Lab 6: Visualizing the Bayesian Workflow

J. Arturo Esquivel

18/02/23

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

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)
```

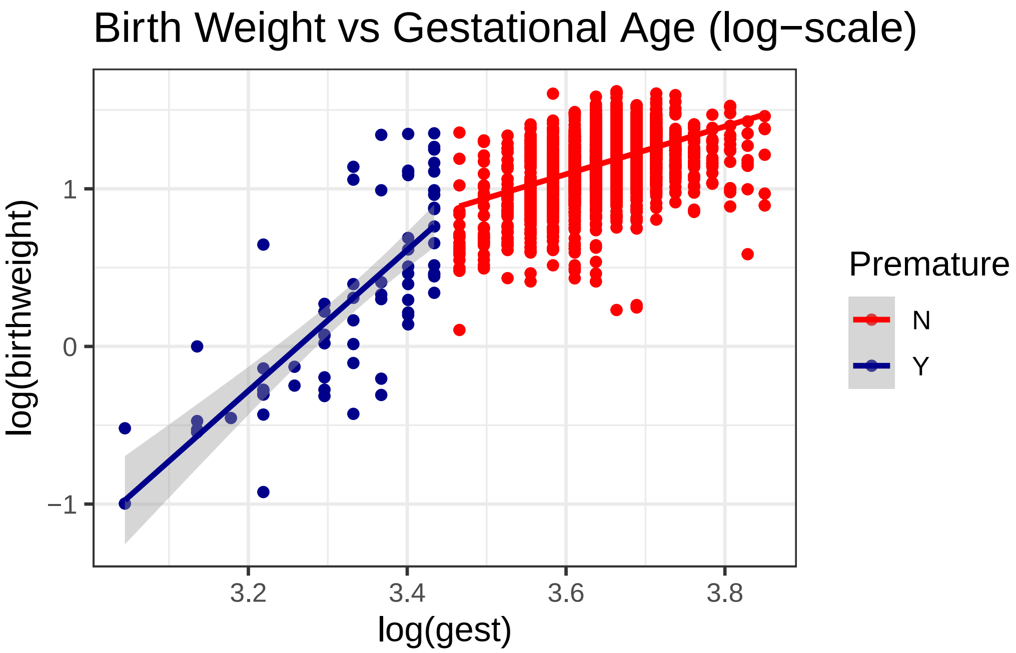
Question 1

The plot below shows the relationship between gestational age and birth weight (log-scale), using color to identify premature cases. In the plot in can be seen that:

1 there seems to be a strong linear relationship between both variables;

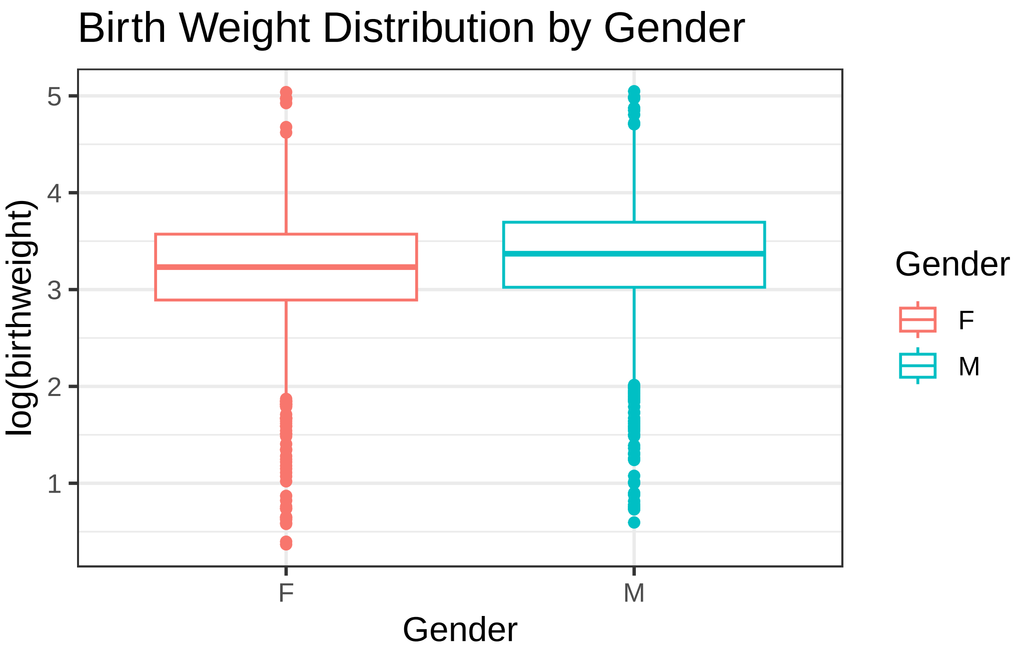
2 such relationship changes considerably between premature and non-premature babies.

