Lab 5:

Toufic Ayoub

09/02/23

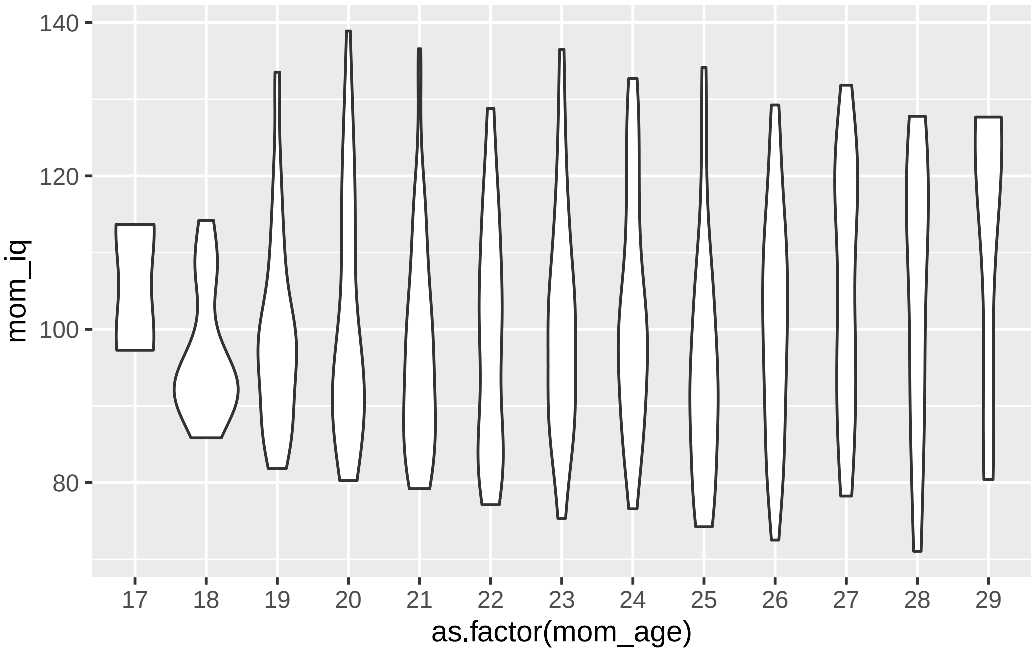
```
library(tidyverse)
library(rstan)
library(tidybayes)
library(here)
```

```
kidiq <- read_rds(here("kidiq.RDS"))
```

Question 1

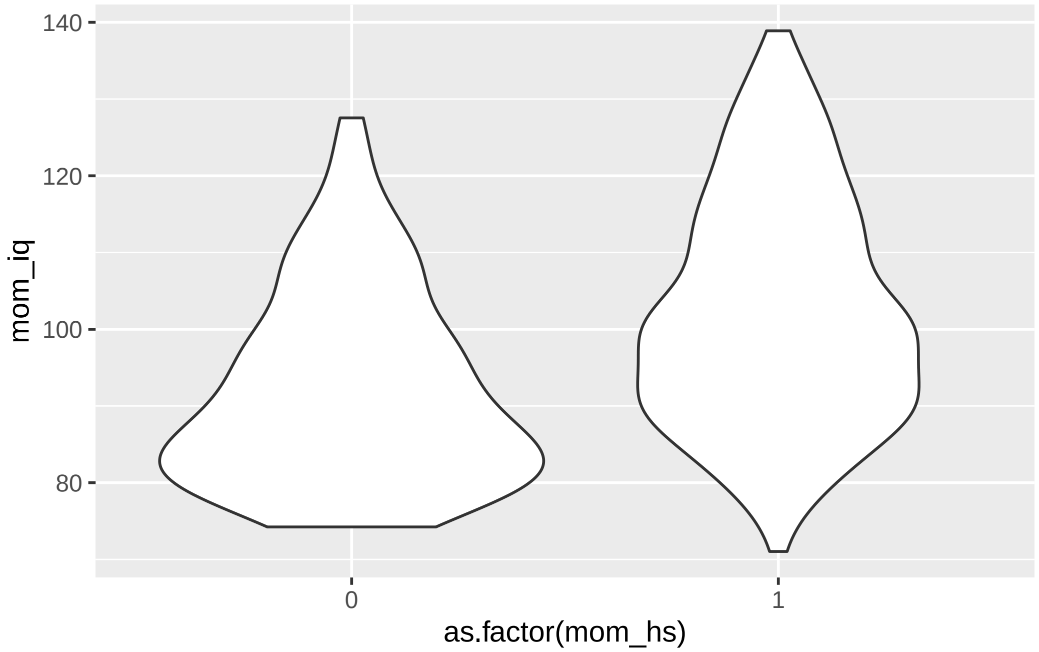
- Use plots or tables to show three interesting observations about the data. Remember:
  - Explain what your graph/ tables show
  - Choose a graph type that's appropriate to the data type

```
kidiq |> ggplot(aes(x=as.factor(mom_age), y=mom_iq)) + geom_violin()
```



