Lab 5

Yaqi Shi, 1003813180

2023-02-12

Introduction

Today we will be starting off using Stan, looking at the kid's test score data set (available in resources for the Gelman Hill textbook).

The data look like this:

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

```
## # 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
```

```
kidiq$mom_hs <- as.character(kidiq$mom_hs)
```

As well as the kid's test scores, we have a binary variable indicating whether or not the mother completed high school, the mother's IQ and age.

Descriptives

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

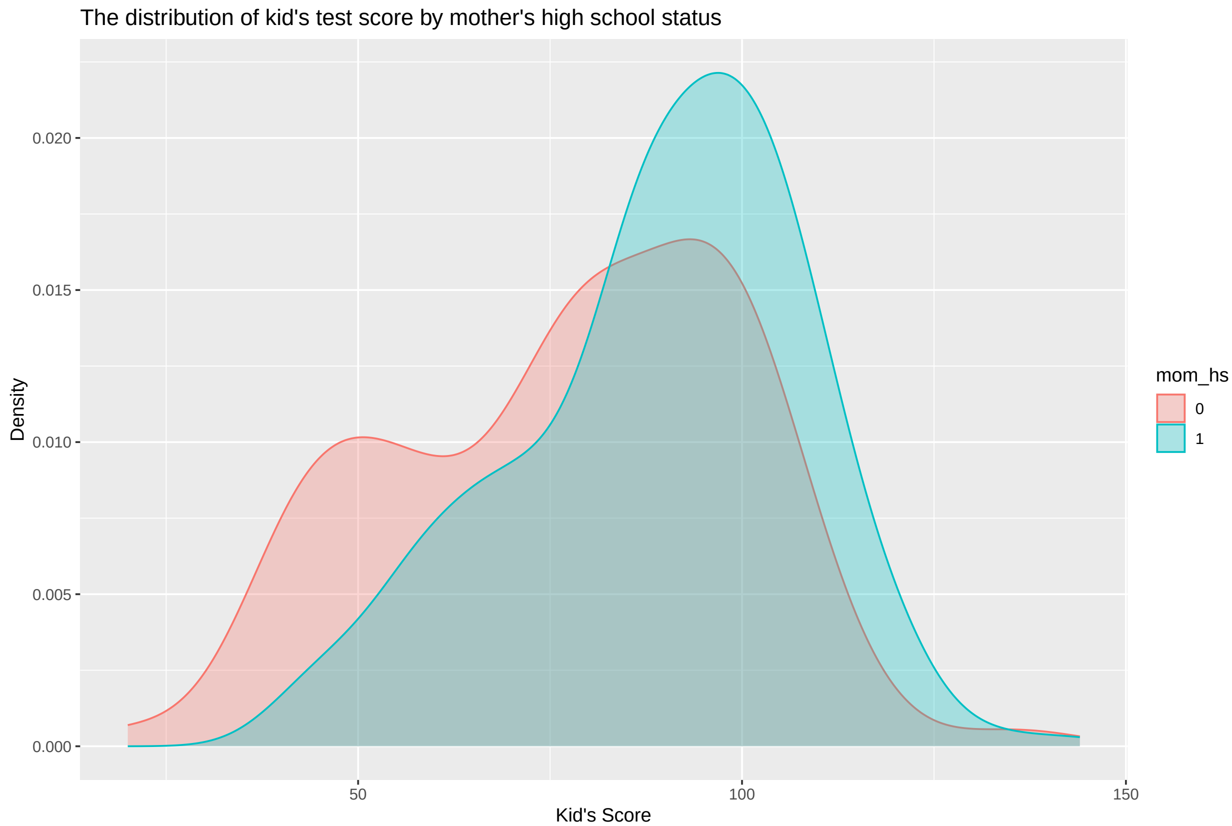
Answer

Since we have the kid's test score as the response variable, we will explore the dataset with the focus on that. Here are three plots that demonstrate the distribution:

1 - The distribution of kid's test score with different mother's high school status.

The type of the plot is a density plot and the curve are differentiated by mother's high school status. The purpose of this plot is to explore the difference in distribution for kid's test score with different mother's high school status. We can see that the kid's test score tends to be higher for mother who completed high school, which is reasonable in general.

