Lab9

Jaya Veluri

12/6/2021

Exercise 1

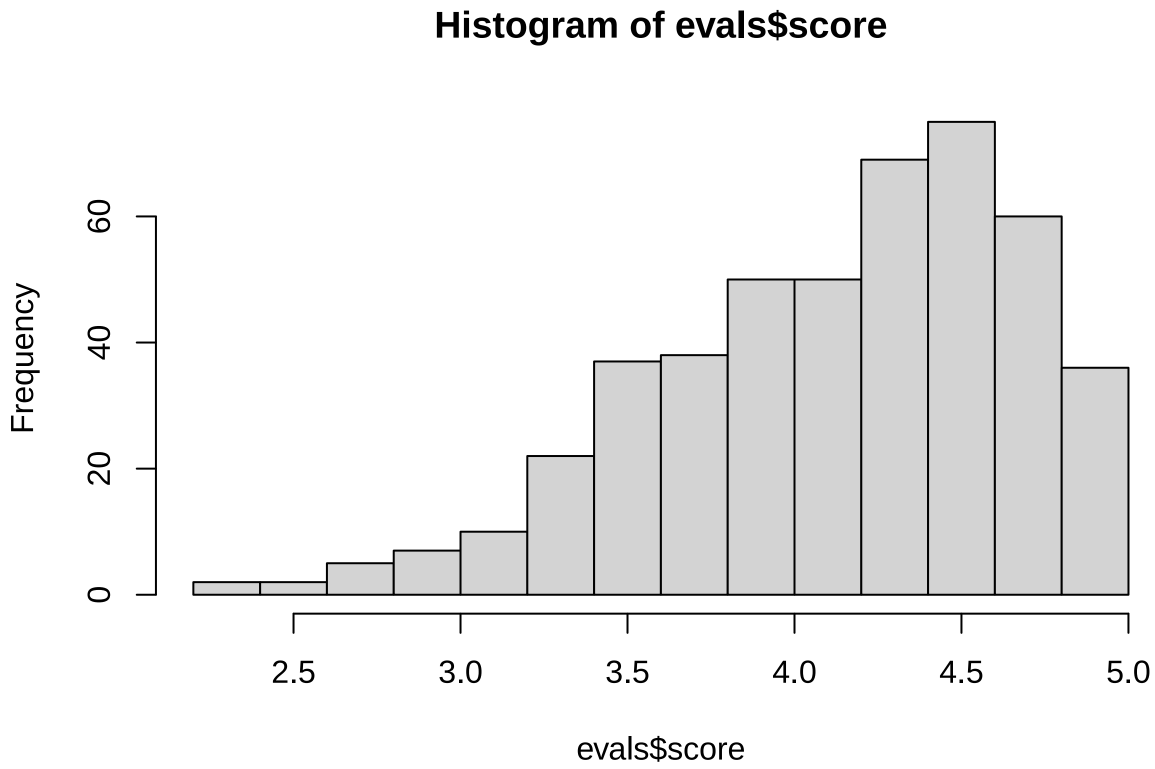
Is this an observational study or an experiment? The original research question posed in the paper is whether beauty leads directly to the differences in course evaluations. Given the study design, is it possible to answer this question as it is phrased? If not, rephrase the question.

```
## This is an observational study since there are no control and experimental groups. Since, this is on
```

Exercise 2

Describe the distribution of score. Is the distribution skewed? What does that tell you about how students rate courses? Is this what you expected to see? Why, or why not?

```
hist(evals$score)
```

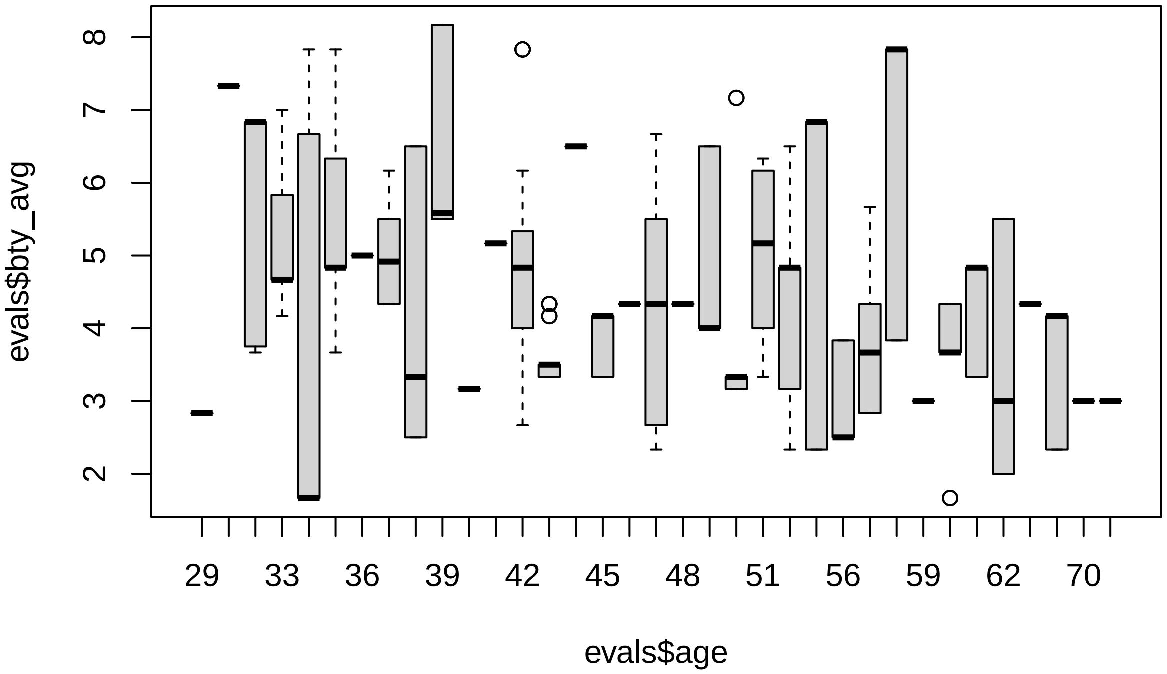


```
## Yes, the evaluation scores are skewed to the left. Students have far more positive evaluations than
```

Exercise 3

Excluding score, select two other variables and describe their relationship using an appropriate visualization (scatterplot, side-by-side boxplots, or mosaic plot).

```
boxplot(evals$bty_avg ~ evals$age)
```

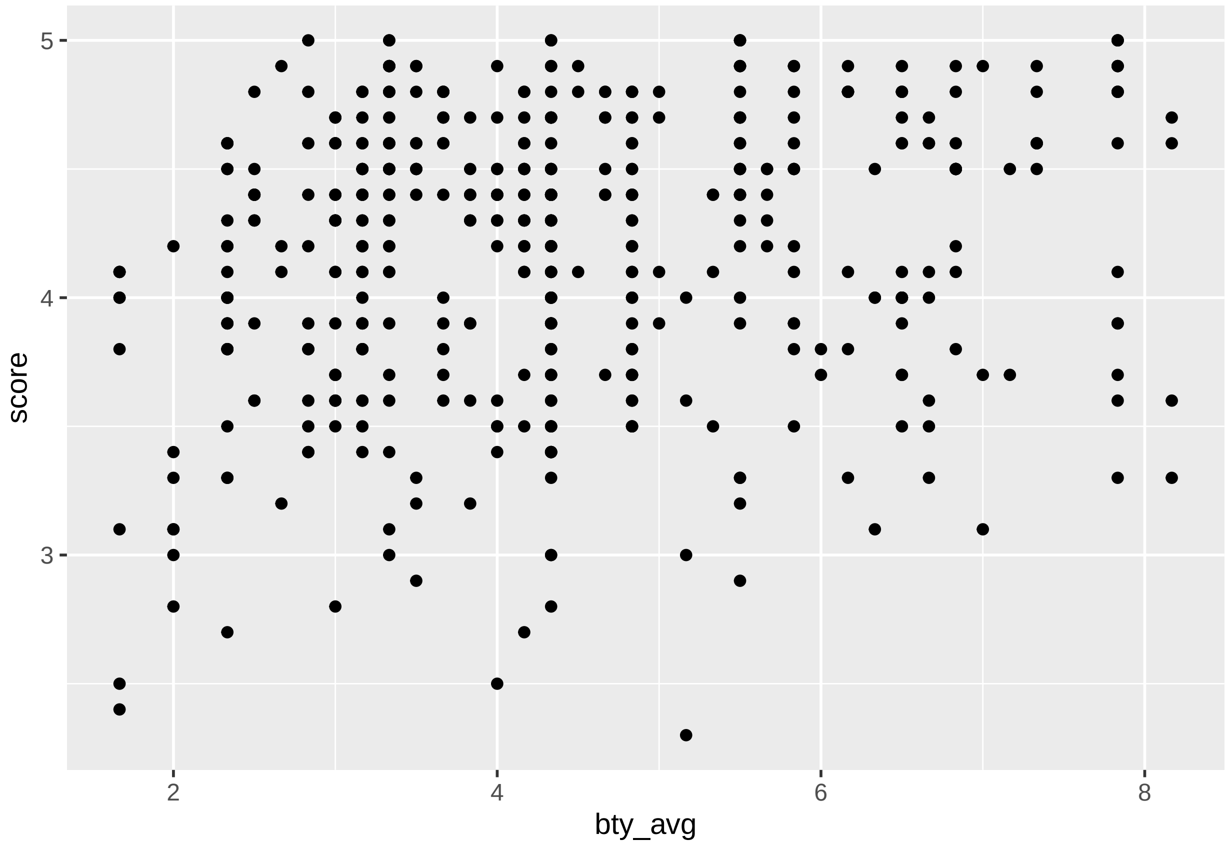


```
## Purely through visualization of the side-by-side boxplot, there doesn't seem to be a relationship be
```

Simple linear regression

The fundamental phenomenon suggested by the study is that better looking teachers are evaluated more favorably. Let's create a scatterplot to see if this appears to be the case:

```
ggplot(data = evals, aes(x = bty_avg, y = score)) +
  geom_point()
```



Before you draw conclusions about the trend, compare the number of observations in the data frame with the approximate number of points on the scatterplot. Is anything awry?

```
nrow(evals)
```

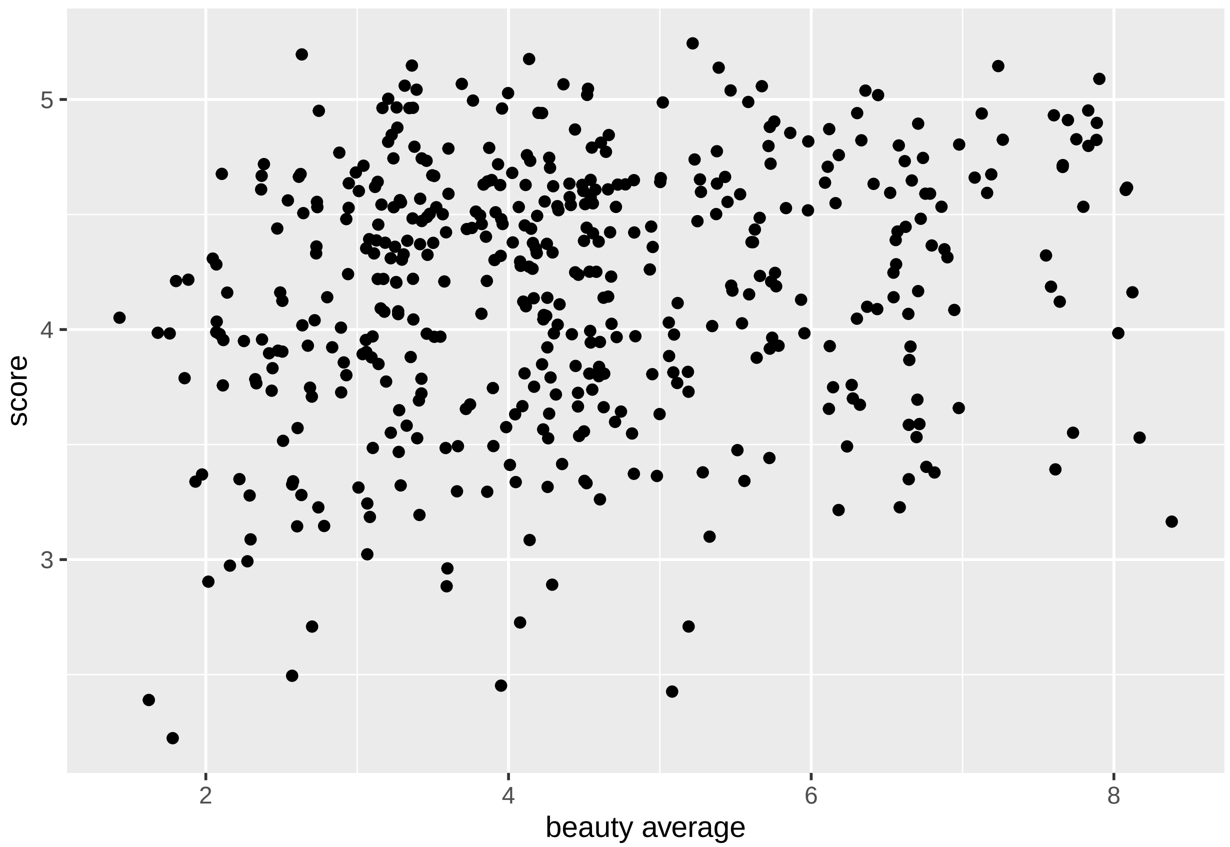
```
## [1] 463
```

```
## There seem to be more observations than the approximate number of points on the scatterplot
```

Exercise 4

Replot the scatterplot, but this time use the function jitter() on the y- or the x-coordinate. (Use ?jitter to learn more.) What was misleading about the initial scatterplot?

```
ggplot(evals, aes(bty_avg, score)) + geom_point(position = position_jitter(w = 0.3, h = 0.3)) + ylab("s
```