```
ds |>
  ggplot(aes(log(gest), log(birthweight), color = preterm)) +
  geom_point() +
  geom_smooth(method = "lm") +
  scale_color_manual(name = "Premature", values = c("Y" = "darkblue", "N" = "red"))  +
  theme_bw(base_size = 13) +
  ggtitle("Birth Weight vs Gestational Age (log-scale)")
```



The plot below shows birth weight distribution by gender. It looks like on average male babies weight slightly more than female ones.

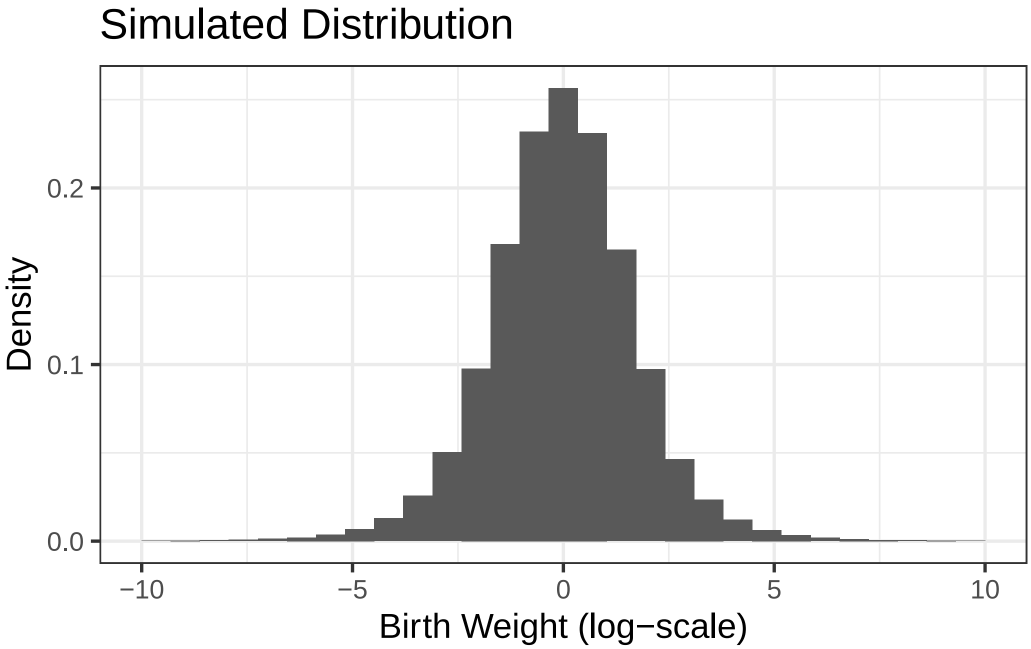
```
ggplot(ds, aes(sex, birthweight, color = sex)) +
  geom_boxplot() +
  labs(title = "Birth Weight Distribution by Gender",
       x = "Gender", y = "log(birthweight)",
       color = "Gender") +
  theme_bw(base_size = 13)  +
  scale_color_discrete( labels = c('F', 'M'))
```



Question 2

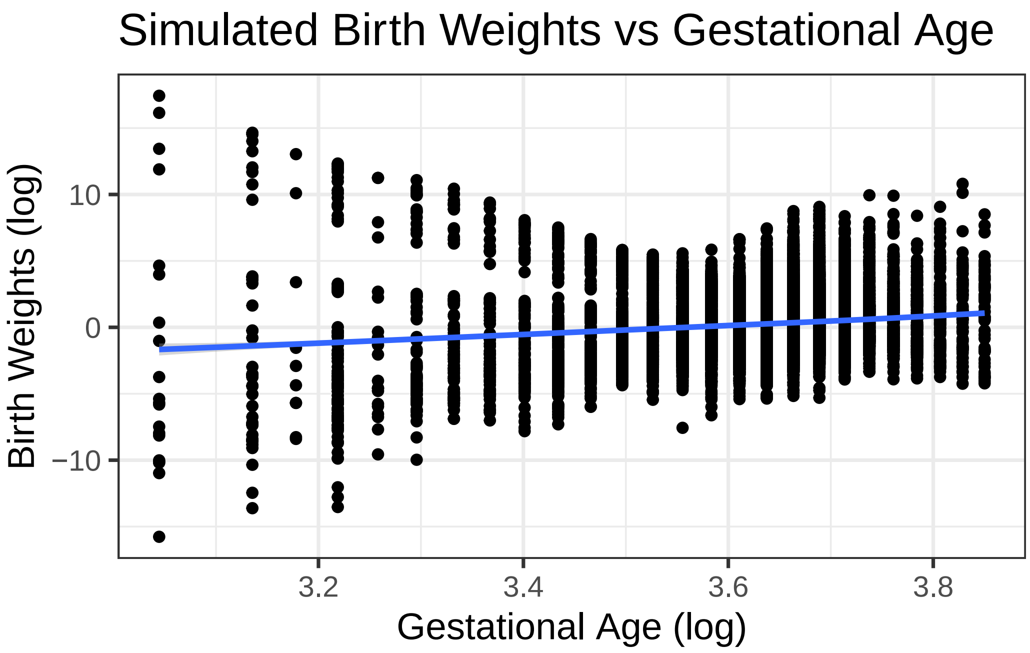
The resulting distribution of simulated (log) birth weights is shown below.

