Applied Bayesian modeling - HW3, age at marriage

Background: The dataset "marriage.csv'' contains (simulated) data on the age of first marriage for a number of women in Kenya. Information on the ethnic group is provided as well. For the questions below, let y_i denote the age of first marriage for the i-th woman, and $j[i]$ her ethnic group. The goal is to learn more about the mean age of marriage within ethnic groups, using a Bayesian hierarchical model.

To answer the questions below, fit the following Bayesian hierarchical model to the marriage data:

$$\begin{aligned} y_i|\alpha_{j[i]}, \sigma_y &\overset{i.i.d}{\sim} N(\alpha_{j[i]}, \sigma_y^2), \\ \alpha_j|\mu_\alpha, \sigma_\alpha &\overset{i.i.d}{\sim} N(\mu_\alpha, \sigma_\alpha^2), \end{aligned}$$

using default priors for μ_α, σ_α, σ_y as set in the brm function. In the questions below, if you present default brm-output, please indicate what brm-parameter name refers to what greek letter in the equations above.

Data and model fitting

```
dat <- read.csv("../data/marriage.csv") %>%
  rename(y = agemarried)
```

- y_i is agemarried (already renamed in 1st chunk)
- groups refer to ethnicgroup

Create data set with info for each group

```
# to plot observations and county means ~ sample sizes,
# easier to see if sample sizes are slighly jittered
set.seed(12345)

datgroup <- dat %>%
  group_by(ethnicgroup) %>%
  summarize(nunits = n(), ybar = mean(y), ethnicgroup = ethnicgroup[1]) %>%
  mutate(nunits_jitter = nunits*
           exp(runif (length(nunits), -.1, .1)))

datgroup
```

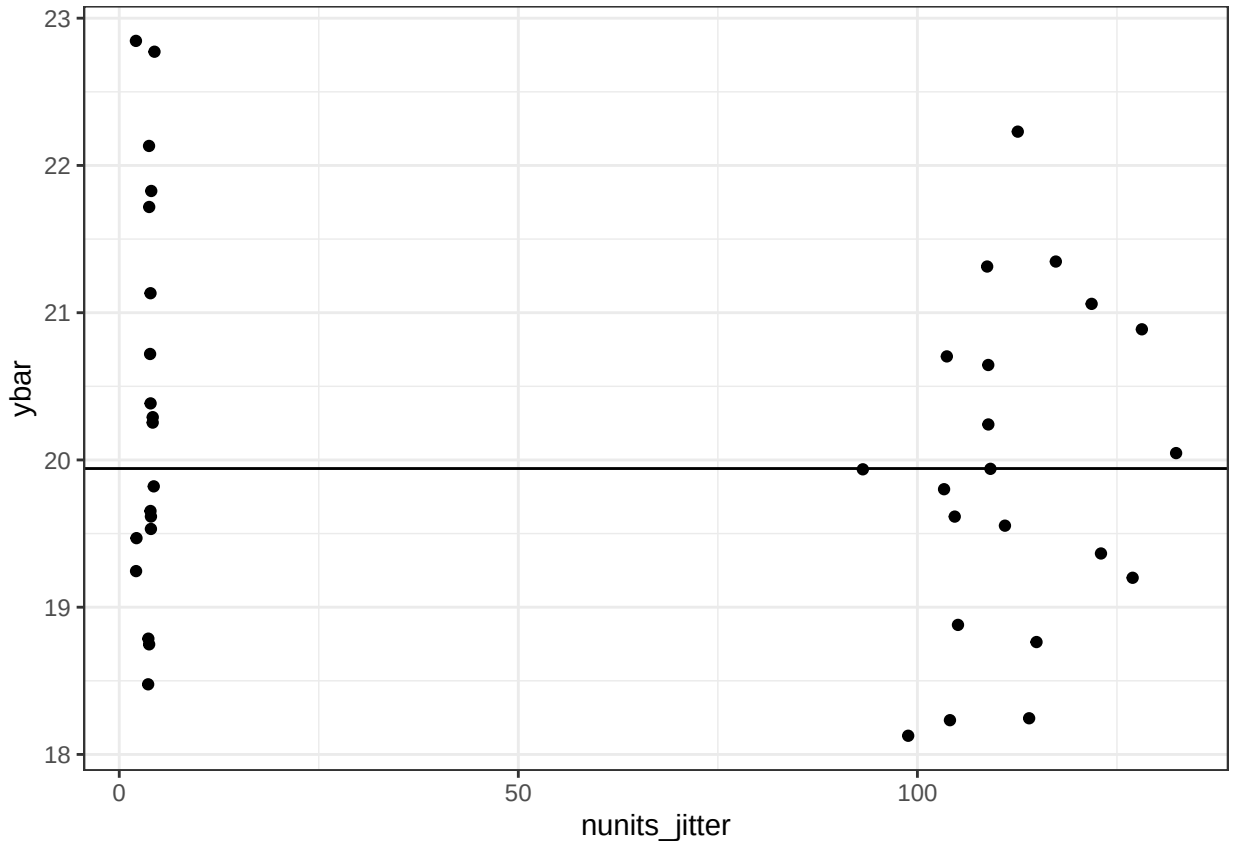
```
## # A tibble: 40 x 4
##    ethnicgroup nunits  ybar nunits_jitter
##          <int>  <int> <dbl>         <dbl>
##  1           1      2  22.8          2.09
##  2           2      2  19.5          2.16
##  3           3      2  19.2          2.11
##  4           4      4  19.8          4.32
```

```
##  5           5      4  19.5          3.97
##  6           6      4  18.7          3.74
##  7           7      4  20.7          3.86
##  8           8      4  21.8          4.01
##  9           9      4  20.3          4.19
## 10          10      4  22.8          4.41
## # ... with 30 more rows
```

```
ngroups <- dim(datgroup)[1]
```

```
# state mean, n.j and county means
ybarbar <- mean(dat$y) # state mean
```

```
datgroup %>%
  ggplot(aes(x = nunits_jitter, y = ybar)) +
  geom_point() +
  geom_hline(mapping = aes(yintercept = ybarbar)) +
  theme_bw()
```




```
fit <- brm(y ~ (1|ethnicgroup), family = gaussian(), data = dat,
                      chains = 4,
           seed = 1234, # need to add seed here to make this reproducible
          iter = 2000,
           thin = 1,
           cores = getOption("mc.cores", 4))
```

```
#saveRDS(fit, "fit_hw.rds")
#fit <- readRDS("fit_hw.rds")
```

