Week 5: Bayesian linear regression and introduction to Stan

12/02/23

```r
library(tidyverse)
library(rstan)
library(tidybayes)
library(here)
library(broom)
theme_set(theme_bw())
```

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

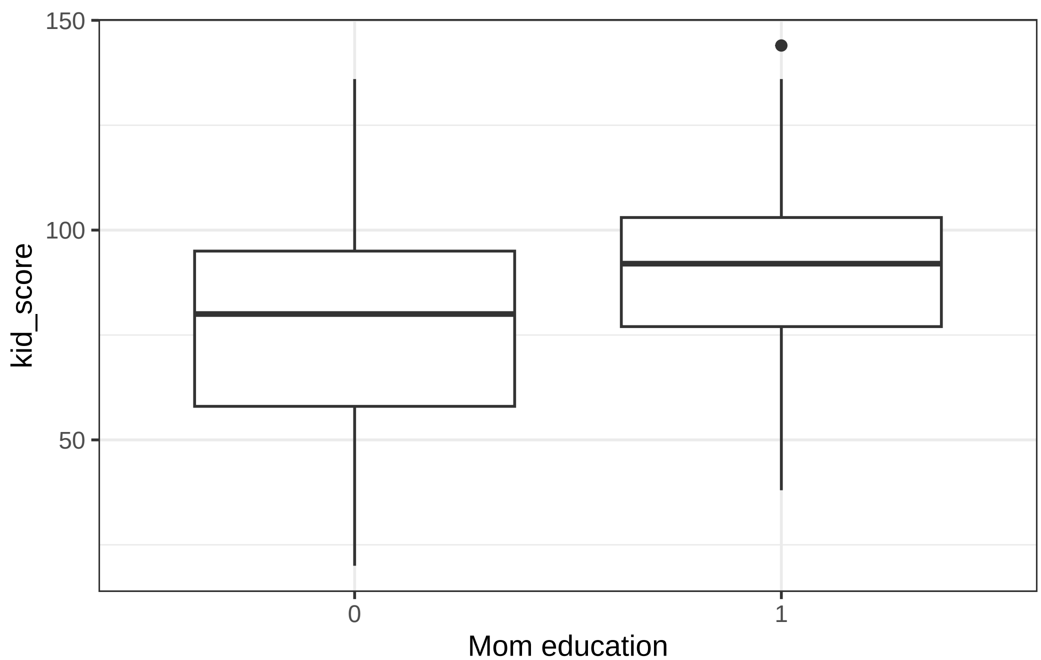
```
# A tibble: 434 x 4
   kid_score mom_hs mom_iq mom_age
       <int>  <dbl>  <dbl>   <int>
 1        65      1  121.       27
 2        98      1   89.4      25
 3        85      1  115.       27
 4        83      1   99.4      25
 5       115      1   92.7      27
 6        98      0  108.       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
```

