STA2201 Lab5

Ichiro Hashimoto

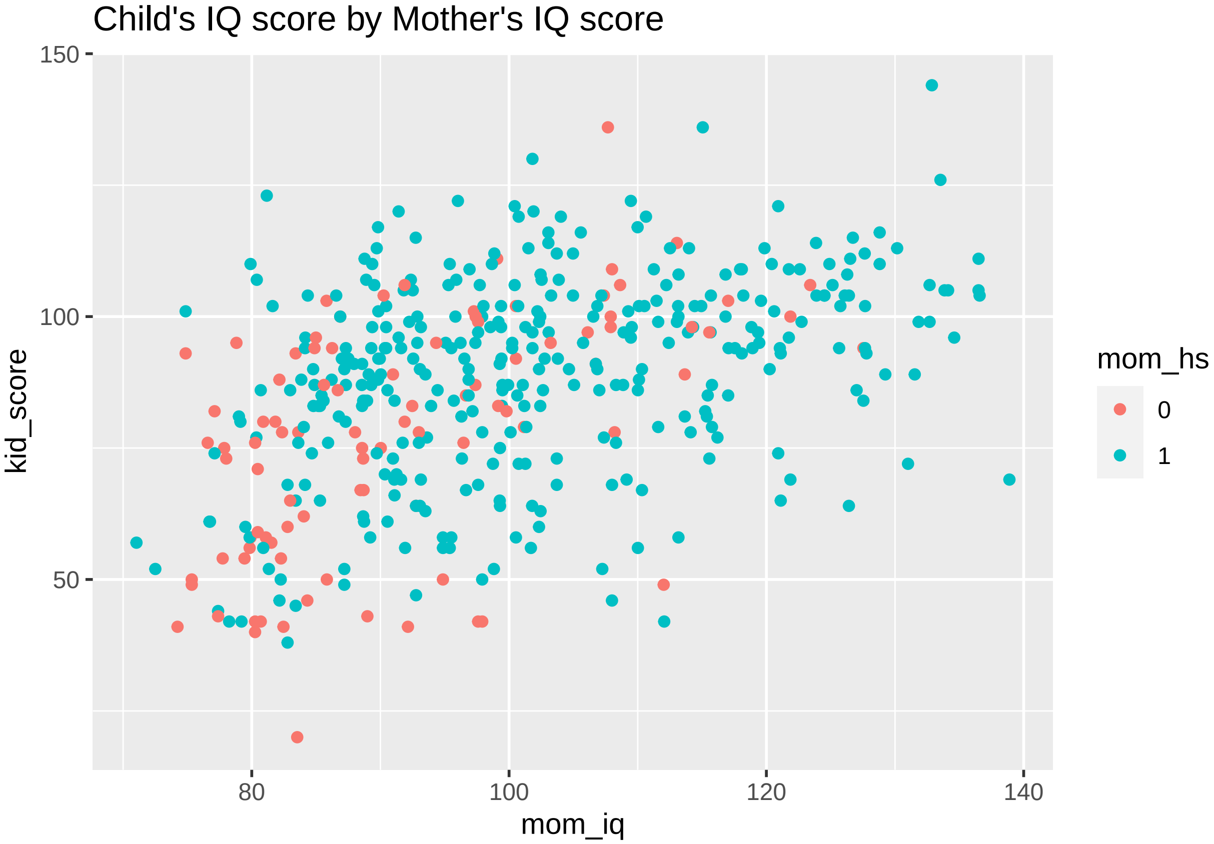
Question 1

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

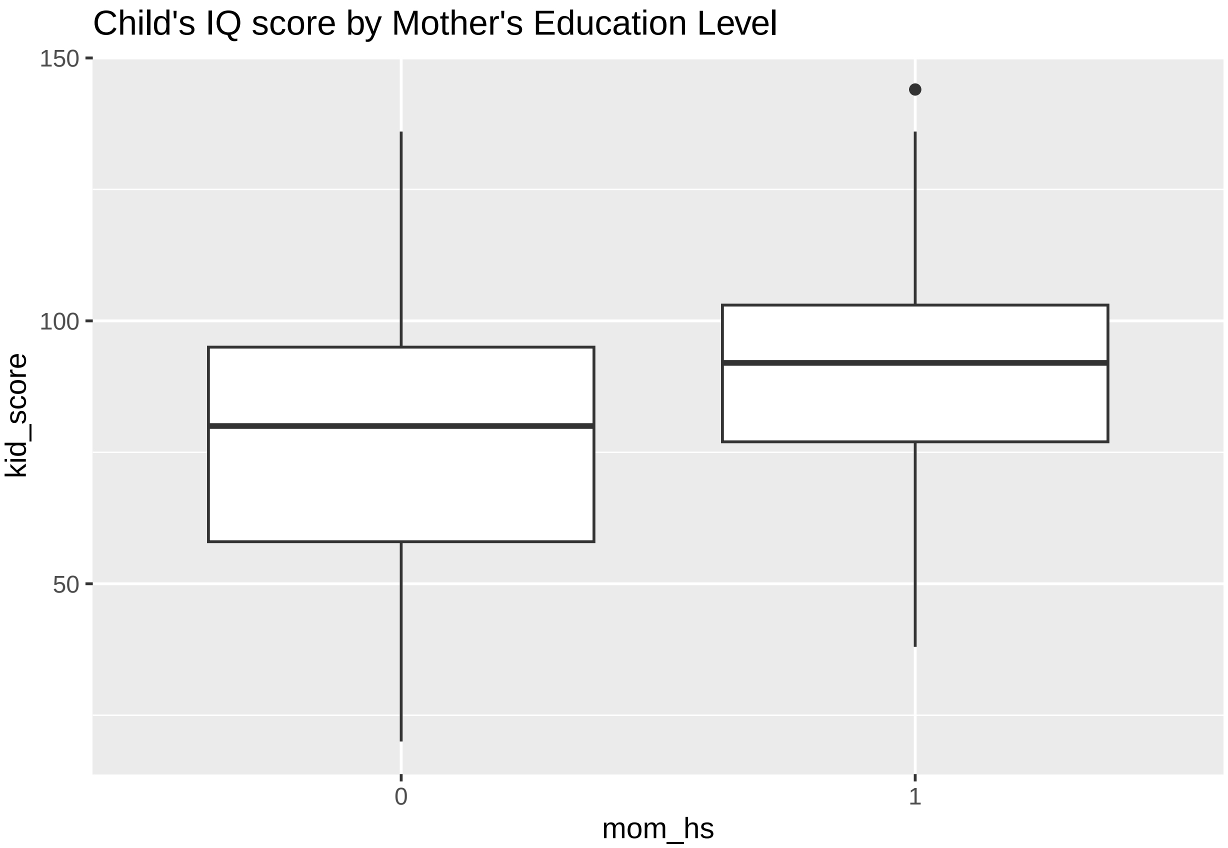
- Explain what your graph/ tables show
- Choose a graph type that's appropriate to the data type

Answer

