Lab 6

21/02/23

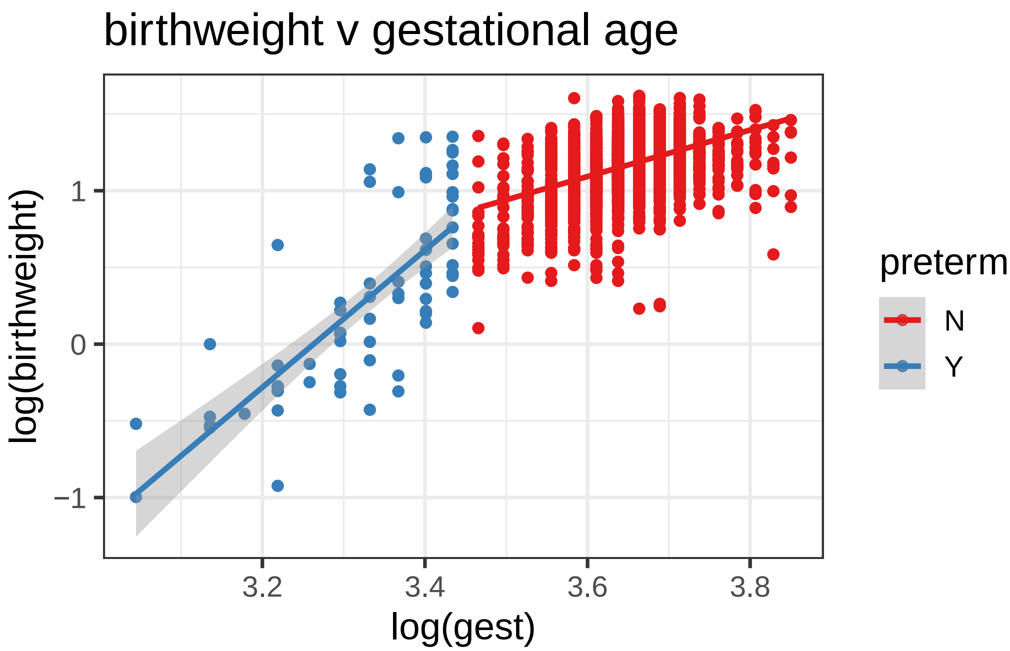
```
library(tidyverse)
library(here)
# for bayes stuff
library(rstan)
library(bayesplot)
library(loo)
library(tidybayes)

ds <- read_rds(here("births_2017_sample.RDS"))
```

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

Question 1

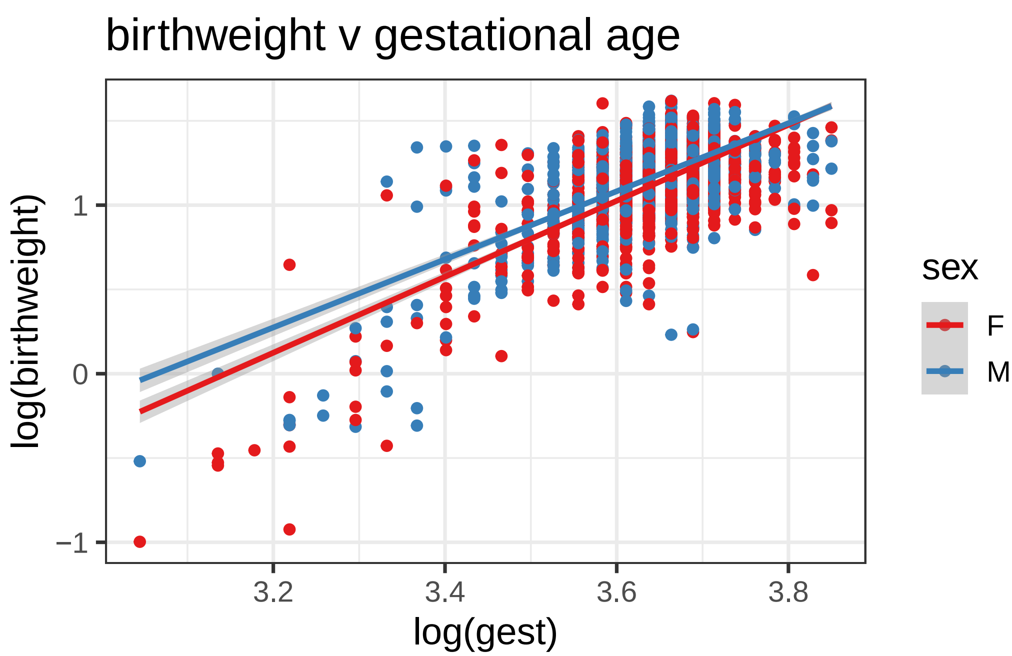
The following plot shows a scatterplot of the log gestational age and log birth weight, split by whether the baby was born prematurely. We can see some evidence of a relationship between log gestational age and log birth weight, and of interaction between log gestational age and whether the baby was born prematurely.

```
ds %>%
  ggplot(aes(log(gest), log(birthweight), color = preterm)) +
  geom_point() + geom_smooth(method = "lm") +
  scale_color_brewer(palette = "Set1") +
  theme_bw(base_size = 14) +
  ggtitle("birthweight v gestational age")
```