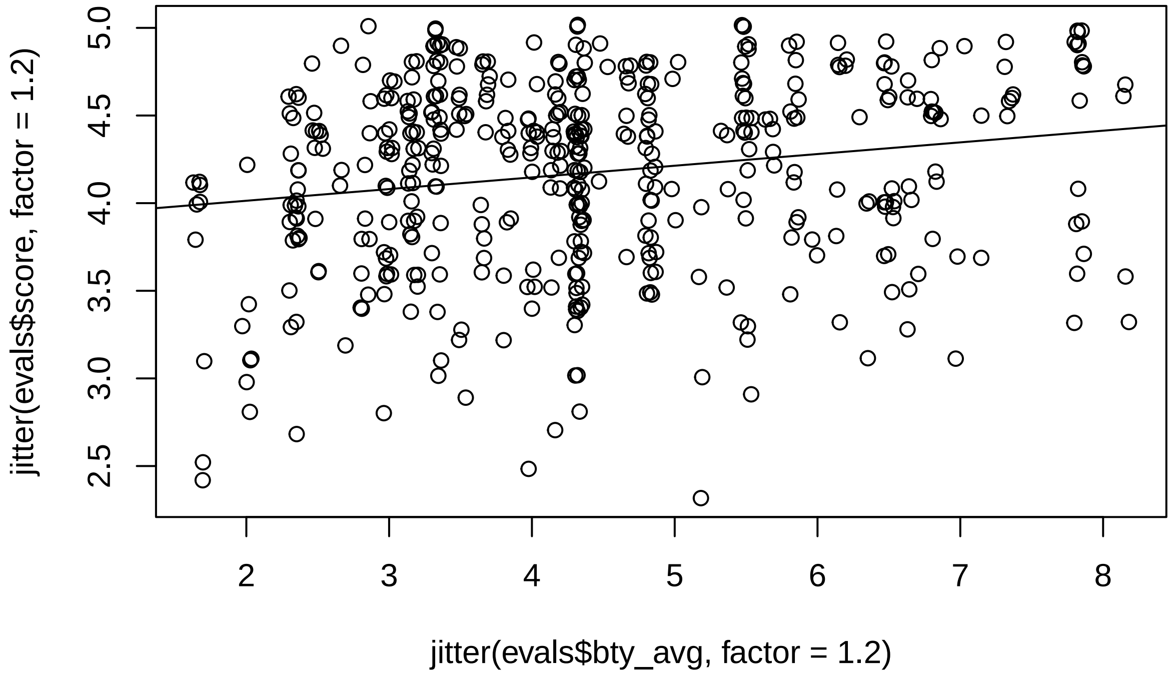


```
## It wasn't able to show the relationship between beauty average and score for the teacher because of
```

Exercise 5

Let's see if the apparent trend in the plot is something more than natural variation. Fit a linear model called m_bty to predict average professor score by average beauty rating and add the line to your plot using abline(m_bty). Write out the equation for the linear model and interpret the slope. Is average beauty score a statistically significant predictor? Does it appear to be a practically significant predictor?

```
m_bty <- lm(evals$score ~ evals$bty_avg)
plot(jitter(evals$score,factor=1.2) ~ jitter(evals$bty_avg,factor=1.2))
abline(m_bty)
```



```
cor(evals$score, evals$bty_avg)
```

```
## [1] 0.1871424
```

```
summary(m_bty)
```

```
##
## Call:
## lm(formula = evals$score ~ evals$bty_avg)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.9246 -0.3690  0.1420  0.3977  0.9309
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    3.88034    0.07614   50.96  < 2e-16 ***
## evals$bty_avg  0.06664    0.01629    4.09 5.08e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5348 on 461 degrees of freedom
## Multiple R-squared:  0.03502,    Adjusted R-squared:  0.03293
## F-statistic: 16.73 on 1 and 461 DF,  p-value: 5.083e-05
## Yes, bty_avg is a statistically significant predictor of evaluation score with p-value close of 0. I
```

Exercise 6

Use residual plots to evaluate whether the conditions of least squares regression are reasonable. Provide plots and comments for each one (see the Simple Regression Lab for a reminder of how to make these).

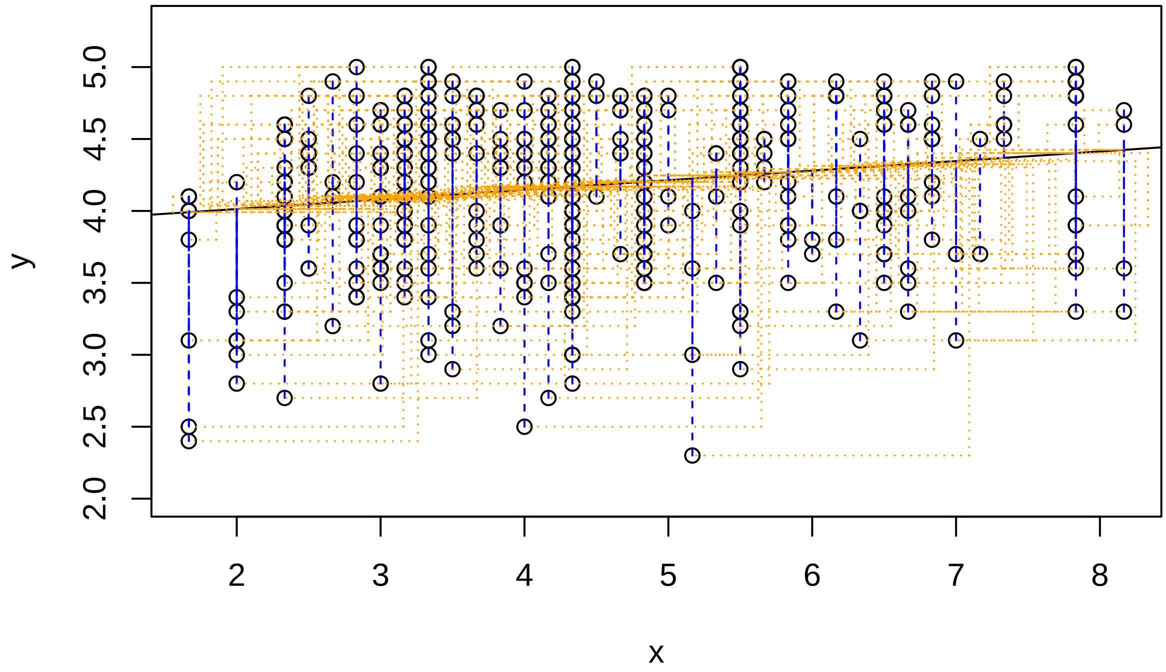
```
plot_ss <- function(x, y, showSquares = FALSE, leastSquares = FALSE){
  plot(y~x, asp = 1)# xlab = paste(substitute(x)), ylab = paste(substitute(y)))

  if(leastSquares){
    m1 <- lm(y~x)
    y.hat <- m1$fit
  } else{
    cat("Click two points to make a line.")
    pt1 <- locator(1)
    points(pt1$x, pt1$y, pch = 4)
    pt2 <- locator(1)
    points(pt2$x, pt2$y, pch = 4)
    pts <- data.frame("x" = c(pt1$x, pt2$x),"y" = c(pt1$y, pt2$y))
    m1 <- lm(y ~ x, data = pts)
    y.hat <- predict(m1, newdata = data.frame(x))
  }
  r <- y - y.hat
  abline(m1)

  oSide <- x - r
  LLim <- par()$usr[1]
  RLim <- par()$usr[2]
  oSide[oSide < LLim | oSide > RLim] <- c(x + r)[oSide < LLim | oSide > RLim] # move boxes to avoid mar

  n <- length(y.hat)
  for(i in 1:n){
    lines(rep(x[i], 2), c(y[i], y.hat[i]), lty = 2, col = "blue")
    if(showSquares){
    lines(rep(oSide[i], 2), c(y[i], y.hat[i]), lty = 3, col = "orange")
    lines(c(oSide[i], x[i]), rep(y.hat[i],2), lty = 3, col = "orange")
    lines(c(oSide[i], x[i]), rep(y[i],2), lty = 3, col = "orange")
    }
  }

  SS <- round(sum(r^2), 3)
  cat("\r                                ")
  print(m1)
  cat("Sum of Squares: ", SS)
}
plot_ss(x = evals$bty_avg, y = evals$score, showSquares = TRUE)
```



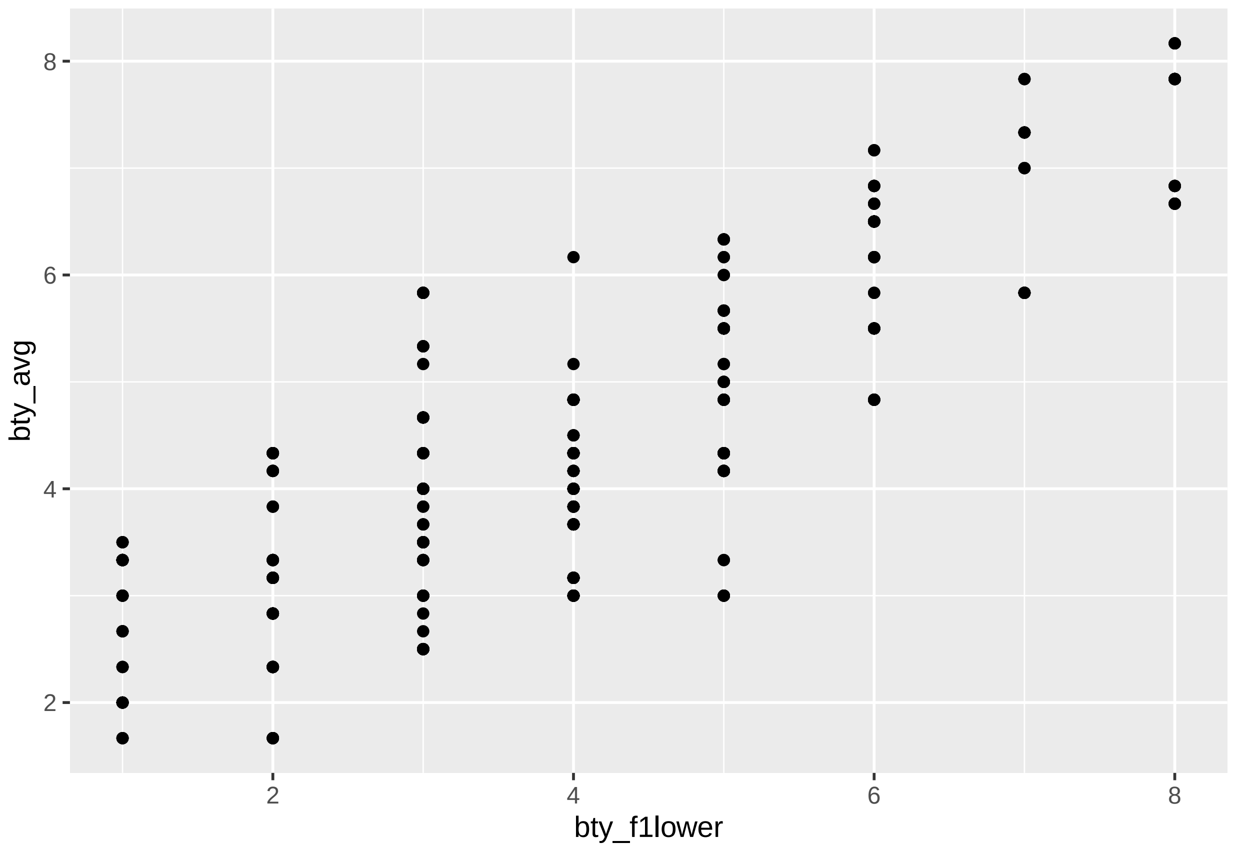
```
## Click two points to make a line.
## Call:
## lm(formula = y ~ x, data = pts)
##
## Coefficients:
## (Intercept)            x
##     3.88034      0.06664
##
## Sum of Squares:  131.868
```

```
## Probably not. There are too many outliers and the distribution is not normal.
```

Multiple linear regression

The data set contains several variables on the beauty score of the professor: individual ratings from each of the six students who were asked to score the physical appearance of the professors and the average of these six scores. Let's take a look at the relationship between one of these scores and the average beauty score.

```
ggplot(data = evals, aes(x = bty_f1lower, y = bty_avg)) +
  geom_point()
```

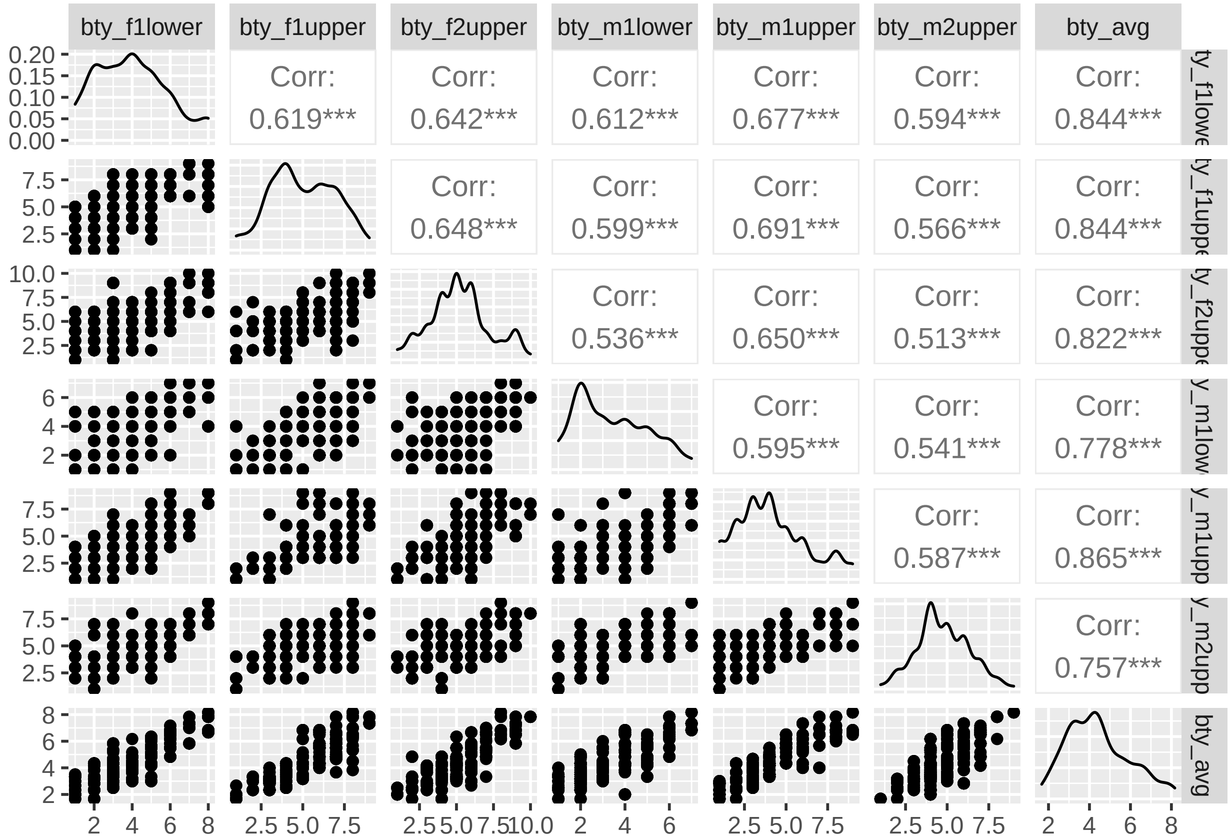


```
evals %>%
  summarise(cor(bty_avg, bty_f1lower))
```

```
## # A tibble: 1 x 1
##   ‘cor(bty_avg, bty_f1lower)‘
##                         <dbl>
## 1                       0.844
```

As expected the relationship is quite strong - after all, the average score is calculated using the individual scores. We can actually take a look at the relationships between all beauty variables (columns 13 through 19) using the following command:

```
evals %>%
  select(contains("bty")) %>%
  ggpairs()
```



These variables are collinear (correlated), and adding more than one of these variables to the model would not add much value to the model. In this application and with these highly-correlated predictors, it is reasonable to use the average beauty score as the single representative of these variables.

