CUNY SPS DATA606 Lab9 - Multiple linear regression

Chinedu Onyeka

November 1st 2021

Grading the professor

Many college courses conclude by giving students the opportunity to evaluate the course and the instructor anonymously. However, the use of these student evaluations as an indicator of course quality and teaching effectiveness is often criticized because these measures may reflect the influence of non-teaching related characteristics, such as the physical appearance of the instructor. The article titled, "Beauty in the classroom: instructors' pulchritude and putative pedagogical productivity" by Hamermesh and Parker found that instructors who are viewed to be better looking receive higher instructional ratings.

Here, you will analyze the data from this study in order to learn what goes into a positive professor evaluation.

Getting Started

Load packages

In this lab, you will explore and visualize the data using the **tidyverse** suite of packages. The data can be found in the companion package for OpenIntro resources, **openintro**.

Let's load the packages.

```
library(tidyverse)
library(openintro)
library(GGally)
```

This is the first time we're using the `GGally` package. You will be using the `ggpairs` function from this package later in the lab.

The data

The data were gathered from end of semester student evaluations for a large sample of professors from the University of Texas at Austin. In addition, six students rated the professors' physical appearance. The result is a data frame where each row contains a different course and columns represent variables about the courses and professors. It's called `evals`.

```
glimpse(evals)
```

```
## Rows: 463
## Columns: 23
## $ course_id     <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 1~
## $ prof_id       <int> 1, 1, 1, 1, 2, 2, 2, 3, 3, 4, 4, 4, 4, 4, 4, 4, 4, 5, 5,~
## $ score         <dbl> 4.7, 4.1, 3.9, 4.8, 4.6, 4.3, 2.8, 4.1, 3.4, 4.5, 3.8, 4~
## $ rank          <fct> tenure track, tenure track, tenure track, tenure track, ~
## $ ethnicity     <fct> minority, minority, minority, minority, not minority, no~
## $ gender        <fct> female, female, female, female, male, male, male, male, ~
## $ language      <fct> english, english, english, english, english, english, en~
```

```
## $ age           <int> 36, 36, 36, 36, 59, 59, 59, 51, 51, 40, 40, 40, 40, 40, ~
## $ cls_perc_eval <dbl> 55.81395, 68.80000, 60.80000, 62.60163, 85.00000, 87.500~
## $ cls_did_eval  <int> 24, 86, 76, 77, 17, 35, 39, 55, 111, 40, 24, 24, 17, 14,~
## $ cls_students  <int> 43, 125, 125, 123, 20, 40, 44, 55, 195, 46, 27, 25, 20, ~
## $ cls_level     <fct> upper, upper, upper, upper, upper, upper, upper, upper, ~
## $ cls_profs     <fct> single, single, single, single, multiple, multiple, mult~
## $ cls_credits   <fct> multi credit, multi credit, multi credit, multi credit, ~
## $ bty_f1lower   <int> 5, 5, 5, 5, 4, 4, 4, 5, 5, 2, 2, 2, 2, 2, 2, 2, 2, 7, 7,~
## $ bty_f1upper   <int> 7, 7, 7, 7, 4, 4, 4, 2, 2, 5, 5, 5, 5, 5, 5, 5, 5, 9, 9,~
## $ bty_f2upper   <int> 6, 6, 6, 6, 2, 2, 2, 5, 5, 4, 4, 4, 4, 4, 4, 4, 4, 9, 9,~
## $ bty_m1lower   <int> 2, 2, 2, 2, 2, 2, 2, 2, 2, 3, 3, 3, 3, 3, 3, 3, 3, 7, 7,~
## $ bty_m1upper   <int> 4, 4, 4, 4, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 6, 6,~
## $ bty_m2upper   <int> 6, 6, 6, 6, 3, 3, 3, 3, 3, 2, 2, 2, 2, 2, 2, 2, 2, 6, 6,~
## $ bty_avg       <dbl> 5.000, 5.000, 5.000, 5.000, 3.000, 3.000, 3.000, 3.333, ~
## $ pic_outfit    <fct> not formal, not formal, not formal, not formal, not form~
## $ pic_color     <fct> color, color, color, color, color, color, color, color, ~
```

We have observations on 21 different variables, some categorical and some numerical. The meaning of each variable can be found by bringing up the help file:

```
?evals
```

Exploring the data

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.

Solution 1:

This is an observational study. Given the study design, it would not be possible to answer the question whether beauty leads directly to the differences in course evaluations as it is not an experiment. A better rephrased question for this study would be "Is there a relationship between beauty and course evaluations?"

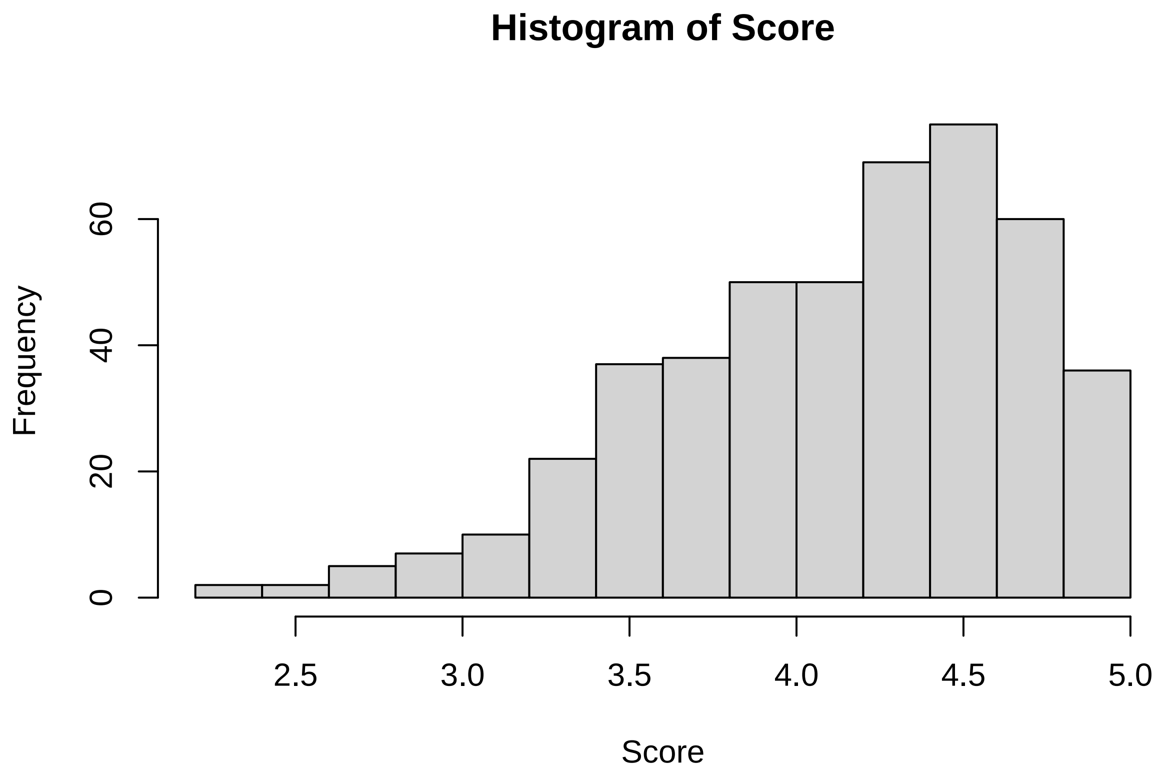
2. Describe the distribution of `score`. Is the distribution skewed? What does that tell you about how students rate courses? Is this what you expected to see? Why, or why not?

Solution 2:

```
# summary statistics for Scores
summary(evals$score)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   2.300   3.800   4.300   4.175   4.600   5.000
```

```
# histogram for score
hist(evals$score, xlab = "Score", main = "Histogram of Score")
```



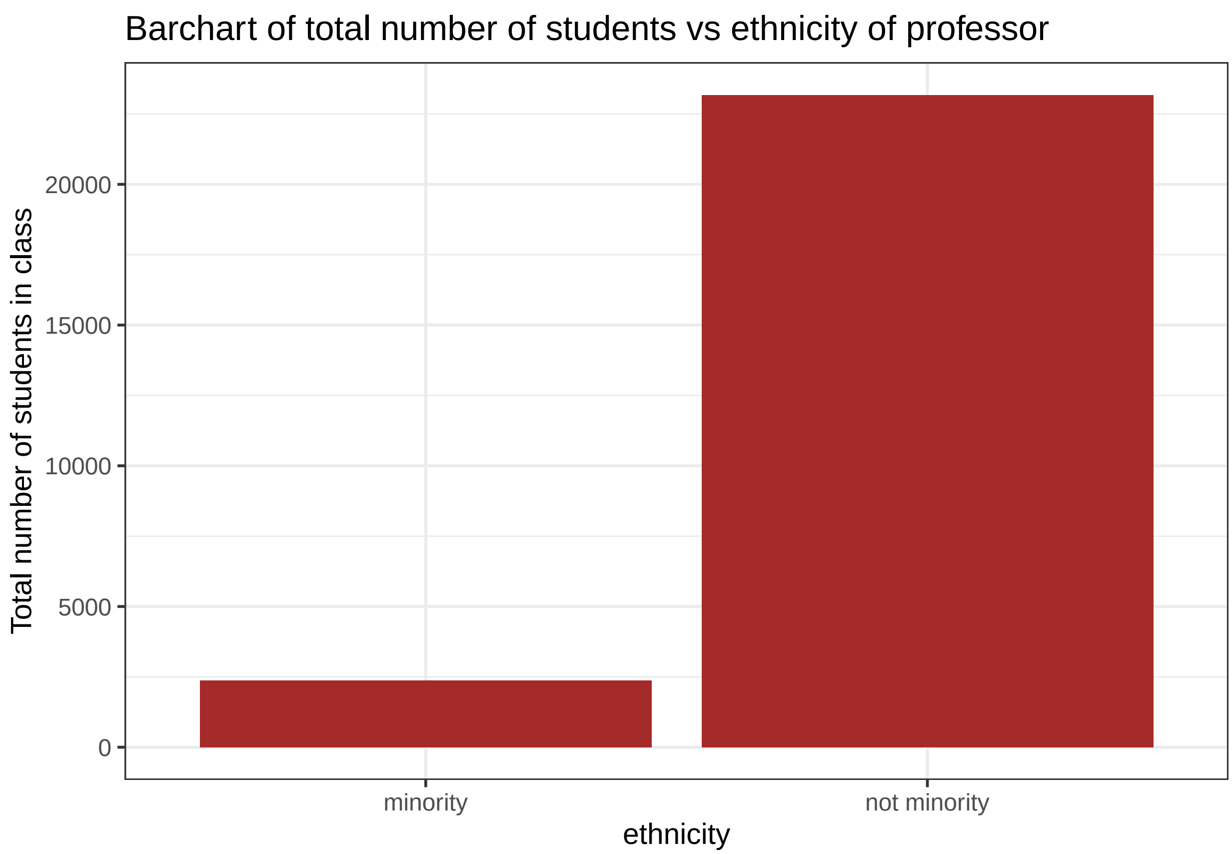
From the histogram, it can be seen that the distribution for `score` *is unimodal and left skewed. The distribution for scores has a mean of about 4.175 and median of 4.3. To a great extent, that is what I expect because I do not expect a lot of students at UT Austin to give low evaluation scores for their professors considering the reputation of the school. However, if the professors sampled in this study are the best professors, the data may not show the true picture of a greater portion of not so good professors.*

3. Excluding `score`, select two other variables and describe their relationship with each other using an appropriate visualization.

