STA2201 Lab #5

Timothy Jordan Regis

13/02/2023

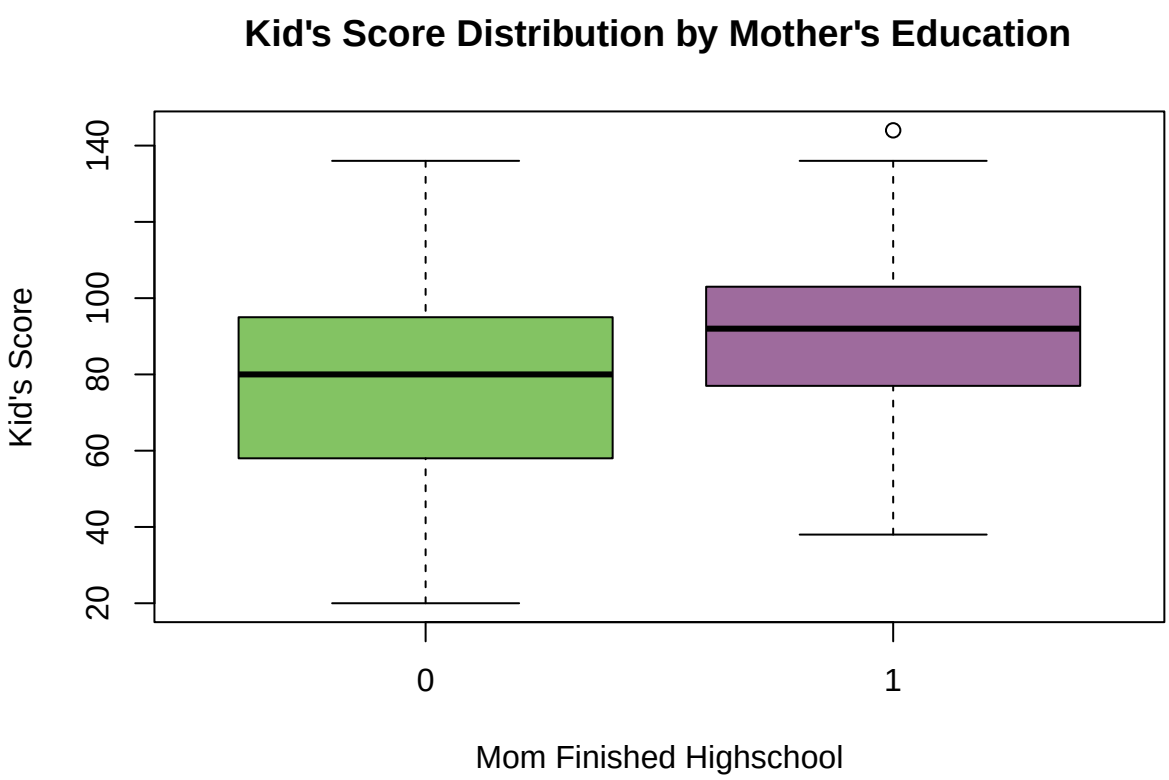
```
## # A tibble: 6 x 4
##   kid_score mom_hs mom_iq mom_age
##       <int>  <dbl>  <dbl>   <int>
## 1        65      1  121.       27
## 2        98      1   89.4      25
## 3        85      1  115.       27
## 4        83      1   99.4      25
## 5       115      1   92.7      27
## 6        98      0  108.       18
```

```
## 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       86.74    0.04 1.02    84.82    86.00    86.74    87.40    88.64   746
## sigma    20.40    0.04 0.72    19.10    19.88    20.37    20.90    21.86   356
## lp__  -1525.85    0.06 1.02 -1528.66 -1526.29 -1525.55 -1525.08 -1524.80   300
##       Rhat
## mu    1.00
## sigma 1.00
## lp__  1.01
##
## Samples were drawn using NUTS(diag_e) at Mon Feb 13 21:43:00 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).
```

