Methods 2 – Portfolio Assignment 3

- *Type:* Group assignment
- *Due:* 30 April 2023, 23:59
- *Instructions:* All problems are exercises from *Regression and Other Stories*. Please edit this file here and add your solutions.

1. Exercise 10.5

Regression modeling and prediction: The folder `KidIQ` contains a subset of the children and mother data discussed earlier in the chapter. You have access to children's test scores at age 3, mother's education, and the mother's age at the time she gave birth for a sample of 400 children.

```
#getwd()
data <- read.csv("data/child_iq.csv")
```

(a) Fit a regression of child test scores on mother's age, display the data and fitted model, check assumptions, and interpret the slope coefficient.

Based on this analysis, when do you recommend mothers should give birth? What are you assuming in making this recommendation?

Since the slope is positive, for every increase in mom age the ppvt increases by 0.85, indicating that giving birth at a later age is better due to the correlation of the values.

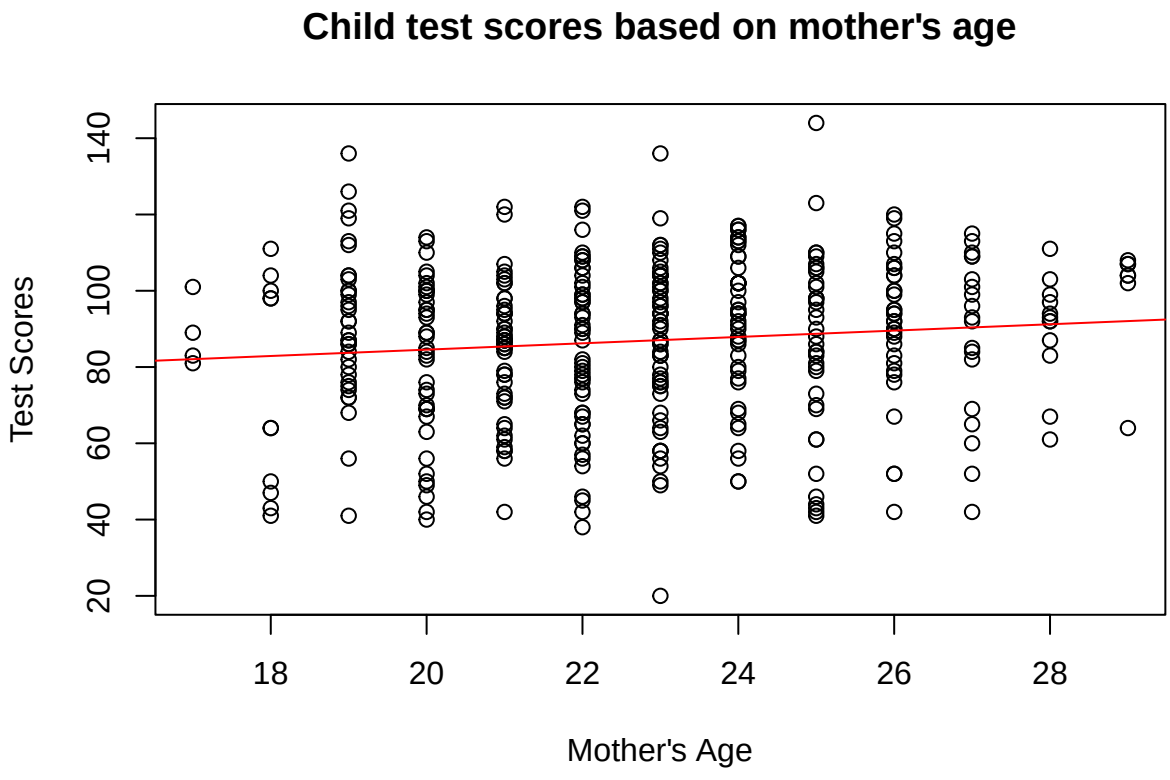
((when using lm, the significance of the mom age variable is however only * and the R-squareds offered are very low, indicating that the variable is not a good explanation for the variance of the ppvt variable. This would imply that the correlation is likely not causation. I'm not aware of how to easily get this information out of stan_glm))

```
model <- stan_glm(ppvt ~ momage, data = data, refresh = 0)
summary(model, digits = 2)
```

```
##
## Model Info:
##  function:     stan_glm
##  family:       gaussian [identity]
##  formula:      ppvt ~ momage
##  algorithm:    sampling
##  sample:       4000 (posterior sample size)
##  priors:       see help('prior_summary')
##  observations: 400
##  predictors:   2
##
## Estimates:
```

```
##                mean   sd    10%   50%   90%
## (Intercept) 67.87   8.90 56.23 67.92 79.25
## momage       0.84   0.39  0.35  0.83  1.34
## sigma       20.39   0.71 19.48 20.36 21.29
##
## Fit Diagnostics:
##            mean   sd    10%   50%   90%
## mean_PPD 86.95   1.44 85.09 86.97 88.79
##
## The mean_ppd is the sample average posterior predictive distribution of the outcome variable (for de
##
## MCMC diagnostics
##               mcse Rhat n_eff
## (Intercept)   0.14 1.00 3934
## momage        0.01 1.00 3995
## sigma         0.01 1.00 3645
## mean_PPD      0.02 1.00 3907
## log-posterior 0.03 1.00 2028
##
## For each parameter, mcse is Monte Carlo standard error, n_eff is a crude measure of effective sample
```

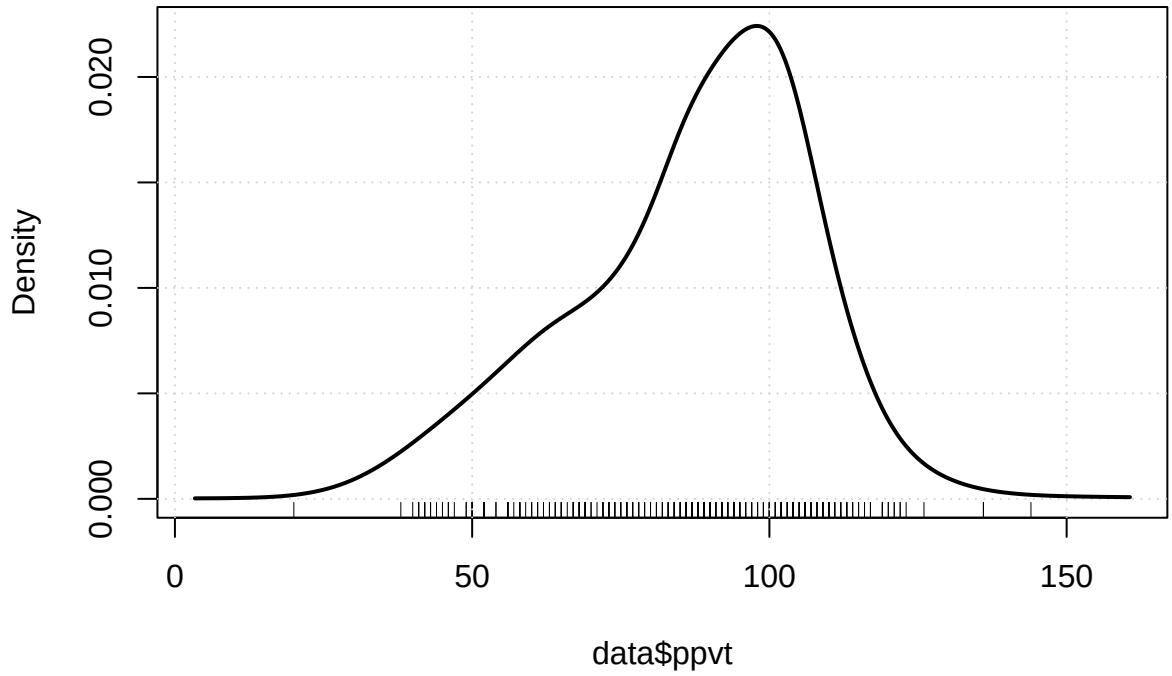
```
plot(data$momage, data$ppvt, xlab = "Mother's Age", ylab = "Test Scores", main = "Child test scores bas
abline(model, col = "red")
```



