DATA606 Final Project

Alexis Mekueko

11/26/2020

Contents

Github Link: https://github.com/asmozo24/DATA607_Final_Project

Web link: https://rpubs.com/amekueko/697306

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.

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.

Introduction

Many students failed in school not because of thier intelligence. There are numerous factors that contribute to students success. In other words, students success in school relies upon on the ability of the school education system to take appropriate measures on these factor. These factors are : weekly studying time, extra-curricular activities, travel time to school, family educational support, student desire to pursue higher education, companionship, parents'job type, etc. Therefore, in this project, we interested in studying these factors to determine any corroletion that could lead to students failure. If none, then we would like to determine the factors which contribute for the most to success. This is done in order for the school education system to keep track of success and improve the factors that negatively impact students success.

Benefits

The interest in experimental study related to school will have the advantage to help schools' officials in decision making in term of improving school education system. This project is seeking to make the collected data about xx school speaks or reveal useful information. I plan to become a consultant using my skills as data scientist in various domain of the society to present meaningful report to government entities, companies, and organizations to help them in decision making. So, this project will contribute to building skills necessary for one to be successful in data science.

Research question

There are some study out there suggesting that study time likely affects students performance. Let's verify that in this study. Do students studing at least 10hrs weekly do well in class than those with single parent? We could also explore the corelation between study time and students performance. Is there a linear relationship between study time and students performance? In another words, do students putting more hours in study their lecons get better grades than those with few hours of in study time? How does study time impact students grades?

1. Obtain Data ## Data collection Data is collected or made available by archive.ics.uci.edu: The UCI Machine Learning Repository is a collection of databases, domain theories, and data generators that are used by the machine learning community for the empirical analysis of machine learning algorithms. The archive was created as an ftp archive in 1987 by David Aha and fellow graduate students at UC Irvine. The current version of the web site was designed in 2007 by Arthur Asuncion and David Newman, and this project is in collaboration with Rexa.info at the University of Massachusetts Amherst. Funding support from the National Science Foundation is gratefully acknowledged.

##Data source

```
 We found some interesting dataset from -> data source: https://archive.ics.uci.edu/ml/machine-learning-

- Unstructured Data
- SQL database
```

2. Scrub Data
   - organizing data
   - Tidying up data
3. Explore Data
   - Inspect data and understand the characteristic of the data
   - Looking for relationship, patterns and values,
4. Model Data
   - Create a predictive models (I don't think we learn data modeling in this class)
5. Interpret Results
   - Explaining findings (Answering the research question)
   - Understanding the audience
   - Actionable information

Creating a reproducible lab report

To create your new lab report, in RStudio, go to New File -> R Markdown… Then, choose From Template and then choose Lab Report for OpenIntro Statistics Labs from the list of templates.

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.

```r
data(evals)
glimpse(evals)
```

```
## Rows: 463
## Columns: 21
## $ score         <dbl> 4.7, 4.1, 3.9, 4.8, 4.6, 4.3, 2.8, 4.1, 3.4, 4.5, 3.8...
## $ rank          <fct> tenure track, tenure track, tenure track, tenure trac...
## $ ethnicity     <fct> minority, minority, minority, minority, not minority,...
## $ gender        <fct> female, female, female, female, male, male, male, mal...
## $ language      <fct> english, english, english, english, english, english,...
## $ age           <int> 36, 36, 36, 36, 59, 59, 59, 51, 51, 40, 40, 40, 40, 4...
## $ cls_perc_eval <dbl> 55.81395, 68.80000, 60.80000, 62.60163, 85.00000, 87....
## $ cls_did_eval  <int> 24, 86, 76, 77, 17, 35, 39, 55, 111, 40, 24, 24, 17, ...
## $ cls_students  <int> 43, 125, 125, 123, 20, 40, 44, 55, 195, 46, 27, 25, 2...
## $ cls_level     <fct> upper, upper, upper, upper, upper, upper, upper, uppe...
## $ cls_profs     <fct> single, single, single, single, multiple, multiple, m...
## $ cls_credits   <fct> multi credit, multi credit, multi credit, multi credi...
## $ bty_f1lower   <int> 5, 5, 5, 5, 4, 4, 4, 5, 5, 2, 2, 2, 2, 2, 2, 2, 2, 7,...
## $ bty_f1upper   <int> 7, 7, 7, 7, 4, 4, 4, 2, 2, 5, 5, 5, 5, 5, 5, 5, 5, 9,...
## $ bty_f2upper   <int> 6, 6, 6, 6, 2, 2, 2, 5, 5, 4, 4, 4, 4, 4, 4, 4, 4, 9,...
## $ bty_m1lower   <int> 2, 2, 2, 2, 2, 2, 2, 2, 2, 3, 3, 3, 3, 3, 3, 3, 3, 7,...
## $ bty_m1upper   <int> 4, 4, 4, 4, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 6,...
## $ bty_m2upper   <int> 6, 6, 6, 6, 3, 3, 3, 3, 3, 2, 2, 2, 2, 2, 2, 2, 2, 6,...
## $ bty_avg       <dbl> 5.000, 5.000, 5.000, 5.000, 3.000, 3.000, 3.000, 3.33...
## $ pic_outfit    <fct> not formal, not formal, not formal, not formal, not f...
## $ pic_color     <fct> color, color, color, color, color, color, color, colo...
```

```r
#?evals
```

