Lab 6

Yaqi Shi, 1003813180

2023-02-22

Introduction

This lab will be looking at trying to replicate some of the visualizations in the lecture notes, involving prior and posterior predictive checks, and LOO model comparisons.

The dataset is a 0.1% of all births in the US in 2017. I've pulled out a few different variables, but as in the lecture, we'll just focus on birth weight and gestational age.

The data

Read it in, along with all our packages.

```
library(tidyverse)
library(here)
library(rstan)
library(bayesplot)
library(loo)
library(tidybayes)

# Read in the dataset
ds <- read_rds("births_2017_sample.RDS")

# Clean the dataset
ds <- ds %>%
  rename(birthweight = dbwt, gest = combgest) %>%
  mutate(preterm = ifelse(gest<32, "Y", "N")) %>%
  filter(ilive=="Y",gest< 99, birthweight<9.999)
```

Brief overview of variables:

- `mager` mum's age
- `mracehisp` mum's race/ethnicity see here for codes: https://data.nber.org/natality/2017/natl2017.pdf (https://data.nber.org/natality/2017/natl2017.pdf) page 15
- `meduc` mum's education see here for codes: https://data.nber.org/natality/2017/natl2017.pdf (https://data.nber.org/natality/2017/natl2017.pdf) page 16
- `bmi` mum's bmi
- `sex` baby's sex
- `combgest` gestational age in weeks
- `dbwt` birth weight in kg
- `ilive` alive at time of report y/n/ unsure

\(\;\)

\(\;\)

Question 1

Use plots or tables to show three interesting observations about the data. Remember:

- Explain what your graph/ tables show
- Choose a graph type that's appropriate to the data type
- If you use `geom_smooth`, please also plot the underlying data

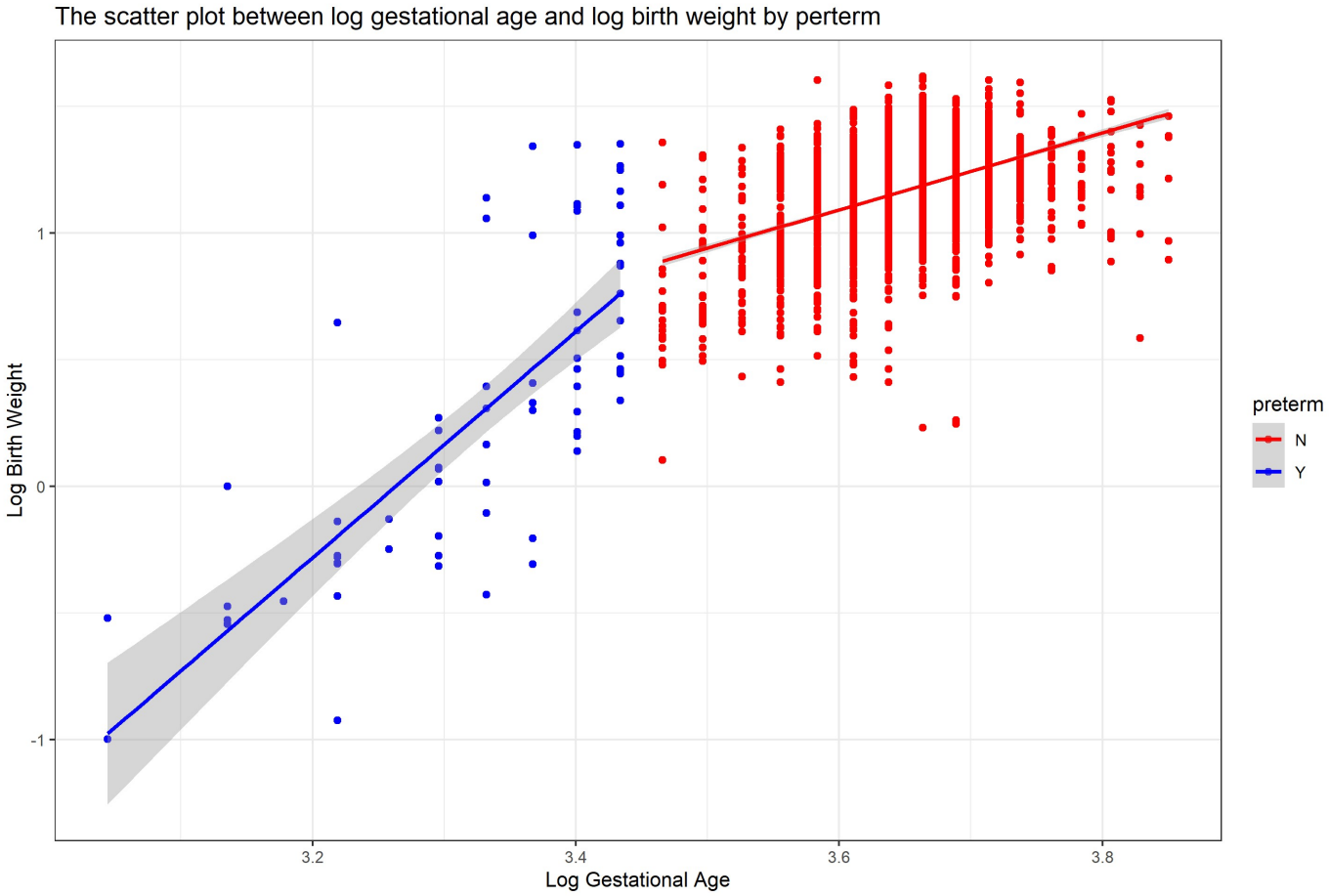
Feel free to replicate one of the scatter plots in the lectures as one of the interesting observations, as those form the basis of our models.

Answer

1 - The scatter plot between log gestational age and log birth weight.

This is a replication of the scatter plot we have in the lecture. The type of the plot is a scatter plot and the color is differentiated by whether the kid is born preterm. The purpose of the plot is to explore the relationship between he gestational age and the birth weight and compare the difference in preterm and non-preterm period. In this plot we can see that as the gestational age increases, the body weight of the baby increases and the weight of a baby grows faster in the preterm period.

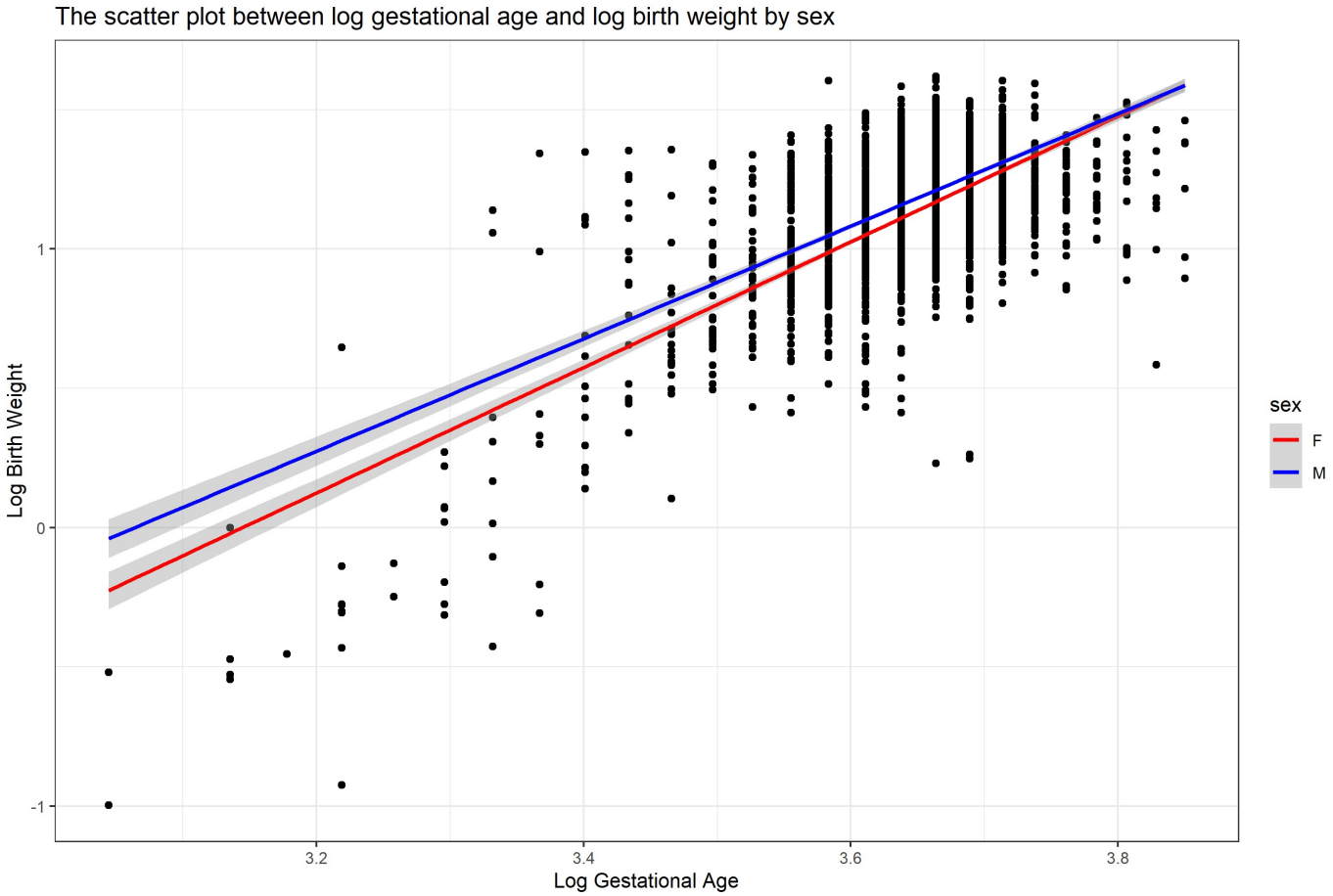
```
ggplot(ds, aes(x=log(gest), y = log(birthweight), color = preterm)) +
  geom_point() +
  geom_smooth(method = lm) +
  theme_bw() +
  scale_color_manual(values=c("red", "blue")) +
  labs(x = "Log Gestational Age", y="Log Birth Weight", title = "The scatter plot between log
gestational age and log birth weight by perterm")
```



2 - The difference in the relationship between the gestational age and birth weight by baby sex.

The type of the plot is scatter plot and the smoothed curve is differentiated by baby's sex. The purpose of this plot is to discover whether baby boys and baby girls show a different relationship between birth weight and gestational age. From this plot, we can see that for both baby boys and baby girls, the body weight increases with gestational age. The slope for baby girls is slightly higher than that for baby boys and the weight of baby boys is higher than the weight of baby girls with the same gestational age.