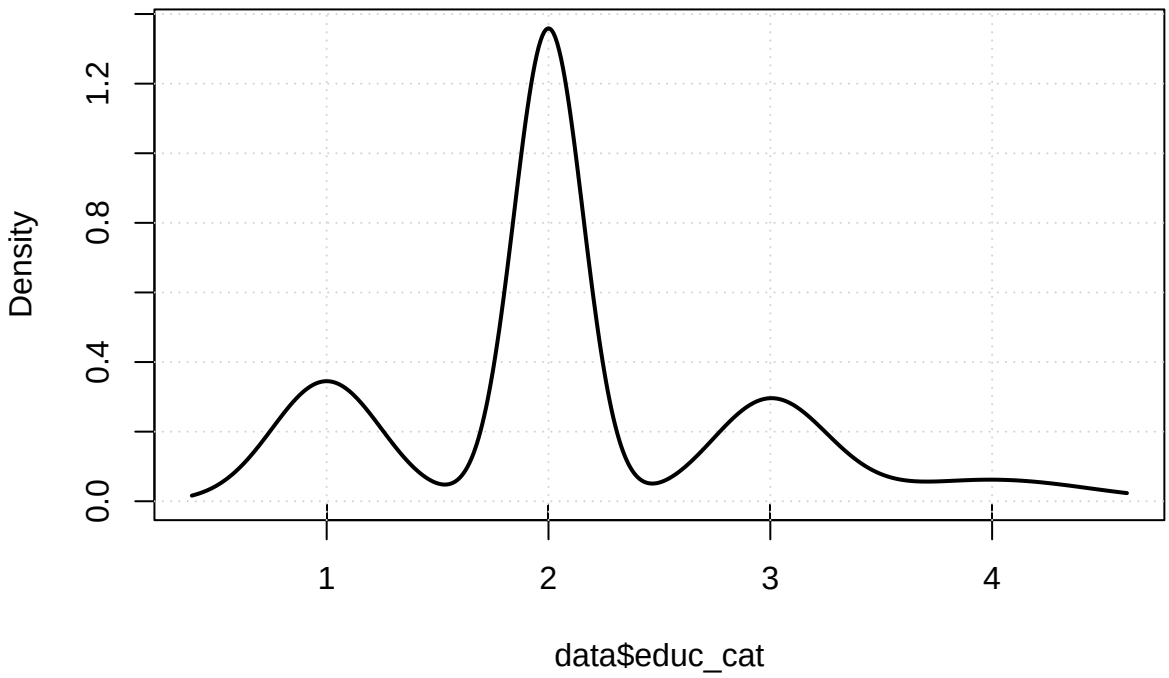
```
#spread of the data itself
summary(data)
```

```
##       ppvt           educ_cat         momage
##  Min.   : 20.00   Min.   :1.000   Min.   :17.00
##  1st Qu.: 74.00   1st Qu.:2.000   1st Qu.:21.00
##  Median : 90.00   Median :2.000   Median :23.00
##  Mean   : 86.93   Mean   :2.112   Mean   :22.79
##  3rd Qu.:102.00   3rd Qu.:3.000   3rd Qu.:25.00
##  Max.   :144.00   Max.   :4.000   Max.   :29.00
```

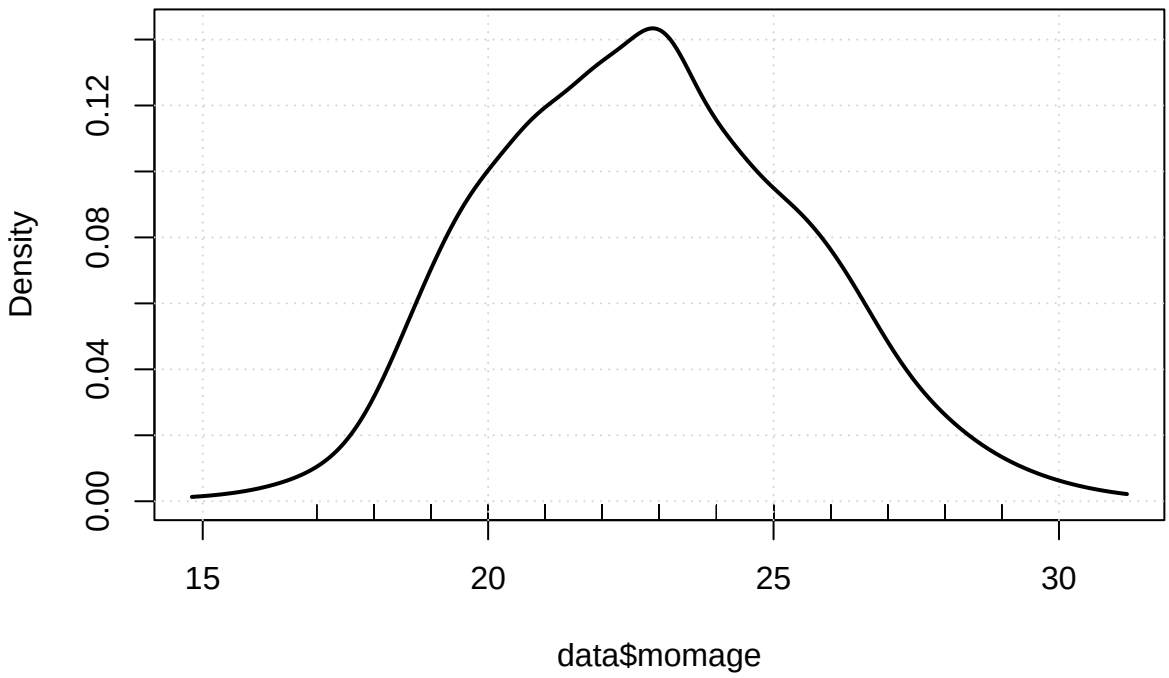
```
densityPlot(data$ppvt)
```




```
densityPlot(data$educ_cat)
```

