HW5

Alexandra Gibbons

2022-09-30

SR Chapter 4

4E1

The likelihood is the first line, $y_i \sim \text{Normal}(\mu, \sigma)$

4E2

There are two parameters, mu and sigma.

4E3

$\Pr(\mu, \sigma \mid y_i) = \frac{\Pi_i \text{Normal}(y_i|\mu,\sigma)\text{Normal}(\mu|0,10)\text{Exponential}(\sigma|1)}{\int\int \Pi_i \text{Normal}(y_i|\mu,\sigma)\text{Normal}(\mu|0,10)\text{Exponential}(\sigma|1) d\mu d\sigma}$

4E4

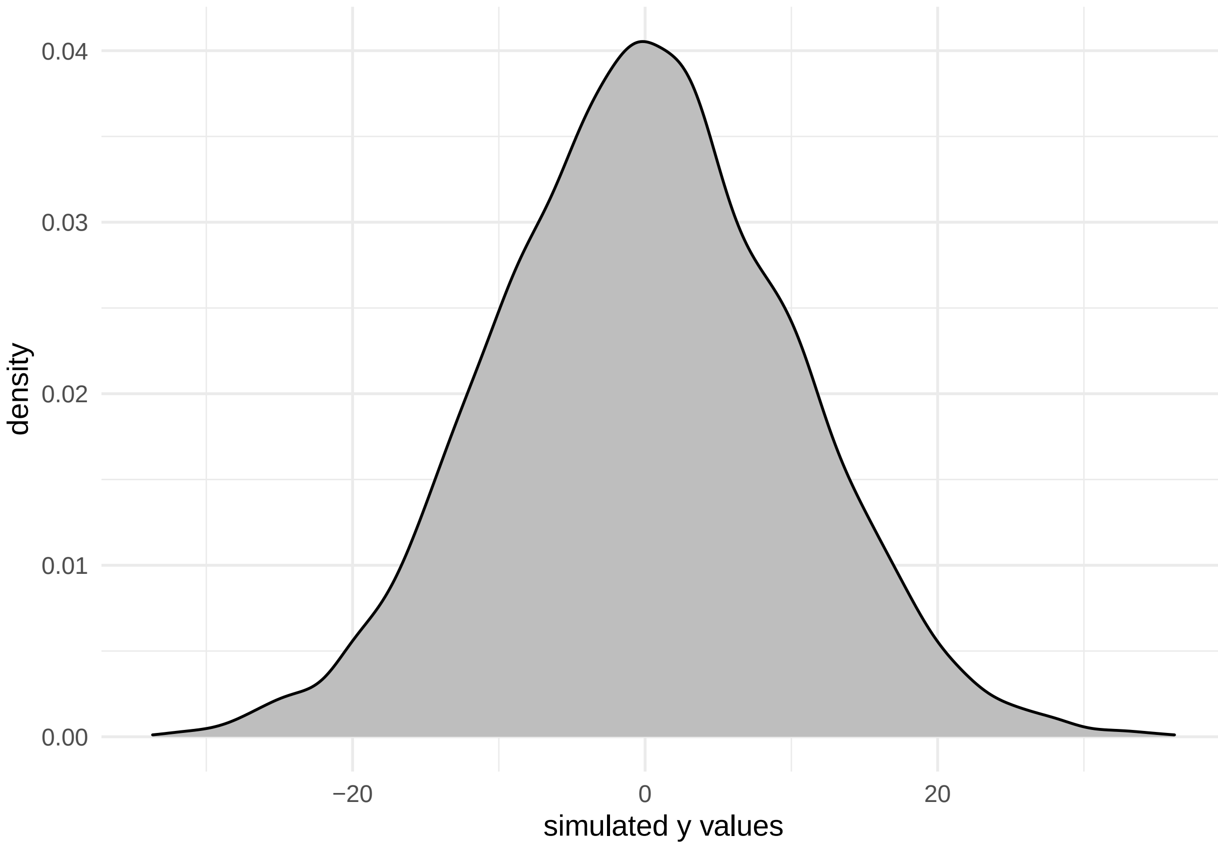
The linear model is the second line, $\mu_i = \alpha + \beta x_i$

4E5

There are four parameters in the posterior distribution: mu, sigma, beta, and alpha.