In order to see if beauty is still a significant predictor of professor score after you've accounted for the professor's gender, you can add the gender term into the model.

```
m_bty_gen <- lm(score ~ bty_avg + gender, data = evals)
summary(m_bty_gen)
```

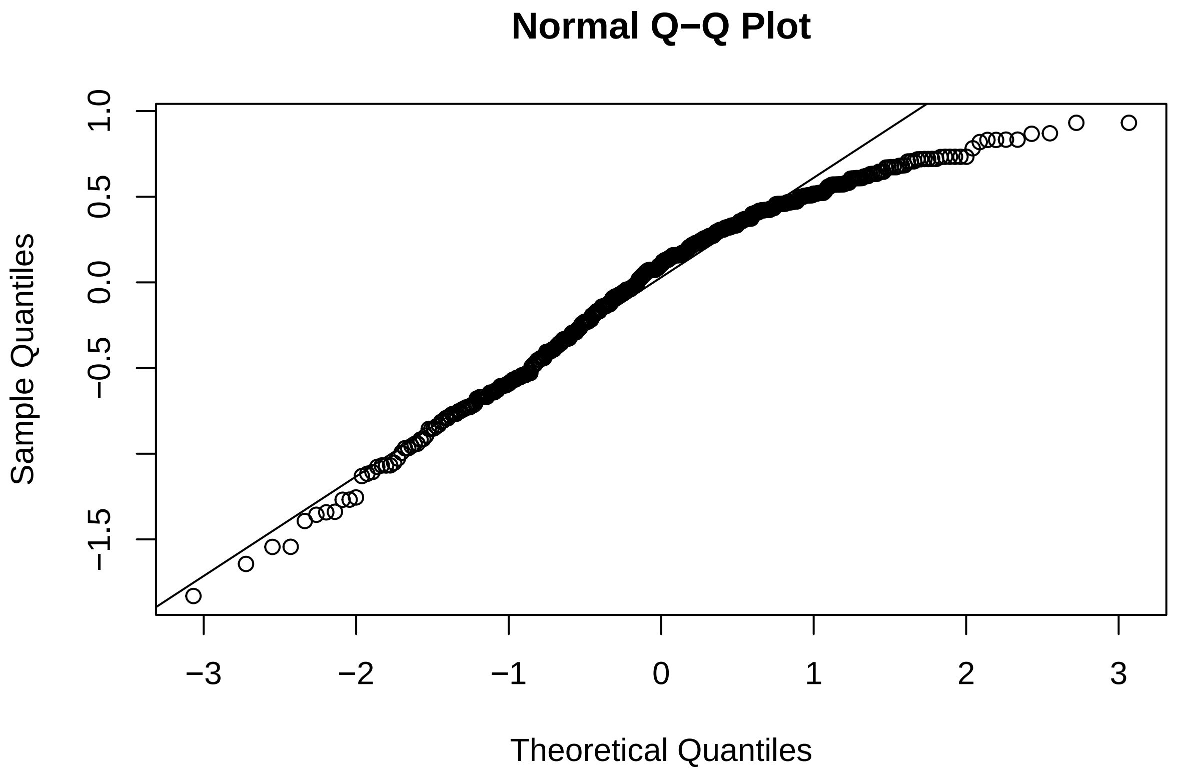
```
##
## Call:
## lm(formula = score ~ bty_avg + gender, data = evals)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.8305 -0.3625  0.1055  0.4213  0.9314
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.74734    0.08466  44.266  < 2e-16 ***
## bty_avg      0.07416    0.01625   4.563 6.48e-06 ***
## gendermale   0.17239    0.05022   3.433 0.000652 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.5287 on 460 degrees of freedom
## Multiple R-squared:  0.05912,    Adjusted R-squared:  0.05503
## F-statistic: 14.45 on 2 and 460 DF,  p-value: 8.177e-07
```

Exercise 7

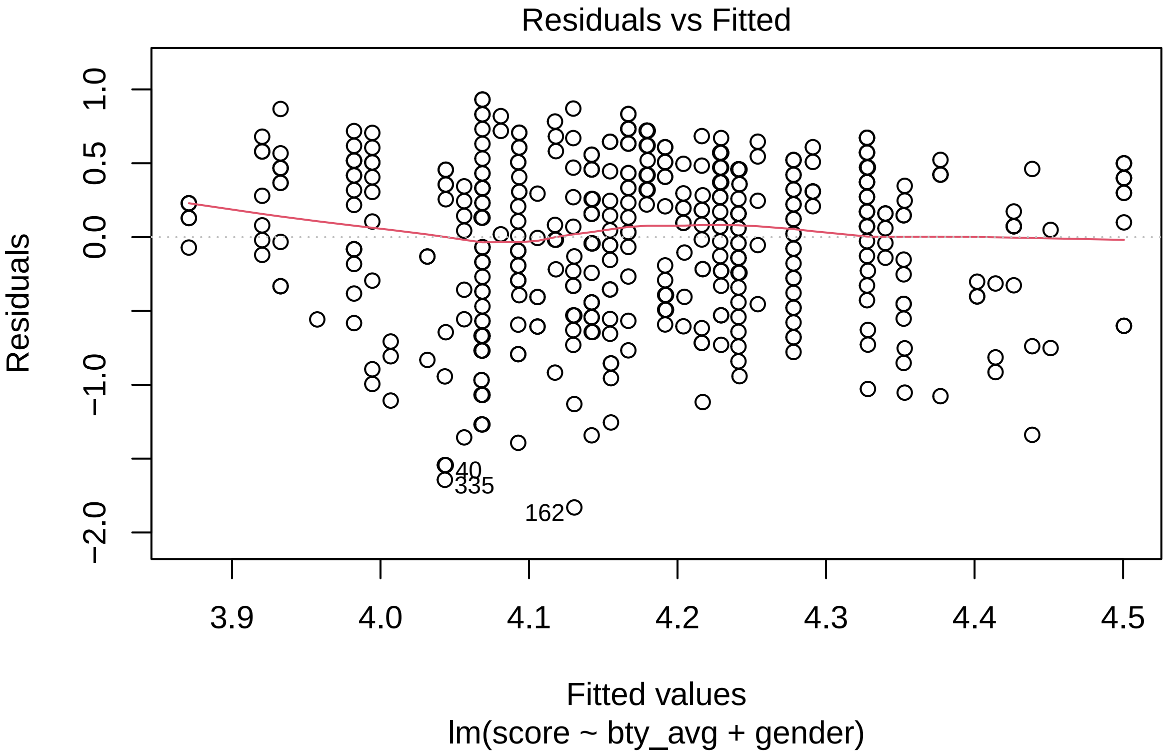
P-values and parameter estimates should only be trusted if the conditions for the regression are reasonable. Verify that the conditions for this model are reasonable using diagnostic plots.

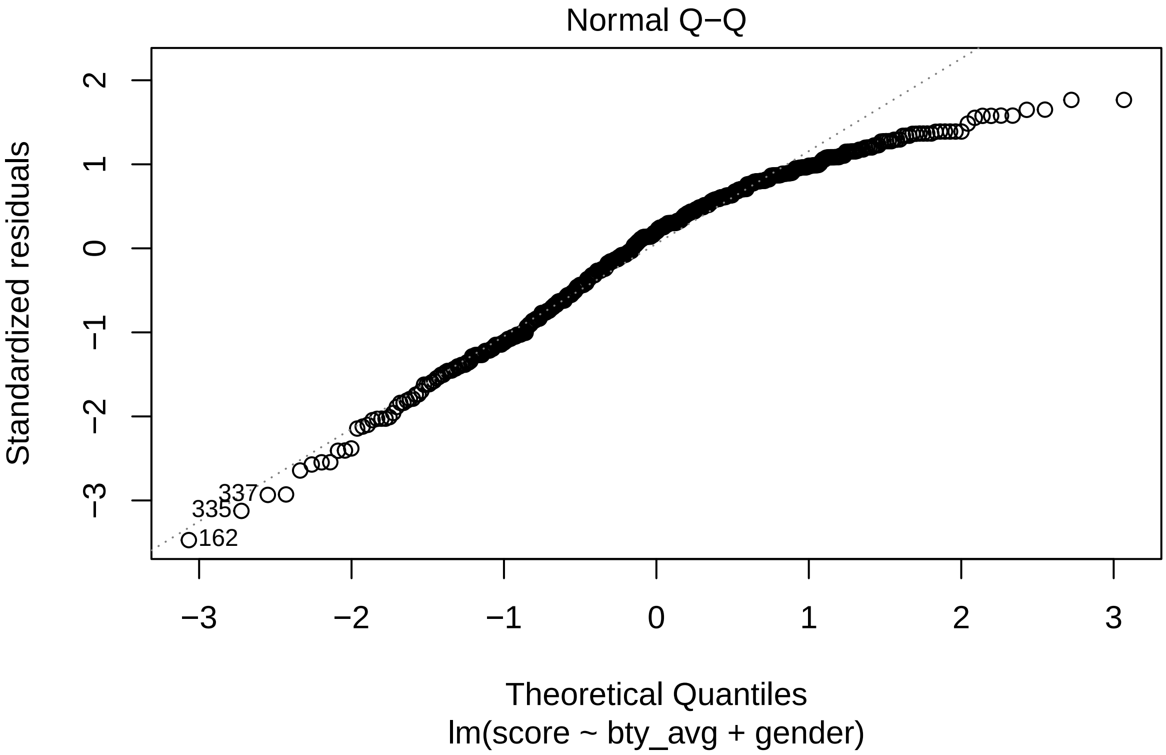
1. the residuals of the model are nearly normal

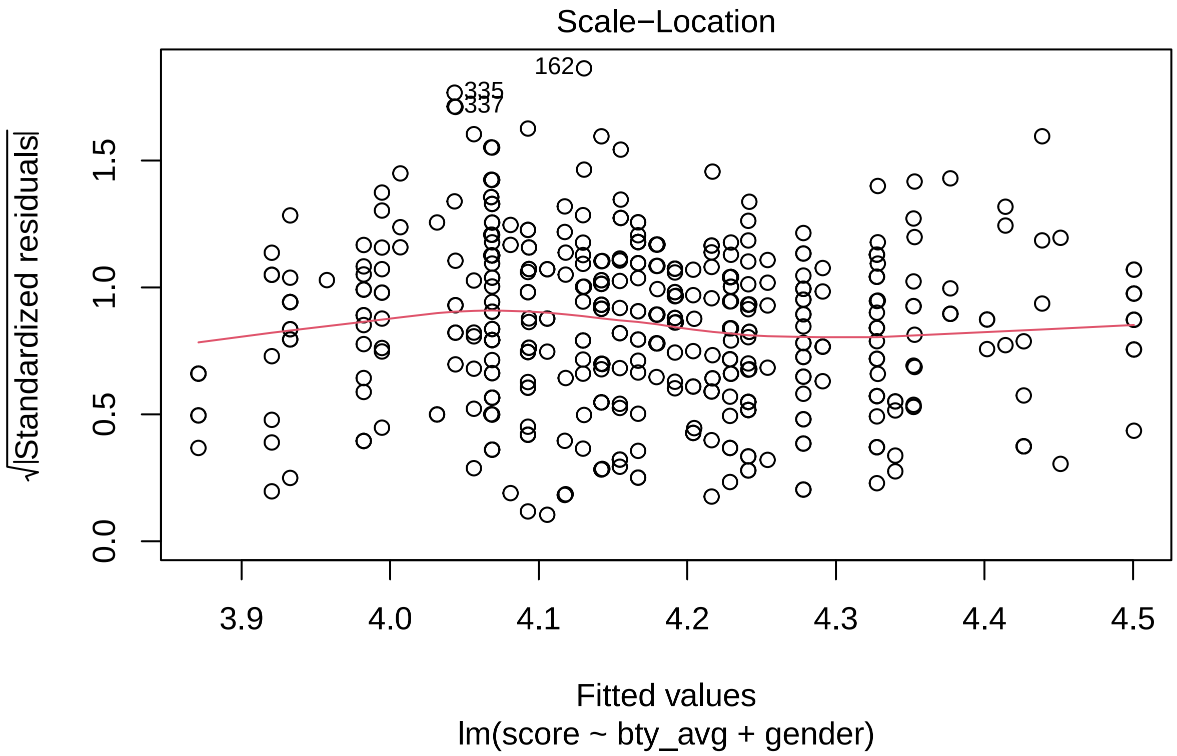
```
# Normal Probability Plot
qqnorm(m_bty_gen$residuals)
qqline(m_bty_gen$residuals)
```

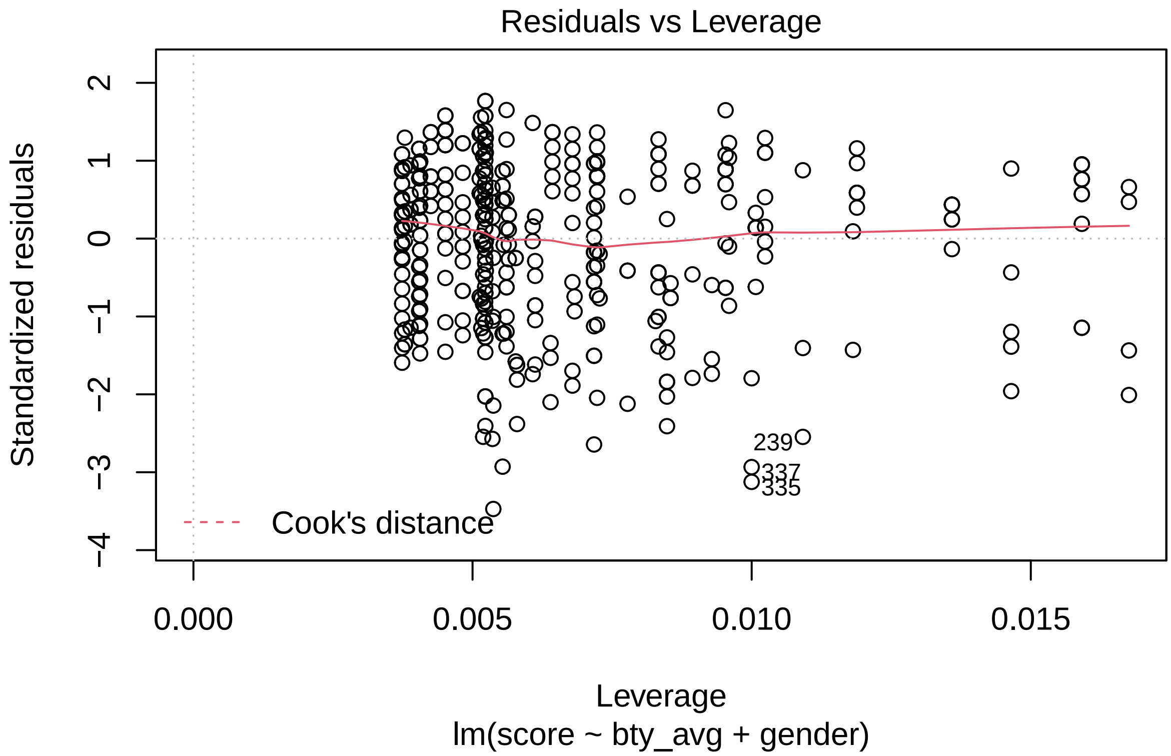


```
#Resiual vs Fitted, Normal Probability Plot, Scale-Location, Residual vs Leverage
plot(m_bty_gen)
```