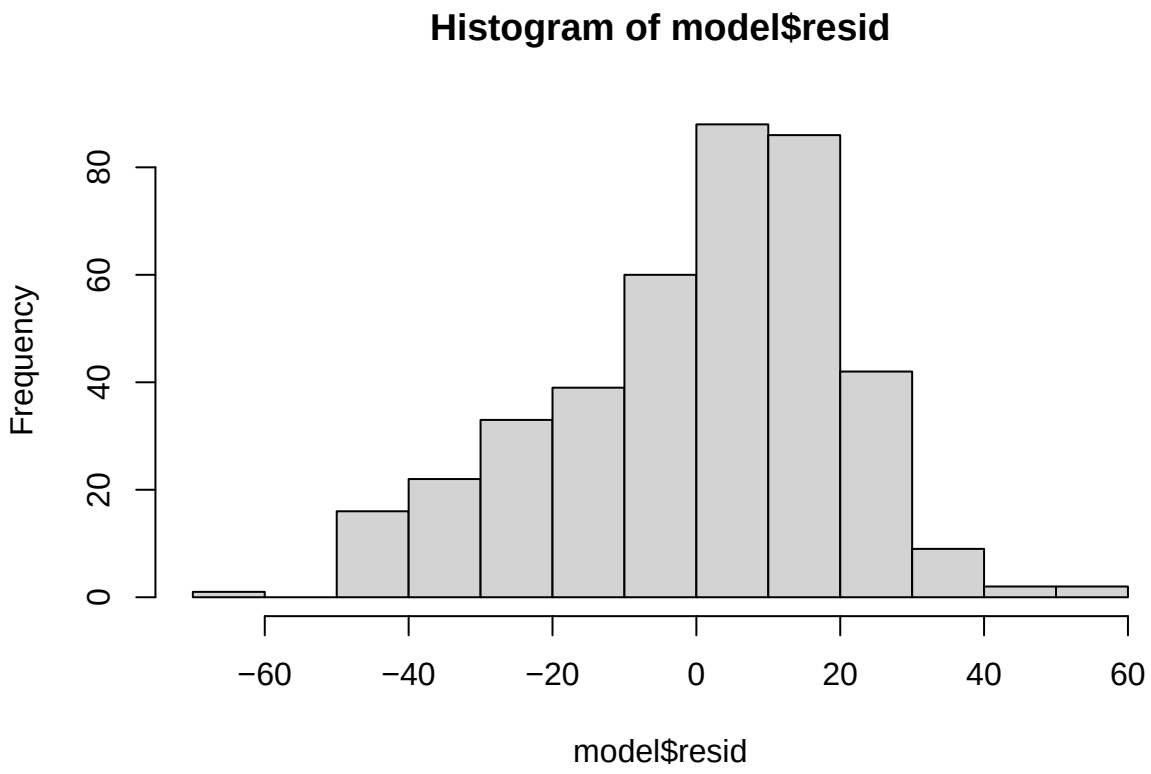



```
densityPlot(data$momage)
```

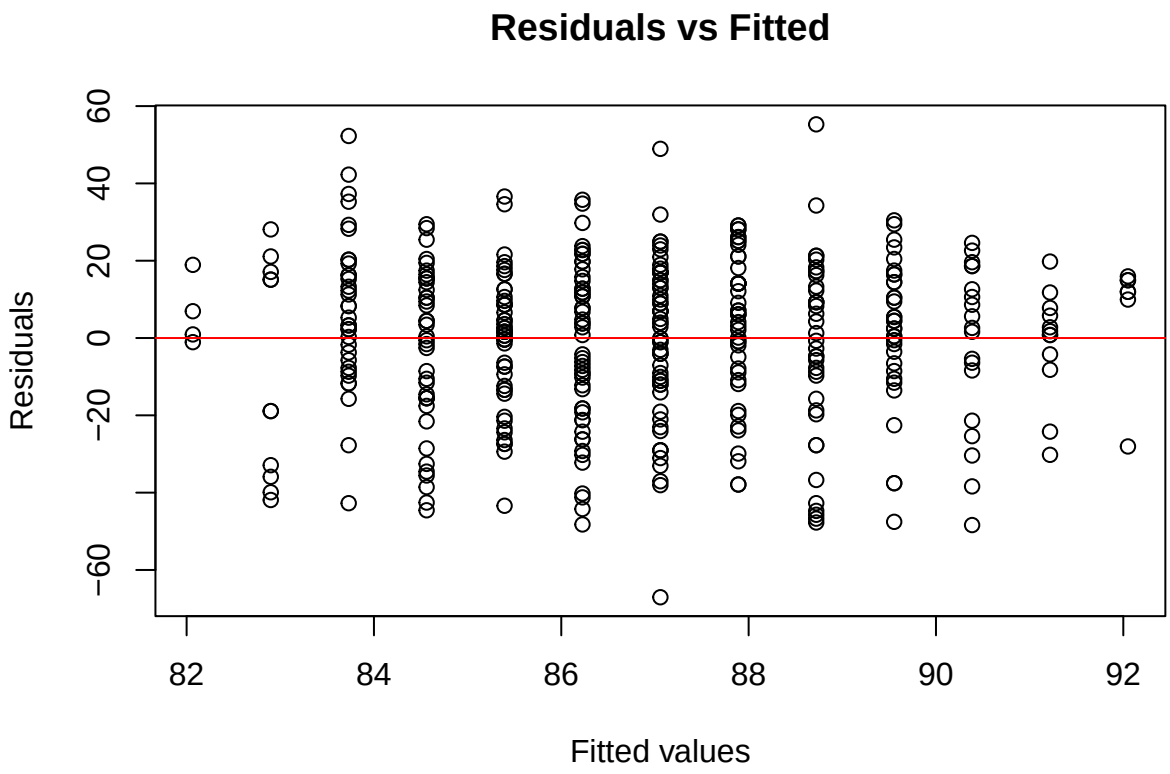


The level of mother education is particularly skewed towards level 2, with there being much fewer mothers with an education at level 4. The ppvt scores and the mother ages appear to be approximately normally distributed, meaning that if there is something that skews the data, this is very possibly the spread of education between mothers, with level 2 being particularly over-represented. That said, ppvt does have some skew towards lower scores, which may affect the results.

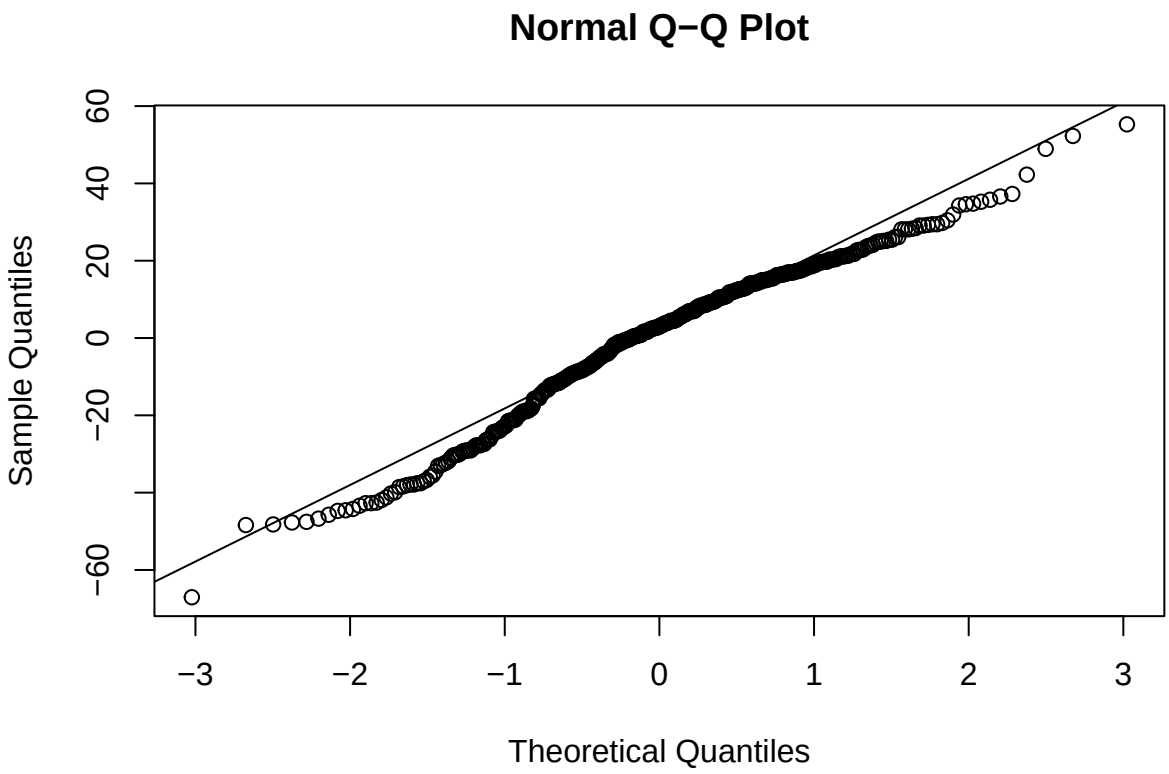
```
#assumptions of the model
hist(model$resid) #a bit skewed but could probably be worse,likely due to all the mothers with educatio
```




```
plot(fitted(model), resid(model), main = "Residuals vs Fitted",
     xlab = "Fitted values", ylab = "Residuals")
abline(h = 0, col = "red") #looks fairly pattern-less
```




```
qqnorm(model$resid)
qqline(model$resid) #a bit skewed but could probably be worse
```



Appears approximately normal enough with somewhat of a skew towards lower values, which may be due to the skew present in the data itself, in particular the education level of many of the mothers being at level 2 and a negative skew being present in the pptv scores of the babies.

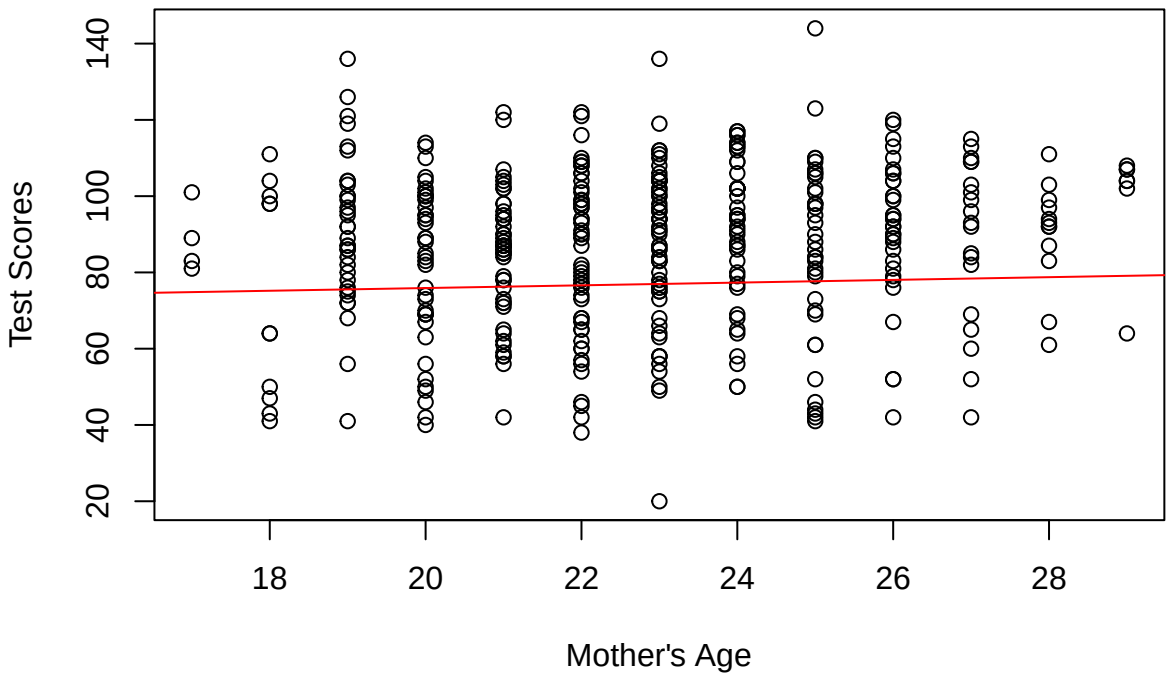
(b) Repeat this for a regression that further includes mother's education, interpreting both slope coefficients in this model. Have your conclusions about the timing of birth changed?

The slope for education level of mother is much higher than that of her age, indicating that it has a much higher effect/correlation on the test scores of the baby than the age at which the mother gave birth, with one increase in education level of mother changing the ppvt by 4.7. Comparably, the age of the mother now has a very low positive slope of 0.34, indicating that it does not have a high effect on the score.