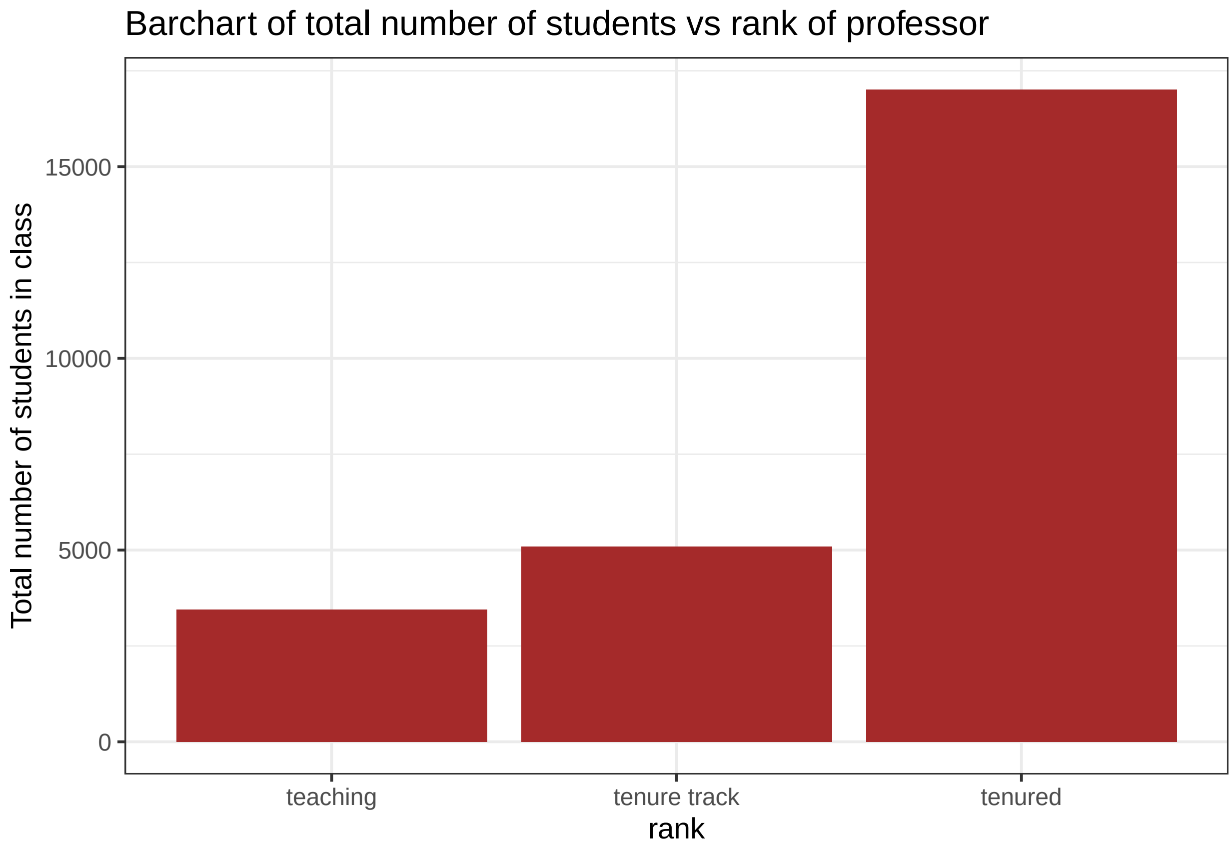
Solution 3:

```
#barplot(x = evals$rank, y = evals$ethnicity, color = "blue")
```

```
# barchart of total number of students vs ethnicity of professor
evals %>% ggplot(aes(x = ethnicity, y = cls_students)) + geom_col(fill = "brown") +
  theme_bw() + labs(title = "Barchart of total number of students vs ethnicity of professor") +
  ylab("Total number of students in class")
```



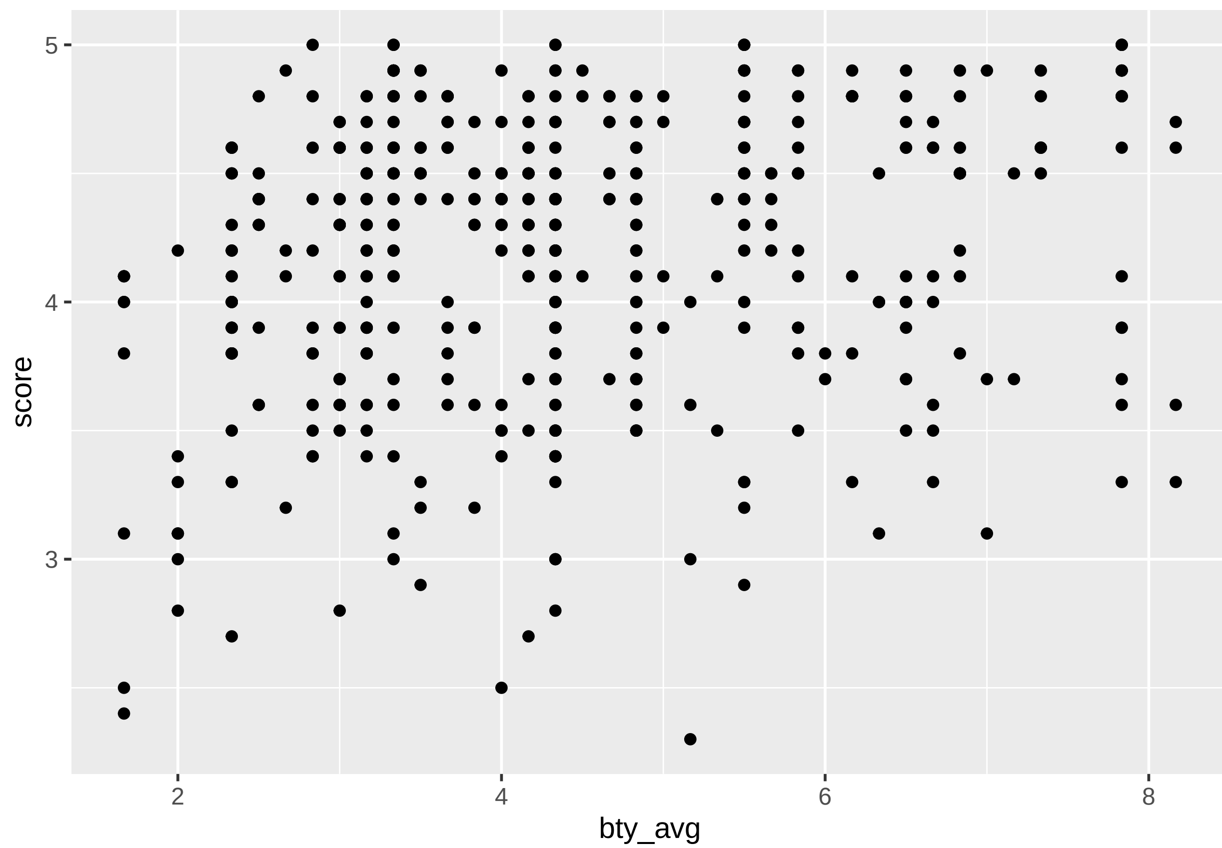

```
# barchart of total number of students vs rank of professor
evals %>% ggplot(aes(x = rank, y = cls_students)) + geom_col(fill = "brown") +
  theme_bw() + labs(title = "Barchart of total number of students vs rank of professor") +
  ylab("Total number of students in class")
```



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()
```

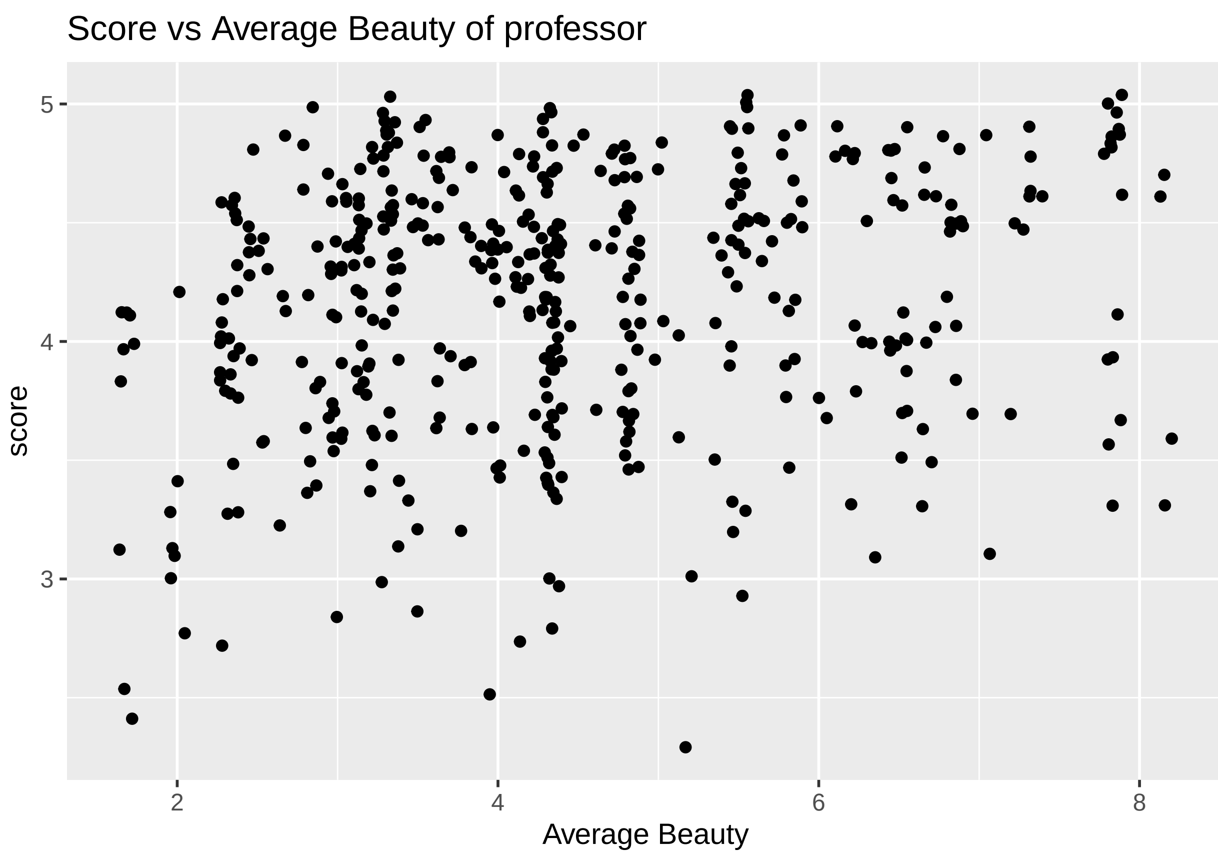


Before you draw conclusions about the trend, compare the number of observations in the data frame with the approximate number of points on the scatterplot. Is anything awry?

4. Replot the scatterplot, but this time use `geom_jitter` as your layer. What was misleading about the initial scatterplot?

Solution 4:

```
# plot with jitter
ggplot(data = evals, aes(x = bty_avg, y = score)) + geom_jitter() +
  labs(title = "Score vs Average Beauty of professor") + xlab("Average Beauty")
```



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. 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?

Solution 5:

```
m_bty <- lm(evals$score ~ evals$bty_avg) # fit a linear model of score vs bty_avg
summary(m_bty) # show the summary of the linear model
```

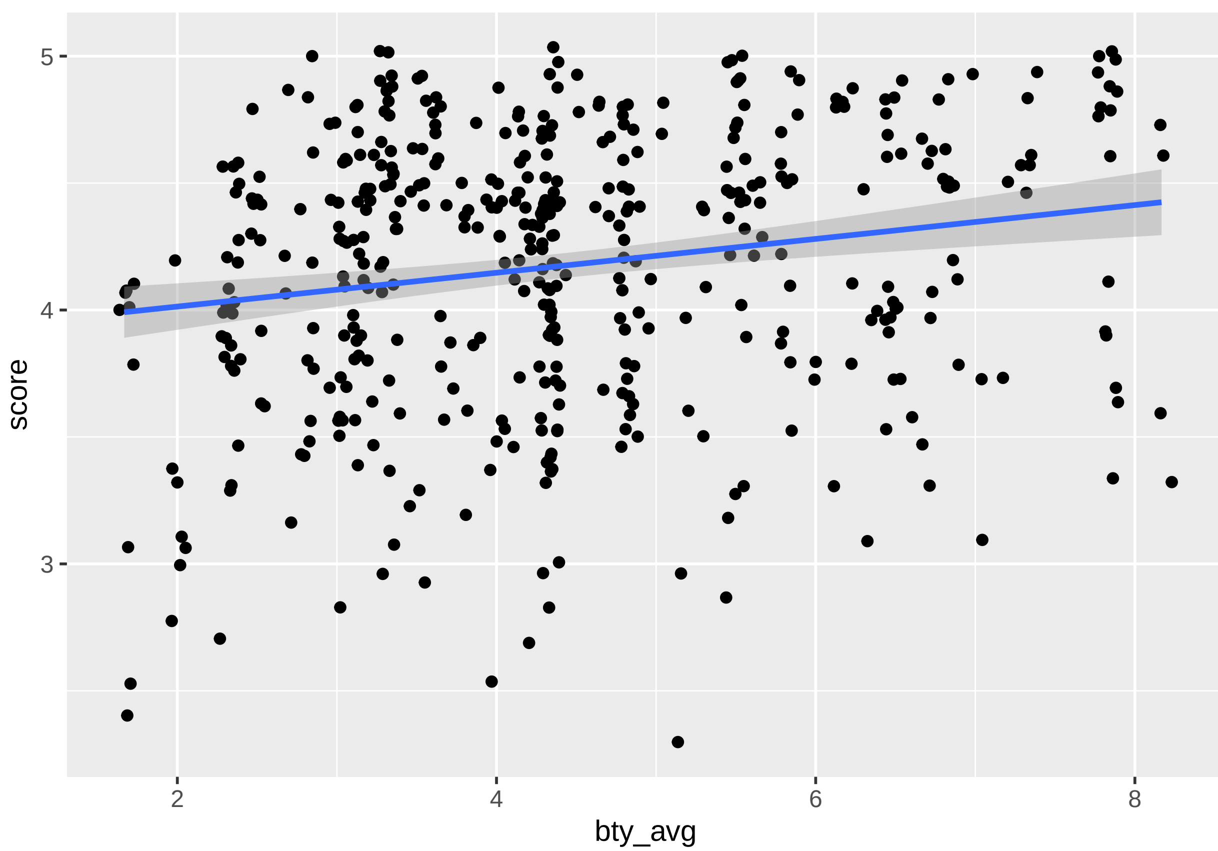
```
##
## Call:
## lm(formula = evals$score ~ evals$bty_avg)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.9246 -0.3690  0.1420  0.3977  0.9309
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    3.88034    0.07614   50.96  < 2e-16 ***
## evals$bty_avg  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
```

The equation of the linear model is $score = 3.88034 + 0.06664 * averageBeauty$
The slope of the linear model is the amount by which the score of the professor will increase for every unit increase in the average beauty of the professor according to the linear model.
The p value is less than zero, so we will reject the null hypothesis that the slope of the linear mode is zero. Hence, we can conclude that the slope of the model is a number other than zero which means that there is statistically significant relationship between the two variables. Therefore, the average beauty (bty_avg) appears to be a practically significant predictor of the professor score.

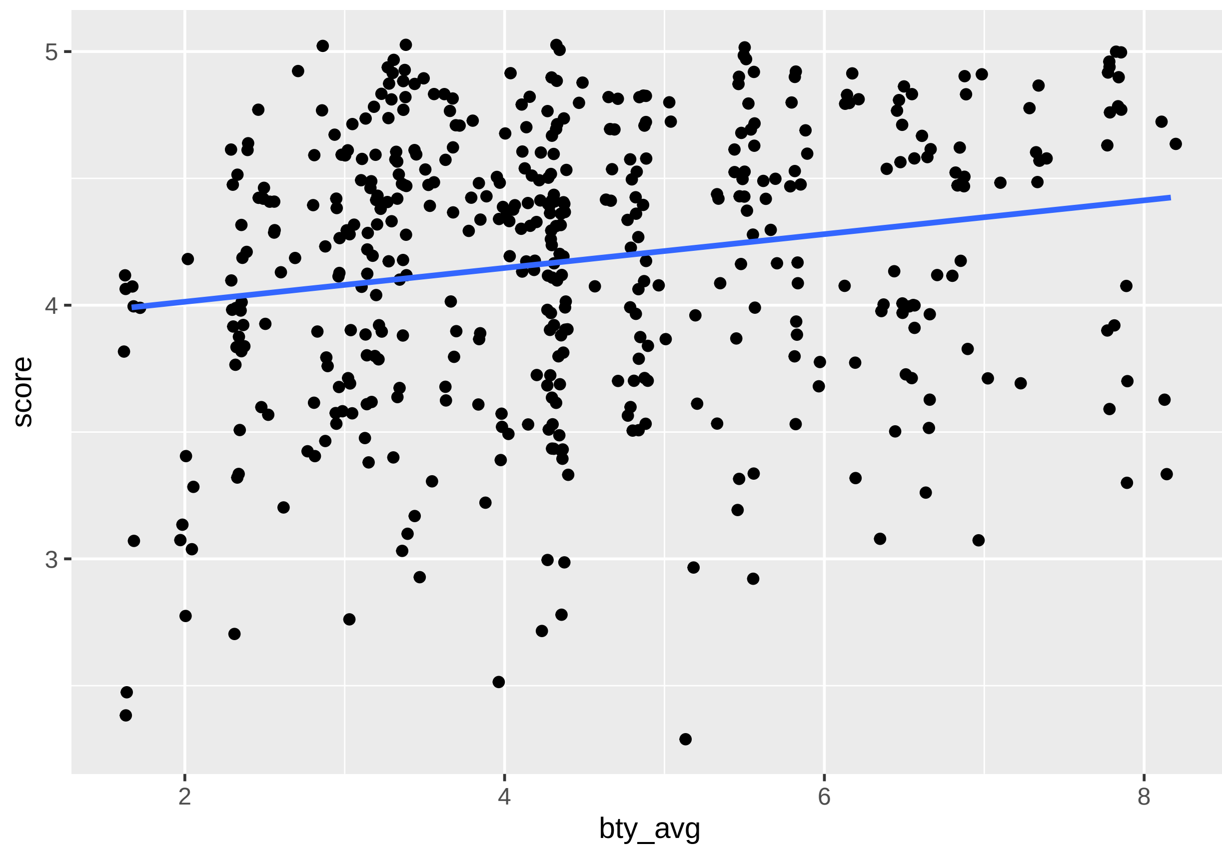
Add the line of the bet fit model to your plot using the following:

```
ggplot(data = evals, aes(x = bty_avg, y = score)) +
  geom_jitter() +
  geom_smooth(method = "lm")
```



The blue line is the model. The shaded gray area around the line tells you about the variability you might expect in your predictions. To turn that off, use `se = FALSE`.

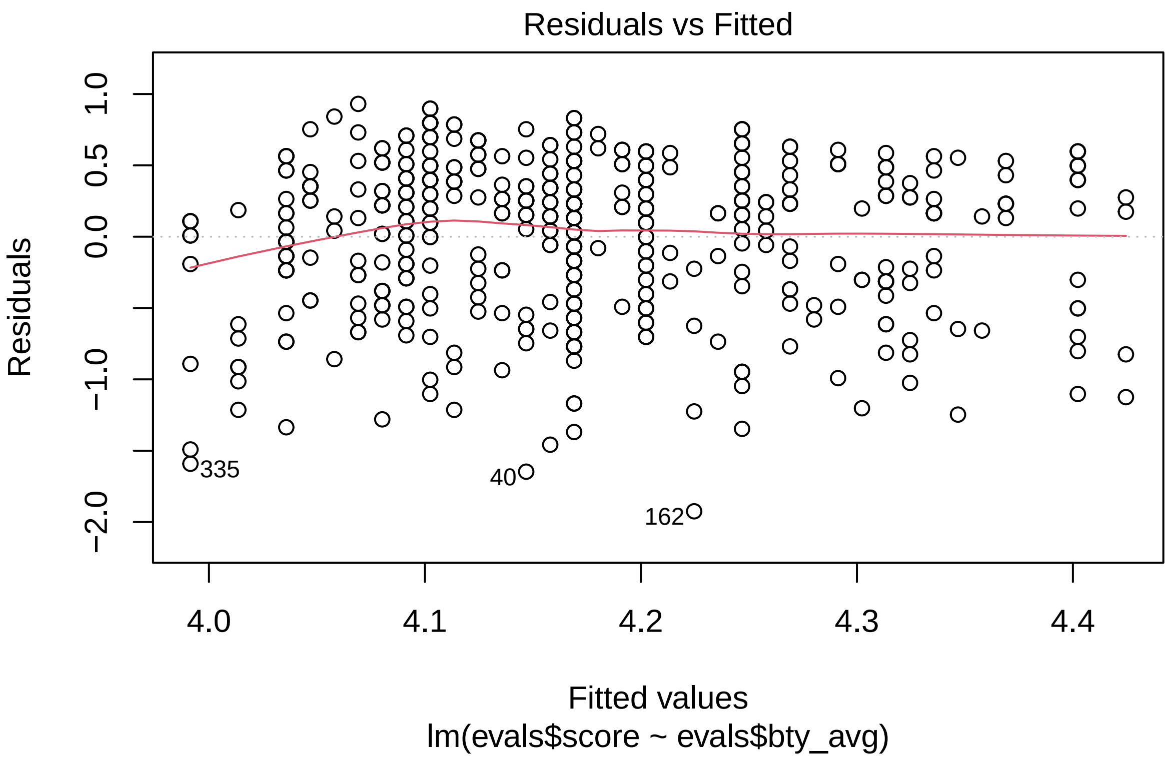
```
ggplot(data = evals, aes(x = bty_avg, y = score)) +
  geom_jitter() +
  geom_smooth(method = "lm", se = FALSE)
```



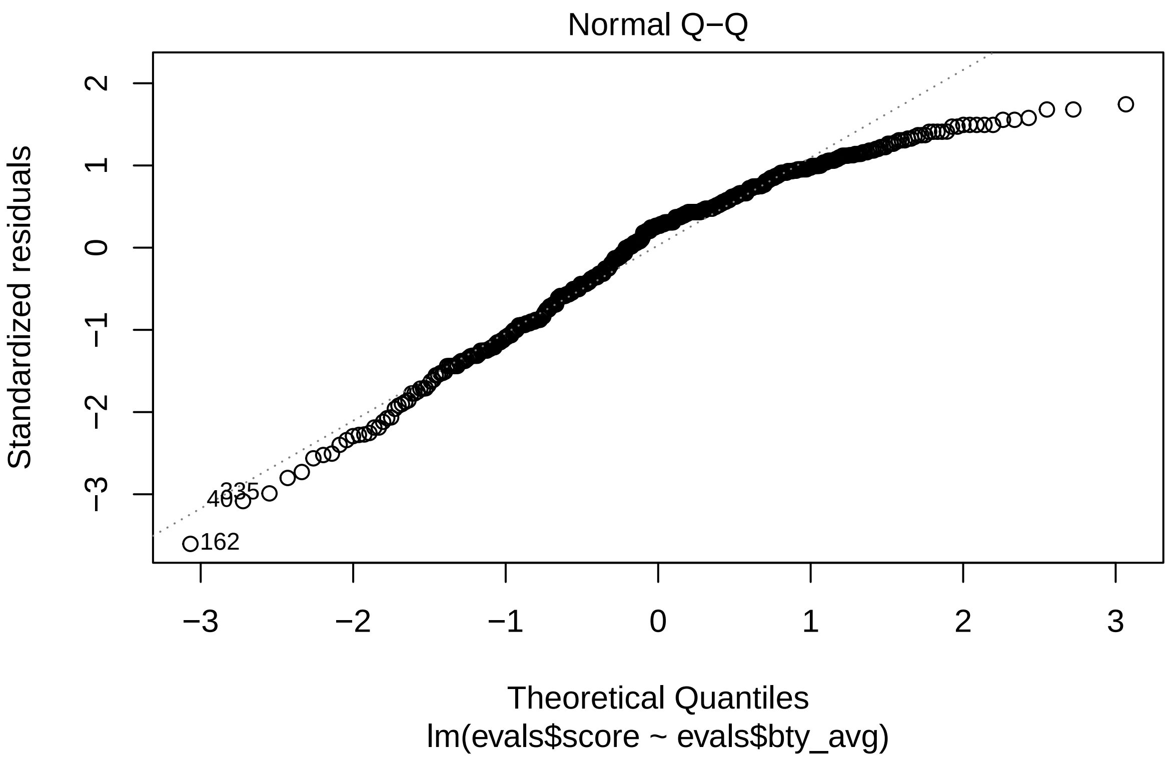
6. Use residual plots to evaluate whether the conditions of least squares regression are reasonable. Provide plots and comments for each one (see the Simple Regression Lab for a reminder of how to make these).

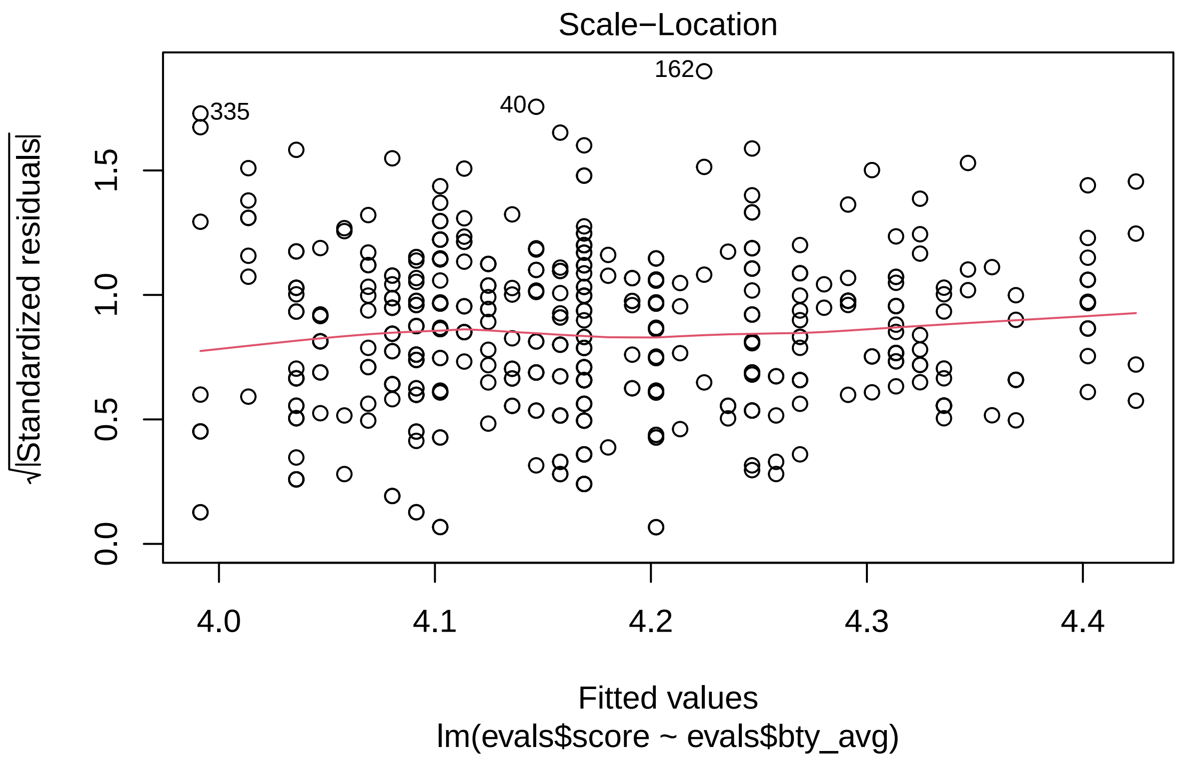
Solution 6:

```
par(mfrow = c(1,1))
plot(m_bty)
```

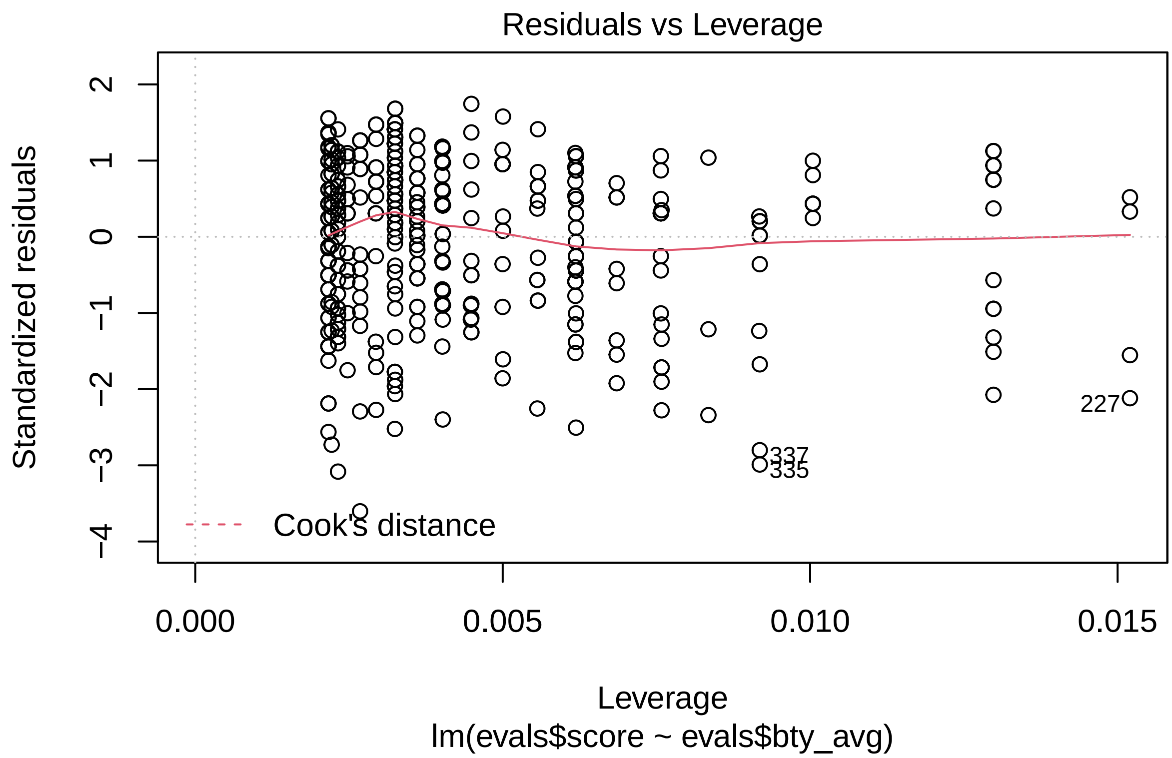


Residuals vs Fitted
Residuals
1.0
0.5
0.0
−1.0
−2.0
335
40
162
4.0
4.1
4.2
4.3
4.4
Fitted values
lm(evals$score ~ evals$bty_avg)


Normal Q−Q
Standardized residuals
2
1
0
−1
−2
−3
335
40
162
−3
−2
−1
0
1
2
3
Theoretical Quantiles
lm(evals$score ~ evals$bty_avg)



Scale−Location
162
40
335
√|Standardized residuals|
0.0
0.5
1.0
1.5
4.0
4.1
4.2
4.3
4.4
Fitted values
lm(evals$score ~ evals$bty_avg)

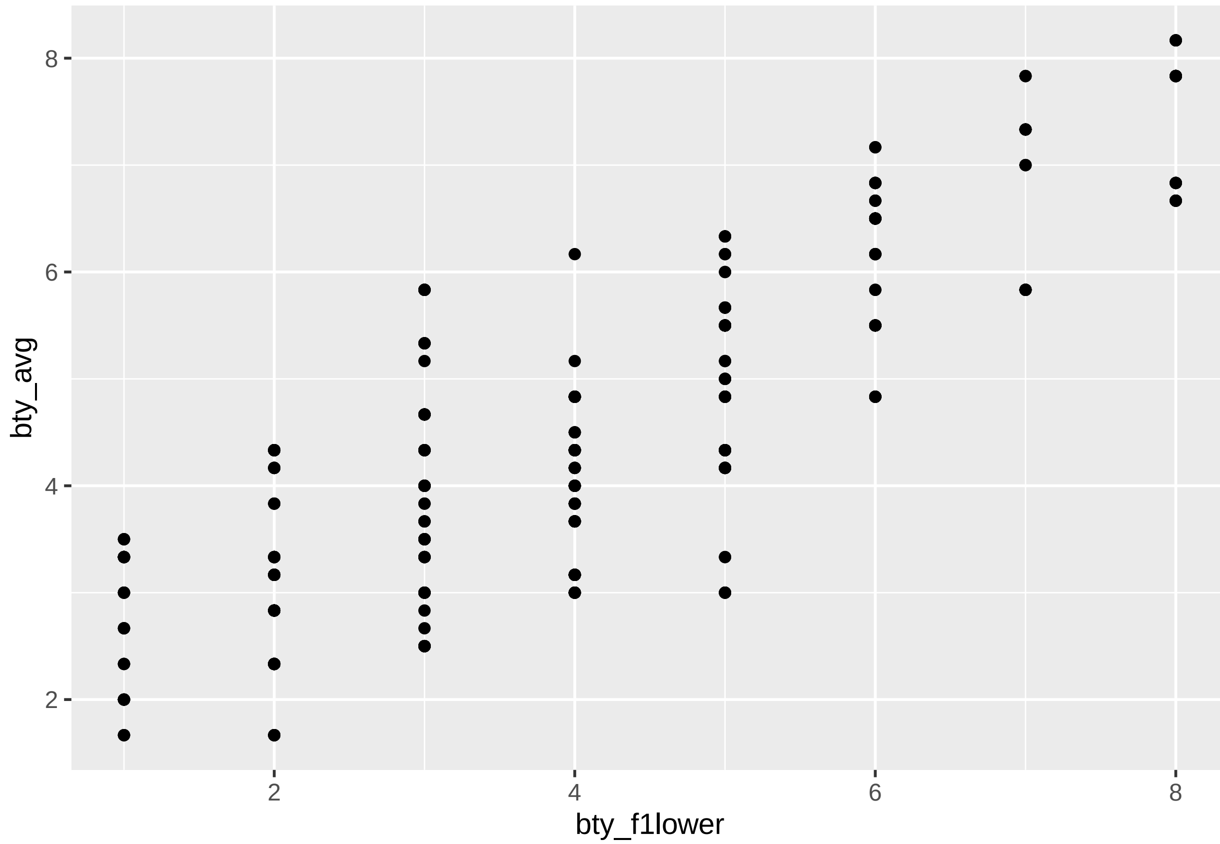


There are four major conditions for the least squares regression.
1. **Linearity:** *The data should follow a linear trend. From the scatter plot, we can see that the data follows a linear trend.*
2. **Nearly Normal residuals:** *From the Q-Q plot, we can see that the residuals are nearly normal*
3. **Constant Variability:** *It can be seen from the Residuals vs Fitted curve that the variability around the least square line is fairly constant and there are no major outliers.*
4. **Independent observations:** *The data was gathered for a large sample of professors from the University of Texas at Austin and we can safely assume that the observations are independent because the professors are likely different and independent of one another.*

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()
```

