Rethinking Statistics: Chapter 5

Easy Questions

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) $\mu_i = \alpha + \beta_x x_i + \beta_z z_i$

Answer: 2 and 4
(1) only one predictor (xi), so not a multiple regression
(2) +(4) two predictors: xi and zi
(3) not a multiple regression model, since only the difference of xi and zi enters the model (with slope β)

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.

Answer: $\mu_i = \alpha + \beta_A A_i + \beta_P P_i$
Latitude = outcome
Animal diversity (A)
Plant diversity (P)

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.

Answer: $\mu_i = \alpha + \beta_F F_i + \beta_S S_i$
Amounf of funding (F)
Size of laboratory (S)
Since it says in the question that *"these variables are both positively associated with time to degree"*, the slope parameters for both should be positive

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) $\mu_i = \alpha_A(1 - B_i - C_i - D_i) + \alpha_B B_i + \alpha_C C_i + \alpha_D D_i$

Answer: (1), (3), (4), and (5) are inferentially equivalent because they each allow the computation of each other's posterior distribution

(1) the model includes a single intercept (for category C) and slopes for A, B, and D.
(3) the model includes a single intercept (for category A) and slopes for B, C, and D.
(4) the model uses the unique index approach to provide a separate intercept for each category (and no slopes).
(5) the model uses the reparameterized approach to multiply the intercept for category A times 1 when in category A and times 0 otherwise.

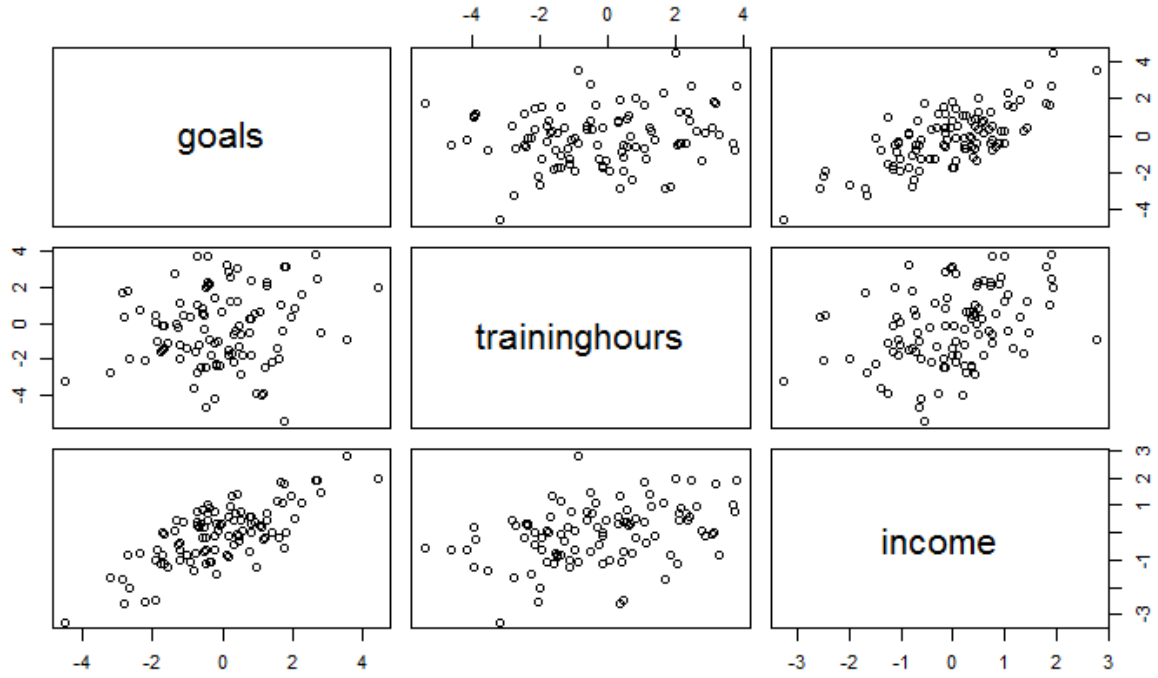
(2) the model is non-identifiable because it includes a slope for all possible categories.

Medium Questions

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).

#5M1
#example spurious correlation between the number hours training and goals during season that vanishes when player income is included
#Use simulation from page 144

```
N <- 100 # number of cases
income <- rnorm(n = 100, mean = 0, sd = 1) # x_real as Gaussian with mean 0 and stddev 1
traininghours <- rnorm(n = N, mean = income, sd = 2) # x_spur as Gaussian with mean=x_real
goals <- rnorm(n = N, mean = income, sd = 1) # y as Gaussian with mean=x_real
d <- data.frame(goals, traininghours, income)# bind all together in data frame
pairs(d)
```



#show that there is a (spurious) association between number of hours training and goals scored.