Exercise 3

(a)

State the point estimates and 95% CIs of each of the following parameters and interpret (explain what information is given by) these estimates: $\%\hat{\beta}$, μ_α, σ_α, σ_y.

Solution

```
summary(fit)
```

```
##  Family: gaussian
##   Links: mu = identity; sigma = identity
## Formula: y ~ (1 | ethnicgroup)
##    Data: dat (Number of observations: 2400)
##   Draws: 4 chains, each with iter = 2000; warmup = 1000; thin = 1;
##          total post-warmup draws = 4000
##
## Group-Level Effects:
## ~ethnicgroup (Number of levels: 40)
##               Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## sd(Intercept)     1.10      0.15     0.84     1.43 1.01      704     1192
##
## Population-Level Effects:
##           Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## Intercept    20.06      0.20    19.67    20.47 1.01      454      787
##
## Family Specific Parameters:
##       Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## sigma     2.01      0.03     1.96     2.07 1.00     4329     2893
##
## Draws were sampled using sampling(NUTS). For each parameter, Bulk_ESS
## and Tail_ESS are effective sample size measures, and Rhat is the potential
## scale reduction factor on split chains (at convergence, Rhat = 1).
```

The point estimates (means or medians) are given below, as well as the 95% CIs:\

Parameter	Posterior mean	95% CI
μ_α; Intercept	20.1	(19.7, 20.5)
σ_α; sd(Intercept)	1.10	(0.84, 1.43)
σ_y; sigma	2.01	(1.96, 2.07)

\ Interpretation:

- $\hat{\mu}_\alpha = 20.1$, thus the estimated mean of ethnic-group specific mean age of marriage is 20.1 years.
- $\hat{\sigma}_\alpha = 1.10$: The across-ethnic group standard deviation of mean ages of marriage is 1.10.

- $\hat{\sigma}_y = 2.01$: the standard deviation of observed ages of first marriage across women within ethnic groups is 2.01.

b

State the point estimate and 95% CIs for α_1 and interpret (explain what information is given by) the estimate.