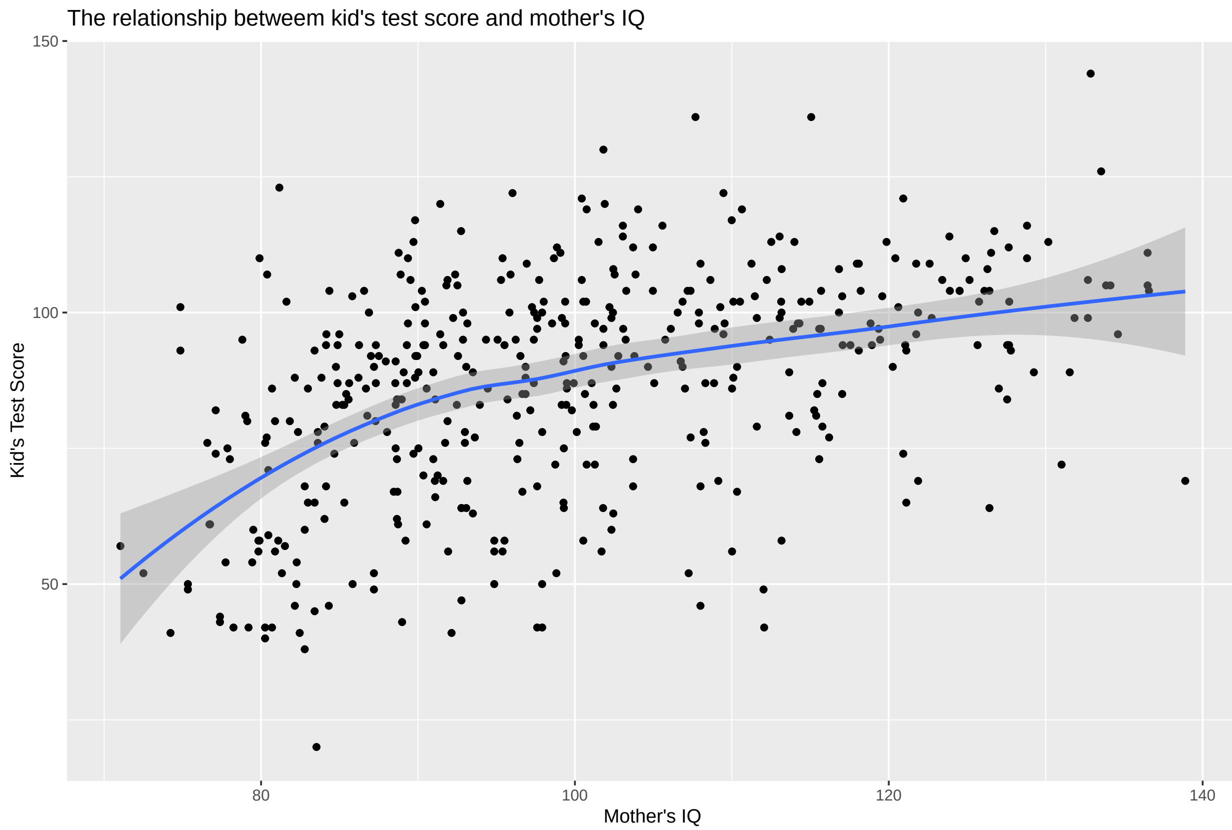
```r
ggplot(kidiq, aes(x=kid_score, color = mom_hs, fill=mom_hs)) +
  geom_density(alpha=0.3) +
  labs(x = "Kid's Score", y="Density", title = "The distribution of kid's test score by mother's high s
```



2 - The relationship between the kid's test score and the mother's IQ

The type of the plot is a scatter plot with a smoothed curve. The purpose of this plot is to explore the relationship between the kid's test score and the mother's IQ. We can see that there is an increasing trend indicating that as the kid's test score is likely to increase with mother's IQ.

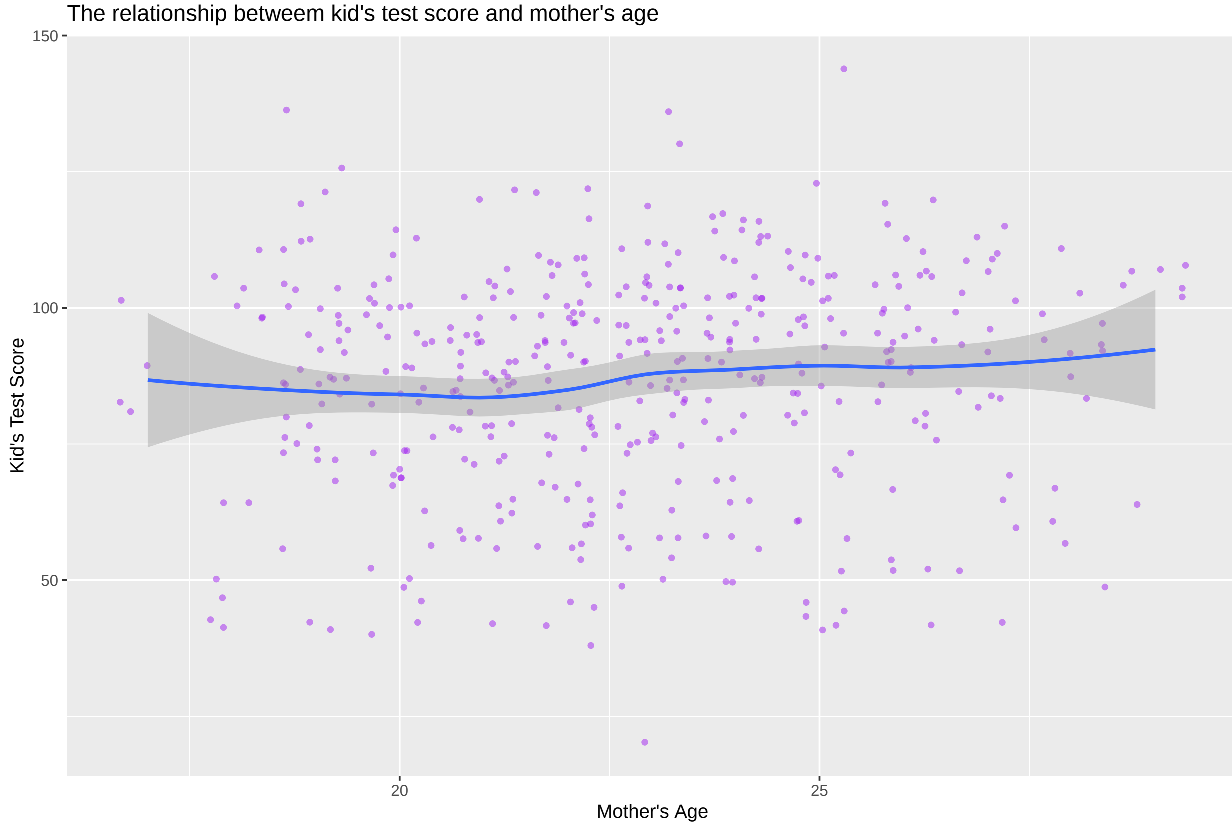
```r
ggplot(kidiq, aes(x=mom_iq, y=kid_score)) +
  geom_point() +
  geom_smooth() +
  labs(x = "Mother's IQ", y="Kid's Test Score", title = "The relationship betweem kid's test score and m
```