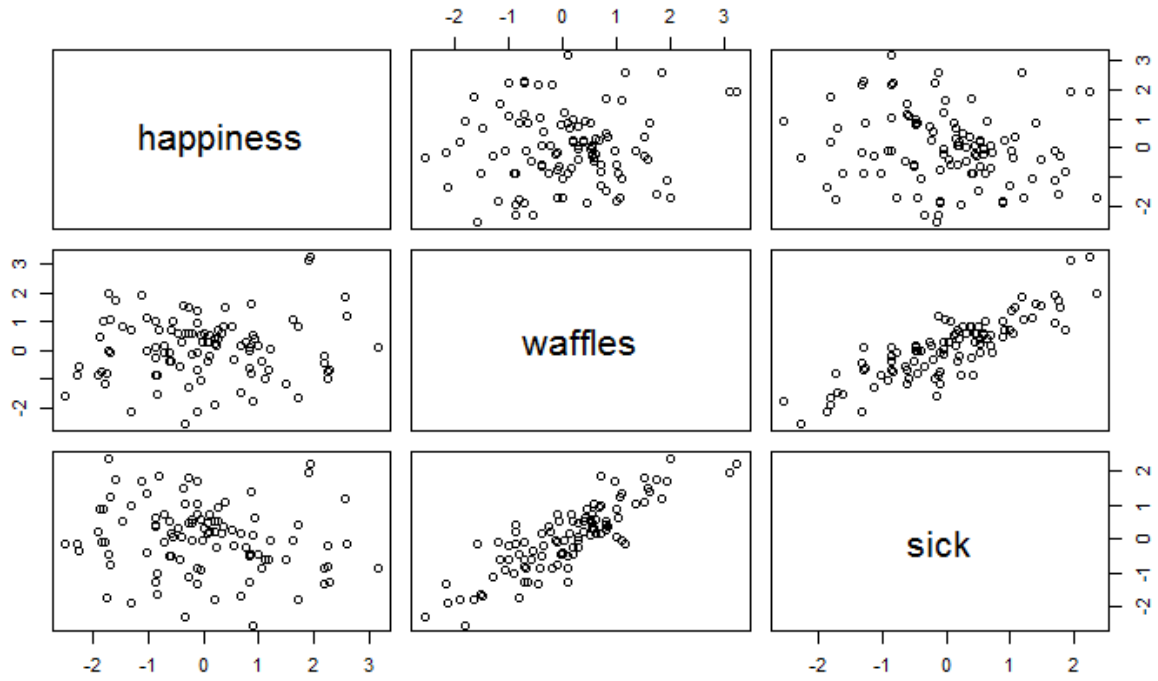
```
m <- quap(
 alist(
   goals ~ dnorm(mu, sigma),
   mu <- a + bo * traininghours,
   a ~ dnorm(0, 5),
   bo ~ dnorm(0, 5),
```

```
   sigma ~ dunif(0, 5)
  ),
  data = d
)
precis(m)

        mean   sd  5.5% 94.5%
a      -0.05 0.15 -0.29  0.19
bo      0.11 0.07 -0.01  0.22
sigma   1.48 0.10  1.31  1.65

#show that this association vanishes when player income is added to the model
m <- quap(
  alist(
    goals ~ dnorm(mu, sigma),
    mu <- a + bo * traininghours + bi * income,
    a ~ dnorm(0, 5),
    bo ~ dnorm(0, 5),
    bi ~ dnorm(0, 5),
    sigma ~ dunif(0, 5)
  ),
  data = d
)
precis(m)

        mean   sd  5.5% 94.5%
a      -0.06 0.10 -0.22  0.10
bo     -0.11 0.05 -0.19 -0.03
bi      1.12 0.10  0.96  1.28
sigma   0.99 0.07  0.88  1.10
```

In this model, player income predicts goals scored during the season when the number training hours is already known, but the number of training hours seen does not predict goals during season when player income is already known. Thus, the bivariate association between goals scored and training hours is spurious.

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.

```
#example of a masked relationship involving the prediction of happiness ratings from the amount of waffles one eats and the amount of time feeling sick due to overeating.
#Use simulation from page 156
N <- 100
rho <- 0.8
waffles <- rnorm(n = N, mean = 0, sd = 1)
sick <- rnorm(n = N, mean = rho * waffles, sd = sqrt(1 - rho^2))
happiness <- rnorm(n = N, mean = waffles - sick, sd = 1)
d <- data.frame(happiness, waffles, sick)
pairs(d)
```




```
# Due to the masking relationship, the bivariate associations should be weak/widely variable.
m <- quap(
 alist(
  happiness ~ dnorm(mu, sigma),
  mu <- a + ba * waffles,
  a ~ dnorm(0, 5),
  ba ~ dnorm(0, 5),
  sigma ~ dunif(0, 5)
 ),
 data = d
)
precis(m)
```

```
        mean   sd  5.5% 94.5%
a      -0.01 0.12 -0.20  0.18
ba      0.10 0.11 -0.08  0.28
sigma   1.22 0.09  1.08  1.35
```

```
m <- quap(
 alist(
  happiness ~ dnorm(mu, sigma),
  mu <- a + bi * sick,
  a ~ dnorm(0, 5),
  bi ~ dnorm(0, 5),
  sigma ~ dunif(0, 5)
 ),
 data = d
```

```
)
precis(m)
```

```
       mean   sd  5.5% 94.5%
a      0.00 0.12 -0.19  0.20
bi    -0.19 0.12 -0.39  0.00
sigma  1.20 0.09  1.07  1.34
```

```
#In the multivariate regression, the masking relationship should be resolved.
m <- quap(
 alist(
  happiness ~ dnorm(mu, sigma),
  mu <- a + ba * waffles + bi * sick,
  a ~ dnorm(0, 5),
  ba ~ dnorm(0, 5),
  bi ~ dnorm(0, 5),
  sigma ~ dunif(0, 5)
 ),
 data = d
)
precis(m)
```

```
       mean   sd  5.5% 94.5%
a     -0.05 0.11 -0.23  0.12
ba     0.92 0.20  0.61  1.23
bi    -1.02 0.21 -1.36 -0.69
sigma  1.09 0.08  0.97  1.21
```