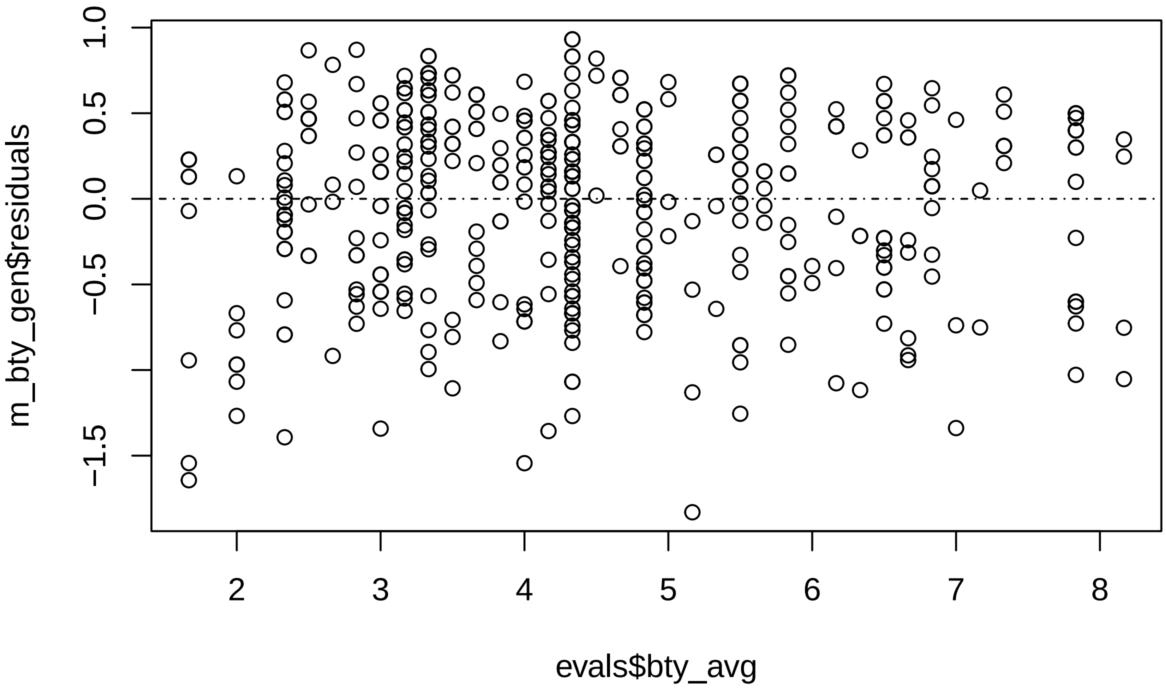

Residuals vs Fitted
Residuals
1.0
0.5
0.0
−1.0
−2.0
3.9
4.0
4.1
4.2
4.3
4.4
4.5
40
335
162
Fitted values
lm(score ~ bty_avg + gender)


Normal Q−Q
Standardized residuals
2
1
0
−1
−2
−3
337
335
162
−3
−2
−1
0
1
2
3
Theoretical Quantiles
lm(score ~ bty_avg + gender)

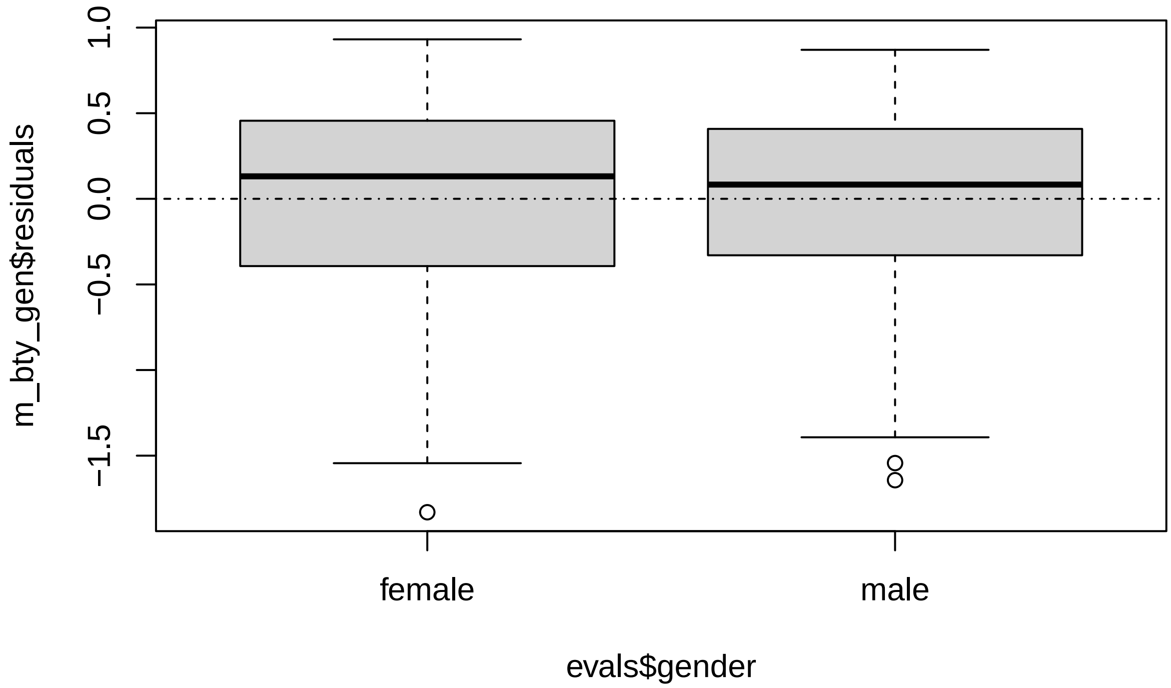

Scale−Location
162
335
337
√|Standardized residuals|
0.0
0.5
1.0
1.5
3.9
4.0
4.1
4.2
4.3
4.4
4.5
Fitted values
lm(score ~ bty_avg + gender)



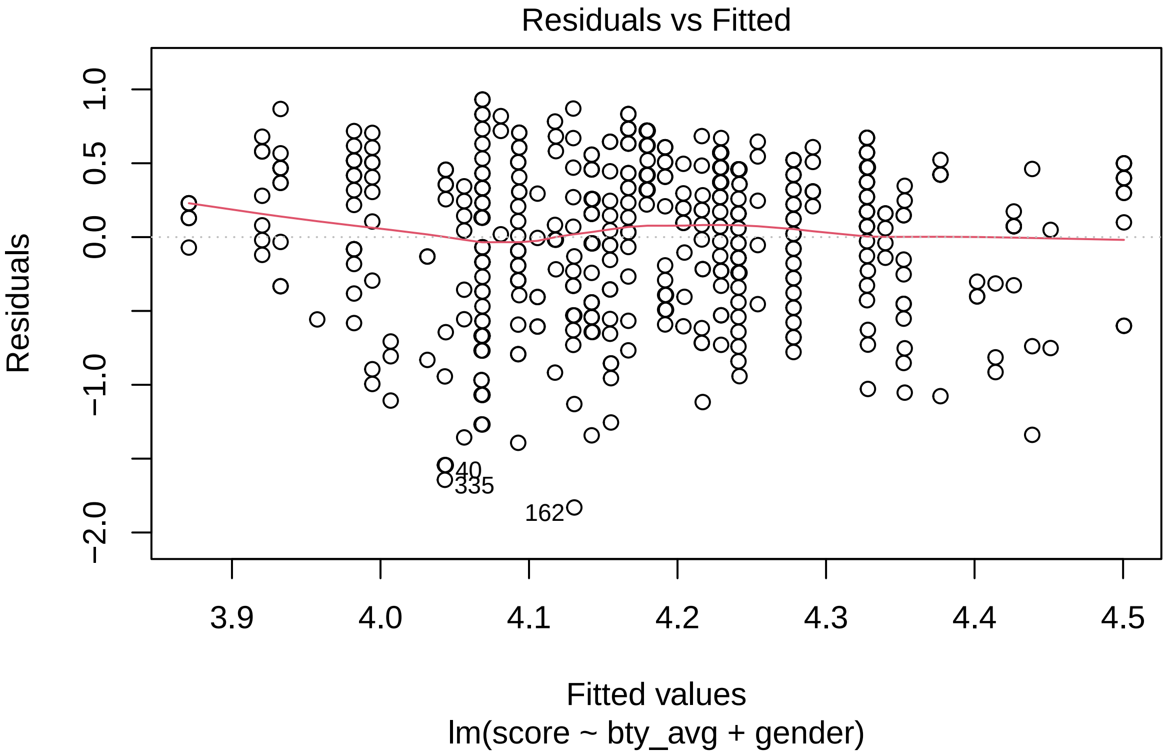
```
# residual plot against each predictor variable
plot(m_bty_gen$residuals ~ evals$bty_avg)
abline(h = 0, lty = 4)  # adds a horizontal dashed line at y = 0
```

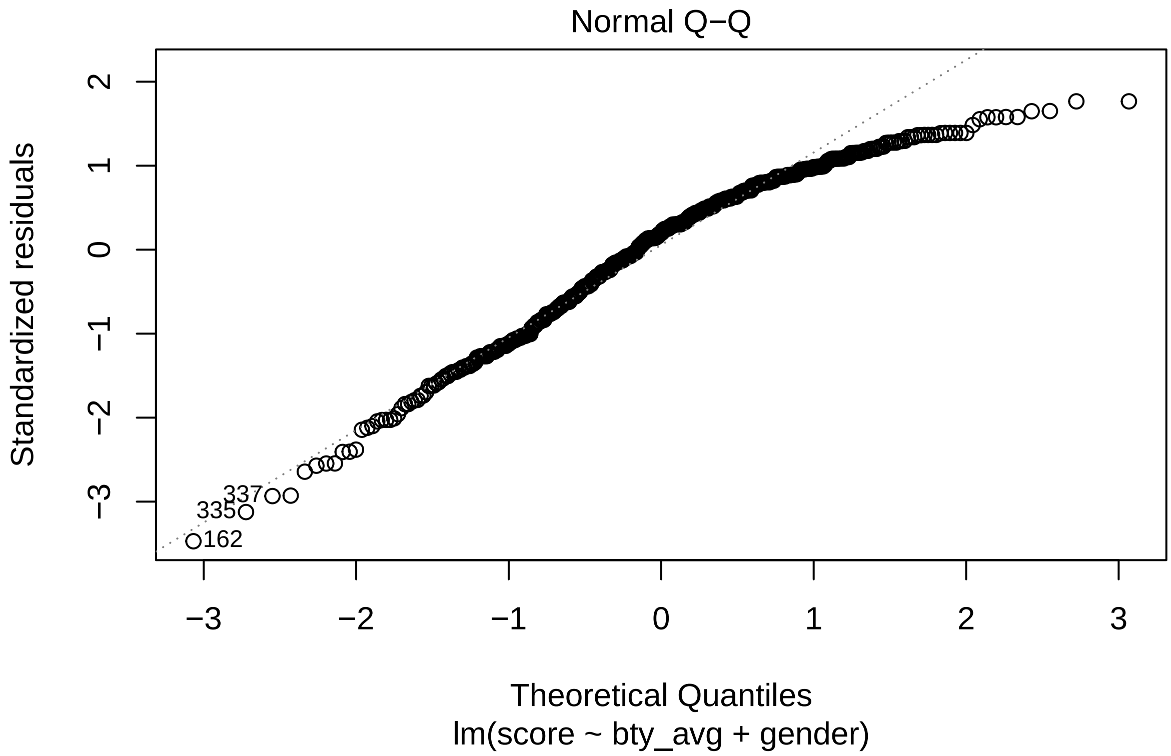


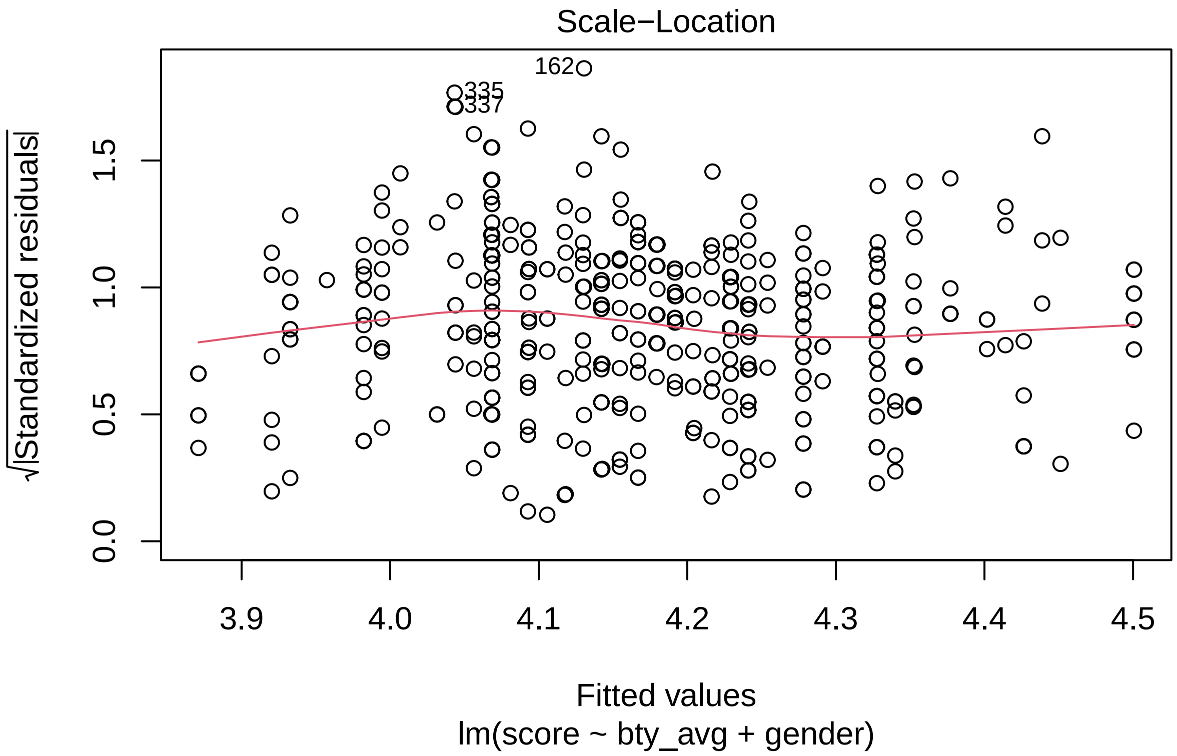
```
plot(m_bty_gen$residuals ~ evals$gender)
abline(h = 0, lty = 4)  # adds a horizontal dashed line at y = 0
```