Solution

Creating outputs first:

eta = alpha - mu_alpha (as compared to notation in slides), labeled here as random effects

```
eta <- as_tibble(ranef(fit)$ethnicgroup[,,"Intercept"], rownames = "ethnicgroup")
head(eta)
```

```
## # A tibble: 6 x 5
##   ethnicgroup Estimate Est.Error   Q2.5 Q97.5
##   <chr>          <dbl>     <dbl>  <dbl> <dbl>
## 1 1              1.06      0.877 -0.634 2.81
## 2 2             -0.201     0.877 -1.90  1.49
## 3 3             -0.320     0.881 -2.10  1.37
## 4 4             -0.127     0.755 -1.65  1.37
## 5 5             -0.284     0.749 -1.76  1.17
## 6 6             -0.702     0.726 -2.12  0.740
```

To get the alpha = eta + mu_alpha, we can use the following call

```
alphas <-
  coef(fit, summary = T)$ethnicgroup %>%
  as_tibble(rownames = "ethnicgroup") %>%
  rename(alph = Estimate.Intercept)
alphas
```

```
## # A tibble: 40 x 5
##    ethnicgroup  alph Est.Error.Intercept Q2.5.Intercept Q97.5.Intercept
##    <chr>       <dbl>               <dbl>          <dbl>           <dbl>
##  1 1            21.1               0.887           19.4            22.9
##  2 2            19.9               0.882           18.2            21.6
##  3 3            19.7               0.886           18.0            21.4
##  4 4            19.9               0.753           18.4            21.4
##  5 5            19.8               0.747           18.3            21.2
##  6 6            19.4               0.721           18.0            20.8
##  7 7            20.4               0.751           19.0            21.9
##  8 8            21.0               0.741           19.6            22.5
##  9 9            20.2               0.771           18.6            21.7
## 10 10           21.5               0.765           20.0            23.1
## # ... with 30 more rows
```

Information for the first group:

```
alphas %>% filter(ethnicgroup ==1)
```

```
## # A tibble: 1 x 5
##   ethnicgroup  alph Est.Error.Intercept Q2.5.Intercept Q97.5.Intercept
##   <chr>       <dbl>               <dbl>          <dbl>           <dbl>
## 1 1            21.1               0.887           19.4            22.9
```

The point estimate is given below, as well as the 95% CIs:\

Parameter	Posterior mean	Posterior median	95% CI
α_1	21.1	21.1	(19.4, 22.9)

\ Interpretation: The mean age at marriage for women in the first ethnic group (with $j = 1$) is given by α_1. Posterior mean is given by 21.1 years with 95% CI ranging from 19.4 to 22.9 years.