The slopes for waffles and feeling sick became much larger in magnitude in the multivariate model. So, because waffles increases happiness and feeling sick due to overeating, and feelings sick due to overeating decreases happiness, the bivariate relationships of waffles and feeling sick due to overeating to happiness are masked.

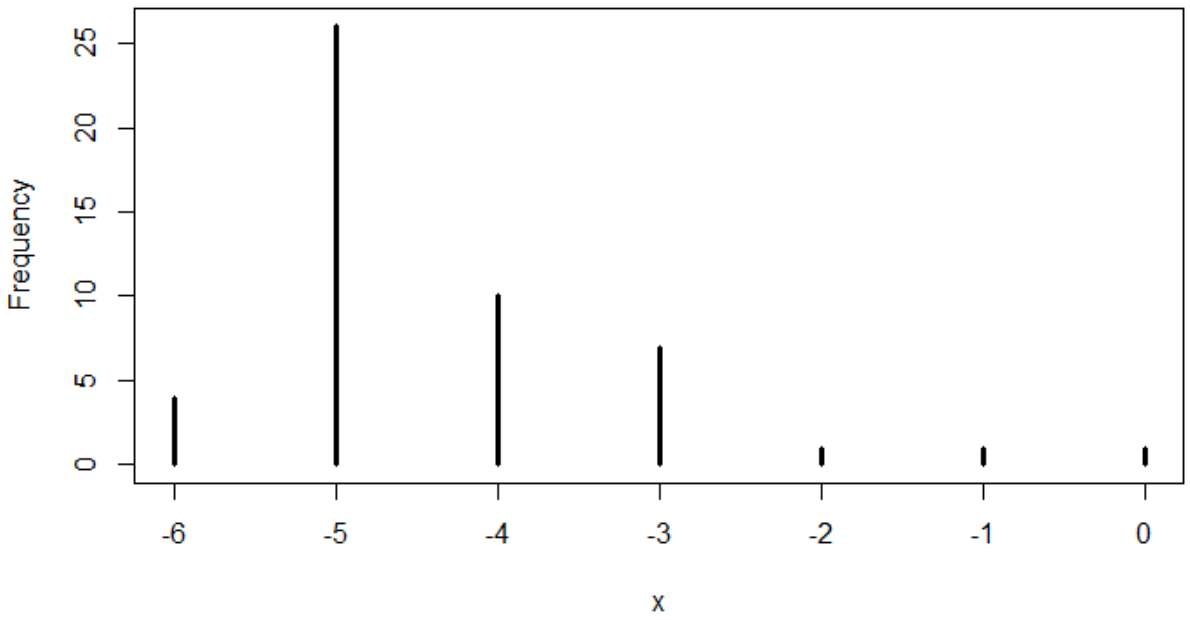
5M3. It is sometimes observed that the best predictor of fire risk is the presence of firefighters— States and localities with many firefighters also have more fires. Presumably firefighters do not *cause* fires. Nevertheless, this is not a spurious correlation. Instead fires cause firefighters. Consider 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?

A high divorce rate might cause a higher marriage rate by introducing more unmarried individuals into the dating pool. Who in turn then want to get married again. This could be evaluated using multiple regression by regressing marriage rate on both divorce rate and re-marriage rate (i.e., the rate of non-first marriages or marriages following divorces). If divorce rate no longer predicts marriage rate even when the re-marriage rate is known, this would support the hypothesis.