4M1

```
n <- 1e4 # want 10k samples

set.seed(13)

sim_m1 <- tibble(sample_mu = rnorm(n, mean=0, sd=10),
                 sample_sigma = rexp(n, rate=1)) %>%
  mutate(sim_y = rnorm(n, mean=sample_mu, sd=sample_sigma))

sim_m1 %>% ggplot(aes(x=sim_y)) +
  geom_density(fill="grey") +
  theme_minimal() +
  xlab("simulated y values")
```



4M2

```
flist <- alist(
  y ~ dnorm(mu, sigma),
  mu ~ dnorm(0,10),
  sigma ~ dexp(1))
```

4M3

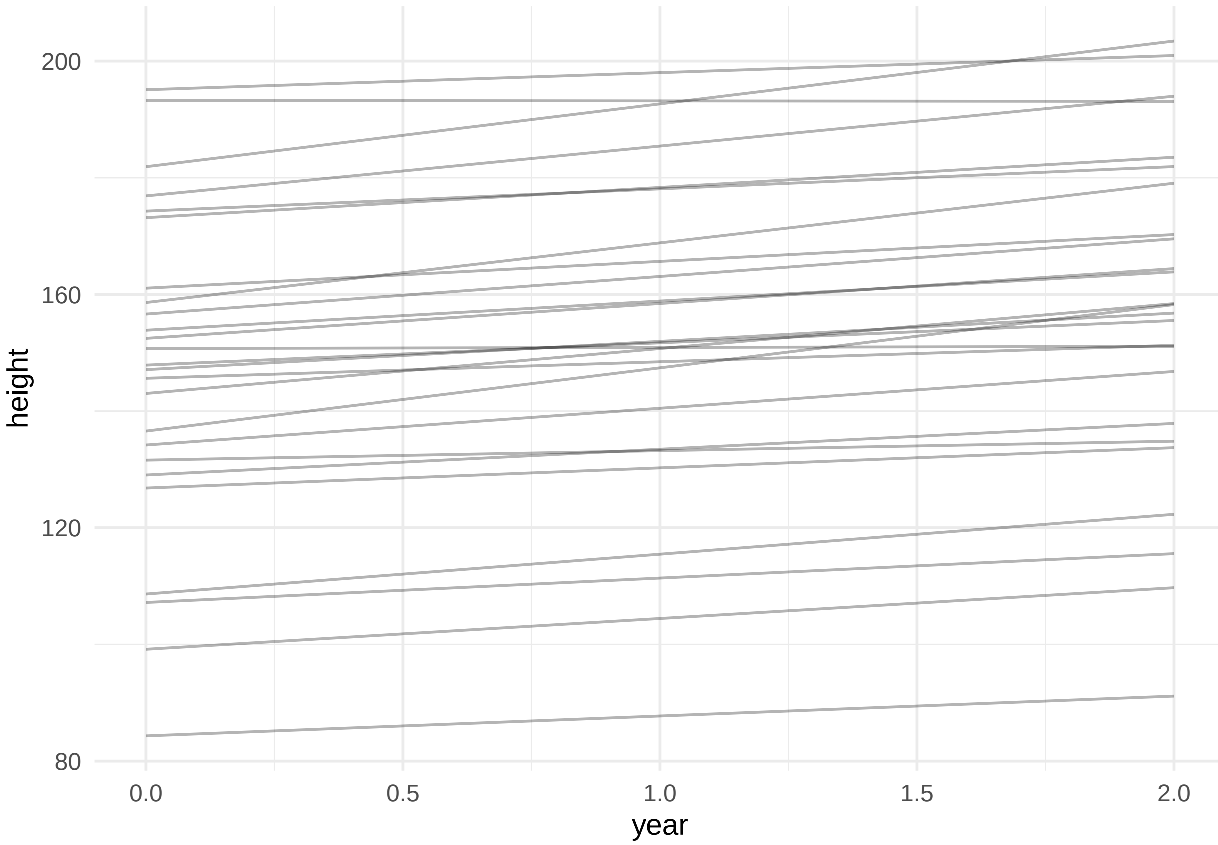
$$
\begin{aligned}
y &\sim \text{Normal}(\mu_i, \sigma) \\
\mu_i &= \alpha + \beta x_i \\
\alpha &\sim \text{Normal}(0, 10) \\
\beta &\sim \text{Uniform}(0, 1) \\
\sigma &\sim \text{Exponential}(1)
\end{aligned}
$$

