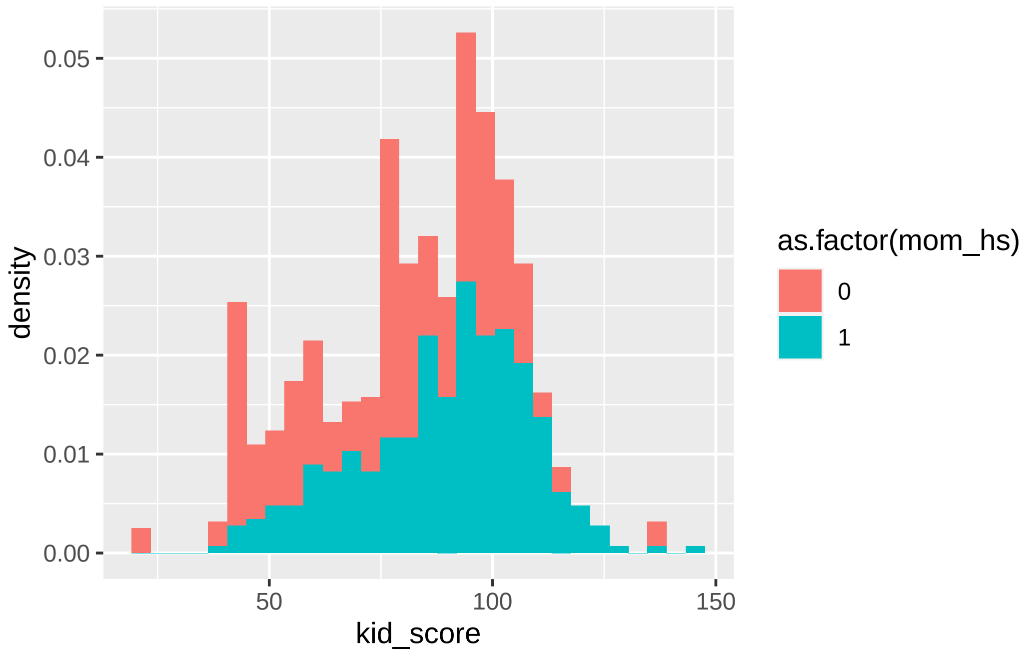
lab5

```r
library(tidyverse)
library(rstan)
library(tidybayes)
library(here)
```

Question 1

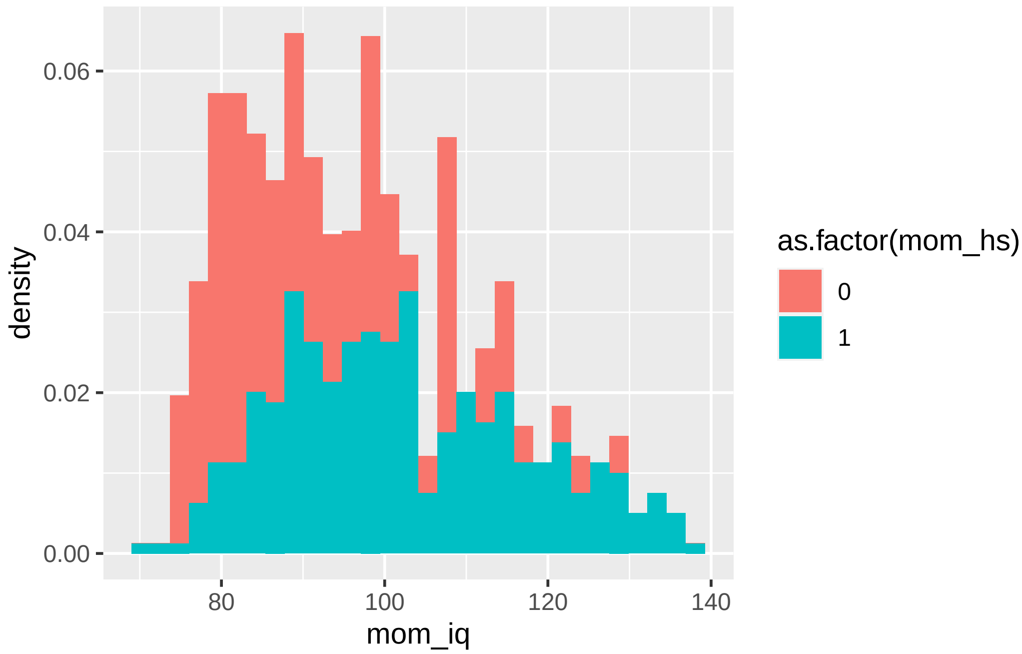
The first plot shows a histogram of the test scores filled with red if the mom completed high school and with blue if the mom did not. For the most part, the proportion of moms completing high school did not change much with test scores. However, for high test scores, the majority of moms completed high school.

```r
kidiq <- read_rds(here("kidiq.RDS"))
ggplot(data=kidiq) +
  geom_histogram(aes(x = kid_score, y = ..density..,fill=as.factor(mom_hs)))
```