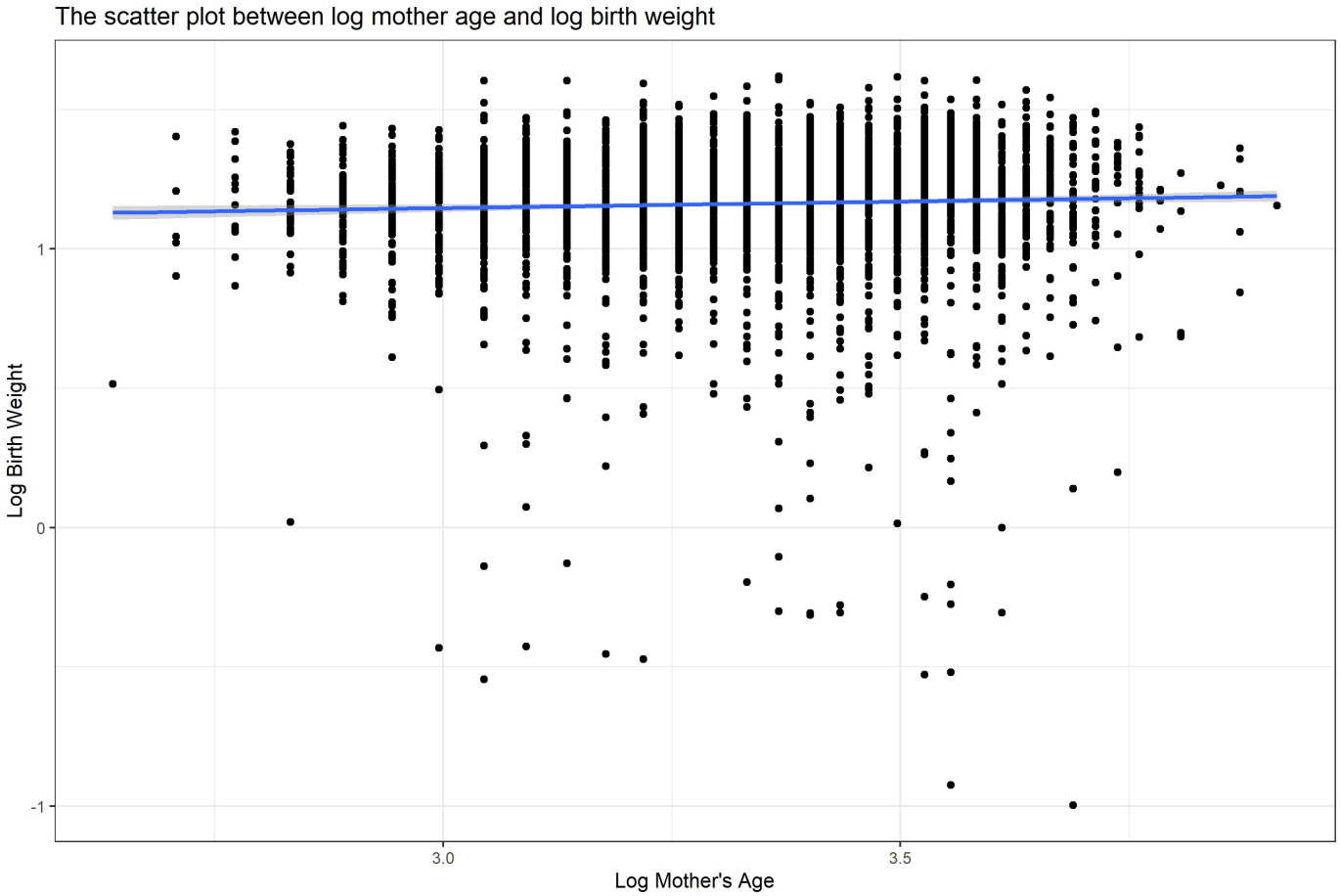
```
ggplot(ds, aes(x=log(gest), y = log(birthweight))) +
  geom_point() +
  geom_smooth(method = lm,aes(color=sex)) +
  theme_bw() +
  scale_color_manual(values=c("red", "blue")) +
  labs(x = "Log Gestational Age", y="Log Birth Weight", title = "The scatter plot between log
gestational age and log birth weight by sex")
```



3 - The relationship between the mother's age and the birth weight

The type of the plot is scatter plot and the smoothed curve is fitted curve using linear regression. The purpose of this plot is to discover the relationship between the mother's age and the baby's birth weight. From this plot, we can see that as the mother's age increase, there is no clear change in the birth weight of a baby. However we do notice that the likelihood of having small birth weight increases with the mother's age. That is as the mother's age increases, it has a higher chance that the baby has a lower body weight which possibly due to preterm labor.

```
ggplot(ds, aes(x=log(mager), y = log(birthweight))) +
  geom_point() +
  geom_smooth(method = lm) +
  theme_bw() +
  labs(x = "Log Mother's Age", y="Log Birth Weight", title = "The scatter plot between log mo
ther age and log birth weight")
```



From the three plots above, we believe that gestational age is highly related to the birth weight of a baby. Preterm and Sex are possible explanatory variables to explain the trends better. Mother's age may not contribute to the change in the birth weight that much.

\(\;\)

\(\;\)

The model

As in lecture, we will look at two candidate models

Model 1 has log birth weight as a function of log gestational age

\[\log(y_i) \sim N(\beta_1 + \beta_2\log(x_i), \sigma^2) \]

Model 2 has an interaction term between gestation and prematurity

\[\log(y_i) \sim N(\beta_1 + \beta_2\log(x_i) + \beta_2 z_i + \beta_3\log(x_i) z_i, \sigma^2) \]

- \(y_i\) is weight in kg
- \(x_i\) is gestational age in weeks, CENTERED AND STANDARDIZED
- \(z_i\) is preterm (0 or 1, if gestational age is less than 32 weeks)

\(\;\)

\(\;\)

Prior predictive checks

Let's put some weakly informative priors on all parameters i.e. for the βs

$$ \beta \sim N(0, 1) $$

and for σ

$$ \sigma \sim N^+(0,1) $$ where the plus means positive values only i.e. Half Normal.

Let's check to see what the resulting distribution of birth weights look like given Model 1 and the priors specified above, assuming we had no data on birth weight (but observations of gestational age).

$\;$

$\;$

Question 2

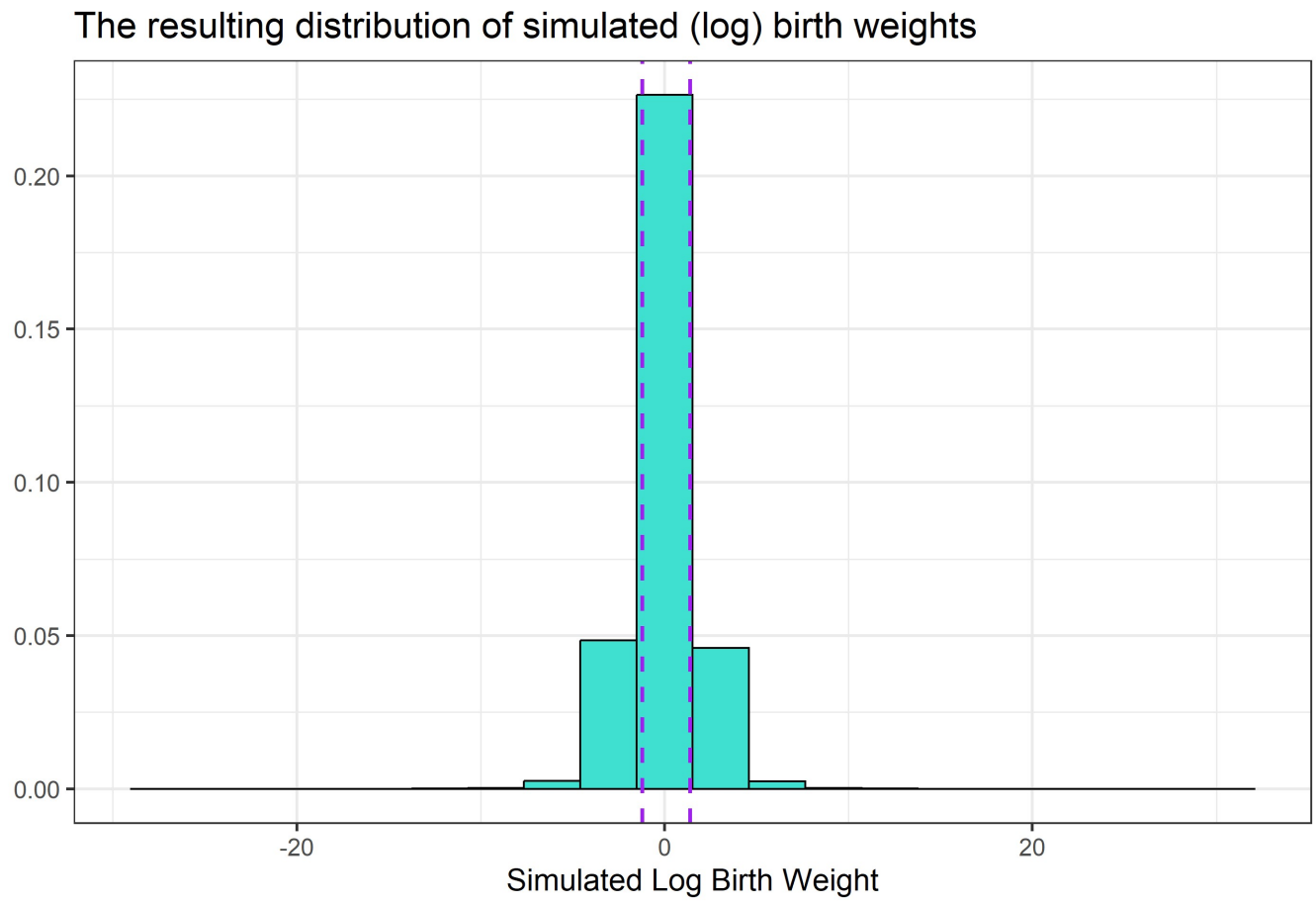
For Model 1, simulate values of βs and σ based on the priors above. Do 1000 simulations. Use these values to simulate (log) birth weights from the likelihood specified in Model 1, based on the set of observed gestational weights. **Remember the gestational weights should be centered and standardized**.

- Plot the resulting distribution of simulated (log) birth weights.
- Plot ten simulations of (log) birthweights against gestational age.

