Week 5: Bayesian linear regression and introduction to Stan

13/02/23

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).

```
knitr::opts_chunk$set(warning = FALSE,
                      echo = FALSE)
library(tidyverse)
library(rstan)
library(tidybayes)
library(here)
```

The data look like this:

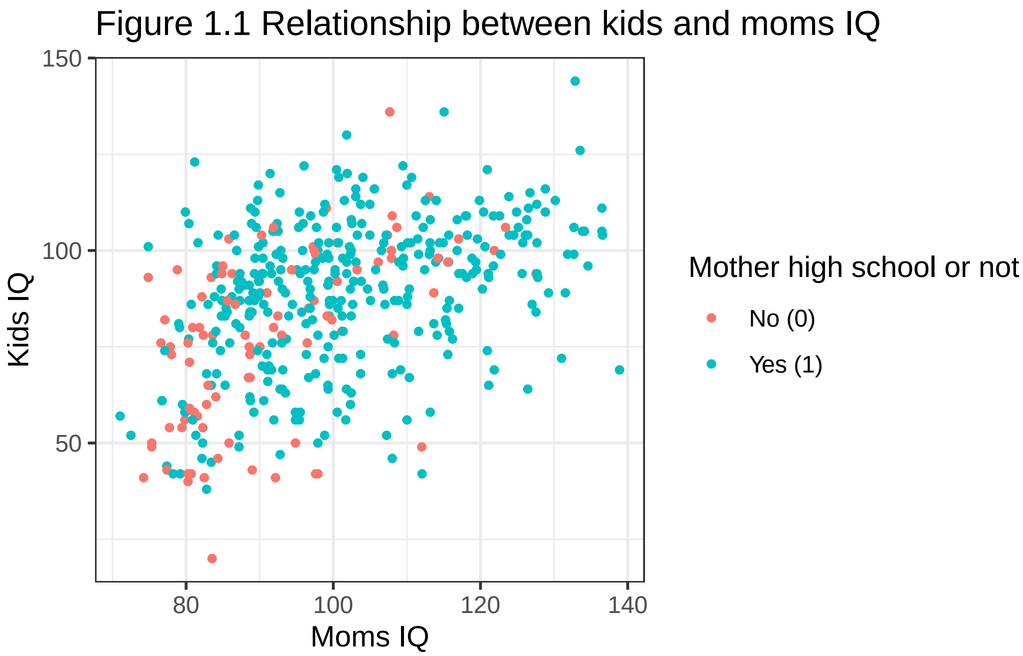
As well as the kid's test scores, we have a binary variable indicating whether or not the mother completed high school, the mother's IQ and age.