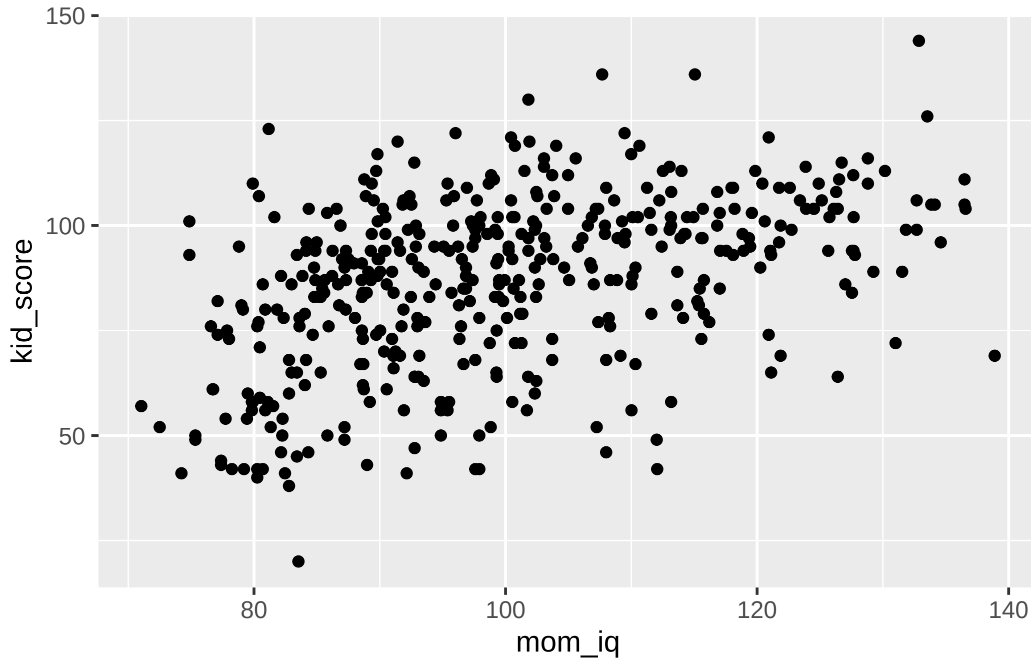
The next graph shows a histogram of the moms' IQ scores filled with red if the mom completed high school and with blue if the mom did not. Here, we can see that there is a higher proportion of high school completion for moms with higher IQ scores.

```
ggplot(data=kidiq) +
  geom_histogram(aes(x = mom_iq, y = ..density..,fill=as.factor(mom_hs)))
```



The following plot shows a scatterplot for the test score and mom's IQ variables. It shows that there may be a relationship between the variables where test scores increase as moms' IQ increase.

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



Question 2

```
y <- kidiq$kid_score
mu0 <- 80
sigma0 <- 10
sigma1=0.1

# named list to input for stan function
data <- list(y = y,
             N = length(y),
             mu0 = mu0,
             sigma0 = sigma0)
fit <- stan(file = here("kids2.stan"),
            data = data,
            chains = 3,
            iter = 500)
data2=list(y = y,
             N = length(y),
             mu0 = mu0,
             sigma0 = sigma1)
fitb=stan(file = here("kids2.stan"),
            data = data2,
```

```r
             chains = 3,
             iter = 500)
```

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

```
             mean    se_mean        sd        2.5%         25%         50%
mu       86.76921 0.03400568 0.9802036    84.98299    86.03436    86.72623
sigma    20.38740 0.04549366 0.7254108    19.14264    19.85033    20.36109
lp__  -1525.82034 0.06481364 1.0735399 -1529.12517 -1526.13177 -1525.51008
              75%       97.5%    n_eff      Rhat
mu       87.46399    88.67849 830.8636 0.9969309
sigma    20.83472    21.81735 254.2531 1.0044391
lp__  -1525.08601 -1524.78598 274.3492 1.0076055
```

```r
summary(fitb)$summary
```

