Week 5 Lab

J. Arturo Esquivel

10/02/23

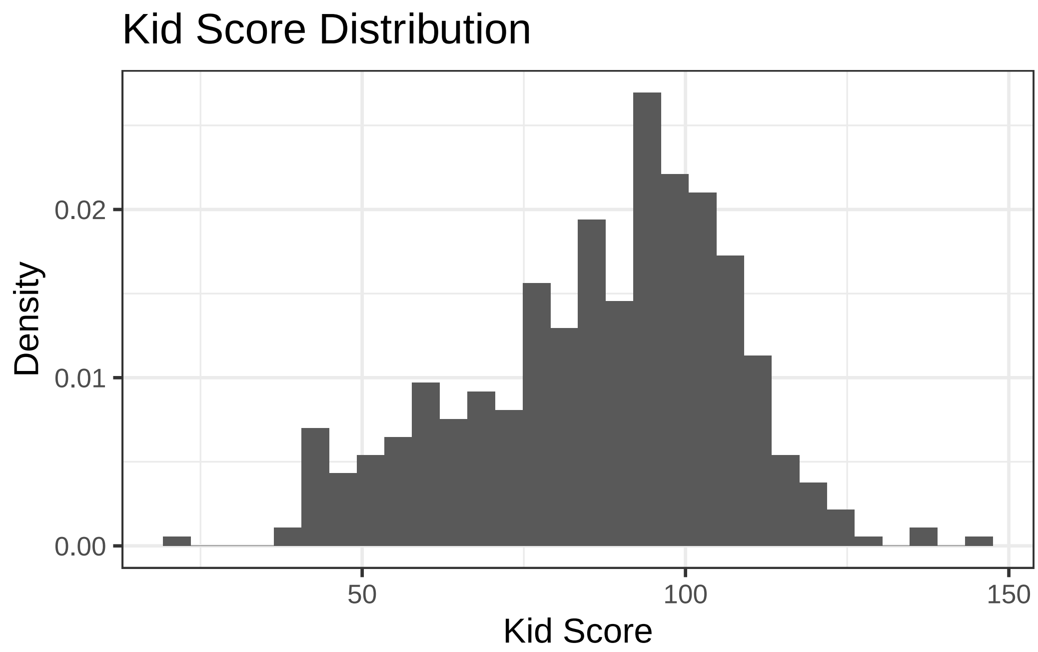
```
library(tidybayes)
library(tidyverse)
library(rstan)
library(here)
```

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

Question 1

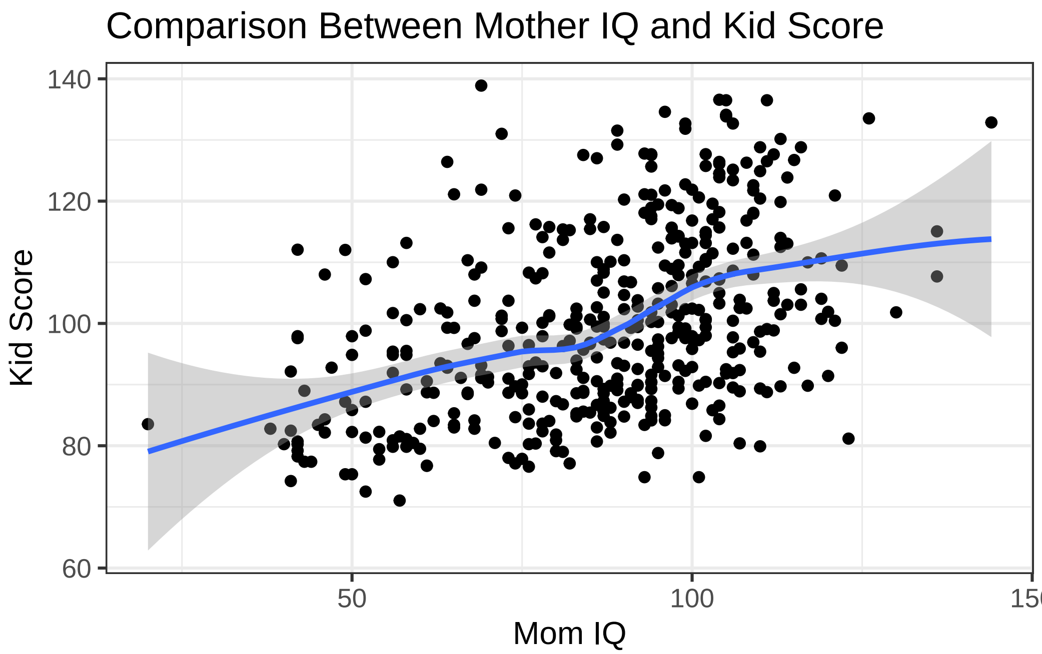
First we show the kids' scores distribution. They seem to approach a normal distribution, centered around a score of 95. There also seem to be a couple of extremely low observations, maybe due to measurement problems, or incomplete data.

```
ggplot(kidiq) +
  geom_histogram(aes(x = kid_score, y = after_stat(density)), position = 'dodge')  +
  labs(title = "Kid Score Distribution",
       x = "Kid Score",
       y = "Density") +
  theme_bw(base_size = 13)
```