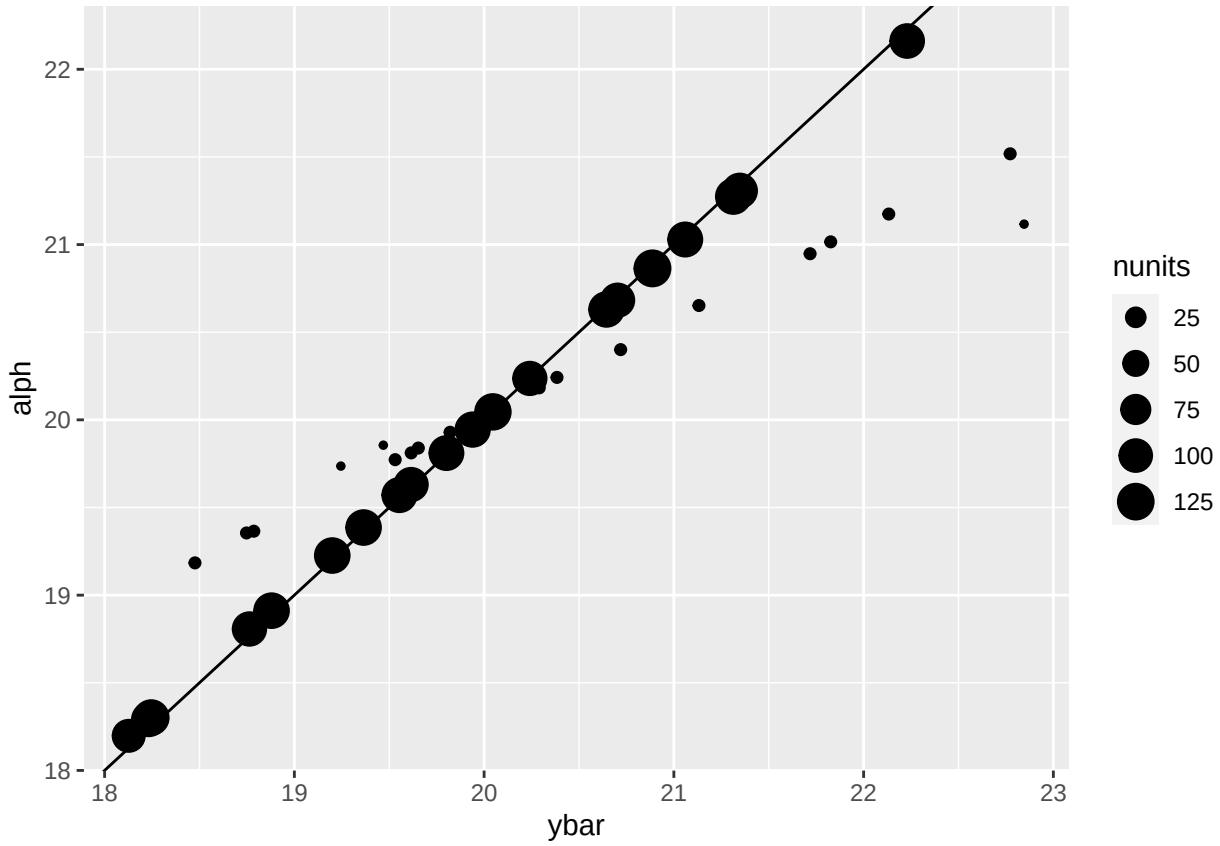
c

Construct two plots: (a) plot of $\hat{\alpha}_j - \bar{y}_j$ against the within-ethnic-group sample size n_j and (b) plot $\hat{\alpha}_j$ against $\bar{y}_j$, with the identity line added. Explain what information these plots provide regarding the comparison of $\bar{y}_j$ and $\hat{\alpha}_j$ for estimating the mean age of marriage within ethnic groups.

See plots below. We see that the partially pooled estimates $\hat{\alpha}_j$ are in between the complete pooling estimate/overall mean $\bar{y}$ and the no-pooling estimates/group means $\bar{y}_j$. When the sample size in the group is large, the estimate $\hat{\alpha}_j$ is similar to the group mean $\bar{y}_j$ but as the sample size decreases, $\hat{\alpha}_j$ gets closer to (is shrunk towards) the overall mean $\bar{y}$. \

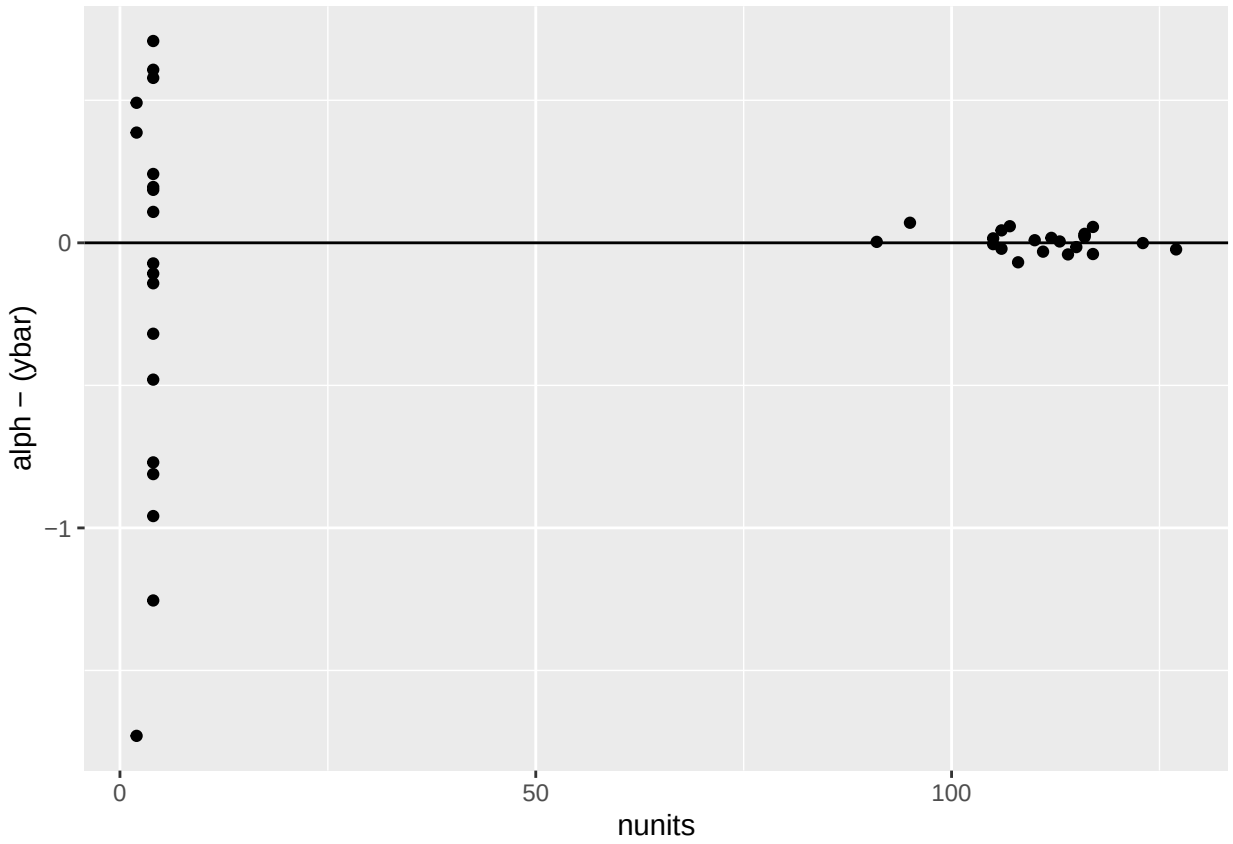
```
datgroup <-
  datgroup %>% mutate(ethnicgroup = as.character(ethnicgroup))

alphas %>%
  left_join(datgroup) %>%
  ggplot(aes(y = alph, x = ybar, size = nunits)) +
  geom_point() +
  geom_abline(slope = 1, intercept = 0)
```



