Lab Week 5

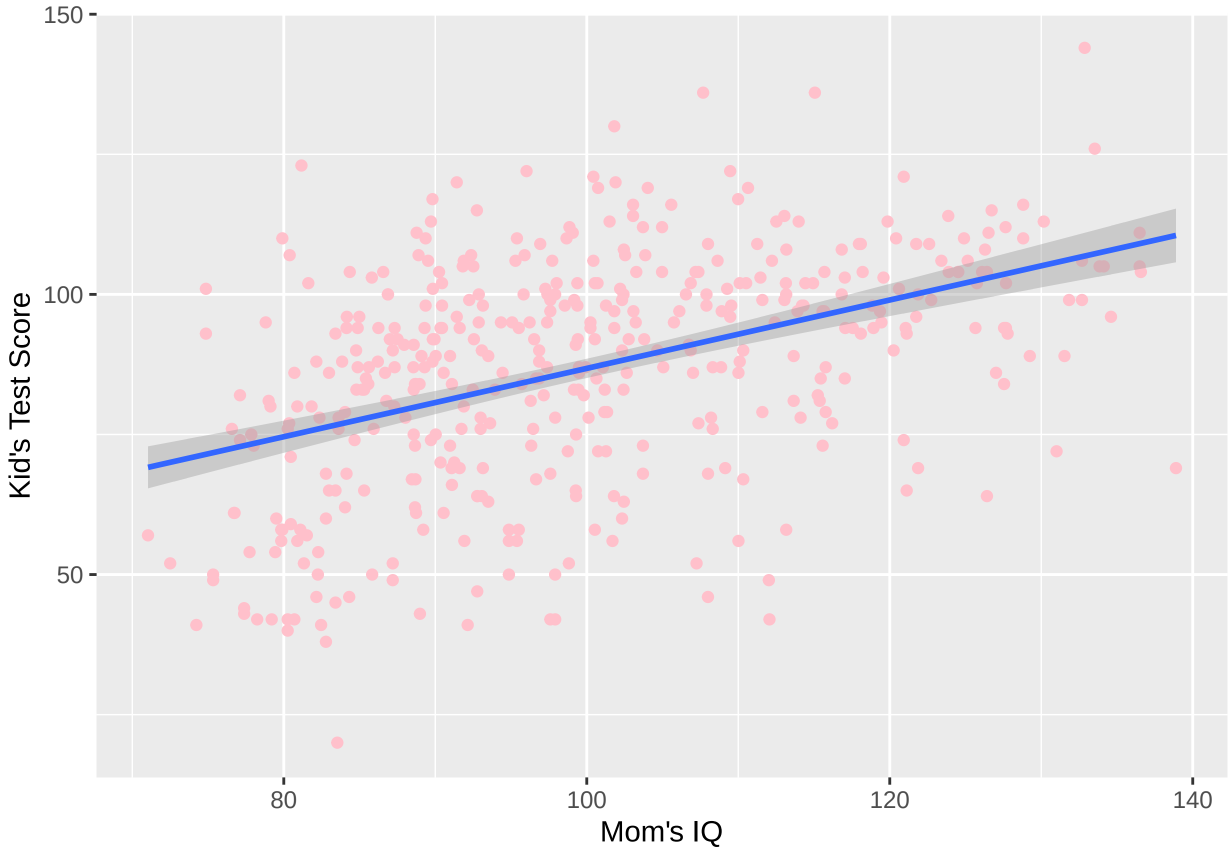
Rosa Fallahpour

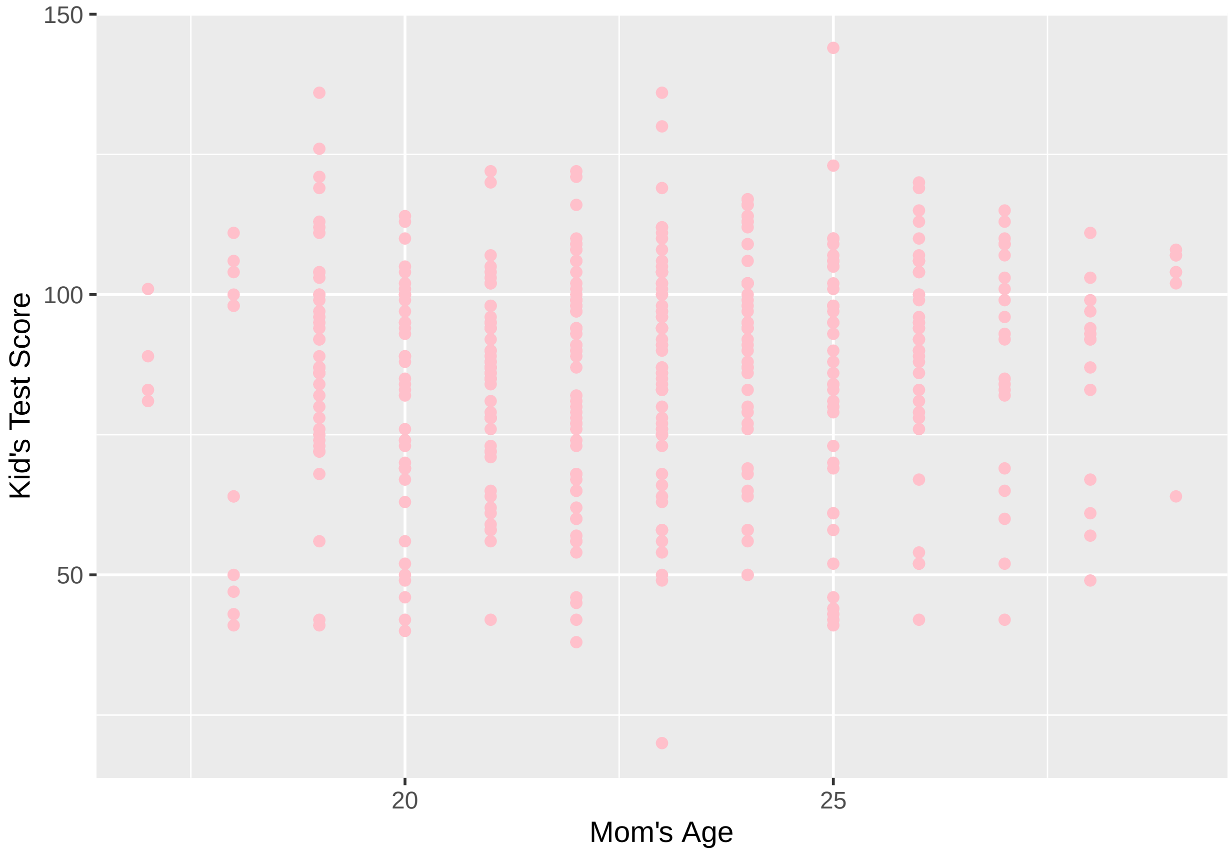
Question 1

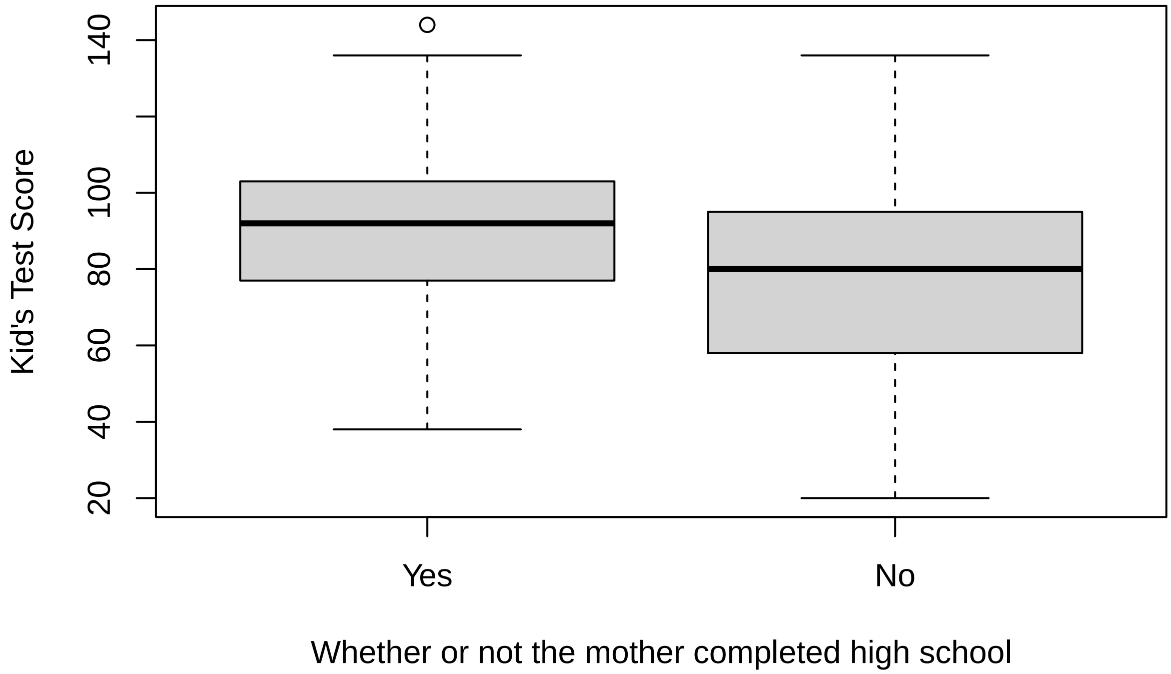
Three following graphs represent the kid's test score with respect to mom's IQ, mom's age and whether or not the mother completed high school, respectively. The first plot shows that the kid's score has the positive relationship with mom's IQ, as we can see that the increase in mom's IQ results in increasing kid's score. The second graph displays the relationship between kid's score and mom's age. It is interesting to see that not a significant change is observed in kid's score in different ages of mothers. Kid's scores are almost in the same range for mothers with different ages. The last plot shows the relationship between kid's score and mom's high school completion. It suggests that the average kid's score for the group whose moms completed their high school is higher than those kid's whose mothers did not complete the high school.

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

```
## # A tibble: 434 x 4
##    kid_score mom_hs mom_iq mom_age
##        <int>  <dbl>  <dbl>   <int>
##  1        65      1  121.       27
##  2        98      1   89.4      25
##  3        85      1  115.       27
##  4        83      1   99.4      25
##  5       115      1   92.7      27
##  6        98      0  108.       18
##  7        69      1  139.       20
##  8       106      1  125.       23
##  9       102      1   81.6      24
## 10        95      1   95.1      19
## # ... with 424 more rows
```