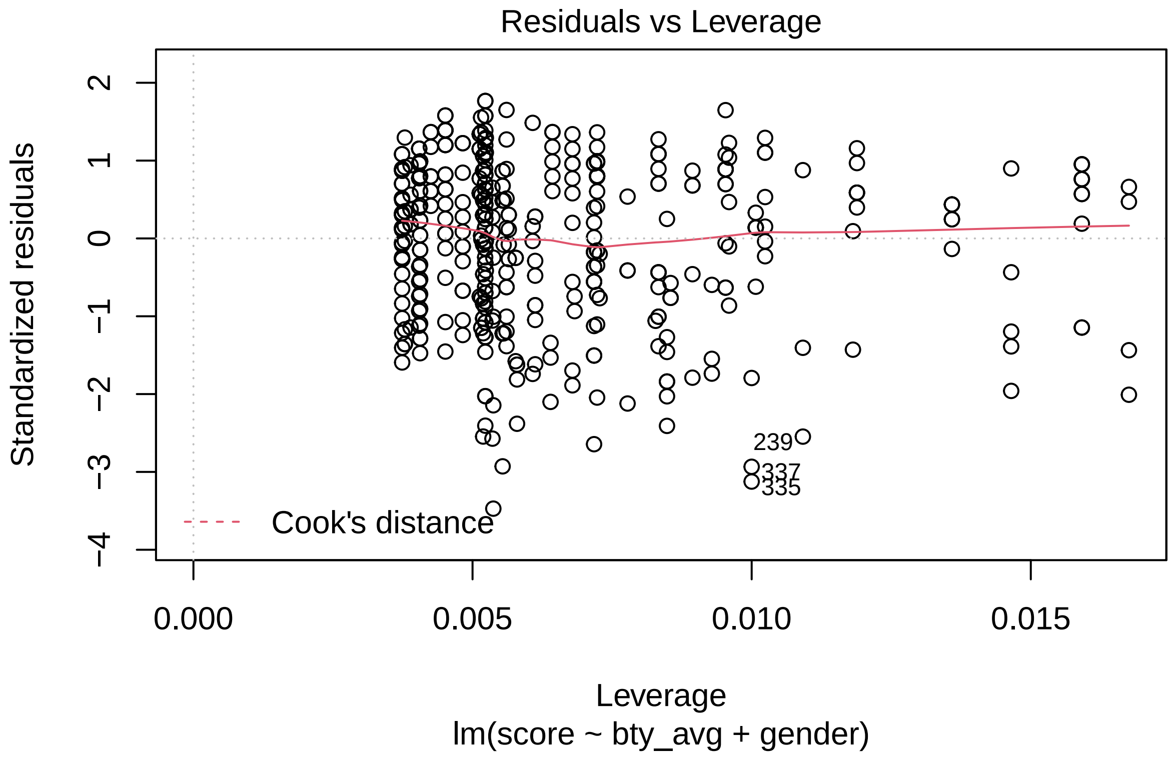


```
#Resiual vs Fitted, Normal Probability Plot, Scale-Location, Residual vs Leverage
plot(m_bty_gen)
```

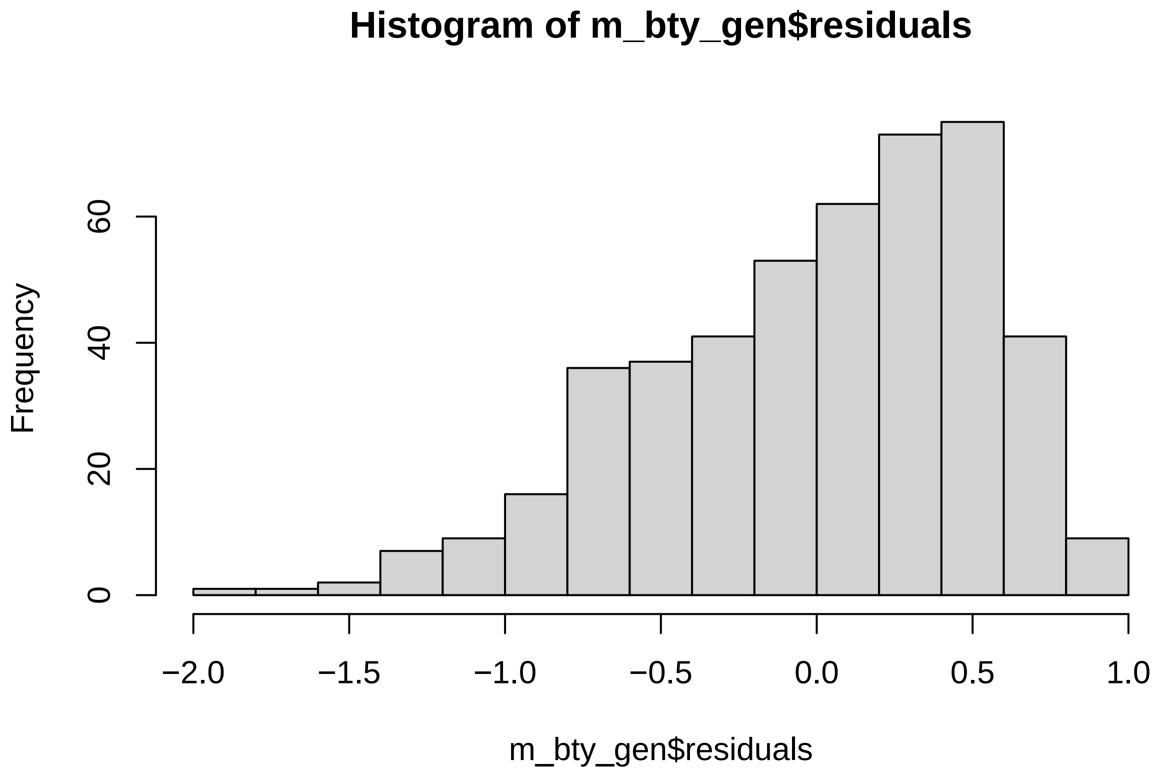

Residuals vs Fitted
Residuals
1.0
0.5
0.0
−1.0
−2.0
3.9
4.0
4.1
4.2
4.3
4.4
4.5
40
335
162
Fitted values
lm(score ~ bty_avg + gender)


Normal Q−Q
Standardized residuals
2
1
0
−1
−2
−3
337
335
162
−3
−2
−1
0
1
2
3
Theoretical Quantiles
lm(score ~ bty_avg + gender)

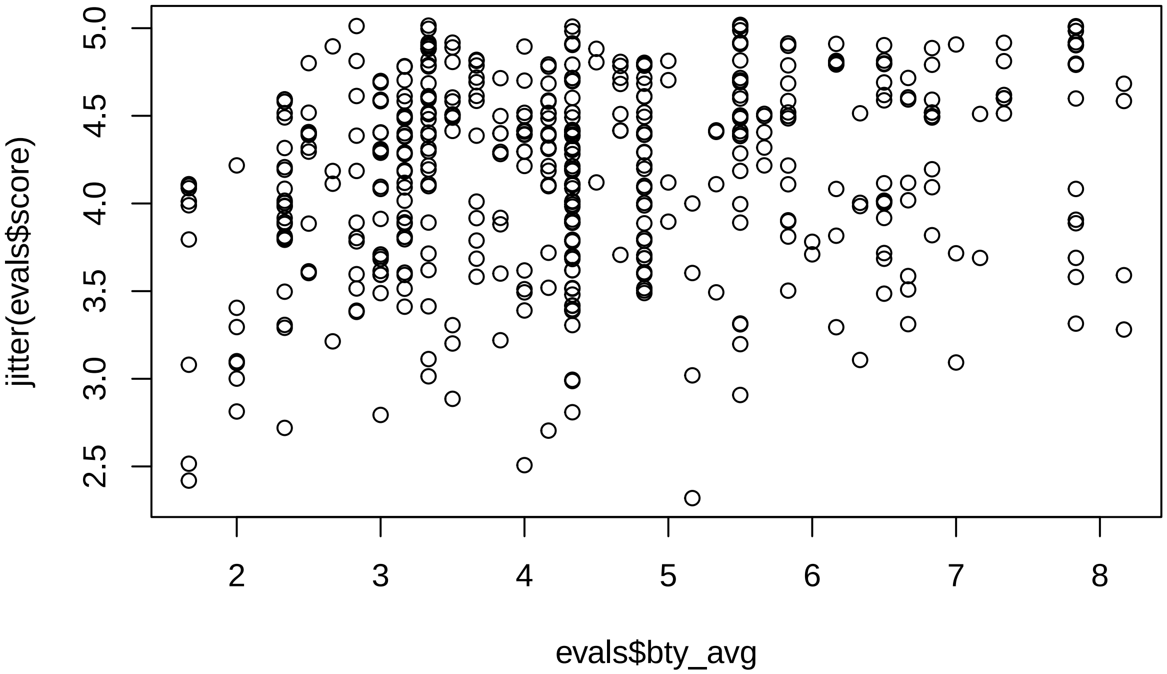

Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
162
335
337
3.9
4.0
4.1
4.2
4.3
4.4
4.5
Fitted values
lm(score ~ bty_avg + gender)



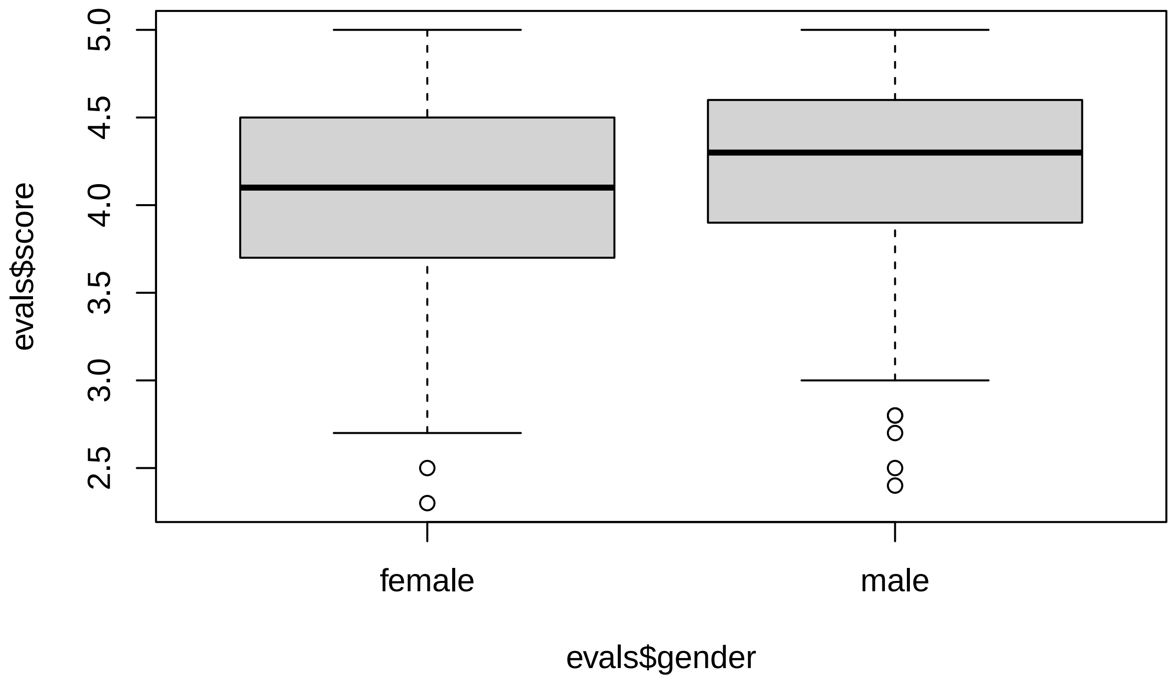
```
#Historgream
hist(m_bty_gen$residuals)
```



```
# Checking linearlidity
plot(jitter(evals$score) ~ evals$bty_avg)
```



```
plot(evals$score ~ evals$gender)
```



The histogram of residuals suggests that the residuals distribution is slightly skewed to the left.

The residuals do not follow the lines for upper quadriles in the Normal Probability Plot for residuals, .

Residuals vs Fitted, show that it appears to be constant variability for residuals. But as was established in the previous exercises, there is a linear relationship between beauty average and teaching evaluation score.

Exercise 8

Is bty_avg still a significant predictor of score? Has the addition of gender to the model changed the parameter estimate for bty_avg?

```
## Note that the estimate for gender is now called gendermale. You’ll see this name change whenever you

## As a result, for female professors, the parameter estimate is multiplied by zero, leaving the interc
```

Exercise 9

What is the equation of the line corresponding to males? (Hint: For males, the parameter estimate is multiplied by 1.) For two professors who received the same beauty rating, which gender tends to have the higher course evaluation score?

```
## score =3.74734+0.07416 * beauty_avg+0.17239*gender_male

## For gender = Male, we will evaluate the equation with gender_male = 1. In case, of female gender, we
## score=3.74734+0.07416*beauty_avg+0.17239

## Male professor will have a evaluation score higher by 0.17239 all other things being equal.
```

Exercise 10

Create a new model called m_bty_rank with gender removed and rank added in. How does R appear to handle categorical variables that have more than two levels? Note that the rank variable has three levels: teaching, tenure track, tenured.

```
m_bty_rank <- lm(score ~ bty_avg + rank, data = evals)
summary(m_bty_rank)
```

```
##
## Call:
## lm(formula = score ~ bty_avg + rank, data = evals)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.8713 -0.3642  0.1489  0.4103  0.9525
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       3.98155    0.09078  43.860  < 2e-16 ***
## bty_avg           0.06783    0.01655   4.098 4.92e-05 ***
## ranktenure track -0.16070    0.07395  -2.173   0.0303 *
## ranktenured      -0.12623    0.06266  -2.014   0.0445 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5328 on 459 degrees of freedom
## Multiple R-squared:  0.04652,    Adjusted R-squared:  0.04029
## F-statistic: 7.465 on 3 and 459 DF,  p-value: 6.88e-05
```

Exercise 11

Which variable would you expect to have the highest p-value in this model? Why? Hint: Think about which variable would you expect to not have any association with the professor score.

```
m_full <- lm(score ~ rank + ethnicity + gender + language + age + cls_perc_eval
             + cls_students + cls_level + cls_profs + cls_credits + bty_avg
             + pic_outfit + pic_color, data = evals)
summary(m_full)
```

```
##
## Call:
## lm(formula = score ~ rank + ethnicity + gender + language + age +
##     cls_perc_eval + cls_students + cls_level + cls_profs + cls_credits +
##     bty_avg + pic_outfit + pic_color, data = evals)
##
```

```
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -1.77397 -0.32432  0.09067  0.35183  0.95036 
## 
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)    
## (Intercept)            4.0952141  0.2905277  14.096  < 2e-16 ***
## ranktenure track      -0.1475932  0.0820671  -1.798  0.07278 .  
## ranktenured           -0.0973378  0.0663296  -1.467  0.14295    
## ethnicitynot minority  0.1234929  0.0786273   1.571  0.11698    
## gendermale             0.2109481  0.0518230   4.071 5.54e-05 ***
## languagenon-english   -0.2298112  0.1113754  -2.063  0.03965 *  
## age                   -0.0090072  0.0031359  -2.872  0.00427 ** 
## cls_perc_eval          0.0053272  0.0015393   3.461  0.00059 ***
## cls_students           0.0004546  0.0003774   1.205  0.22896    
## cls_levelupper         0.0605140  0.0575617   1.051  0.29369    
## cls_profssingle       -0.0146619  0.0519885  -0.282  0.77806    
## cls_creditsone credit  0.5020432  0.1159388   4.330 1.84e-05 ***
## bty_avg                0.0400333  0.0175064   2.287  0.02267 *  
## pic_outfitnot formal  -0.1126817  0.0738800  -1.525  0.12792    
## pic_colorcolor        -0.2172630  0.0715021  -3.039  0.00252 ** 
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.498 on 448 degrees of freedom
## Multiple R-squared:  0.1871, Adjusted R-squared:  0.1617 
## F-statistic: 7.366 on 14 and 448 DF,  p-value: 6.552e-14
```

