Chapter 4 Exercises

Emily Maloney

January 28, 2019

Chapter 4

```
library(tidyverse)
library(brms)
library(tidybayes)
```

Easy Problems

4E1

In this model, the likelihood is defined by $y\ i \sim \text{Normal}(\mu, \sigma)$.

4E2

There are 2 parameters in the posterior distribution of this model.

4E3

omit

4E4

The line describing the linear model is $\mu\ i = \alpha + \beta\ x\ i$.

4E5

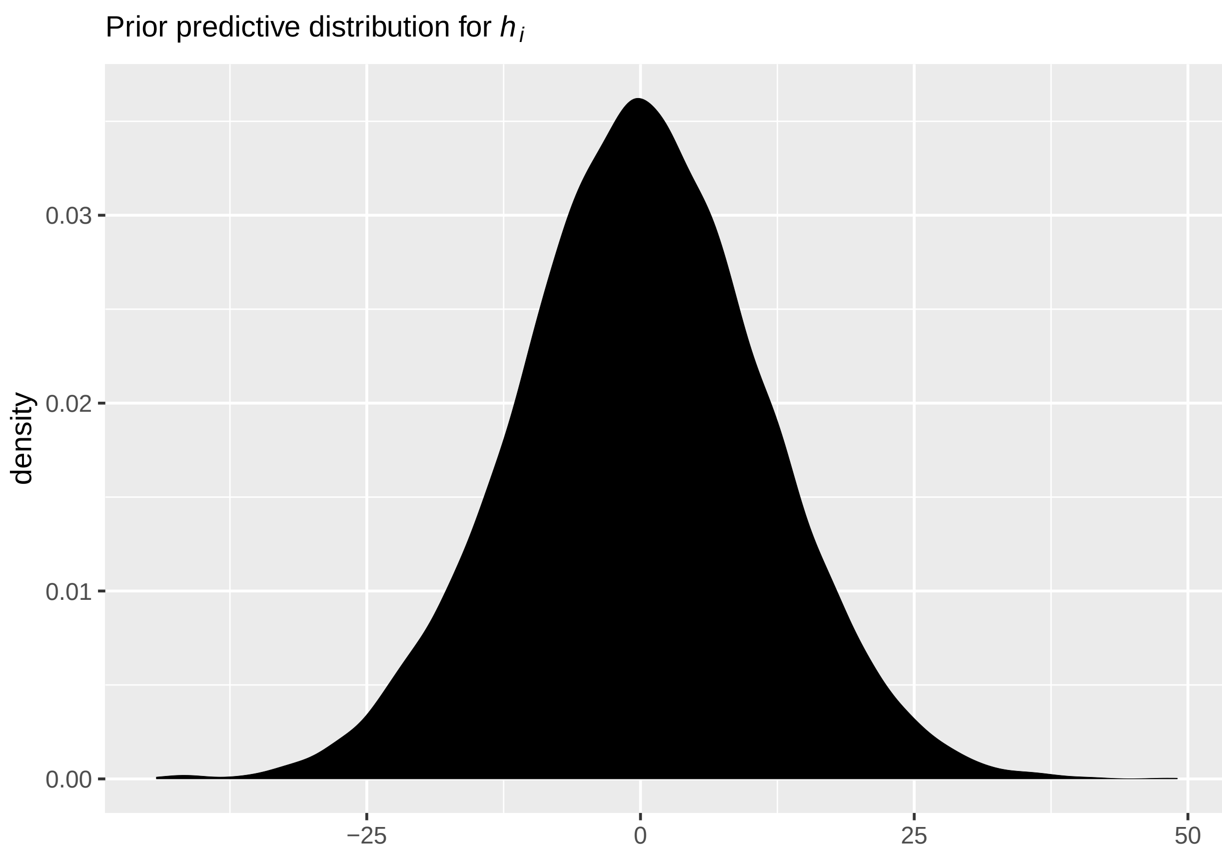
There are 3 parameters in the posterior distribution of this model.