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:

Exploring the data

Exercice 1

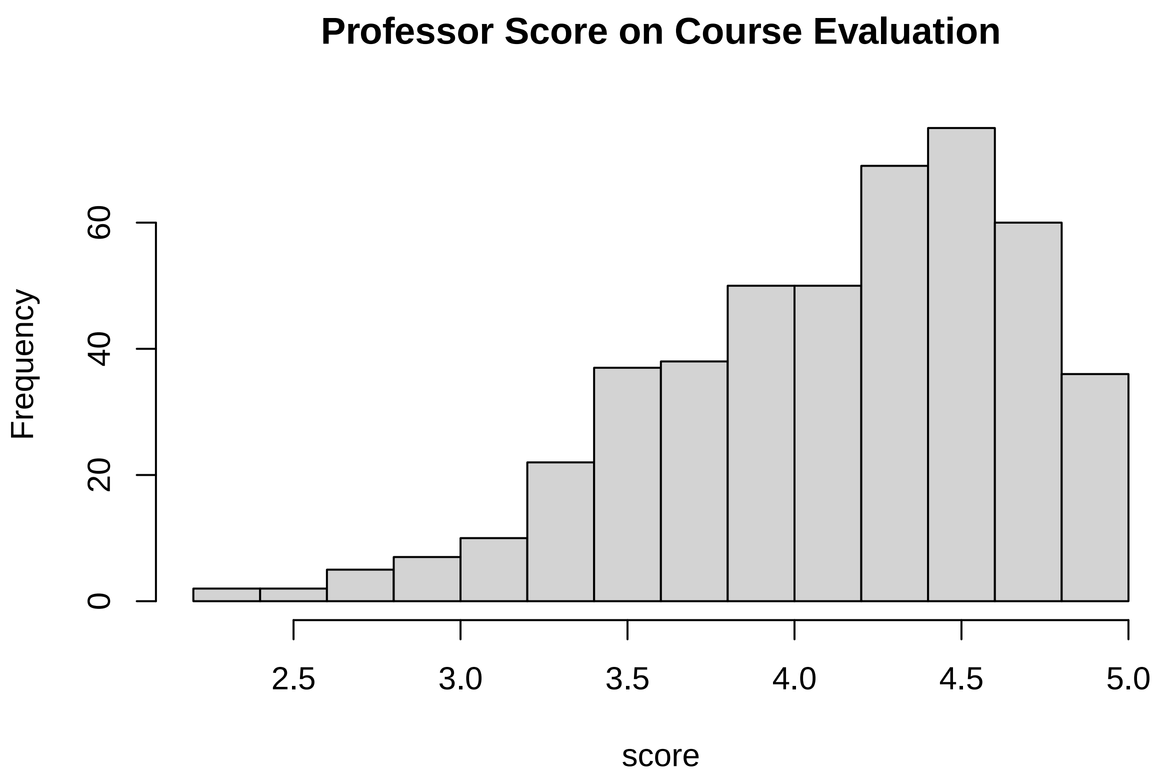
1. Is this an observational study or an experiment? The original research question posed in the paper is whether beauty leads directly to the differences in course evaluations. Given the study design, is it possible to answer this question as it is phrased? If not, rephrase the question.

Answer: this is an observational study (glimpse of the data shows observed variables such as rank [tenure track…], gender[female, male], age). We could rephrase the research question as, is there a linear corelation between the beauty of the professor and the score he or she received?