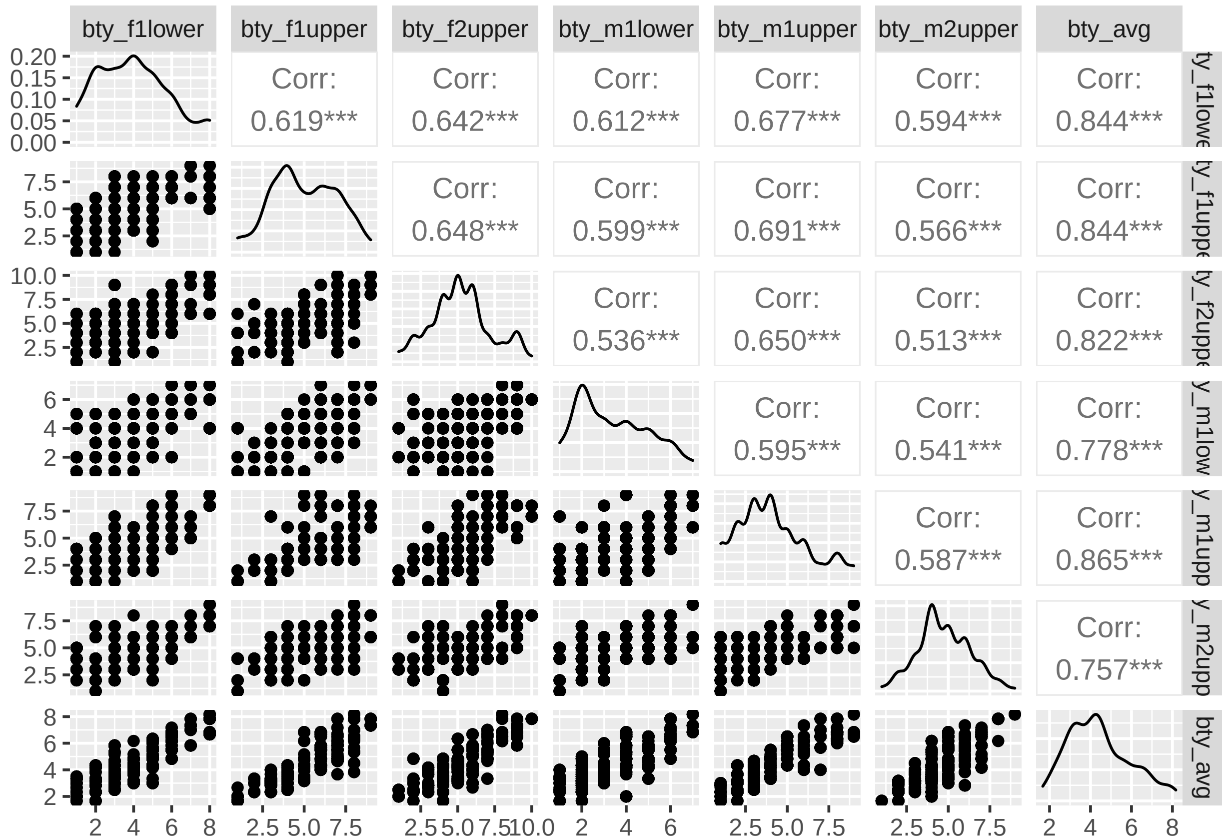



```
evals %>%
  summarise(cor(bty_avg, bty_f1lower))
```

```
## # 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. You can actually 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)
```

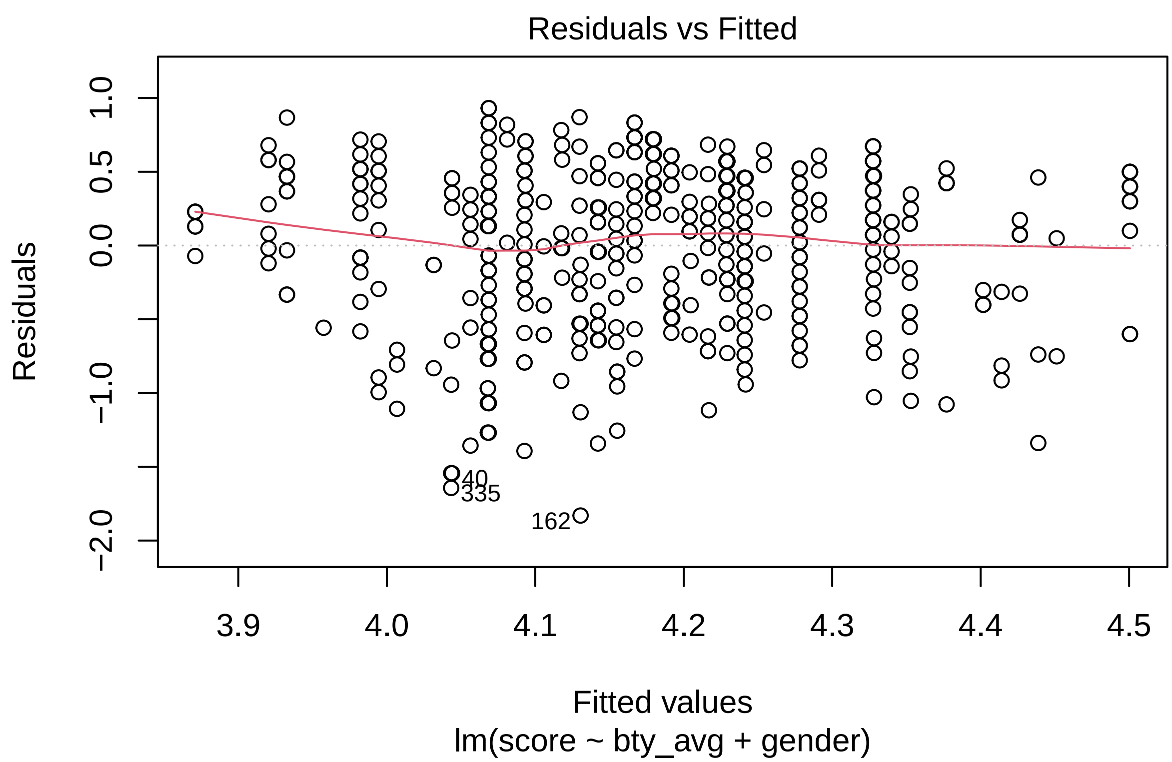
```
##
## Call:
## lm(formula = score ~ bty_avg + gender, data = evals)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.8305 -0.3625  0.1055  0.4213  0.9314
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.74734    0.08466  44.266  < 2e-16 ***
## 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
##
```

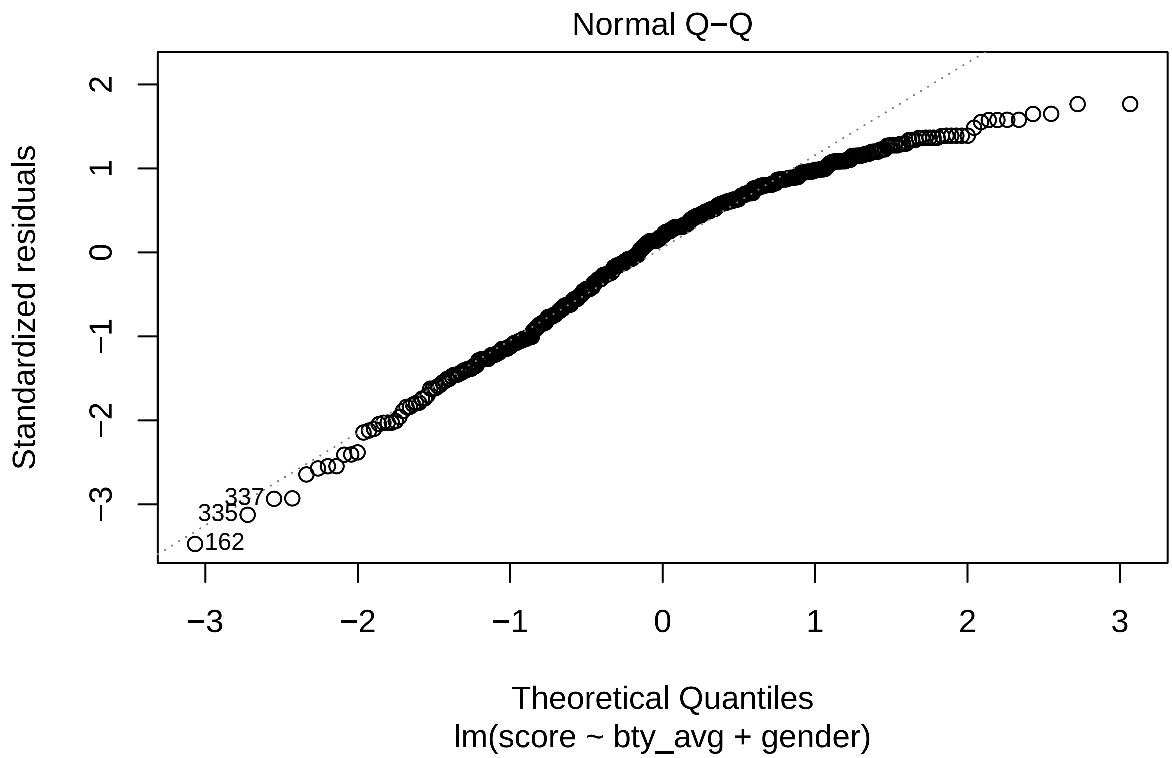
```
## Residual standard error: 0.5287 on 460 degrees of freedom
## Multiple R-squared:  0.05912,    Adjusted R-squared:  0.05503
## F-statistic: 14.45 on 2 and 460 DF,  p-value: 8.177e-07
```