Question 1

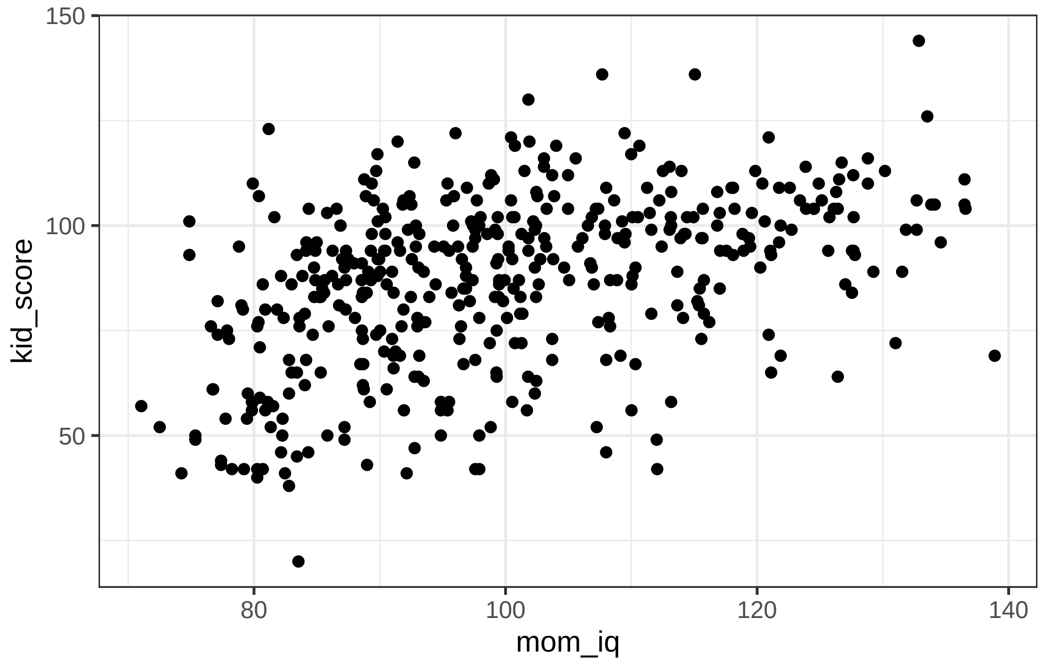
Use plots or tables to show three interesting observations about the data. Remember:

- Explain what your graph/ tables show
- Choose a graph type that's appropriate to the data type

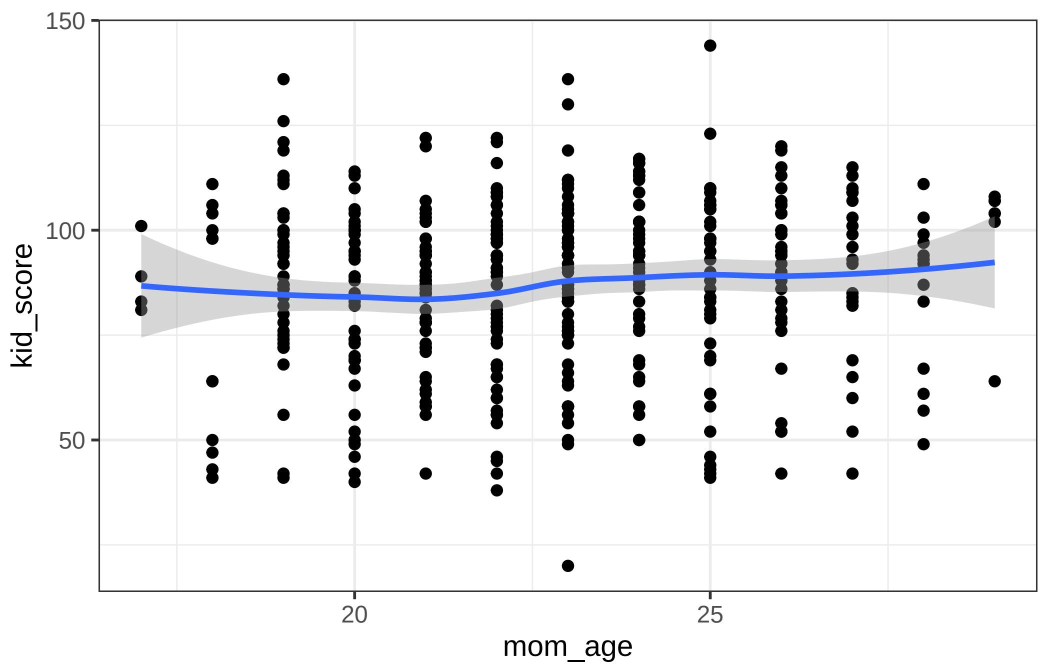
```
#colnames(kidiq)
kidiq |> ggplot(aes(x = factor(mom_hs), y = kid_score)) +
  geom_boxplot() + labs(x = "Mom education")
```




```
kidiq |> ggplot(aes(x = mom_iq, y = kid_score)) +
  geom_point()
```




```r
kidiq |> ggplot(aes(x = mom_age, y = kid_score)) +
  geom_point() + geom_smooth()
```