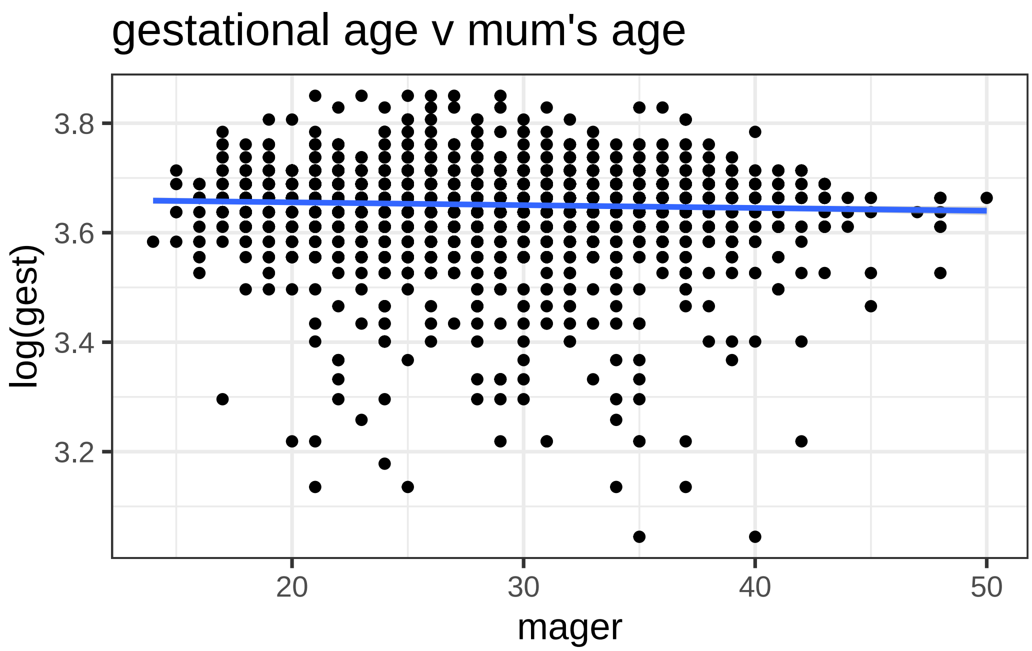
The following plot shows a scatterplot of the log gestational age and log birth weight, split by the sex of the baby. Here, we do not see as much evidence of interaction between log gestational age and the sex of the baby. We see that males with low gestational age may weigh a bit more than females at the same gestational age, but that this difference reduces as gestational age increases.

```
ds %>%
  ggplot(aes(log(gest), log(birthweight), color = sex)) +
  geom_point() + geom_smooth(method = "lm") +
  scale_color_brewer(palette = "Set1") +
  theme_bw(base_size = 14) +
  ggtitle("birthweight v gestational age")
```



The following plot shows a scatterplot of the mum's age and log gestational age. We see a bit of evidence of a relationship between the two variables, since gestational age seems to decrease slightly age mum's age increases.