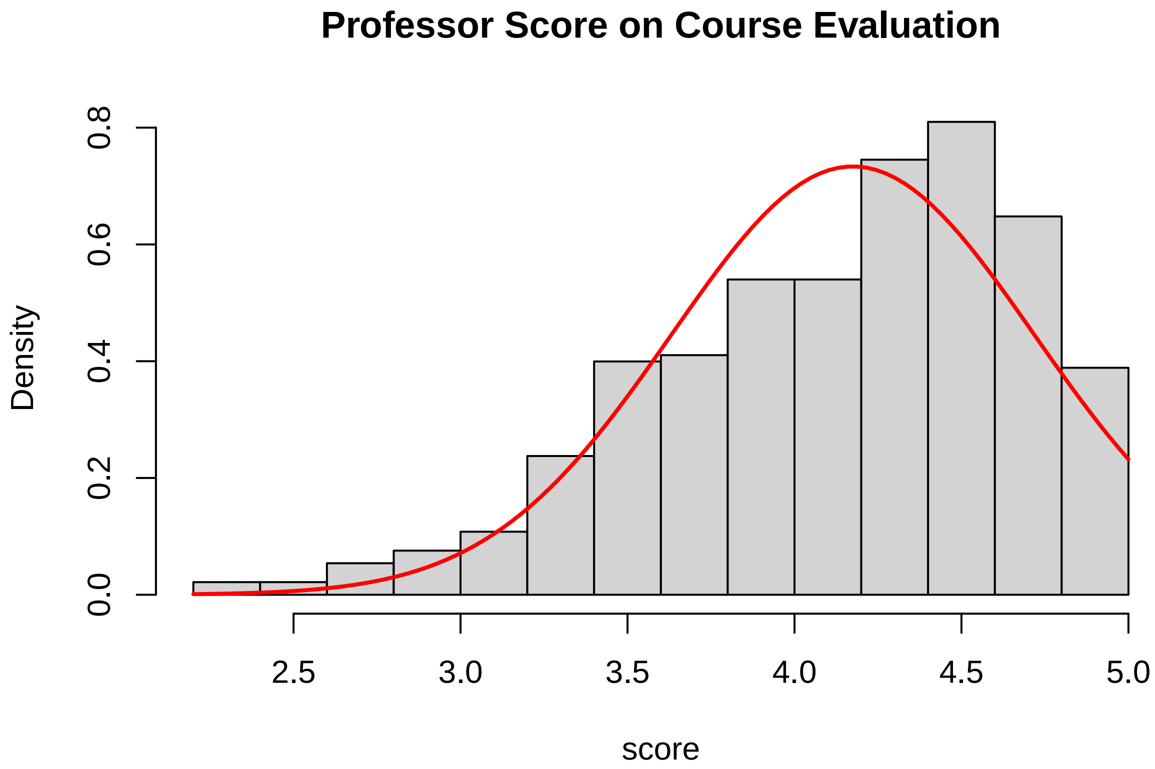
Exercice 2

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?

```r
hist(evals$score,  main = "Professor Score on Course Evaluation", xlab = "score")
```



```r
hist(evals$score, prob = TRUE, breaks = 11,  main = "Professor Score on Course Evaluation", xlab = "sco
x <- seq(from = 0, to = 5, by = 0.5)
curve(dnorm(x, mean = mean(evals$score), sd = sd(evals$score) ), add = TRUE, col = "red", lwd = 2)
```



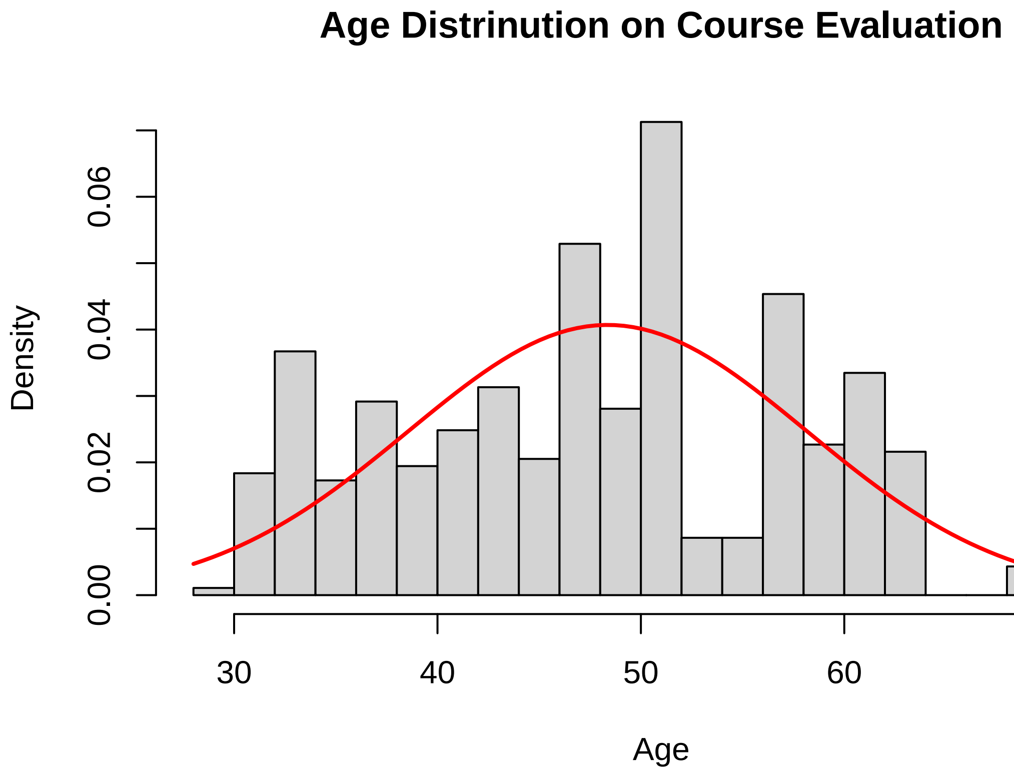
```
#lines(density(evals$score, adjust = 1.8), col = "Red", lwd = 2)
```

Answer : the fitted curve on the density plot shows left skewed distribution, single mode. . Yes, it is what I expect to see since the sample is not random and most scores appear to be above 3.8 which are typical score for most teachers.