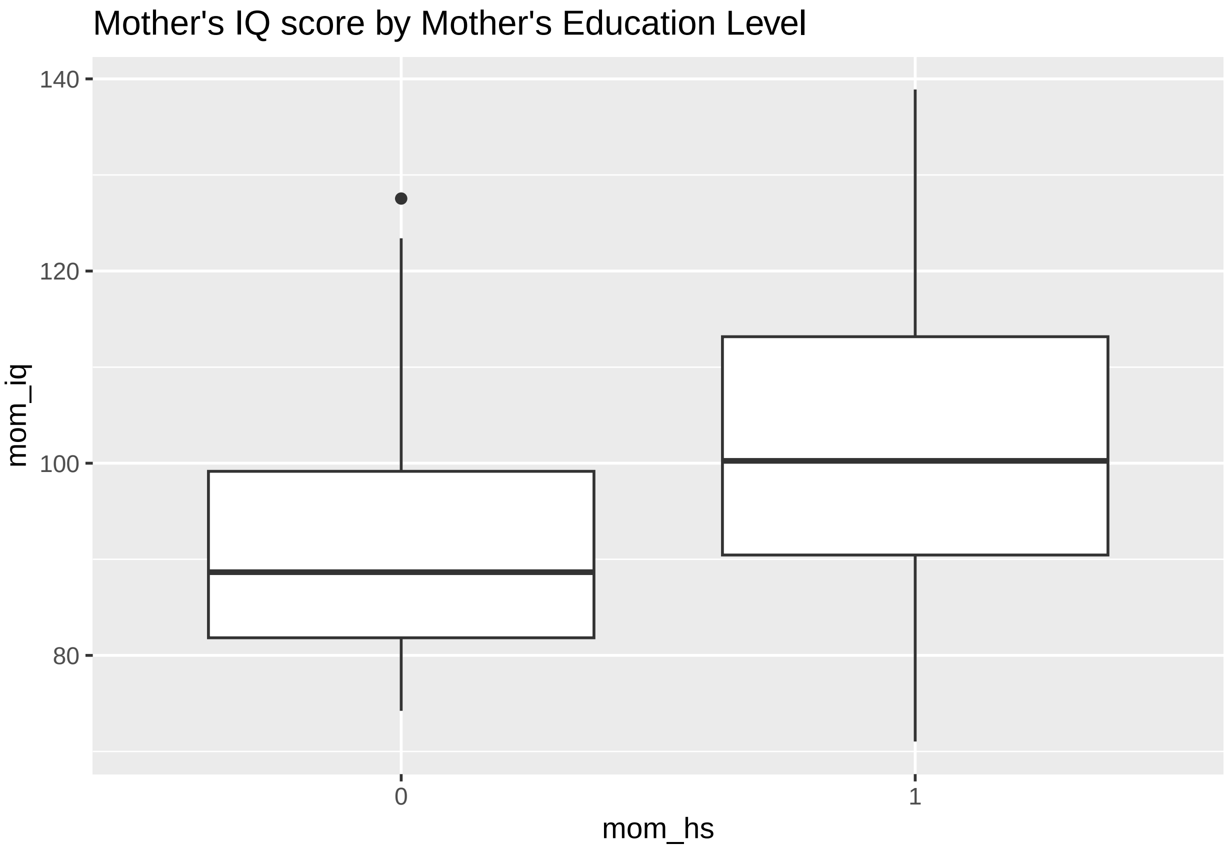
The following plot indicates not only that there are linear relation between child's IQ score and mother's IQ score but also indicates that the distribution seems different for child's IQ score whose mother with a high school degree and whose mother without a high school.