((Furthermore, lm summary has marked the variable as highly statistically significant with a low p-value (***, p ~ 0) and the R squared value has increased significantly by the addition of educ_cat, indicating that it explains a considerable amount of the variance of the baby test score variable.))

```
model2 <- stan_glm(ppvt ~ momage + educ_cat, data = data, refresh = 0)

plot(data$momage, data$ppvt, xlab = "Mother's Age", ylab = "Test Scores")
abline(model2, col = "red")
```

```
## Warning in abline(model2, col = "red"): only using the first two of 3 regression
## coefficients
```



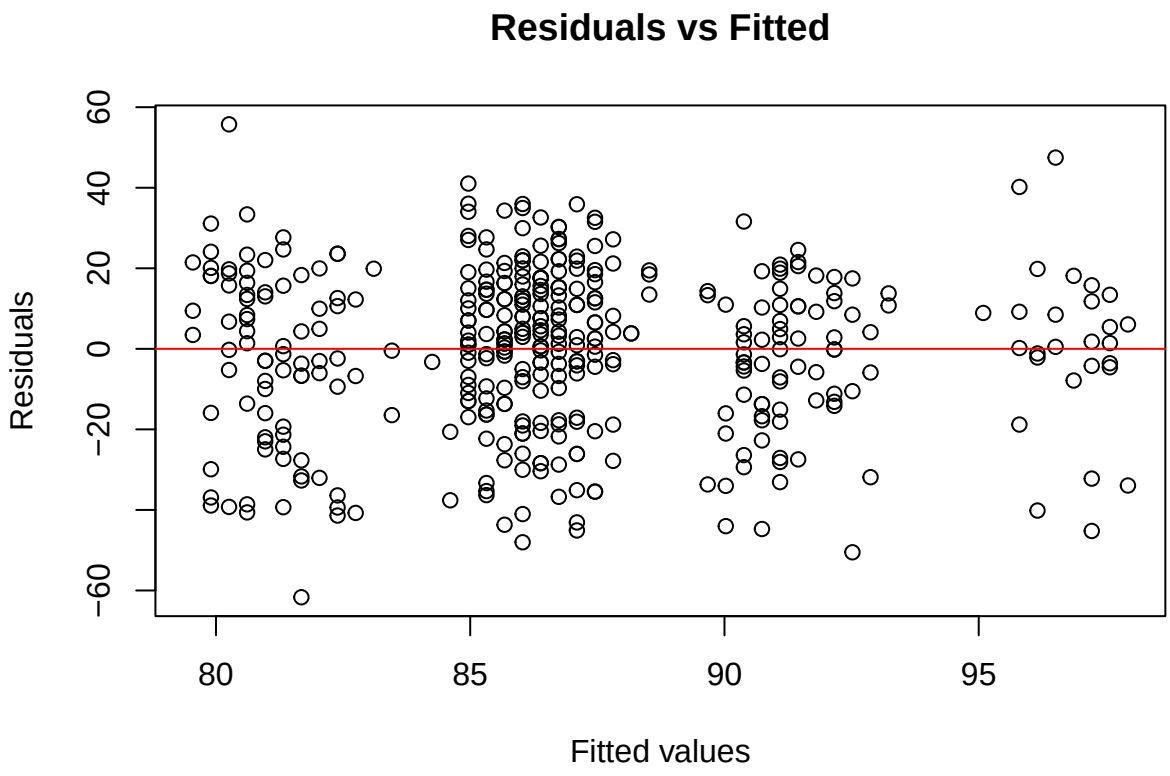

```
summary(model2, digits = 2)
```

```
##
## Model Info:
##  function:     stan_glm
##  family:       gaussian [identity]
##  formula:      ppvt ~ momage + educ_cat
##  algorithm:    sampling
##  sample:       4000 (posterior sample size)
##  priors:       see help('prior_summary')
##  observations: 400
##  predictors:   3
##
## Estimates:
##               mean   sd    10%   50%   90%
## (Intercept) 68.90   8.73 57.88 68.79 80.03
## momage       0.35   0.41 -0.17  0.36  0.88
## educ_cat     4.71   1.36  2.96  4.71  6.48
## sigma       20.06   0.70 19.18 20.05 20.97
##
## Fit Diagnostics:
##            mean   sd    10%   50%   90%
## mean_PPD 86.91   1.41 85.14 86.93 88.71
##
## The mean_ppd is the sample average posterior predictive distribution of the outcome variable (for de
```

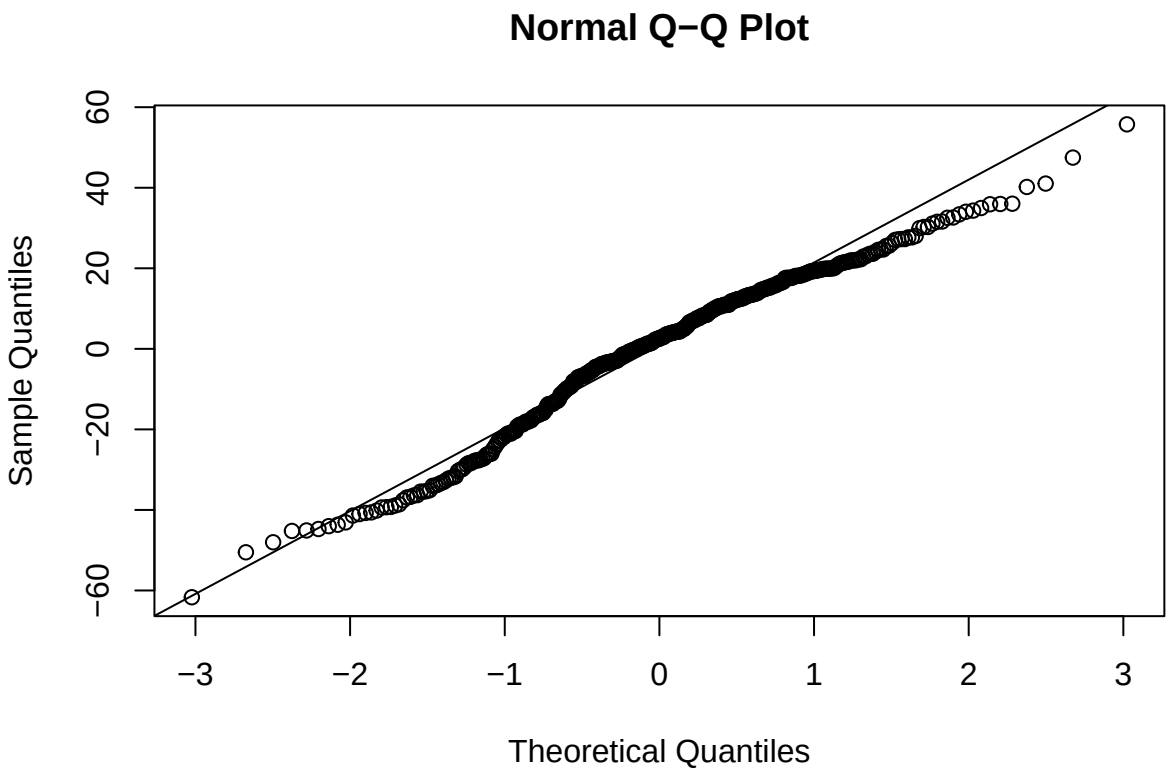
```
##
## MCMC diagnostics
##               mcse Rhat n_eff
## (Intercept)   0.13 1.00 4274
## momage        0.01 1.00 3754
## educ_cat      0.02 1.00 4016
## sigma         0.01 1.00 4882
## mean_PPD      0.02 1.00 3958
## log-posterior 0.03 1.00 1730
##
## For each parameter, mcse is Monte Carlo standard error, n_eff is a crude measure of effective sample
```

```
#assumptions
hist(model2$resid)#a bit skewed but could probably be worse, likely due to all the mothers with educati
```


```
plot(fitted(model2), resid(model2), main = "Residuals vs Fitted",
     xlab = "Fitted values", ylab = "Residuals")
abline(h = 0, col = "red") #looks fairly patternless, even if the datapoints are somewhat concentrated
```




```
qqnorm(model2$resid)
qqline(model2$resid) #a bit skewed but could probably be worse
```