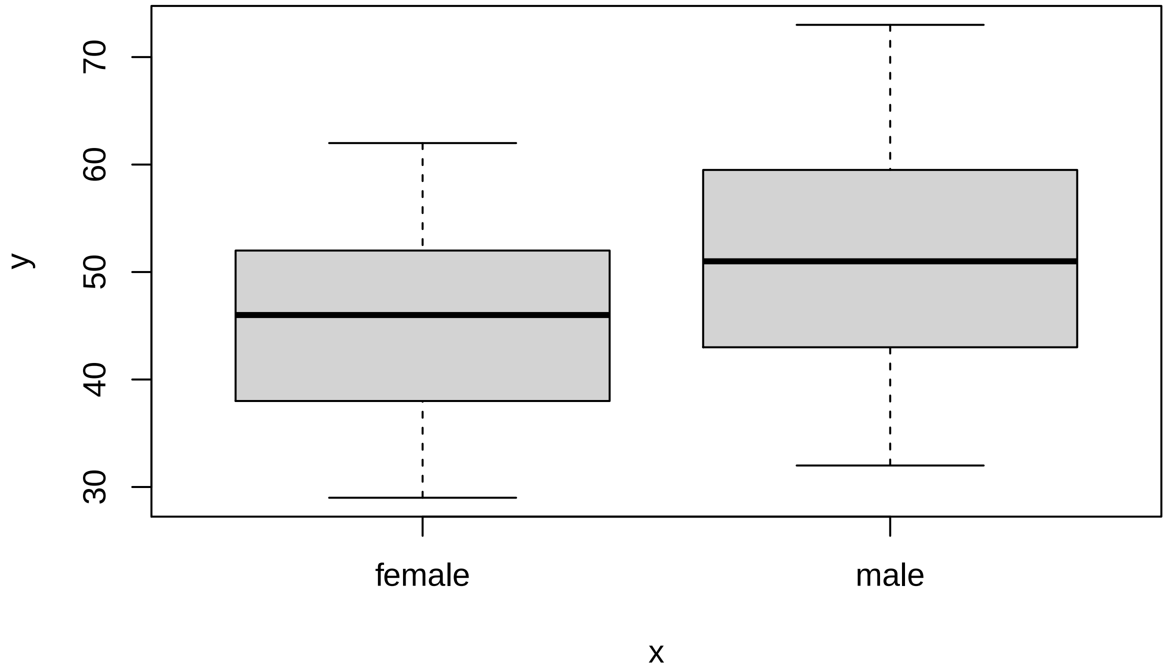
Exercice 3

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

```
hist(evals$age, prob = TRUE, breaks = 16, main = "Age Distrinution on Course Evaluation", xlab = "Age")
x <- seq(from = 20, to =95, by = 10)
curve(dnorm(x, mean = mean(evals$age), sd = sd(evals$age) ), add = TRUE, col = "red", lwd = 2)
```



```
plot(evals$gender, evals$age)
```

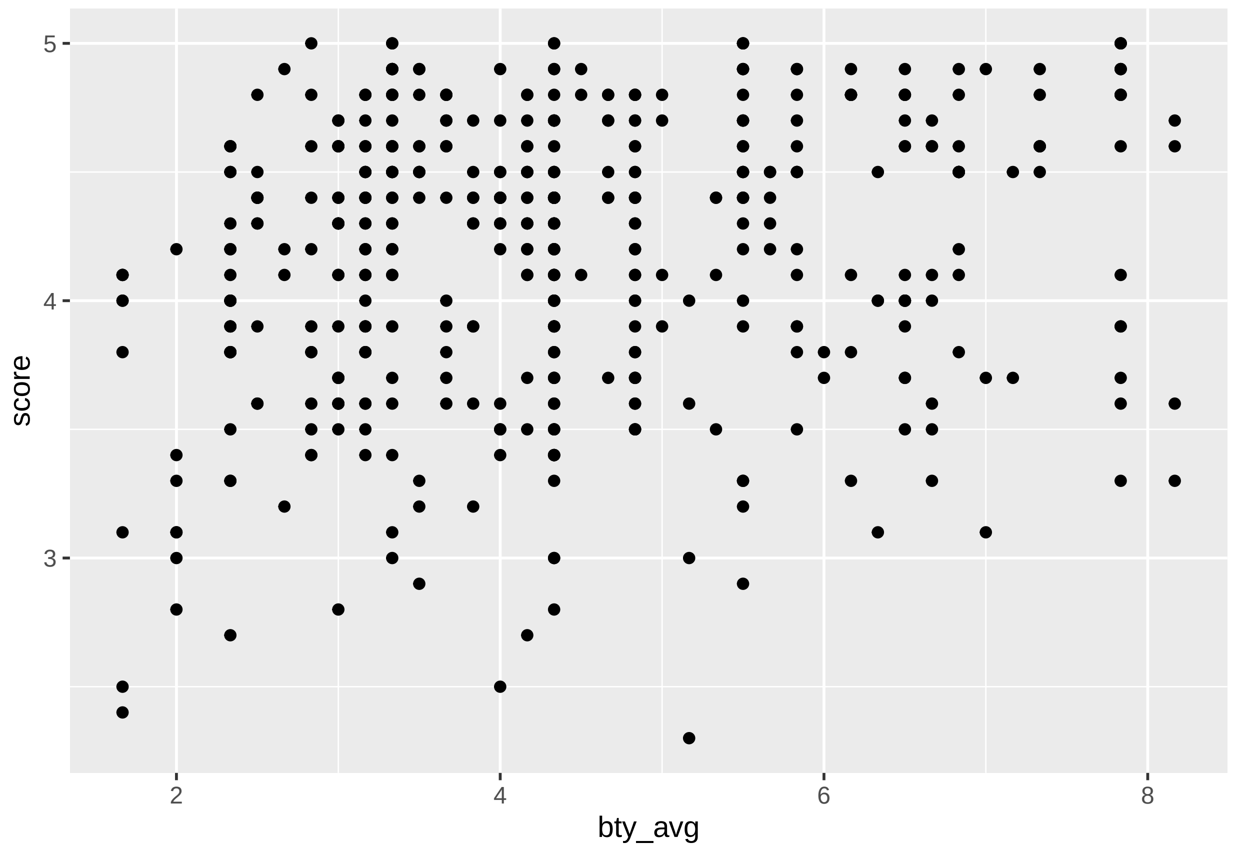


Answer: the age distribution is in this professor course evaluation shows a normal , unimodal distribution with mean around 50, futhermore the age by gender shows male older than female.