Plot of alpha - ybar

```
alphas %>%
  left_join(datgroup) %>%
  ggplot(aes(y = alph - (ybar), x = nunits)) +
  geom_point() +
  geom_hline(yintercept = 0)
```



Optional/just for fun exercise

As stated in the slides, for the multilevel model under consideration here, the full conditional distribution for the j-th state mean is given by:

$$\begin{aligned}
\alpha_j|\boldsymbol{y}, \mu_\alpha, \sigma_y, \sigma_\alpha &\sim N(m, v), \\
v &= \left(n_j/\sigma_y^2 + 1/\sigma_\alpha^2\right)^{-1}, \\
m &= v \cdot \left(\frac{n_j}{\sigma_y^2}\bar{y}_j + \frac{1}{\sigma_\alpha^2}\mu_\alpha\right) = \frac{\frac{n_j}{\sigma_y^2}\bar{y}_j + \frac{1}{\sigma_\alpha^2}\mu_\alpha}{\frac{n_j}{\sigma_y^2} + \frac{1}{\sigma_\alpha^2}}.
\end{aligned}$$

Give the derivation of the full conditional. Hint: Check slides in module 4 for the derivation of the normal distribution for μ in the normal-normal setting.

Start by using Bayes rule to go to a product of two densities that are know:

$$\begin{aligned}
&p(\alpha_j|\boldsymbol{y}, \mu_\alpha, \sigma_y, \sigma_\alpha) \propto p(\boldsymbol{y}|\alpha_j, \sigma_y) p(\alpha_j|\mu_\alpha, \sigma_\alpha) \text{ (Bayes' rule)} \\
&\propto \exp\left(\frac{-1}{2\sigma_y^2}\sum_{i=1}^{n}(y_i - \alpha_{j[i]})^2\right) \cdot \exp\left(\frac{-1}{2\sigma_\alpha^2}(\alpha_j - \mu_\alpha)^2\right) \\
&\propto \exp\left(-\frac{1}{2}f(\alpha_j)\right),
\end{aligned}$$

where

$$\begin{aligned} f(\alpha_j) &= 1/\sigma_y^2 \sum_{i=1}^{n} (y_i - \alpha_{j[i]})^2 + 1/\sigma_\alpha^2 (\alpha_j - \mu_\alpha)^2, \\ &\propto 1/\sigma_y^2 (-2 n_j \bar{y}_j \cdot \alpha_j + n_j \alpha_j^2) + 1/\sigma_\alpha^2 (\alpha_j^2 - 2\mu_\alpha \cdot \alpha_j), \\ &\propto (1/\sigma_\alpha^2 + n_j/\sigma_y^2)\alpha_j^2 - B \cdot \alpha_j, \\ &\propto (1/\sigma_\alpha^2 + n_j/\sigma_y^2) \left(\alpha_j - \frac{1/2 \cdot B}{1/\sigma_\alpha^2 + n_j/\sigma_y^2} \right)^2, \end{aligned}$$

with $B = 2\left(\frac{n_j}{\sigma_y^2}\bar{y}_j + \frac{1}{\sigma_\alpha^2}\mu_\alpha\right)$. Remember that if $Z \sim N(m, s^2)$, $p(z) \propto \exp(-1/2 \cdot 1/s^2 (z-m)^2)$. Thus

$$\alpha_j | \boldsymbol{y}, \sigma^2 \sim N\left(\frac{1/2 \cdot B}{\left(n_j/\sigma_y^2 + 1/\sigma_\alpha^2\right)}, \left(n_j/\sigma_y^2 + 1/\sigma_\alpha^2\right)^{-1} \right),$$

which completes the derivation.