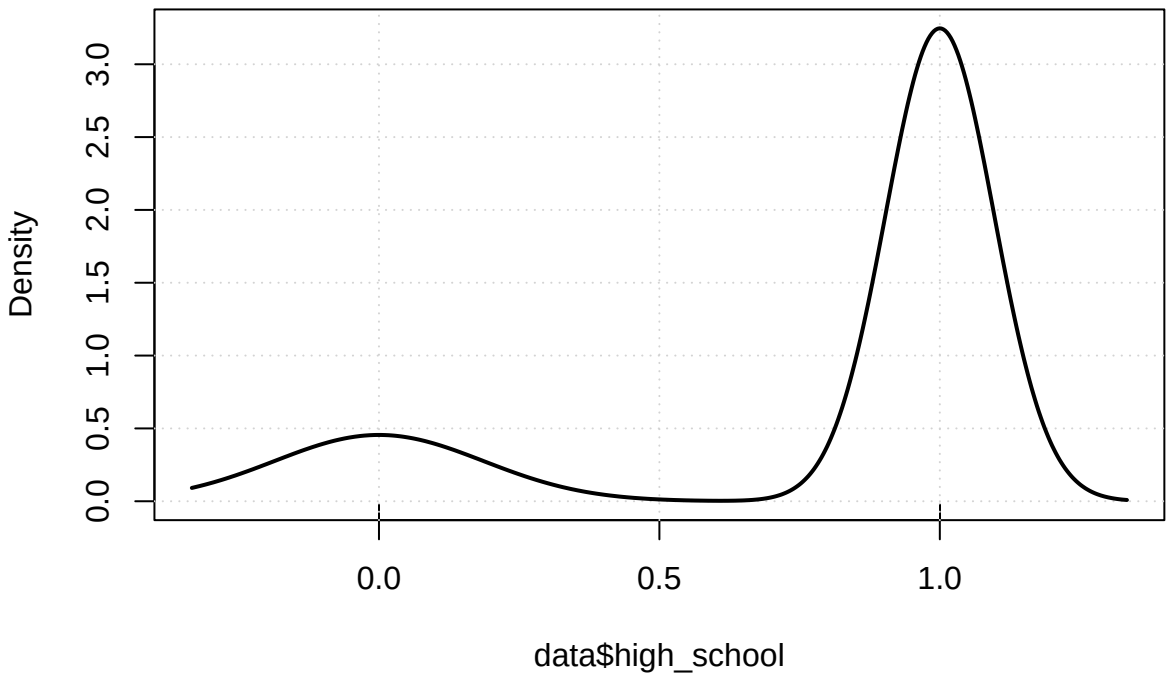


Appears approximately normal enough with somewhat of a skew towards lower values, which may be due to the skew present in the data itself, in particular the education level of many of the mothers being at level 2 and a negative skew being present in the pptv scores of the babies.

(c) Now create an indicator variable reflecting whether the mother has completed high school or not. Consider interactions between high school completion and mother's age. Also create a plot that shows the separate regression lines for each high school completion status group.

HS completion appears to be more influential than age as an indicator of a higher test score for the baby, at least when viewed visually.

```
data$high_school <- ifelse(data$educ_cat >= 2,1,0)
densityPlot(data$high_school) #most have graduated hs
```



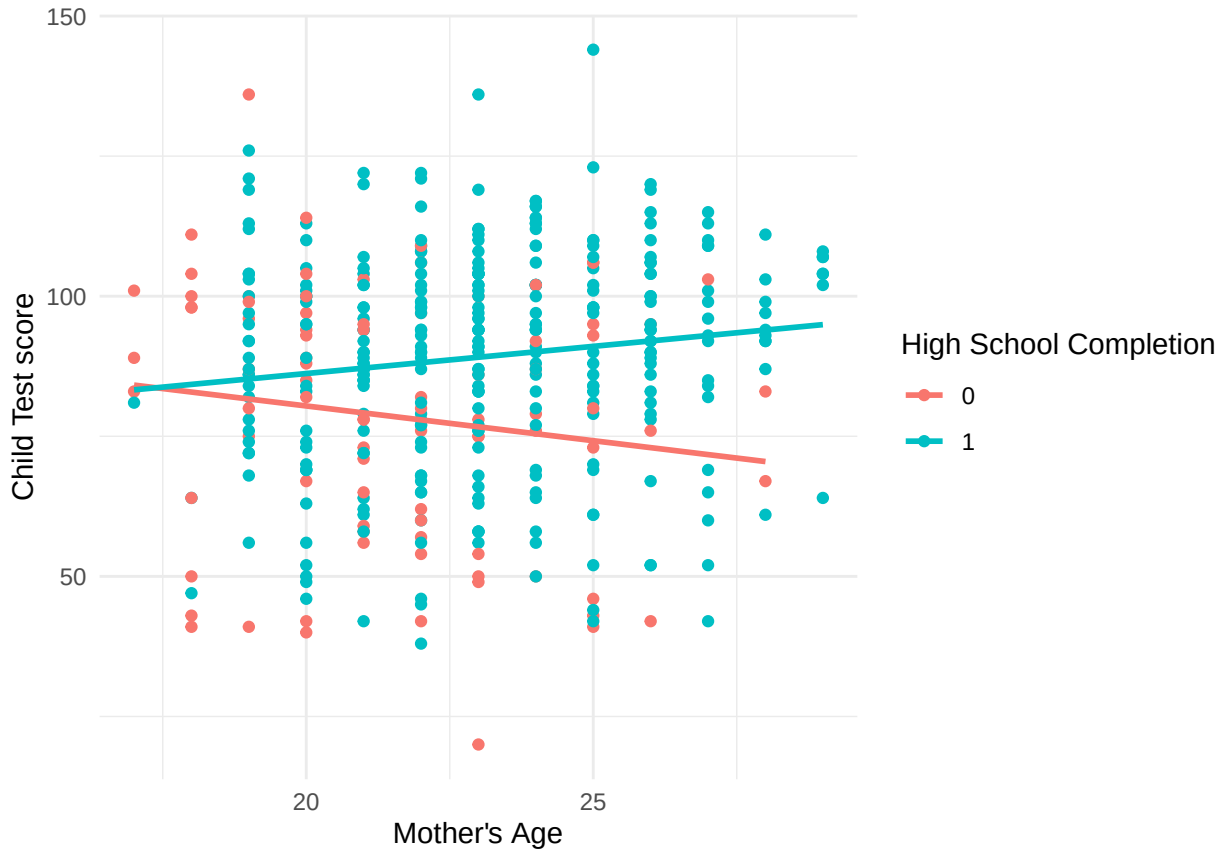

```
#interaction between hs completion and age
model3 <- stan_glm(ppvt ~ momage * high_school, data = data, refresh = 0)
summary(model3, digits = 2)
```

```
##
## Model Info:
##  function:     stan_glm
##  family:       gaussian [identity]
##  formula:      ppvt ~ momage * high_school
##  algorithm:    sampling
##  sample:       4000 (posterior sample size)
##  priors:       see help('prior_summary')
##  observations: 400
##  predictors:   4
##
## Estimates:
##                       mean   sd     10%    50%    90%
## (Intercept)         103.03  17.77  80.53 103.05 125.80
## momage               -1.14   0.81  -2.18  -1.14  -0.10
## high_school         -35.53  20.15 -61.66 -36.10  -9.77
## momage:high_school    2.08   0.91   0.91   2.09   3.27
## sigma                19.88   0.71  18.98  19.85  20.82
##
## Fit Diagnostics:
##            mean   sd    10%   50%   90%
## mean_PPD 86.96   1.42 85.12 86.97 88.78
```

```
##
## The mean_ppd is the sample average posterior predictive distribution of the outcome variable (for de
##
## MCMC diagnostics
##                     mcse Rhat n_eff
## (Intercept)         0.52 1.00 1170
## momage              0.02 1.00 1160
## high_school         0.60 1.00 1119
## momage:high_school  0.03 1.00 1111
## sigma               0.02 1.00 2159
## mean_PPD            0.03 1.00 2675
## log-posterior       0.05 1.00 1196
##
## For each parameter, mcse is Monte Carlo standard error, n_eff is a crude measure of effective sample
```

```r
ggplot(data, aes(x = momage, y = ppvt, color = as.factor(high_school))) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, aes(group = as.factor(high_school))) +
  labs(x = "Mother's Age", y = "Child Test score", color = "High School Completion") +
  theme_minimal()
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



(d) Finally, fit a regression of child test scores on mother's age and education level for the first 200 children and use this model to predict test scores for the next 200. Graphically display comparisons of the predicted and actual scores for the final 200 children.