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()
```



```
min(evals$score)
```

```
## [1] 2.3
```

```
summary(evals$score)
```

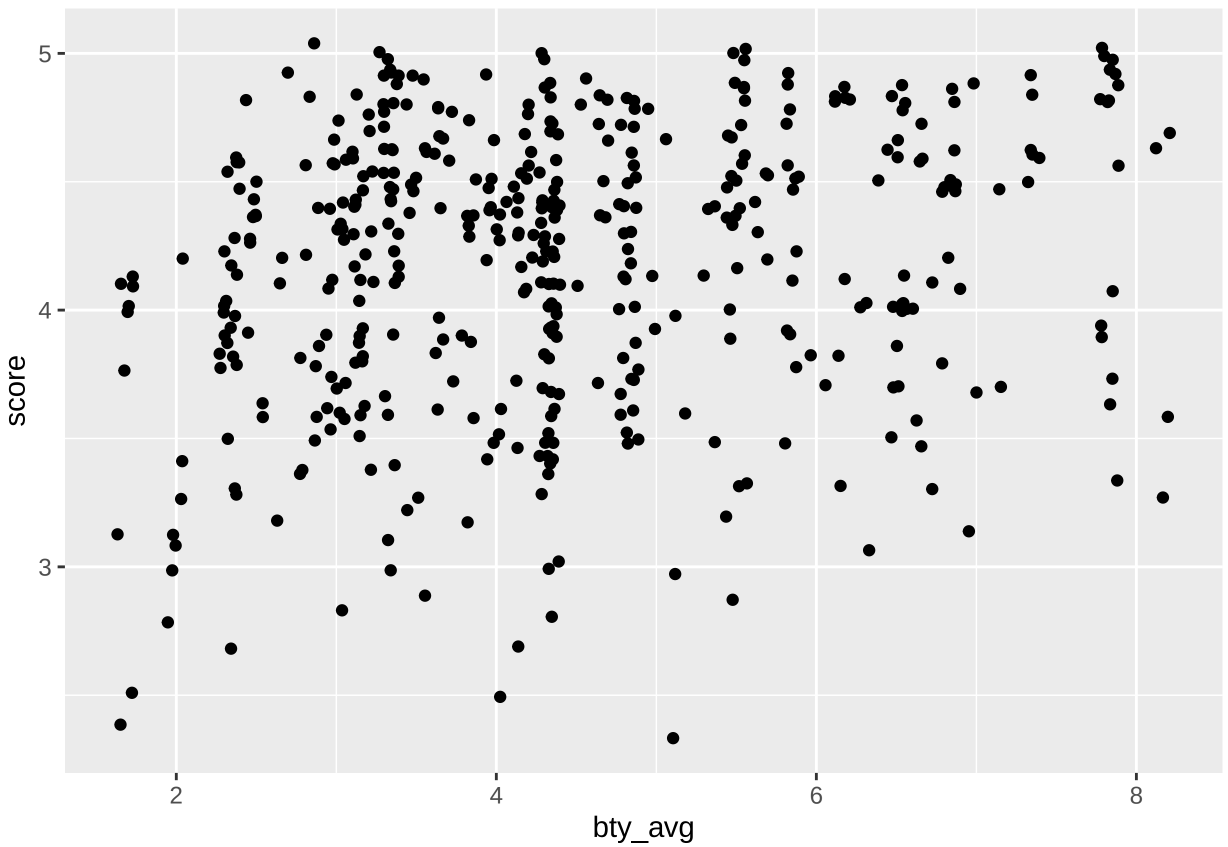
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   2.300   3.800   4.300   4.175   4.600   5.000
```

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? ### Answer: the dataframe (evals) has 463 rows , the score goes from 2.3 to 5, not sure what is awry here if not , the scatterplot shows no pattern.

Exercice 4

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

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



Answer: Now, we can can a typical scatterplot with agregate points in some area, so the initial scatterplot was showing overlapping point.

Exercice 5

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?