4M4

$H_i \sim$ Normal (μ_i, σ)
$\mu_i = \alpha + \beta year_i$
$\alpha \sim$ Normal$(140, 30)$ Chose this mean for alpha because it is the average-ish height in cm for 10 year olds. Alpha represents expected height at year 0 because my years are coded as 0, 1, 2.
$\beta \sim$ Normal$(6, 3)$ Kids grow about 6cm/year, and beta represents growth rate per year.
$\sigma \sim$ Exponential(1) Unassuming, positive because sd must be a positive number.

Let's see what the prior predictive might look like.

```r
set.seed(13)

years <- c(0, 1, 2) # 3 years but starting with 0

n_lines <- 25

lines_m4 <- tibble(n=1:n_lines,
                   a = rnorm(n_lines, 140, 30),
                   b = rnorm(n_lines, 6, 3)) %>%
  expand(nesting(n, a, b), year=range(years)) %>%
  mutate(height= a + b * year)

lines_m4 %>%
  ggplot(aes(x = year, y = height, group = n)) +
  geom_line(alpha = .3) +
  theme_minimal()
```

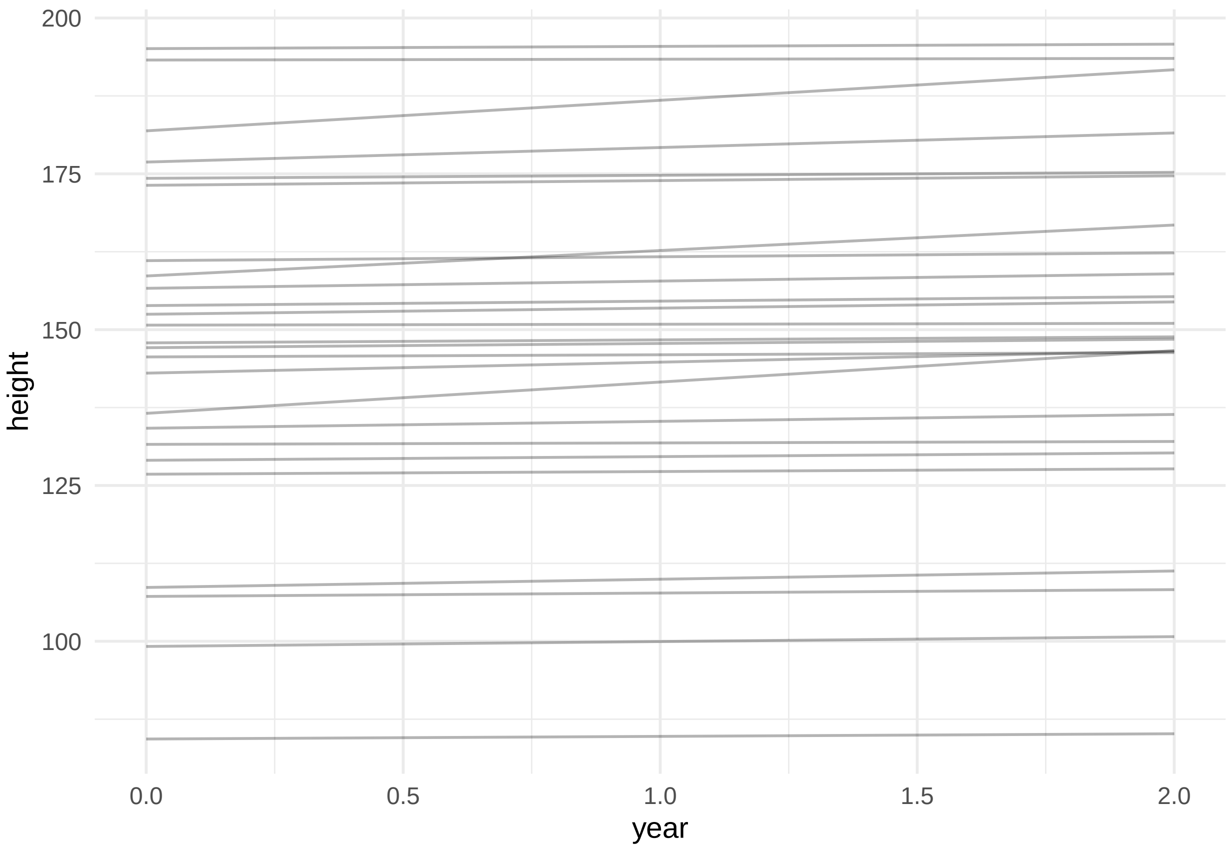


4M5

With this additional information, the only prior that I will change is beta. I will instead use a log-normal prior to ensure that beta is positive. $H_i \sim \text{Normal}\ (\mu_i, \sigma)$
$\mu_i = \alpha + \beta year_i$
$\alpha \sim \text{Normal}(140, 30)$
$\beta \sim \text{LogNormal}(0, 1)$
$\sigma \sim \text{Exponential}(1)$

Visualizing again to ensure slopes (aka growth rates) are all positive.

```
set.seed(13)