Note that this expression also follows from generalizing the finding in module 4. In that module, we concluded that if $y_i | \mu, \sigma^2 \sim N(\mu, \sigma^2)$ (independent) and if $\mu \sim N(\mu_0, s_{\mu 0}^2)$, then $\mu | \boldsymbol{y}, \sigma^2 \sim N\left(\frac{\mu_0/s_{\mu 0}^2 + n \cdot \bar{y}/\sigma^2}{1/s_{\mu 0}^2 + n/\sigma^2}, \frac{1}{1/s_{\mu 0}^2 + n/\sigma^2}\right)$.

We can generalize that finding to include unknown parameters in the density for μ, as long as we condition on these parameters:
if $y_i | \mu, \sigma^2 \sim N(\mu, \sigma^2)$ (independent) and if $\mu | \gamma, \delta \sim N(\gamma, \delta^2)$, then $\mu | \boldsymbol{y}, \sigma^2, \gamma, \delta \sim N\left(\frac{\gamma/\delta^2 + n \cdot \bar{y}/\sigma^2}{1/\delta^2 + n/\sigma^2}, \frac{1}{1/\delta^2 + n/\sigma^2}\right)$.

In this HW question, we are in the same setting; We observe n_j observations in group $j[i] = h$ where, conditional on $\alpha_{j[i]}$ and σ_y, the y's in that group are independent with $y_i | \alpha_{j[i]}, \sigma_y \sim N(\alpha_{j[i]}, \sigma_y^2)$. The mean parameter $\alpha_{j[i]}$ is assigned a normal density with parameters μ_α and σ_α. Hence we find that

$$\begin{aligned} \alpha_j | \boldsymbol{y}, \mu_\alpha, \sigma_y, \sigma_\alpha &\sim N(m, v), \\ v &= \left(n_j/\sigma_y^2 + 1/\sigma_\alpha^2\right)^{-1}, \\ m &= v \cdot \left(\frac{n_j}{\sigma_y^2}\bar{y}_j + \frac{1}{\sigma_\alpha^2}\mu_\alpha\right) = \frac{\frac{n_j}{\sigma_y^2}\bar{y}_j + \frac{1}{\sigma_\alpha^2}\mu_\alpha}{\frac{n_j}{\sigma_y^2} + \frac{1}{\sigma_\alpha^2}}. \end{aligned}$$

Exercise 4

Continue with the marriage data set from exercise 3.
The goal in this exercise is to predict the age at first marriage for a randomly sampled woman in an ethnic group for which we have not yet observed any data, using the model and data from exercise 3.

(a) Obtain samples from the predictive posterior density for the age at first marriage and visualize the samples in a histogram. In your answer, include R code (that does NOT use the predict function from brms) as well as a write up in equations how you obtained the samples. Make sure to introduce notation first to explain what you're sampling.

(b) Use the samples to construct a point prediction and 95% prediction interval for age at first marriage. In your answer, include the expression used for calculating the point prediction from the samples.

(c) What is the probability that the observed age at first marriage will be greater than $\bar{y}$? In your answer, include the expression used for calculating this probability from the samples.

Solutions

Goal: get samples for age of marriage for a woman in ethnic group j, denoted by $\tilde{y}_k^{(s)} \sim p(\tilde{y}_k|\boldsymbol{y})$ with $j[k] = h$ referring to a new group.