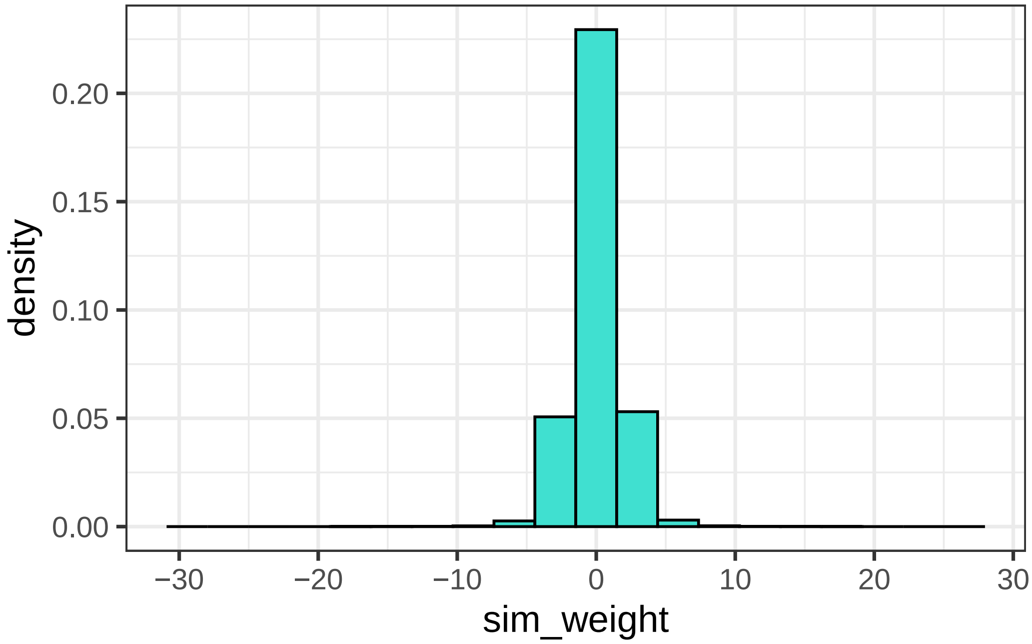
```
ds %>%
  ggplot(aes(mager, log(gest))) +
  geom_point() + geom_smooth(method = "lm") +
  scale_color_brewer(palette = "Set1") +
  theme_bw(base_size = 14) +
  ggtitle("gestational age v mum's age")
```



Question 2

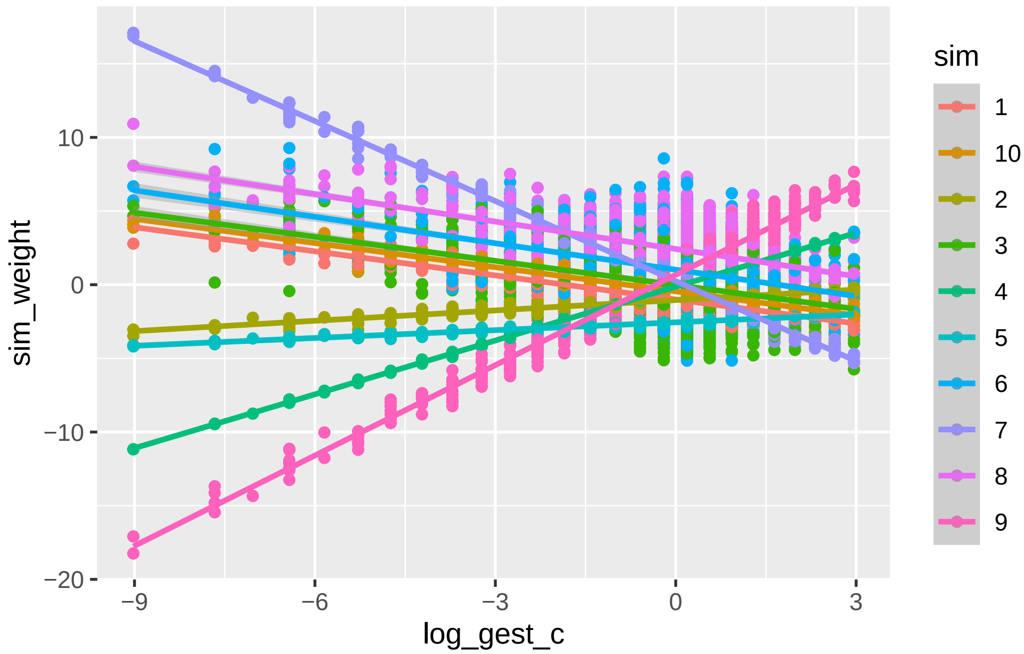
```
set.seed(123)
nsims <- 1000
sigma <- abs(rnorm(nsims, 0, 1))
beta0 <- rnorm(nsims, 0, 1)
beta1 <- rnorm(nsims, 0, 1)

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])
}

dsl <- dsims %>%
  pivot_longer(`1`:`10`, names_to = "sim", values_to = "sim_weight")

dsl1 <- dsims %>%
  pivot_longer(`1`:`1000`, names_to = "sim", values_to = "sim_weight")

dsl1 %>%
```

```
ggplot(aes(sim_weight)) + geom_histogram(aes(y = ..density..), bins = 20,
                                         fill = "turquoise", color = "black") +
theme_bw(base_size = 14)
```