3 - The relationship between the kid's test score and the mother's age

The type of the plot is a scatter plot with a smoothed curve. The purpose of this plot is to explore the relationship between the kid's test score and the mother's Age. I use the jittered plot here for the points to better capture the distribution of the points and reduce overlapping of the points. We can see that there is no clear inidation that these two varibles are related. The distribution of the kid's test score is relatively random across mother's age.

```
ggplot(kidiq, aes(x=mom_age, y=kid_score)) +
  geom_point(position = "jitter", alpha=0.5, shape = 16, color="purple") +
  geom_smooth() +
  labs(x = "Mother's Age", y="Kid's Test Score", title = "The relationship betweem kid's test score and
```



Estimating mean, no covariates

In class we were trying to estimate the mean and standard deviation of the kid's test scores. The `kids2.stan` file contains a Stan model to do this. If you look at it, you will notice the first `data` chunk lists some inputs that we have to define: the outcome variable `y`, number of observations `N`, and the mean and standard deviation of the prior on `mu`. Let's define all these values in a `data` list.

```
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 = here("Lab_5/kids2.stan"),
            data = data,
            chains = 3,
            iter = 500)
```

Here is a summary of the fit:

```
summary(fit)$summary
```

```
##              mean    se_mean        sd        2.5%         25%         50%
## mu       86.73327 0.04362959 1.0261262    84.64654    86.05123    86.74434
## sigma    20.43631 0.03256942 0.6602705    19.32290    19.96468    20.39760
## lp__  -1525.76139 0.05935378 1.1291114 -1528.98517 -1526.07011 -1525.39849
##               75%       97.5%    n_eff      Rhat
## mu       87.42196    88.75577 553.1454 0.9991073
## sigma    20.81975    21.78554 410.9827 1.0014728
## lp__  -1525.01492 -1524.78582 361.8902 0.9988979
```

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

Here we change the prior to be more informative by changing the standard deviation to be 0.1 and rerun the model:

```r
# Set up the parameter
y <- kidiq$kid_score
mu0 <- 80

# Change the standard deviation to be 0.1
sigma1 <- 0.1

# named list to input for stan function
data1 <- list(y = y,
             N = length(y),
             mu0 = mu0,
             sigma0 = sigma1)

# Fit the model
fit1 <- stan(file = here("Lab_5/kids2.stan"),
            data = data1,
            chains = 3,
            iter = 500)
```

Here is a summary of this fit:

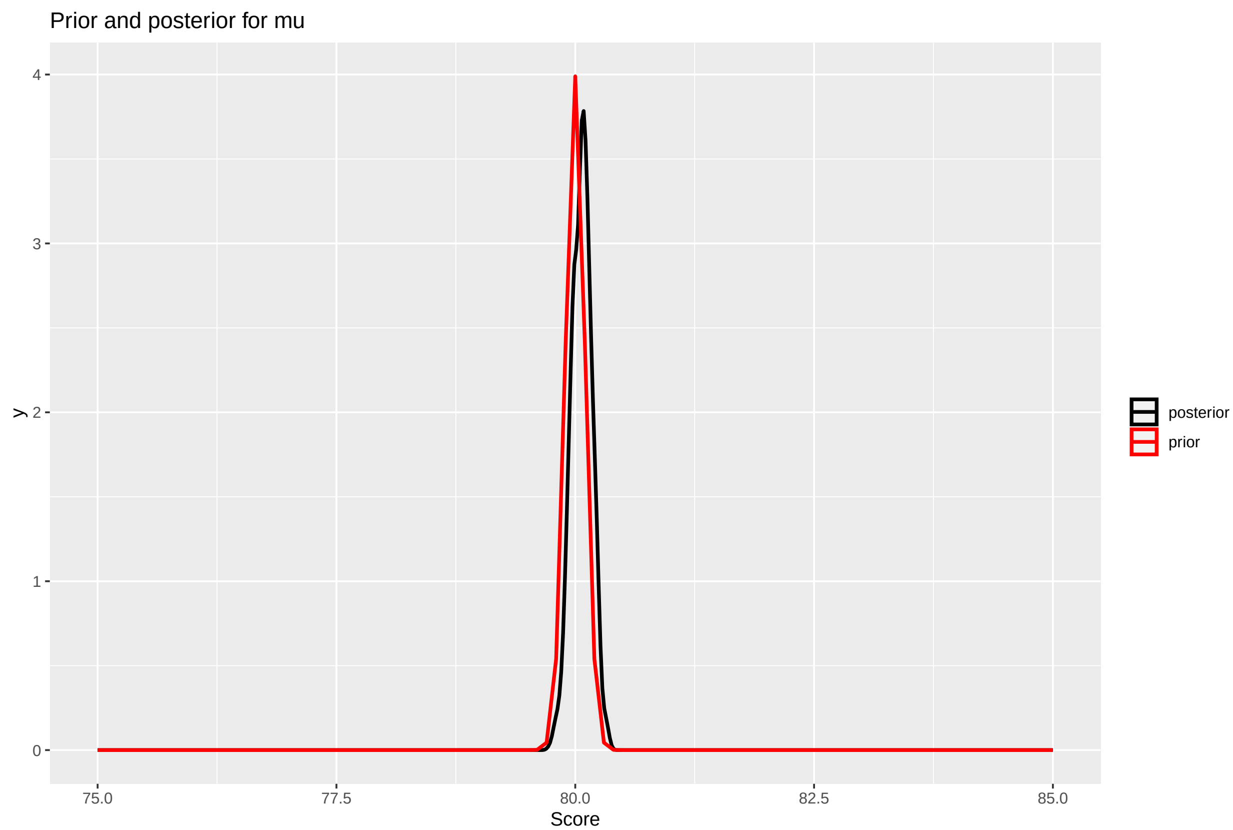
```r
summary(fit1)$summary
```

```
##              mean     se_mean        sd        2.5%         25%         50%
## mu       80.06523 0.004421338 0.1034365    79.86109    79.99238    80.06853
## sigma    21.41273 0.029555210 0.7126718    20.00502    20.93913    21.40734
## lp__  -1548.39264 0.053691827 1.0374908 -1551.26132 -1548.71956 -1548.10973
##               75%       97.5%    n_eff     Rhat
## mu       80.13738    80.25393 547.3188 1.001870
## sigma    21.89598    22.86060 581.4483 1.001244
## lp__  -1547.66259 -1547.38327 373.3809 1.001963
```

Based on the estimation results, we can see that the estimate changed but not in a very big scale compared to the first fit.

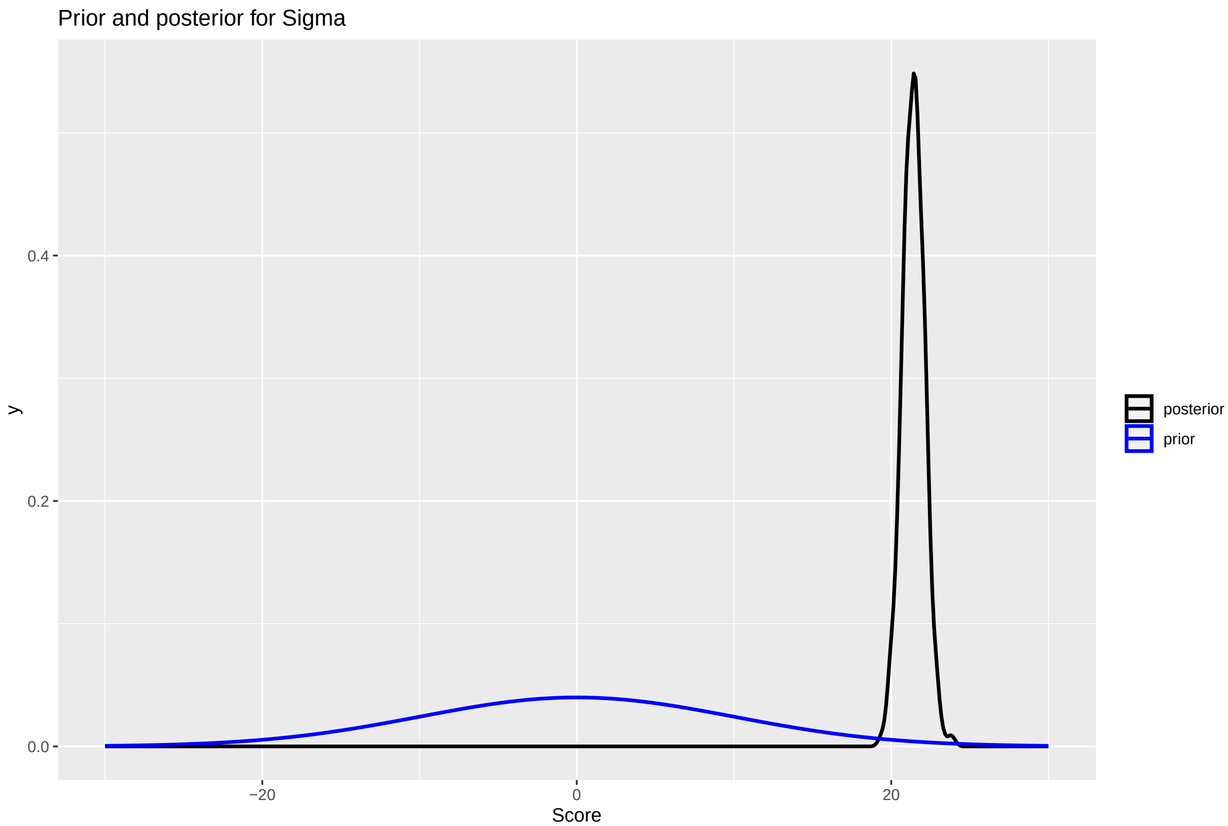
Now we move on and plot the prior and posterior densities for both mu and sigma:

```r
# Recall the parameter
y <- kidiq$kid_score
mu0 <- 80

# Change the standard deviation to be 0.1
sigma1 <- 0.1

# Retrive the data
dsamples1 <- fit1   %>%
  gather_draws(mu, sigma) # gather = long format

# Plot for mu
dsamples1  %>%
  filter(.variable == "mu")  %>%
  ggplot(aes(.value, color = "posterior")) + geom_density(size = 1) +
  xlim(c(75, 85)) +
  stat_function(fun = dnorm,
        args = list(mean = mu0,
                    sd = sigma1),
        aes(colour = 'prior'), size = 1) +
  scale_color_manual(name = "", values = c("prior" = "red", "posterior" = "black")) +
  ggtitle("Prior and posterior for mu") +
  xlab("Score")
```




```r
# Plot for sigma
dsamples1  %>%
  filter(.variable == "sigma")  %>%
```

```
ggplot(aes(.value, color = "posterior")) + geom_density(size = 1) +
xlim(c(-30, 30)) +
stat_function(fun = dnorm,
      args = list(mean = 0,
                  sd = 10),
      aes(colour = 'prior'), size = 1) +
scale_color_manual(name = "", values = c("prior" = "blue", "posterior" = "black")) +
ggtitle("Prior and posterior for Sigma") +
xlab("Score")
```