150
100
50
80
100
120
140
Kid's Test Score
Mom's IQ


150
100
50
Kid's Test Score
20
25
Mom's Age



Estimating mean, no covariates

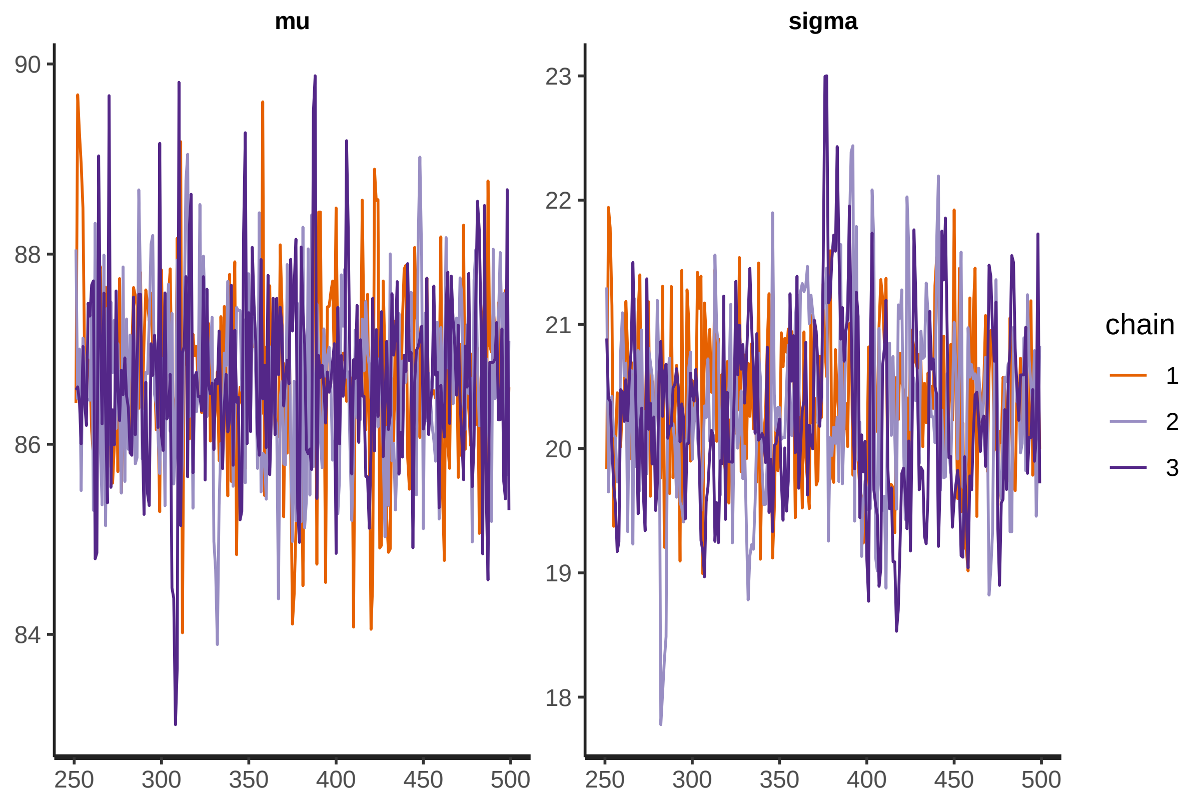
here is the summary:

```
fit
```

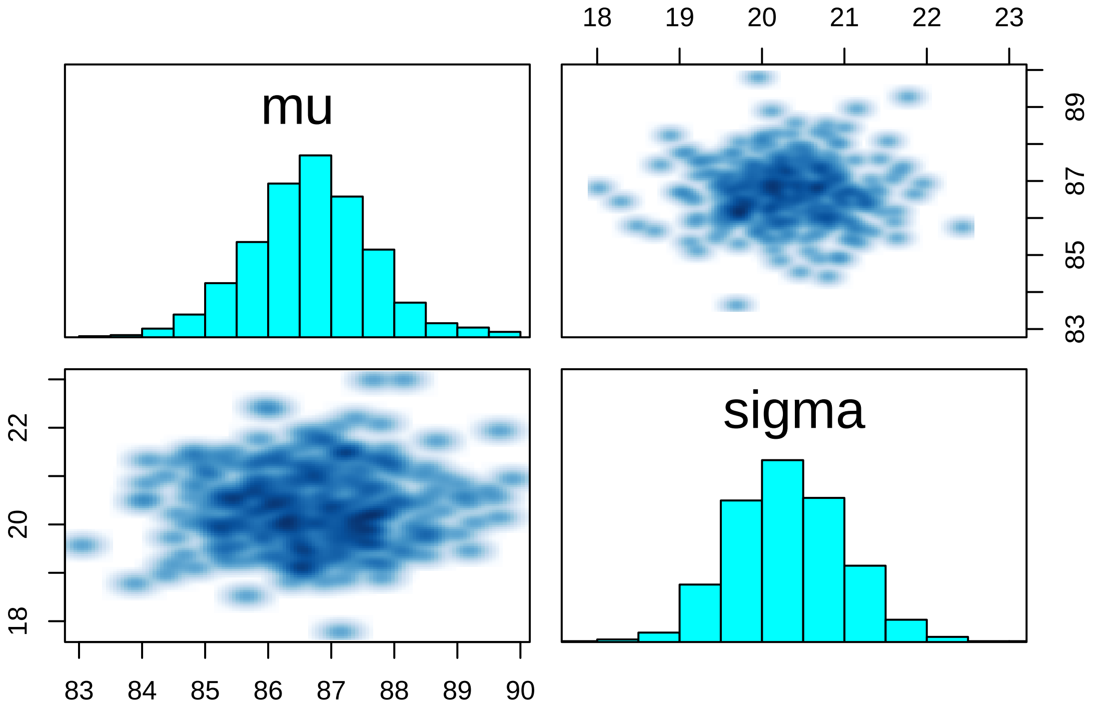
```
## Inference for Stan model: anon_model.
## 3 chains, each with iter=500; warmup=250; thin=1;
## post-warmup draws per chain=250, total post-warmup draws=750.
##
##           mean se_mean   sd     2.5%      25%      50%      75%    97.5% n_eff
## mu       86.69    0.04 0.97    84.83    86.07    86.69    87.31    88.68   721
## sigma    20.32    0.05 0.71    19.04    19.84    20.28    20.79    21.73   218
## lp__  -1525.79    0.07 1.17 -1529.09 -1526.18 -1525.40 -1525.05 -1524.78   286
##       Rhat
## mu    1.00
## sigma 1.00
## lp__  1.01
##
## Samples were drawn using NUTS(diag_e) at Sun Feb 12 19:26:50 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).
```