years <- c(0, 1, 2) # 3 years but starting with 0

n_lines <- 25

lines_m5 <- tibble(n=1:n_lines,
                   a = rnorm(n_lines, 140, 30),
                   b = rlnorm(n_lines, 0, 1)) %>%
  expand(nesting(n, a, b), year=range(years)) %>%
  mutate(height= a + b * year)

lines_m5 %>%
  ggplot(aes(x = year, y = height, group = n)) +
  geom_line(alpha = .3) +
  theme_minimal()
```




```
min(lines_m5$b)
```

```
## [1] 0.1317243
```

Yep, all of the slopes are positive (checked manually as well to be sure).

4M6

The variance is equal to the the standard deviation, σ, squared. So, if the maximum variance is 64cm, then the maximum standard deviation in 8cm. I adjust the prior for σ accordingly below. $H_i \sim \text{Normal}\ (\mu_i, \sigma)$

$\mu_i = \alpha + \beta year_i$
$\alpha \sim \text{Normal}(140, 30)$
$\beta \sim \text{LogNormal}(0, 1)$
$\sigma \sim \text{Uniform}(0, 8)$

4M7

```
data("Howell1")
set.seed(13)

d <- Howell1 %>%
  filter(age>=18)

m7flist <- alist(
  height ~ dnorm(mu, sigma),
  mu <- a + b*weight,
  a ~ dnorm(178, 20),
  b ~ dlnorm(0, 1),
  sigma ~ dunif(0, 50))

m4.3 <- quap(m7flist, data=d)

precis(m4.3)
```

```
##              mean         sd        5.5%       94.5%
## a     114.5427830 1.89758848 111.5100701 117.5754959
## b       0.8905365 0.04175451   0.8238047   0.9572682
## sigma   5.0722515 0.19120529   4.7666686   5.3778345
```

```
vcov(m4.3, 3)
```