Medium Problems

4M1

```
#sampling from both priors to get simulation of observed heights
n <- 1e4
set.seed(432)
tibble(sample_mu = rnorm(n, mean = 0, sd = 10),
       sample_sigma = runif(n, min = 0, max = 10)) %>%
       mutate(x = rnorm(n, mean = sample_mu, sd = sample_sigma)) %>%

       ggplot(aes(x = x)) +
       geom_density(fill = "black", size = 0) +
       labs(subtitle = expression(paste("Prior predictive distribution for ", italic(h[i]))),
            x = NULL)
```



Simulating observed heights from the prior information results in a distribution centered at 0, which is expected, given that the prior specification for μ is a normal distribution with a mean of 0.

4M2

The model translated into a map formula is:
flist <- alist (
y ~ dnorm(mu, sigma),
mu ~ dnorm(0, 10),
sigma ~ dunif(0, 10)
)

4M3

The map model formula translated into a mathematical model definition is:
$y\ i$ ~ Normal(μ, σ)
$\mu\ i = \alpha + \beta\ x\ i$
α ~ Normal(0, 50)
β ~ Uniform(0, 10)
σ ~ Uniform(0, 50)

4M4

The mathematical model definitions for predicting height using year as a predictor I would use is:
$y\ i$ ~ Normal(μ, σ)
$\mu\ i = \alpha + \beta\ x\ i$

$\alpha \sim$ Normal(107, 10)
$\beta \sim$ Normal(7, 2)
$\sigma \sim$ Uniform(0, 25)

For the specification of alpha, I was assuming that the students would be kindergarteners during the first year of observation, so I guessed that the mean would be around 3.5 feet, which is ~107 centimeters, and a standard deviation of 10 centimeters, because that allows for a fair amount of variation in heights of kindergarteners. Considering that children grow fairly quickly, I then decided that the prior for beta should have a mean of 3 inches, which is around 7 centimeters, and a standard deviation of 1, because most kids grow at fairly similar rates at that yonug age. Finally, to specify the prior for sigma, I did not have strong assumptions or knowledge about what sigma overall would be, so I specified a uniform distribution going from 0 to 25.

4M5

In this situation, I would not change my prior, because this information is the actual data and will instead be used to fit the model with the priors as they are.

4M6

In this case, this new information is additional prior information and not data from the sample, so I would change the prior specification for σ (standard deviation) to $\sigma \sim$ Uniform(0, 8).

Hard Problems

4H1

```r
library(rethinking)
library(tidyverse)
library(knitr)
data(Howell1) # load in data

d <- Howell1
#d2 <- d %>% filter(age >= 18) # filter to only adults

#fit model
mhw.1 <- rethinking::map(alist(
             height ~ dnorm(mu, sigma),
             mu <- a + b*weight,
             a ~ dnorm(156, 100),
             b ~ dnorm(0, 10),
             sigma ~ dunif(0, 50)
            ),
         data = d)
precis(mhw.1)
```

```
##        Mean StdDev  5.5% 94.5%
## a     75.45   1.05 73.77 77.12
## b      1.76   0.03  1.72  1.81
## sigma  9.35   0.28  8.89  9.80
```

```r
N <- 1e4 # sample size

# Get predictive means and data
preds <-
  as_tibble(MASS::mvrnorm(mu = mhw.1@coef,
```

```
                          Sigma = mhw.1@vcov , n = N )) %>%      # rather than extract.samples
  mutate(weight = sample(c(46.95, 43.72, 64.78, 32.59, 54.63), N, replace = T),
         predmean = a + b * weight ,                              # line uncertainty
         predverb = rnorm(N, a + b*weight, sigma )) %>%           # data uncertainty
  group_by(weight) %>%
  mutate(lb_mu = rethinking::HPDI(predmean, prob = .89)[1],
         ub_mu = rethinking::HPDI(predmean, prob = .89)[2],
         lb_ht = rethinking::HPDI(predverb, prob = .89)[1],
         ub_ht = rethinking::HPDI(predverb, prob = .89)[2]) %>%
  slice(1) %>%
  mutate(yhat = mhw.1@coef["a"] + mhw.1@coef["b"] * weight) %>%       # yhat for reg line
  select(weight, yhat,lb_ht, ub_ht)

kable(preds, type = "pandoc", caption = "!Kung Predicted Heights")
```

Table 1: !Kung Predicted Heights

weight	yhat	lb_ht	ub_ht
32.59	132.9363	117.5893	147.0679
43.72	152.5704	137.6683	167.1974
46.95	158.2684	143.2099	173.5524
54.63	171.8164	156.3160	186.5088
64.78	189.7218	174.6773	204.7158

Using the model's specification of $\beta = 1.76$, $\alpha = 75.44$, and $\sigma = 9.35$, the expected heights and 89% intervals for these individuals were produced by simulating from the posterior distribution of the model and are shown in the table above.