From the following boxplot, we can further confirm that child's IQ score is distributed higher if their mother has a high school degree.



At the same time, the following boxplot also indicated that mother's IQ score is higher if she has a high school degree.



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.

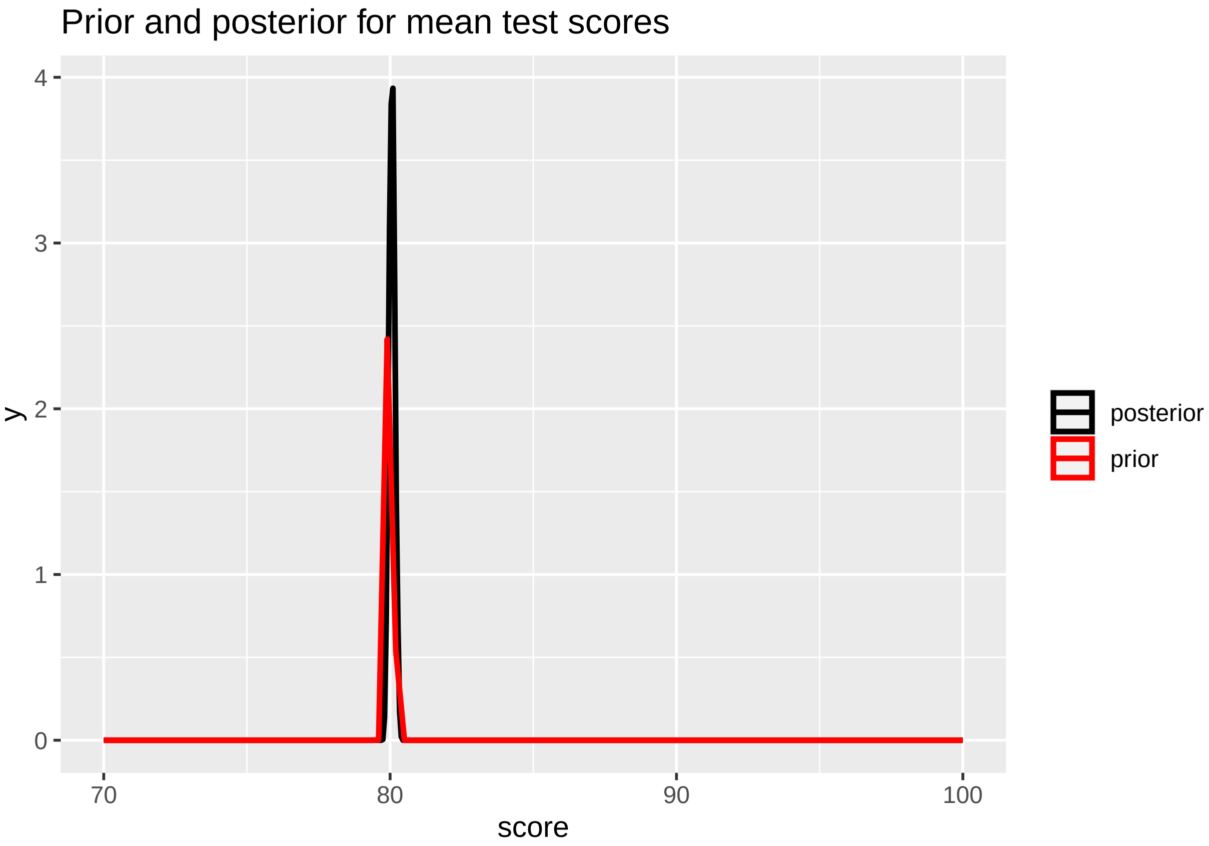
Answer

From the following summary of the new model, we can see that the estimate of 'mu' has changed quite a bit. As expected, it was pulled to the prior information (mu = 80).

```
## 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    79.88    79.99    80.07    80.13    80.26   627
## sigma    21.41    0.03 0.76    20.04    20.84    21.41    21.89    22.88   705
## lp__  -1548.39    0.05 0.93 -1550.69 -1548.90 -1548.13 -1547.66 -1547.40   334
##       Rhat
## mu    1.00
## sigma 1.00
## lp__  1.01
##
## Samples were drawn using NUTS(diag_e) at Mon Feb 13 00:37:27 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 prior and posterior densities are plotted as follows:

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



Question 3

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

Answer

Here are the summary of the new model where `mom_hs` was added as a covariate:

```