Comments:

Here we can see that with a very informative prior, the posterior distribution changed a lot in terms of shape and spread for both mu and sigma. For mu specifically, the prior and posterior looks similar as the prior is too informative and it dominates the results. Similarly for sigma, even though the informative prior is for mu, the posterior distribution for sigma is still impacted. In terms of the mean, the result for mu and sigma do not change too much compared to the previous fit.

Adding covariates

Now let's see how kid's test scores are related to mother's education. We want to run the simple linear regression

$$Score = \alpha + \beta X$$

where $X = 1$ if the mother finished high school and zero otherwise.

kid3.stan has the stan model to do this. Notice now we have some inputs related to the design matrix X and the number of covariates (in this case, it's just 1).

Let's get the data we need and run the model.

```
kidiq <- read_rds("kidiq.RDS")
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("Lab_5/kids3.stan"),
            data = data,
            iter = 1000)
```

Question 3

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

Answer

Here we run the lm model using the same data and compare the results for the coefficient estimates:

```
# Stan result
summary(fit2)$summary
```

```
##                 mean     se_mean       sd         2.5%          25%         50%
## alpha       78.07571 0.08040835 2.095272    74.132558    76.622385    78.02926
## beta[1]     11.11214 0.09533244 2.387059     6.274035     9.519877    11.15948
## sigma       19.84833 0.02163771 0.660732    18.550389    19.405386    19.85218
## lp__     -1514.46707 0.05643531 1.323446 -1517.926726 -1515.026709 -1514.10852
##                 75%       97.5%    n_eff      Rhat
## alpha       79.49658    82.27625 679.0134 1.002136
## beta[1]     12.78793    15.57975 626.9670 1.002460
## sigma       20.27297    21.16540 932.4558 1.000881
## lp__     -1513.51075 -1512.99085 549.9343 1.007059
```

```
# lm model result
model2 <- lm(kidiq$kid_score ~ kidiq$mom_hs)
summary(model2)
```

```
##
## Call:
## lm(formula = kidiq$kid_score ~ kidiq$mom_hs)
##
## 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 ***
## kidiq$mom_hs1    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
```

```
# Compare intercept and slope
# Stan
summary(fit2)$summary[1:2,1]
```

```
##    alpha  beta[1]
## 78.07571 11.11214
```

```
# lm
summary(model2)$coefficients[,"Estimate"]
```

```
##   (Intercept) kidiq$mom_hs1
##      77.54839      11.77126
```