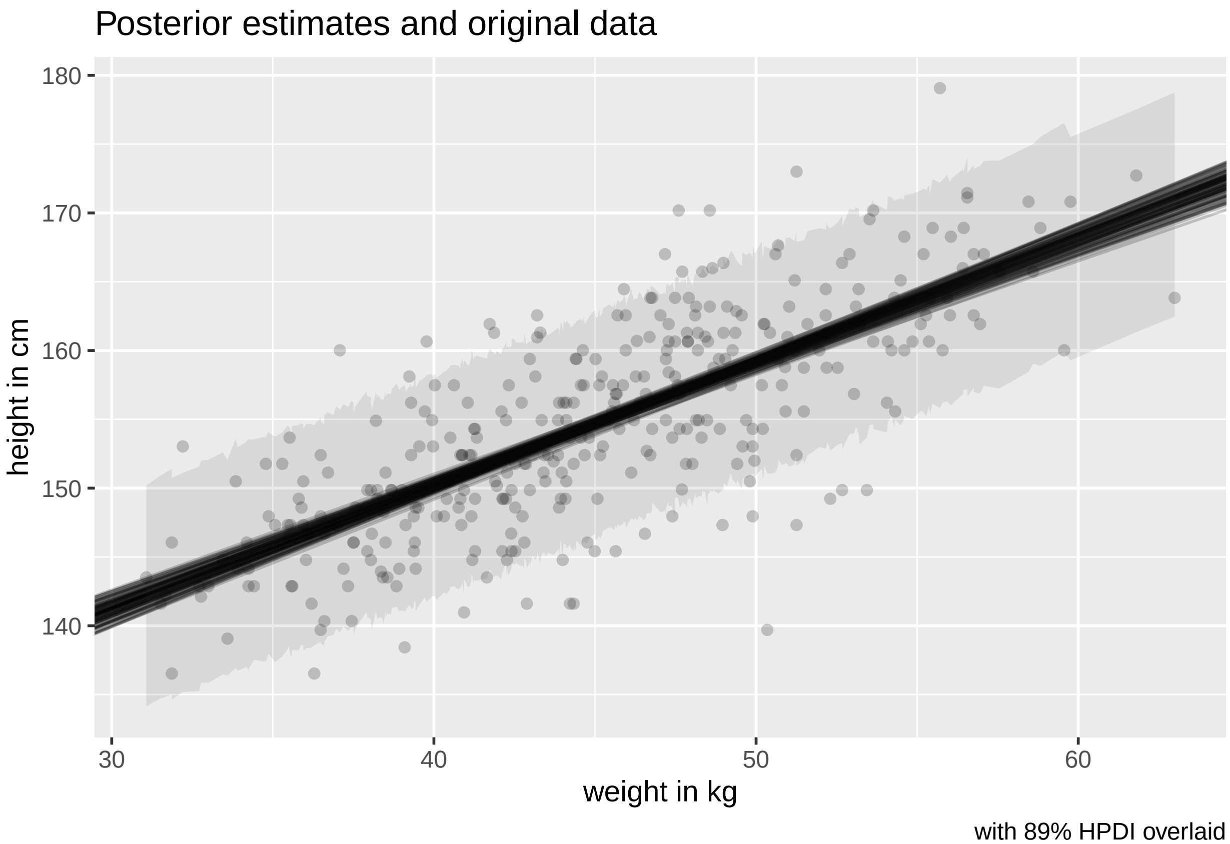
```
##                  a             b         sigma
## a      3.600842055 -0.0784247491  0.0094746644
## b     -0.078424749  0.0017434387 -0.0002069919
## sigma  0.009474664 -0.0002069919  0.0365594646
```

The covariance among parameters is much larger (well, in absolute value, the correlation is negative) when we use a raw measure of weight instead of a centered weight, as expected.

Next, let's look at the posterior prediction.

```
set.seed(13)

draws <- tidy_draws(m4.3, n=100)

p <- ggplot(draws) +
  geom_abline(aes(intercept = a,
                  slope = b),
              alpha = .2) +
  geom_point(
    data = d,
    mapping = aes(x = weight,
                  y = height),
    alpha = .2) +
  labs(x = "weight in kg",
```

```
       y = "height in cm",
       title = "Posterior estimates and original data")

sim <- predicted_draws(m4.3,
                       newdata = d,
                       draws=1000)

sim <- sim %>%
  group_by(.row) %>%
  mutate(lo_bound = HPDI(.prediction)[1],
         up_bound = HPDI(.prediction)[2])

p + geom_ribbon(data = sim,
                mapping = aes(
                  x = weight,
                  ymax = up_bound,
                  ymin = lo_bound),
                alpha = .1) +
  labs(caption = "with 89% HPDI overlaid")
```