Descriptives

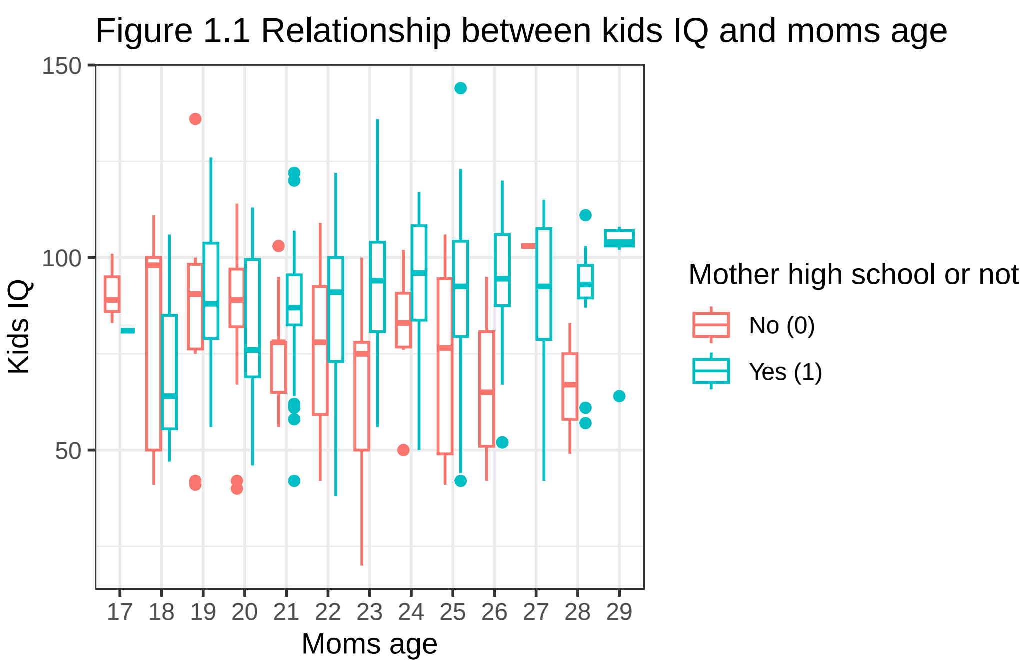
Question 1

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

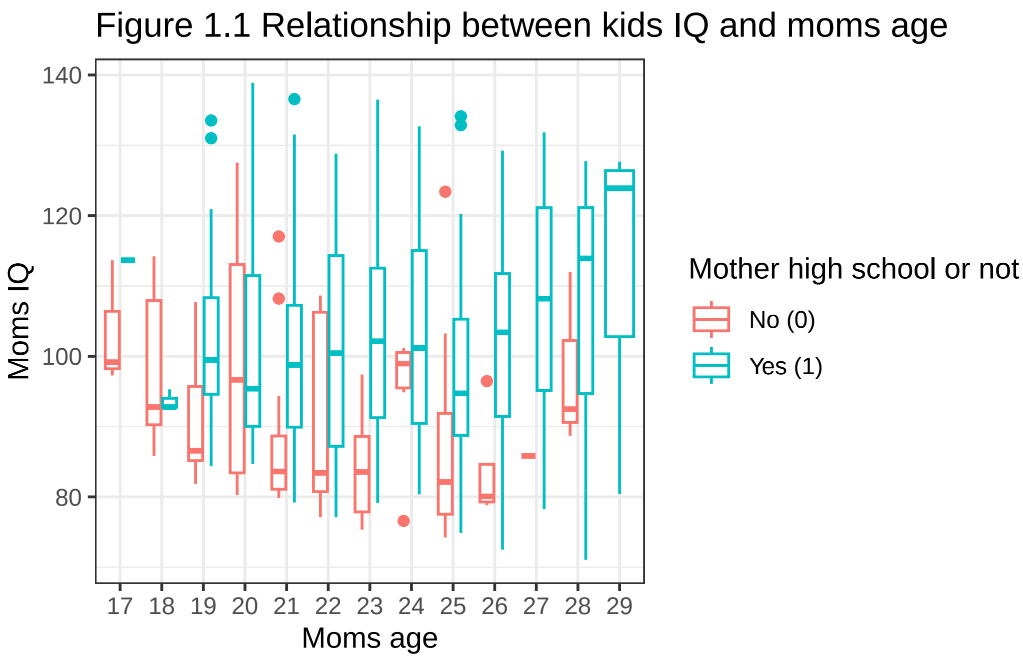
- Explain what your graph/ tables show
- Choose a graph type that's appropriate to the data type



- There appears to be a positive relationship between moms IQ and kids IQ. Also, mother with high school or above degree seems to be related to both higher moms IQ and kids IQ.



- As for the relationship between kids' IQ and mothers' age, overall the relationship is not clear, but it does appear that for mothers age 20 or less, not having a high school degree is positively correlated with higher kids IQ, while for mothers age 21 or more, the reverse is true. Of course, this is without observing the kids' age, which can be an impactful factor.



- There appears to be a positive relationship between moms IQ and they finished high school or not.

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.