```
l_norm_gest <- (log(ds$gest)-mean(log(ds$gest)))/sd(log(ds$gest))

n <- 1000
m <- length(l_norm_gest)
set.seed(29)
sigma <- abs(rnorm(n, 0, 1))
beta_0 <- rnorm(n, 0, 1)
beta_1 <- rnorm(n, 0, 1)

l_bw <- matrix(0, n, m)
for(i in 1:n){
  l_bw[i, ] <- rnorm(m, rep(beta_0[i], m) + beta_1[i]*l_norm_gest, rep(sigma[i], m))
}

data.frame(l_bw) |>
  pivot_longer(cols = everything()) |>
  ggplot(aes(value)) +
  geom_histogram(aes(y = after_stat(density))) +
  theme_bw(base_size = 13) +
  xlim(c(-10, 10)) +
  labs(title = "Simulated Distribution",
       x = "Birth Weight (log-scale)", y = "Density")
```



A plot of ten simulations of (log) birth weights against gestational age is shown below. With such vague priors, it is clear that the relationship between birth weight and gestational age is lost due to the variance of the distribution. Hopefully, that will be corrected in the posteriors.

```
ten_sim <- data.frame(l_bw[1:10, ]) |>
  pivot_longer(cols = everything())

cbind(ten_sim, rep(log(ds$gest), 10)) |>
  ggplot(aes(`rep(log(ds$gest), 10)`, value)) +
  geom_point() +
  geom_smooth() +
  labs(title = "Simulated Birth Weights vs Gestational Age",
       x = "Gestational Age (log)",
       y = "Birth Weights (log)") +
  theme_bw(base_size = 14)
```



```
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)

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

Question 3

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

pos_estimates <- mod1 |>
  gather_draws(beta[condition]) |>
  group_by(.draw) |>
  mutate(.value = ifelse(condition==2, .value*val,  .value)) |>