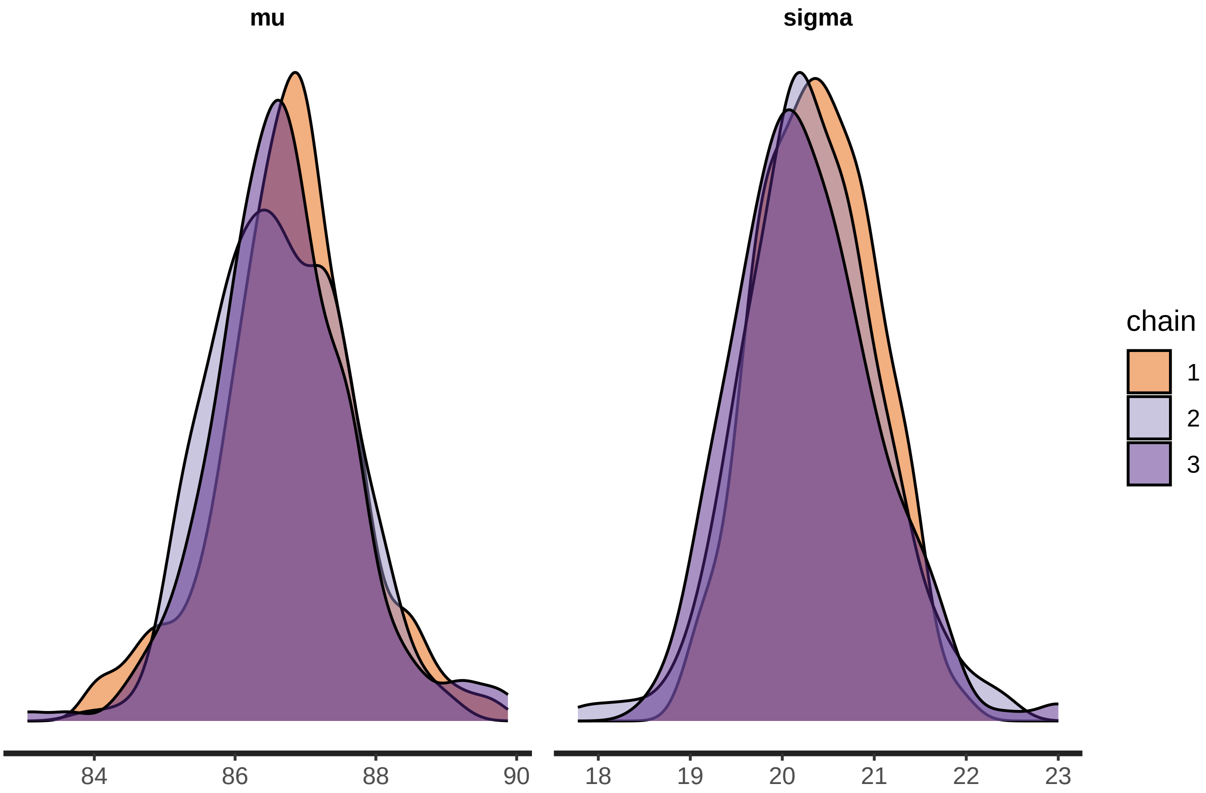
Traceplot:

```
traceplot(fit)
```



```
pairs(fit, pars = c("mu", "sigma"))
```


mu
sigma
18
19
20
21
22
23
83
85
87
89
18
20
22
83
84
85
86
87
88
89
90



Understanding output

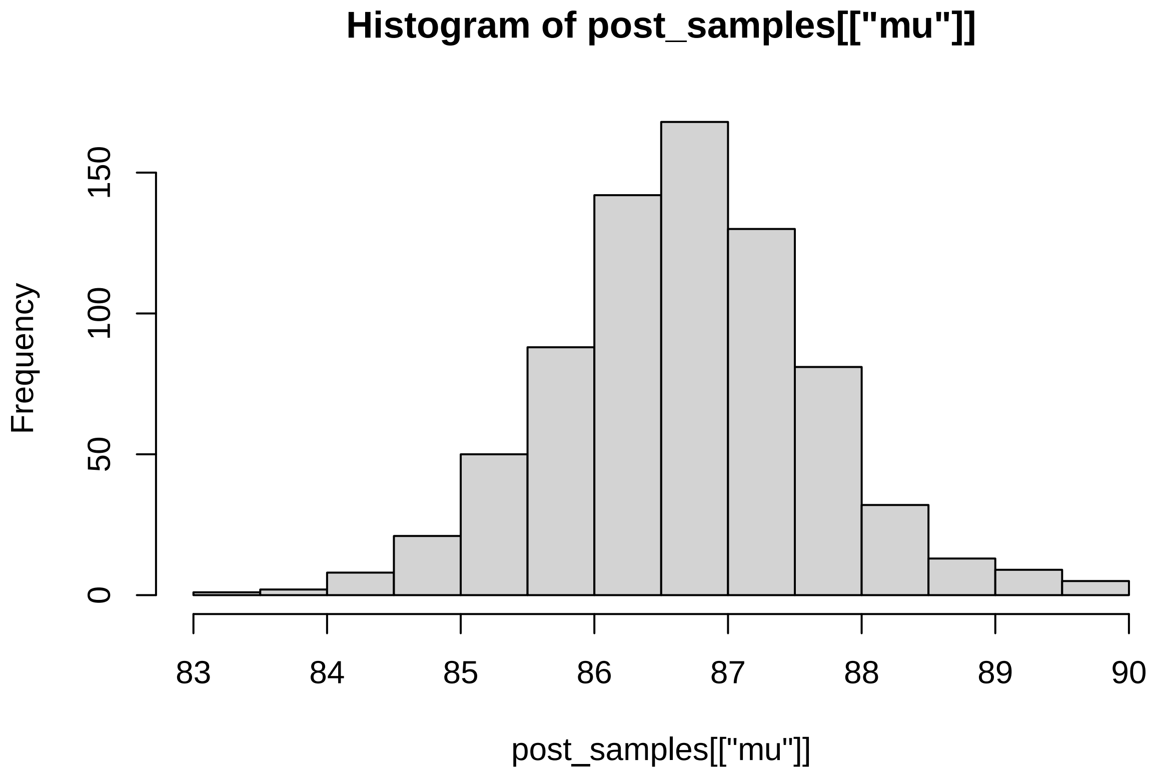
Samples from the posteriors:

```
post_samples <- rstan::extract(fit)
head(post_samples[["mu"]])
```

```
## [1] 86.34436 89.27382 86.55427 85.81895 87.78656 86.76803
```

Histogram of mu:

```
hist(post_samples[["mu"]])
```



```
median(post_samples[["mu"]])
```

```
## [1] 86.6928
```

```
quantile(post_samples[["mu"]], 0.025)
```

```
##     2.5%
## 84.82652
```

```
quantile(post_samples[["mu"]], 0.975)
```

```
##    97.5%
## 88.67626
```

Plot estimates

Get the posterior samples for mu and sigma in long format:

```
library(tidybayes)
dsamples <- fit  |>
  gather_draws(mu, sigma) # gather = long format
dsamples
```

```
## # A tibble: 1,500 x 5
## # Groups:   .variable [2]
##    .chain .iteration .draw .variable .value
##     <int>      <int> <int> <chr>      <dbl>
##  1      1          1     1 mu          86.4
##  2      1          2     2 mu          89.7
##  3      1          3     3 mu          89.3
##  4      1          4     4 mu          89.0
##  5      1          5     5 mu          88.5
##  6      1          6     6 mu          86.8
##  7      1          7     7 mu          86.9
##  8      1          8     8 mu          86.9
##  9      1          9     9 mu          86.7
## 10      1         10    10 mu          86.1
## # ... with 1,490 more rows
```