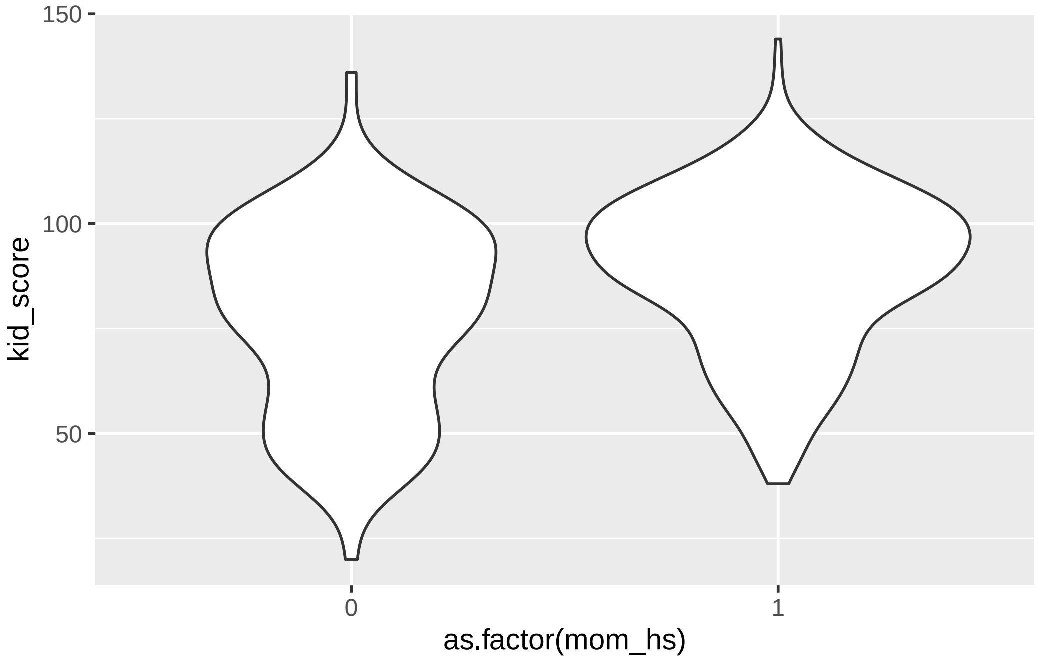
There is a relationship between the age of the mothers are their own IQ scores. We see in the first plot that as age increases, the concentration of IQ is pushed more towards the top of the violin plot. This trend is well pronounced when comparing the age 24 and age 29 plots.

```
kidiq |> ggplot(aes(x=as.factor(mom_hs), y=mom_iq)) + geom_violin()
```



There is also a relationship between a mother's IQ score and their completion of HS. The violin plot representing the completion of HS is taller and has more mass concentration further up than the concentration of the other group, indicating that completion of HS is associated with higher IQ scores.

```r
kidiq |> ggplot(aes(x=as.factor(mom_hs), y=kid_score)) + geom_violin()
```



Finally we look at the relationship between a child's IQ score and their mother's completion of HS. Here we see a similar mass distribution as with the mothers IQ scores, but now the majority of the mass between the two plots more closely overlaps. This indicates that, yes there is an association, likely extended through the IQ of the mother, there are some other factors also at play.

Question 2

- Change the prior to be much more informative (by changing the standard deviation to be 0.1). Rerun the model. Do the estimates change? Plot the prior and posterior densities.