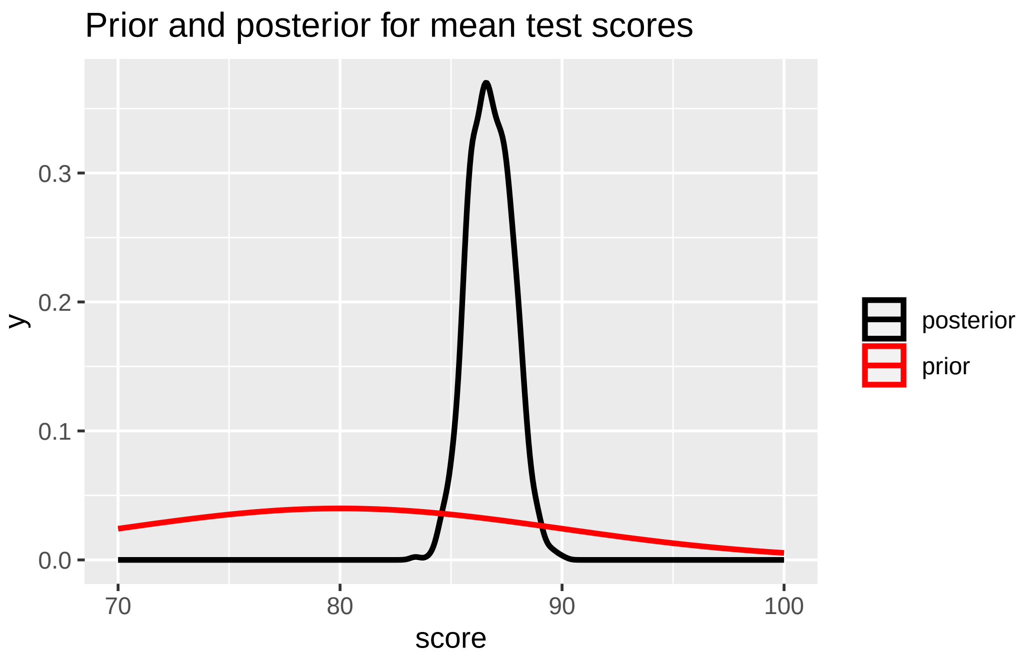
```
             mean     se_mean        sd        2.5%         25%         50%
mu       80.06702 0.004353507 0.1031311    79.86544    79.99923    80.06940
sigma    21.44652 0.031278861 0.7270843    20.10146    20.89449    21.44089
lp__  -1548.40783 0.048695794 0.9603889 -1551.17045 -1548.82372 -1548.14118
              75%       97.5%    n_eff      Rhat
mu       80.13763    80.26695 561.1784 1.0012787
sigma    21.94170    22.88597 540.3408 0.9982958
lp__  -1547.68447 -1547.39120 388.9658 1.0091044
```

From the summaries of the fits, we can see that with the more informative prior, the mu estimate decreased to be closer to the mu0 value 80. The standard error of this estimate also decreased. The estimate for sigma did not change much however.

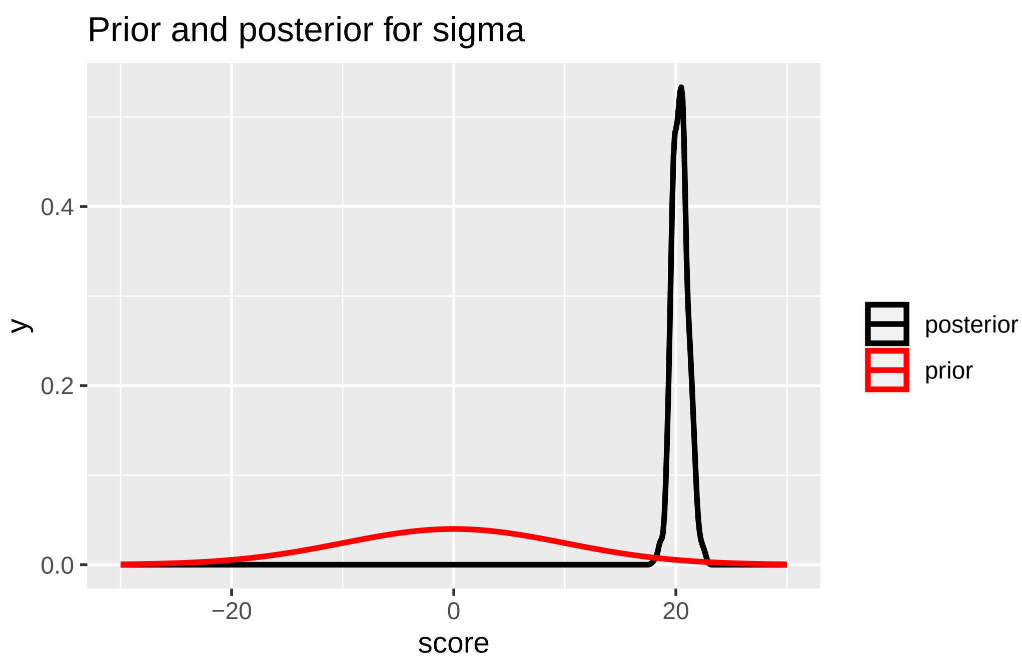
The following plots show the prior and posterior densities for the mean test scores and sigma.

```r
dsamples <- fit  |>
  gather_draws(mu, sigma) # gather = long format
dsamples |>
  filter(.variable == "mu") |>
  ggplot(aes(.value, color = "posterior")) + geom_density(size = 1) +
  xlim(c(70, 100)) +
  stat_function(fun = dnorm,
                args = list(mean = mu0,
```

```
                    sd = sigma0),
              aes(colour = 'prior'), size = 1) +
  scale_color_manual(name = "",
  values = c("prior" = "red", "posterior" = "black")) +
  ggtitle("Prior and posterior for mean test scores") +
  xlab("score")
```




```
dsamples |>
  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" = "red", "posterior" = "black")) +
  ggtitle("Prior and posterior for sigma") +
  xlab("score")
```



