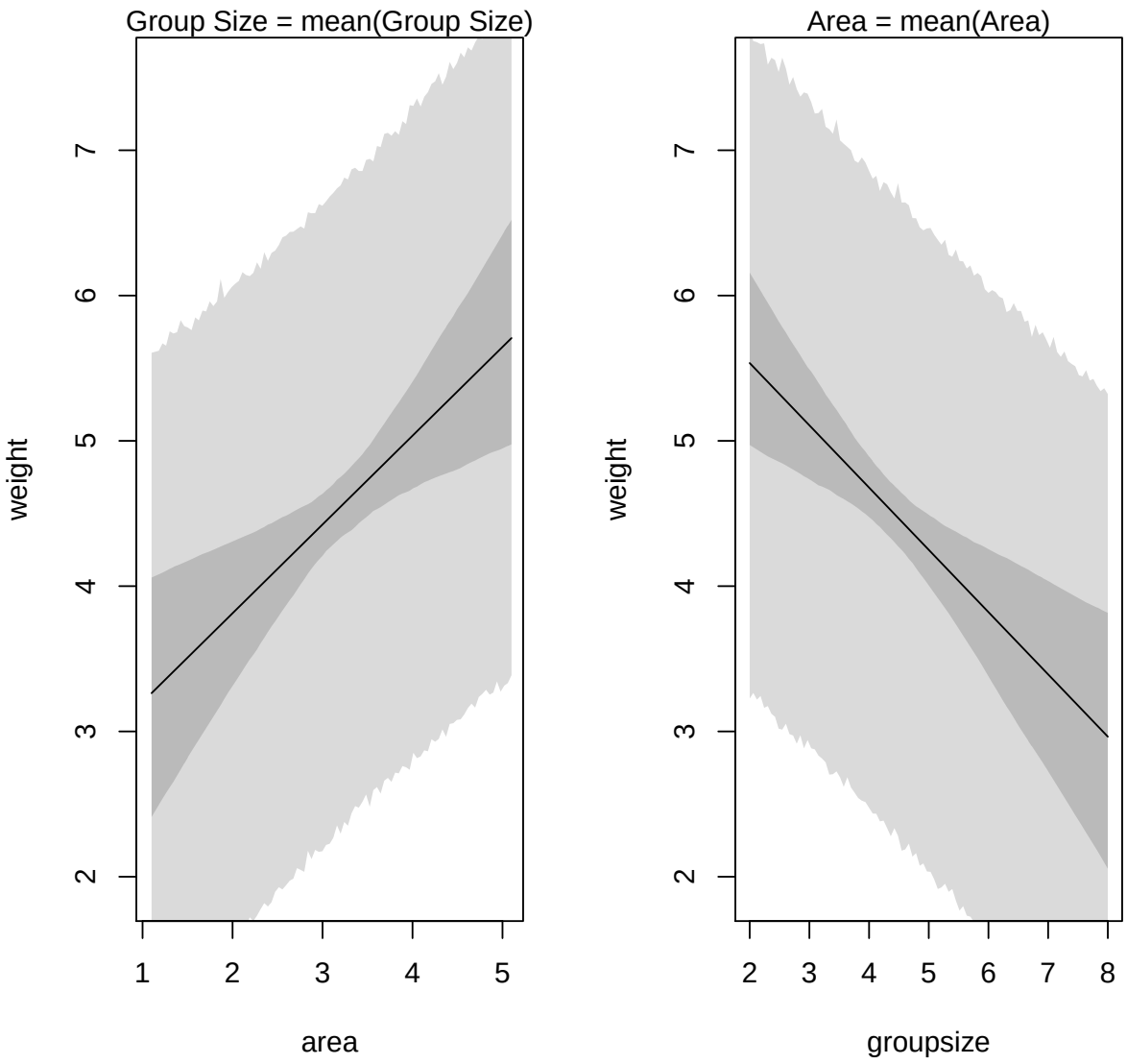
```r
mu2.mean <- apply(mu2, 2, mean)
mu2.PI <- apply(mu2, 2, PI, 0.95)

A.sim <- sim(model3, data = pred.data1, n = 1e4)
```

```
## [ 1000 / 10000 ]
[ 2000 / 10000 ]
[ 3000 / 10000 ]
[ 4000 / 10000 ]
[ 5000 / 10000 ]
[ 6000 / 10000 ]
[ 7000 / 10000 ]
[ 8000 / 10000 ]
[ 9000 / 10000 ]
[ 10000 / 10000 ]
```

```r
G.sim <- sim(model3, data = pred.data2, n = 1e4)
```

```
## [ 1000 / 10000 ]
[ 2000 / 10000 ]
[ 3000 / 10000 ]
[ 4000 / 10000 ]
[ 5000 / 10000 ]
[ 6000 / 10000 ]
[ 7000 / 10000 ]
[ 8000 / 10000 ]
[ 9000 / 10000 ]
[ 10000 / 10000 ]
```

```r
A.PI <- apply(A.sim, 2, PI, 0.95)
G.PI <- apply(G.sim, 2, PI, 0.95)

par(mfrow=c(1,2))

plot(weight ~ area, data = d, type = "n")
mtext("Group Size = mean(Group Size)")
lines(area.seq, mu1.mean)
shade(mu1.PI, area.seq)
shade(A.PI, area.seq)

plot(weight ~ groupsize, data = d, type = "n")
mtext("Area = mean(Area)")
lines(groupsize.seq, mu2.mean)
shade(mu2.PI, groupsize.seq)
shade(G.PI, groupsize.seq)
```