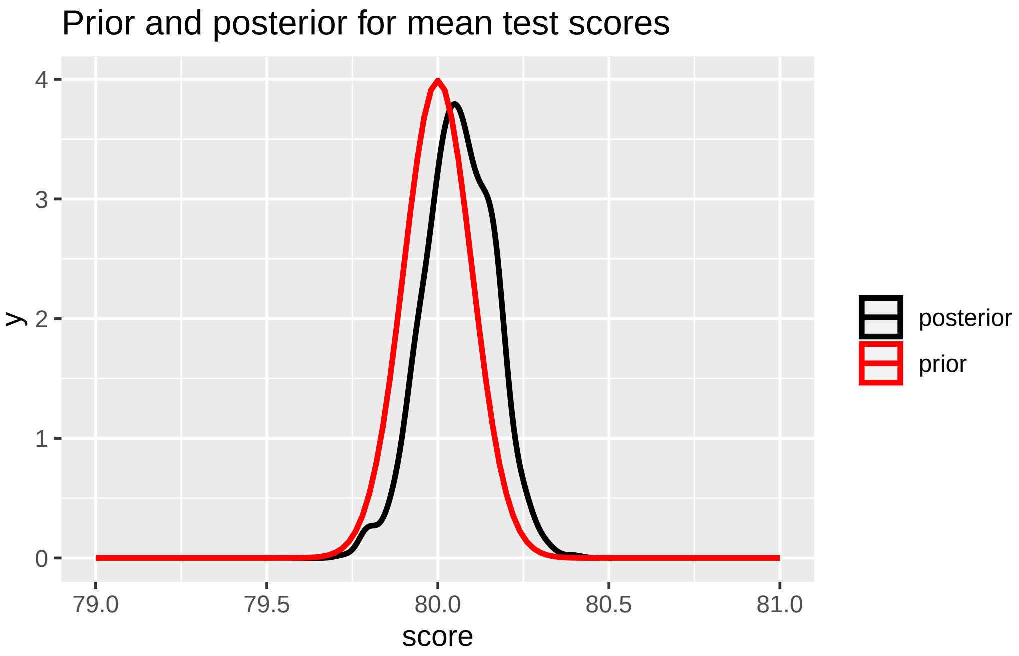
```
y <- kidiq$kid_score
mu0 <- 80
sigma0 <- 0.1
# named list to input for stan function
data <- list(y = y,
             N = length(y),
             mu0 = mu0,
             sigma0 = sigma0)

fit <- stan(file = here("kids2.stan"),
            data = data,
            chains = 3,
            iter = 500)
```

```
fit
```

```
Inference for Stan model: anon_model.
3 chains, each with iter=500; warmup=250; thin=1; 
post-warmup draws per chain=250, total post-warmup draws=750.

          mean se_mean   sd     2.5%      25%      50%      75%    97.5% n_eff
mu       80.06    0.00 0.10    79.85    80.00    80.06    80.13    80.25   619
sigma    21.41    0.04 0.77    20.04    20.85    21.39    21.90    23.00   436
lp__  -1548.45    0.06 1.08 -1551.40 -1548.82 -1548.14 -1547.70 -1547.40   322
      Rhat
mu       1
sigma    1
lp__     1

Samples were drawn using NUTS(diag_e) at Thu Feb  9 20:50:47 2023.
For each parameter, n_eff is a crude measure of effective sample size,
and Rhat is the potential scale reduction factor on split chains (at 
convergence, Rhat=1).
```

Yes, we see that the estimates do change. Now $\mu = 80.07$ and $\sigma = 21.38$, whereas we had $\mu = 86.74$ and $\sigma = 20.38$ previously. Plotting the prior and posterior densities:

```
fit  |>
  gather_draws(mu, sigma) |>
  filter(.variable == "mu") |>
  ggplot(aes(.value, color = "posterior")) + geom_density(size = 1) +
  xlim(c(79, 81)) +
  stat_function(fun = dnorm,
        args = list(mean = mu0,
                    sd = sigma0),
        aes(colour = 'prior'), size = 1) +
  scale_color_manual(name = "", values = c("prior" = "red", "posterior" = "black")) +
  ggtitle("Prior and posterior for mean test scores") +
  xlab("score")
```



Question 3

```
X <- as.matrix(kidiq$mom_hs, ncol = 1) # force this to be a matrix
K <- 1
data <- list(y = y, N = length(y),
             X =X, K = K)
fit2 <- stan(file = here("kids3.stan"),
            data = data,
            iter = 1000)
```