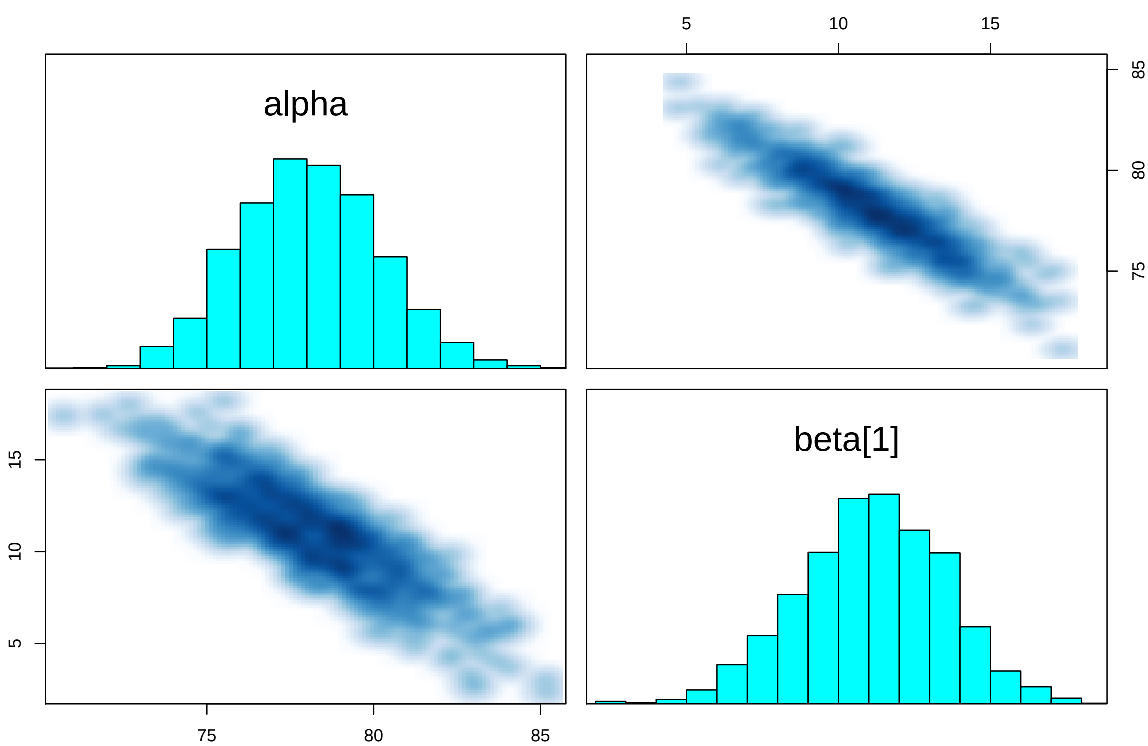
Based on the above results, we can see that the coefficient estimates are close between the two methods. Thus we conclude that the estimates of the intercept and slope from the stan model are comparable to results from `lm()`.

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?

Answer

Here is the pairs plot:

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



Comment:

Based on the above pair plot, we can see that when the slope gets larger, the intercepts gets smaller. The plot shows a strong negative relationship between the intercept and slope. The correlation between the intercept and the slope should close to -1. A small change in slope can change the intercept in the same scale. This is a potential problem as it will make it harder to interpret the intercept and make it harder to sample. People may consider centering to solve this problem.

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

```
# Create a new variable of IQ for mean centering
kidiq$mom_iq_meanadj <- kidiq$mom_iq - mean(kidiq$mom_iq)

# Generate data for the model
X <- cbind(as.matrix(kidiq$mom_hs, ncol = 1),
           as.matrix(kidiq$mom_iq_meanadj, ncol = 1)) # force this to be a matrix
K <- 2
data3 <- list(y = y, N = length(y),
             X =X, K = K)
```

```
# Fit the model
fit3 <- stan(file = here("Lab_5/kids3.stan"),
            data = data3,
            iter = 1000)
```

Here is a summary of the model:

```
summary(fit3)$summary
```

```
##                 mean     se_mean         sd          2.5%          25%
## alpha      82.310493 0.056369698 1.88467807    78.5974533    81.053852
## beta[1]     5.682124 0.064598716 2.14711813     1.6331092     4.115403
## beta[2]     0.569271 0.001661967 0.06121331     0.4470788     0.527536
## sigma      18.103729 0.015359981 0.62794679    16.8966596    17.679418
## lp__    -1474.473982 0.048806654 1.39950948 -1477.8465074 -1475.177502
##                  50%           75%         97.5%     n_eff      Rhat
## alpha      82.2970434    83.6023228    85.8720039 1117.8485 1.0004424
## beta[1]     5.7225128     7.1922346    10.0519580 1104.7503 1.0004640
## beta[2]     0.5704792     0.6122566     0.6875759 1356.5842 0.9998079
## sigma      18.0870380    18.5086701    19.4128825 1671.3381 0.9996980
## lp__    -1474.1852782 -1473.4053246 -1472.6776623  822.2306 1.0041006
```

Interpretation:

The coefficient of the mean centered IQ is 0.57. This means that with all other variables being the same (the high school status remain unchanged) if the mum's IQ is one unit higher than the mean IQ, the kid's test score will increase by 0.57 points compared to the base test score. Similarly, if the mum's IQ is one unit lower than the mean IQ, the kid's test score will decrease by 0.57 points compared to the base test score. For each unit of increase in the mum's IQ, the kids test score will show a 0.57 points increase in the test score. The intercept(alpha) represents the base test score that a kid will have with no high school mom and a mean IQ.