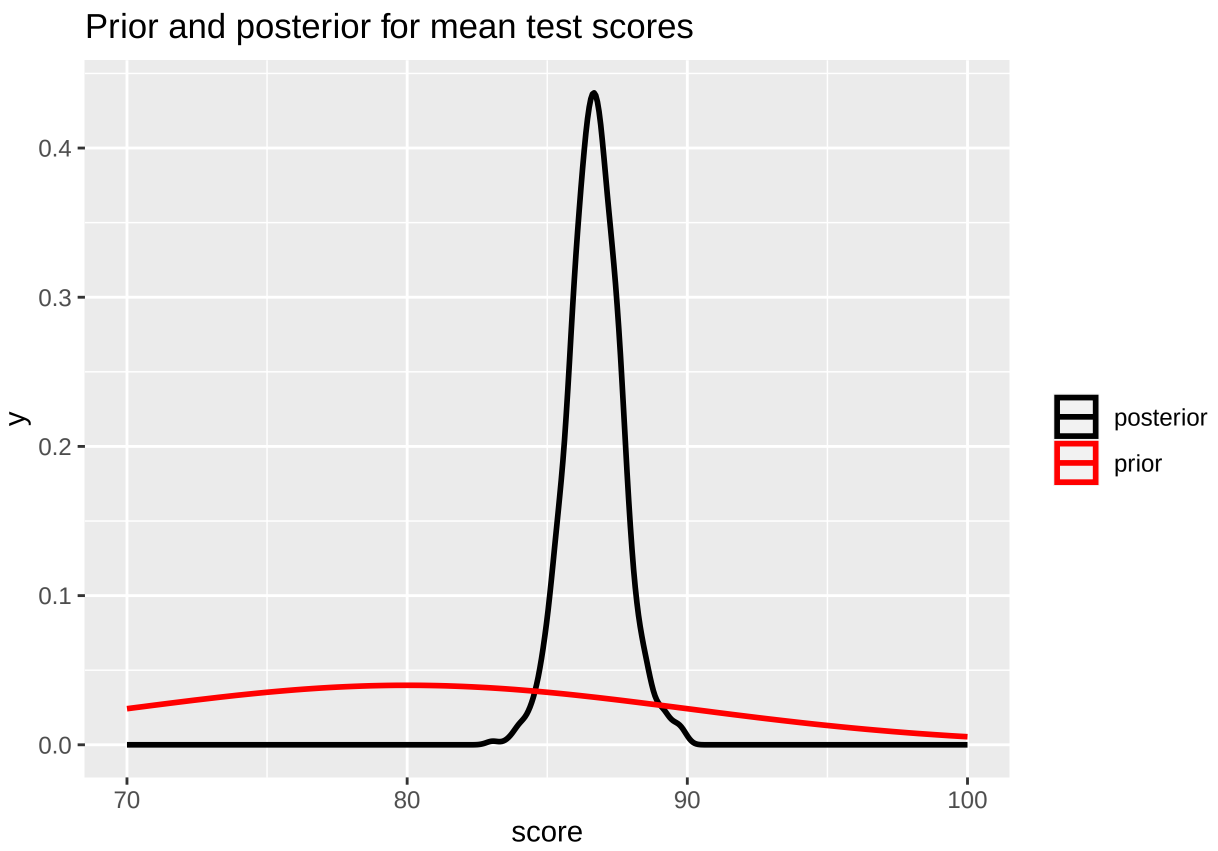
```
# wide format
fit  |>  spread_draws(mu, sigma)
```

```
## # A tibble: 750 x 5
##    .chain .iteration .draw    mu sigma
##     <int>      <int> <int> <dbl> <dbl>
##  1      1          1     1  86.4  19.8
##  2      1          2     2  89.7  21.9
##  3      1          3     3  89.3  21.8
##  4      1          4     4  89.0  21.2
##  5      1          5     5  88.5  19.4
##  6      1          6     6  86.8  20.0
##  7      1          7     7  86.9  20.4
##  8      1          8     8  86.9  20.4
##  9      1          9     9  86.7  20.0
## 10      1         10    10  86.1  20.4
## # ... with 740 more rows
```

```
# quickly calculate the quantiles using
dsamples |>
  median_qi(.width = 0.8)
```

```
## # A tibble: 2 x 7
##   .variable .value .lower .upper .width .point .interval
##   <chr>      <dbl>  <dbl>  <dbl>  <dbl> <chr>  <chr>
## 1 mu          86.7   85.5   87.8    0.8 median qi
## 2 sigma       20.3   19.5   21.2    0.8 median qi
```

Let's plot the density of the posterior samples for mu and add in the prior distribution:



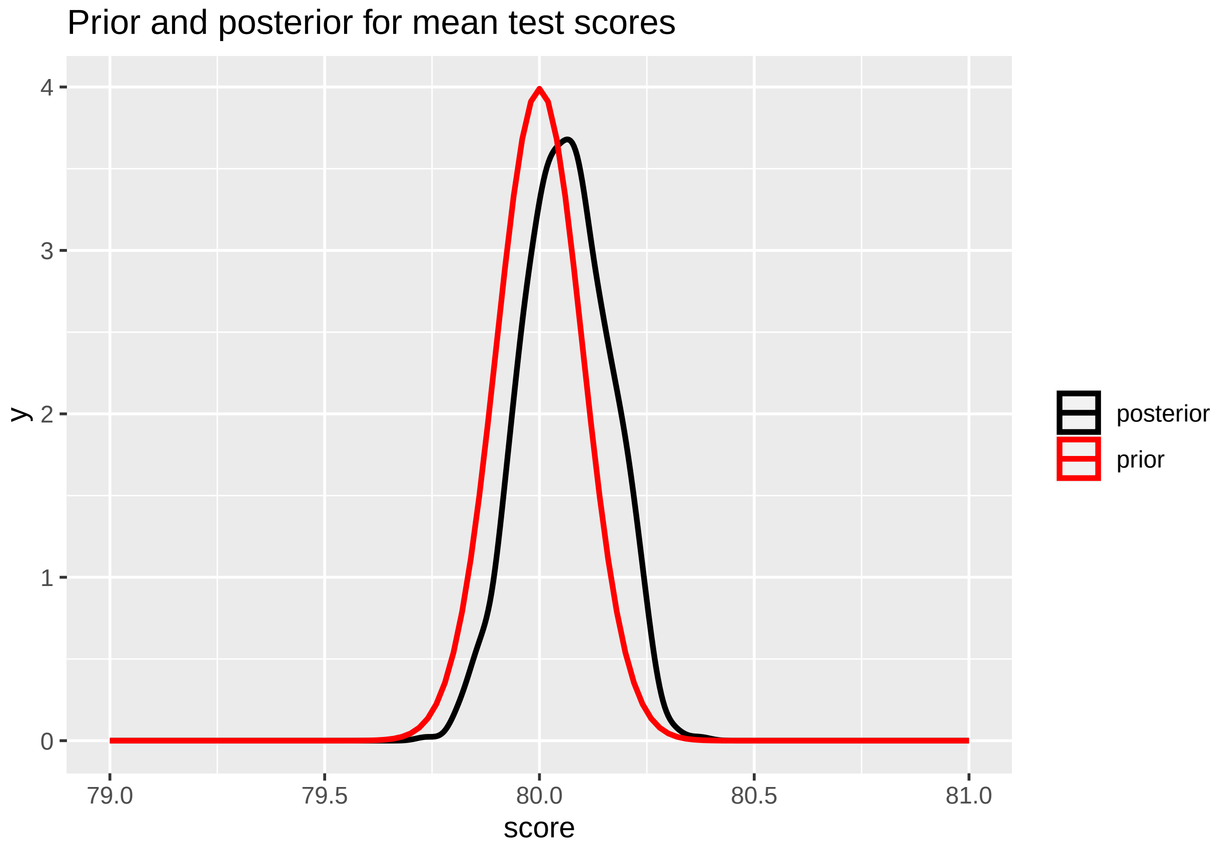
Question 2

In the model with more informative prior, mu estimate and it's standard error has decreased. However, it shows a slight increase in the sigma value in the more informative model.

```
fit_inform
```

```
## 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    79.86    79.99    80.06    80.13    80.24   648
## sigma     21.37    0.03 0.69    20.18    20.88    21.34    21.80    22.83   698
## lp__   -1548.34    0.05 0.90 -1550.59 -1548.74 -1548.08 -1547.66 -1547.40   362
##        Rhat
## mu        1
## sigma     1
## lp__      1
##
## Samples were drawn using NUTS(diag_e) at Sun Feb 12 19:26:52 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).
```

Plotting the prior and posterior densities:



Adding Covariates

```
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      77.94    0.08 1.96    74.31    76.59    77.92    79.21    81.86
## beta[1]    11.24    0.09 2.19     6.90     9.77    11.28    12.78    15.43
## sigma      19.82    0.02 0.65    18.56    19.39    19.83    20.26    21.07
## lp__    -1514.31    0.04 1.15 -1517.25 -1514.81 -1514.04 -1513.46 -1512.97
##         n_eff Rhat
## alpha     639 1.00
## beta[1]   631 1.00
## sigma    1081 1.00
## lp__      723 1.01
##
## Samples were drawn using NUTS(diag_e) at Sun Feb 12 19:27:30 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).
```

Question 3

- **(a)** As we can see in the following summaries which are related to lm model and model fit2, the estimates are very close to each other.

```
model_lm <- lm(kid_score~mom_hs, data=kidiq)
summary(model_lm)$`coefficient`
```

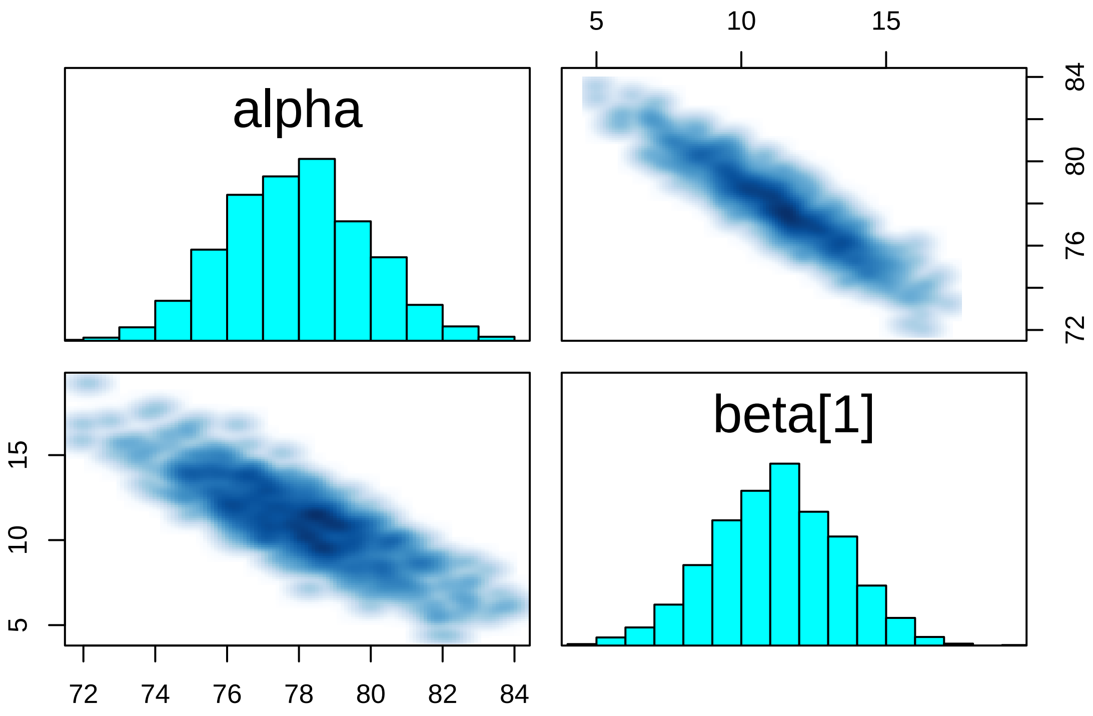
```
##             Estimate Std. Error   t value      Pr(>|t|)
## (Intercept) 77.54839   2.058612 37.670231 1.392224e-138
## mom_hs      11.77126   2.322427  5.068516  5.956524e-07
```

```
summary(fit2)$summary[c("alpha", "beta[1]"), ]
```