Given that

$$\begin{aligned} p(\tilde{y}_k|\boldsymbol{y}) &= \int\int\int p(\tilde{y}_k|\mu_\alpha, \sigma_\alpha, \sigma_y) p(\mu_\alpha, \sigma_\alpha, \sigma_y|\boldsymbol{y}) d\mu_\alpha d\sigma_\alpha d\sigma_y, \\ &= \int\int\int \left(\int p(\tilde{y}_k|\tilde{\alpha}_h, \sigma_y) p(\tilde{\alpha}_h|\mu_\alpha, \sigma_\alpha) d\tilde{\alpha}_h \right) p(\mu_\alpha, \sigma_\alpha, \sigma_y|\boldsymbol{y}) d\mu_\alpha d\sigma_\alpha d\sigma_y, \end{aligned}$$

We can sample $\tilde{y}_k^{(s)} \sim p(\tilde{y}|\boldsymbol{y})$ in three steps:

(1) Sample $(\mu_\alpha^{(s)}, \sigma_\alpha^{(s)}, \sigma_y^{(s)}) \sim p(\mu_\alpha, \sigma_\alpha, \sigma_y|\boldsymbol{y})$,

(2) Sample $\tilde{\alpha}_h^{(s)} \sim p(\tilde{\alpha}_h|\mu_\alpha^{(s)}, \sigma_\alpha^{2(s)})$.

(3) Sample $\tilde{y}_k^{(s)} \sim p(\tilde{y}_k|\tilde{\alpha}_h^{(s)}, \sigma_y^{2(s)})$

We already have samples $(\mu_\alpha^{(s)}, \sigma_\alpha^{(s)}, \sigma_y^{(s)})$ from fitting the model, so just need to draw the $\tilde{\alpha}_h^{(s)}$ and $\tilde{y}$s.

We first need to get samples $\tilde{\alpha}_h \sim p(\tilde{\alpha}_h|\boldsymbol{y})$ where $\tilde{\alpha}_j$ refers to the mean age of marriage for the new ethnic group. We obtain samples from it as follows:

- For each posterior sample $(\mu_\alpha^{(s)}, \sigma_\alpha^{2(s)}) \sim p((\mu_\alpha, \sigma_\alpha^2)|\boldsymbol{y})$, sample $\tilde{\alpha}_h^{(s)}|(\mu_\alpha^{(s)}, \sigma_\alpha^{2(s)}) \sim N(\mu_\alpha^{(s)}, (\sigma_\alpha^2)^{(s)})$.

Continuing with the $\tilde{\alpha}_h^{(s)}$s from the previous step, we then obtain samples from $p(\tilde{y}_k|\boldsymbol{y})$ with $j[k] = h$ as follows:

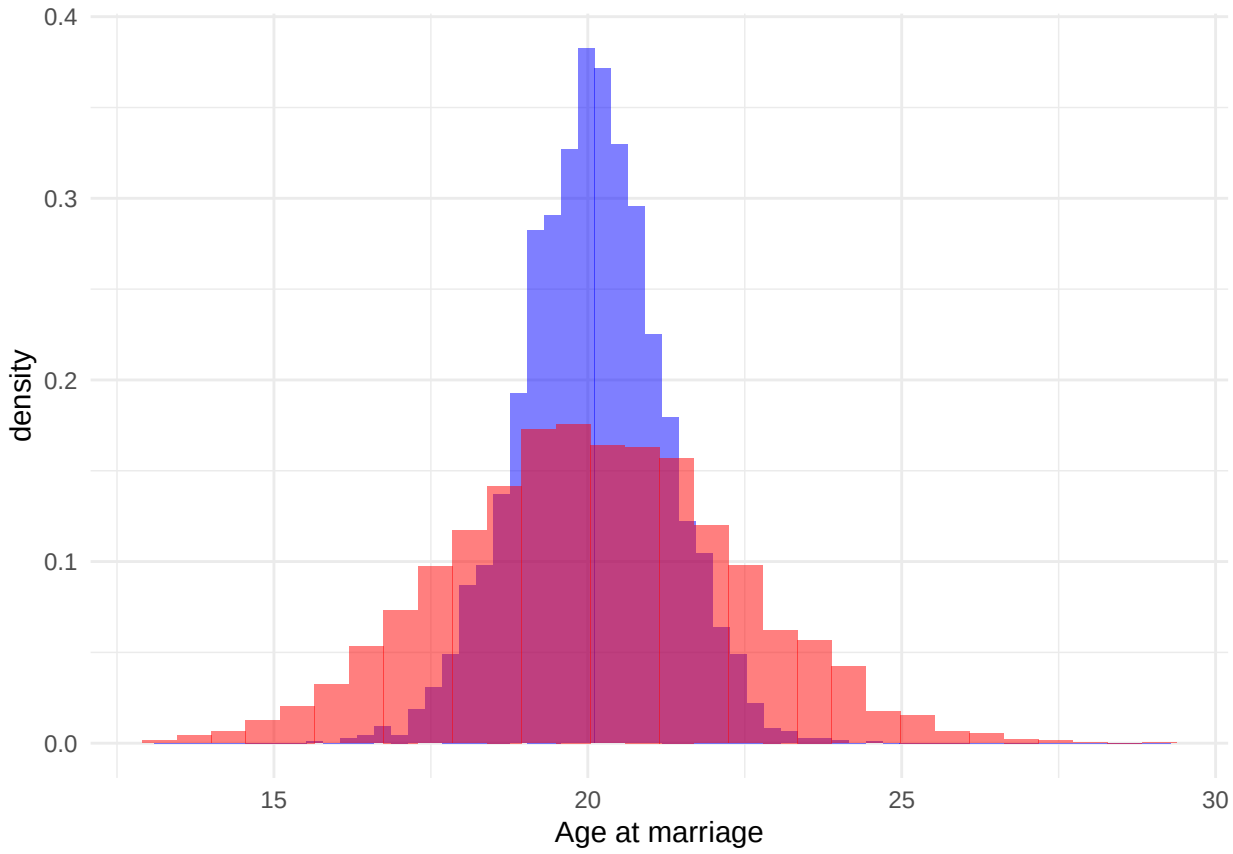
- Sample $\tilde{y}_k^{(s)} \sim p(\tilde{y}|\tilde{\alpha}_h^{(s)}, \sigma_y^{2(s)})$, here $\tilde{y}_k|\tilde{\alpha}_h^{(s)}, \sigma_y^{2(s)} \sim N\left(\alpha_{j[k]}^{(s)}, (\sigma_y^2)^{(s)}\right)$,