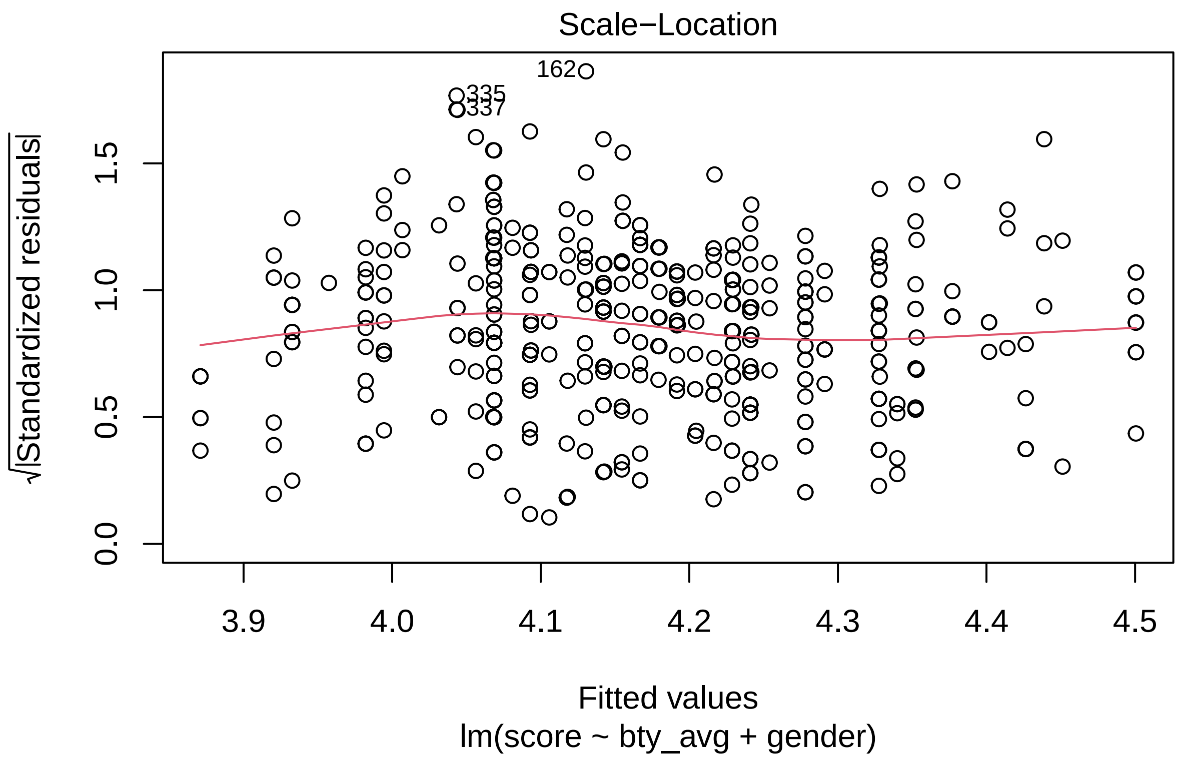
7. P-values and parameter estimates should only be trusted if the conditions for the regression are reasonable. Verify that the conditions for this model are reasonable using diagnostic plots.

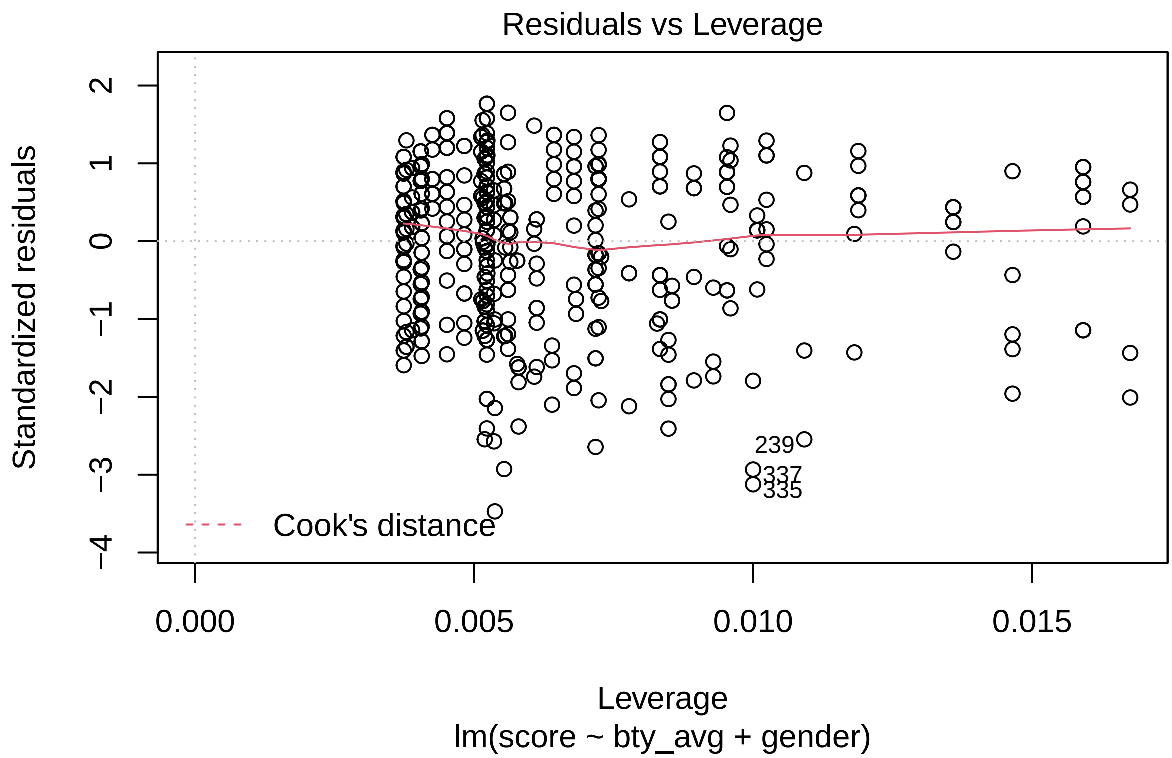
Solution 7:

```r
# check conditions for regression
par(mfrow = c(1,1))
plot(m_bty_gen)
```




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


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)



There are four major conditions for the least squares regression.
1. **Linearity:** *The data should follow a linear trend. From the scatter plot, we can see that the data follows a linear trend.*
2. **Nearly Normal residuals:** *From the Q-Q plot, we can see that the residuals are nearly normal*
3. **Constant Variability:** *It can be seen from the Residuals vs Fitted curve that the variability around the least square line is fairly constant and there are no major outliers.*
4. **Independent observations:** *The data was gathered for a large sample of professors from the University of Texas at Austin and we can safely assume that the observations are independent because the professors are likely different and independent of one another.*

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`?

Solution 8:

```
# summary of the m_bty_gen linear model
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
```

```r
# summary of the m_bty linear model
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
```

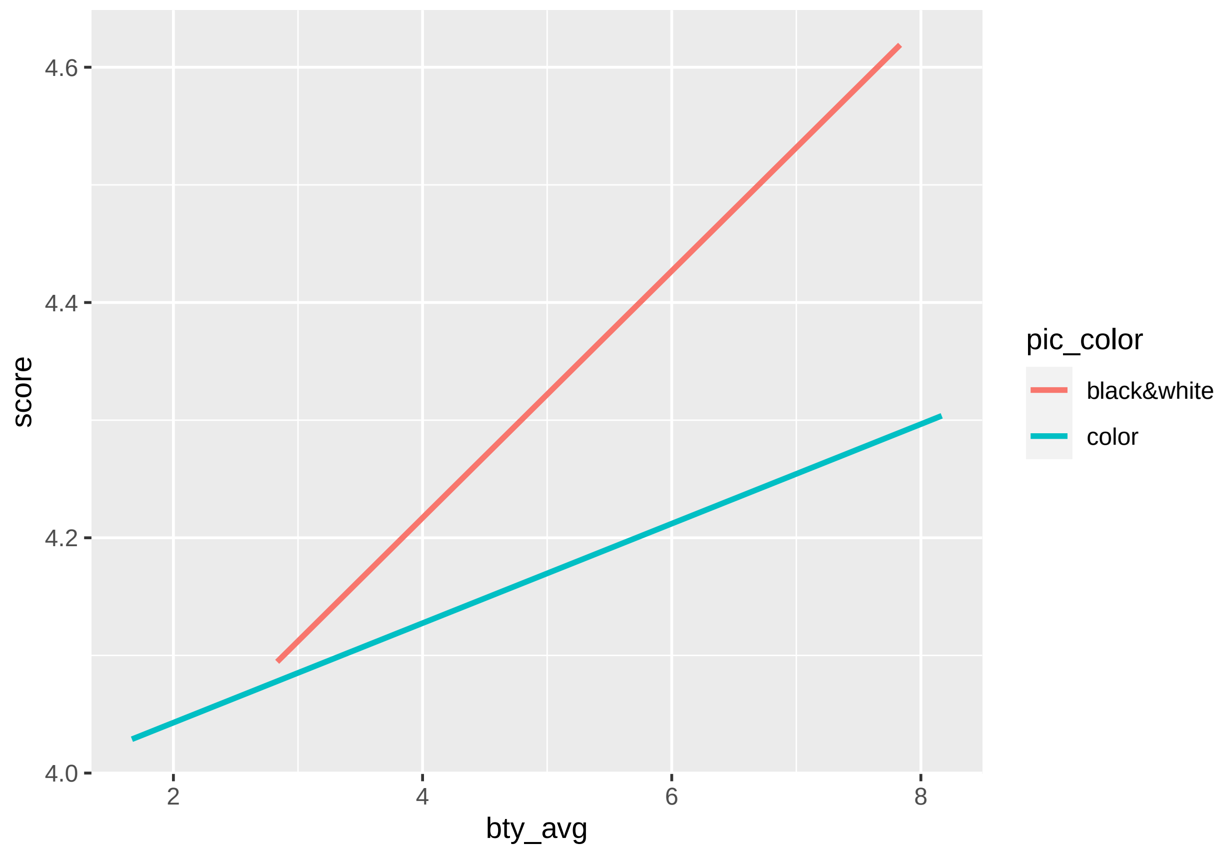
Yes the `bty_avg` is still a significant predictor of `score`. The addition of `gender` has slightly improved the parameter estimate for `bty_avg`.

Note that the estimate for `gender` is now called `gendermale`. You'll see this name change whenever you introduce a categorical variable. The reason is that R recodes `gender` from having the values of `male` and `female` to being an indicator variable called `gendermale` that takes a value of 0 for female professors and a value of 1 for male professors. (Such variables are often referred to as "dummy" variables.)