Exercise 12

Check your suspicions from the previous exercise. Include the model output in your response.

```
## The "number of professors" (cls_profs) as the variable to have the least assoication with the profes
```

Exercise 13

Check your suspicions from the previous exercise. Include the model output in your response.

```
## All other things being equal, Evaluation for professor that not minority tends to be 0.1234929 highe
```

Exercise 14

Drop the variable with the highest p-value and re-fit the model. Did the coefficients and significance of the other explanatory variables change? (One of the things that makes multiple regression interesting is that coefficient estimates depend on the other variables that are included in the model.) If not, what does this say about whether or not the dropped variable was collinear with the other explanatory variables?

```
m_full_1 <- lm(score ~ rank + ethnicity + gender + language + age + cls_perc_eval
             + cls_students + cls_level + cls_credits + bty_avg
             + pic_outfit + pic_color, data = evals)
summary(m_full_1)
```

```
##
## Call:
## lm(formula = score ~ rank + ethnicity + gender + language + age +
##     cls_perc_eval + cls_students + cls_level + cls_credits +
##     bty_avg + pic_outfit + pic_color, data = evals)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.7836 -0.3257  0.0859  0.3513  0.9551
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)            4.0872523  0.2888562  14.150  < 2e-16 ***
## ranktenure track      -0.1476746  0.0819824  -1.801 0.072327 .
## ranktenured           -0.0973829  0.0662614  -1.470 0.142349
## ethnicitynot minority  0.1274458  0.0772887   1.649 0.099856 .
## gendermale             0.2101231  0.0516873   4.065 5.66e-05 ***
## languagenon-english   -0.2282894  0.1111305  -2.054 0.040530 *
## age                   -0.0089992  0.0031326  -2.873 0.004262 **
## cls_perc_eval          0.0052888  0.0015317   3.453 0.000607 ***
## cls_students           0.0004687  0.0003737   1.254 0.210384
## cls_levelupper         0.0606374  0.0575010   1.055 0.292200
## cls_creditsone credit  0.5061196  0.1149163   4.404 1.33e-05 ***
## bty_avg                0.0398629  0.0174780   2.281 0.023032 *
## pic_outfitnot formal  -0.1083227  0.0721711  -1.501 0.134080
## pic_colorcolor        -0.2190527  0.0711469  -3.079 0.002205 **
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.4974 on 449 degrees of freedom
## Multiple R-squared:  0.187,  Adjusted R-squared:  0.1634
## F-statistic: 7.943 on 13 and 449 DF,  p-value: 2.336e-14
```

Exercise 15

Using backward-selection and p-value as the selection criterion, determine the best model. You do not need to show all steps in your answer, just the output for the final model. Also, write out the linear model for predicting score based on the final model you settle on.

```
m_full_best <- lm(score ~ ethnicity + gender + language + age + cls_perc_eval
             +   cls_credits + bty_avg + pic_color, data = evals)
summary(m_full_best)
```

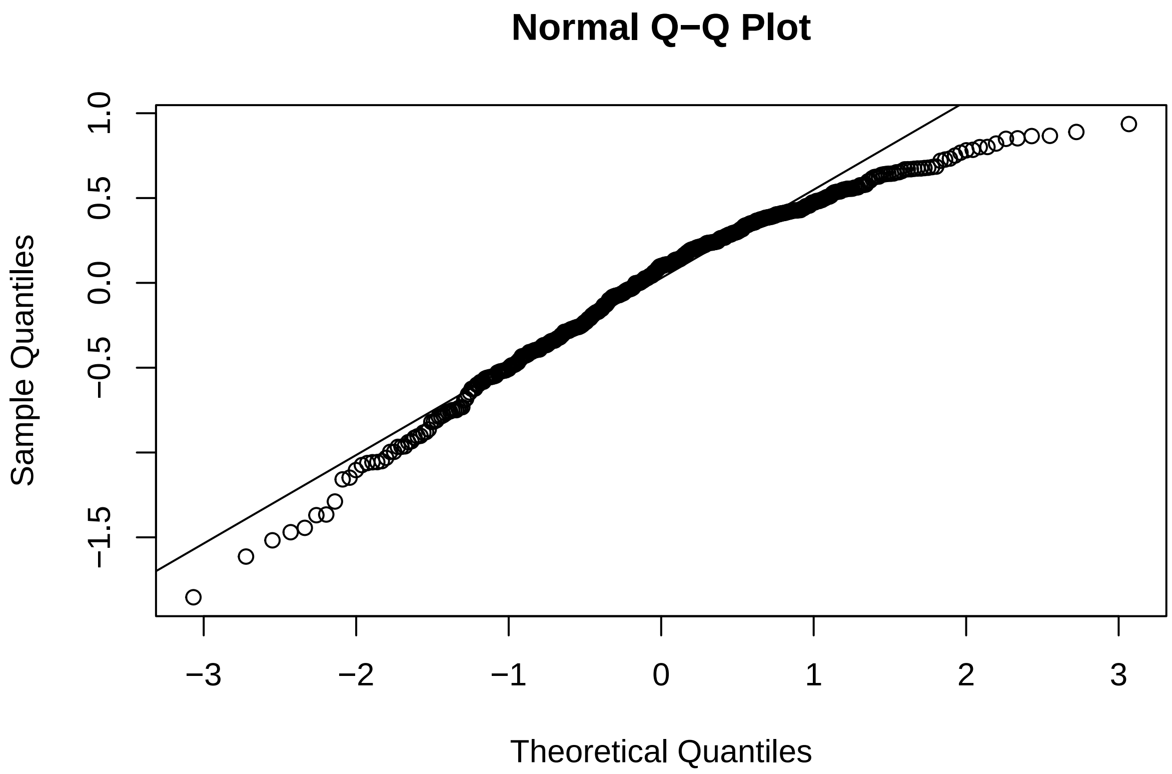
```
##
## Call:
## lm(formula = score ~ ethnicity + gender + language + age + cls_perc_eval +
##     cls_credits + bty_avg + pic_color, data = evals)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.85320 -0.32394  0.09984  0.37930  0.93610
##
## Coefficients:
```

```
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)             3.771922   0.232053  16.255  < 2e-16 ***
## ethnicitynot minority   0.167872   0.075275   2.230  0.02623 *
## gendermale              0.207112   0.050135   4.131 4.30e-05 ***
## languagenon-english    -0.206178   0.103639  -1.989  0.04726 *
## age                    -0.006046   0.002612  -2.315  0.02108 *
## cls_perc_eval           0.004656   0.001435   3.244  0.00127 **
## cls_creditsone credit   0.505306   0.104119   4.853 1.67e-06 ***
## bty_avg                 0.051069   0.016934   3.016  0.00271 **
## pic_colorcolor         -0.190579   0.067351  -2.830  0.00487 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4992 on 454 degrees of freedom
## Multiple R-squared:  0.1722, Adjusted R-squared:  0.1576
## F-statistic:  11.8 on 8 and 454 DF,  p-value: 2.58e-15
```

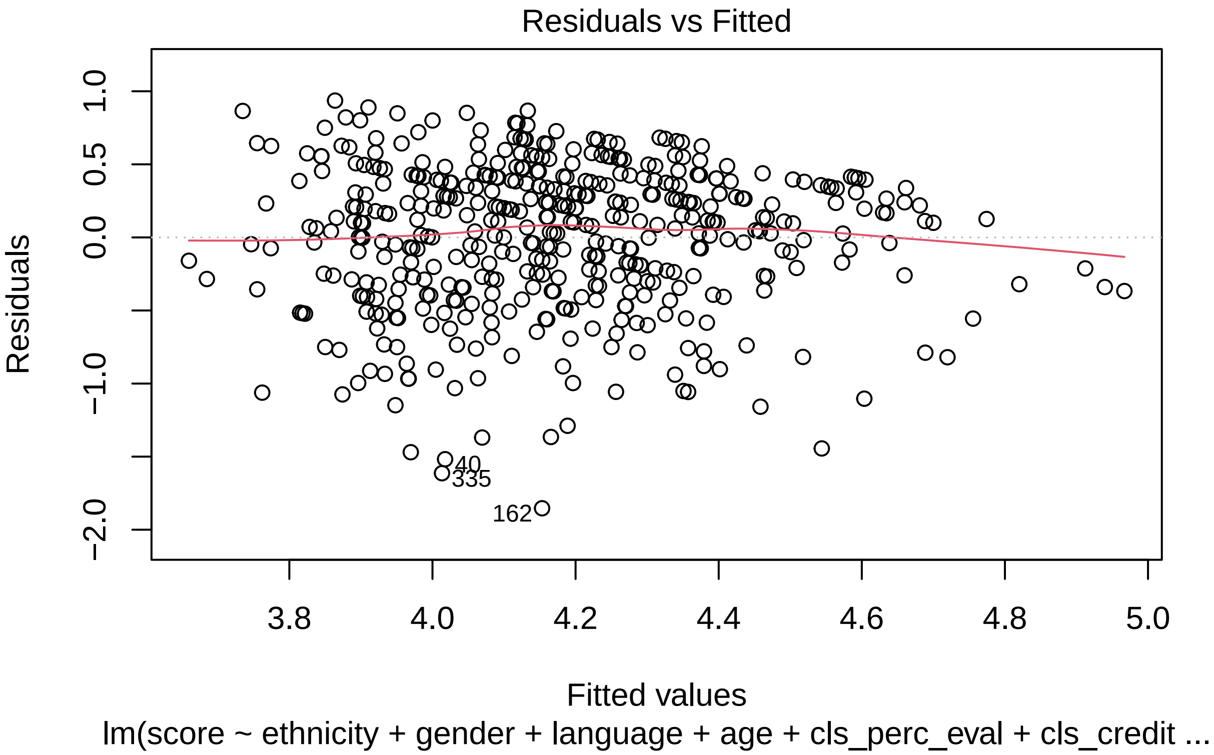
Exercise 16

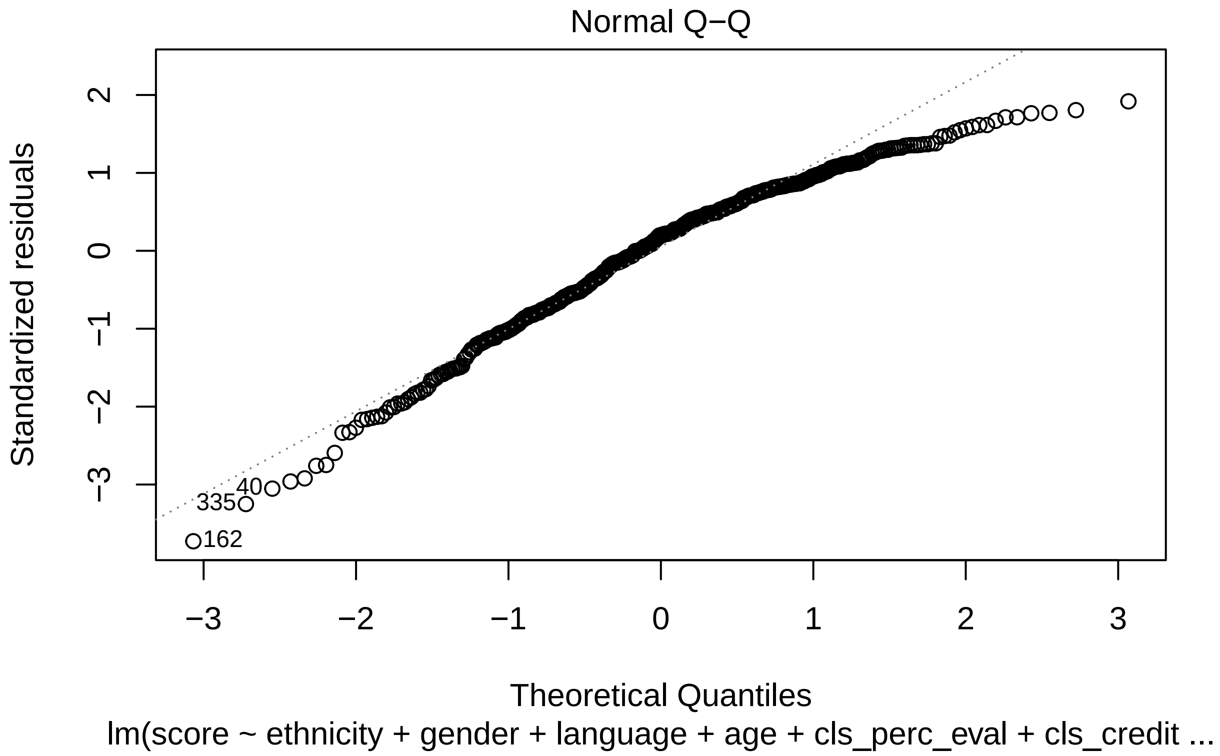
Verify that the conditions for this model are reasonable using diagnostic plots