In each plot, the darker shaded region shows 95% percentile intervals of the means, and the lighter shaded region shows 95% prediction intervals.

When holding each variable constant at its mean, we can see that the other variable has a strong correlation with weight. In the case where area varies while group size remains constant, the correlation is strongly positive. In the opposite case, the correlation is strongly negative. The 95% percentile intervals of the means demonstrate that a horizontal line is not included in the interval, again demonstrating that the relationship for each variable is important.

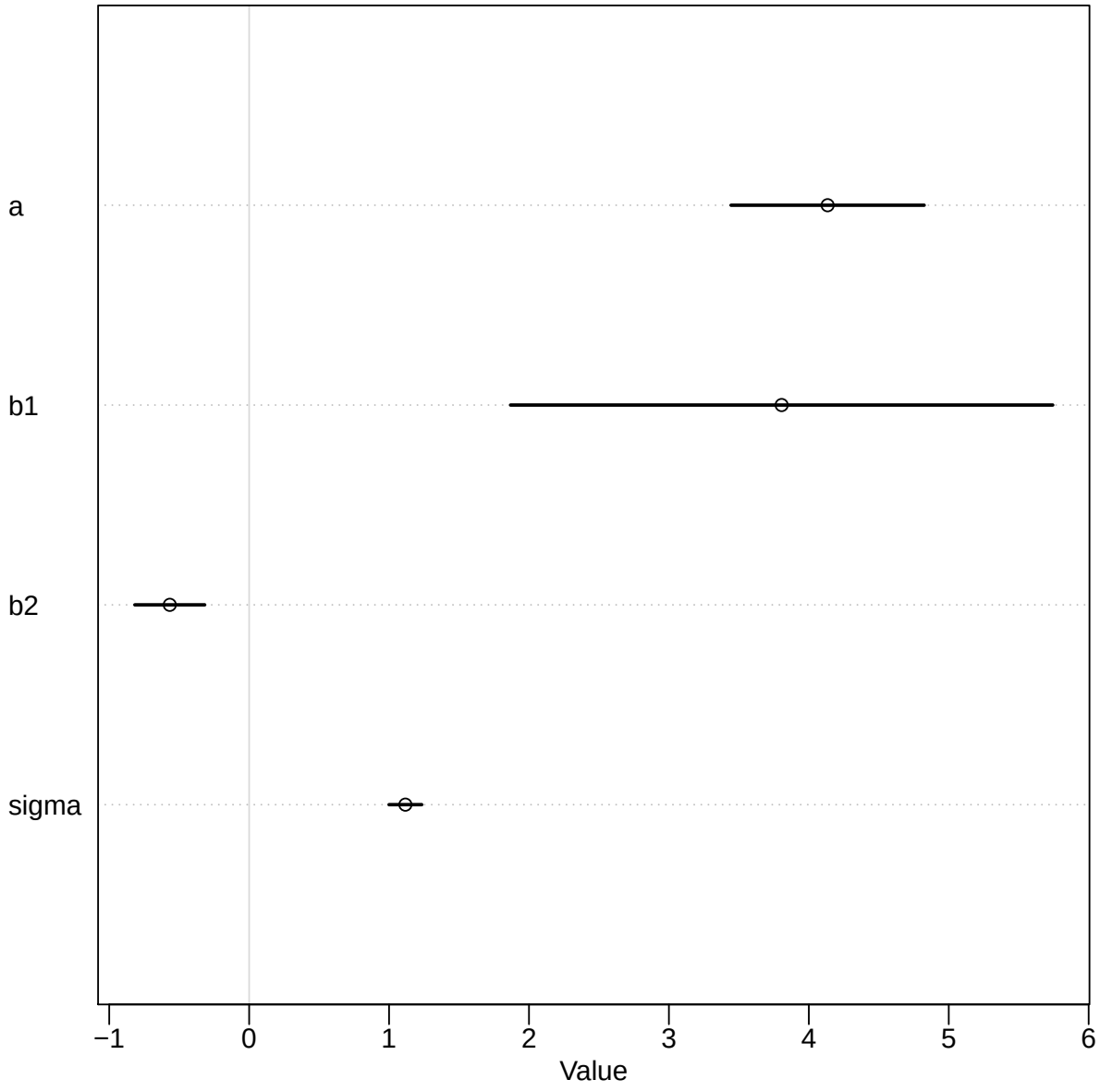
Problem 5H3.
Finally, consider the avgfood variable. Fit two more multiple regressions: (1) body weight as an additive function of avgfood and groupsize, and (2) body weight as an additive function of all three variables, avgfood and groupsize and area. Compare the results of these models