Question 5

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

Answer

Here we run the lm model using the same data and compare the results for the coefficient estimates:

```
# Stan result
summary(fit3)$summary
```

```
##                  mean     se_mean         sd          2.5%          25%
## alpha       82.310493 0.056369698 1.88467807    78.5974533    81.053852
## beta[1]      5.682124 0.064598716 2.14711813     1.6331092     4.115403
## beta[2]      0.569271 0.001661967 0.06121331     0.4470788     0.527536
## sigma       18.103729 0.015359981 0.62794679    16.8966596    17.679418
## lp__     -1474.473982 0.048806654 1.39950948 -1477.8465074 -1475.177502
##                   50%           75%         97.5%     n_eff      Rhat
## alpha       82.2970434    83.6023228    85.8720039 1117.8485 1.0004424
## beta[1]      5.7225128     7.1922346    10.0519580 1104.7503 1.0004640
## beta[2]      0.5704792     0.6122566     0.6875759 1356.5842 0.9998079
## sigma       18.0870380    18.5086701    19.4128825 1671.3381 0.9996980
## lp__     -1474.1852782 -1473.4053246 -1472.6776623  822.2306 1.0041006
```

```
# lm model result
model3 <- lm(kidiq$kid_score ~ kidiq$mom_hs + kidiq$mom_iq_meanadj)
summary(model3)
```

```
##
## Call:
## lm(formula = kidiq$kid_score ~ kidiq$mom_hs + kidiq$mom_iq_meanadj)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -52.873 -12.663   2.404  11.356  49.545
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)            82.12214    1.94370  42.250  < 2e-16 ***
## kidiq$mom_hs1           5.95012    2.21181   2.690  0.00742 **
## kidiq$mom_iq_meanadj   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
```

```
# Compare intercept and slope
# Stan
summary(fit3)$summary[1:3,1]
```

```
##     alpha   beta[1]   beta[2]
## 82.310493  5.682124  0.569271
```

```
# lm
summary(model3)$coefficients[,"Estimate"]
```

```
##          (Intercept)        kidiq$mom_hs1 kidiq$mom_iq_meanadj
```

```
##              82.122143              5.950117              0.563906
```

Based on the above results, we can see that the coefficient estimates are close and the standard error are similar as well. Thus the lm model results confirm what we get using stan model.

Question 6

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

Answer

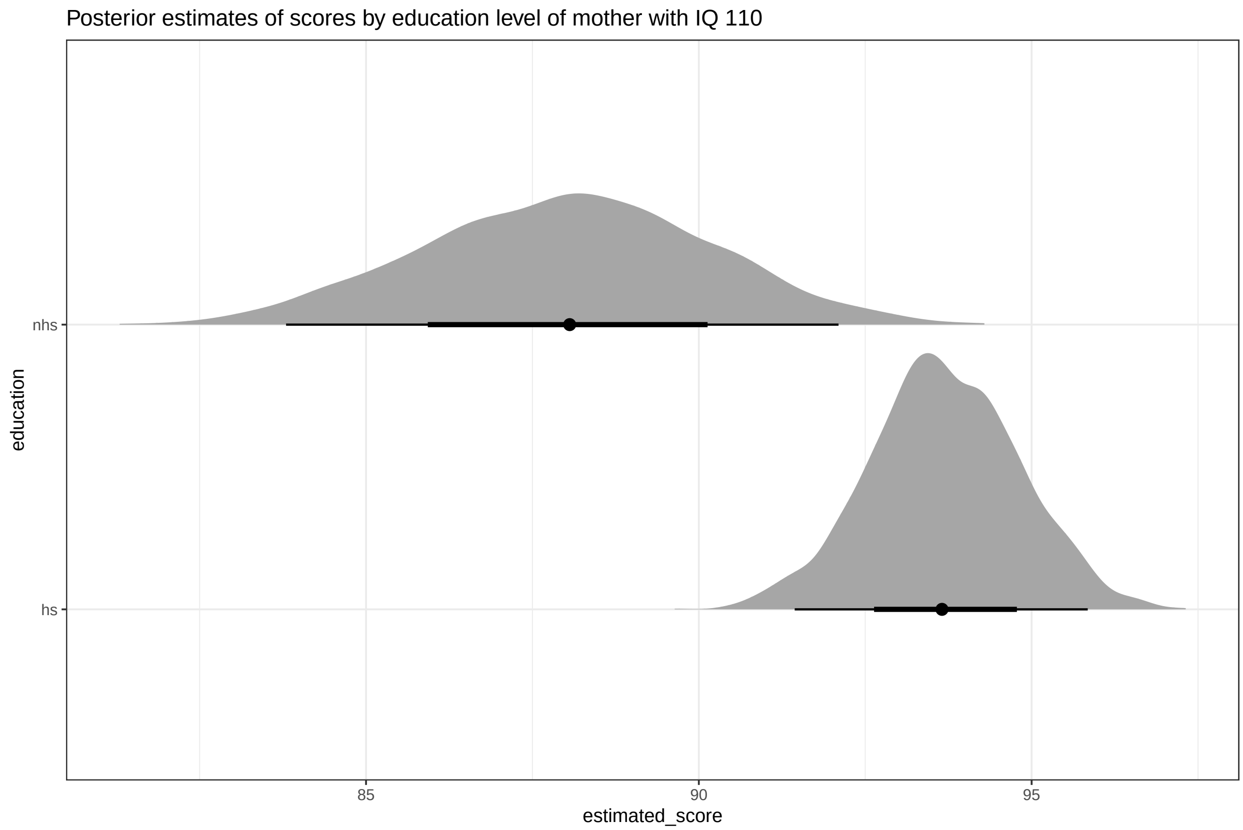
First we find out the input for the mother IQ.

```
110-mean(kidiq$mom_iq)
```

```
## [1] 10
```