```
#data for training
data_train <- head(data, 200)
nrow(data_train)
```

```
## [1] 200
```

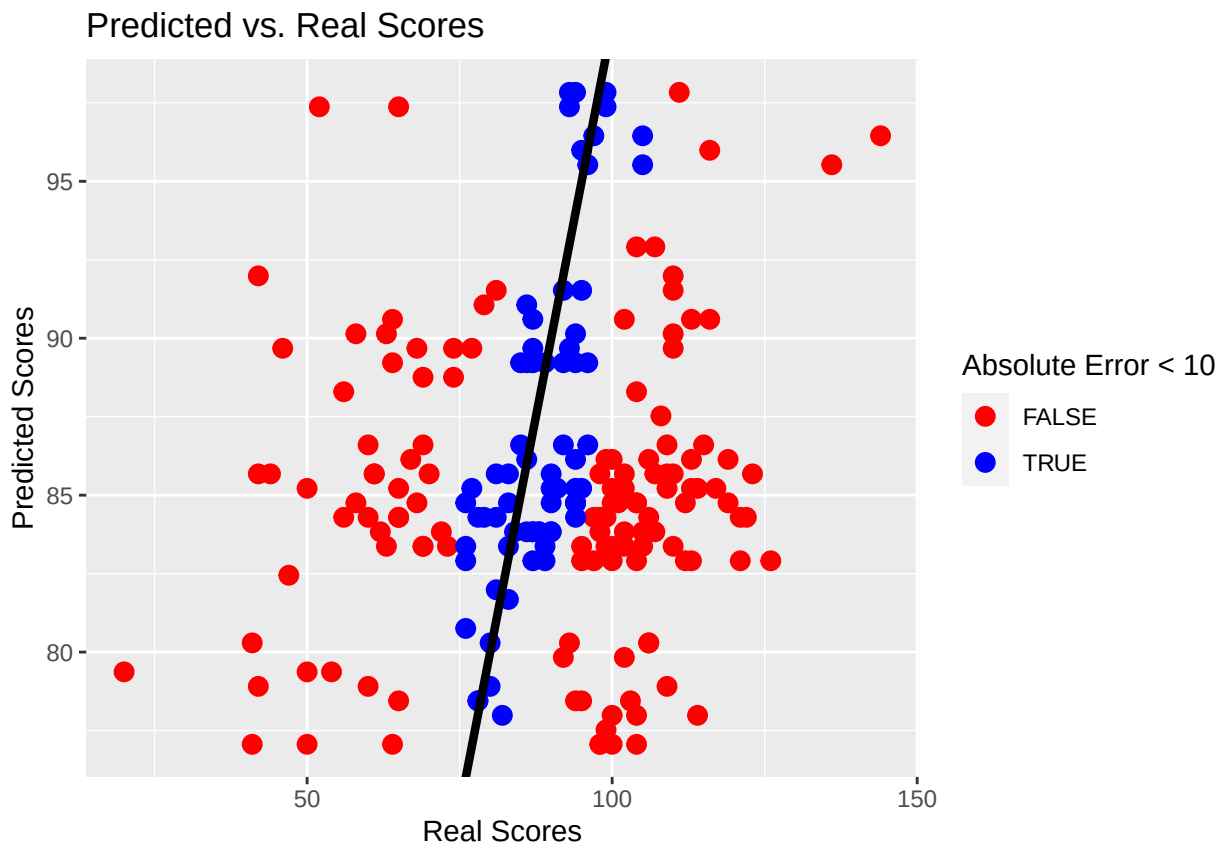
```
#data for testing
data_test <- tail(data, 200)
nrow(data_test)
```

```
## [1] 200
```

```
#just incase checking if there are common rows
intersect(row.names(data_train), row.names(data_test))
```

```
## character(0)
```

```
m_predict <- stan_glm(ppvt ~ momage + educ_cat, data = data_train, refresh = 0)
#summary(m_predict, digits = 2)

#predicting on the data_test
predicted_scores <- predict(m_predict, newdata = data_test)

df <- data.frame(real_scores = data_test$ppvt, predicted_scores = predicted_scores)

ggplot(df, aes(x = real_scores, y = predicted_scores)) +
  geom_point(aes(color = abs(real_scores - predicted_scores) < 10), size = 3) +
  scale_color_manual(values = c("red", "blue")) +
  labs(color = "Absolute Error < 10") +
  geom_abline(intercept = 0, slope = 1, linewidth = 1.5) +
  ggtitle("Predicted vs. Real Scores") +
  xlab("Real Scores") +
  ylab("Predicted Scores")
```




```
#its not very good at predicting? not a good model for it? did i do it right?
```

2. Exercise 10.6

Regression models with interactions: The folder `Beauty` contains data (use file `beauty.csv`) from Hamermesh and Parker (2005) on student evaluations of instructors' beauty and teaching quality for several courses at the University of Texas. The teaching evaluations were conducted at the end of the semester, and the beauty judgments were made later, by six students who had not attended the classes and were not aware of the course evaluations.

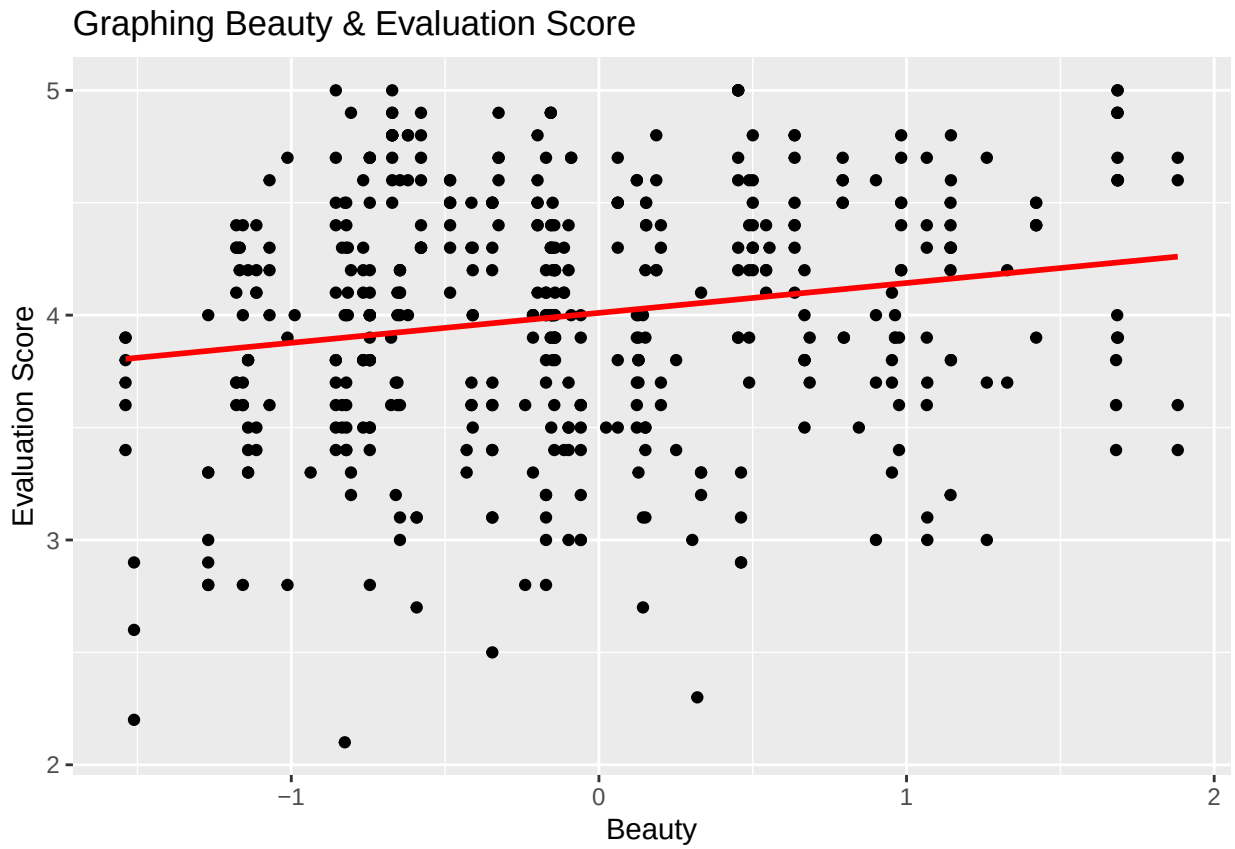
(a) Run a regression using beauty (the variable `beauty`) to predict course evaluations (`eval`), adjusting for various other predictors. Graph the data and fitted model, and explain the meaning of each of the coefficients along with the residual standard deviation. Plot the residuals versus fitted values.

```
beauty <- read.csv("data/beauty.csv")
```

```
#putting in all the possible predictors
b1 <- stan_glm(eval ~ beauty + female + age + minority + nonenglish + lower + course_id, data = beauty,
#b11 <- lm(eval ~ beauty + female + age + minority + nonenglish + lower + course_id, data = beauty) #i

summary(b1)
```

```
##
## Model Info:
##  function:     stan_glm
##  family:       gaussian [identity]
##  formula:      eval ~ beauty + female + age + minority + nonenglish + lower +
##     course_id
##  algorithm:    sampling
##  sample:       4000 (posterior sample size)
##  priors:       see help('prior_summary')
##  observations: 463
##  predictors:   8
##
## Estimates:
##               mean   sd   10%   50%   90%
## (Intercept)  4.2    0.1  4.0   4.2   4.4
## beauty       0.1    0.0  0.1   0.1   0.2
## female      -0.2    0.1 -0.3  -0.2  -0.1
## age          0.0    0.0  0.0   0.0   0.0
## minority    -0.1    0.1 -0.2  -0.1   0.0
## nonenglish  -0.3    0.1 -0.4  -0.3  -0.1
## lower        0.1    0.1  0.0   0.1   0.2
## course_id    0.0    0.0  0.0   0.0   0.0
## sigma        0.5    0.0  0.5   0.5   0.6
##
## Fit Diagnostics:
##            mean   sd   10%   50%   90%
## mean_PPD 4.0    0.0  4.0   4.0   4.0
##
## The mean_ppd is the sample average posterior predictive distribution of the outcome variable (for de
##
## MCMC diagnostics
##               mcse Rhat n_eff
## (Intercept)   0.0  1.0  3947
## beauty        0.0  1.0  4499
## female        0.0  1.0  4433
## age           0.0  1.0  3958
## minority      0.0  1.0  4125
## nonenglish    0.0  1.0  4097
## lower         0.0  1.0  4930
## course_id     0.0  1.0  4185
## sigma         0.0  1.0  5353
## mean_PPD      0.0  1.0  4233
## log-posterior 0.1  1.0  1633
##
## For each parameter, mcse is Monte Carlo standard error, n_eff is a crude measure of effective sample
```