1. From the box plot, we note that the median IQ score is higher when mom possesses high-school education level. However the distributions overlap.
2. From the first scatter plot, we note that as the mom's IQ score increases, the child's score also increases i.e. there is a positive relationship between those variables.
3. The second scatter plot does not show any relationship between the mom's age and the child's IQ score.

```
# Fit first model with uninformative prior

y <- kidiq$kid_score
mu0 <- 80
sigma0 <- 10

# named list to input for stan function
data <- list(y = y,
             N = length(y),
             mu0 = mu0,
             sigma0 = sigma0)
fit <- stan(file = ("kids2.stan"),
            data = data,
            chains = 3,
            iter = 500)
```

```
fit
```

```
Inference for Stan model: kids2.
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      86.73    0.04 0.97    84.81    86.06    86.68    87.43    88.52   547
sigma   20.34    0.03 0.71    19.05    19.83    20.32    20.78    21.75   747
lp__ -1525.79    0.05 1.06 -1528.28 -1526.20 -1525.47 -1525.05 -1524.79   465
      Rhat
mu       1
sigma    1
lp__     1

Samples were drawn using NUTS(diag_e) at Sun Feb 12 19:11:05 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 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.

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

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

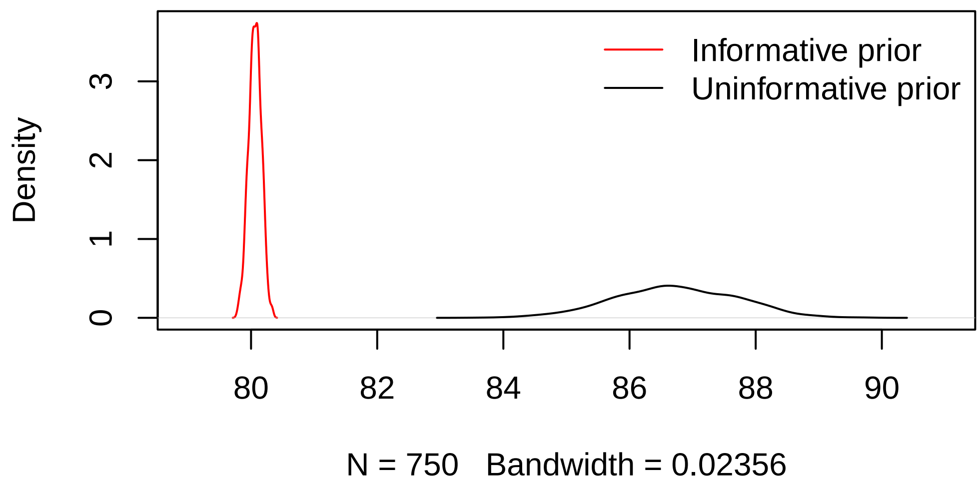
```r
fit.informative
```

```
Inference for Stan model: kids2.
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    80.00    80.06    80.13    80.25   544
sigma    21.46    0.03 0.74    20.06    20.97    21.40    21.93    22.96   693
lp__  -1548.39    0.05 0.98 -1550.89 -1548.77 -1548.13 -1547.67 -1547.39   368
      Rhat
mu       1
sigma    1
lp__     1