- Previous fit:

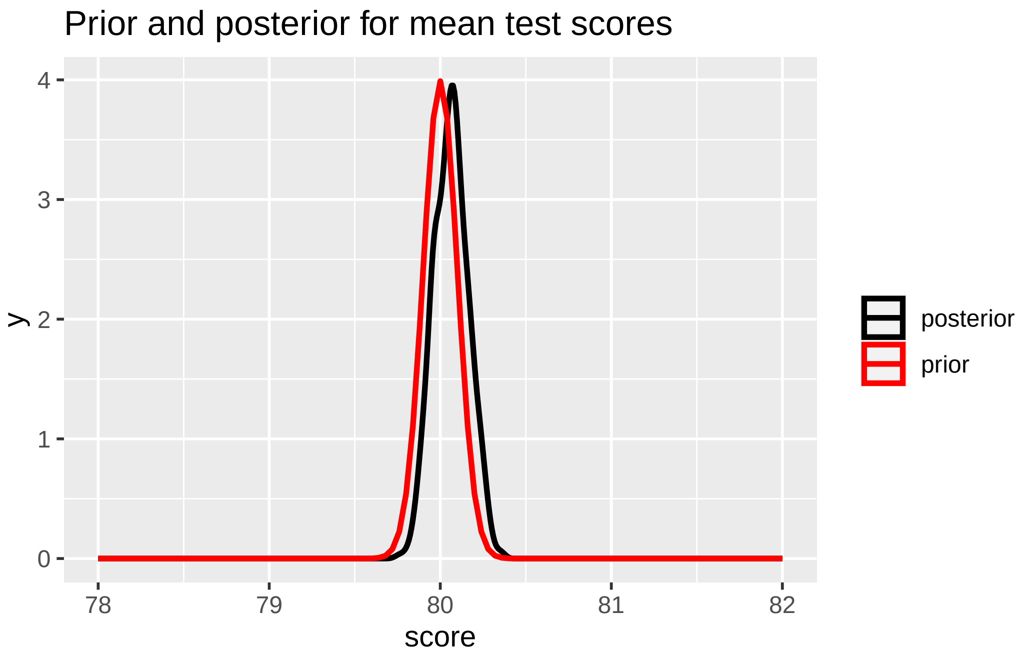
```
Inference for Stan model: anon_model.
3 chains, each with iter=500; warmup=250; thin=1;
post-warmup draws per chain=250, total post-warmup draws=750.

         mean se_mean   sd     2.5%      25%      50%      75%    97.5% n_eff
mu      86.75    0.04 1.02    84.79    86.03    86.80    87.48    88.75   708
sigma   20.38    0.03 0.71    19.09    19.89    20.35    20.86    21.86   761
```

```
lp__  -1525.84    0.05 1.04 -1528.43 -1526.37 -1525.52 -1525.06 -1524.78   392
      Rhat
mu       1
sigma    1
lp__     1

Samples were drawn using NUTS(diag_e) at Mon Feb 13 00:51:49 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).
```

- Current fit:

```
Inference for Stan model: anon_model.
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    79.87    79.99    80.06    80.13    80.26   802
sigma    21.35    0.03 0.71    20.01    20.86    21.34    21.82    22.82   600
lp__  -1548.37    0.08 0.95 -1550.96 -1548.74 -1548.09 -1547.67 -1547.40   157
      Rhat
mu    1.00
sigma 1.01
lp__  1.01

Samples were drawn using NUTS(diag_e) at Mon Feb 13 00:51:50 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).
```

- The estimate for mu is way closer to the prior mean 80 than the previous model (almost at 80 comparing to 86+). The estimate for sigma is also increased slightly by around 1.