As a result, for female professors, the parameter estimate is multiplied by zero, leaving the intercept and slope form familiar from simple regression.

$$\begin{aligned}\widehat{score} &= \hat{\beta}_0 + \hat{\beta}_1 \times bty\_avg + \hat{\beta}_2 \times (0) \\ &= \hat{\beta}_0 + \hat{\beta}_1 \times bty\_avg\end{aligned}$$

```r
ggplot(data = evals, aes(x = bty_avg, y = score, color = pic_color)) +
 geom_smooth(method = "lm", formula = y ~ x, se = FALSE)
```



9. What is the equation of the line corresponding to those with color pictures? (*Hint:* For those with color pictures, the parameter estimate is multiplied by 1.) For two professors who received the same beauty rating, which color picture tends to have the higher course evaluation score?

Solution 9:

```
m_bty_color <- lm(score ~ bty_avg + pic_color, data = evals)
summary(m_bty_color)
```

```
##
## Call:
## lm(formula = score ~ bty_avg + pic_color, data = evals)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.8892 -0.3690  0.1293  0.4023  0.9125
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       4.06318    0.10908  37.249  < 2e-16 ***
## bty_avg           0.05548    0.01691   3.282  0.00111 **
## pic_colorcolor -0.16059    0.06892  -2.330  0.02022 *
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5323 on 460 degrees of freedom
## Multiple R-squared:  0.04628,	Adjusted R-squared:  0.04213
## F-statistic: 11.16 on 2 and 460 DF,  p-value: 1.848e-05
```

The equation of the line corresponding to those with color pictures is given by:
$score = 4.06318 + 0.05548 * `btyavg` - 0.16059 * `piccolorcolor`$
For those with color pictures, the parameter estimate is multiplied by 1 while for those with black and white, the parameter estimate is multiplies by 0. For two professors with the same beauty rating, the one with a black and white picture will tend to have higher course evaluation score because the parameter estimate will be multiplied by 0 which will eliminate that portion whereas for the one with a color picture, the parameter estimate will be multiplied by 1 which will decrease the evaluation score by 0.16059

The decision to call the indicator variable `gendermale` instead of `genderfemale` has no deeper meaning. R simply codes the category that comes first alphabetically as a 0. (You can change the reference level of a categorical variable, which is the level that is coded as a 0, using the`relevel()` function. Use `?relevel` to learn more.)

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`.

Solution 10:

```
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
```

R handles categorical variables with k levels by returning a model with k - 1 categorical variables. In this case, there are 3 levels for the rank, but R returned a model with 2 categorical variables as the last variable not returned is considered a reference variable.

The interpretation of the coefficients in multiple regression is slightly different from that of simple regression. The estimate for `bty_avg` reflects how much higher a group of professors is expected to score if they have a beauty rating that is one point higher *while holding all other variables constant*. In this case, that translates into considering only professors of the same rank with `bty_avg` scores that are one point apart.

The search for the best model

We will start with a full model that predicts professor score based on rank, gender, ethnicity, language of the university where they got their degree, age, proportion of students that filled out evaluations, class size, course level, number of professors, number of credits, average beauty rating, outfit, and picture color.

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.

Solution 11:

Without looking at details from the model, I would expect the number of professors `cls_profs` to have the highest p-value because I do not think that the number of professors should impact how students rate their professors.

```
# linear model of score vs all other characteristics
m_full_model <- lm(score ~ rank + gender + ethnicity + language + age + cls_perc_eval +
                     cls_students + cls_level + cls_profs + cls_credits + bty_avg + pic_outfit +
                     pic_color, data = evals)
summary(m_full_model)
```

```
##
## Call:
## lm(formula = score ~ rank + gender + ethnicity + 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
## gendermale            0.2109481  0.0518230   4.071 5.54e-05 ***
## ethnicitynot minority  0.1234929  0.0786273   1.571  0.11698
## 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
```

The number of professors teaching sections in course has the highest p -value.

Let's run the model…

```
m_full <- lm(score ~ rank + gender + ethnicity + 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 + gender + ethnicity + 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
## gendermale             0.2109481  0.0518230   4.071 5.54e-05 ***
## ethnicitynot minority  0.1234929  0.0786273   1.571  0.11698
## 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
```

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

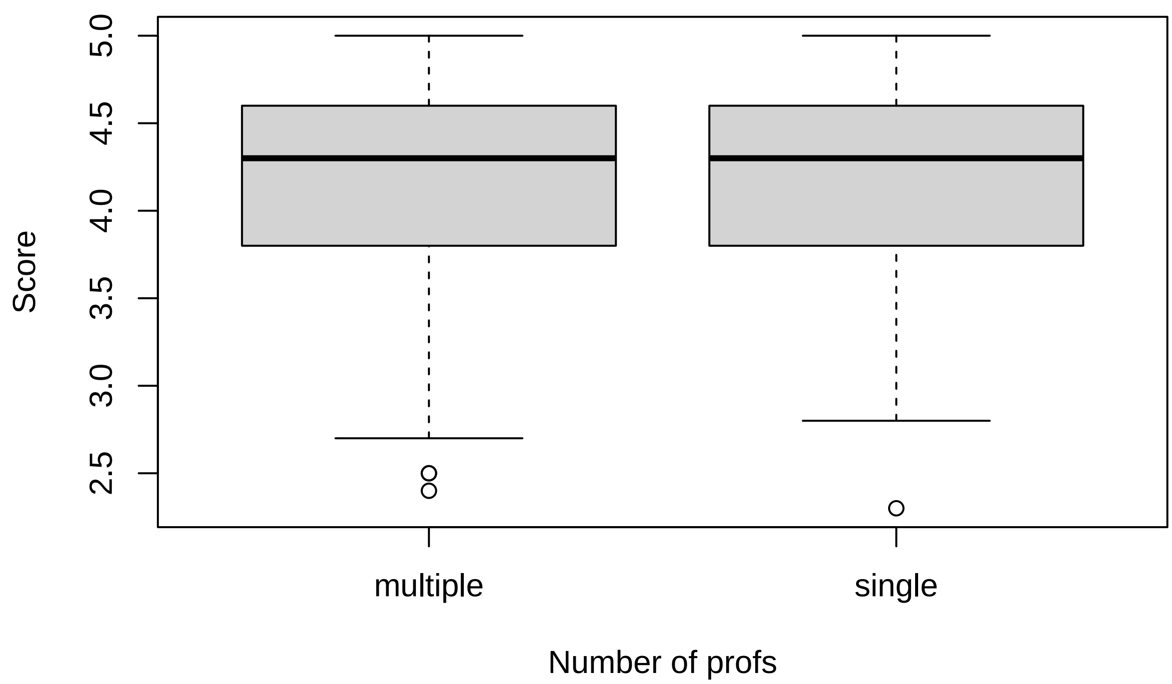
Solution 12:

From the previous exercise, my suspicion appeared to be correct.

```r
m_cls_profs <- lm(score ~ cls_profs, data =  evals)
summary(m_cls_profs)
```

```
##
## Call:
## lm(formula = score ~ cls_profs, data = evals)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.8554 -0.3846  0.1154  0.4154  0.8446
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      4.18464    0.03111 134.493   <2e-16 ***
## cls_profssingle -0.02923    0.05343  -0.547    0.585
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5443 on 461 degrees of freedom
## Multiple R-squared:  0.0006486,  Adjusted R-squared:  -0.001519
## F-statistic: 0.2992 on 1 and 461 DF,  p-value: 0.5847
```

```r
plot(evals$cls_profs, evals$score, xlab = "Number of profs", ylab = "Score")
```



13. Interpret the coefficient associated with the ethnicity variable.

Solution 13:

The coefficient of associated with the ethnicity variable (0.1234929) is the amount that the evaluation score of the professor will increase by if the professor is not minority assuming that other variables are held constant.

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?

Solution 14:

Since the `cls_prof` is the variable with the highest p-value, we drop it.

```
m_full_drop1 <- lm(score ~ rank + gender + ethnicity + language + age + cls_perc_eval
             + cls_students + cls_level + cls_credits + bty_avg
             + pic_outfit + pic_color, data = evals)
summary(m_full_drop1)
```

```
##
## Call:
## lm(formula = score ~ rank + gender + ethnicity + 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
## gendermale            0.2101231  0.0516873   4.065 5.66e-05 ***
## ethnicitynot minority  0.1274458  0.0772887   1.649 0.099856 .
## 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
```

Yes. The coefficients and significance of the other variables changed slightly.

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.

Solution 15:

I did backward elimination for all features whose p-values are greater than 0.05 until no p-value is below 0.05 to settle on the model with the features below:

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

```
##
## Call:
## lm(formula = score ~ gender + ethnicity + 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 ***
## gendermale             0.207112   0.050135   4.131 4.30e-05 ***
## ethnicitynot minority  0.167872   0.075275   2.230  0.02623 *
## 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
```

Therefore the linear model for predicting the score based on the final model that I settled on is given by:

score = 3.771922 + 0.207112 * gendermale + 0.167872 * ethnicity_not_minority - 0.206178 * language_non_english - 0.006046 * age
+ 0.004656 * cls_perc_eval + 0.505306 * cls_credits_one_credit + 0.051069 * bty_avg - 0.190579 * pic_colorcolor

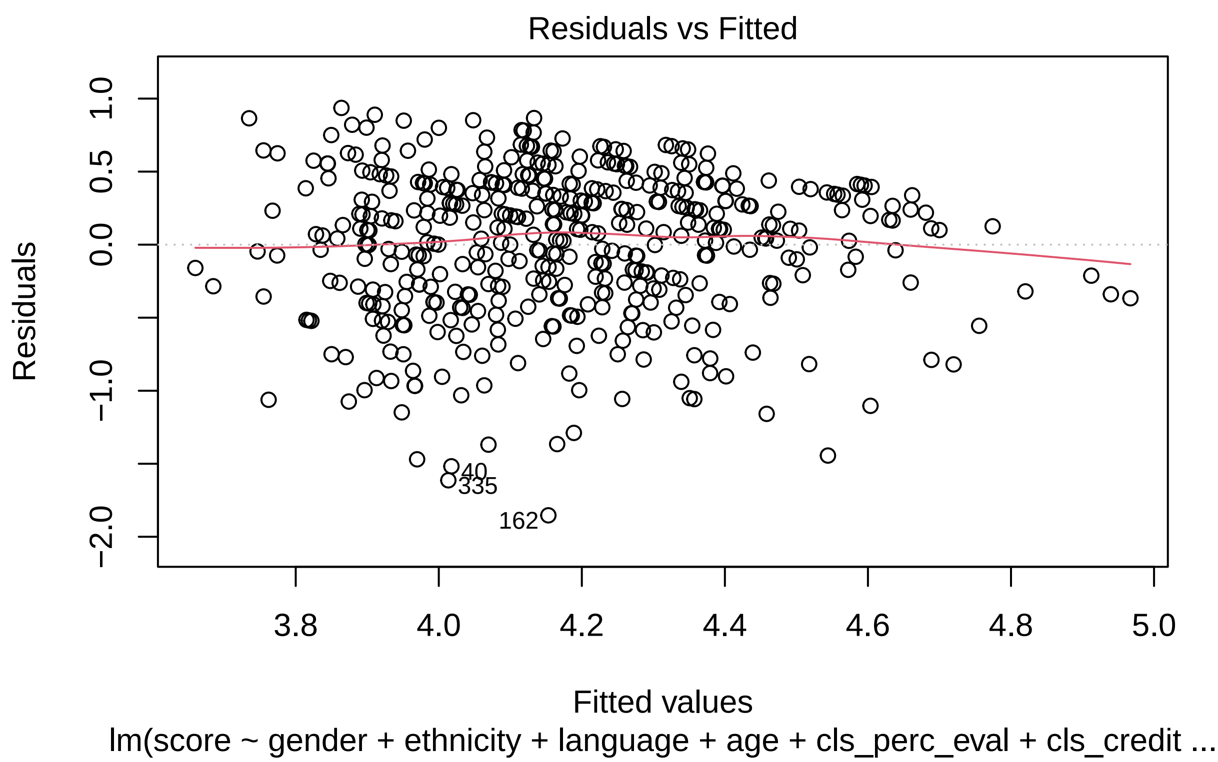
16. Verify that the conditions for this model are reasonable using diagnostic plots.