4H2

a)

```
#filter data to only children
d3 <- d %>% filter(age < 18)

#fit model
mhw.2 <- rethinking::map(alist(
            height ~ dnorm(mu, sigma),
            mu <- a + b*weight,
            a ~ dnorm(156, 100),
            b ~ dnorm(0, 10),
            sigma ~ dunif(0, 50)
           ),
        data = d3)

#summary call for what's in the model
precis(mhw.2)
```

```
##        Mean StdDev  5.5% 94.5%
## a     58.25   1.40 56.02 60.48
## b      2.72   0.07  2.61  2.83
## sigma  8.43   0.43  7.75  9.12
```

For every 10 units increase in weight, the model predicts that a child will get 27.2 cm taller.

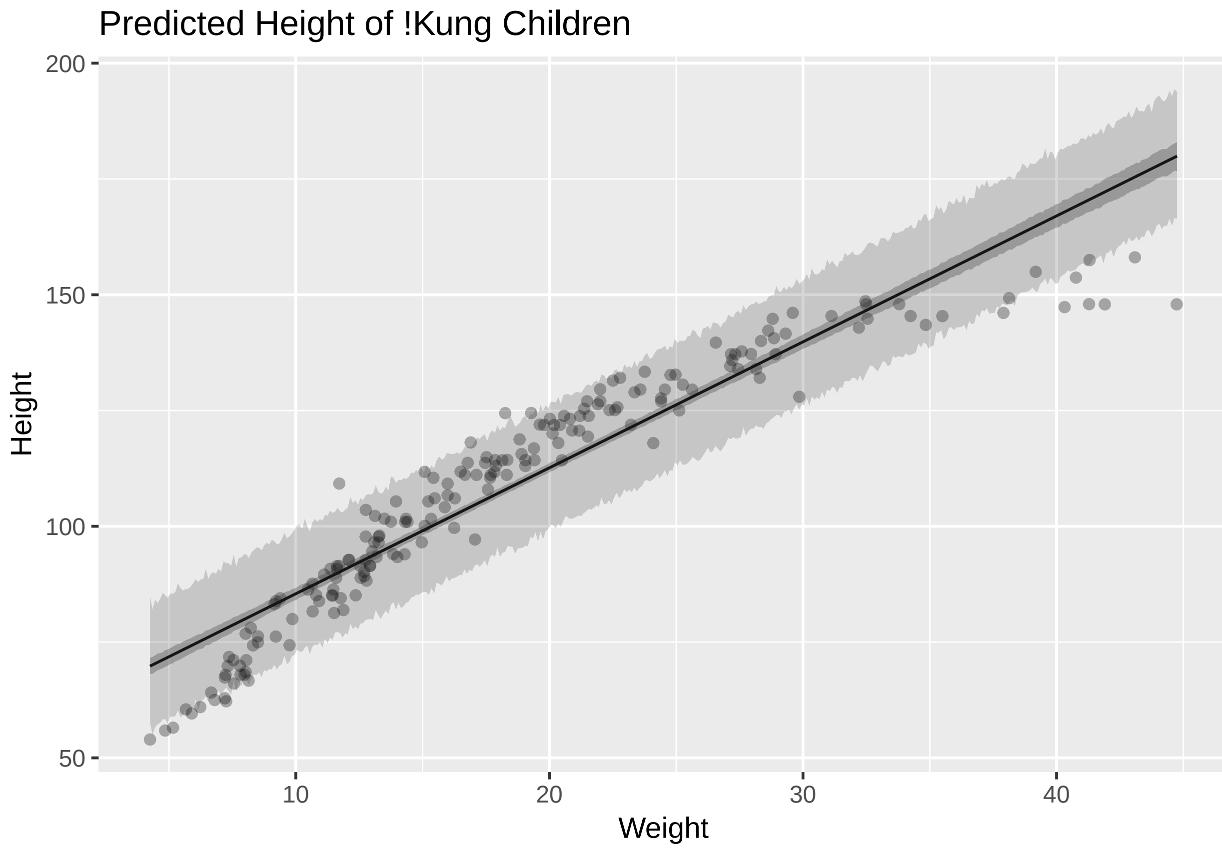
b)

```
N <- 1e6 # sample size

# Get predictive means and data
preds <-
  as.tibble(MASS::mvrnorm(mu = mhw.2@coef,
                          Sigma = mhw.2@vcov , n = N )) %>%      # rather than extract.samples
  mutate(weight = sample(seq(from = 4.25, to = 44.75, by = 0.1), N, replace = T),
         predmean = a + b * weight ,                              # line uncertainty
         predverb = rnorm(N, a + b*weight, sigma )) %>%           # data uncertainty
  group_by(weight) %>%
  mutate(lb_mu = rethinking::HPDI(predmean, prob = .89)[1],
         ub_mu = rethinking::HPDI(predmean, prob = .89)[2],
         lb_ht = rethinking::HPDI(predverb, prob = .89)[1],
         ub_ht = rethinking::HPDI(predverb, prob = .89)[2]) %>%
  slice(1) %>%
  mutate(yhat = mhw.2@coef["a"] + mhw.2@coef["b"] * weight) %>%      # yhat for reg line
  select(weight, yhat, lb_mu, ub_mu, lb_ht, ub_ht)
```