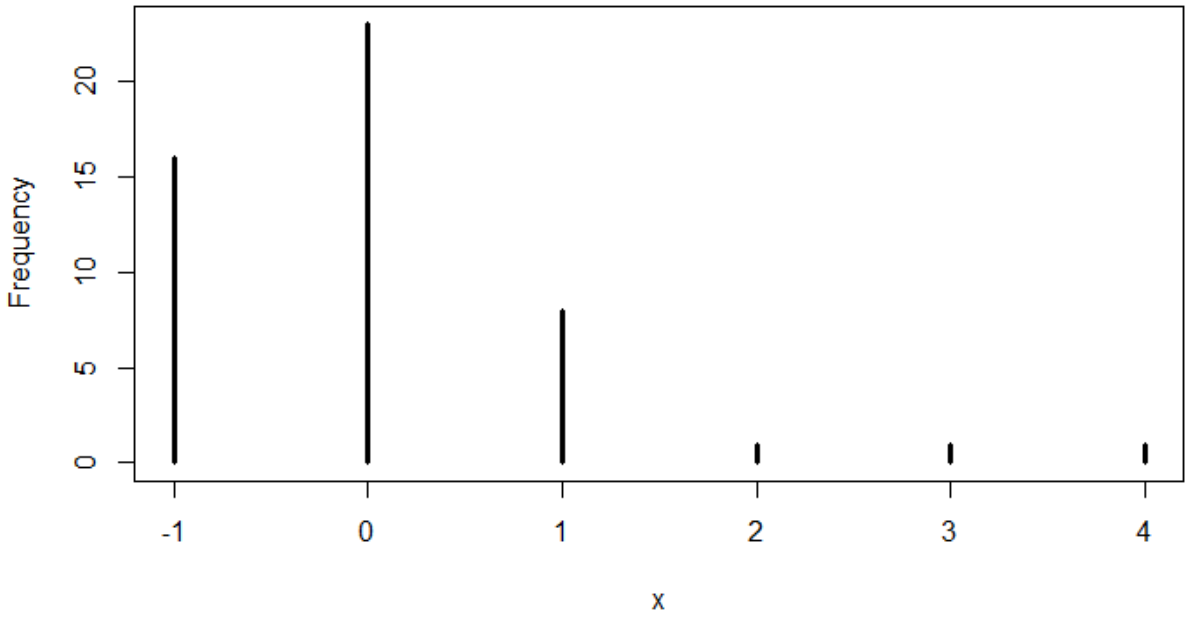
5M4. In the divorce data, States with high numbers of Mormons (members of The Church of Jesus Christ of Latter-day Saints, LDS) have much lower divorce rates than the regression models expected. Find a list of LDS population by State and use those numbers as a predictor variable, predicting divorce rate 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.

```
#Add new column in WaffleDivorce data frame with percent Mormons per State
data(WaffleDivorce)
d <- WaffleDivorce
d$LDS <- c(0.0077, 0.0453, 0.0610, 0.0104, 0.0194, 0.0270, 0.0044, 0.0057, 0.0041, 0.0075, 0.0082, 0.0520, 0.2623,
0.0045, 0.0067, 0.0090, 0.0130, 0.0079, 0.0064, 0.0082, 0.0072, 0.0040, 0.0045, 0.0059, 0.0073, 0.0116, 0.0480,
0.0130, 0.0065, 0.0037, 0.0333, 0.0041, 0.0084, 0.0149, 0.0053, 0.0122, 0.0372, 0.0040, 0.0039, 0.0081, 0.0122,
0.0076, 0.0125, 0.6739, 0.0074, 0.0113, 0.0390, 0.0093, 0.0046, 0.1161)
d$logLDS <- log(d$LDS)
d$logLDS.s <- (d$logLDS - mean(d$logLDS)) / sd(d$logLDS)
simplehist(d$LDS)
```


```
#log transform these data
simplehist(d$logLDS)
```



#standardize
simplehist(d$logLDS.s)