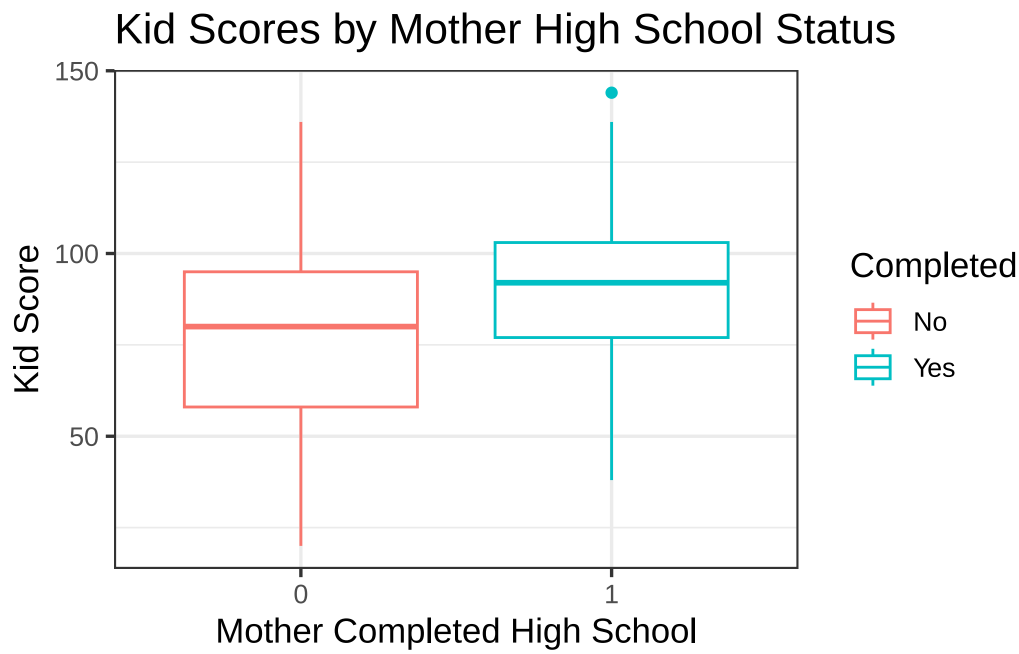
Then, we can inspect the relationship between kids' scores, and that of their mothers. There seems to be a slight direct relationship between the two.

```
kidiq |>
  ggplot(aes(kid_score, mom_iq)) +
  geom_point() +
  geom_smooth() +
  labs(title = "Comparison Between Mother IQ and Kid Score",
       x = "Mom IQ", y = "Kid Score") +
  theme_bw(base_size = 12)
```



Finally, we can assess how kids' IQ relates to High School completion by their mothers. The plot below shows that, overall, kids which mothers completed High School tend to have higher scores.

```
ggplot(kidiq, aes(as.factor(mom_hs), kid_score, color = as.factor(mom_hs))) +
  geom_boxplot() +
  labs(title = "Kid Scores by Mother High School Status",
       x = "Mother Completed High School", y = "Kid Score",
       color = "Completed") +
  theme_bw(base_size = 13)  +
  scale_color_discrete( labels = c('No', 'Yes'))
```




```
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("kids2.stan"),
            data = data,
            chains = 3,
            iter = 500,
            verbose = FALSE,
            refresh = 0)
```