```
#summary(b11) #the R squared is pathetically small, beauty and female are according to summary statisti

#graphing the data
ggplot(beauty, aes(x=beauty, y=eval)) +
  geom_point() +
  geom_smooth(method="lm", se=FALSE, color="red", formula = y~x) +
  labs(x="Beauty", y="Evaluation Score") +
  ggtitle("Graphing Beauty & Evaluation Score")
```



I fit a model which included all the predictors with no interactions. The intercept is 4.2, meaning that when all the predictors are at their "default" value, or constant, the evaluation is 4.2. The "beauty" coefficient being at 0.1 indicates that when other predictors are held constant and the beauty value is increased by 1 unit, then the evaluation goes up by 0.1. The predictor "lower", whether the instructor is an instructor of a lower-class division course, has a similar effect due to the same coefficient 0.1 . "age" and "course_id" appear to have no effect on course evaluation due to the coefficients being 0, at least not according to this model. Whether the instructor is female, a minority and a non-english speaker has a negative effect on the course evaluation, with the coefficients being -0.2, -0.1, -0.3 respectively (when the other predictors are held constant). The sigma value is 0.5, meaning that the average difference between observed values and predicted values is 0.5 units.

(b) Fit some other models, including beauty and also other predictors. Consider at least one model with interactions. For each model, explain the meaning of each of its estimated coefficients.

See also Felton, Mitchell, and Stinson (2003) for more on this topic.

```
m1 <- stan_glm(eval ~ beauty + female + minority + nonenglish, data = beauty, refresh = 0)
m2 <- stan_glm(eval ~ beauty + female + nonenglish + age, data = beauty, refresh = 0)
m3 <- stan_glm(eval ~ beauty*female + minority + nonenglish, data = beauty, refresh = 0)
m4 <- stan_glm(eval ~ beauty + beauty*female + minority*nonenglish, data = beauty, refresh = 0)

#lm for my own sake
# l1 <- lm(eval ~ beauty + female + minority + nonenglish, data = beauty)
# l2 <- lm(eval ~ beauty + female + nonenglish + age, data = beauty)
```

```
# l3 <- lm(eval ~ beauty*female + minority + nonenglish, data = beauty)
# l4 <- lm(eval ~ beauty + beauty*female + minority*nonenglish, data = beauty)

summary(m1)
```

```
##
## Model Info:
##  function:     stan_glm
##  family:       gaussian [identity]
##  formula:      eval ~ beauty + female + minority + nonenglish
##  algorithm:    sampling
##  sample:       4000 (posterior sample size)
##  priors:       see help('prior_summary')
##  observations: 463
##  predictors:   5
##
## Estimates:
##               mean   sd   10%   50%   90%
## (Intercept)  4.1    0.0  4.1   4.1   4.2
## beauty       0.2    0.0  0.1   0.2   0.2
## female      -0.2    0.1 -0.3  -0.2  -0.1
## minority     0.0    0.1 -0.1   0.0   0.1
## nonenglish  -0.3    0.1 -0.5  -0.3  -0.2
## sigma        0.5    0.0  0.5   0.5   0.6
##
## Fit Diagnostics:
##            mean   sd   10%   50%   90%
## mean_PPD 4.0    0.0  4.0   4.0   4.0
##
## The mean_ppd is the sample average posterior predictive distribution of the outcome variable (for de
##
## MCMC diagnostics
##               mcse Rhat n_eff
## (Intercept)   0.0  1.0  6441
## beauty        0.0  1.0  5555
## female        0.0  1.0  5344
## minority      0.0  1.0  4882
## nonenglish    0.0  1.0  4358
## sigma         0.0  1.0  6003
## mean_PPD      0.0  1.0  4298
## log-posterior 0.0  1.0  1828
##
## For each parameter, mcse is Monte Carlo standard error, n_eff is a crude measure of effective sample
```

```
summary(m2)
```

```
##
## Model Info:
##  function:     stan_glm
##  family:       gaussian [identity]
##  formula:      eval ~ beauty + female + nonenglish + age
##  algorithm:    sampling
```

```
##  sample:       4000 (posterior sample size)
##  priors:       see help('prior_summary')
##  observations: 463
##  predictors:   5
##
## Estimates:
##               mean   sd   10%   50%   90%
## (Intercept)  4.2    0.1  4.1   4.2   4.4
## beauty       0.1    0.0  0.1   0.1   0.2
## female      -0.2    0.1 -0.3  -0.2  -0.1
## nonenglish  -0.3    0.1 -0.5  -0.3  -0.2
## age          0.0    0.0  0.0   0.0   0.0
## sigma        0.5    0.0  0.5   0.5   0.6
##
## Fit Diagnostics:
##            mean   sd   10%   50%   90%
## mean_PPD 4.0    0.0  4.0   4.0   4.0
##
## The mean_ppd is the sample average posterior predictive distribution of the outcome variable (for de
##
## MCMC diagnostics
##               mcse Rhat n_eff
## (Intercept)   0.0  1.0  4495
## beauty        0.0  1.0  4174
## female        0.0  1.0  5068
## nonenglish    0.0  1.0  5089
## age           0.0  1.0  4494
## sigma         0.0  1.0  4678
## mean_PPD      0.0  1.0  4036
## log-posterior 0.0  1.0  2019
##
## For each parameter, mcse is Monte Carlo standard error, n_eff is a crude measure of effective sample
```

```
summary(m3)
```

```
##
## Model Info:
##  function:     stan_glm
##  family:       gaussian [identity]
##  formula:      eval ~ beauty * female + minority + nonenglish
##  algorithm:    sampling
##  sample:       4000 (posterior sample size)
##  priors:       see help('prior_summary')
##  observations: 463
##  predictors:   6
##
## Estimates:
##                  mean   sd   10%   50%   90%
## (Intercept)     4.1    0.0  4.1   4.1   4.2
## beauty          0.2    0.0  0.1   0.2   0.3
## female         -0.2    0.1 -0.3  -0.2  -0.1
## minority        0.0    0.1 -0.1   0.0   0.1
## nonenglish     -0.3    0.1 -0.5  -0.3  -0.2
## beauty:female -0.1    0.1 -0.2  -0.1   0.0
```

```
## sigma          0.5    0.0  0.5   0.5   0.6
##
## Fit Diagnostics:
##            mean   sd   10%   50%   90%
## mean_PPD 4.0    0.0  4.0   4.0   4.0
##
## The mean_ppd is the sample average posterior predictive distribution of the outcome variable (for de
##
## MCMC diagnostics
##               mcse Rhat n_eff
## (Intercept)   0.0  1.0  4153
## beauty        0.0  1.0  2629
## female        0.0  1.0  4544
## minority      0.0  1.0  4154
## nonenglish    0.0  1.0  4084
## beauty:female 0.0  1.0  2659
## sigma         0.0  1.0  4074
## mean_PPD      0.0  1.0  3659
## log-posterior 0.0  1.0  1998
##
## For each parameter, mcse is Monte Carlo standard error, n_eff is a crude measure of effective sample
```

```
summary(m4)
```

```
##
## Model Info:
##  function:     stan_glm
##  family:       gaussian [identity]
##  formula:      eval ~ beauty + beauty * female + minority * nonenglish
##  algorithm:    sampling
##  sample:       4000 (posterior sample size)
##  priors:       see help('prior_summary')
##  observations: 463
##  predictors:   7
##
## Estimates:
##                        mean   sd   10%   50%   90%
## (Intercept)           4.1    0.0  4.1   4.1   4.2
## beauty                0.2    0.0  0.1   0.2   0.3
## female               -0.2    0.1 -0.3  -0.2  -0.1
## minority              0.0    0.1 -0.1   0.0   0.1
## nonenglish           -0.2    0.2 -0.4  -0.2  -0.1
## beauty:female        -0.1    0.1 -0.2  -0.1   0.0
## minority:nonenglish -0.2    0.2 -0.5  -0.2   0.1
## sigma                 0.5    0.0  0.5   0.5   0.6
##
## Fit Diagnostics:
##            mean   sd   10%   50%   90%
## mean_PPD 4.0    0.0  4.0   4.0   4.0
##
## The mean_ppd is the sample average posterior predictive distribution of the outcome variable (for de
##
## MCMC diagnostics
##                      mcse Rhat n_eff
```