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

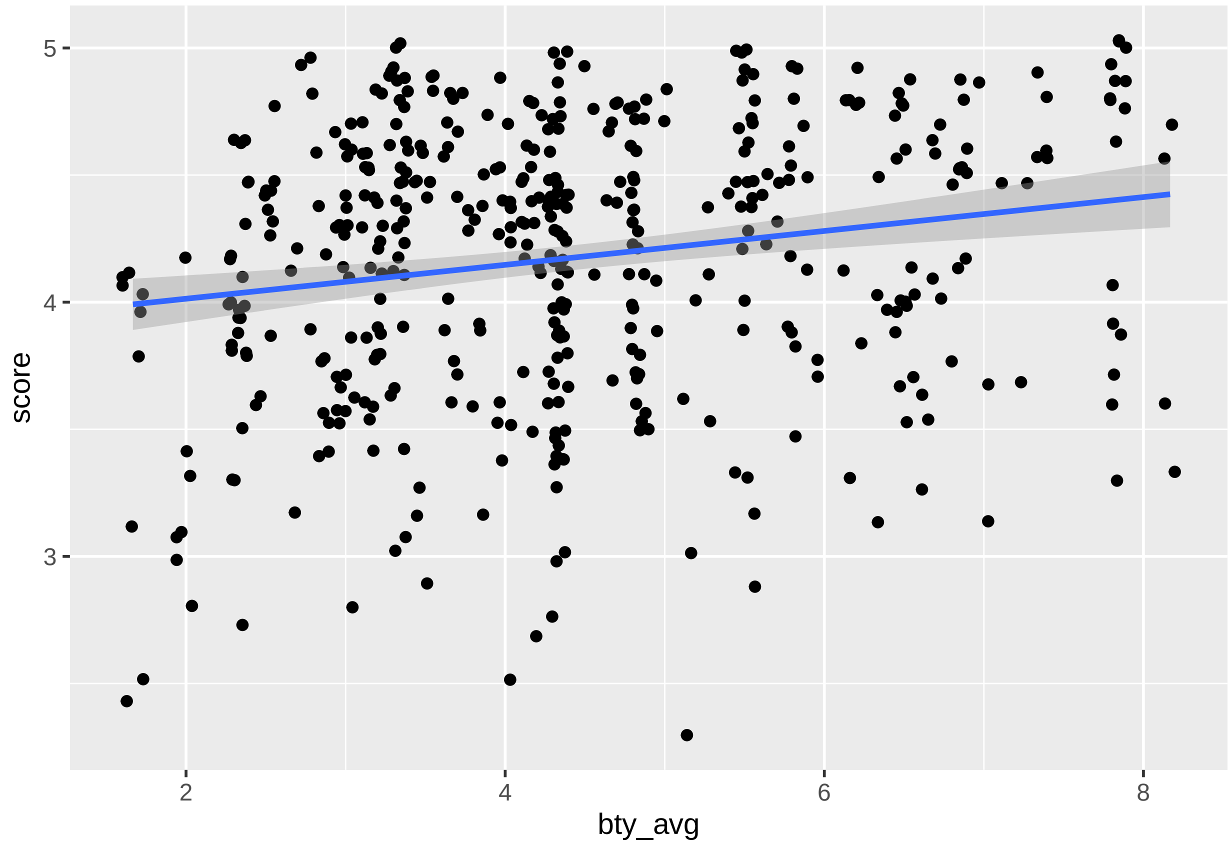
```
m_bty <- lm(score ~ bty_avg, data = evals)
summary(m_bty)
```

```
##
## Call:
## lm(formula = score ~ bty_avg, data = evals)
##
## 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 ***
## 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
```

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

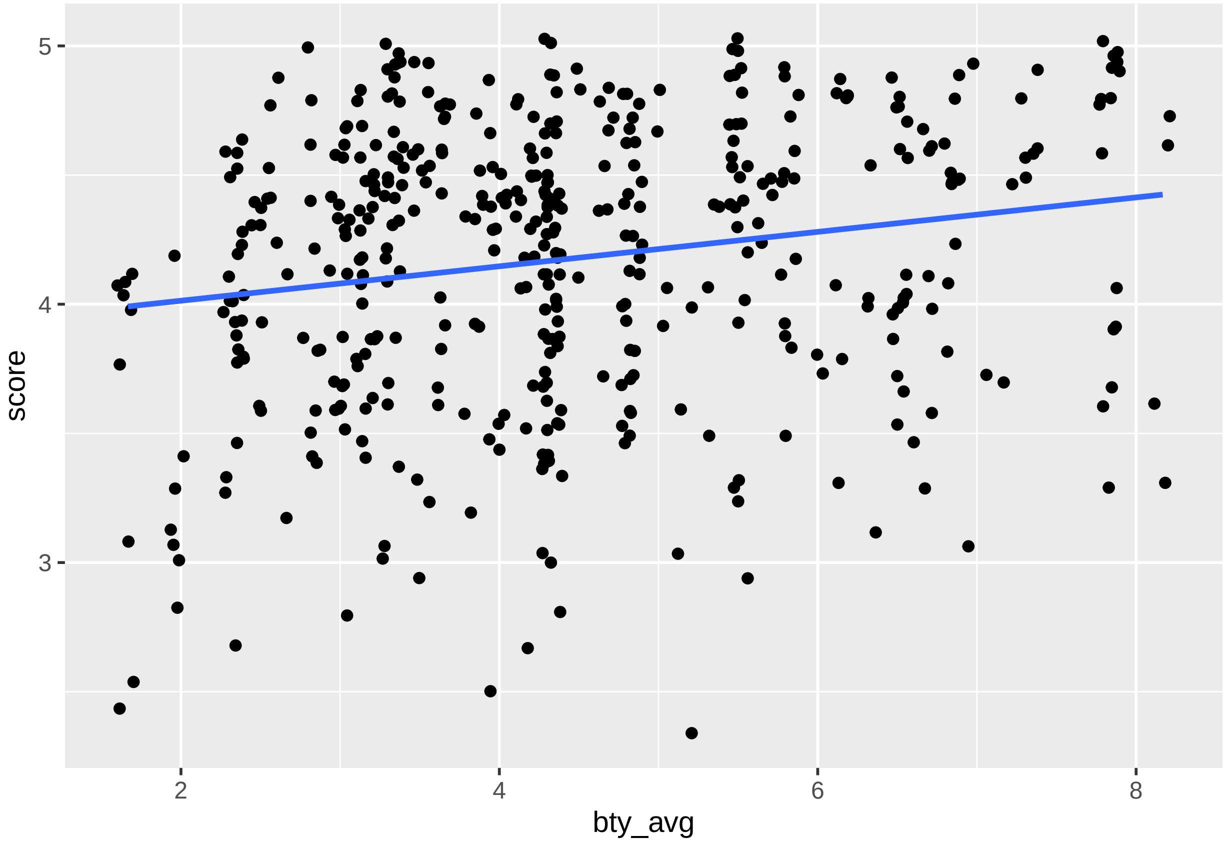
```
## `geom_smooth()` using formula 'y ~ x'
```