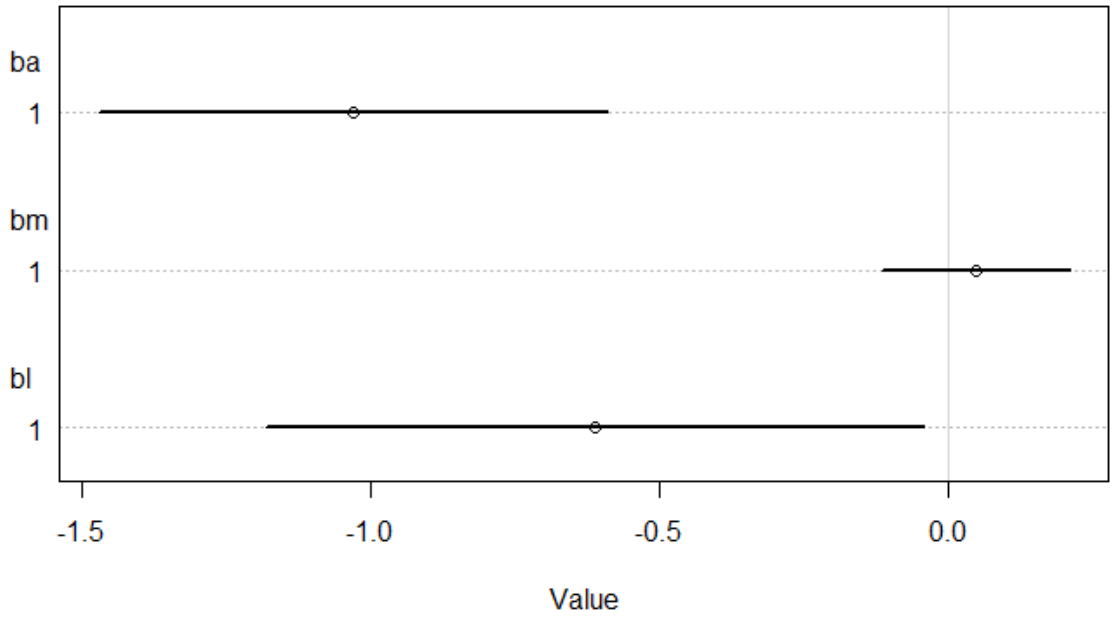


#multiple regression model.

```
m <- quap(
 alist(
  Divorce ~ dnorm(mu, sigma),
  mu <- a + bm * Marriage + ba * MedianAgeMarriage + bl * logLDS.s,
  a ~ dnorm(10, 20),
  bm ~ dnorm(0, 10),
  ba ~ dnorm(0, 10),
  bl ~ dnorm(0, 10),
  sigma ~ dunif(0, 5)
 ),
 data = d
)
precis(m)
```

```
       mean   sd  5.5% 94.5%
a     35.46 6.77 24.63 46.28
bm     0.05 0.08 -0.08  0.19
ba    -1.03 0.22 -1.39 -0.67
bl    -0.61 0.29 -1.07 -0.14
sigma  1.38 0.14  1.16  1.60
```

```
#plot posterior distributions
plot( coeftab(m), par=c("ba","bm", "bl") )
```



Conclusion: The slope of marriage was widely variable and its interval includes zero. However, the slopes of both median age at marriage and percentage of LDS population were negative and their intervals did not include zero. Thus, states with older median age at marriage or higher percentages of Mormons had lower divorce rates.

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.

Option 1:
Mechanism

(1) Variable for calories burnt while exercising
(2) Variable for calories ingested at restaurant

Model

$\mu_i=\alpha+\beta_G G_i+\beta_B B_i+\beta_R R_i$

Price of gasoline (G)
Calories burnt (B)
Calories ingested at restaurant (R)

Option 2:
Mechanism

(1)Variable for time spent exercising (minutes/week)
(2)Variable for frequency of eating at restaurant (times/week)

Model

$\mu_i=\alpha+\beta_G G_i+\beta_E E_i+\beta_R R_i$

Price of gasoline (G)
Time spent exercising (E)
Eating at restaurant (R)

Hard Questions

All three 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?

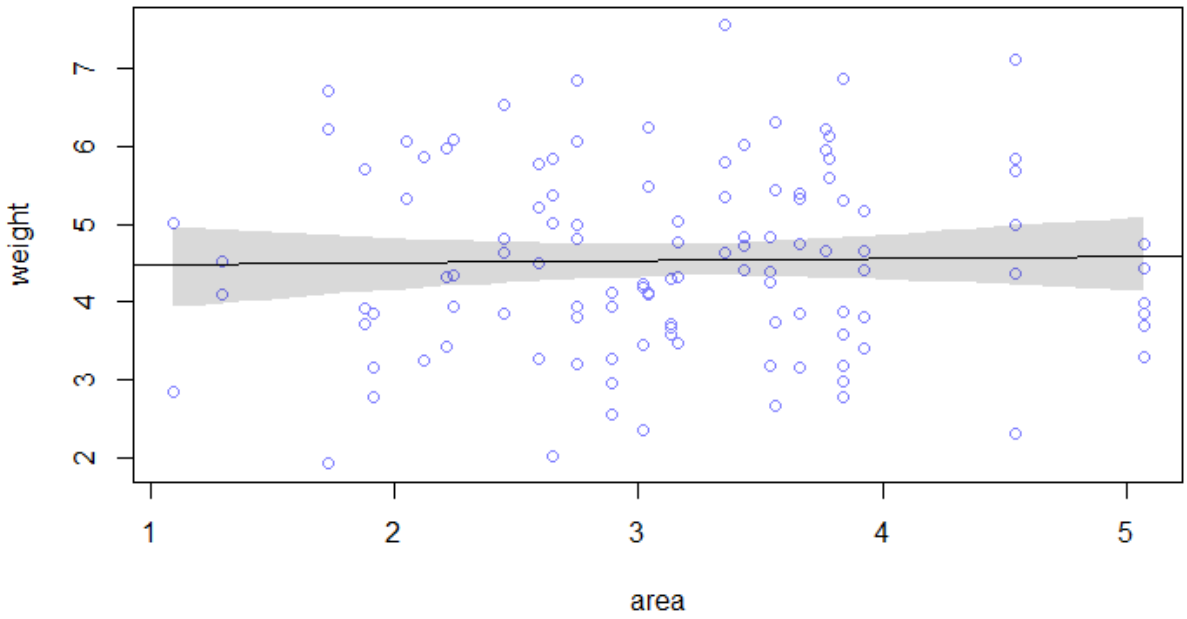
OPTION 1: body weight as a linear function of territory size

```
#use model on p.139
d <- foxes
ma <- quap(
```

```
 alist(
   weight ~ dnorm(mu, sigma),
   mu <- a + ba * area,
   a ~ dnorm(5, 5),
   ba ~ dnorm(0, 5),
   sigma ~ dunif(0, 5)
 ),
 data = d
)
precis(ma)
```

```
mean    sd  5.5% 94.5%
a      4.45 0.39  3.83  5.08
ba     0.02 0.12 -0.16  0.21
sigma 1.18 0.08  1.06  1.30
```

```
#plot regression + MAP regression line + the 95% interval of the mean
area.seq <- seq(from = min(d$area), to = max(d$area), length.out = 30)
mu <- link(ma, data = data.frame(area = area.seq))

mu.PI <- apply(mu, 2, PI, prob = 0.95)
plot(weight ~ area, data = d, col = rangi2)
abline(ma)
shade(mu.PI, area.seq)
```