Samples were drawn using NUTS(diag_e) at Sun Feb 12 19:11:06 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).
```

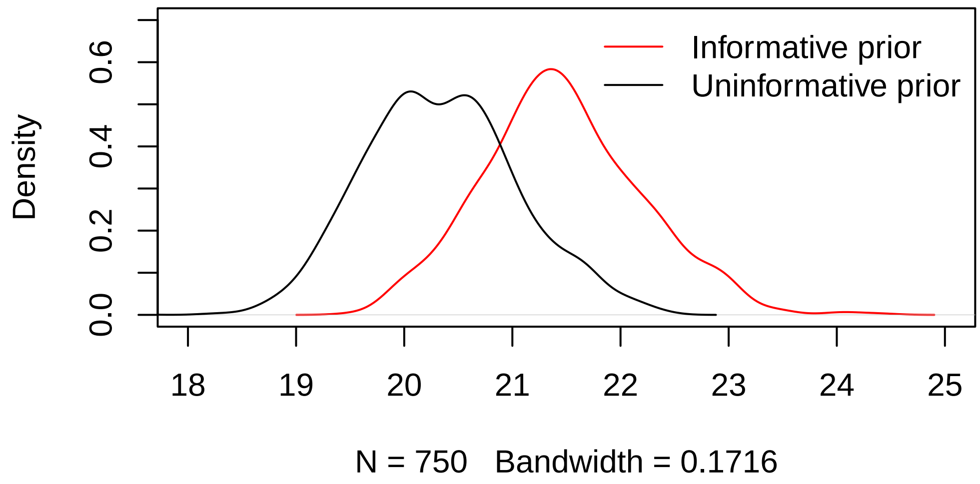
```r
post = extract(fit.informative)
post_samples_fit = extract(fit)
```

The estimate for the mean parameter decreases with the informative prior and the variance increases slightly. The standard deviation with the informative prior is much lower. This can be shown in the plots below where the densities of the posterior samples are flatter with the uninformative prior.

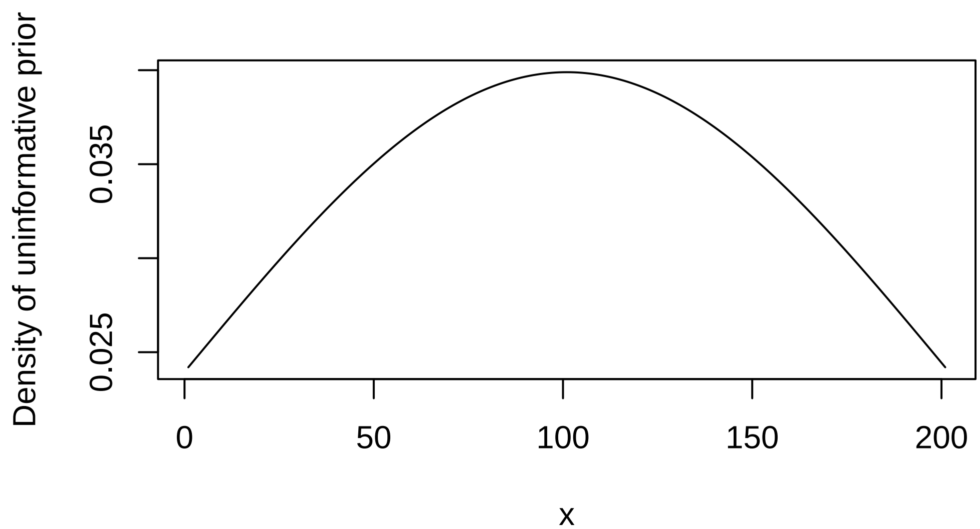
```
# Posterior densities