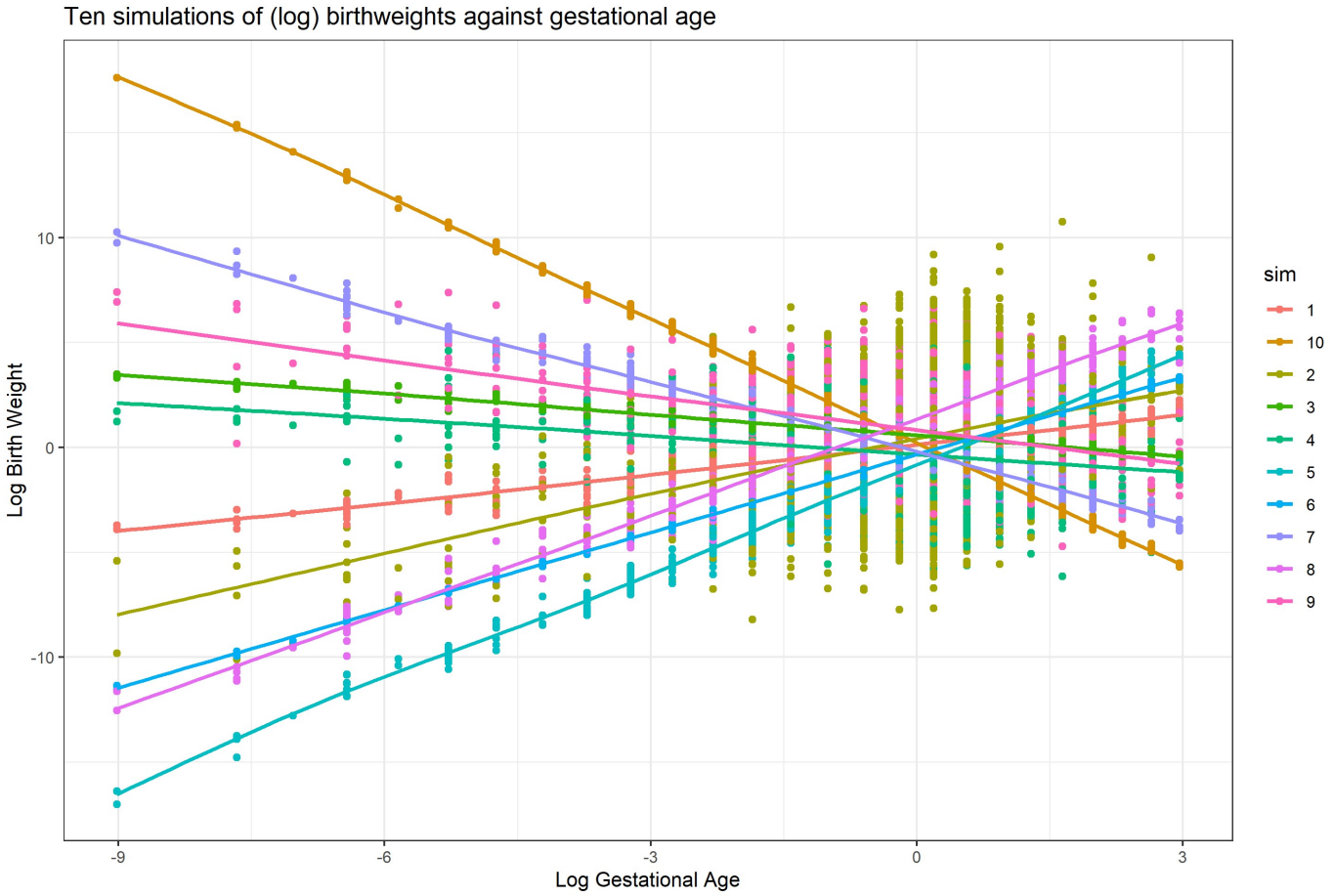
Answer

Before I start here, I do admit that I read the blog in Monica's blog [https://www.monicaalexander.com/posts/2020-28-02-bayes_viz/#fn1 (https://www.monicaalexander.com/posts/2020-28-02-bayes_viz/#fn1)] so if your find any of the code look familiar in the simulation part, this is the reason why :)

```
# Simulate value for betas and sigma
set.seed(37)
nsims <- 1000
beta0 <- rnorm(nsims,0,1)
beta1 <- rnorm(nsims,0,1)
sigma <- abs(rnorm(nsims,0,1))

# Simulate the log birth weight
dsims <- tibble(log_gest_c = (log(ds$gest)-mean(log(ds$gest)))/sd(log(ds$gest)))

for(i in 1:nsims){
  this_mu <- beta0[i] + beta1[i]*dsims$log_gest_c
  dsims[paste0(i)] <- this_mu + rnorm(nrow(dsims), 0, sigma[i])
}

# Calculate and plot the resulting log birth weight
dsl <- dsims %>%
  pivot_longer(`1`:`1000`, names_to = "sim", values_to = "sim_weight")

dsl %>%
  ggplot(aes(sim_weight)) + geom_histogram(aes(y = ..density..), bins = 20, fill = "turquois
e", color = "black") +
  theme_bw(base_size = 16) +
  geom_vline(xintercept = log(0.3), color = "purple", lwd = 1, lty = 2)+
  geom_vline(xintercept = log(4), color = "purple", lwd = 1, lty = 2)+
  labs(x = "Simulated Log Birth Weight", y="", title = "The resulting distribution of simulat
ed (log) birth weights")
```




```
# Calculate and plot ten simulations of (log) birthweights against gestational age
dsl10 <- dsims %>%
  pivot_longer(`1`:`10`, names_to = "sim", values_to = "sim_weight") %>%
  select(log_gest_c, sim,sim_weight)

dsl10 %>%
  ggplot(aes(x = log_gest_c, y = sim_weight, color = sim)) +
  geom_point()+
  geom_smooth(se = FALSE)+
  theme_bw()+
  labs(x = "Log Gestational Age", y="Log Birth Weight", title = "Ten simulations of (log) bir
thweights against gestational age")
```



Observation:

Based on online information, the baby's birth weight should between 0.3 to 4 kg (assuming from 20 week to full term). This means that the log birth weight should between -1.20 to 1.38 as indicated in the purple line in the first plot. Based on the plot, we can see that there are still many log birth weight that fall out of this range. which is concerning and unrealistic. In the second plot, some simulation shows negative relationship between gestational age and birth body weight, it is contradicting to our EDA as well.

\(\;\)

\(\;\)

Run the model

Now we're going to run Model 1 in Stan. The stan code is in the `code/models` folder.

First, get our data into right form for input into stan.

```
ds$log_weight <- log(ds$birthweight)
ds$log_gest_c <- (log(ds$gest) - mean(log(ds$gest)))/sd(log(ds$gest))

# put into a list
stan_data <- list(N = nrow(ds),
                  log_weight = ds$log_weight,
                  log_gest = ds$log_gest_c)
```

Now fit the model

```
mod1 <- stan(data = stan_data,
             file = here("Lab_6/simple_weight.stan"),
             iter = 500,
             seed = 243)
```

```
summary(mod1)$summary[c("beta[1]", "beta[2]", "sigma"),]
```

```
##              mean      se_mean          sd      2.5%       25%       50%
## beta[1] 1.1624783 8.160385e-05 0.002856578 1.1570200 1.1604786 1.1625011
## beta[2] 0.1437529 8.295075e-05 0.002912236 0.1381284 0.1416970 0.1436747
## sigma   0.1690330 1.113724e-04 0.001902828 0.1652694 0.1677842 0.1690763
##               75%     97.5%     n_eff      Rhat
## beta[1] 1.1644669 1.1681028 1225.3801 0.9978044
## beta[2] 0.1456716 0.1495180 1232.5721 0.9998714
## sigma   0.1702528 0.1727953  291.9066 1.0146111
```

\(\;\)

\(\;\)

Question 3

Based on model 1, give an estimate of the expected birthweight of a baby who was born at a gestational age of 37 weeks.

Answer

Since we use the centered and standardized gestational age here, we first perform the same thing for the new gestational age 37 weeks. Also, since the model is built for log birth weight, we also need to exponentiate the new result to get the birth weight in the original scale.

```
# Calculate the new input into the model
new_gest_age <- (log(37) - mean(log(ds$gest)))/sd(log(ds$gest))
new_gest_age
```

```
## [1] -0.5945826
```

```
# Calculate the expected birth weight of the baby
est_beta0 <- summary(mod1)$summary[c("beta[1]", "beta[2]", "sigma"),][1,1]
est_beta1 <- summary(mod1)$summary[c("beta[1]", "beta[2]", "sigma"),][2,1]
exp_birthweight <- exp(est_beta0 + est_beta1 * new_gest_age)
exp_birthweight
```

```
## [1] 2.935874
```

Thus we get the estimated expected birth weight of a baby who was born at a gestational age of 37 weeks is

2.935874 kg.

\(\;\)

\(\;\)

Question 4

Write a stan model to run Model 2, and run it.

Answer

The stan scripts is in the folder as well under “simple_weight_model2.stan”.

First we prepare the data:

```
# Create indicator variable
ds$preterm_ind <- ifelse(ds$preterm == "Y", 1,0)

# Put into a list
stan_data_model2 <- list(N = nrow(ds),
                         log_weight = ds$log_weight,
                         log_gest = ds$log_gest_c,
                         preterm = ds$preterm_ind,
                         intercept = ds$preterm_ind * ds$log_gest_c)
```

Now we fit the model2

```
mod2 <- stan(data = stan_data_model2,
             file = here("Lab_6/simple_weight_model2.stan"),
             iter = 500,
             seed = 243)

# Save the model result for reloading during comparing result in Q5
save(mod2, file = "mod2_my.Rdata")
```

The summary of coefficient estimation will show up in Q5

\(\;\)

\(\;\)

Question 5

For reference I have uploaded some model 2 results. Check your results are similar.

```
load("mod2.Rda")
summary(mod2)$summary[c(paste0("beta[", 1:4, "]"), "sigma"),]
```

```
##              mean      se_mean          sd       2.5%        25%       50%
## beta[1] 1.1697241 1.385590e-04 0.002742186 1.16453578 1.16767109 1.1699278
## beta[2] 0.5563133 5.835253e-03 0.058054991 0.43745504 0.51708255 0.5561553
## beta[3] 0.1020960 1.481816e-04 0.003669476 0.09459462 0.09997153 0.1020339
## beta[4] 0.1967671 1.129799e-03 0.012458398 0.17164533 0.18817091 0.1974114
## sigma   0.1610727 9.950037e-05 0.001782004 0.15784213 0.15978020 0.1610734
##               75%     97.5%     n_eff      Rhat
## beta[1] 1.1716235 1.1750167 391.67359 1.0115970
## beta[2] 0.5990427 0.6554967  98.98279 1.0088166
## beta[3] 0.1044230 0.1093843 613.22428 0.9978156
## beta[4] 0.2064079 0.2182454 121.59685 1.0056875
## sigma   0.1623019 0.1646189 320.75100 1.0104805
```

Answer

Here is a summary of the results from my version of model 2

```
# Reload my model result
load("mod2_my.Rdata")
summary(mod2)$summary[c("beta[1]", "beta[2]","beta[3]","beta[4]", "sigma"),]
```

```
##              mean      se_mean          sd       2.5%        25%       50%
## beta[1] 1.1696329 8.021297e-05 0.002705139 1.16410383 1.16791913 1.1695478
## beta[2] 0.1018545 1.111662e-04 0.003424916 0.09508969 0.09961365 0.1019319
## beta[3] 0.5620695 3.406560e-03 0.062560942 0.43112646 0.52265217 0.5614275
## beta[4] 0.1982641 6.964438e-04 0.012807594 0.17144797 0.18979854 0.1986269
## sigma   0.1611971 8.785429e-05 0.001825790 0.15774991 0.15994557 0.1611909
##               75%     97.5%     n_eff     Rhat
## beta[1] 1.1714725 1.1748162 1137.3388 1.000638
## beta[2] 0.1040358 0.1087724  949.1923 1.002232
## beta[3] 0.6039584 0.6839901  337.2675 1.015352
## beta[4] 0.2062635 0.2232005  338.1917 1.013325
## sigma   0.1623667 0.1649513  431.8927 1.004553
```

Based on my output (displayed in Q4) and the given output, my results and the given results are similar. Noticed that the beta2 and beta3 are flipped in the given results as clarified in Piazza.