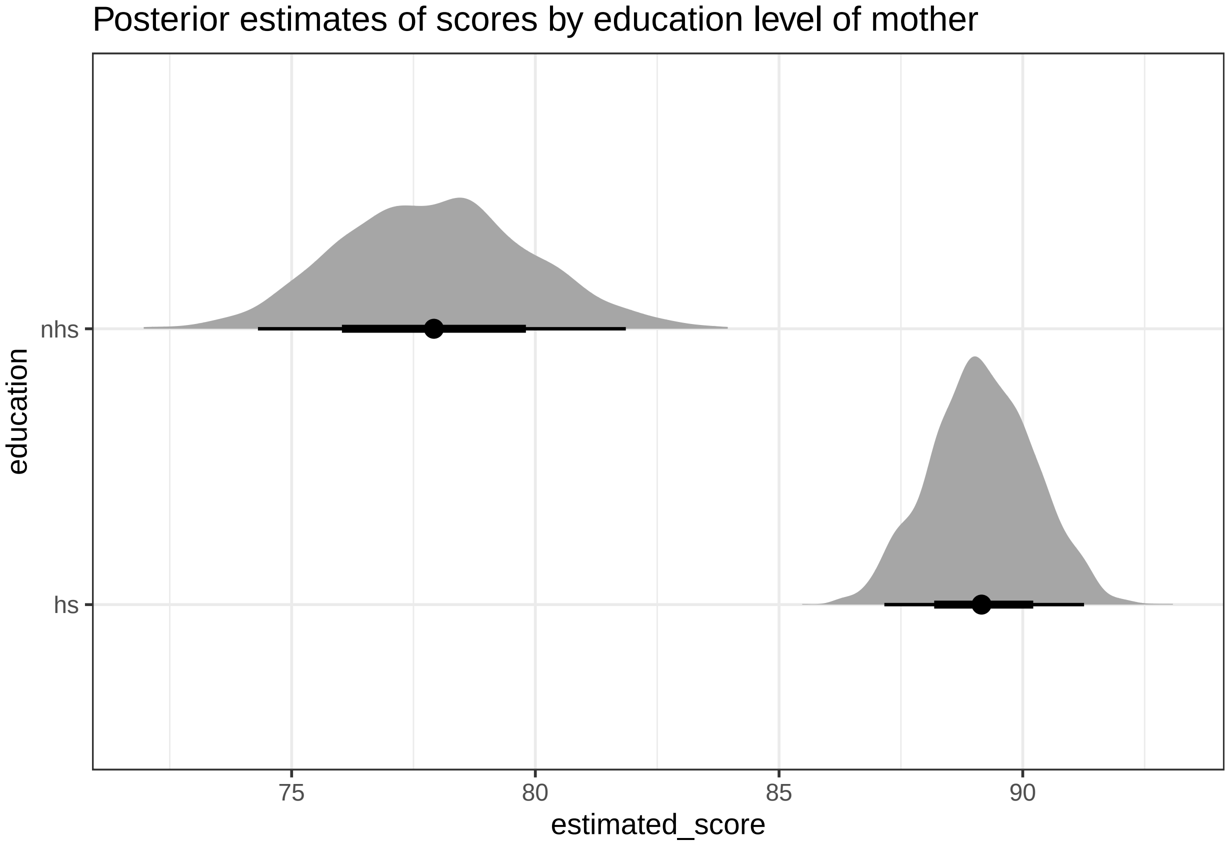
```
##             mean    se_mean       sd      2.5%       25%      50%      75%
## alpha   77.93599 0.07741812 1.956677 74.309334 76.589552 77.91817 79.20661
## beta[1] 11.23684 0.08712886 2.188108  6.895808  9.765203 11.28365 12.78110
##            97.5%    n_eff     Rhat
## alpha   81.85649 638.7826 1.004810
## beta[1] 15.43086 630.6862 1.004007
```

- **(b)** As the figure shows, the slope variation includes the opposite variation of the intercept. Thus, the intercept interpretation and sampling would be harder.

```
pairs(fit2, pars = c("alpha", "beta[1]"))
```



Plotting the results



Question 4

```
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      82.26    0.07 1.95    78.53    80.91    82.26    83.57    86.04
## beta[1]     5.74    0.07 2.20     1.49     4.27     5.67     7.23    10.00
## beta[2]     0.56    0.00 0.06     0.45     0.52     0.56     0.60     0.68
## sigma      18.13    0.02 0.63    16.97    17.69    18.10    18.54    19.41
## lp__    -1474.45    0.05 1.42 -1478.01 -1475.16 -1474.12 -1473.40 -1472.68
##         n_eff Rhat
## alpha     801    1
## beta[1]   857    1
## beta[2]  1316    1
## sigma    1537    1
## lp__      705    1
##
```

```
## Samples were drawn using NUTS(diag_e) at Sun Feb 12 19:27:31 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).
```

For a given mother's high school completion, one unit increase of mom's IQ, results in the posterior mean of the kid's score to increase by 0.57.