  summarise(est = sum(.value))

estimate <- exp(mean(pos_estimates$est))
```

The posterior (mean) estimate for the expected birth weight of a baby born after 37 weeks of gestation is 2.94 kg.

Question 4

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

mod_2 <- stan(data = stan_data,
             file = here("simple_weight2.stan"),
             iter = 500,
             seed = 16)
```

The summary for model 2 is shown below.

```
summary(mod_2)[["summary"]][c(paste0("beta[",1:4, "]"), "sigma"),]
```

```
             mean      se_mean          sd       2.5%        25%       50%
beta[1] 1.1696488 7.853910e-05 0.002860062 1.16374780 1.16771561 1.1697053
beta[2] 0.1019725 1.322962e-04 0.003387189 0.09539317 0.09960448 0.1019775
beta[3] 0.5668000 3.947972e-03 0.061128139 0.44844101 0.52771631 0.5663092
beta[4] 0.1991842 8.338113e-04 0.012695467 0.17438074 0.19064932 0.1993908
sigma   0.1613575 7.285451e-05 0.001754763 0.15814043 0.16012919 0.1612583
              75%     97.5%     n_eff      Rhat
beta[1] 1.1716422 1.1751566 1326.1081 0.9980518
beta[2] 0.1043939 0.1083681  655.5172 1.0012470
beta[3] 0.6039692 0.6959524  239.7365 1.0196105
beta[4] 0.2070213 0.2257584  231.8258 1.0176058
sigma   0.1625793 0.1649615  580.1284 1.0113129
```

Question 5

```
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
```

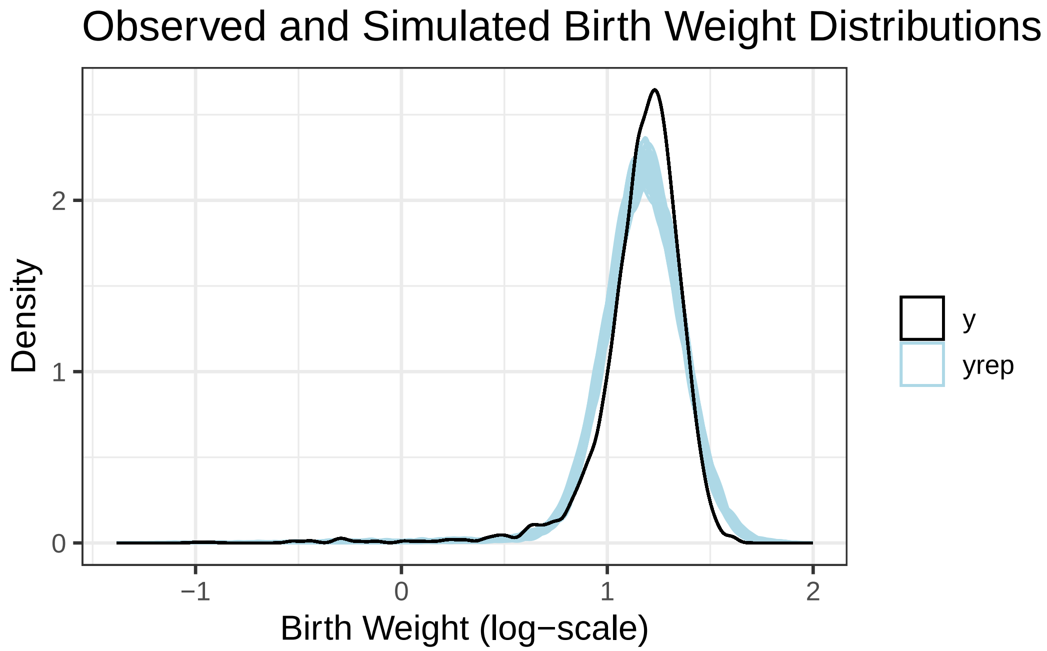
The results coincide with the summary provided.

```
set.seed(1856)
y <- ds$log_weight
yrep1 <- extract(mod1)[["log_weight_rep"]]
yrep2 <- extract(mod_2)[["log_weight_rep"]]
samp100 <- sample(nrow(yrep1), 100)
```