The next plot shows ten simulations of (log) birthweights plotted against gestational age.

```
dsl %>%
  ggplot(aes(x=log_gest_c,y=sim_weight,color=sim))+geom_point()+
  geom_smooth(method = "lm")
```



Run the model

First, we run model 1.

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

N <- nrow(ds)
log_weight <- ds$log_weight
log_gest_c <- ds$log_gest_c
preterm <- ifelse(ds$preterm=="Y", 1, 0)

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

```
mod1 <- stan(data = stan_data,
             file = here("simple_weight.stan"),
             iter = 1000,
```

```
              seed = 243)
```

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

```
             mean      se_mean          sd      2.5%       25%       50%
beta[1] 1.1625735 5.794303e-05 0.002748364 1.1569407 1.1608332 1.1625648
beta[2] 0.1437137 5.192697e-05 0.002701544 0.1383608 0.1419081 0.1437506
sigma   0.1689425 7.055689e-05 0.001917996 0.1649500 0.1676547 0.1690340
              75%     97.5%     n_eff      Rhat
beta[1] 1.1643470 1.1677751 2249.8106 0.9996791
beta[2] 0.1455499 0.1488306 2706.6873 0.9991123
sigma   0.1702690 0.1725518  738.9525 1.0118298
```

Question 3

Since the model is given by

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

First, we standardize the log gestational age of 37:

```
(log(37) - mean(log(ds$gest)))/sd(log(ds$gest))
```

```
[1] -0.5945826
```

The log of the estimate of the expected birthweight of a baby who was born at a gestational age of 37 weeks is given by $1.16 + (0.14 * (-0.59)) = 1.08$, so the estimate is $e^{1.08} = 2.94$ kg.

Question 4

```
mod2a <- stan(data = stan_data,
              file = "simple_weight_preterm_int.stan",
              iter = 1000,
              seed = 263)
```

```
summary(mod2a)$summary[c(paste0("beta[", 1:4, "]"), "sigma"),]
```

```
             mean      se_mean          sd       2.5%        25%       50%
beta[1] 1.1696558 5.392043e-05 0.002652912 1.16458836 1.16790317 1.1697285
beta[2] 0.1019343 9.559925e-05 0.003667535 0.09468693 0.09939467 0.1019442
beta[3] 0.5608750 2.422406e-03 0.064685262 0.43300277 0.51777804 0.5618463
beta[4] 0.1980600 5.094155e-04 0.013396208 0.17111575 0.18922796 0.1982794
sigma   0.1613514 4.872986e-05 0.001831730 0.15763277 0.16012143 0.1613597
              75%     97.5%     n_eff      Rhat
beta[1] 1.1713899 1.1747510 2420.6903 0.9991869
beta[2] 0.1043464 0.1091868 1471.7685 1.0007776
beta[3] 0.6058877 0.6850282  713.0450 1.0020175
beta[4] 0.2072074 0.2239998  691.5434 1.0010371
sigma   0.1625821 0.1648373 1412.9686 0.9997262
```

Question 5

From the summary statistics below, we can see that the results are similar, except it seems like beta[2] and beta[3] have been switched between the two models.