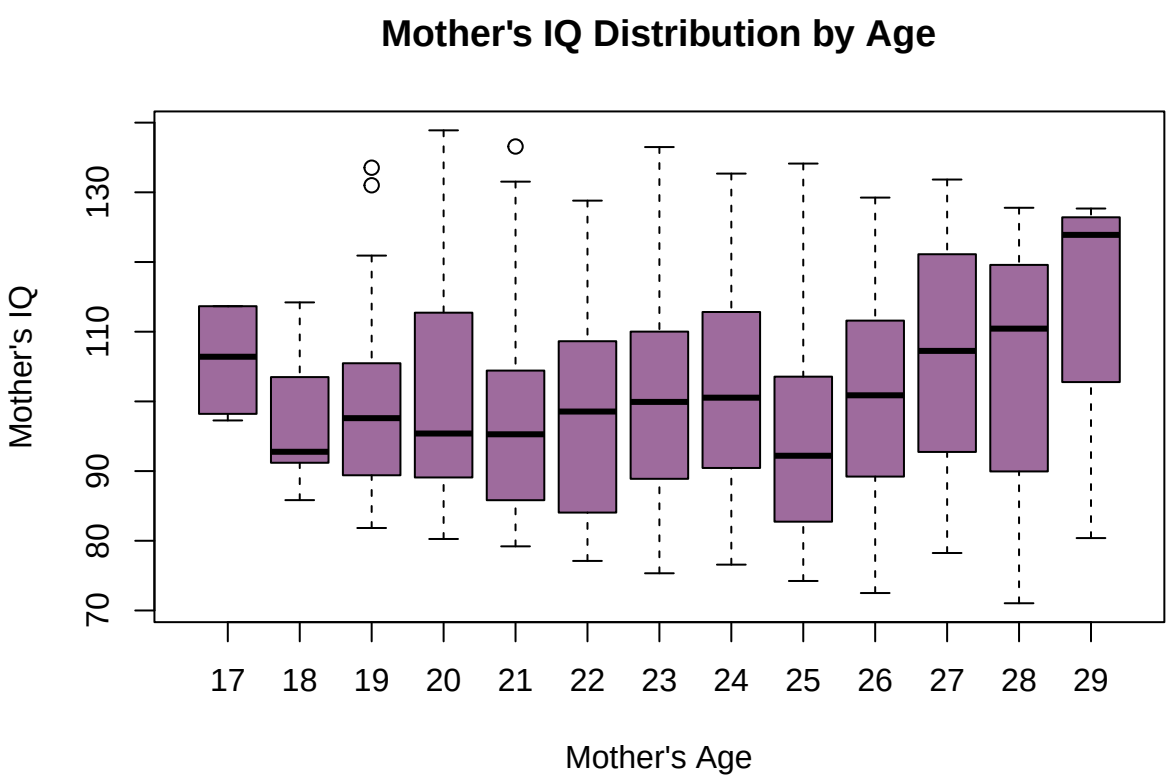
Question 1

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

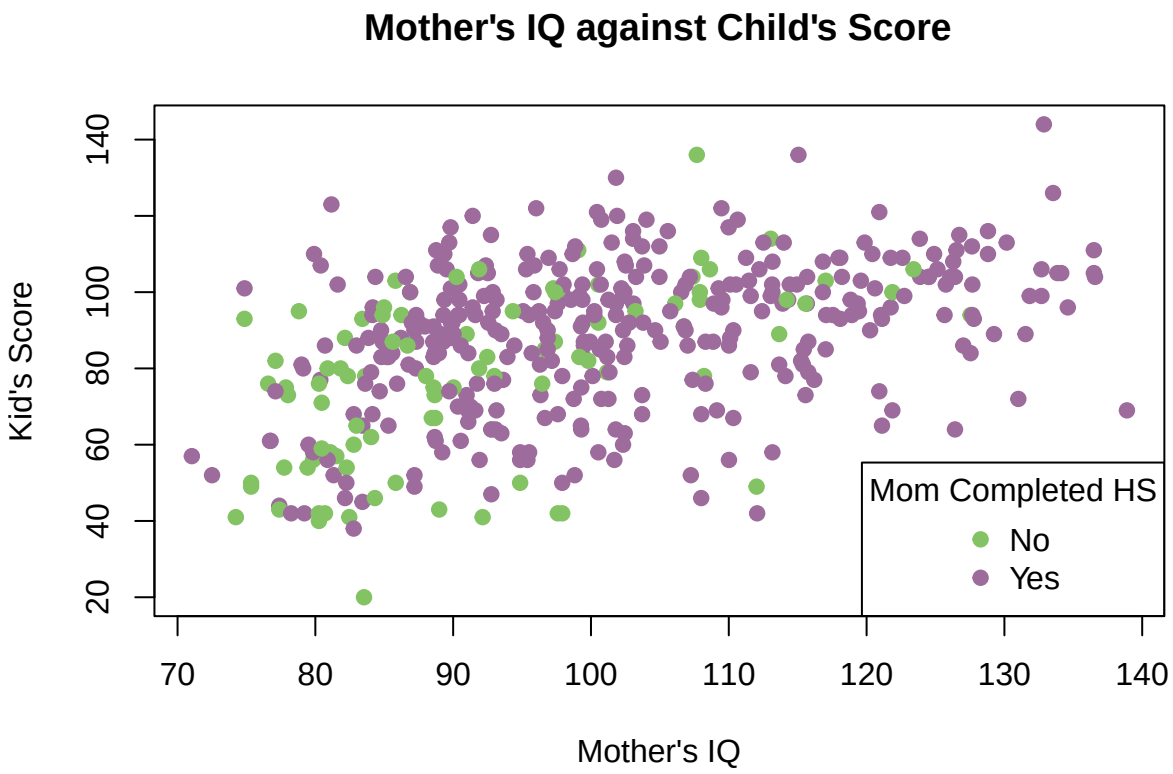
- Explain what your graph/ tables show
- Choose a graph type that's appropriate to the data type