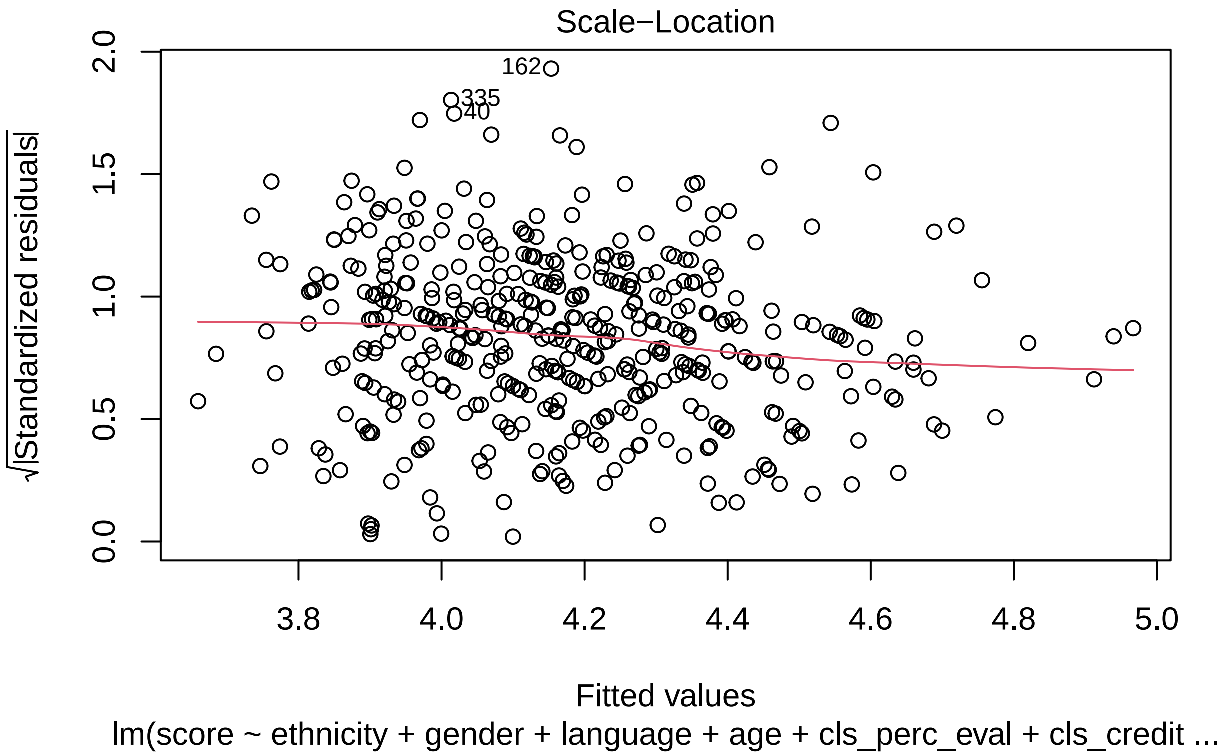
```
# Normal Probability Plot
qqnorm(m_full_best$residuals)
qqline(m_full_best$residuals)
```

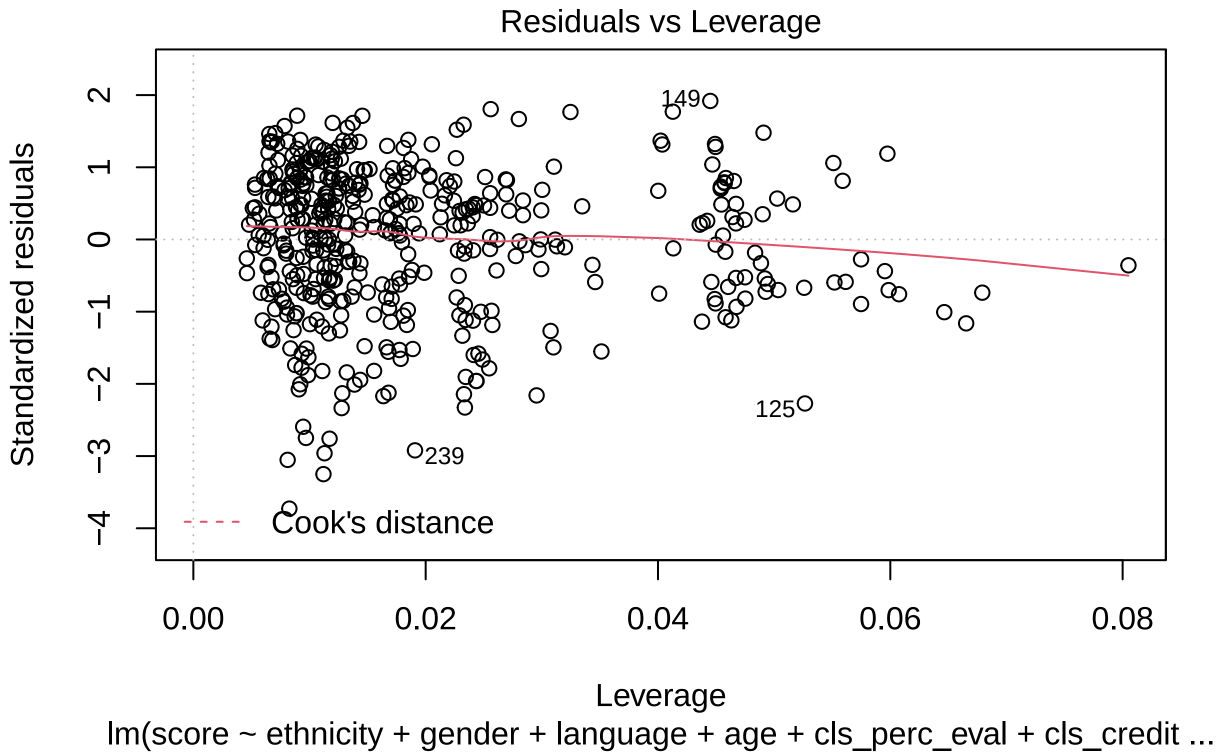


```
## Resiual vs Fitted, Normal Probability Plot, Scale-Location, Residual vs Leverage
plot(m_full_best)
```

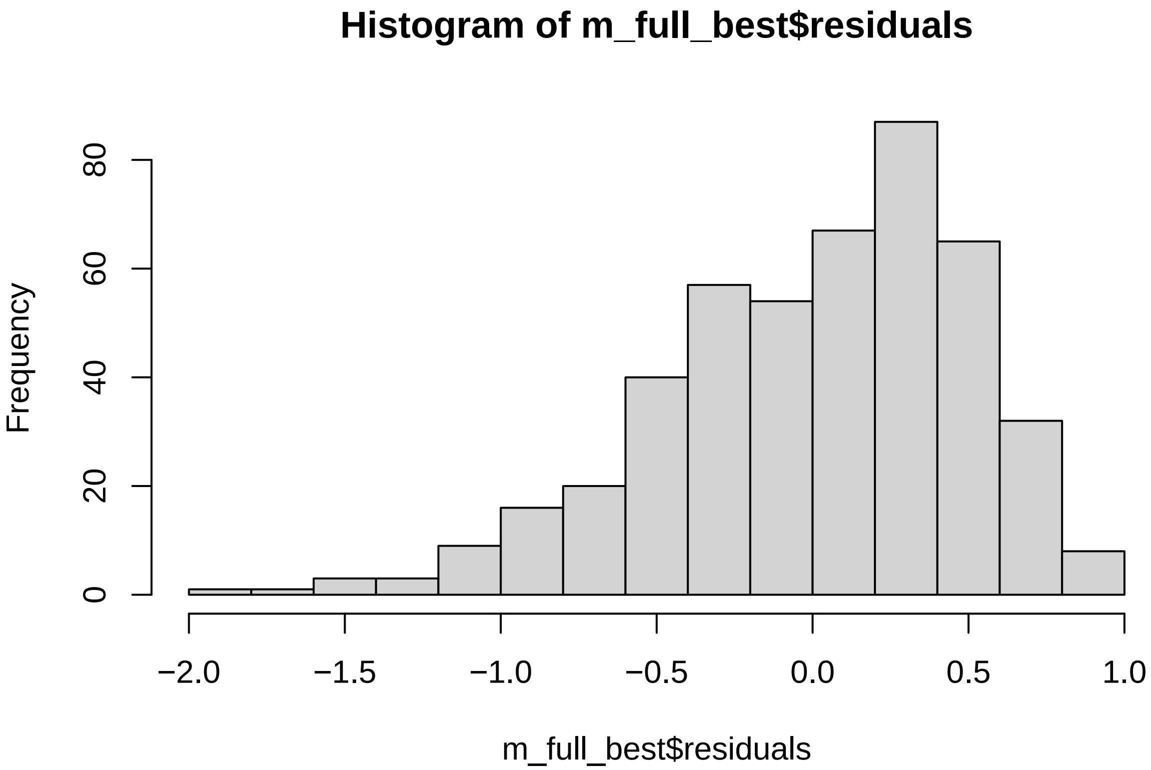



Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(score ~ ethnicity + gender + language + age + cls_perc_eval + cls_credit ...
−3
−2
−1
0
1
2
3
40
335
162

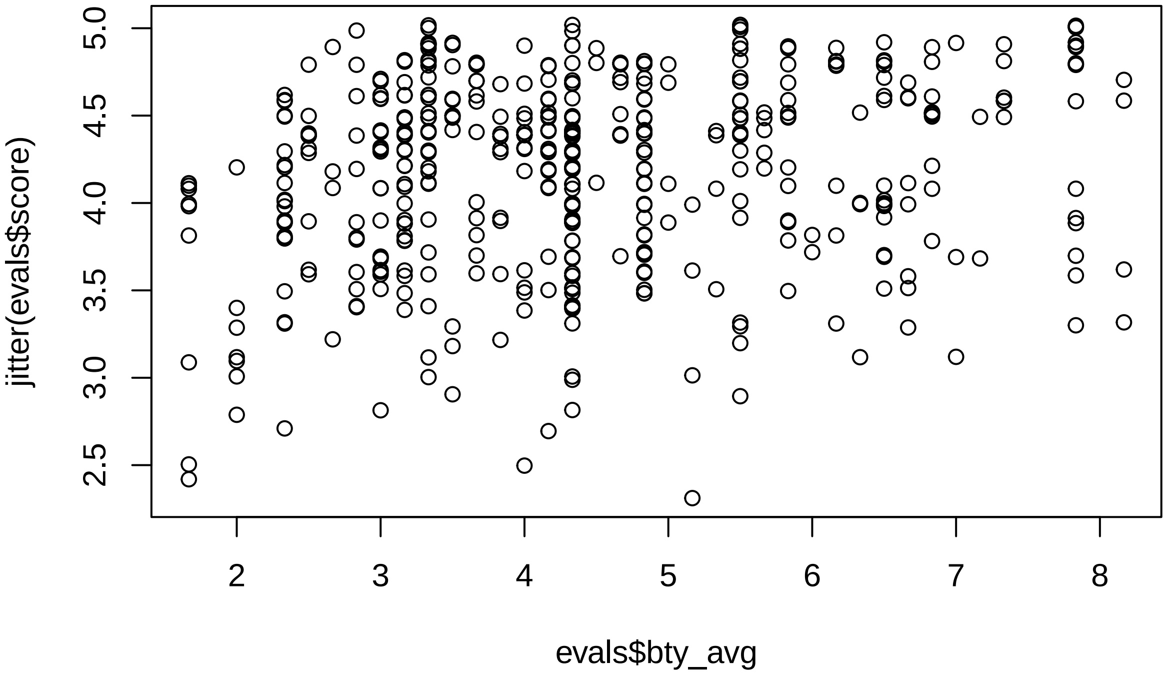

Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
3.8
4.0
4.2
4.4
4.6
4.8
5.0
162
335
40
Fitted values
lm(score ~ ethnicity + gender + language + age + cls_perc_eval + cls_credit ...



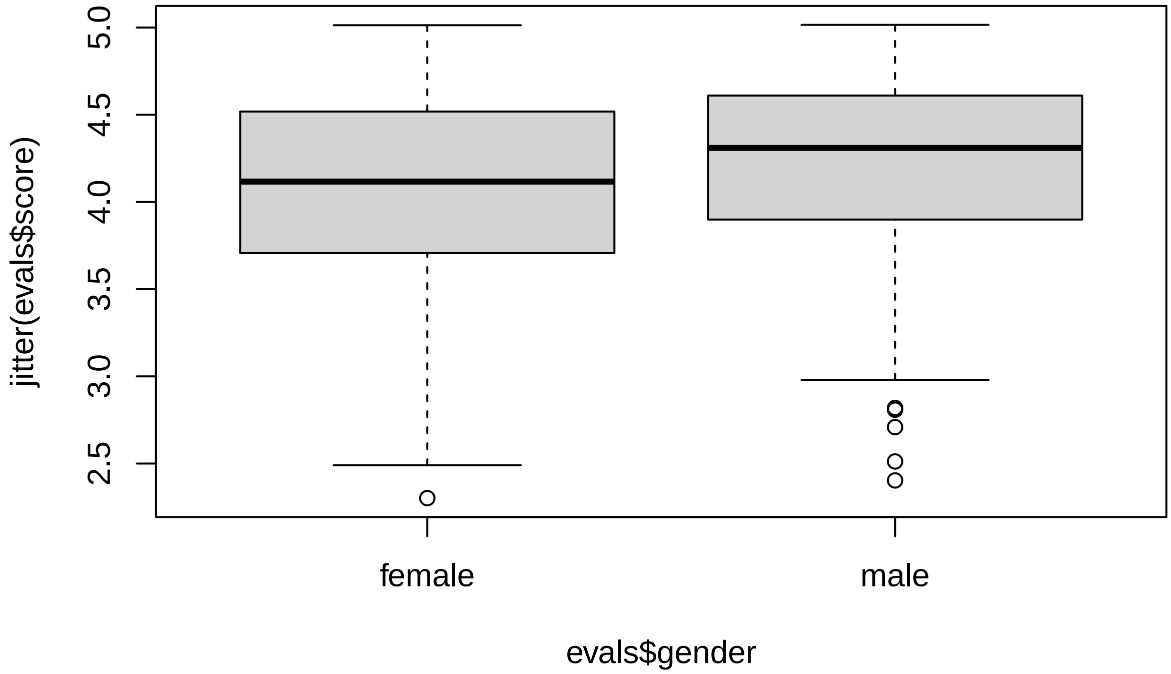
```
#Historgream
hist(m_full_best$residuals)
```



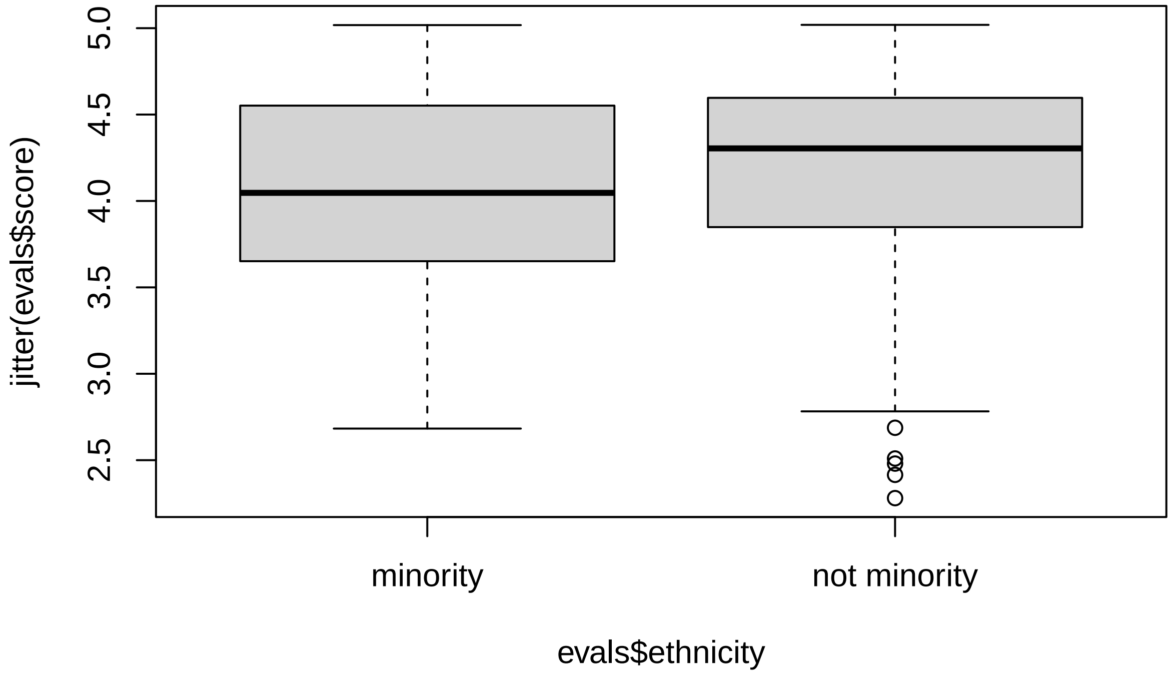
```
# Checking linearity
plot(jitter(evals$score) ~ evals$bty_avg)
```