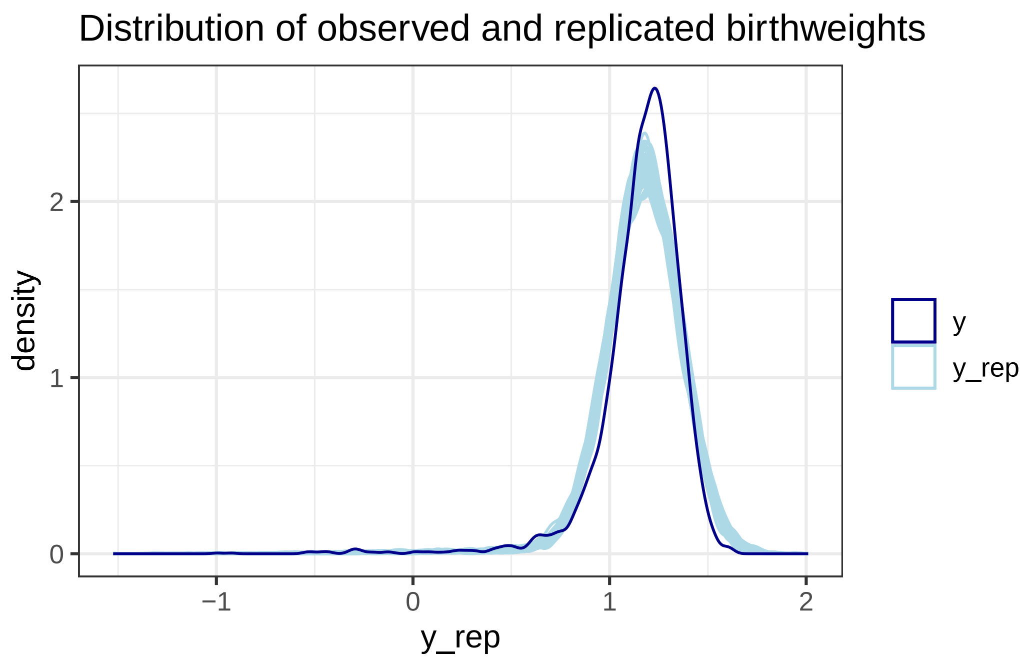
```r
load(here("mod2.Rda"))
summary(mod2)$summary[c(paste0("beta[", 1:4, "]"), "sigma"),]
```

```
             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
```

Question 6

```r
set.seed(1856)
yrep1 <- extract(mod1)[["log_weight_rep"]]
yrep2 <- extract(mod2a)[["log_weight_rep"]]
```

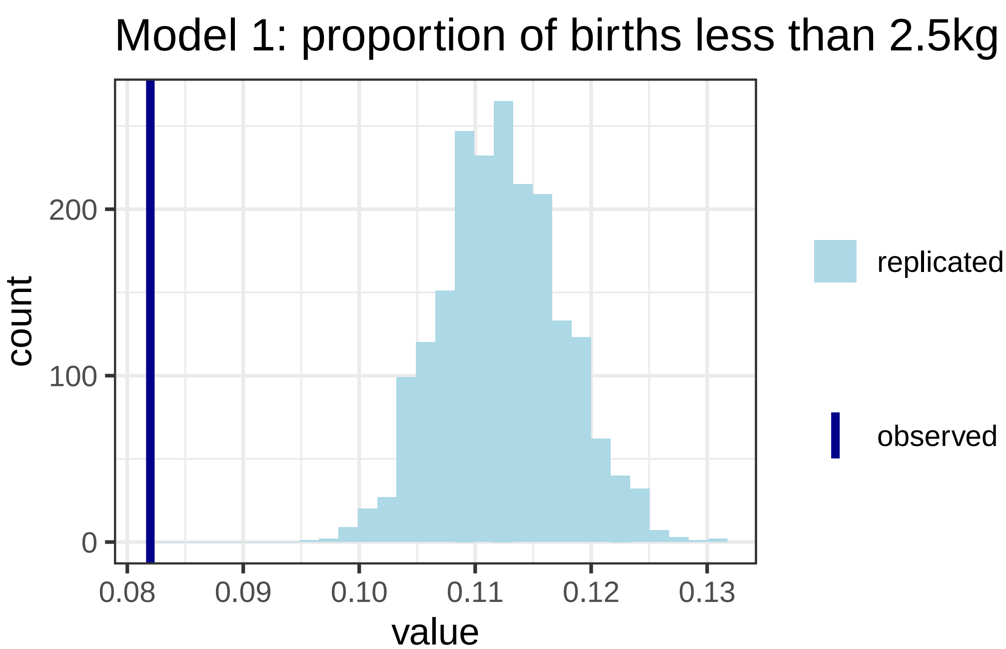
```r
samp100 <- sample(nrow(yrep2), 100)
# first, get into a tibble
rownames(yrep2) <- 1:nrow(yrep2)
dr <- as_tibble(t(yrep2))
dr <- dr %>% bind_cols(i = 1:N, log_weight_obs = log(ds$birthweight))

# turn into long format; easier to plot
dr <- dr %>%
  pivot_longer(-(i:log_weight_obs), names_to = "sim", values_to ="y_rep")

# filter to just include 100 draws and plot!
dr %>%
  filter(sim %in% samp100) %>%
  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")) +
  ggtitle("Distribution of observed and replicated birthweights") +
  theme_bw(base_size = 12)
```