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?

```
model4 <- map(
  alist(
    weight ~ dnorm(mu, sigma),
    mu <- a + b1 * avgfood + b2 * groupsize,
    a ~ dnorm(10, 10),
    b1 ~ dnorm(0, 100),
    b2 ~ dnorm(0, 100),
    sigma ~ dunif(0, 10)
  ), data = d)

precis(model4)
```

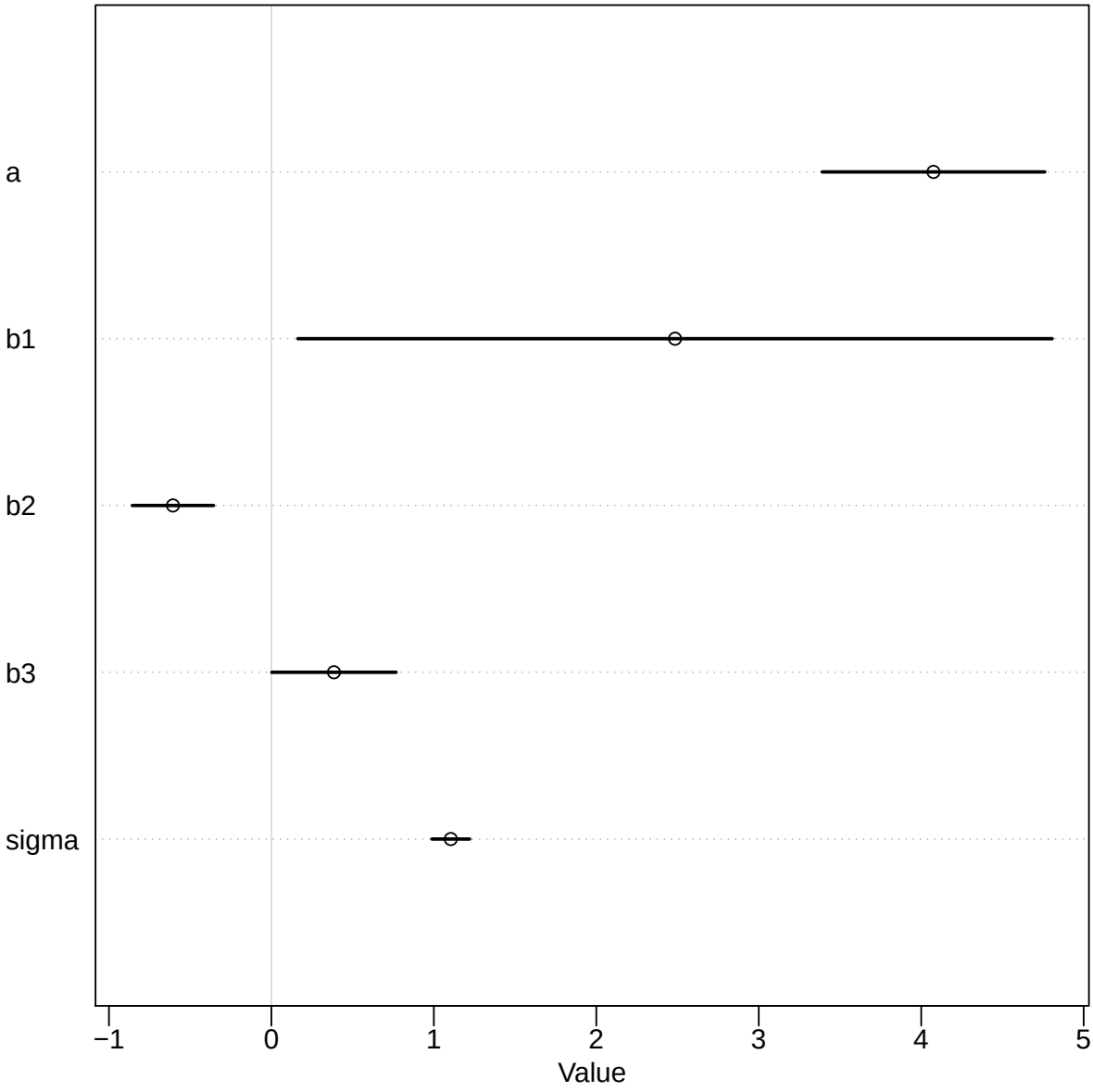
```
##        Mean StdDev  5.5% 94.5%
## a      4.13   0.43  3.44  4.82
## b1     3.81   1.21  1.87  5.74
## b2    -0.57   0.16 -0.82 -0.32
## sigma  1.12   0.07  1.00  1.23
```

```
plot(precis(model4))
```




```
model5 <- map(
  alist(
    weight ~ dnorm(mu, sigma),
     mu <- a + b1 * avgfood + b2 * groupsize + b3 * area,
    a ~ dnorm(10, 10),
    b1 ~ dnorm(0, 100),
    b2 ~ dnorm(0, 100),
    b3 ~ dnorm(0, 100),
    sigma ~ dunif(0, 10)
  ), data = d)
```