plot(density(post[["mu"]]), xlim = c(79,91),
     main = "", col = 'red')
lines(density(post_samples_fit[["mu"]]))
legend("topright", bty = 'n', col = c('red', 1),
       legend = c("Informative prior", "Uninformative prior"),
       lty =1)
```



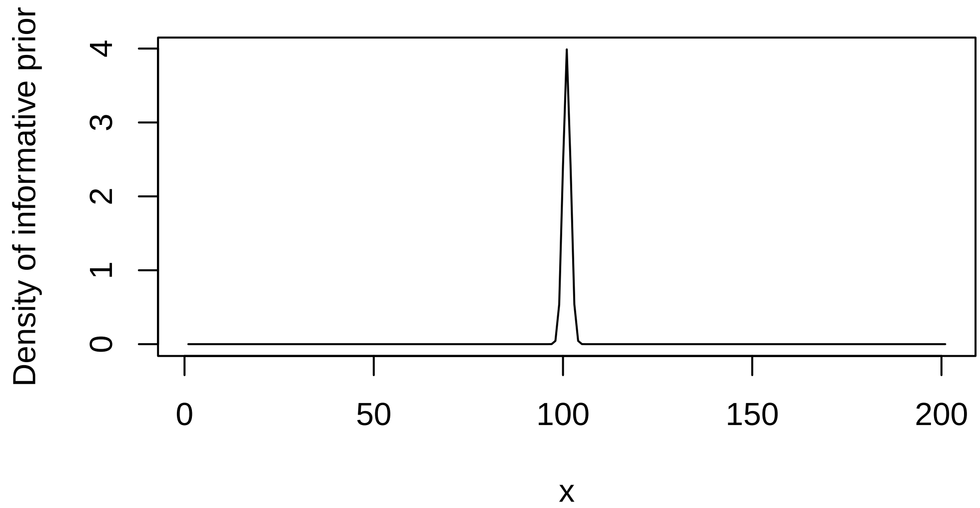

```
plot(density(post[["sigma"]]), xlim = c(18,25),
     main ="", ylim = c(0, .7), col = 'red')
lines(density(post_samples_fit[["sigma"]]))
legend("topright", bty = 'n', col = c('red', 1),
       legend = c("Informative prior", "Uninformative prior"),
       lty =1)
```




```
# Prior densities

plot(dnorm(seq(mu0-10,mu0 + 10, by =.1),
           mean = mu0, sd = 10), type ='l',
     ylab = "Density of uninformative prior", xlab = "x")
```




```
plot(dnorm(seq(mu0-10,mu0 + 10, by =.1),
            mean = mu0, sd = 0.1), type ='l',
     ylab = "Density of informative prior", xlab = "x")
```




```
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 = ("kids3.stan"),
            data = data,
            iter = 1000)
```

Question 3

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

```
m1 <- lm(kid_score ~ factor(mom_hs), kidiq)
summary(m1)
```

```
Call:
lm(formula = kid_score ~ factor(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 ***
factor(mom_hs)1   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
```

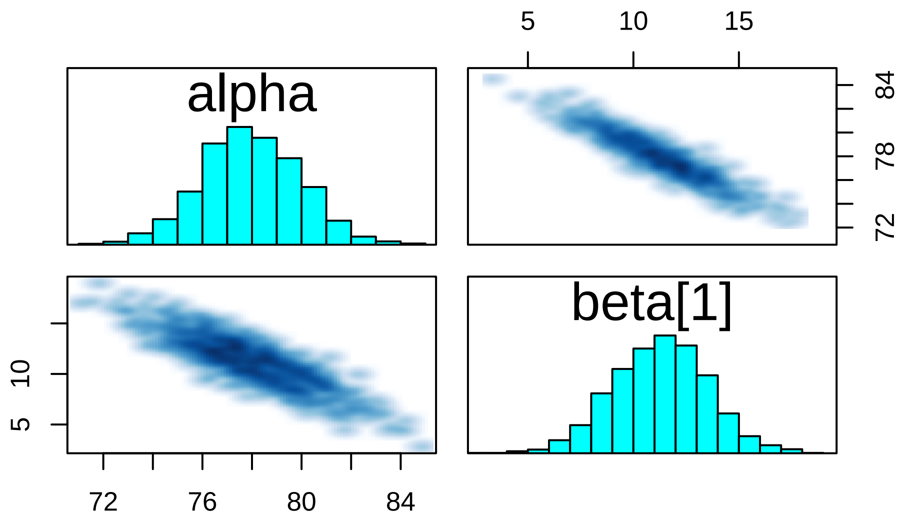
```r
fit2
```