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)
```

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



Answer

Equation for the linear model. y = ax + b, where y = score, a = 0.06664, x = average beauty (bty_avg), b = 3.88034.

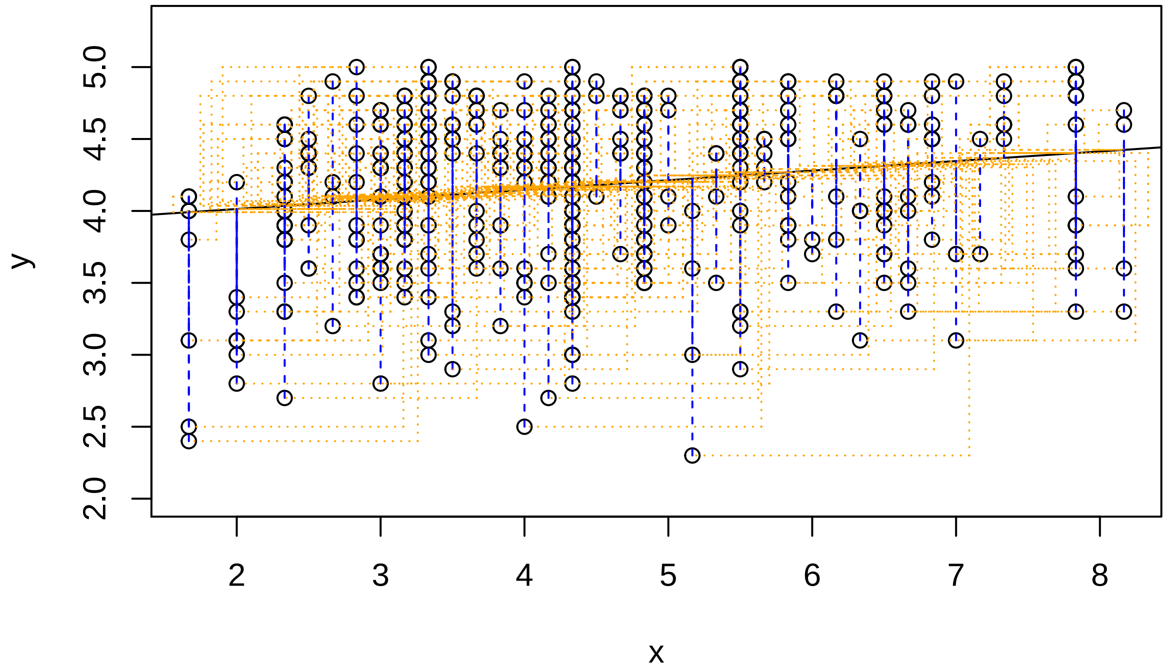
```
slope (a) = the ratio of professor score to professor beauty. this ratio being less than 1 means the in
The intercept means no matter how beautifull a professor is, he or she received a score of 3.88.
Based on the R-dquared (0.03502) , adjusted R-squared (0.03293) and p-value (0) , the average beauty sc
```

Exercice 6

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).

```r
library(statsr)
evals2 <- select(evals, bty_avg, score)
head(evals2)
```

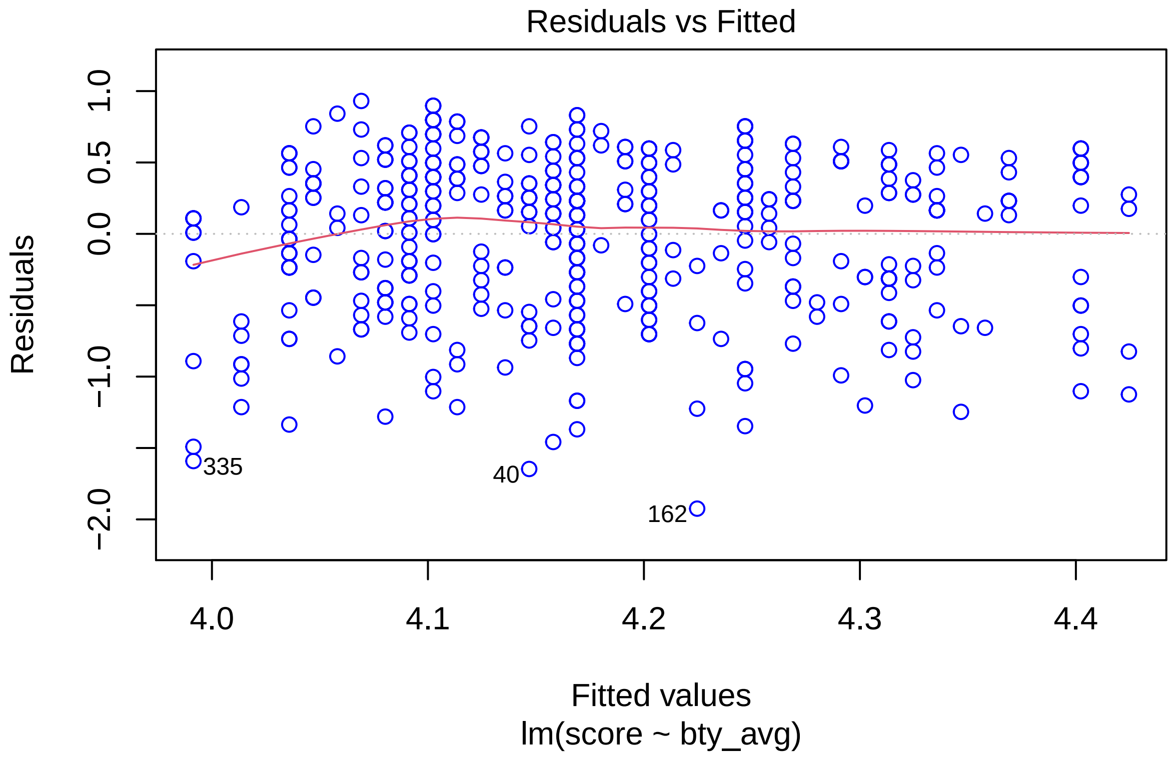
```r
plot_ss(x = evals2$bty_avg, y = evals2$score , showSquares = TRUE)
```