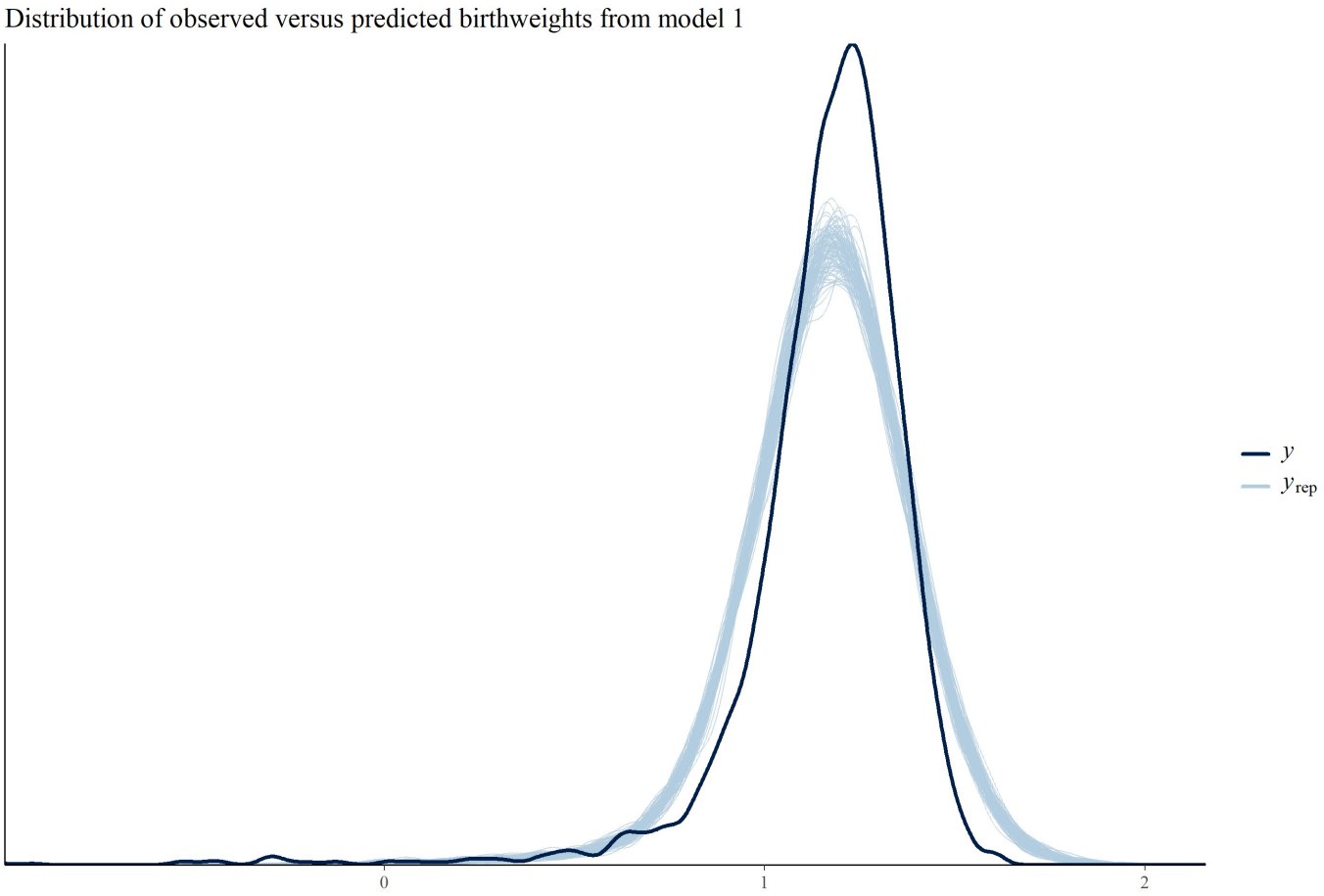
\(\;\)

\(\;\)

PPCs

Now we've run two candidate models let's do some posterior predictive checks. The `bayesplot` package has a lot of inbuilt graphing functions to do this. For example, let's plot the distribution of our data (y) against 100 different datasets drawn from the posterior predictive distribution:

```
set.seed(1856)
y <- ds$log_weight
yrep1 <- extract(mod1)[["log_weight_rep"]]
samp100 <- sample(nrow(yrep1), 100)
ppc_dens_overlay(y, yrep1[samp100, ])  + ggtitle("Distribution of observed versus predicted b
irthweights from model 1")
```



\(\;\)

\(\;\)

Question 6

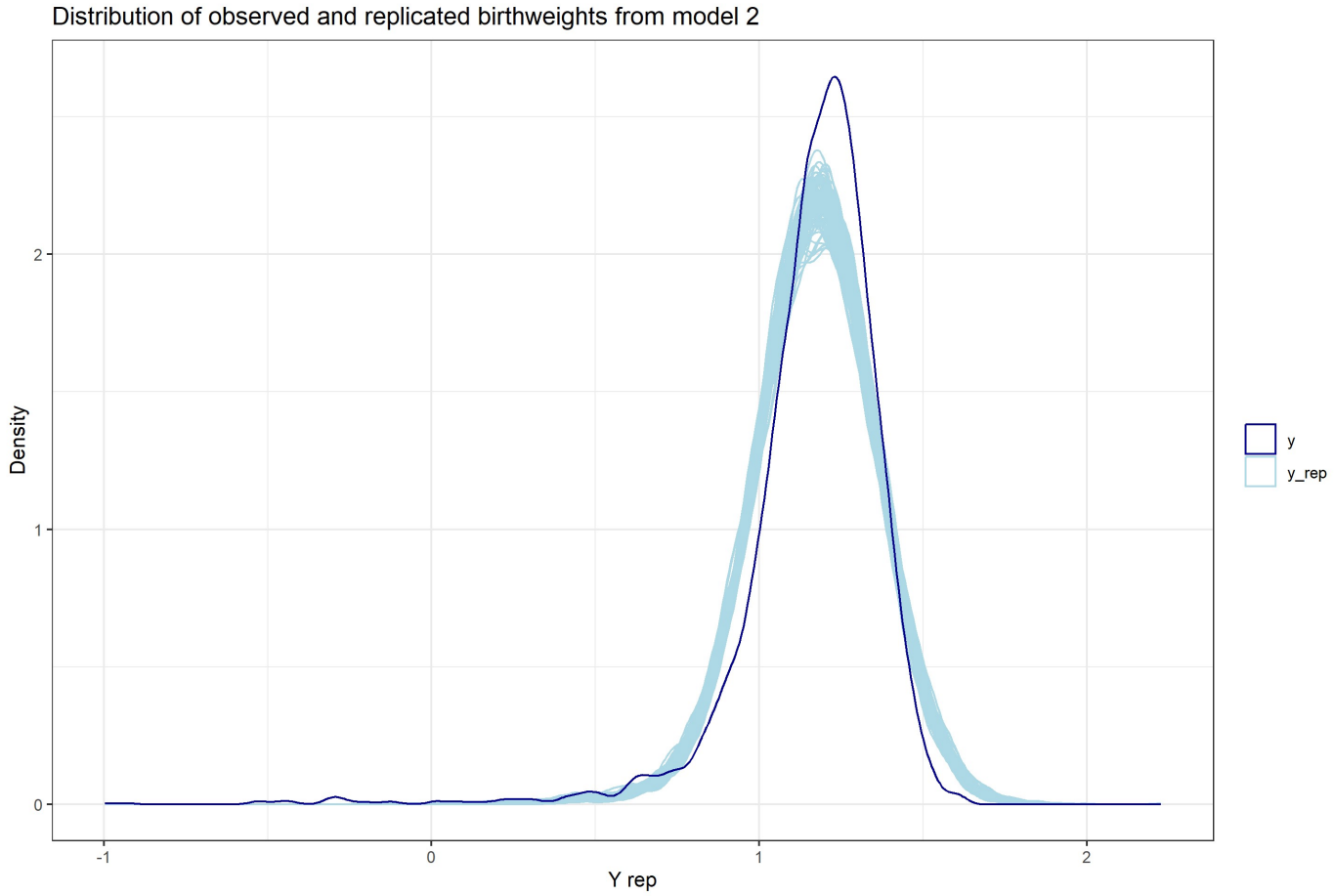
Make a similar plot to the one above but for model 2, and **not** using the bayes plot in built function (i.e. do it yourself just with `geom_density`)

Answer

Note, the model 2 results we are using here is the on we ran rather than the reference results.

Here is a similar plot for model 2 using geom_density:

```
# Set up the dataset from model2
set.seed(3729)
y <- ds$log_weight
yrep2 <- extract(mod2)[["log_weight_rep"]]
samp2100 <- sample(nrow(yrep2), 100)
rownames(yrep2) <- 1:nrow(yrep2)
dr <- as_tibble(t(yrep2))
dr <- dr %>% bind_cols(i = 1:nrow(ds), log_weight_obs = log(ds$birthweight))

# Turn into long format for plotting
dr <- dr %>%
  pivot_longer(-(i:log_weight_obs), names_to = "sim", values_to ="y_rep")

# Filter the sample and plot the density
dr %>%
  filter(sim %in% samp2100) %>%
  ggplot(aes(y_rep, group = sim)) +
  geom_density(alpha = 0.2, aes(color = "y_rep")) +
  geom_density(data = ds %>% mutate(sim = 1),
               aes(x = log(birthweight), col = "y")) +
  scale_color_manual(name = "",
                     values = c("y" = "darkblue",
                                "y_rep" = "lightblue")) +
  labs(x = "Y rep", y="Density",
       title = "Distribution of observed and replicated birthweights from model 2") +
  theme_bw()
```



\(\;\)