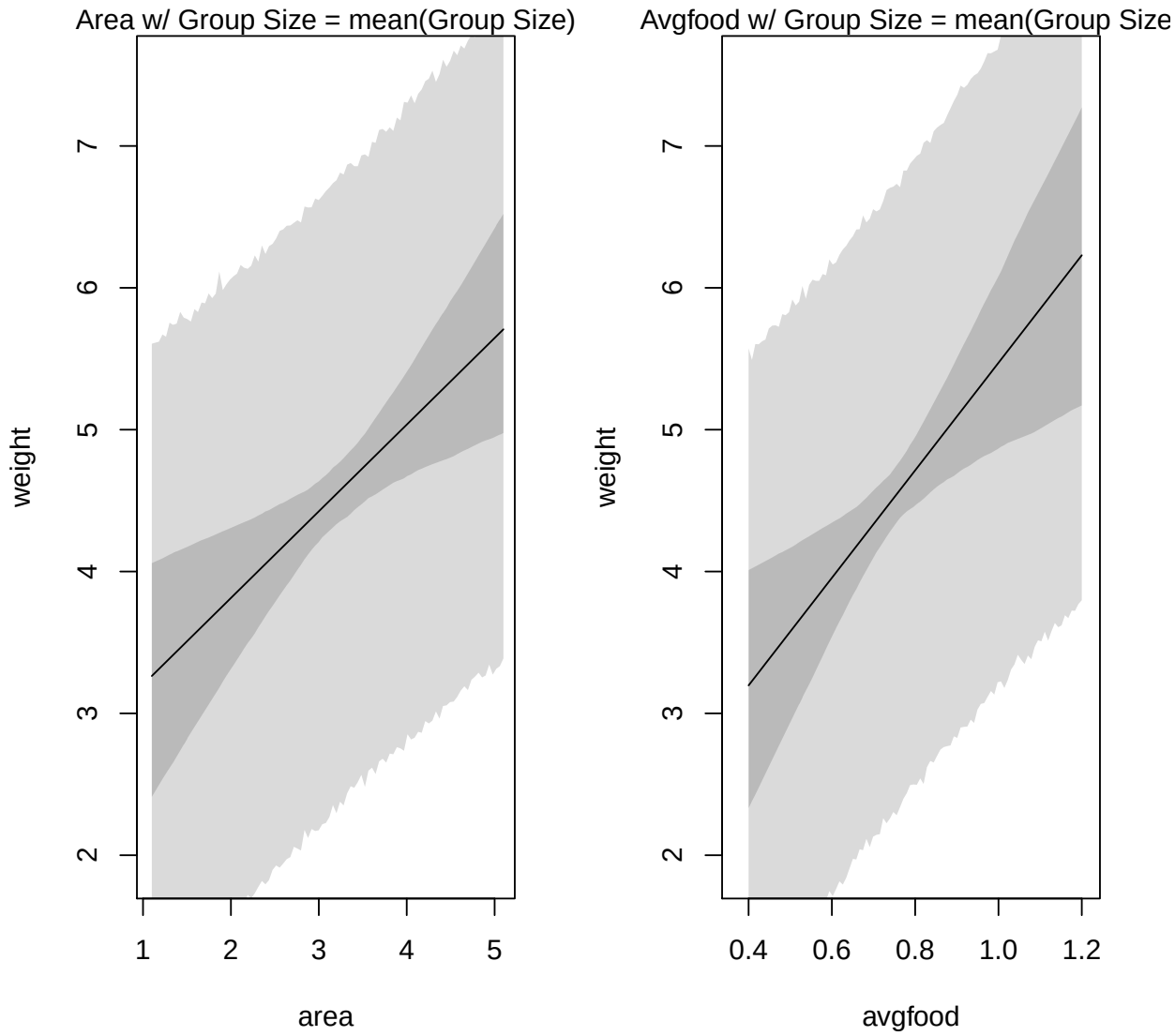
```
precis(model5)

##         Mean StdDev  5.5% 94.5%
## a       4.08   0.43  3.39  4.76
## b1      2.48   1.45  0.16  4.81
## b2     -0.61   0.16 -0.86 -0.36
## b3      0.38   0.24  0.00  0.77
## sigma   1.10   0.07  0.99  1.22

plot(precis(model5))
```