Question 7

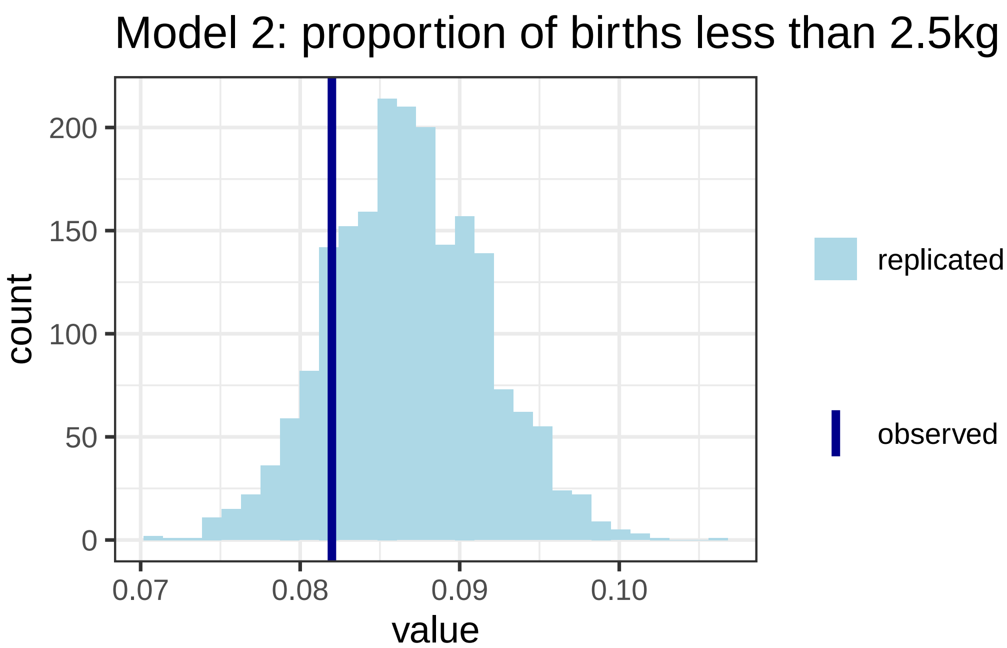
We plot the test statistic of the proportion of births under 2.5kg for the data and the posterior predictive samples for model 1.

```
y <- log_weight
t_y <- mean(y<=log(2.5))
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)))

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 = 14) +
  scale_color_manual(name = "",
                     values = c("observed" = "darkblue"))+
  scale_fill_manual(name = "",
                     values = c("replicated" = "lightblue"))
```



We do the same thing for model 2.

```r
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 = 14) +
  scale_color_manual(name = "",
                     values = c("observed" = "darkblue"))+
  scale_fill_manual(name = "",
                     values = c("replicated" = "lightblue"))
```



Question 8

We add a term for the sex of the baby and an interaction between the sex and gestation of the baby to the model:

$$\log(y_i) \sim N(\beta_1 + \beta_2 \log(x_i) + \beta_2 s_i + \beta_3 \log(x_i) s_i, \sigma^2)$$

- y_i is weight in kg
- x_i is gestational age in weeks, centered and standardized
- s_i is sex (0 for female, 1 for male)

We run the model in Stan:

```r
sex <- ifelse(ds$sex=="M", 1, 0)

# put into a list
stan_data3 <- list(N = N,
                   log_weight = log_weight,
                   log_gest = log_gest_c,
                   sex=sex)
```