Question 6

The plot below shows the distribution of the data (y) and 100 datasets drawn from the posterior predictive distribution of model 2.

```
data.frame(t(yrep2[1:100, ])) |>
  bind_cols(log_weight_obs = ds$log_weight) |>
  pivot_longer(-(log_weight_obs), names_to = "Iter", values_to = "yrep") |>
  ggplot(aes(yrep, group = Iter)) +
  geom_density(alpha = 0.2, aes(color = "yrep")) +
  geom_density(aes(x = log_weight_obs, col = "y")) +
  scale_color_manual(name = "", values = c("y" = "black", "yrep" = "lightblue")) +
  labs(title = "Observed and Simulated Birth Weight Distributions",
       x = "Birth Weight (log-scale)", y = "Density") +
  theme_bw(base_size = 13)
```

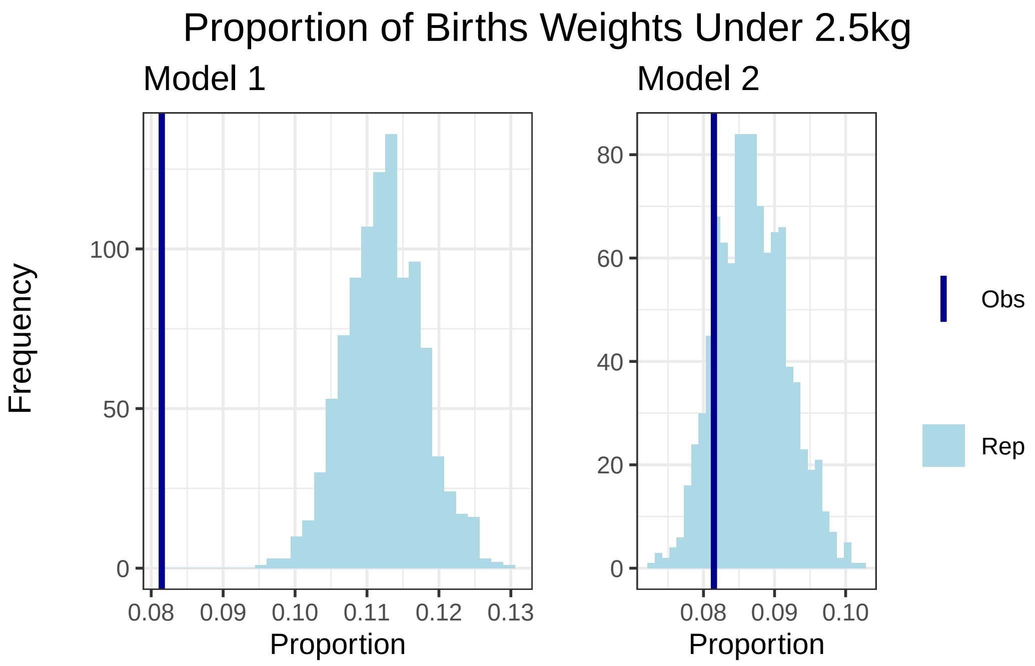


Question 7

A figure comparing the proportion of replicated births under 2.5 kg of weight for both models is shown below.

```
stat_1 <- sapply(1:nrow(yrep1), function(i) mean(yrep1[i, ] < log(2.5)))
stat_2 <- sapply(1:nrow(yrep2), function(i) mean(yrep2[i, ] < log(2.5)))

p1 <- data.frame(stat_1) |> ggplot(aes(stat_1)) +
  geom_histogram(aes(fill = "Rep")) +
  geom_vline(aes(xintercept = mean(ds$log_weight < log(2.5)), color = "Obs"), lwd = 1.1) +
  labs(title = "Model 1", x = "Proportion", y = "") +
  theme_bw(base_size = 11) +
  theme(legend.position= "none") +