## 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    67.01    0.18 4.37  58.13  64.14  67.01  69.93  75.50
## beta[1]  11.09    0.10 2.40   6.43   9.51  11.09  12.65  15.90
## sigma    19.81    0.02 0.70  18.45  19.35  19.77  20.27  21.23
## lp__    -1514.38    0.05 1.26 -1517.60 -1515.02 -1514.06 -1513.44 -1512.88
##         n_eff Rhat
## alpha     621 1.00
## beta[1]   612 1.00
## sigma     899 1.00
## lp__      545 1.01
##
## Samples were drawn using NUTS(diag_e) at Mon Feb 13 00:37:52 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).
```

In addition, the following is a summary from a simple linear regression. Comparing estimates from these two models, we can confirm that both give similar estimates.

```
##
## 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)   65.777      4.255  15.458  < 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
```

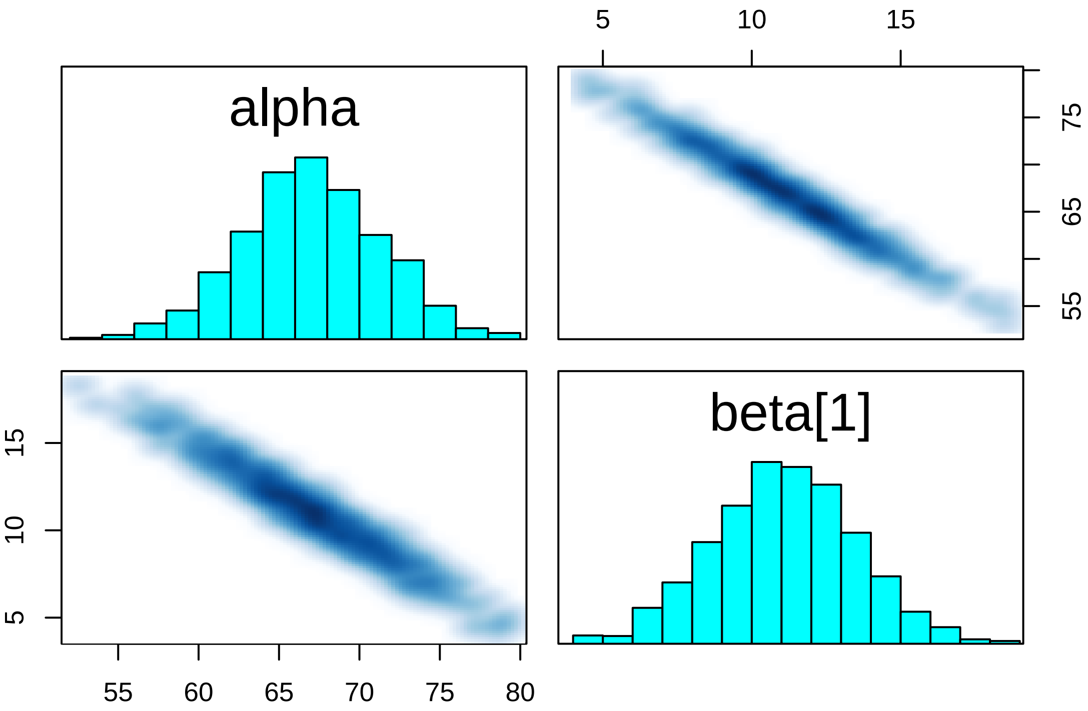
b) Do a pairs plot to investigate the joint sample distributions of the slope and intercept. Comment briefly on what you see. Is this potentially a problem?

Answer

From the joint sample distributions of the slope and intercept, we find that they have a clear linear relation where they should be independently distributed. This is because we did not do centering.