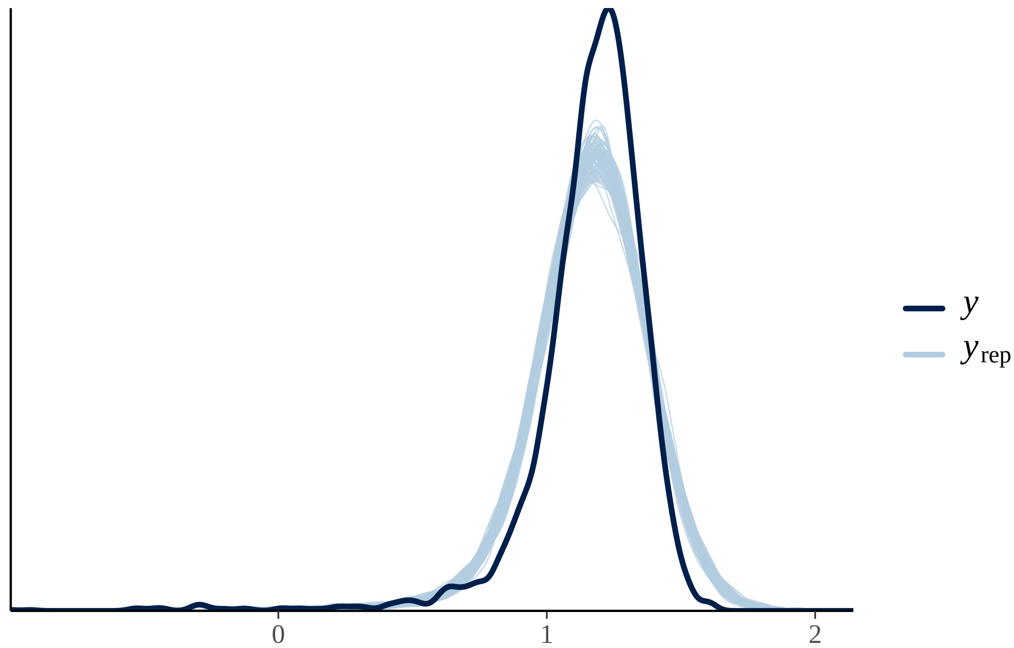
```r
mod3 <- stan(data = stan_data3,
             file = "lab6q8.stan",
             iter = 1000,
             seed = 263)
```

```r
summary(mod3)$summary[c(paste0("beta[", 1:4, "]"), "sigma"),]
```

```
               mean      se_mean          sd        2.5%         25%
beta[1]  1.14029158 9.511273e-05 0.003759362  1.13309747  1.13771225
beta[2]  0.15158030 9.364352e-05 0.003638240  0.14432112  0.14916156
beta[3]  0.04449608 1.361263e-04 0.005345387  0.03422208  0.04073658
beta[4] -0.01597786 1.433960e-04 0.005335531 -0.02647306 -0.01968030
sigma    0.16729286 4.774567e-05 0.001845547  0.16373324  0.16604456
                50%         75%        97.5%    n_eff      Rhat
beta[1]  1.14021471  1.14288620  1.147565763 1562.252 1.0005749
beta[2]  0.15159357  0.15398402  0.158606000 1509.480 1.0000357
beta[3]  0.04448984  0.04821316  0.054799279 1541.965 1.0016636
beta[4] -0.01597486 -0.01243044 -0.005501541 1384.462 0.9995216
sigma    0.16722373  0.16857334  0.170899757 1494.108 0.9999597
```

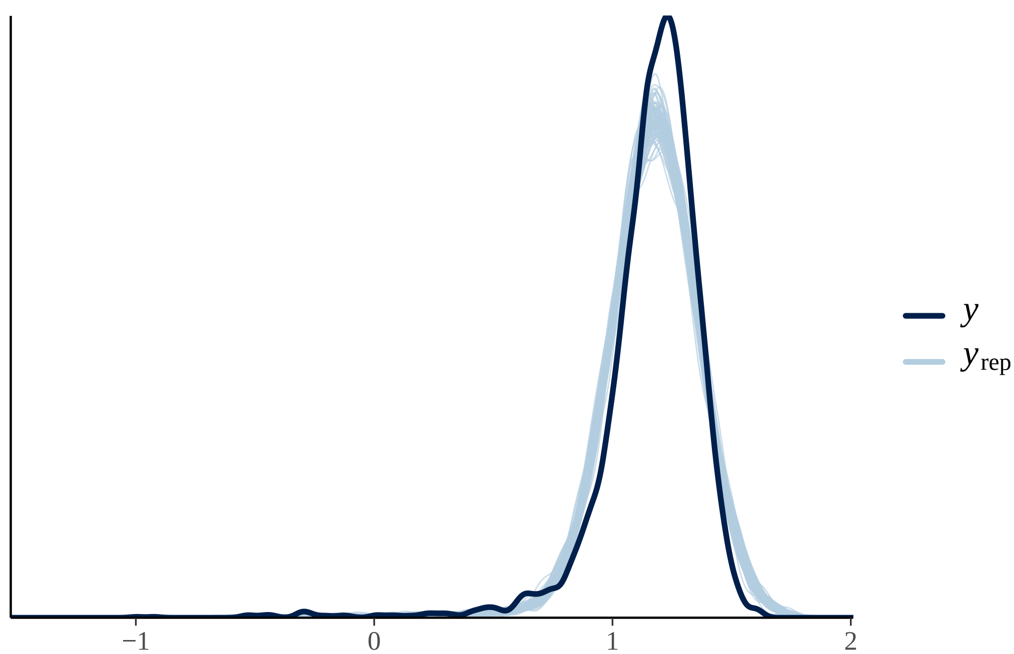
First, we extract the samples from the posterior predictive distribution and compare the densities of 100 sampled datasets to the actual data.

```r
yrep3 <- extract(mod3)[["log_weight_rep"]]
ppc_dens_overlay(y, yrep3[samp100, ])
```