Question 2

The results for such a model are shown below.

```r
sigma0 <- 0.1

data <- list(y = y,
             N = length(y),
             mu0 = mu0,
             sigma0 = sigma0)

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

```r
summary(fit)$summary[,1]
```

```
       mu       sigma        lp__
 86.72464    20.37545 -1525.72641
```

```r
summary(fit_i)$summary[,1]
```

```
       mu       sigma        lp__
 80.06334    21.40419 -1548.41887
```

The results do change considerably. Including a highly informative prior leads to a posterior mean virtually equal to the one specified in the prior. The posterior mean went from 86.7 when the prior was weakly informative, to 80.1 for the highly informative prior.

Looking at the prior and posterior distributions we can see that the posterior distribution remained pretty much the same as the prior specified.

```r
dsamples <- fit_i  |>
  gather_draws(mu, sigma)

dsamples |>
  filter(.variable == "mu") |>
  ggplot(aes(.value, color = "posterior")) + geom_density(size = 1) +
  xlim(c(75, 90)) +
  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")
```


```
X <- as.matrix(kidiq$mom_hs, ncol = 1)
K <- 1

data <- list(y = y, N = length(y),
             X = X, K = K)
fit2 <- stan(file = here("kids3.stan"),
            data = data,
            iter = 1000)
```

Question 3

a)

We can see that both, the intercept and $\hat{\beta}$ from the simple linear regression model are similar to the corresponding posterior means from the Bayesian model.

```r
fit_lm <- lm(kid_score ~ mom_hs, kidiq)

summary(fit2)$summary[,1]
```

```
    alpha     beta[1]       sigma         lp__
 77.96305    11.28886    19.82740 -1514.43481
```

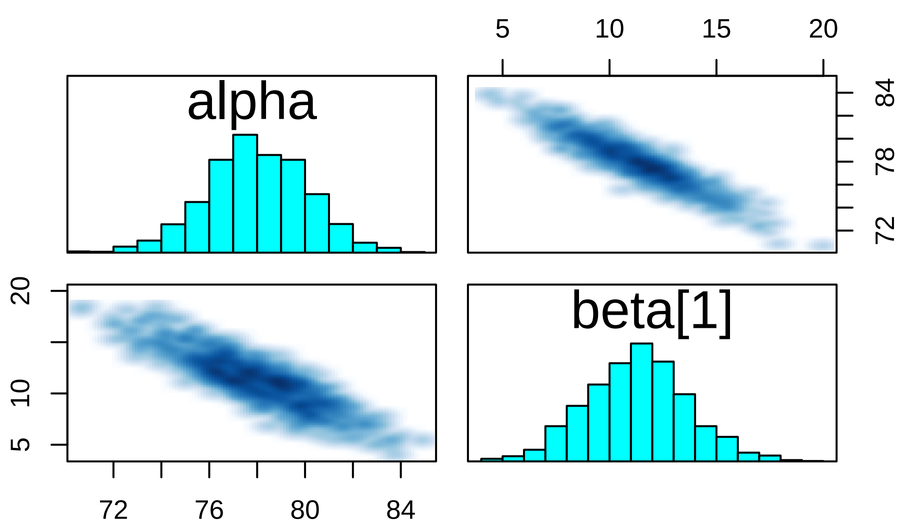
```r
fit_lm$coefficients
```

```
(Intercept)      mom_hs
   77.54839    11.77126
```

b)

The `pairs` plot is shown below. We see that the joint distribution for the parameters follows the line-like pattern common to linear regression. This can only be a problem in terms of efficiency, since exploration across the distribution of the parameters is limited to values along the line pattern.

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



Question 4

```
kidiq$mom_iq_c <- kidiq$mom_iq - mean(kidiq$mom_iq)
X <- as.matrix(cbind(kidiq$mom_hs, kidiq$mom_iq_c), ncol = 2)
K <- 2

data <- list(y = y, N = length(y),
             X = X, K = K)
fit3 <- stan(file = here("kids4.stan"),
            data = data,
            iter = 1000)
```

The results for such a model are below. The coefficient for the mother's IQ centered suggests that for every unit the IQ of the mother is above average, the kid's expected score increases in 0.56.