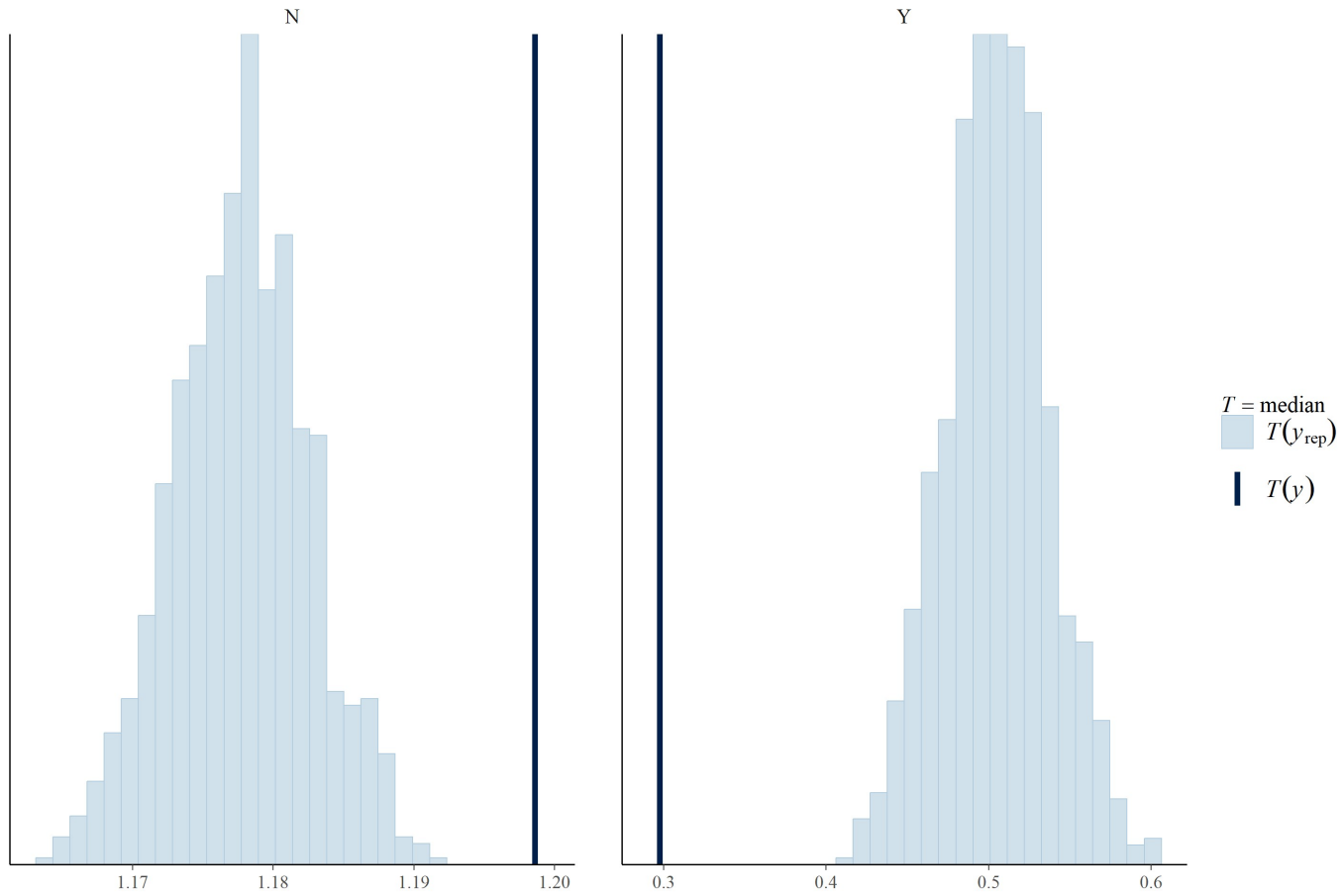
\(\;\)

Test statistics

We can also look at some summary statistics in the PPD versus the data, again either using `bayesplot` – the function of interest is `ppc_stat` or `ppc_stat_grouped` – or just doing it ourselves using ggplot.

E.g. medians by prematurity for Model 1

```
ppc_stat_grouped(ds$log_weight, yrep1, group = ds$preterm, stat = 'median')
```



\(\;\)

\(\;\)

Question 7

Use a test statistic of the proportion of births under 2.5kg. Calculate the test statistic for the data, and the posterior predictive samples for both models, and plot the comparison (one plot per model).

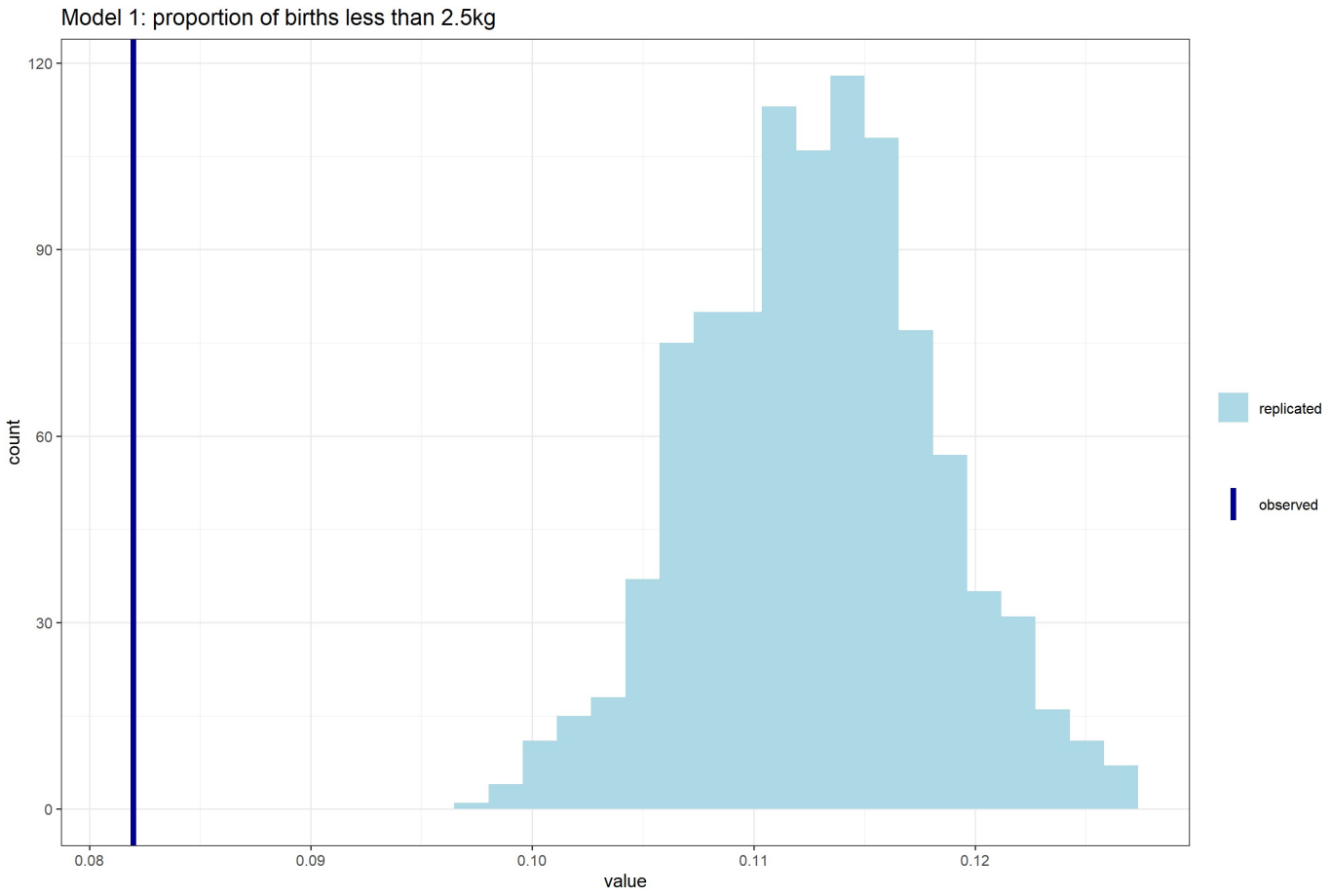
Answer

We use the test statistics of the proportion of births under 2.5kg and calculate it for original data, model 1 and model 2 data. Then we plot the results:

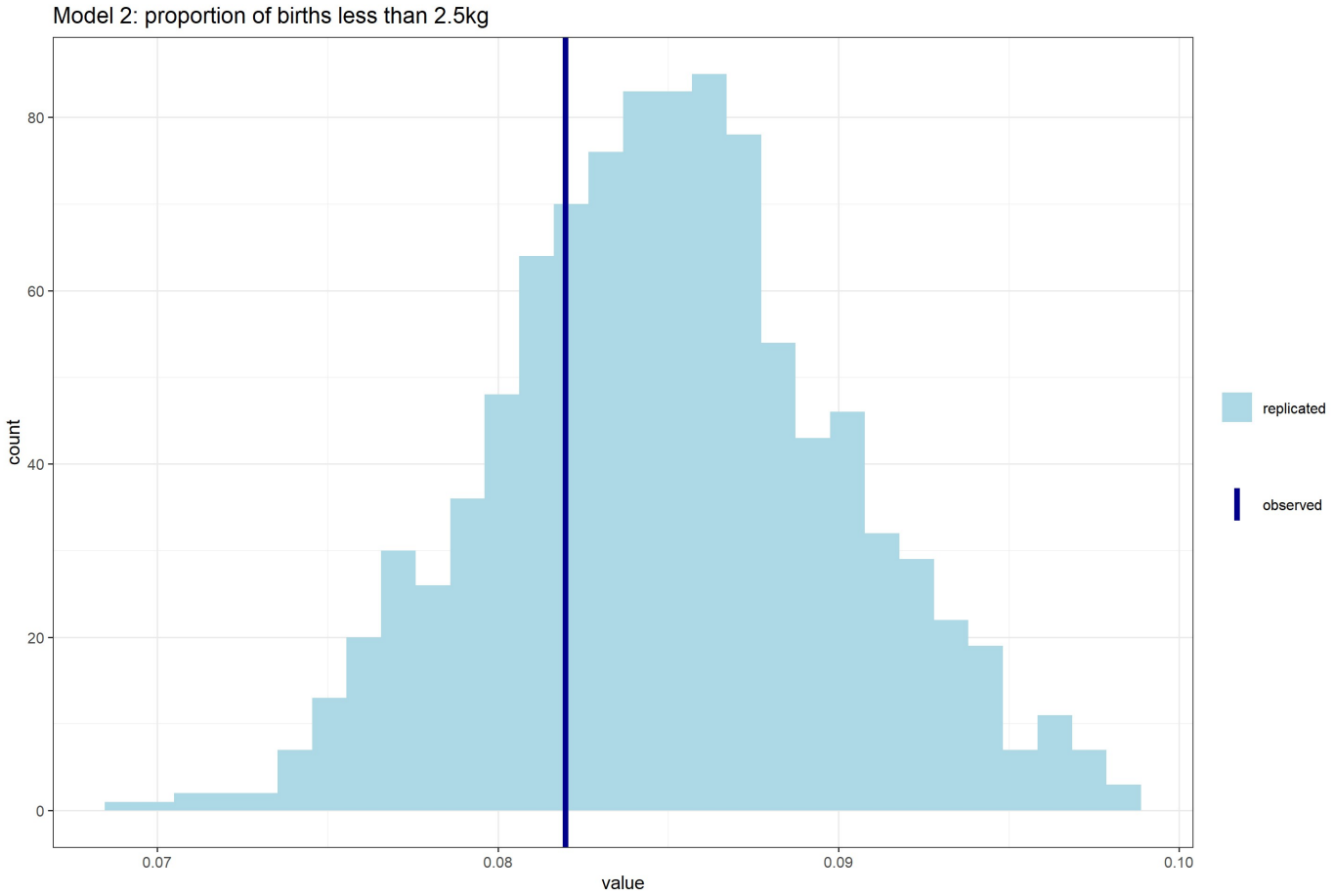
```
# The test statistics for the data
y <- ds$log_weight
t_y <- mean(y<=log(2.5))
t_y
```

```
## [1] 0.08198855
```

```
# Calculate the test statistics for both models
t_y_rep <- sapply(1:nrow(yrep1), function(i) mean(yrep1[i,]<=log(2.5)))
t_y_rep_2 <- sapply(1:nrow(yrep2), function(i) mean(yrep2[i,]<=log(2.5)))

# Plot for Model 1
ggplot(data = as_tibble(t_y_rep), aes(value)) +
    geom_histogram(aes(fill = "replicated")) +
    geom_vline(aes(xintercept = t_y, color = "observed"), lwd = 1.5) +
  ggtitle("Model 1: proportion of births less than 2.5kg") +
  theme_bw(base_size = 10) +
  scale_color_manual(name = "",
                     values = c("observed" = "darkblue"))+
  scale_fill_manual(name = "",
                     values = c("replicated" = "lightblue"))
```



```
# Plot for Model 2
ggplot(data = as_tibble(t_y_rep_2), aes(value)) +
    geom_histogram(aes(fill = "replicated")) +
    geom_vline(aes(xintercept = t_y, color = "observed"), lwd = 1.5) +
  ggtitle("Model 2: proportion of births less than 2.5kg") +
  theme_bw(base_size = 10) +
  scale_color_manual(name = "",
                     values = c("observed" = "darkblue"))+
  scale_fill_manual(name = "",
                     values = c("replicated" = "lightblue"))
```



Observation:

Here we get the test statistics for the data is 0.08198855, that is the proportion of babies that has a body weight under 2.5kg is around 8.2%. Based on the two plots, we can see that the observed proportion is more likely coming from the model2 based on the distribution of the proportion of model 2. The observed results is much smaller than the results predicted using model 1.