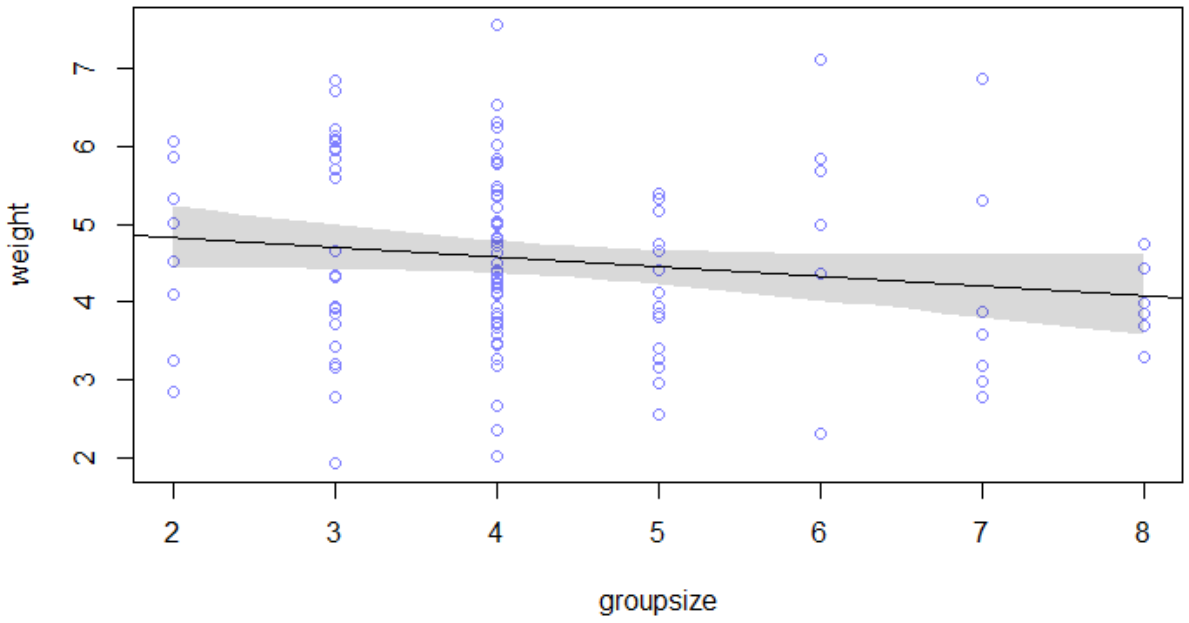
OPTION 2: body weight as a linear function of group size

```
mg <- quap(
 alist(
```

```
  weight ~ dnorm(mu, sigma),
  mu <- a + bg * groupsize,
  a ~ dnorm(5, 5),
  bg ~ dnorm(0, 5),
  sigma ~ dunif(0, 5)
 ),
 data = d
)
precis(mg)
```

```
        mean   sd  5.5% 94.5%
a       5.07 0.32  4.55  5.59
bg     -0.12 0.07 -0.24 -0.01
sigma  1.16 0.08  1.04  1.29
```

```
#plot regression + MAP regression line + the 95% interval of the mean
groupsize.seq <- seq(from = min(d$groupsize), to = max(d$groupsize), length.out = 30)
mu <- link(mg, data = data.frame(groupsize = groupsize.seq))
mu.PI <- apply(mu, 2, PI, prob = 0.95)
plot(weight ~ groupsize, data = d, col = rangi2)
abline(mg)
shade(mu.PI, groupsize.seq)
```



Conclusion: Group size may have slightly negative relationship with body weight but very small. So it seems that group size and territory area are not important for the prediction of body weight in foxes.

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?