Question 3

a)

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

```
summary(fit2)$summary
```

```
               mean    se_mean        sd        2.5%         25%         50%
alpha      78.02433 0.07432390 2.0439261   74.003862   76.631508    77.98186
beta[1]    11.18866 0.08439735 2.2883669    6.835207    9.623019    11.19421
sigma      19.81223 0.02120888 0.6807241   18.500672   19.362543    19.79395
lp__    -1514.40017 0.05278990 1.2868115 -1517.758794 -1514.924555 -1514.06885
                75%        97.5%      n_eff      Rhat
```

```
alpha      79.40725    81.98679  756.2641 1.004373
beta[1]    12.67876    15.80118  735.1800 1.005427
sigma      20.26716    21.17511 1030.1646 1.001713
lp__     -1513.47645 -1512.97611  594.1938 1.004480
```

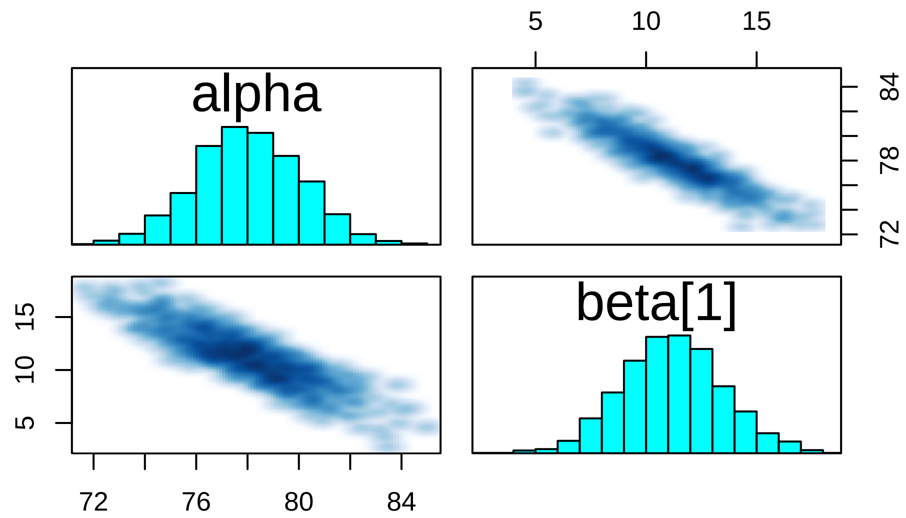
```r
summary(lm(kid_score~mom_hs,data=kidiq))
```

```
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
```

From the summaries of the fits above, we can see that the estimates of the intercept and slope are comparable.

b)

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



From the `pairs` plot, we can see that changes in the slope would induce the opposite change in the intercept, which would make it hard to interpret what the intercepts mean. The correlation makes it harder to sample.