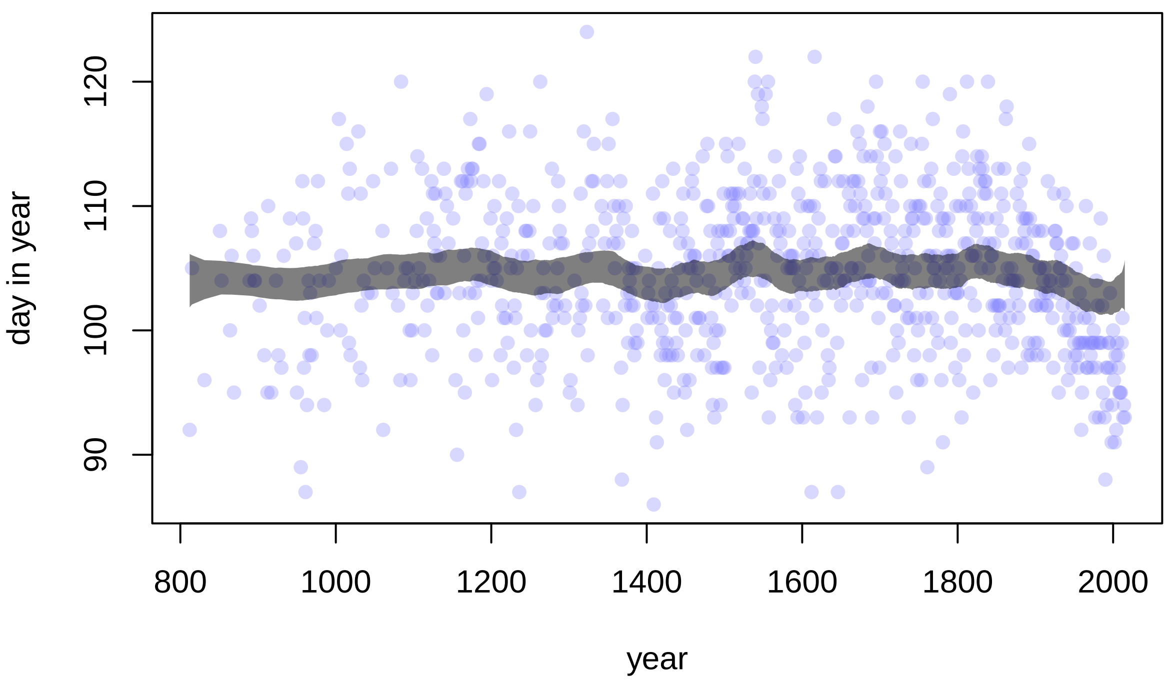


The posterior distributions look very similar for both models (this one and the one in the book), which is good.