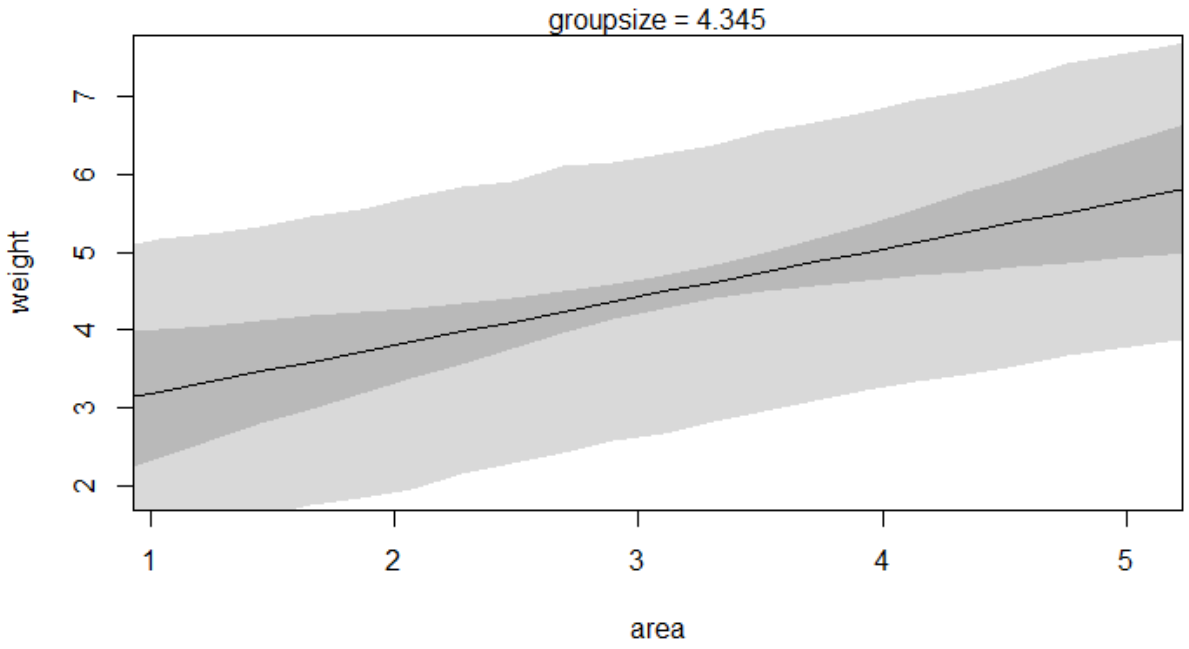
First start with regression model

```
mag <- quap(
 alist(
  weight ~ dnorm(mu, sigma),
  mu <- a + ba * area + bg * groupsize,
  a ~ dnorm(5, 5),
  ba ~ dnorm(0, 5),
  bg ~ dnorm(0, 5),
  sigma ~ dunif(0, 5)
 ),
 data = d
)
precis(mag)
```

```
       mean   sd  5.5% 94.5%
a      4.45 0.37  3.86  5.05
ba     0.62 0.20  0.30  0.93
bg    -0.43 0.12 -0.62 -0.24
sigma  1.12 0.07  1.00  1.24
```

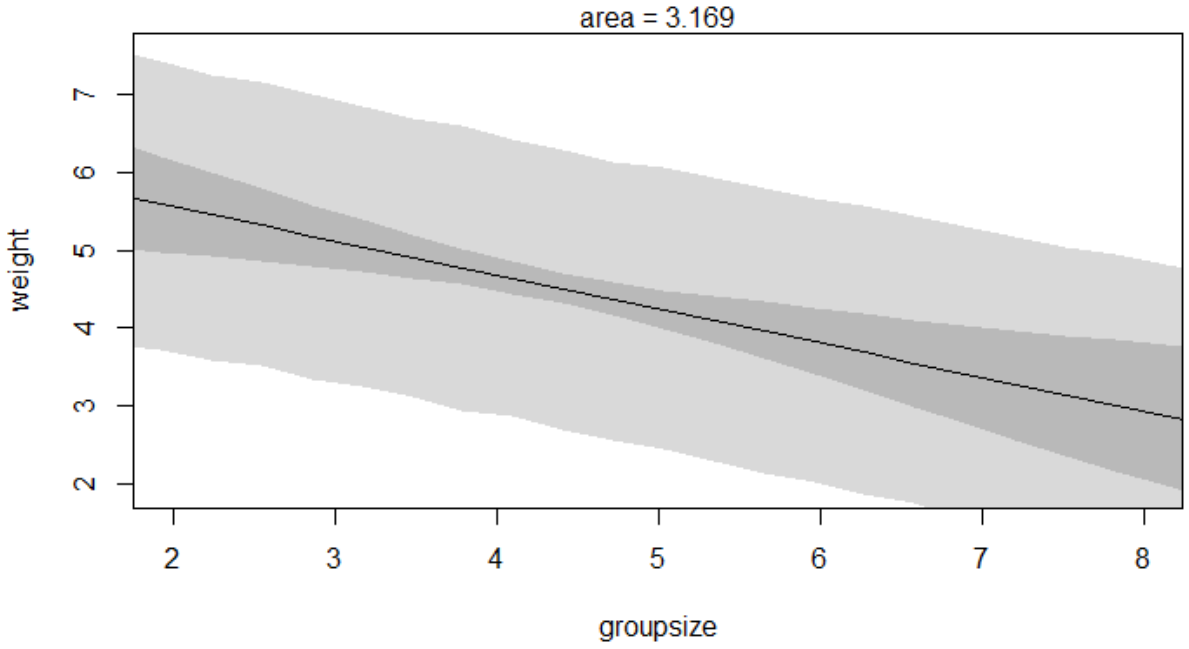
Counterfactual plot for territory area model (p.145)

```
G.avg <- mean(d$groupsize)
A.seq <- seq(from = 0, to = 6, length.out = 30)
pred.data <- data.frame(
 groupsize = G.avg,
 area = A.seq
)
mu <- link(mag, data = pred.data)

mu.mean <- apply(mu, 2, mean)
mu.PI <- apply(mu, 2, PI, prob = 0.95)
A.sim <- sim(mag, data = pred.data, n = 1e4)

A.PI <- apply(A.sim, 2, PI)
plot(weight ~ area, data = d, type = "n")
mtext("groupsize = 4.345")
lines(A.seq, mu.mean)
shade(mu.PI, A.seq)
shade(A.PI, A.seq)
```