Question 4

```
data <- list(y = y, N = length(y),
             X =cbind(as.matrix(kidiq$mom_hs),
                      as.matrix(kidiq$mom_iq - mean(kidiq$mom_iq))), K = 2)
fit3 <- stan(file = here("kids3.stan"),
             data = data,
             iter = 1000)
```

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

```
                mean      se_mean         sd          2.5%           25%
alpha     82.2498389 0.061093467 1.95311487    78.6228931    80.8914985
beta[1]    5.7765745 0.067849706 2.19900428     1.3365284     4.2665660
beta[2]    0.5632544 0.001626245 0.06030683     0.4456434     0.5218089
sigma     18.1320399 0.015914628 0.62378391    16.9660701    17.7101031
lp__   -1474.4448840 0.050257174 1.40365511 -1477.9338898 -1475.1305740
                 50%          75%        97.5%    n_eff      Rhat
alpha     82.1653674    83.635274   86.0021514 1022.036 1.0009676
beta[1]    5.7954874     7.326548    9.9903209 1050.404 1.0012308
```

```
beta[2]      0.5643056     0.602966     0.6823863 1375.186 0.9996500
sigma       18.1127938    18.547967    19.3988265 1536.298 0.9998132
lp__     -1474.1362346 -1473.407582 -1472.6434379  780.054 1.0021370
```

For this fit of the model, we get that for a given outcome of mother's high school completion, each IQ point above the mean IQ score of 100 is associated with a mean increase in test score by 0.56.

Question 5

```r
kidiq5=kidiq %>% mutate(z_mom_iq=mom_iq-mean(mom_iq))
summary(lm(kid_score~mom_hs+z_mom_iq,data=kidiq5))
```

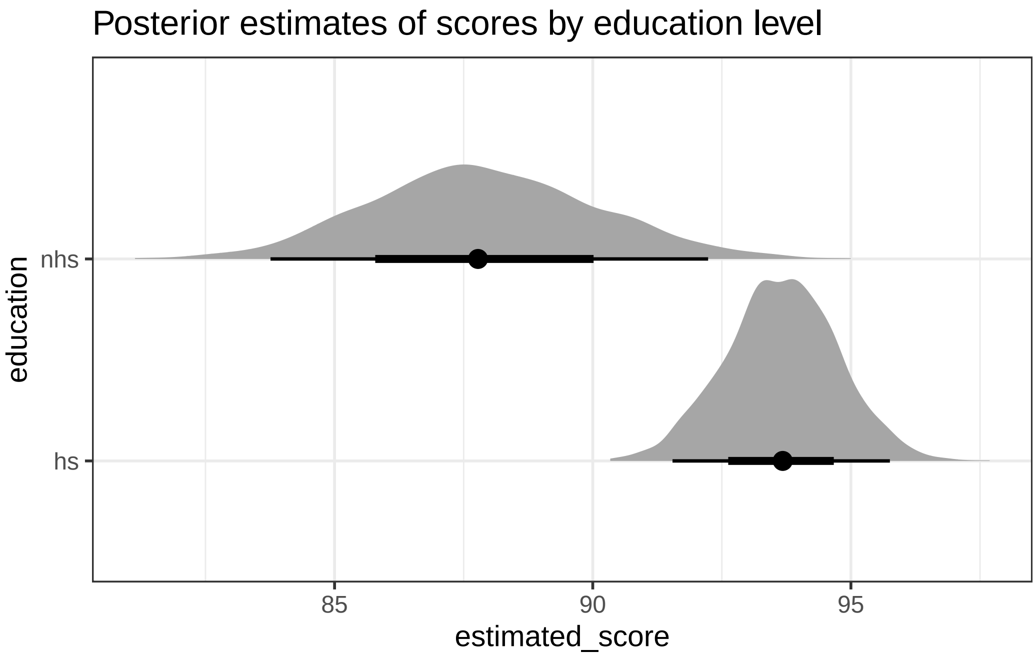
```
Call:
lm(formula = kid_score ~ mom_hs + z_mom_iq, data = kidiq5)

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 ** 
z_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
```

From these results, we can see that the estimates are similar to those obtained in question 4.

Question 6

The following plot shows the posterior estimates of scores by education of mother for mothers who have an IQ of 110.

```
post_samples=extract(fit3)
nhs=post_samples$alpha+10*post_samples$beta[,2]
hs=post_samples$alpha+post_samples$beta[,1]+10*post_samples$beta[,2]
data6=tibble(nhs,hs)
 data6|>
  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")
```



Question 7

The following histogram shows samples from the posterior predictive distribution for a new kid with a mother who graduated high school and has an IQ of 95.

```
sigma=post_samples$sigma
alpha=post_samples$alpha
beta1=post_samples$beta[,1]
beta2=post_samples$beta[,2]
```

```r
lin_pred=alpha+beta1-5*beta2
y_new <- rnorm(n = length(sigma),mean = lin_pred, sd = sigma)
hist(y_new)
```