Since we use the mean of the IQ score in the model and we have the IQ of 110 is 10 unit above the mean. Thus we need to adjust it for both high school and non-high school mother. Here is the process for plotting:

```
# Plot the posterior estimates of scores by education for mothers who have an IQ of 110
fit3 %>%
  spread_draws(alpha, beta[k], sigma)  %>%
  pivot_wider(names_from = k, names_prefix = "beta", values_from = beta)   %>%
  mutate(nhs = alpha + beta2 * 10, hs = alpha + beta1 + beta2 * 10)  %>%
  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 with IQ 110")
```



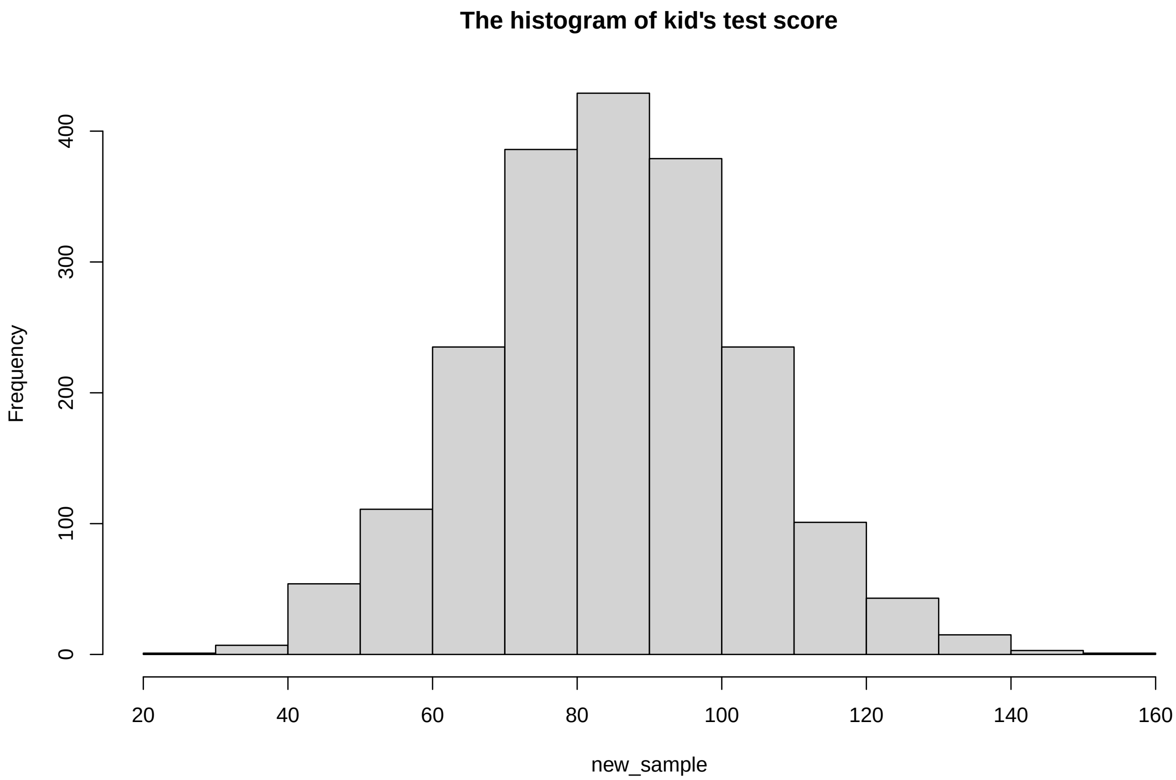
Question 7

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.

Answer

Here we generate the samples and plot the result in the form of histogram.

```
# New value for IQ
x_new <- 95- mean(kidiq$mom_iq)

# Estimated Parameter
post_samples3 <- extract(fit3)
alpha <- post_samples3$alpha
beta1 <- post_samples3$beta[,1]
beta2 <- post_samples3$beta[,2]
sigma <- post_samples3$sigma

# Point Estimation
lin_pred <- alpha + beta1 + beta2 *x_new

# Sampling
new_sample <- rnorm(length(sigma), mean = lin_pred, sd = sigma)

# Histogram
hist(new_sample, main="The histogram of kid's test score")
```


The histogram of kid's test score
Frequency
0
100
200
300
400
20
40
60
80
100
120
140
160
new_sample