```
summary(fit3)$summary[,1]
```

```
      alpha        beta[1]        beta[2]          sigma            lp__
 82.2585578      5.7640324      0.5642052     18.1125205 -1474.4420703
```

Question 5

As the summary below shows, the results obtained are vary similar for both approaches.

```
fit_lm2 <- lm(kid_score ~ mom_hs + mom_iq_c, kidiq)
fit_lm2$coefficients
```

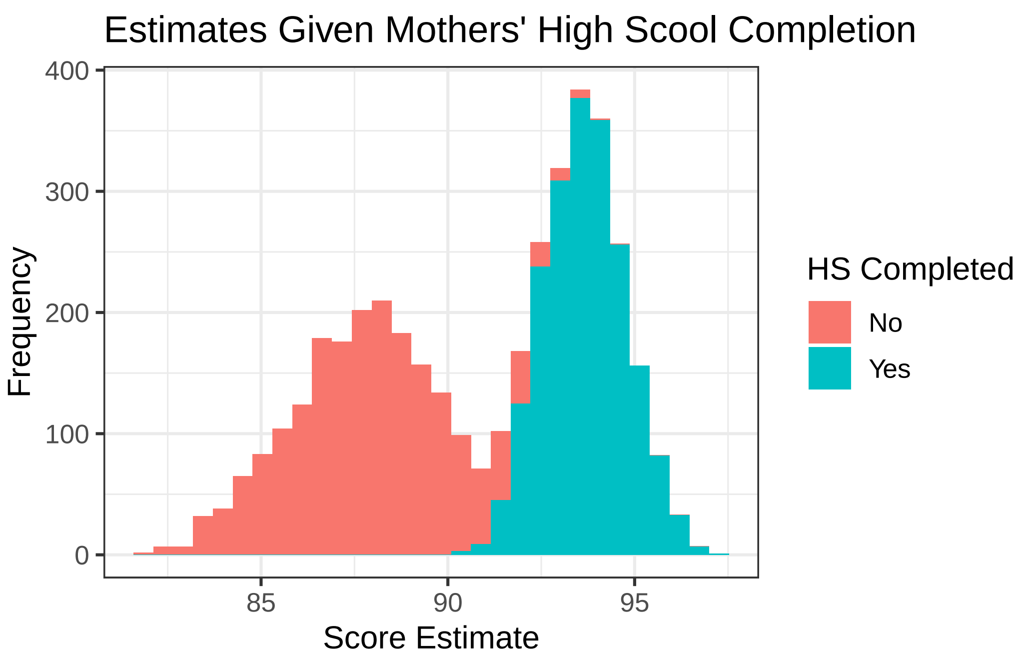
```
(Intercept)      mom_hs    mom_iq_c
  82.122143    5.950117    0.563906
```

Question 6

The plot of the posterior estimates given the education conditions of the mother are shown below.

```
estimate0 <- extract(fit3)[["alpha"]] + extract(fit3)[["beta"]][, 2]*(110-mean(kidiq$mom_i
estimate1 <- extract(fit3)[["alpha"]] + extract(fit3)[["beta"]][, 1] +
  extract(fit3)[["beta"]][, 2]*(110-mean(kidiq$mom_iq))

data.frame("No" = estimate0, "Yes" = estimate1) |>
  pivot_longer(c(No,Yes), names_to = "hs_completed") |>
  ggplot(aes(value, fill = hs_completed)) +
  geom_histogram()  +
  labs(title = "Estimates Given Mothers' High Scool Completion",
       x = "Score Estimate", y = "Frequency",
       fill = "HS Completed") +
  theme_bw(base_size = 12)
```



Question 7

The posterior predictive distribution for a new kid with a mother who graduated high school and has an IQ of 95 is shown below.

```
prediction <- extract(fit3)[["alpha"]] +
  extract(fit3)[["beta"]][, 1] +
```

```
  extract(fit3)[["beta"]][, 2]*(95-mean(kidiq$mom_iq)) +
  rnorm(length(extract(fit3)[["alpha"]]), 0, extract(fit3)[["sigma"]])

data.frame("Pred" = prediction) |>
  ggplot(aes(Pred)) +
  geom_histogram()  +
  labs(title = "Posterior Predictive Distribution",
       x = "Predicted Score", y = "Frequency") +
  theme_bw(base_size = 13)
```