\(\;\)

\(\;\)

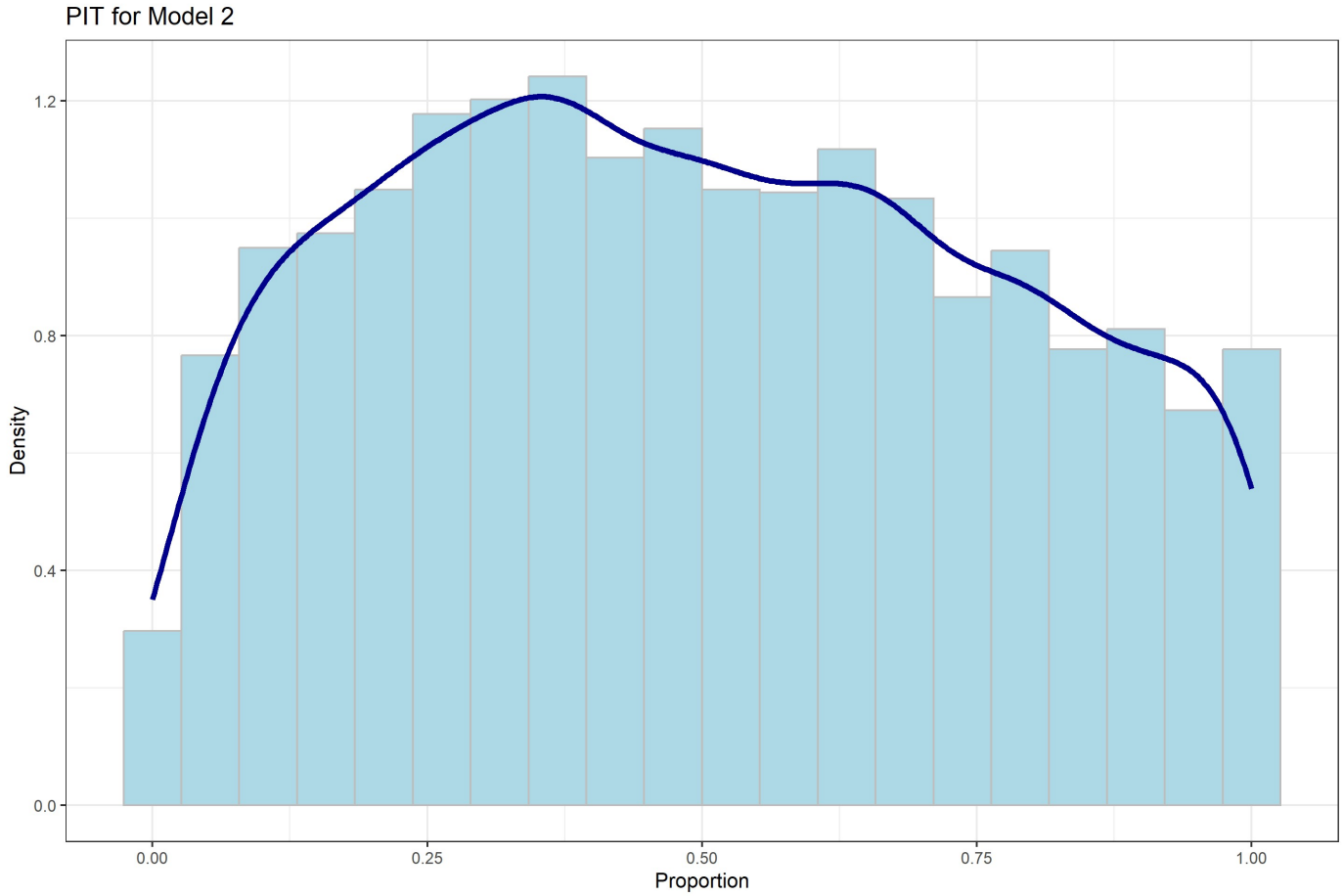
Bonus question (not required)

Create your own PIT histogram “from scratch” for Model 2.

Answer

I use the whole model 2 result and this is what I get:

```
# Set up the dataset from model2
set.seed(3729)
yrep2 <- extract(mod2)[["log_weight_rep"]]
samp2100 <- sample(nrow(yrep2), 100)
rownames(yrep2) <- 1:nrow(yrep2)
d2 <- as_tibble(t(yrep2))
d2 <- d2 %>% bind_cols(i = 1:nrow(ds), log_weight_obs = log(ds$birthweight))

# Turn into long format for plotting
d2 <- d2 %>%
  pivot_longer(-(i:log_weight_obs), names_to = "sim", values_to ="y_rep")

# Create indicator variable for the result
d2$ind <- ifelse(d2$log_weight_obs <= d2$y_rep,1,0)

# Calculate the mean which is the proportion
d2_summary <- d2 %>%
  group_by(i) %>%
  summarize(prop = mean(ind))

# Plotting
d2_summary %>%
  ggplot(aes(x = prop)) +
  geom_histogram(aes(y=..density..), bins = 20, fill = "lightblue", color = "grey")+
  geom_density(color = "darkblue", lwd = 1.5)+
  labs(x = "Proportion", y="Density",
       title = "PIT for Model 2") +
  theme_bw()
```



\(\;\)

\(\;\)

Question 8

Based on the original dataset, choose one (or more) additional covariates to add to the linear regression model. Run the model in Stan, and compare with Model 2 above on at least 2 posterior predictive checks.

Answer

Based on my EDA, I believe it is also reasonable to bring the variable “sex” into the model. There are some differences between baby boys and baby girls and I want to keep that into my model3. Thus this will be the form of my model 3:

\[\log(y_i) \sim N(\beta_1 + \beta_2\log(x_i) + \beta_3 z_i + \beta_4\log(x_i) z_i + \beta_5s_i, \sigma^2) \] where

- \(y_i\) is weight in kg
- \(x_i\) is gestational age in weeks, CENTERED AND STANDARDIZED
- \(z_i\) is preterm (0 or 1, if gestational age is less than 32 weeks)
- \(s_i\) is sex (0 or 1, for female or male)

Now we set up the model in R and stan, the stan script can be found in the folder as well

```
# Create indicator variable for sex
ds$sex_ind <- ifelse(ds$preterm == "M", 1,0)

# Put into a list
stan_data_model3 <- list(N = nrow(ds),
                         log_weight = ds$log_weight,
                         log_gest = ds$log_gest_c,
                         preterm = ds$preterm_ind,
                         intercept = ds$preterm_ind * ds$log_gest_c,
                         sex = ds$sex_ind)
```

Now we fit the model3

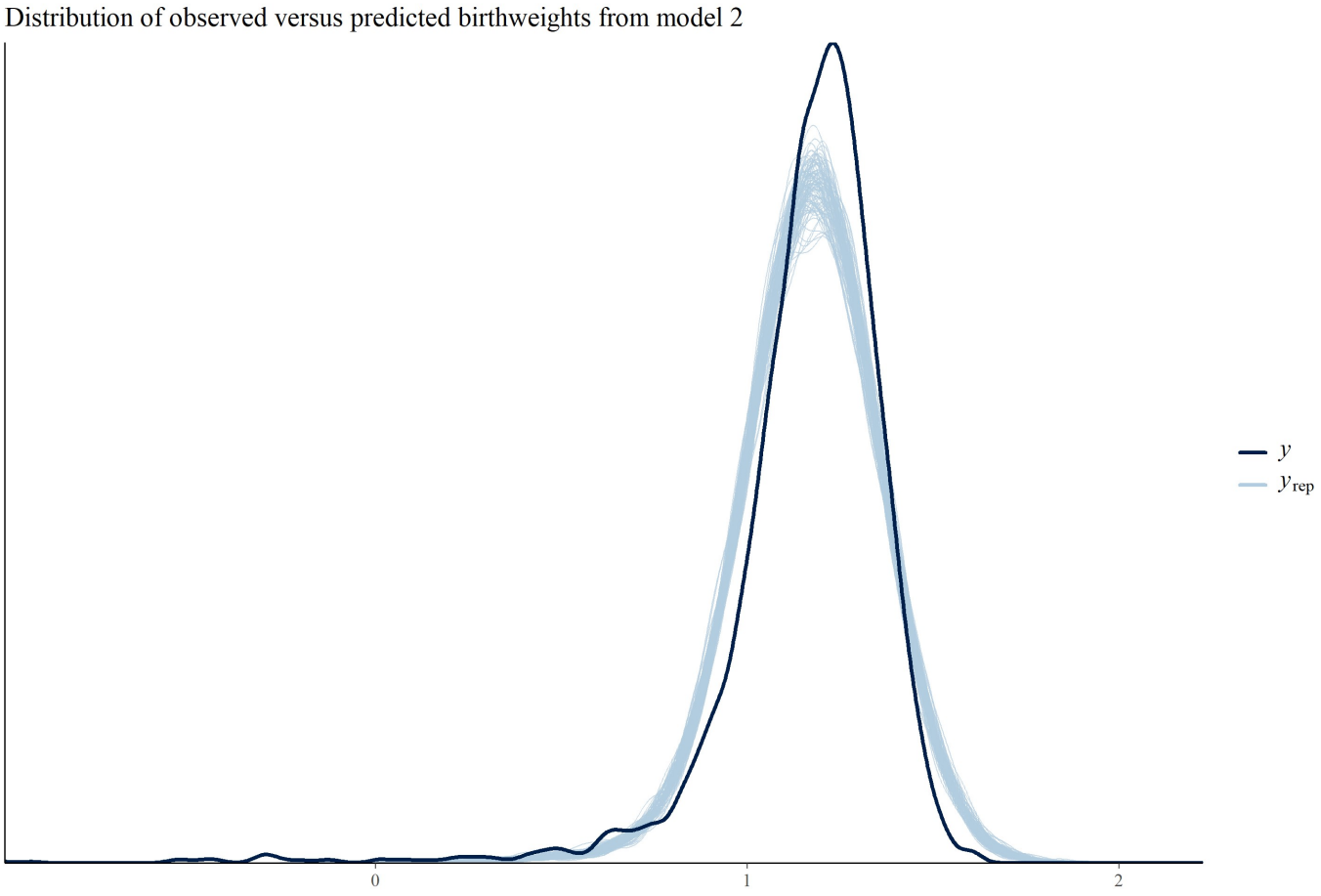
```
mod3 <- stan(data = stan_data_model3,
             file = here("Lab_6/simple_weight_model3.stan"),
             iter = 500,
             seed = 243)
```