- a) Confirm that the estimates of the intercept and slope are comparable to results from `lm()`

```
fit2
```

```
Inference for Stan model: anon_model.
4 chains, each with iter=1000; warmup=500; thin=1;
post-warmup draws per chain=500, total post-warmup draws=2000.

            mean se_mean   sd    2.5%     25%     50%     75%   97.5%
alpha      78.02    0.07 2.08   73.89   76.67   78.07   79.40   82.05
```

```
beta[1]    11.13    0.08 2.31     6.77     9.55    11.08    12.70    15.69
sigma      19.84    0.02 0.67    18.57    19.40    19.82    20.28    21.26
lp__    -1514.43    0.05 1.31 -1517.82 -1514.94 -1514.09 -1513.47 -1512.97
        n_eff Rhat
alpha     848 1.01
beta[1]   859 1.00
sigma    1263 1.00
lp__      806 1.00

Samples were drawn using NUTS(diag_e) at Thu Feb  9 20:51:43 2023.
For each parameter, n_eff is a crude measure of effective sample size,
and Rhat is the potential scale reduction factor on split chains (at
convergence, Rhat=1).
```

From stan we find that $\alpha = 77.85$ and $\beta = 11.31$. Now lets run an linear model and comapre:

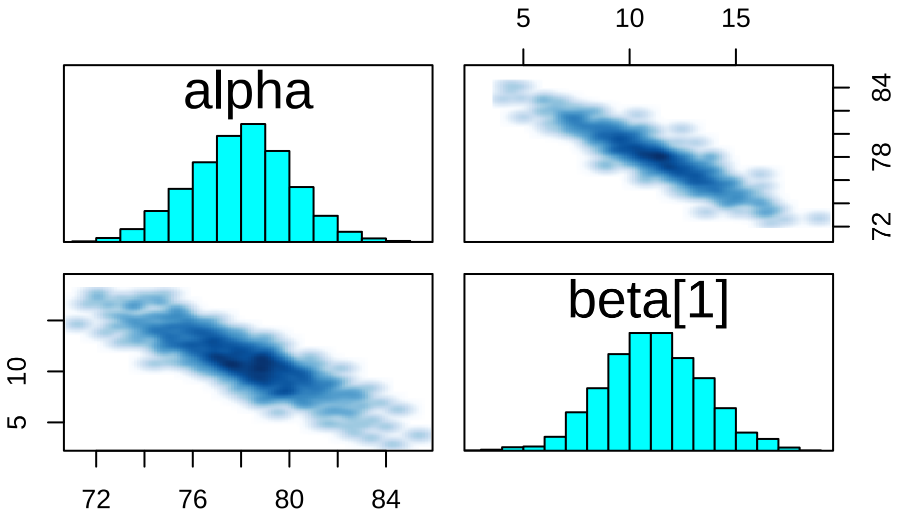
```r
lin_mod = lm(kid_score ~ mom_hs, data = kidiq)
summary(lin_mod)
```

```

Call:
lm(formula = kid_score ~ mom_hs, data = kidiq)

Residuals:
   Min     1Q Median     3Q    Max 
-57.55 -13.32   2.68  14.68  58.45 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   77.548      2.059  37.670  < 2e-16 ***
mom_hs        11.771      2.322   5.069 5.96e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 19.85 on 432 degrees of freedom
Multiple R-squared:  0.05613,	Adjusted R-squared:  0.05394 
F-statistic: 25.69 on 1 and 432 DF,  p-value: 5.957e-07
```

From the linear model we have $\alpha = 77.548$ and $\beta = 11.771$, which are near identical estimates.

- b) Do a `pairs` plot to investigate the joint sample distributions of the slope and intercept. Comment briefly on what you see. Is this potentially a problem?

```
pairs(fit2,  pars = c("alpha", "beta"))
```



We see a negative correlation between the distributions of the slope an intercept. As talked about in class this isn't so much an issue with our model but with the nature of linear regression and how the slope and intercept are intrinsically linked to one other. It is potentially a problem, due to the restrictive nature of the link. Centering the data is a good idea.