Question 3

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

- Stan:

```
Inference for Stan model: anon_model.
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      78.06    0.07 1.93    74.45    76.76    78.01    79.33    81.98
beta[1]    11.14    0.08 2.18     6.56     9.75    11.15    12.60    15.24
sigma      19.80    0.02 0.66    18.61    19.34    19.78    20.24    21.14
lp__    -1514.29    0.04 1.16 -1517.29 -1514.83 -1513.99 -1513.43 -1512.97
        n_eff Rhat
alpha     864    1
beta[1]   801    1
sigma    1169    1
lp__      753    1
```

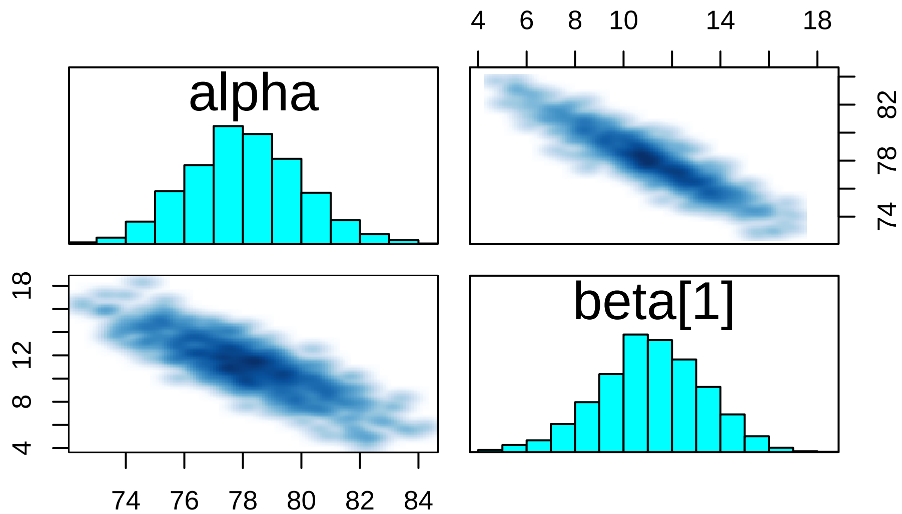
```
Samples were drawn using NUTS(diag_e) at Mon Feb 13 00:53:24 2023.
For each parameter, n_eff is a crude measure of effective sample size,
and Rhat is the potential scale reduction factor on split chains (at
convergence, Rhat=1).
```

- Lm:

```
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
```

- The estimates match fairly well for both intercept and slope.

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?



- The joint distribution has a much narrower when comparing to the previous one, potentially due to the mean of the variables are far away from the origin. This may make the algorithm harder to converge.

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: anon_model.
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      82.36    0.06 1.90    78.68    81.05    82.36    83.69    85.87
beta[1]     5.62    0.07 2.14     1.66     4.09     5.65     7.09     9.79
beta[2]     0.57    0.00 0.06     0.45     0.53     0.57     0.61     0.68
sigma      18.12    0.01 0.60    16.98    17.71    18.11    18.51    19.35
lp__    -1474.41    0.04 1.37 -1477.87 -1475.07 -1474.10 -1473.41 -1472.68
        n_eff Rhat
alpha     944 1.00
beta[1]  1022 1.00
```