Now we compare the results with model 2 on at least two posterior predictive checks:

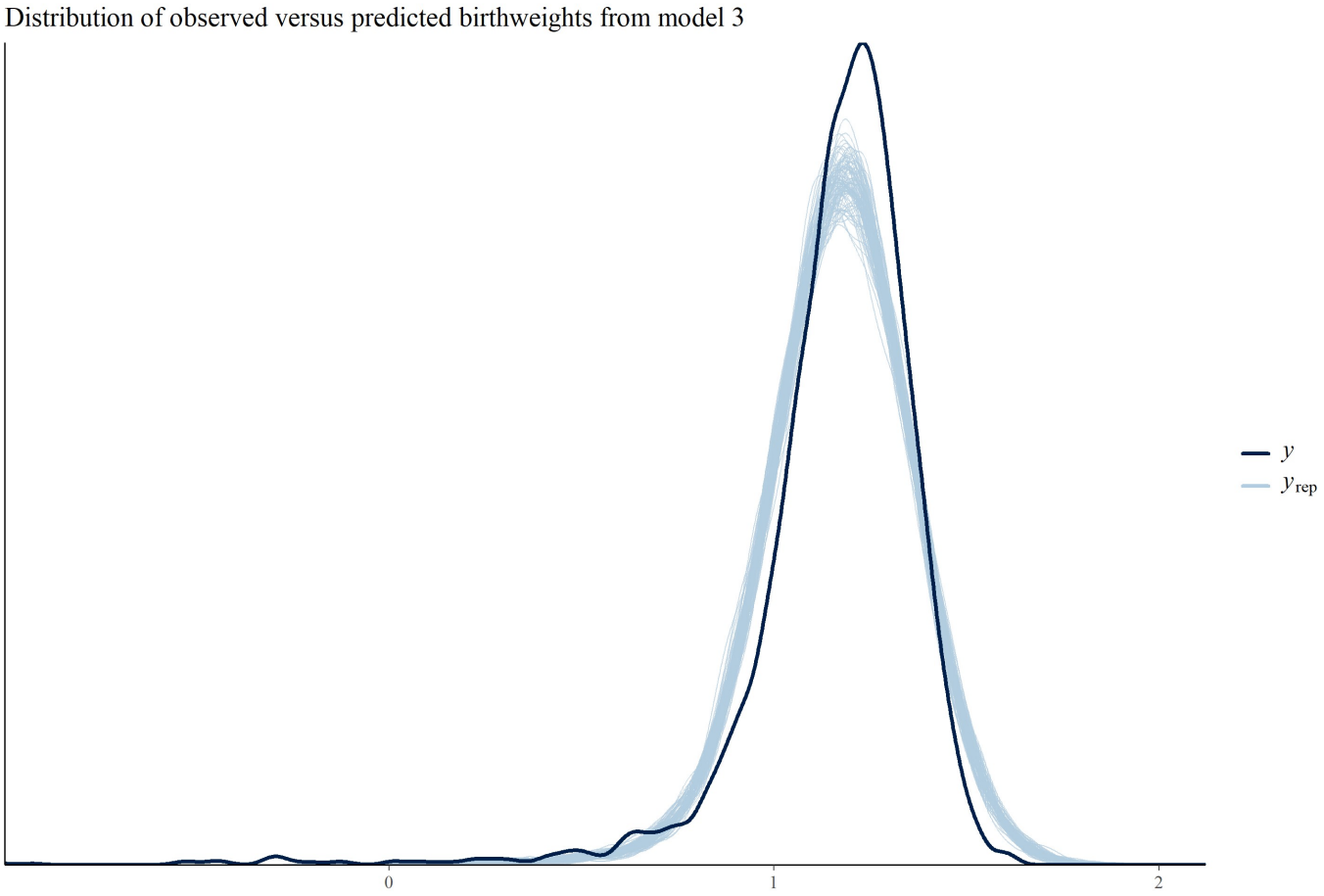
1. The distribution of our data (y) against 100 different datasets drawn from the posterior predictive distribution

Here we provide the plot from both model 2 and model 3

```
set.seed(3729)

# Extract the data
y <- ds$log_weight

yrep2 <- extract(mod2)[["log_weight_rep"]]
samp2100 <- sample(nrow(yrep2), 100)

yrep3 <- extract(mod3)[["log_weight_rep"]]
samp3100 <- sample(nrow(yrep3), 100)

# Plot the data
ppc_dens_overlay(y, yrep2[samp2100, ])  + ggtitle("Distribution of observed versus predicted
birthweights from model 2")
```




```
ppc_dens_overlay(y, yrep3[samp3100, ])  + ggtitle("Distribution of observed versus predicted
birthweights from model 3")
```

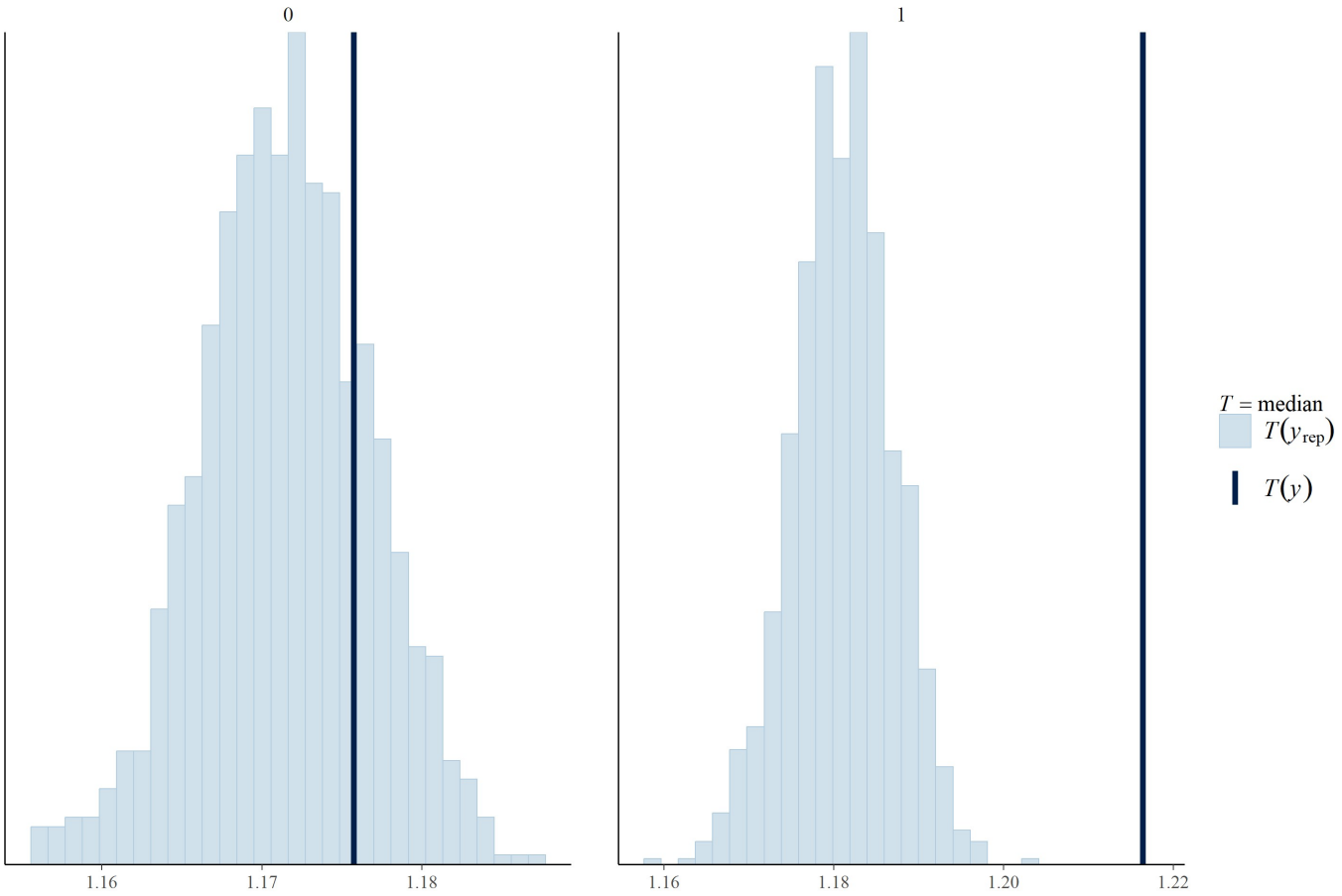


From the plot, we didn't see too much difference between the two model. Both do not deviate from the original data that much.

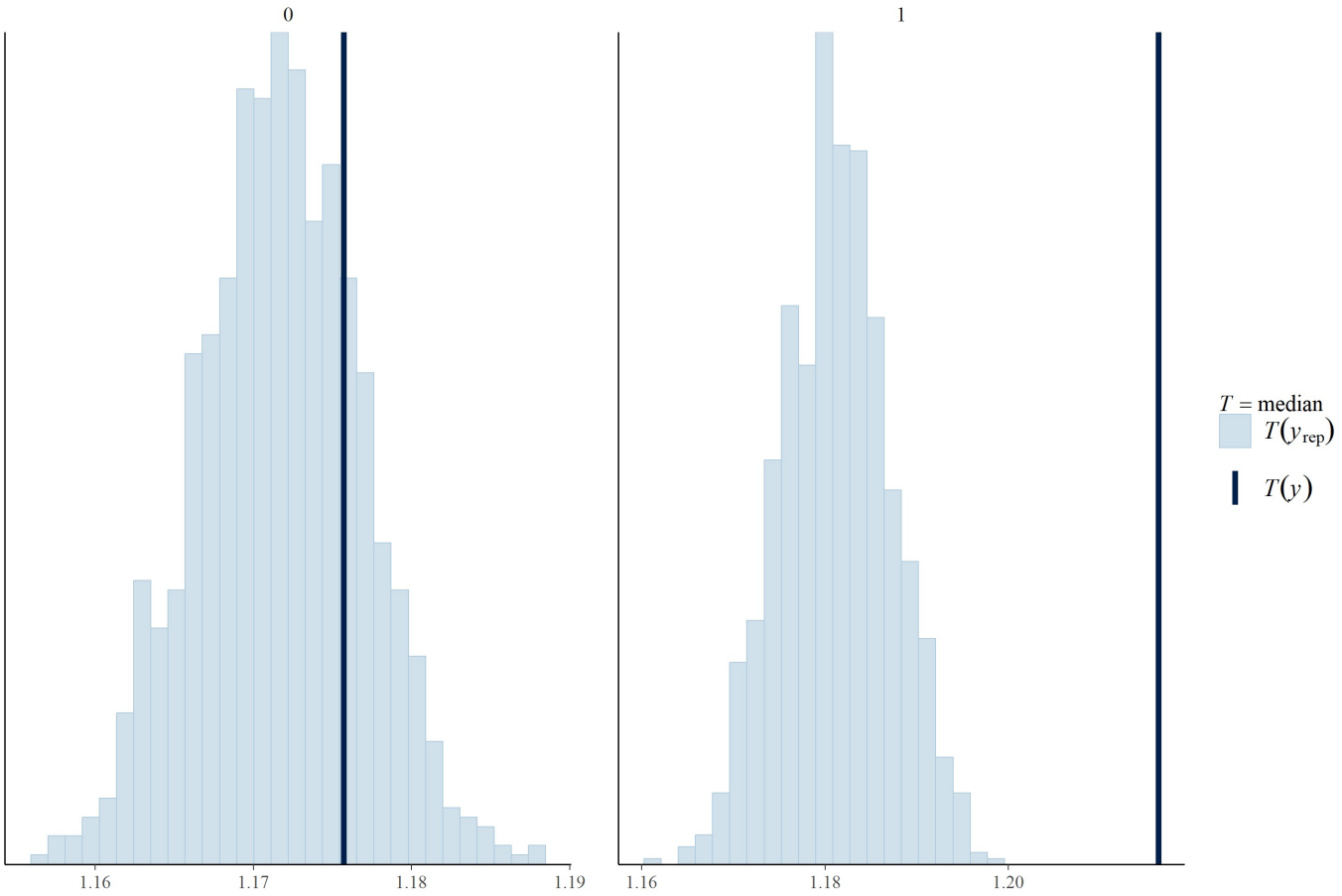
2. Test Statistics: median of different education between the two model