```
Inference for Stan model: kids3.
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    0.07 1.99    73.98    76.57    77.90    79.33    81.83
beta[1]    11.24    0.08 2.24     6.77     9.75    11.29    12.70    15.54
sigma      19.79    0.02 0.65    18.56    19.34    19.77    20.25    21.04
lp__    -1514.30    0.04 1.16 -1517.46 -1514.78 -1514.00 -1513.47 -1512.98
        n_eff Rhat
alpha     778    1
beta[1]   769    1
sigma    1062    1
lp__      787    1

Samples were drawn using NUTS(diag_e) at Sun Feb 12 19:11:48 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).
```

The estimates of the Bayesian model are very similar to the MLE model.

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?

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



We note that the posterior samples between the parameters are highly correlated which indicates that the sampler is inefficient since we note that high slope values lead to low intercept values.

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(cbind(kidiq$mom_hs,
                     kidiq$mom_iq-mean(kidiq$mom_iq)),
               ncol = 1) # force this to be a matrix
K <- 2

data <- list(y = y, N = length(y),
             X =X, K = K)
fit3 <- stan(file = ("kids3.stan"),
             data = data,
             iter = 1000)
```

```
fit3
```

```
Inference for Stan model: kids3.
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     82.37    0.06 1.98    78.53    81.01    82.38    83.68    86.25
beta[1]    5.65    0.07 2.24     1.23     4.14     5.69     7.20     9.93
beta[2]    0.56    0.00 0.06     0.44     0.52     0.56     0.61     0.68
sigma     18.12    0.02 0.61    16.97    17.71    18.12    18.53    19.38
lp__   -1474.50    0.05 1.48 -1478.36 -1475.18 -1474.18 -1473.46 -1472.70
        n_eff Rhat
alpha     961 1.00
beta[1]   973 1.00
beta[2]  1130 1.00
sigma    1203 1.00
lp__      926 1.01

Samples were drawn using NUTS(diag_e) at Sun Feb 12 19:11:51 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).
```

It is expected that as the centered mom's IQ score increases by 1 unit, the child's IQ score increases by 0.56 units.

Question 5

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

```r
m2 <- lm(kid_score ~ factor(mom_hs) + I(mom_iq-mean(mom_iq)), kidiq)
summary(m2)
```

```

Call:
lm(formula = kid_score ~ factor(mom_hs) + I(mom_iq - mean(mom_iq)),
    data = kidiq)

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 ***
```

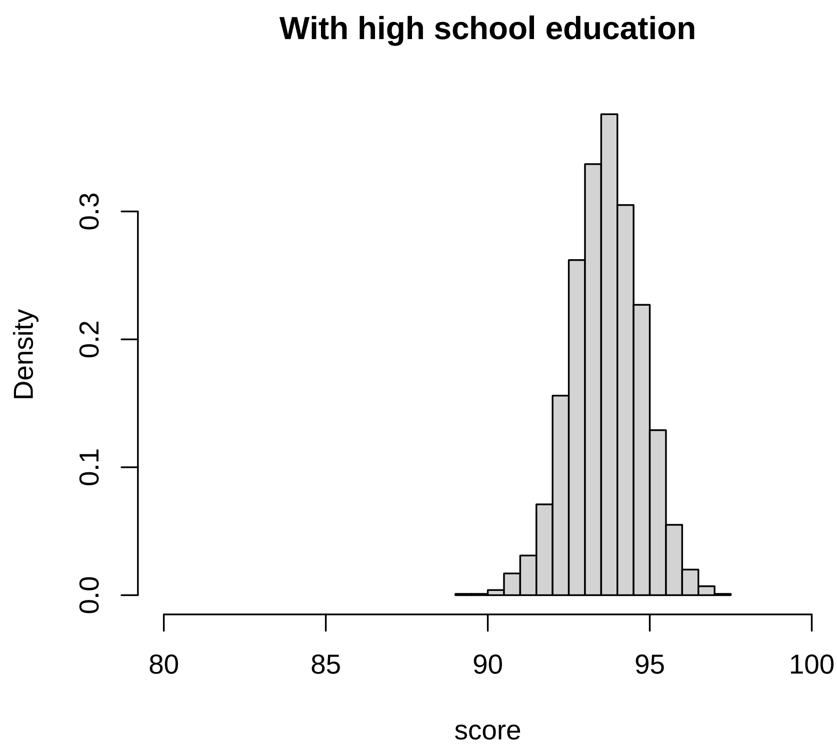
```
factor(mom_hs)1           5.95012    2.21181   2.690  0.00742 **
I(mom_iq - mean(mom_iq))  0.56391    0.06057   9.309  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

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
```