  scale_color_manual(name = "", values = c("Obs" = "darkblue"))+
  scale_fill_manual(name = "", values = c("Rep" = "lightblue"))
```

```
p2 <- data.frame(stat_2) |> ggplot(aes(stat_2)) +
  geom_histogram(aes(fill = "Rep")) +
  geom_vline(aes(xintercept = mean(ds$log_weight < log(2.5)), color = "Obs"), lwd = 1.1) +
  labs(title = "Model 2", x = "Proportion", y = "") +
  theme_bw(base_size = 11) +
  scale_color_manual(name = "", values = c("Obs" = "darkblue"))+
  scale_fill_manual(name = "", values = c("Rep" = "lightblue"))

grid.arrange(p1, p2, ncol=2, nrow =1, left = "Frequency",
             top = textGrob("Proportion of Births Weights Under 2.5kg",
                            gp = gpar(fontsize = 15, font = 11)))
```

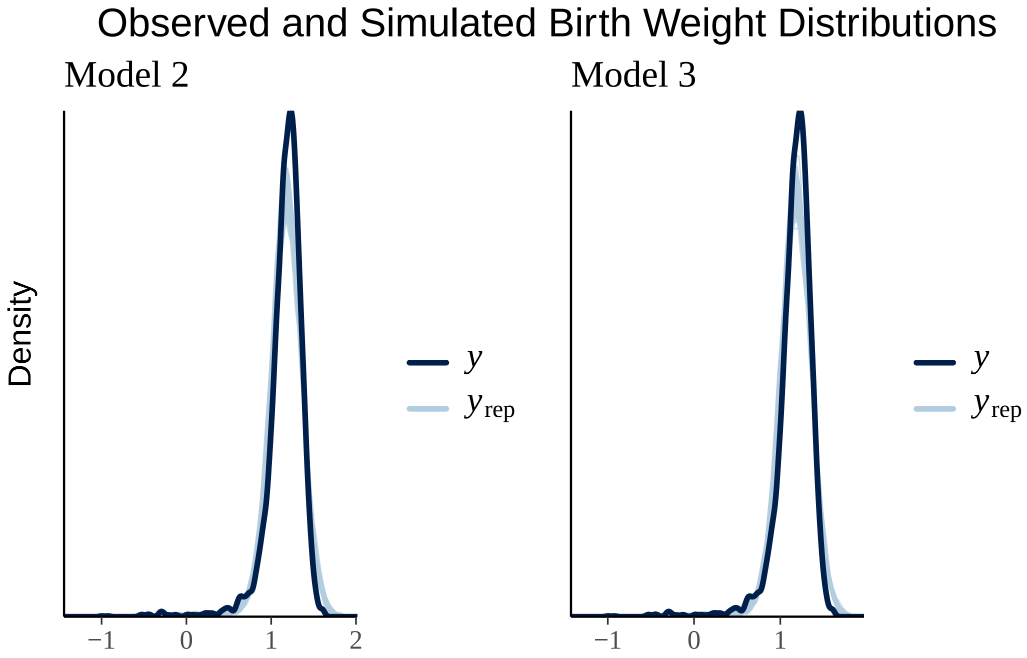


Question 8

The gender of the baby was also added as covariate to the model. This covariate was selected because the EDA suggested it can have a relationship with the weight of the baby. Also, other covariates (such as BMI) were assessed, leading to less significant changes in the results. Two posterior checks are shown below. The first plot shows the distribution of the data (y) and 100 datasets drawn from the posterior predictive distribution of models 2 and 3. Results do not change noticeably between both models.