Here we create a new variable by grouping education into two level, those who completed a higher education and those who didn't. The separation rule is based on the code of education in the column "meduc".

```
# Create the group for education level
ds$edu_ind <- ifelse(ds$meduc > 4, 1,0)

# Compare the results by different education group
ppc_stat_grouped(ds$log_weight, yrep2, group = ds$edu_ind, stat = 'median')
```




```
ppc_stat_grouped(ds$log_weight, yrep3, group = ds$edu_ind, stat = 'median')
```

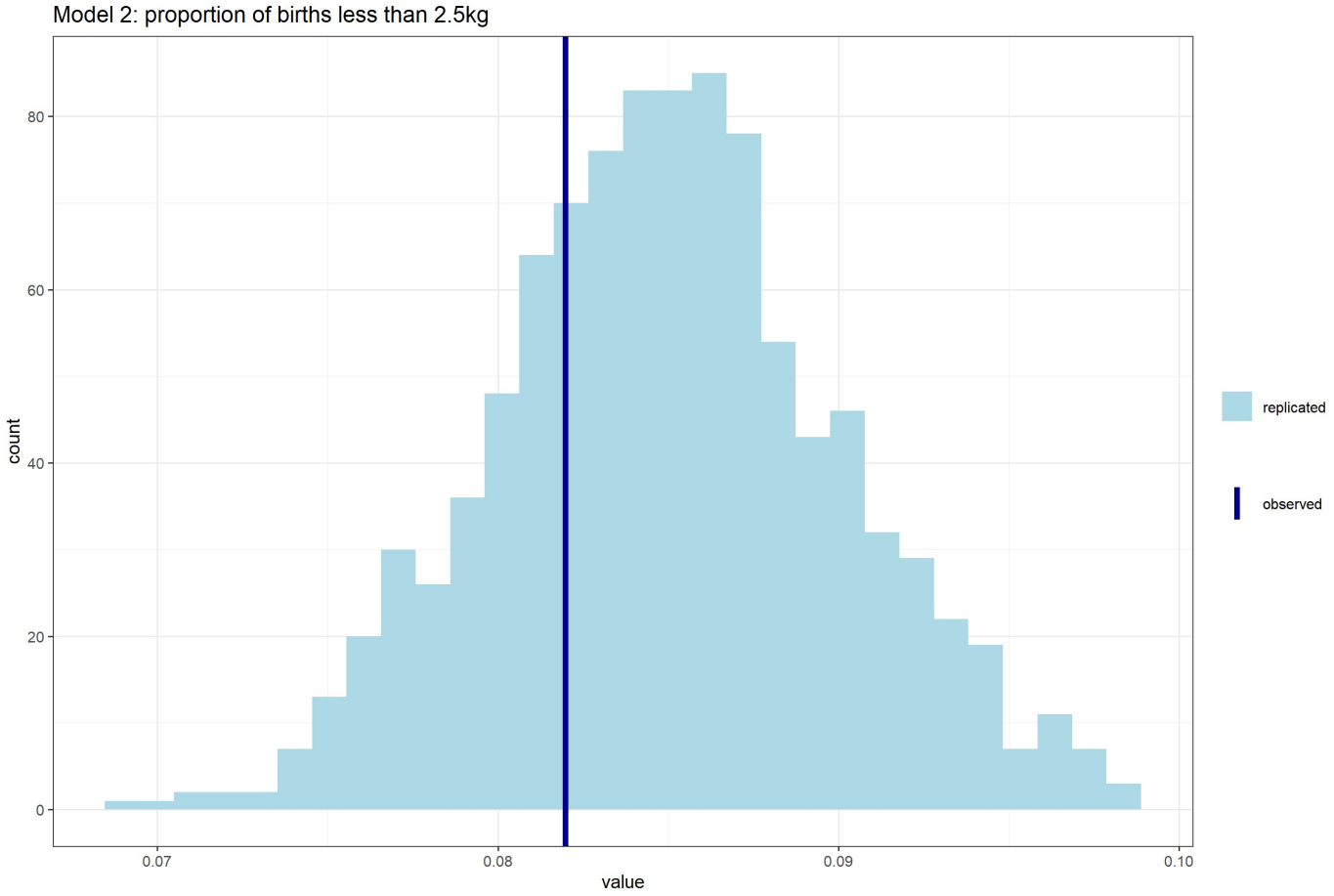


Based on the above plot, we can see that both model does not work well for those who completed higher education, but both are fine for the other category. There is no significant difference between these two models.

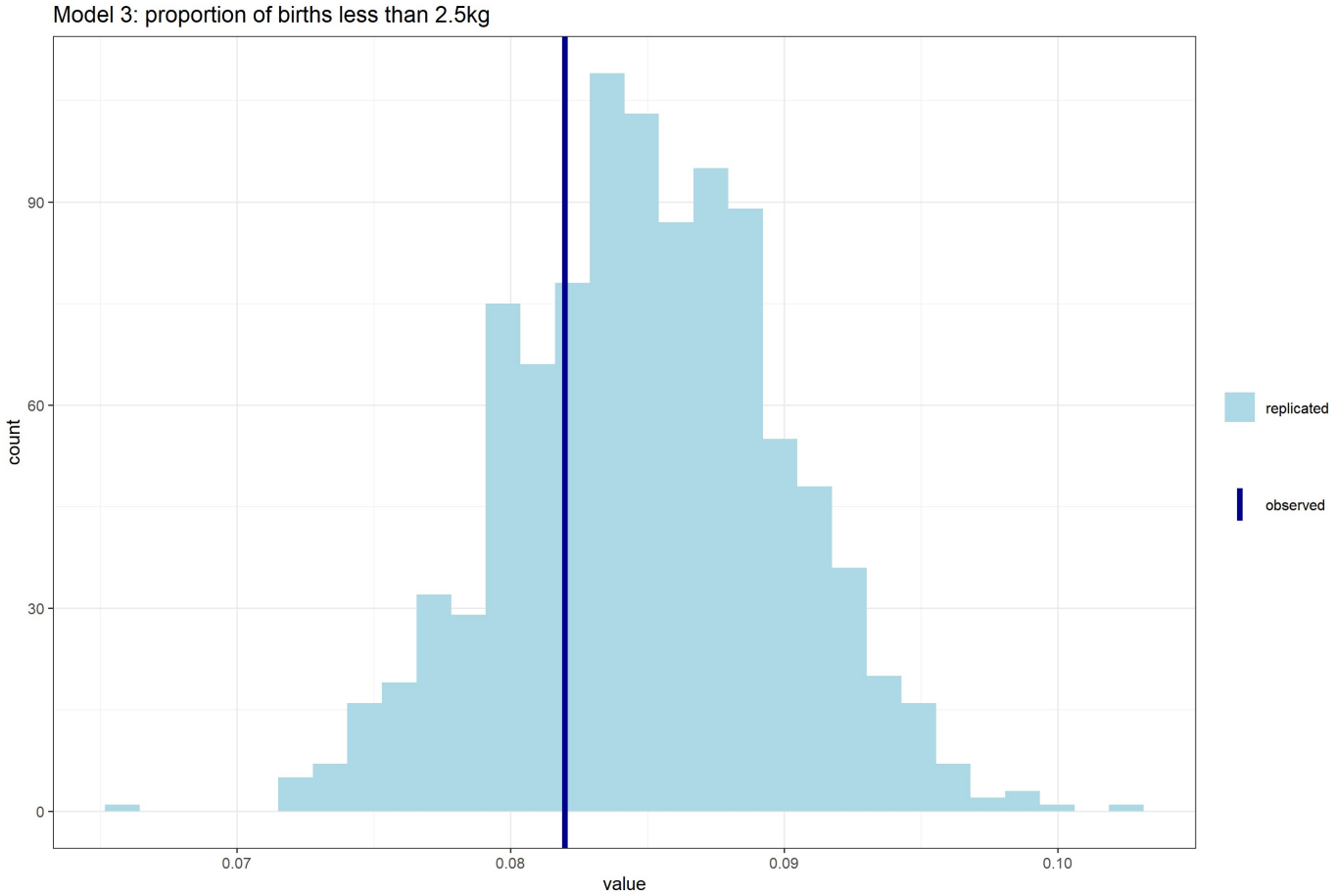
3. Test Statistics: the proportion of births under 2.5kg

This is the test statistics we used before and we want to use it again to compare model 2 and model 3:

```
# The test statistics for the data
y <- ds$log_weight
t_y <- mean(y<=log(2.5))

# Calculate the test statistics for both models
t_y_rep_2 <- sapply(1:nrow(yrep2), function(i) mean(yrep2[i,]<=log(2.5)))
t_y_rep_3 <- sapply(1:nrow(yrep3), function(i) mean(yrep3[i,]<=log(2.5)))


# Plot for Model 2
ggplot(data = as_tibble(t_y_rep_2), aes(value)) +
    geom_histogram(aes(fill = "replicated")) +
    geom_vline(aes(xintercept = t_y, color = "observed"), lwd = 1.5) +
  ggtitle("Model 2: proportion of births less than 2.5kg") +
  theme_bw(base_size = 10) +
  scale_color_manual(name = "",
                     values = c("observed" = "darkblue"))+
  scale_fill_manual(name = "",
                     values = c("replicated" = "lightblue"))
```




```
# Plot for Model 3
ggplot(data = as_tibble(t_y_rep_3), aes(value)) +
    geom_histogram(aes(fill = "replicated")) +
    geom_vline(aes(xintercept = t_y, color = "observed"), lwd = 1.5) +
  ggtitle("Model 3: proportion of births less than 2.5kg") +
  theme_bw(base_size = 10) +
  scale_color_manual(name = "",
                     values = c("observed" = "darkblue"))+
  scale_fill_manual(name = "",
                     values = c("replicated" = "lightblue"))
```



Well, based on the two plots, both works pretty well for this test statistics. We can see that model 3 shows a more concentrated distribution, which is kind of a slight improvement compared to model 2.

Overall, based on all the checks above, my conclusion is that there is no significant difference between model 2 and model 3. Having “sex” as a new addition to model 2 improves the model a little bit and the results are similar with model 2. It would be nice to have it in the model as it helps somehow.