where $\tilde{y}_k$ refers to the age of marriage for a woman in ethnic group j from above.

```
samp <- as_draws_df(fit)
#dim(samp)
#names(samp)[1:3]
sigmay_s <- samp$sigma
mualpha_s <- samp$b_Intercept
sigmaalpha_s <- samp$sd_ethnicgroup__Intercept
S <- length(sigmay_s)

set.seed(1234) # to make the sampling reproducible
alphanew_s <- rnorm(S, mualpha_s, sigmaalpha_s)
ytilde_s <- rnorm(S, alphanew_s, sigmay_s)
```

```
p <- as_tibble(alphanew_s) %>%
  ggplot(aes(alphanew_s, after_stat(density))) +
  geom_histogram(alpha = .5, fill = "blue", bins = 60) +
  theme_minimal() +
  xlab("Age at marriage")
p +
  geom_histogram(as_tibble(ytilde_s), , bins = 30, mapping = aes(ytilde_s, after_stat(density)),
                 alpha = .5, fill = "red", adjust = 1.5, size = 1.5)
```



We create predictions and PIs using the samples, with point prediction given by $1/S \sum_s \tilde{y}^{(s)}$.

```
point_interval(alphanew_s, .point = mean)
```

```
##          y     ymin     ymax .width .point .interval
## 1 20.05392 17.79977 22.26695   0.95   mean        qi
```

```
point_interval(ytilde_s, .point = mean)
```

```
##          y    ymin     ymax .width .point .interval
## 1 20.07729 15.6774 24.52542   0.95   mean        qi
```

Result above for both the predicted group mean as well as a woman. As expected, the point estimates are approximately the same but the PI for the age of marriage of one woman is more uncertain as compared to the mean age of marriage in the ethnic group.

	Point estimate (posterior mean)	95% PI
group mean	20.1	(17.8, 22.3)
woman	20.1	(15.7, 24.5)

The probability that the observed age at first marriage will be greater than $\bar{y}$ is approximated as follows:

$$P(\tilde{y}_k > \bar{y}|\boldsymbol{y}) \approx 1/S \sum_s 1(\tilde{y}_k^{(s)} > \bar{y}).$$

```
mean(ytilde_s > ybarbar)
```

```
## [1] 0.518
```

The value is 0.52.

Compare to brm output

brm for a new house in a new county

```
newdata2 <- data.frame(
  ethnicgroup = c("justsomeotherplace")
)
samples_ytilde2 <- posterior_predict(fit, newdata = newdata2, allow_new_levels = TRUE)
```

```
point_interval(samples_ytilde2, .point = mean)
```

```
##          y     ymin     ymax .width .point .interval
## 1 20.03613 15.56084 24.40675   0.95   mean        qi
```

```
point_interval(ytilde_s, .point = mean)
```

```
##          y    ymin     ymax .width .point .interval
## 1 20.07729 15.6774 24.52542   0.95   mean        qi
```