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

```
#plot
ggplot(d3, aes(x = weight)) +
  geom_jitter(aes(y = height), 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_ht, ymax = ub_ht), alpha = .2) +
  labs(x = "Weight",
       y = "Height",
       title = "Predicted Height of !Kung Children")
```



c) The most concerning aspect of model fit is that a good bit of the data at the highest and lowest weights are not included in the 89% HPDI, and most of the data at middle weights seem to be falling above the MAP regression line although still in the 89% HPDI. Overall, the shape looks more curvilinear than linear, so I hypothesize that adding a squared weight term may result in a better fitting model.

4H3

```
#add variable of log weight
d <- d%>% mutate(logweight = log(weight))

#fit model
mhw.3 <- rethinking::map(alist(
             height ~ dnorm(mu, sigma),
             mu <- a + b*logweight,
             a ~ dnorm(178, 100),
             b ~ dnorm(0, 10),
             sigma ~ dunif(0, 50)
            ),
         data = d)

#summary call for what's in the model
precis(mhw.3)
```

```
##         Mean StdDev   5.5%  94.5%
## a     -23.55   1.33 -25.69 -21.42
## b      47.01   0.38  46.40  47.62
```

```
## sigma   5.13   0.16   4.89   5.38
```

```
47.01*log(101/100)
```

```
## [1] 0.4677651
```