a) Let's compare avgfood and area:

```r
avgfood.seq <- seq(from = round(min(d$avgfood), 1), to = round(max(d$avgfood),1), length

pred.data3 <-data.frame(avgfood = avgfood.seq, groupsize = G.avg)

mu3 <- link(model4, data = pred.data3)

## [ 100 / 1000 ]
[ 200 / 1000 ]
[ 300 / 1000 ]
[ 400 / 1000 ]
[ 500 / 1000 ]
[ 600 / 1000 ]
[ 700 / 1000 ]
[ 800 / 1000 ]
[ 900 / 1000 ]
[ 1000 / 1000 ]


mu3.mean <- apply(mu3, 2, mean)
mu3.PI <- apply(mu3, 2, PI, 0.95)

F.sim <- sim(model4, data = pred.data3, n = 1e4)

## [ 1000 / 10000 ]
[ 2000 / 10000 ]
[ 3000 / 10000 ]
[ 4000 / 10000 ]
[ 5000 / 10000 ]
[ 6000 / 10000 ]
[ 7000 / 10000 ]
[ 8000 / 10000 ]
[ 9000 / 10000 ]
[ 10000 / 10000 ]


F.PI <- apply(F.sim, 2, PI, 0.95)

par(mfrow=c(1,2))

plot(weight ~ area, data = d, type = "n")
mtext("Area w/ Group Size = mean(Group Size)")
lines(area.seq, mu1.mean)
shade(mu1.PI, area.seq)
shade(A.PI, area.seq)
```

```
plot(weight ~ avgfood, data = d, type = "n")
mtext("Avgfood w/ Group Size = mean(Group Size)")
lines(avgfood.seq, mu3.mean)
shade(mu3.PI, avgfood.seq)
shade(F.PI, avgfood.seq)
```