```
beta[2]  1514 1.00
sigma    1632 1.00
lp__      988 1.01

Samples were drawn using NUTS(diag_e) at Mon Feb 13 00:53:26 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).
```

- for every 1 score increase in the mom's IQ, the kid's IQ is expected to increase by 0.57.

Question 5

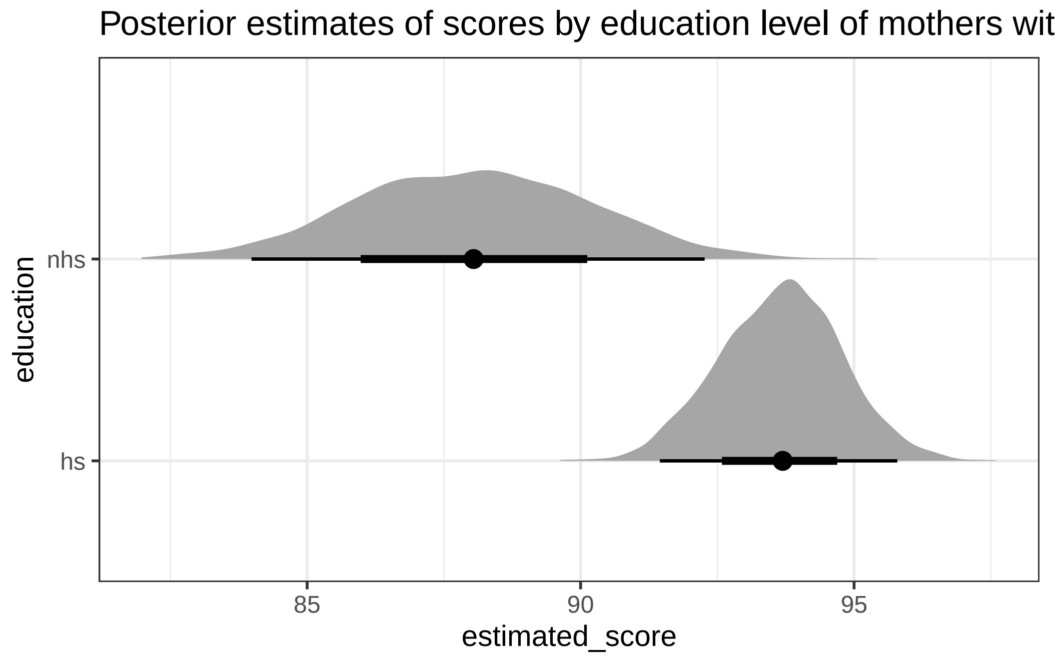
Confirm the results from Stan agree with `lm()`

```
Call:
lm(formula = kid_score ~ mom_hs + mom_iq, data = mutate(kidiq,
    mom_iq = mom_iq - mean(mom_iq)))

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       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
```

- The result indeed match for all beta hats.

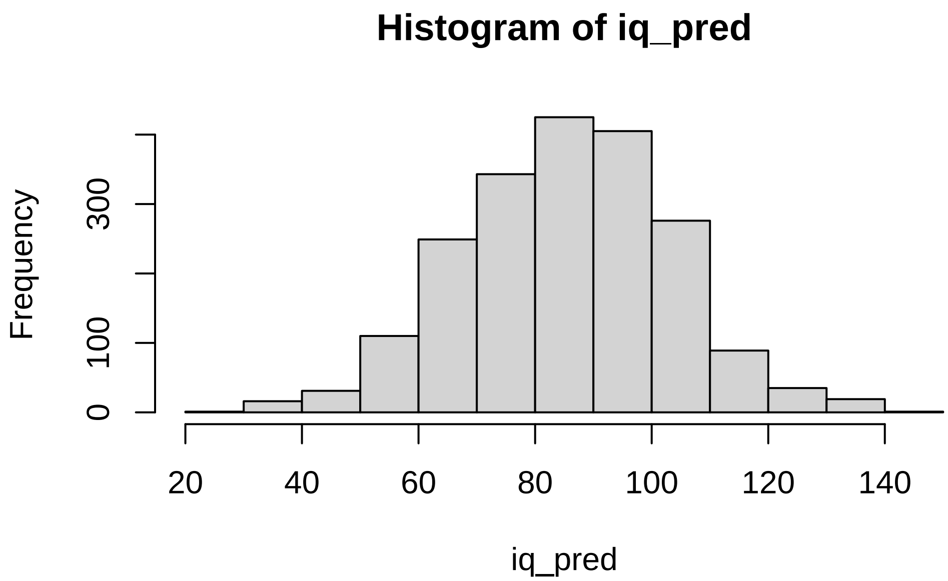
Question 6

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.


Histogram of iq_pred
Frequency
0
100
300
20
40
60
80
100
120
140
iq_pred