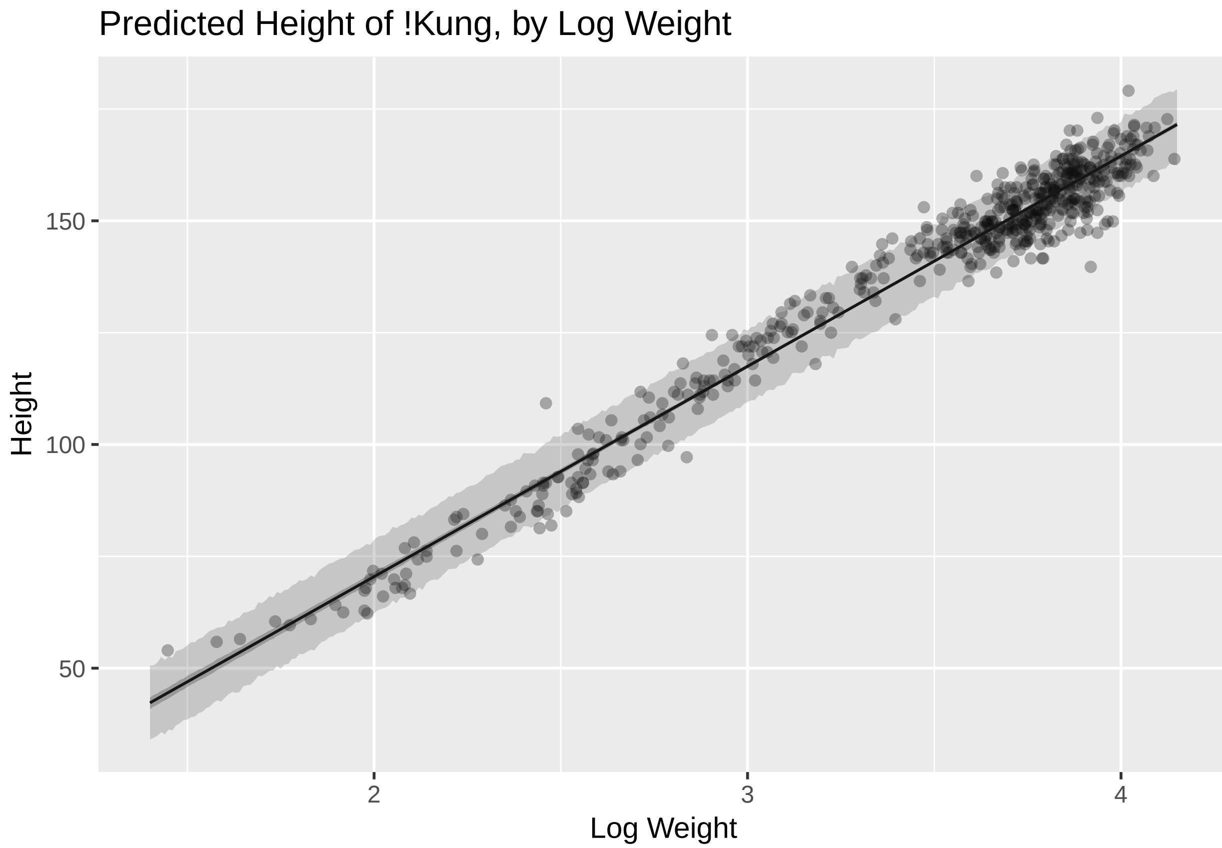
For every 1% increase in weight, we expect a 0.468 centimeter increase in height.

b)

```
N <- 1e6 # sample size

# Get predictive means and data
preds <-
  as.tibble(MASS::mvrnorm(mu = mhw.3@coef,
                          Sigma = mhw.3@vcov , n = N )) %>%      # rather than extract.samples
  mutate(logweight = sample(seq(from = 1.4, to = 4.15, by = 0.01), N, replace = T),
         predmean = a + b * logweight,                           # line uncertainty
         predverb = rnorm(N, a + b*logweight, sigma )) %>%        # data uncertainty
  group_by(logweight) %>%
  mutate(lb_mu = rethinking::HPDI(predmean, prob = .89)[1],
         ub_mu = rethinking::HPDI(predmean, prob = .89)[2],
         lb_ht = rethinking::HPDI(predverb, prob = .89)[1],
         ub_ht = rethinking::HPDI(predverb, prob = .89)[2]) %>%
  slice(1) %>%
  mutate(yhat = mhw.3@coef["a"] + mhw.3@coef["b"] * logweight) %>%       # yhat for reg line
  select(logweight, yhat, lb_mu, ub_mu, lb_ht, ub_ht)

#plot
ggplot(data = d, aes(x = logweight)) +
  geom_jitter(aes(y = height), 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_ht, ymax = ub_ht), alpha = .2) +
  labs(x = "Log Weight",
      y = "Height",
      title = "Predicted Height of !Kung, by Log Weight")
```



This plot of the model's MAP regression line and 89% HPDI interval with the actual data superimposed on top looks like a better fit than the previous model, considering that now the vast majority of the data points fall inside the 89% HPDI and appear to follow the regression line to a greater extent.