In our first graph, we plot boxplots of the distribution of kid scores, grouped by the status of the mother's high school education. As we can see, there is a slightly higher average of scores for kids whose mothers had successfully completed high school education, but this is only a small difference as we can see each box overlapping by a considerable amount. We can also see that the lowest observations are seen in the no-high school group, with the highest obersvations landing in the high school group. The lack of a significant difference here is somewhat surprising as we would likely expect a mother's education to have more impact on a child, but this result could be representative of the fact that a child's intelligence may come more from their environment or school.



In our second plot, we have displayed the distribution of the mothers' IQ scores across each age group to determine if there's a significant relationship between the two variables. As we can see, all bars overlap a considerable amount, with no significant differences being shown. However, we do notice a slight trend, in the first half of the graph, before the age 24, there is little change at all between the IQ distribution, whereaas after age 24, there is a small positive relationship between age and IQ, with the mothers that are 28 years old holding the highest IQs on average. This result is potentially suggesting that mothers who have children later in life tend to be more intelligent, but the significant overlap provents any definitive conclusion from being made.



Finally, we plot the mother's IQ against their kid's score, coloring by the mother's high school education status. As we can see, there is a subtle positive relationship between a mother's IQ and their child's score. Furthermore, this trend holds if we look separately at the no-high school group (green) and the high school group (purple), where we can also see that the no-high school group tends to see lower scores in general, agreeing with our previous boxplot. This result is somewhat expected, however, we were surprised at how well the relationship held in each group, as well as the closeness of points between each high school group.

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.