```
## Warning in par(usr): argument 1 does not name a graphical parameter
```

```
## Warning in par(usr): argument 1 does not name a graphical parameter
```



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.

Answer

The following is the summary of the new model. The estimate indicates that if mum's IQ increase by 1, child's IQ also increases by 0.56.

```
## 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      76.20    0.17 4.01    68.15    73.38    76.12    78.96    84.05
## beta[1]     5.93    0.09 2.20     1.62     4.48     5.98     7.42    10.18
## beta[2]     0.57    0.00 0.06     0.45     0.52     0.56     0.61     0.69
## sigma      18.13    0.02 0.61    17.02    17.69    18.12    18.55    19.36
## lp__    -1474.37    0.05 1.42 -1477.86 -1475.01 -1474.05 -1473.31 -1472.63
##         n_eff Rhat
## alpha     567 1.01
## beta[1]   576 1.01
## beta[2]  1073 1.00
## sigma    1155 1.00
## lp__      808 1.00
##
## Samples were drawn using NUTS(diag_e) at Mon Feb 13 00:37:55 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 5

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

Answer

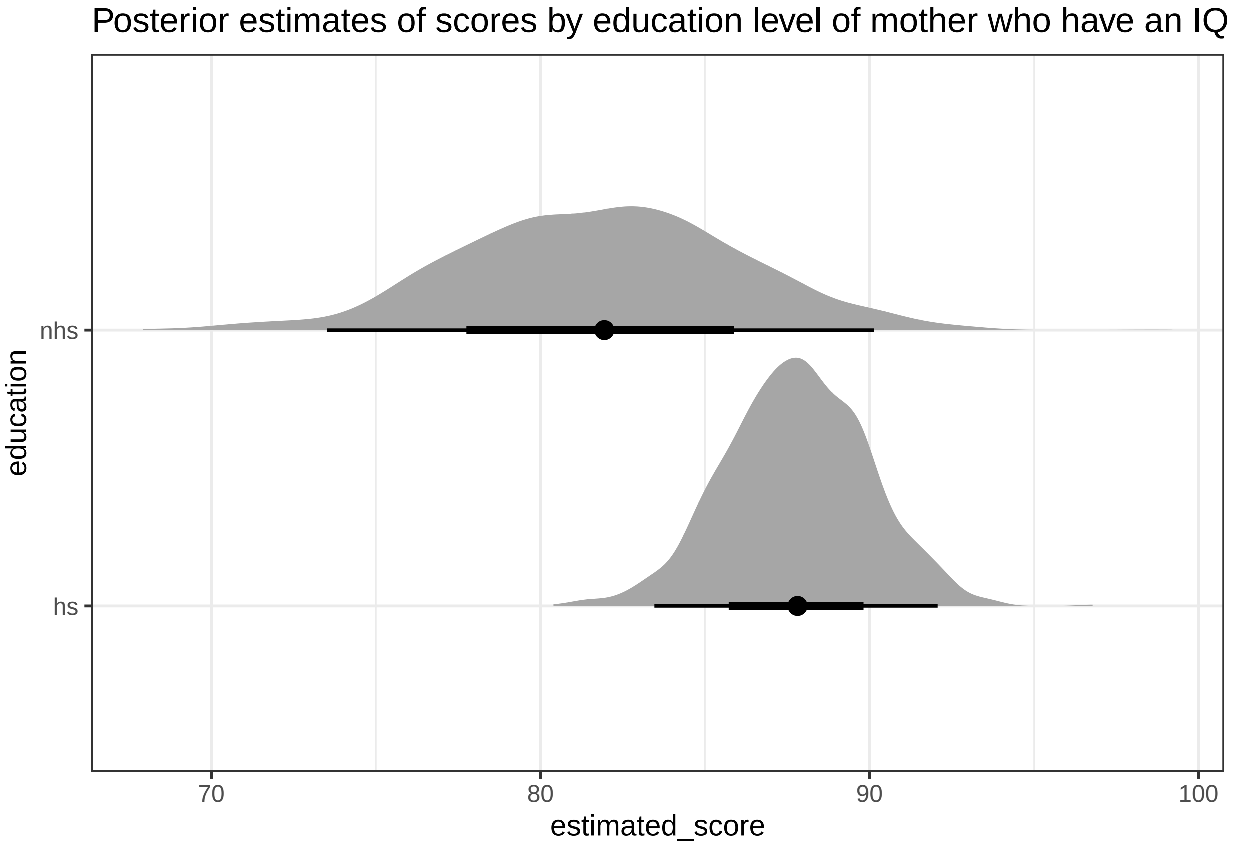
The following summary from ‘lm()’ indicates that the results from Stan is similar to this one.

```
##
## Call:
## lm(formula = kid_score ~ mom_hs + mom_iq, data = kidiq2)
##
## 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) 76.17203    4.04446  18.834  < 2e-16 ***
## mom_hs       5.95012    2.21181   2.690  0.00742 **
## 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
```