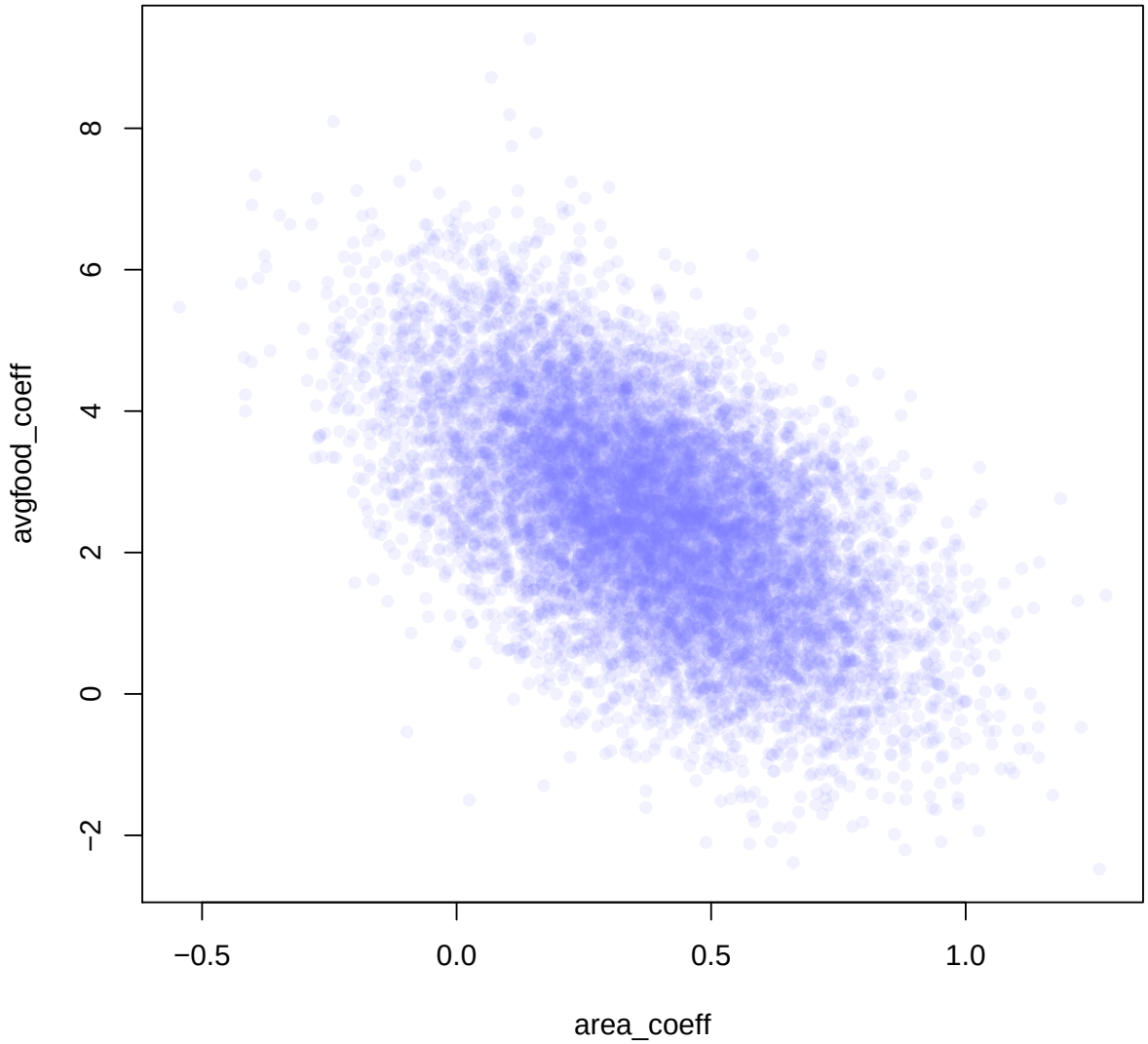
Holding group size constant at its mean, it appears that avgfood has a slightly more positive correlation with weight than area does. As a result, we should use avgfood in place of area.

b) This occurs due to the presence of multicollinearity in our predictors, which we will demonstrate below:

```
print(cor(d$avgfood, d$area))

## [1] 0.8831038
```

We can see that the variables avgfood and area are highly positively correlated. Now to show that the posterior distributions for their coefficients are highly correlated as well:

```
post <- extract.samples(model5)
names(post) <- c("intercept", "avgfood_coeff", "groupsize coeff", "area_coeff", "sigma")
plot(avgfood_coeff ~ area_coeff, post, col=col.alpha(rangi2, 0.1), pch = 16)
```



We can see that the two coefficients are highly negatively correlated. From the correlation

of the variables and coefficients, we can see that both variables contain nearly the same information. Thus, only one should be included in the model.