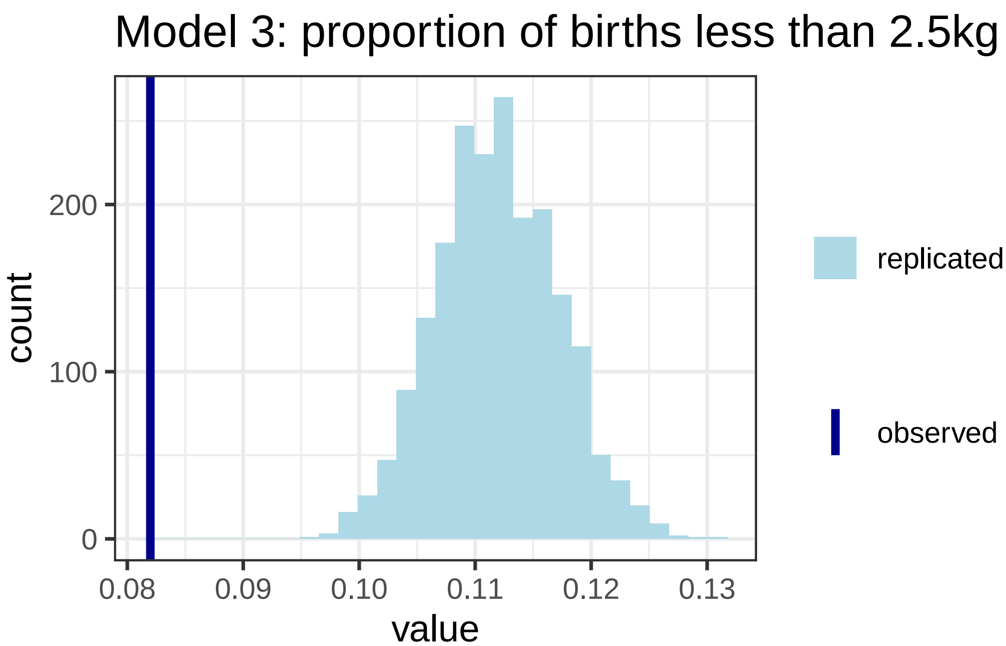
We show the same plot for model 2 and find that the densities of the sampled datasets for model 2 are closer to the actual data then for our new model.

```
ppc_dens_overlay(y, yrep2[samp100, ])
```



Next, we calculate the proportion of babies who have a weight less than 2.5kg (considered low birth weight) in each of the replicated datasets, and compare them to the proportion in the data.

```
t_y_rep_3 <- sapply(1:nrow(yrep3), function(i) mean(yrep3[i,]<=log(2.5)))
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 = 14) +
  scale_color_manual(name = "",
                     values = c("observed" = "darkblue"))+
  scale_fill_manual(name = "",
                     values = c("replicated" = "lightblue"))
```



We do the same thing for model 2 and find that model 2 still does better here.

```
t_y_rep_3 <- sapply(1:nrow(yrep2), function(i) mean(yrep2[i,]<=log(2.5)))
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 = 14) +
```

```
scale_color_manual(name = "",
                   values = c("observed" = "darkblue"))+
scale_fill_manual(name = "",
                  values = c("replicated" = "lightblue"))
```