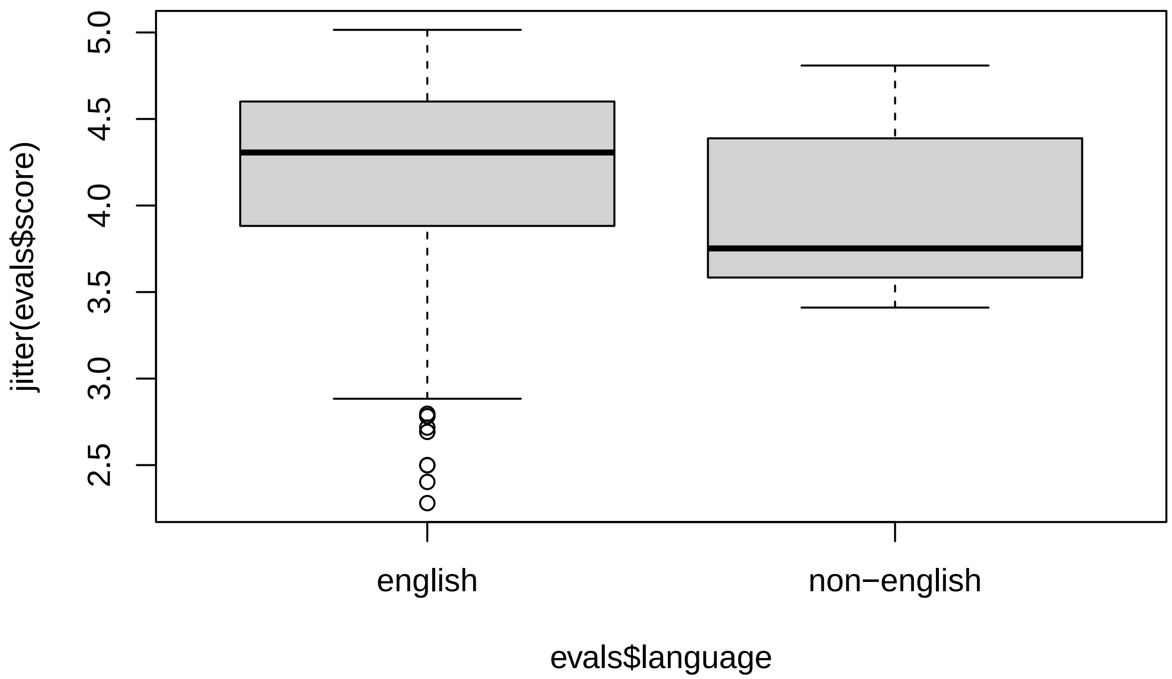
```
plot(jitter(evals$score) ~ evals$gender)
```



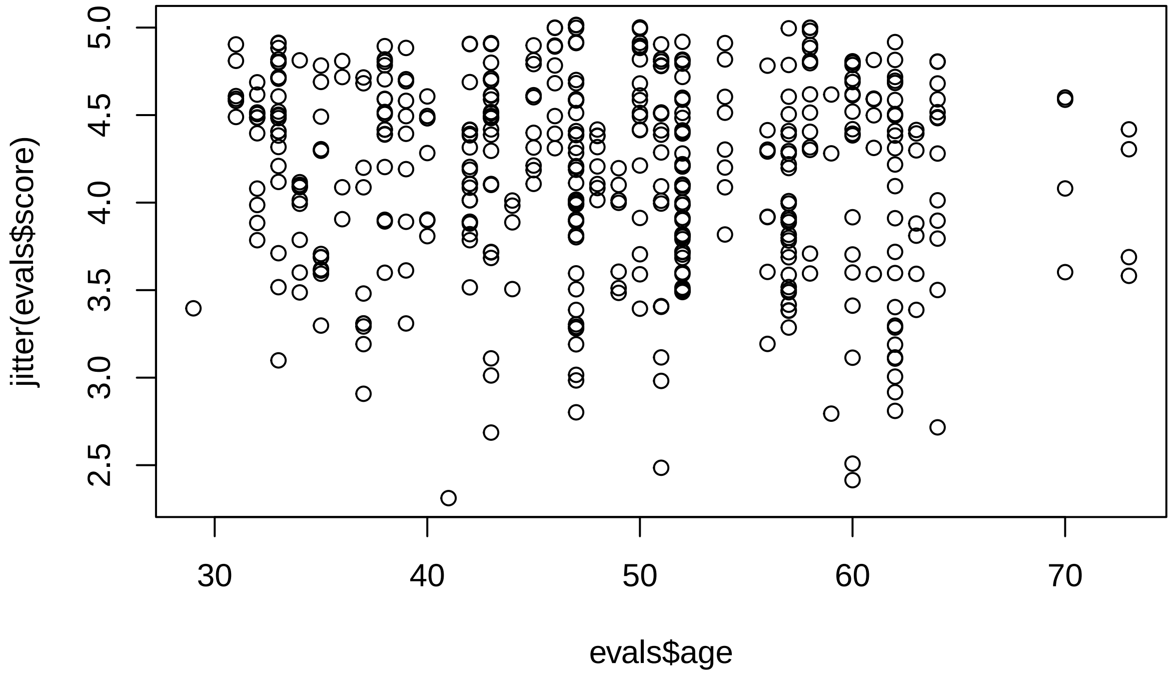
```
plot(jitter(evals$score) ~ evals$ethnicity)
```



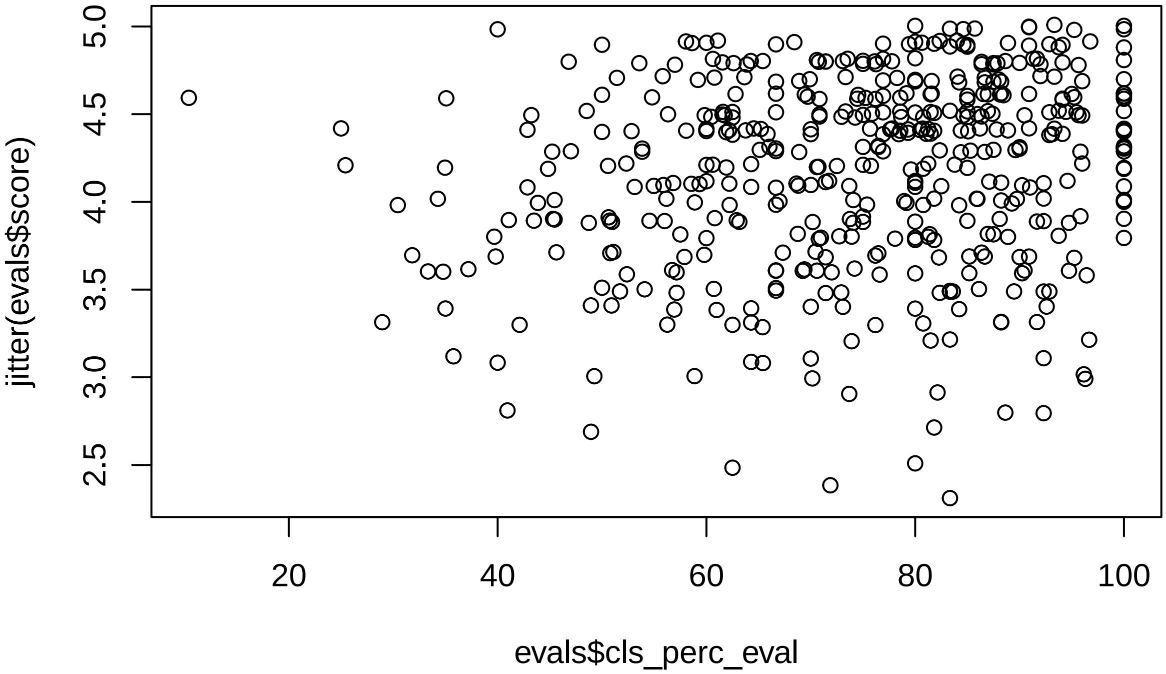
```
plot(jitter(evals$score) ~ evals$language)
```



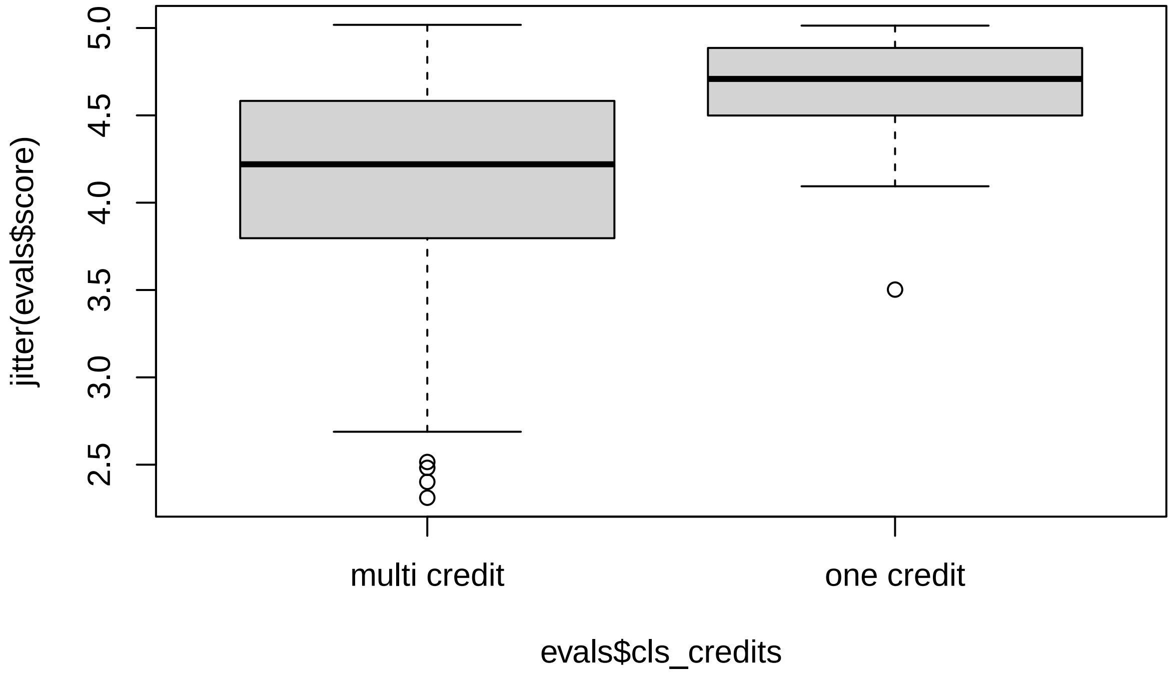
```
plot(jitter(evals$score) ~ evals$age)
```



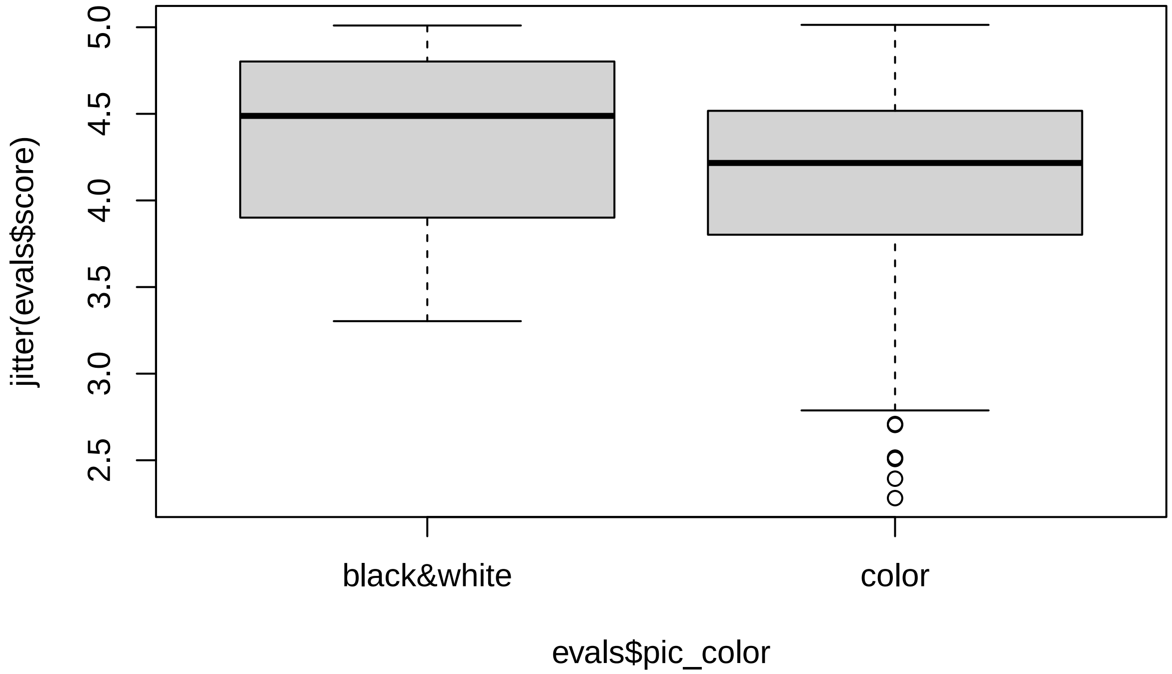
```
plot(jitter(evals$score) ~ evals$cls_perc_eval)
```



```
plot(jitter(evals$score) ~ evals$cls_credits)
```



```
plot(jitter(evals$score) ~ evals$pic_color)
```



## Exercise 17 The original paper describes how these data were gathered by taking a sample of professors from the University of Texas at Austin and including all courses that they have taught. Considering that each row represents a course, could this new information have an impact on any of the conditions of linear regression?

```
## No. Even if the course is being taught by the same professor, Class courses are independent of each
```

Exercise 18

Based on your final model, describe the characteristics of a professor and course at University of Texas at Austin that would be associated with a high evaluation score.

```
## Based on the coefficients Professor would be younger male teaching one credit class, he would not be
```

Exercise 19

Would you be comfortable generalizing your conclusions to apply to professors generally (at any university)? Why or why not?

```
## No, this was not conducted as an experiment but based on a sample in a given university. These resul
```