Question 6

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

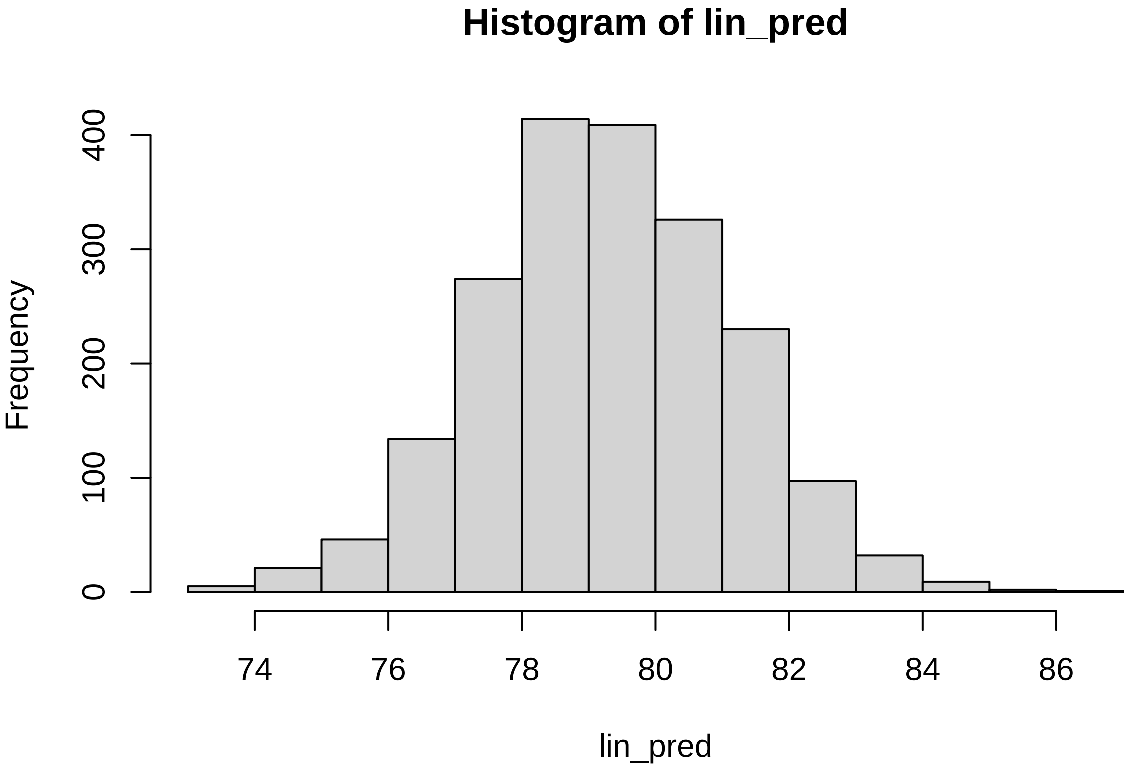
Answer



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.

Answer


Histogram of lin_pred
Frequency
lin_pred