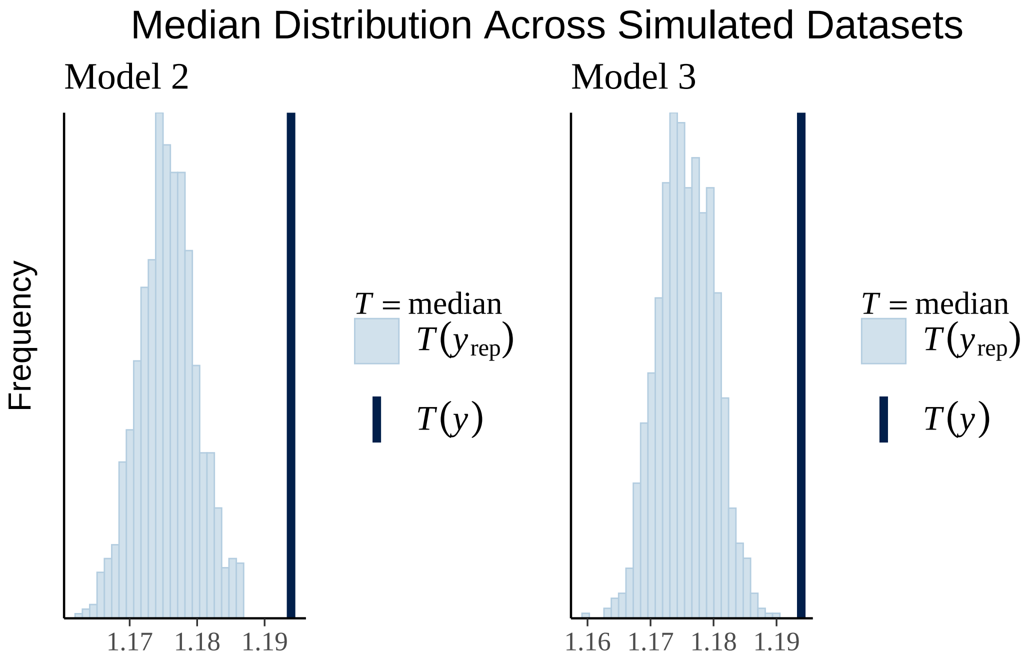
```r
ds$sex <- ifelse(ds$sex == "M", 1, 0)
stan_data <- list(N = nrow(ds),
                  log_weight = ds$log_weight,
                  log_gest = ds$log_gest_c,
                  preterm = ds$preterm,
                  sex = ds$sex)

mod3 <- stan(data = stan_data,
             file = here("simple_weight3.stan"),
             iter = 500,
             seed = 291)
```

```r
yrep3 <- extract(mod3)[["log_weight_rep"]]
samp100 <- sample(nrow(yrep3), 100)
p1 <- ppc_dens_overlay(y, yrep2[samp100, ])  + ggtitle("Model 2")
p2 <- ppc_dens_overlay(y, yrep3[samp100, ])  + ggtitle("Model 3")

grid.arrange(p1, p2, ncol=2, nrow =1, left = "Density",
             top = textGrob("Observed and Simulated Birth Weight Distributions",
                            gp = gpar(fontsize = 15, font = 11)))
```



Similarly, the plot below shows the distribution of the median across the 100 simulated datasets for both models. The distribution for model 3 is slightly closer to the value observed in the data. However, the difference does not seem to be significant.

```r
p1 <- ppc_stat(y, yrep2, stat = 'median') +
  ggtitle("Model 2")
p2 <- ppc_stat(y, yrep3, stat = 'median') +
  ggtitle("Model 3")

grid.arrange(p1, p2, ncol=2, nrow =1, left = "Frequency",
             top = textGrob("Median Distribution Across Simulated Datasets",
                            gp = gpar(fontsize = 15, font = 11)))
```


Median Distribution Across Simulated Datasets
Model 2
Model 3
Frequency
$T = \text{median}$
$T(y_{\text{rep}})$
$T(y)$
$T = \text{median}$
$T(y_{\text{rep}})$
$T(y)$