4M8

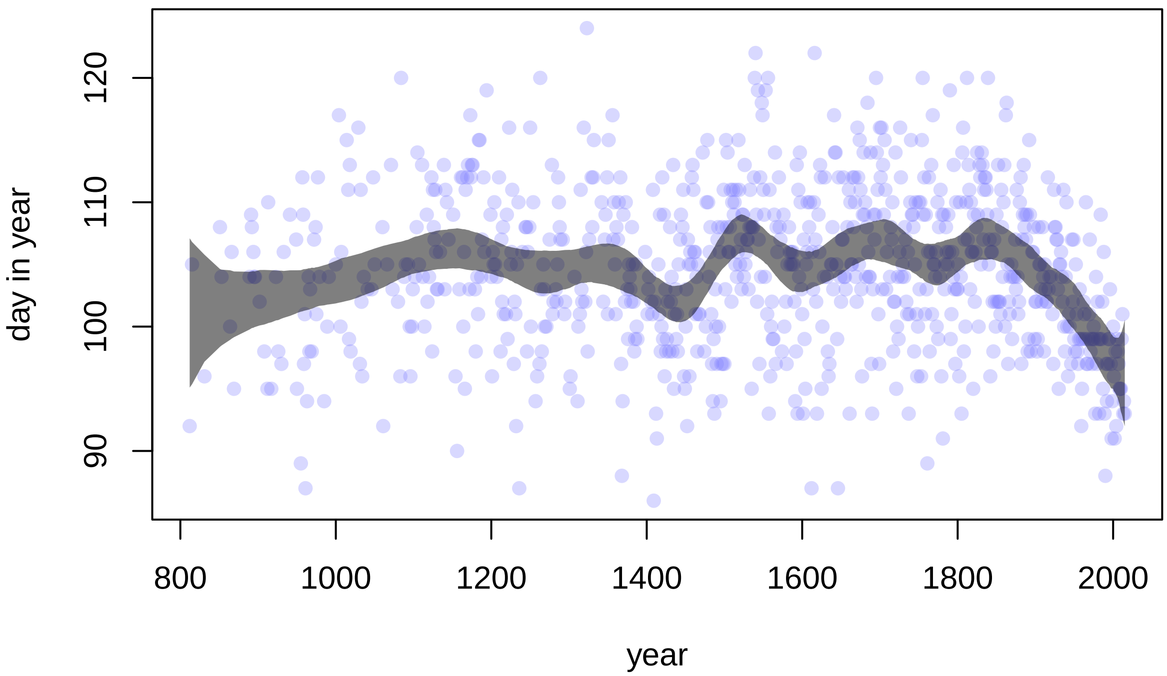
```
data("cherry_blossoms")
d8 <- cherry_blossoms %>%
```

```
  drop_na(doy)

num_knots <- 50
knot_list <- quantile(d8$year, probs = seq(from = 0, to = 1, length.out = num_knots))

B <- bs(d8$year,
        knots=knot_list[-c(1, num_knots)],
        degree=3,
        intercept=TRUE)

m8 <- quap(
  alist(D ~ dnorm(mu, sigma),
        mu <- a + B%*%w,
        a ~ dnorm(100, 10),
        w ~ dnorm(0, 1),
        sigma ~ dexp(1)),
  data=list(D=d8$doy,
            B=B),
  start=list(w=rep(0, ncol(B))))

post <- extract.samples(m8)
w <- apply(post$w, 2, mean)

mu <- link(m8)
mu_PI <- apply(mu, 2, PI, 0.97)

plot( d8$year , d8$doy , col=col.alpha(rangi2,0.3) , pch=16, xlab = "year", ylab = "day in year" )
shade( mu_PI , d8$year , col=col.alpha("black",0.5))
```



This spline is much wigglier and has rougher edges on the posterior interval than the spline with 15 knots in SR, which is to be expected because splines with more knots are more flexible. Next, let's look at a spline with 15 knots but with a larger standard deviation for the weights' prior.

```r
num_knots <- 15
knot_list <- quantile(d8$year, probs = seq(from = 0, to = 1, length.out = num_knots))

B <- bs(d8$year,
        knots=knot_list[-c(1, num_knots)],
        degree=3,
        intercept=TRUE)

m8 <- quap(
  alist(D ~ dnorm(mu, sigma),
        mu <- a + B%*%w,
        a ~ dnorm(100, 10),
        w ~ dnorm(0, 5),
        sigma ~ dexp(1)),
  data=list(D=d8$doy,
            B=B),
  start=list(w=rep(0, ncol(B))))

post <- extract.samples(m8)
w <- apply(post$w, 2, mean)

mu <- link(m8)
mu_PI <- apply(mu, 2, PI, 0.97)

plot( d8$year , d8$doy , col=col.alpha(rangi2,0.3) , pch=16, xlab = "year", ylab = "day in year" )
shade( mu_PI , d8$year , col=col.alpha("black",0.5))
```



This spline is also wigglier than the spline in SR and its wiggles are more pronounced. This, again, is to be expected, because our prior for the weights is more uncertain, so the weights instead are based more on the observed data because they are less limited by the prior.