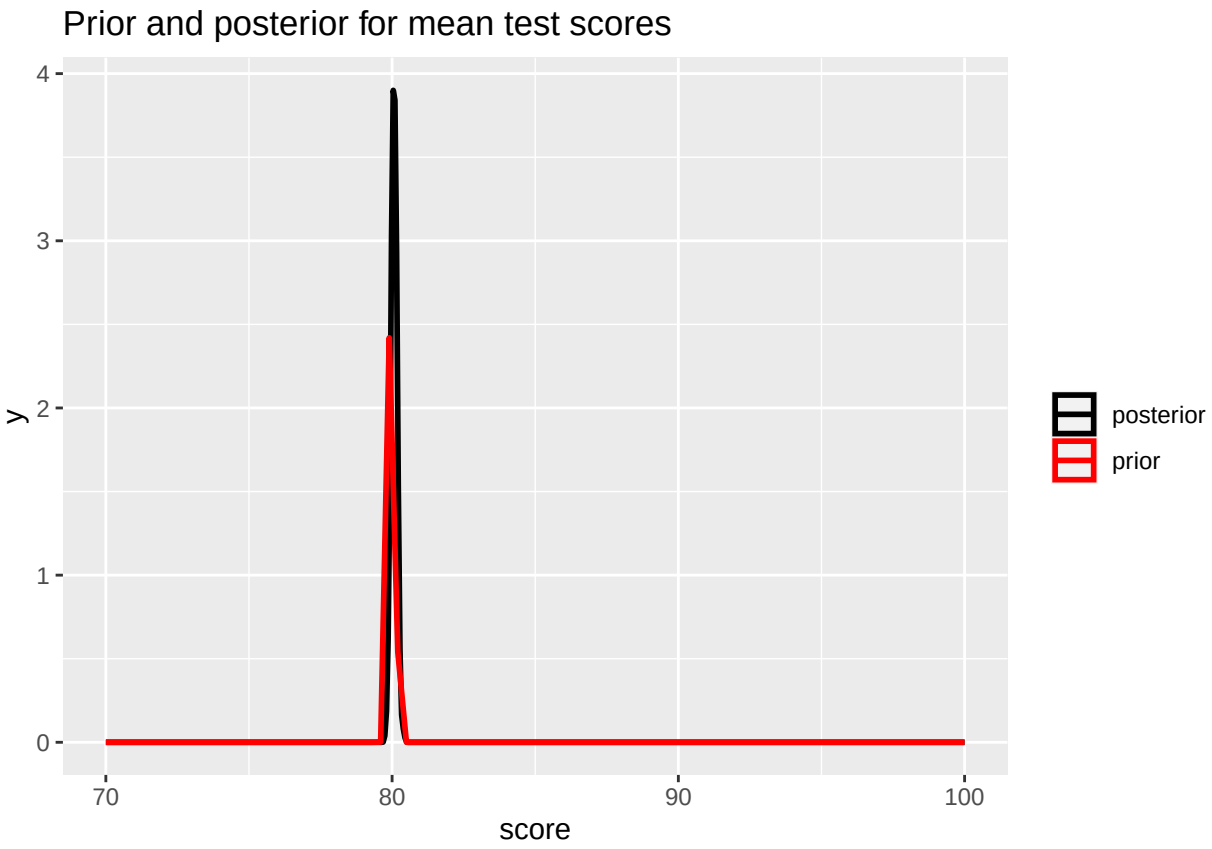
```
## Inference for Stan model: kids2.
## 3 chains, each with iter=500; warmup=250; thin=1;
## post-warmup draws per chain=250, total post-warmup draws=750.
##
##           mean se_mean   sd     2.5%      25%      50%      75%    97.5% n_eff
## mu       86.74    0.04 1.02    84.82    86.00    86.74    87.40    88.64   746
## sigma    20.40    0.04 0.72    19.10    19.88    20.37    20.90    21.86   356
## lp__  -1525.85    0.06 1.02 -1528.66 -1526.29 -1525.55 -1525.08 -1524.80   300
##       Rhat
## mu    1.00
## sigma 1.00
## lp__  1.01
##
```

```
## Samples were drawn using NUTS(diag_e) at Mon Feb 13 21:43:00 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).


## 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       80.07    0.00 0.10    79.88    80.00    80.07    80.14    80.25   585
## sigma    21.42    0.03 0.71    20.14    20.98    21.38    21.84    22.94   686
## lp__  -1548.34    0.05 1.03 -1550.96 -1548.65 -1548.01 -1547.62 -1547.40   365
##       Rhat
## mu    1.00
## sigma 1.01
## lp__  1.01
##
## Samples were drawn using NUTS(diag_e) at Mon Feb 13 21:43:01 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).
```

As we can see, the estimates do change slightly, going from 86.74 to 80.06 on mu, 20.40 to 21.50 on sigma, and -1525.72 to -1548.40 on lp. Furthermore, we also see changes in the se_mean and sd of each estimator. (Example from one run, I can't figure out how to save runs yet)

```
## Warning: Using ‘size‘ aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use ‘linewidth‘ instead.
```



As we can see, the prior is now much more narrow than previously

Question 3

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

a)

```
## 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      77.92    0.08 1.99    73.90    76.67    77.97    79.19    81.71
## beta[1]    11.28    0.08 2.22     6.94     9.77    11.29    12.69    15.74
## sigma      19.81    0.02 0.66    18.58    19.34    19.80    20.28    21.11
## lp__    -1514.35    0.04 1.21 -1517.53 -1514.89 -1514.03 -1513.45 -1512.99
##         n_eff Rhat
## alpha     700 1.00
## beta[1]   692 1.00
## sigma     931 1.00
## lp__      811 1.01
##
```

```
## Samples were drawn using NUTS(diag_e) at Mon Feb 13 21:43:26 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).


##
## 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
```