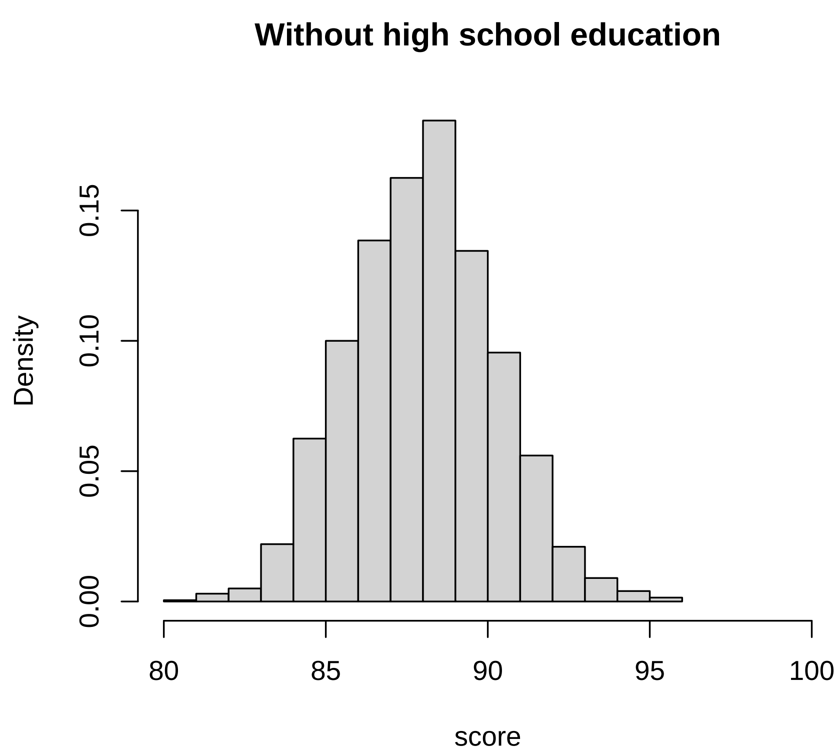
The results are similar.

Question 6

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

```r
m <- mean(kidiq$mom_iq)
post_samples = extract(fit3)

par(mfrow = c(2,1))
post_y1 <- (post_samples[["alpha"]] +
              post_samples[["beta"]][,1] + (110 - m)*post_samples[["beta"]][,2])
hist(post_y1,  freq = F, xlim =
       c(80, 100),
     main = "With high school education",
     xlab  = "score")

post_y0 <- (post_samples[["alpha"]] +
              0*post_samples[["beta"]][,1] + (110 - m)*post_samples[["beta"]][,2])
hist(post_y0 , freq = F,
     xlim = c(80, 100),
     main = "Without high school education",
     xlab  = "score")
```


With high school education
Density
0.0
0.1
0.2
0.3
80
85
90
95
100
score


Without high school education
Density
0.00
0.05
0.10
0.15
80
85
90
95
100
score

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.

```
# Sample from posterior predictive distribution

ynew <- rnorm(n = 1000)*(post_samples[["sigma"]]) +
  post_samples[["alpha"]] +
  1*post_samples[["beta"]][,1] +
  (95 - m)*post_samples[["beta"]][,2]
hist(ynew, freq = F)
```