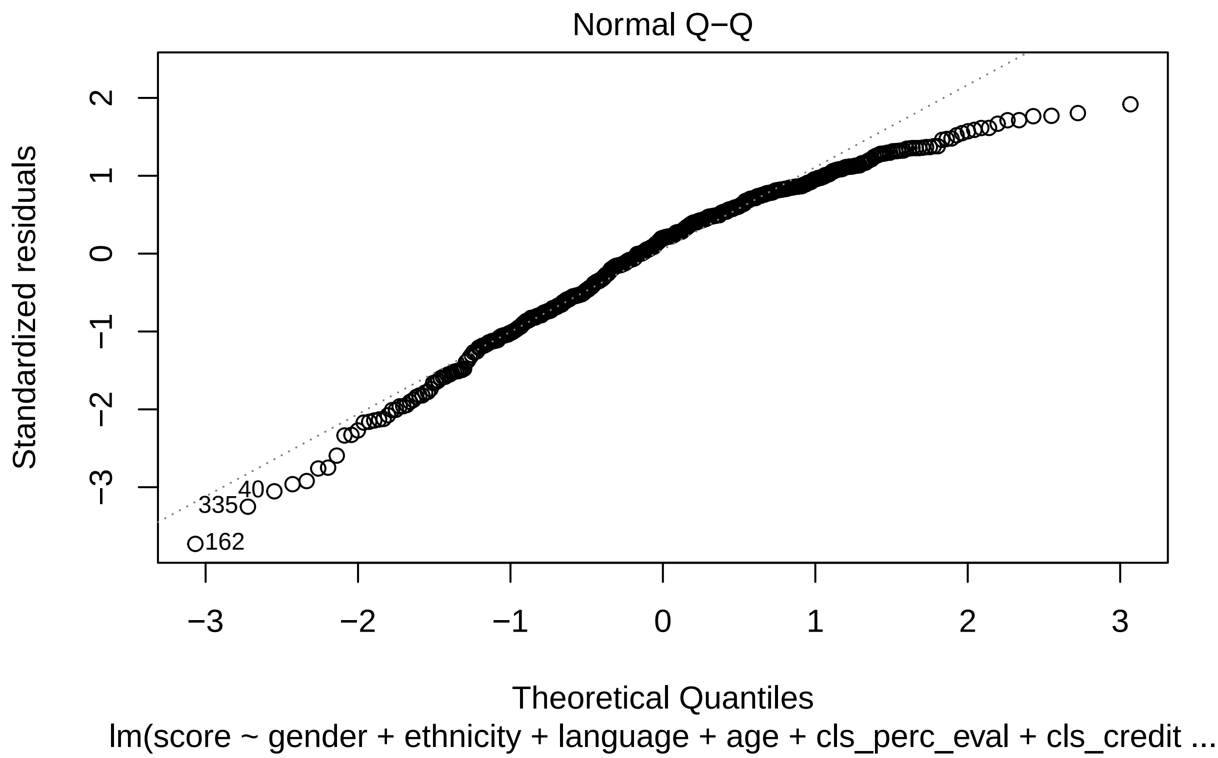
Solution 16:

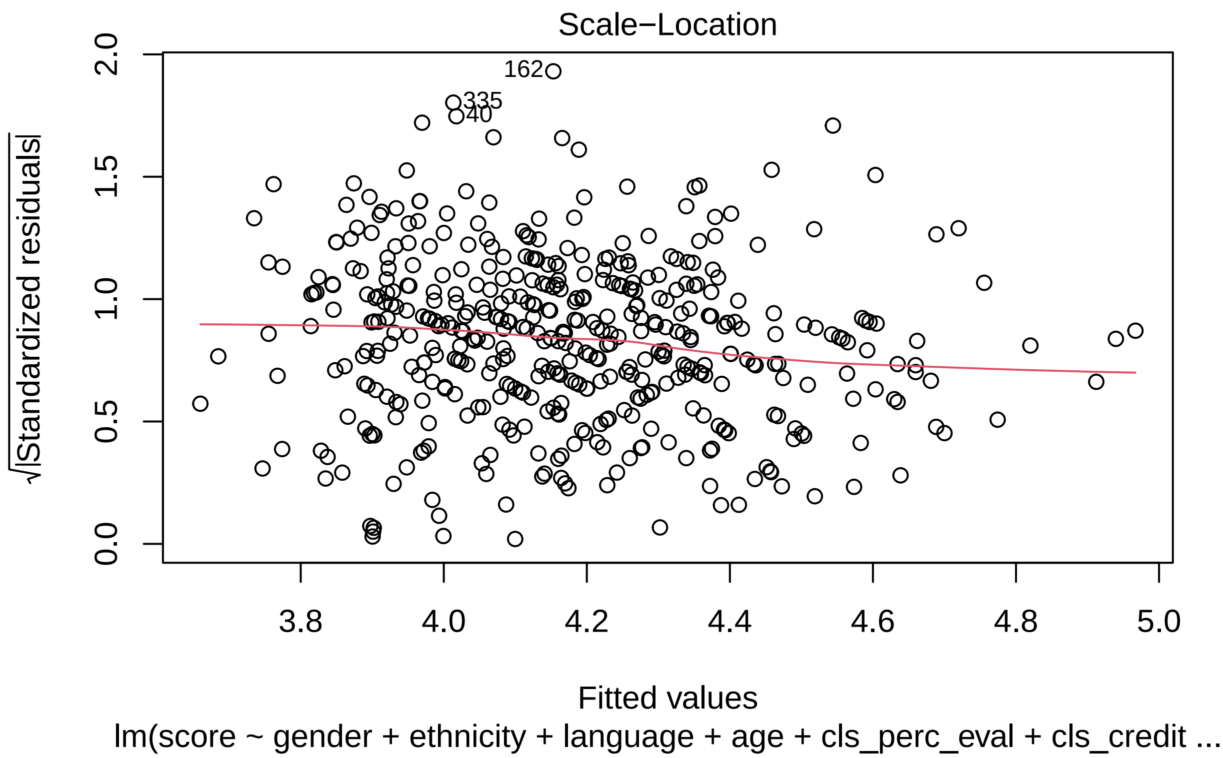
```
# diagnostic plot
par(c(1,1))
```

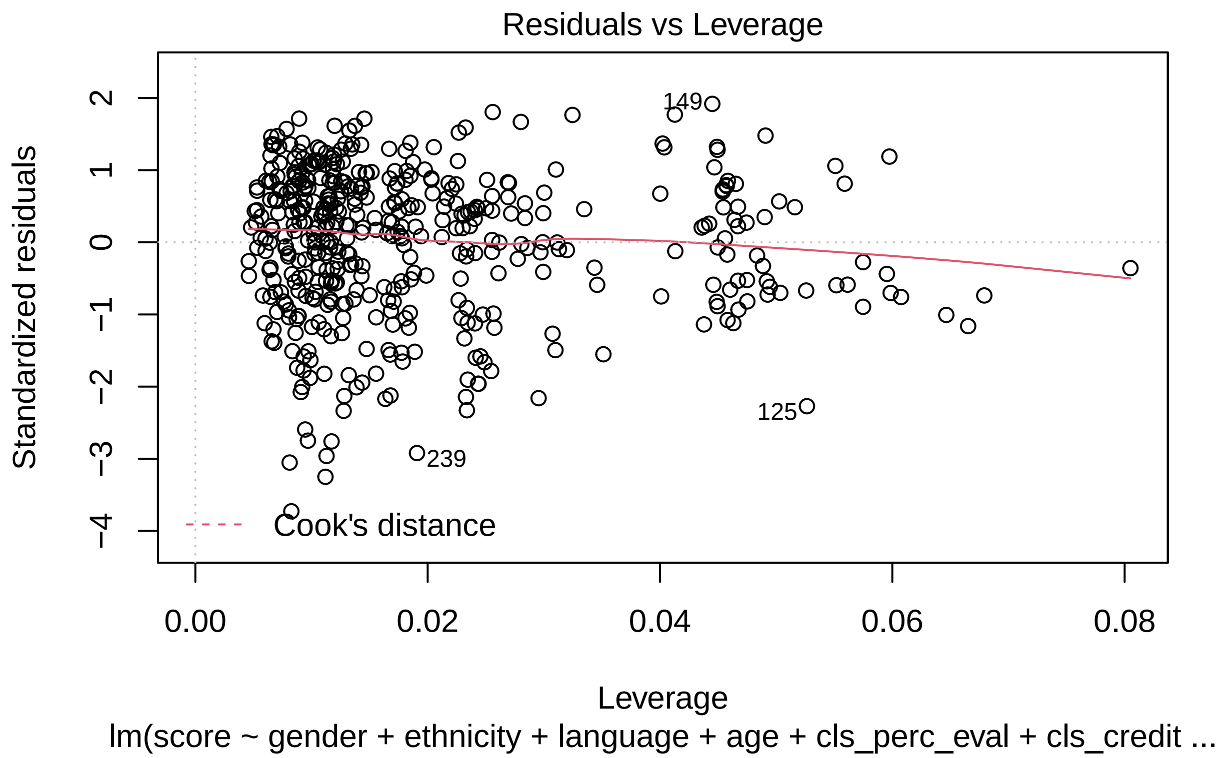
```
## NULL
```

```
plot(m_full_final)
```




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


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



From the plots, the conditions for this model do not appear to be reasonable as there as some notable outliers from the residuals vs fitted curve and there appears to not be constant variability in the residuals. Also, the normality conditions appear to not be met as well.

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?

Solution 17:

Yes this information may have an impact on the "Independent Observations" conditions as the rows (observations) would not be independent of each other which is a necessary ingredient for linear regression.

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.

Solution 18:

Based on my final model, the characteristics of a professor and course at University of Texas at Austin that would be associated with a high evaluation score would be a professor who is male, not minority, speaks English, young, and considered handsome.

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

Solution 19:

No, I would not be comfortable generalizing my conclusions to apply to professors at any University as the situation in UT Austin may be different from that in other Universities. Take for example, a University in France that speaks French (Not English), English proficiency may not improve a professor's evaluation score as that may not be too relevant. Also, In a typical minority school like HBCU, not being minority may not improve a professor's chances of a high evaluation score. Furthermore, the number of observations is 463 from only one University which is not large enough nor sufficient to make a generalization to any University.