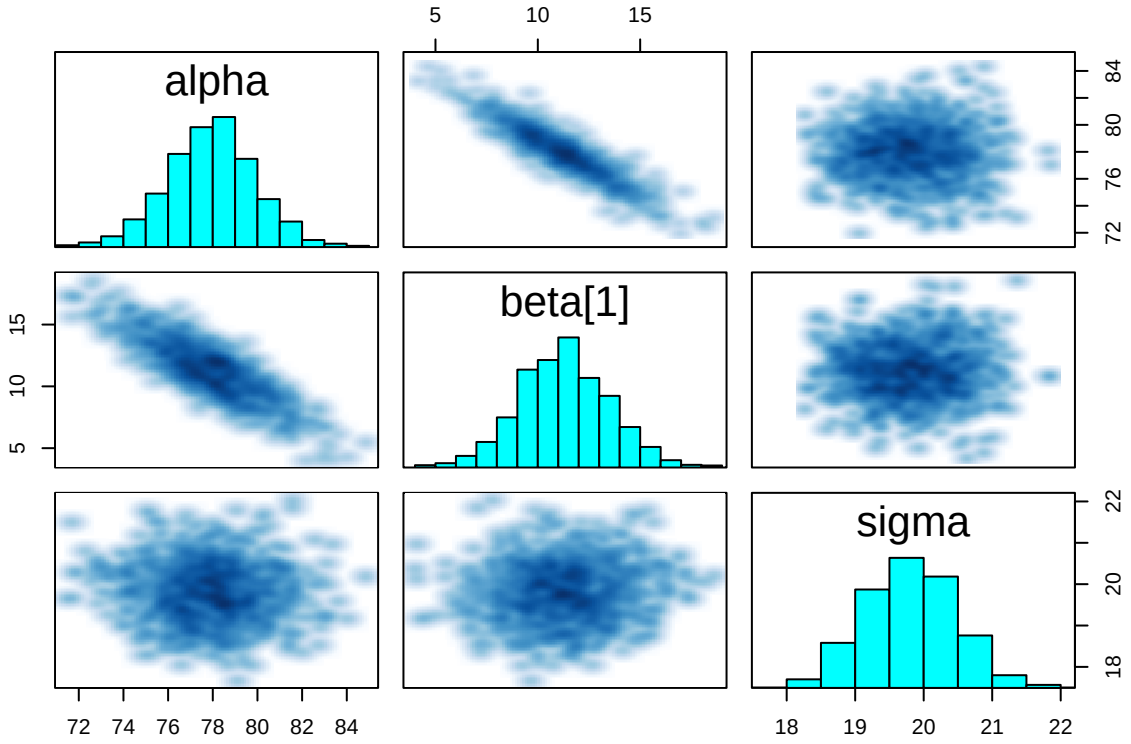
Results are extremely close, only off by a few decimal points:

(Example from one Run (couldn't figure out how to save))

77.92 vs. 77.55 for the Intercept, and

11.30 vs. 11.77 for the mother's high school education

b)



As we can see, while their distributions with sigma are fine, we can see that the joint distribution between alpha and beta[1] is highly correlated by the distinct grouping of points and sharp decreasing slope. This may be a potential problem as it disrupts the model's fit and sampling abilities.

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.

```
## Inference for Stan model: kids3new.
## 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      82.30    0.06 1.93    78.45    81.03    82.26    83.58    86.18
## beta[1]     5.74    0.07 2.21     1.40     4.25     5.73     7.20    10.15
## beta[2]     0.56    0.00 0.06     0.45     0.52     0.56     0.60     0.67
## sigma      18.11    0.02 0.62    16.94    17.69    18.10    18.50    19.45
## lp__    -1474.45    0.05 1.45 -1478.20 -1475.13 -1474.13 -1473.42 -1472.63
##         n_eff Rhat
## alpha     968    1
## beta[1]   948    1
## beta[2]  1408    1
```

```
## sigma    1260    1
## lp__      792    1
##
## Samples were drawn using NUTS(diag_e) at Mon Feb 13 21:43:50 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).
```

With an estimate of 0.57, our model suggests that a 1 unit increase in the mother's mean centered IQ leads to an increase in their child's score of approximately 0.57.