Question 4

- Add in mother's IQ as a covariate and rerun the model. Please mean center the covariate before putting it into the model. Interpret the coefficient on the (centered) mum's IQ.

```
X <- as.matrix(data.frame(kidiq$mom_hs, kidiq$mom_iq - mean(kidiq$mom_iq)), ncol = 2)
K <- 2
data <- list(y = y, N = length(y),
             X = X, K = K)
fit3 <- stan(file = here("kids3.stan"),
            data = data,
            iter = 1000)
```

```
fit3
```

```
Inference for Stan model: anon_model.
4 chains, each with iter=1000; warmup=500; thin=1;
post-warmup draws per chain=500, total post-warmup draws=2000.

            mean se_mean   sd     2.5%      25%      50%      75%    97.5%
alpha      82.25    0.06 1.85    78.67    81.04    82.22    83.49    85.91
beta[1]     5.79    0.07 2.11     1.73     4.38     5.71     7.19     9.98
beta[2]     0.57    0.00 0.06     0.44     0.52     0.57     0.61     0.69
sigma      18.13    0.02 0.62    16.93    17.70    18.12    18.54    19.41
lp__    -1474.45    0.05 1.46 -1478.15 -1475.11 -1474.13 -1473.35 -1472.68
        n_eff Rhat
alpha     976    1
beta[1]  1006    1
beta[2]  1206    1
sigma    1459    1
lp__      850    1

Samples were drawn using NUTS(diag_e) at Thu Feb  9 20:51:45 2023.
For each parameter, n_eff is a crude measure of effective sample size,
and Rhat is the potential scale reduction factor on split chains (at
convergence, Rhat=1).
```

We see that the estimate for $\beta_2 = 0.57$. The fact that we have centered the data should not affect how we interpret the coefficient. What this means is that for every unit increase in the mother's IQ, we expect a 0.57 increase in the child's IQ.

Question 5

- Confirm the results from Stan agree with `lm()`

```r
kidiq_centered = data.frame(kid_score = kidiq$kid_score,
                            mom_hs = kidiq$mom_hs,
                            mom_iq_c = kidiq$mom_iq - mean(kidiq$mom_iq))

lin_mod2 = lm(kid_score ~ mom_hs + mom_iq_c, data = kidiq_centered)
summary(lin_mod2)
```

```
Call:
lm(formula = kid_score ~ mom_hs + mom_iq_c, data = kidiq_centered)
```

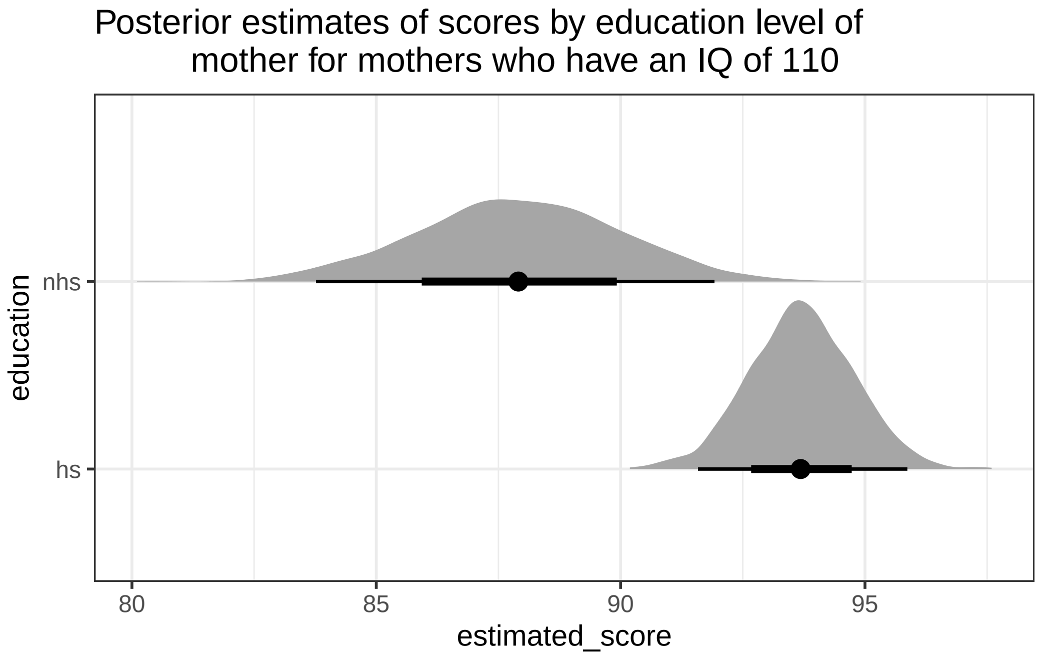
```
Residuals:
    Min      1Q  Median      3Q     Max 
-52.873 -12.663   2.404  11.356  49.545 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 82.12214    1.94370  42.250  < 2e-16 ***
mom_hs       5.95012    2.21181   2.690  0.00742 ** 
mom_iq_c     0.56391    0.06057   9.309  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 18.14 on 431 degrees of freedom
Multiple R-squared:  0.2141,    Adjusted R-squared:  0.2105 
F-statistic: 58.72 on 2 and 431 DF,  p-value: < 2.2e-16
```

Running the linear model above, we see that the coefficinet for $\beta_2 = 0.56391$, which is very close to the to the Stan output.