```
## (Intercept)         0.0  1.0  4285
## beauty              0.0  1.0  2890
## female              0.0  1.0  3632
## minority            0.0  1.0  3474
## nonenglish          0.0  1.0  2932
## beauty:female       0.0  1.0  2922
## minority:nonenglish 0.0  1.0  2616
## sigma               0.0  1.0  4163
## mean_PPD            0.0  1.0  3920
## log-posterior       0.1  1.0  1675
##
## For each parameter, mcse is Monte Carlo standard error, n_eff is a crude measure of effective sample
```

```
# summary(l1)
# summary(l2)
# summary(l3)
# summary(l4)
```

1. model 1 (m1) is a simple model featuring no interactions (eval ~ beauty + female + minority + nonenglish).

It's intercept is 4.1, meaning that when all the predictors are held constant, the evaluation value is 4.1. In this model, "beauty" has a coefficient of 0.2, meaning that by an increase in "beauty" by 1 unit, the evaluation goes up by 0.2, provided all the other predictors are held constant. "female" has a coefficient of -0.2, meaning that an increase in "female" by 1 unit, or rather, if the instructor is female, they are going to have an evaluation lower by -0.2, provided the other predictors are held constant. "nonenglish", -0.3, if the instructor is not a native speaker of english. If other predictors are held constant, then this means that the evaluation will go down by -0.3. In this model minority seems to have no effect. Sigma is 0.5, the average difference between observed values and predicted values is 0.5 units.

2. model 2 (m2) is another simple model featuring no interactions (eval ~ beauty + female + nonenglish + age)

Intercept 4.2, when all predictors are held constant the evaluation value is 4.2. "beauty" 0.2, when beauty increases by one unit and all the other predictors are held constant, evaluation increases by 0.2 "female" -0.2, when female increases by one unit (= instructor is a woman) and all the other predictors are held constant, evaluation decreases by 0.2 "nonenglish" -0.3, when nonenglish increases by one unit (= instructor is not a native speaker), and all the other predicotrs are held constant, evaluation decreases by 0.3. "age" 0, age seems to have no correlation with evaluation value. sigma 0.5, the average difference between observed values and predicted values is 0.5 units.

3. model 3 (m3) is a model featuring an interaction between beauty and gender (eval ~ beauty*female + minority + nonenglish)

Intercept is 4.1 "beauty" 0.2 "female" -0.2 minority 0.0 nonenglish -0.3 beauty:female -0.1, this variable represents the interaction effect between beauty and whether the instructor is a woman on the response variable "eval". By having an interaction like this, we can see whether the effect of beauty on the response variable differs based on whether the instructor is male or female. sigma 0.5

beauty:female represents the interaction effect between the beauty and female predictor variables on the response variable (eval). An interaction effect means that the effect of one predictor on the response variable is different depending on the level of another predictor, allowing for different slopes. Beauty on it's own,

when other predictors are held constant, has an effect/correlation of 0.2 points per increase in unit. When other values are held constant, "female" is at its base value, which is male. Meaning that the interaction effect in this case showcases the effect of beauty when the gender of the instructor is female, the coefficient of which is -0.1 + 2.0 = 1.0. It appears beauty is "more beneficial" towards male teachers when it comes to increasing their evaluation score, at least according to this model.

4. model 4 (m4) is a model featuring an interaction between beauty & gender and minority & nonenglish (eval ~ beauty + beauty*female* + *minority*nonenglish

Intercept is 4.1 "beauty" 0.2 "female" -0.2 minority 0.0 nonenglish -0.2 beauty:female -0.1 minority:nonenglish -0.2 sigma 0.5

minority:nonenglish is an interaction variable which represents the interaction between "minority" and "nonenglish", giving a different slope and intercept for the response variable based on the 4 possible combinations of "minority" and "nonenglish".

The coefficients are to be interpreted as follows:

minority:nonenglish coefficient = difference in the effect of being a non-English speaking minority compared to being a non-English speaking non-minority on the evaluation score –> both a minority and not a native english speaker (-0.2)

minority coefficient = difference in minority value while all other values are held constant –> minority but a native english speaker (0.0, no effect)

nonenglish coefficient = difference in nonenglish value while all other values are held constant –> not a minority but is a non-native english speaker (-0.2)

Intercept = evaluation score for non-minority non-English speaking instructors.

It appears that being a non-native english speaker has a negative effect on evaluation score while whether someone is a minority doesn't appear to matter according to this model.

3. Exercise 10.7

Predictive simulation for linear regression: Take one of the models from the previous exercise.

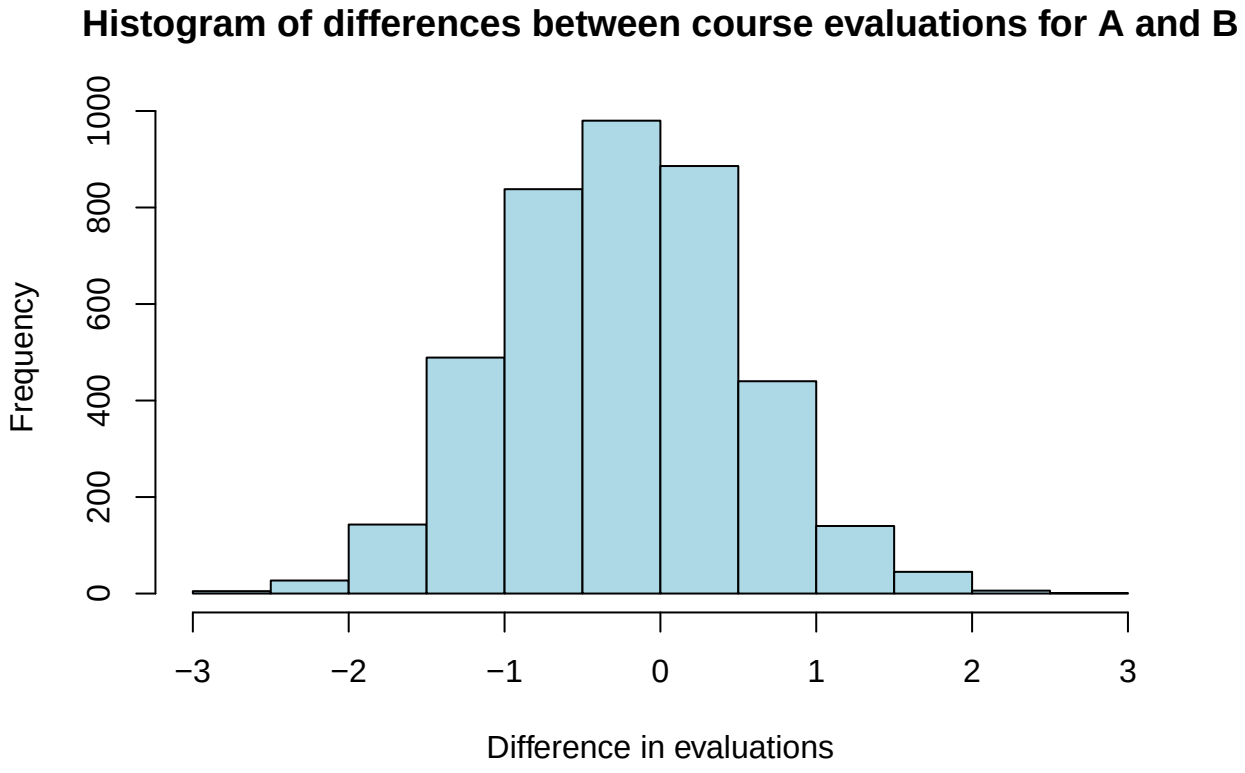
(a) Instructor A is a 50-year-old woman who is a native English speaker and has a beauty score of -1. Instructor B is a 60-year-old man who is a native English speaker and has a beauty score of -0.5. Simulate 1000 random draws of the course evaluation rating of these two instructors. In your simulation, use posterior_predict to account for the uncertainty in the regression parameters as well as predictive uncertainty.

```
#chosen model (since minority is not mentioned in the simulated instructors)
m2 <- stan_glm(eval ~ beauty + female + nonenglish + age, data = beauty, refresh = 0)

instr_a <- data.frame(age = 50, nonenglish = 0, beauty = -1, female = 1)
instr_b <- data.frame(age = 60, nonenglish = 0, beauty = -0.5, female = 0)

set.seed(123)
sim_a <- posterior_predict(m2, newdata = instr_a, n.sims = 1000)
sim_b <- posterior_predict(m2, newdata = instr_b, n.sims = 1000)
```

(b) Make a histogram of the difference between the course evaluations for A and B. What is the probability that A will have a higher evaluation?

37.6%

```
diff_evals <- sim_a - sim_b
hist(diff_evals, breaks = 20, col = "lightblue", xlab = "Difference in evaluations",
     main = "Histogram of differences between course evaluations for A and B")
```




```
prob_a_better <- mean(diff_evals > 0)
cat("The probability that A will have a higher evaluation than B is", (round(prob_a_better, 3))*100, "%
```

```
## The probability that A will have a higher evaluation than B is 38 %
```