Counterfactual plot for group size

```
A.avg <- mean(d$area)
G.seq <- seq(from = 1, to = 10, length.out = 30)
pred.data <- data.frame(
  groupsize = G.seq,
  area = A.avg
)
mu <- link(mag, data = pred.data)

mu.mean <- apply(mu, 2, mean)
mu.PI <- apply(mu, 2, PI, prob = 0.95)
G.sim <- sim(mag, data = pred.data, n = 1e4)

G.PI <- apply(G.sim, 2, PI)
plot(weight ~ groupsize, data = d, type = "n")
mtext("area = 3.169")
lines(G.seq, mu.mean)
shade(mu.PI, G.seq)
shade(G.PI, G.seq)
```



Conclusion: The results from the multiple regression model differ from the bivariate models. It says that territory area is positively related to body weight and that group size is negatively related to body weight. The results differ because this is an example of a masking relationship. Territory area is positively related to body weight and group size is negatively related to body weight, but these effects get cancelled out in the bivariate regressions because territory area and group size are positively related.

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?

MODEL 1: body weight as an additive function of avgfood and groupsize

```
m <- quap(
 alist(
  weight ~ dnorm(mu, sigma),
  mu <- a + bf * avgfood + bg * groupsize,
  a ~ dnorm(5, 5),
  bf ~ dnorm(0, 5),
  bg ~ dnorm(0, 5),
  sigma ~ dunif(0, 5)
 ),
 data = d
)
precis(m)
```

```
mean   sd  5.5% 94.5%
a      4.18 0.43  3.50  4.86
bf     3.60 1.18  1.72  5.48
bg    -0.54 0.15 -0.79 -0.30
sigma  1.12 0.07  1.00  1.23
```

Model 2: body weight as an additive function of all three variables (avgfood, groupsize and area)

```
m <- quap(
 alist(
  weight ~ dnorm(mu, sigma),
  mu <- a + bf * avgfood + bg * groupsize + ba * area,
  a ~ dnorm(5, 5),
  bf ~ dnorm(0, 5),
  bg ~ dnorm(0, 5),
  ba ~ dnorm(0, 5),
  sigma ~ dunif(0, 5)
 ),
 data = d
)
precis(m)
```

```
     mean   sd  5.5% 94.5%
a      4.10 0.42  3.42  4.78
bf     2.30 1.39  0.08  4.53
bg    -0.59 0.15 -0.84 -0.35
ba     0.40 0.24  0.02  0.78
sigma  1.10 0.07  0.99  1.22
```

Compare results of both models to previous models:

Model 1 shows a large positive relationship between average food and body weight and a negative relationship between group size and body weight. The estimate for the group size slope is similar to that in the model with both territory area and group size. Thus, the masking effect on group size was overcome by including the average food variable. Perhaps average food and territory size are positively correlated (to have similar effects).

Model 2 is similar to the previous one in terms of the group size slope, this is again the masking relationship problem. However, the estimates of the average food and territory area slopes have decreased in magnitude compared to previous models. This is likely due to multicollinearity between the two variables.

Question 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.

Average food and territory area are both good predictors of body weight. They are very highly correlated (positively) with each other and both have a similar relationships with body weight. Therefore it would be better to select only one of them. The predictor with the larger standardized slope estimate.

#calculate standard slope estimate for average food

```
d$avgfood.s <- (d$avgfood - mean(d$avgfood)) / sd(d$avgfood)
m <- quap(
 alist(
  weight ~ dnorm(mu, sigma),
  mu <- a + bf * avgfood.s + bg * groupsize,
  a ~ dnorm(5, 5),
  bf ~ dnorm(0, 5),
  bg ~ dnorm(0, 5),
  sigma ~ dunif(0, 5)
 ),
 data = d
)
precis(m)
```

```
       mean   sd  5.5% 94.5%
a      6.95 0.68  5.86  8.04
bf     0.74 0.24  0.36  1.12
bg    -0.56 0.15 -0.80 -0.31
sigma  1.12 0.07  1.00  1.23
```

```
#calculate standard slope estimate for territory area
d$area.s <- (d$area - mean(d$area)) / sd(d$area)
m <- quap(
 alist(
  weight ~ dnorm(mu, sigma),
  mu <- a + ba * area.s + bg * groupsize,
  a ~ dnorm(5, 5),
  ba ~ dnorm(0, 5),
  bg ~ dnorm(0, 5),
  sigma ~ dunif(0, 5)
 ),
 data = d
)
precis(m)
```

```
mean   sd  5.5% 94.5%
a      6.39 0.53  5.54  7.24
ba     0.57 0.18  0.27  0.86
bg    -0.43 0.12 -0.62 -0.24
sigma  1.12 0.07  1.00  1.24
```

Conclusion: average food would be preferable to territory area

Question b: When both average food and territory area are in the same model, their effects are reduced and their standard errors are larger than when they are in separate models due to multicollinearity between average food and territory area. The two variables are highly correlated and have very similar relationships with body weight. Thus, the partial effect of each becomes smaller (when controlling for the other).