Question 6

- Plot the posterior estimates of scores by education of mother for mothers who have an IQ of 110.

```
# Filtering the output from spread_draws to extract the two beta coeff
beta_1s = fit3 |>
            spread_draws(alpha, beta[k], sigma) |> filter(k == 1)

beta_2s = fit3 |>
            spread_draws(alpha, beta[k], sigma) |> filter(k == 2)

# Joining the data so that we have two columsn representing beta1 and
# beta2 for a single draw
merged = left_join(beta_1s, beta_2s, by = c(".draw"))

# Now we can continue as shown above. Note that beta.x = beta1 and beta.y = beta2.
# Also note that since beta.y was fit to centered data, we need to center the 110
# IQ score
merged |>
     mutate(nhs = alpha.x + 0*beta.x + (110-mean(kidiq$mom_iq))*beta.y,
         hs = alpha.x + 1*beta.x + (110-mean(kidiq$mom_iq))*beta.y) |>
  select(nhs, hs) |>
  pivot_longer(nhs:hs, names_to = "education", values_to = "estimated_score") |>
  ggplot(aes(y = education, x = estimated_score)) +
  stat_halfeye() +
  theme_bw() +
  ggtitle("Posterior estimates of scores by education level of
         mother for mothers who have an IQ of 110")
```



We that for mothers with an IQ of 110, the completion of high school positively influences the scores of their children. That is, if the mothers finished HS, their children's scores are generally larger than those children whose mothers did not complete HS.

Question 7

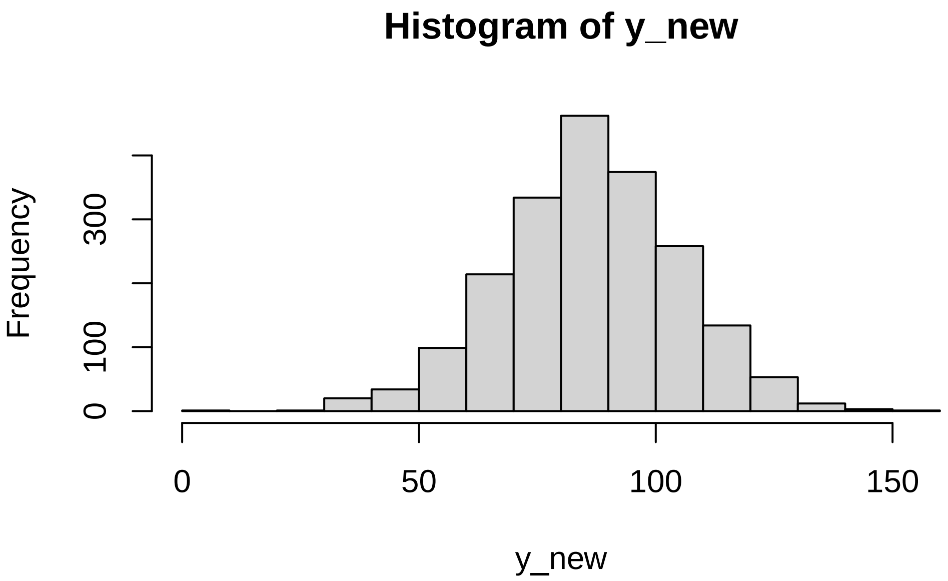
- Generate and plot (as a histogram) samples from the posterior predictive distribution for a new kid with a mother who graduated high school and has an IQ of 95.

```
post_samples <- extract(fit3)
alpha = post_samples[["alpha"]]
beta1 = post_samples[["beta"]][,1]
```

```
beta2 = post_samples[["beta"]][,2]
sigma = post_samples[["sigma"]]
```

If a mother has graduated highschool, we denote that by 1. And again we need to center the IQ of 95. We first compute the linear predictor, then can generate the posterior predictive distribution.

```
linear_pred = alpha + (1 * beta1) + ((95 - mean(kidiq$mom_iq)) * beta2)

y_new = rnorm(n = length(sigma), mean = linear_pred, sd = sigma)
hist(y_new)
```


Histogram of y_new
Frequency
y_new