Question 5

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

```
## Inference for Stan model: kids3new.
## 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      82.30    0.06 1.93    78.45    81.03    82.26    83.58    86.18
## beta[1]     5.74    0.07 2.21     1.40     4.25     5.73     7.20    10.15
## beta[2]     0.56    0.00 0.06     0.45     0.52     0.56     0.60     0.67
## sigma      18.11    0.02 0.62    16.94    17.69    18.10    18.50    19.45
## lp__    -1474.45    0.05 1.45 -1478.20 -1475.13 -1474.13 -1473.42 -1472.63
##         n_eff Rhat
## alpha     968    1
## beta[1]   948    1
## beta[2]  1408    1
## sigma    1260    1
## lp__      792    1
##
## Samples were drawn using NUTS(diag_e) at Mon Feb 13 21:43:50 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).
```

```
##
## Call:
## lm(formula = kid_score ~ mom_hs + mom_iq_c, data = kidiq_q5)
##
## 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
```

As we can see, there is only a small decimal point difference between the lm estimates and Stan's:

(Example from one run)

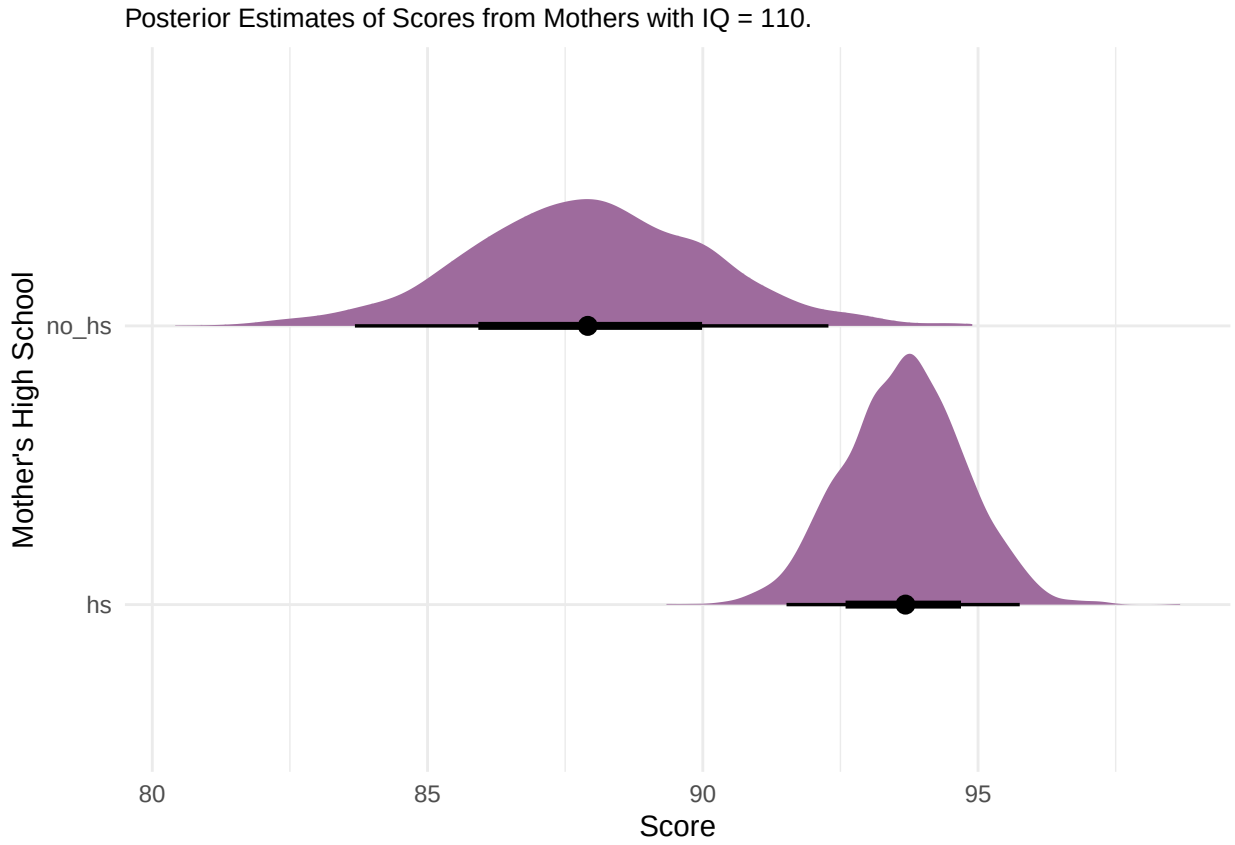
82.32 vs. 82.12 for the Intercept

5.73 vs. 5.95 for the mother's high school education

0.57 vs. 0.56 for the mother's IQ

Question 6

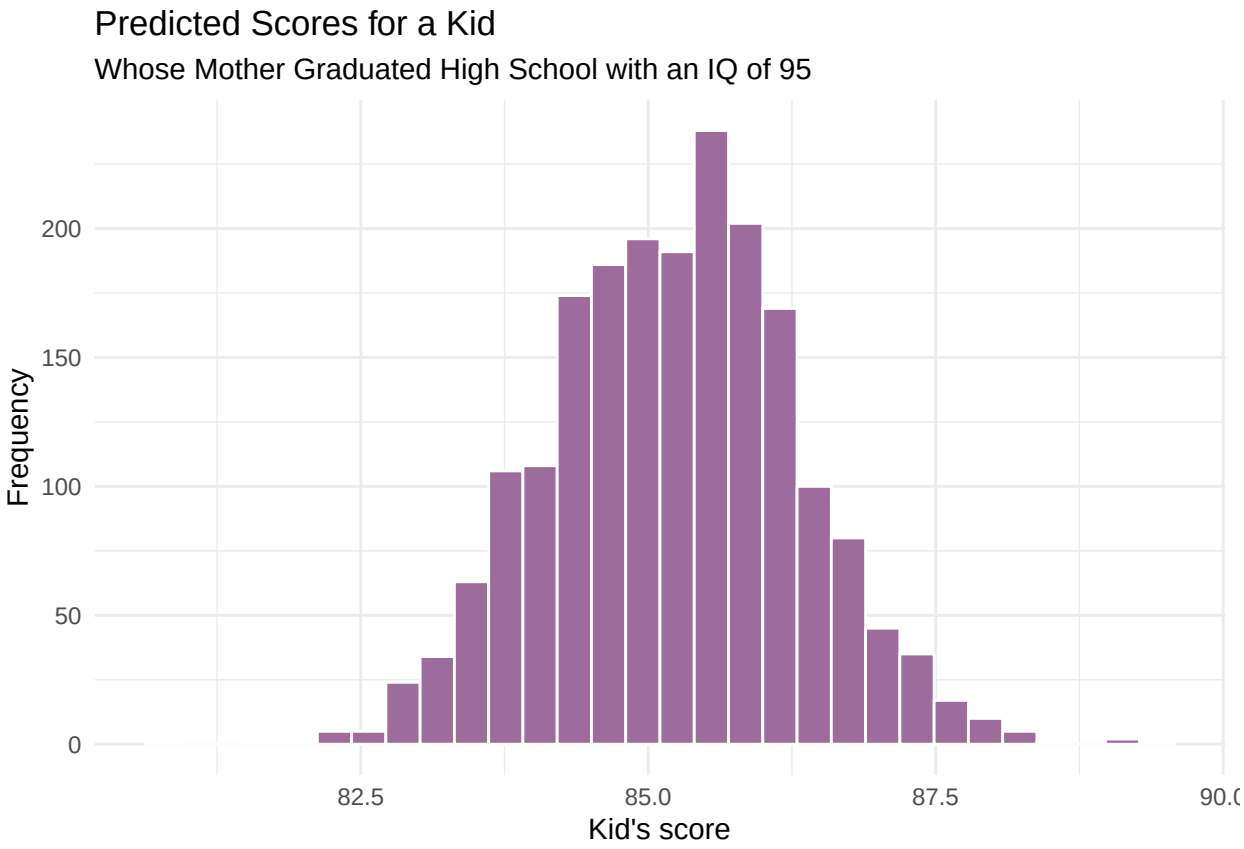
Plot the posterior estimates of scores by education of mother for mothers who have an IQ of 110.



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.

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```


Predicted Scores for a Kid
Whose Mother Graduated High School with an IQ of 95
Frequency
200
150
100
50
0
82.5
85.0
87.5
Kid's score