Question 5

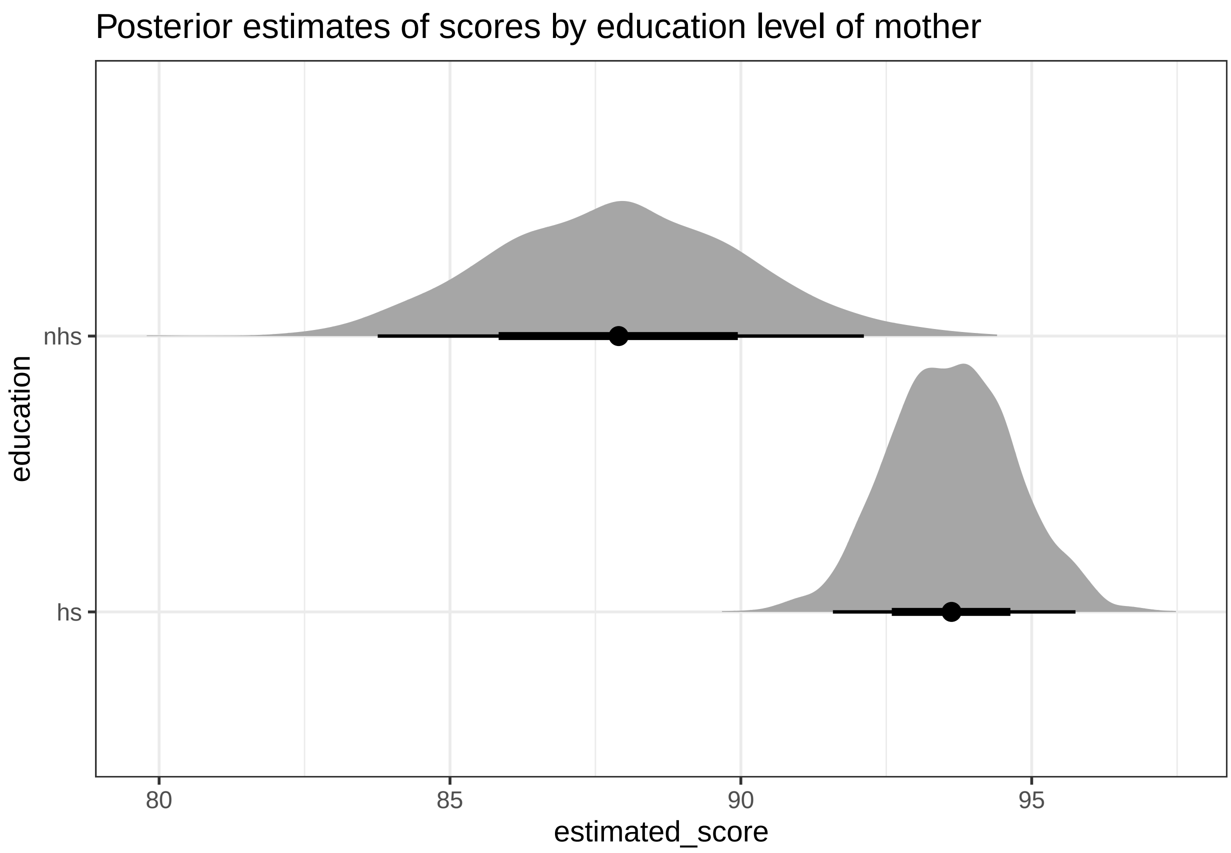
As we can see the estimates of two models are comparable.

```
momiq_2 <- kidiq$mom_iq-mean(kidiq$mom_iq)
model2_lm <- lm(kid_score~mom_hs+momiq_2, data=kidiq)
summary(model2_lm)$`coefficient`
```

```
##              Estimate Std. Error   t value      Pr(>|t|)
## (Intercept) 82.122143 1.94370047 42.250411 2.435765e-155
## mom_hs       5.950117 2.21181218  2.690155  7.419327e-03
## momiq_2      0.563906 0.06057408  9.309362  6.609618e-19
```

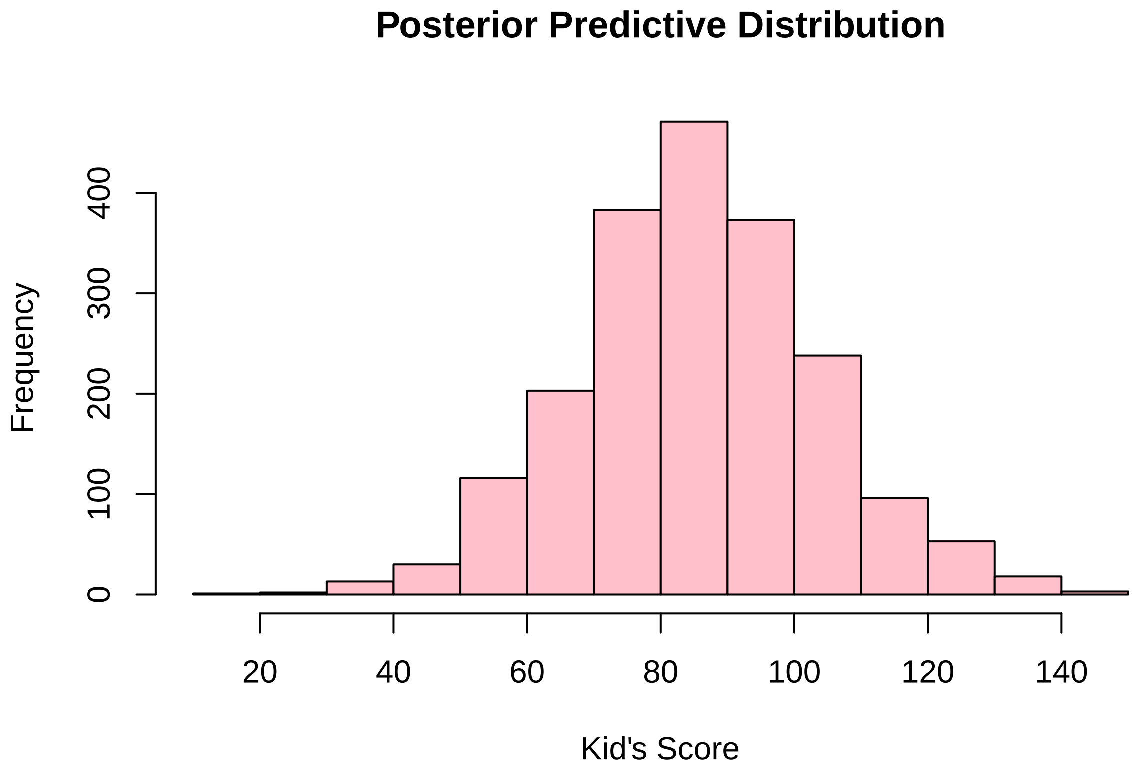
Question 6

```
data <- as.data.frame(fit3 %>% spread_draws(alpha, beta[condition], sigma))
data %>%
  reshape(
    idvar = c(".iteration", ".draw", ".chain"),
    timevar = "condition", v.names = "beta", direction = "wide"
  ) %>% mutate(nhs = alpha + beta.2 * 10, hs = alpha + beta.1 + beta.2 * 10) %>%
  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")
```



Question 7

```
postsample <- rstan ::extract(fit3)
alpha <- postsample[["alpha"]]
beta1 <-postsample[["beta"]][,1]
beta2 <- postsample[["beta"]][,2]
x_new_2 <- 95-mean(kidiq$mom_iq)
lin_pred <- alpha + beta1*1+beta2*-5
sigma <- postsample[["sigma"]]
y_new <- rnorm(n= length(sigma),mean = lin_pred, sd=sigma)
hist(y_new, xlab = "Kid's Score", main="Posterior Predictive Distribution", col="pink")
```


Posterior Predictive Distribution
Frequency
0
100
200
300
400
20
40
60
80
100
120
140
Kid's Score