```
## 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
```

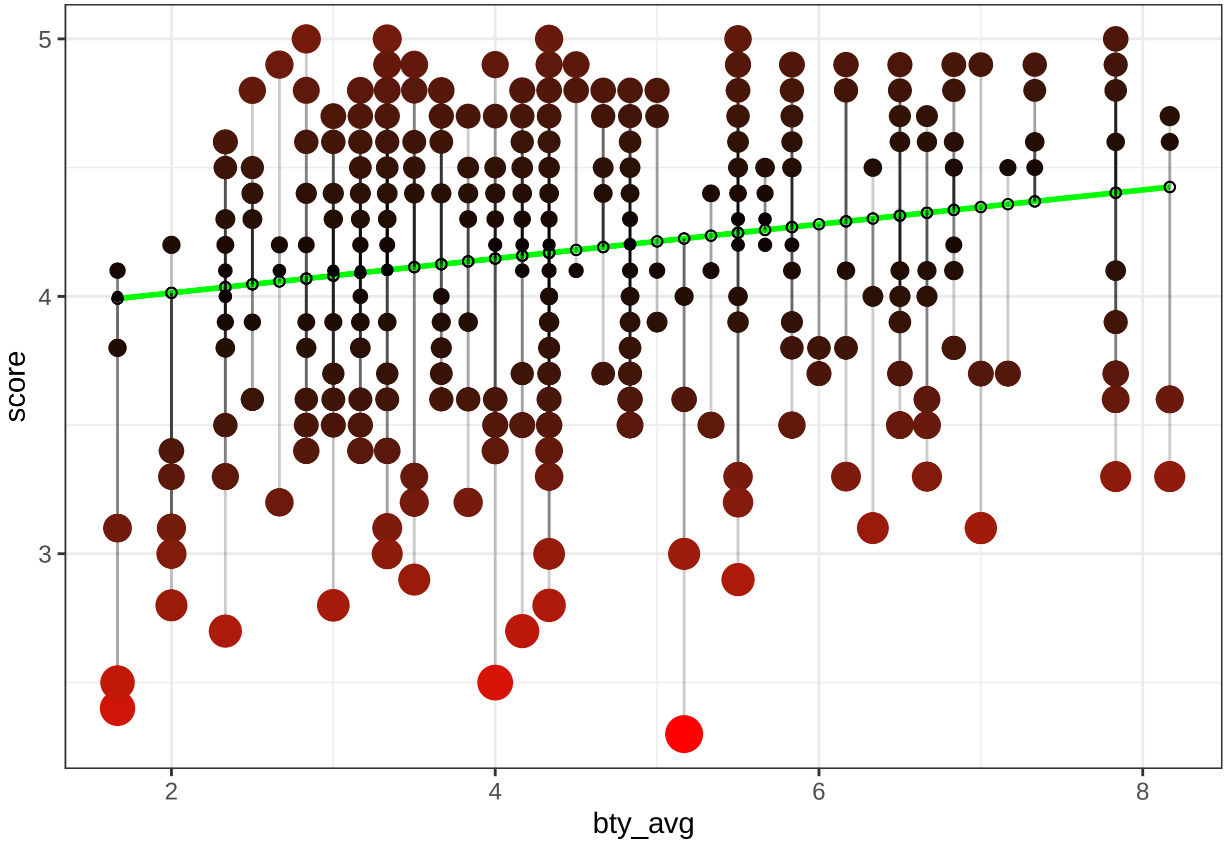
Just by looking at the regression line and squared residuals, we can tell that many points off. Let's recall that the Least Squares Regression Line is the line that makes the vertical distance from the data points to the regression line as small as possible. It's called a "least squares" because the best line of fit is one that minimizes the variance (the sum of squares of the errors).

```
evals1_fit <- lm(score~bty_avg, data = evals2)
plot(evals1_fit, which = 1, col = c("blue"))
```



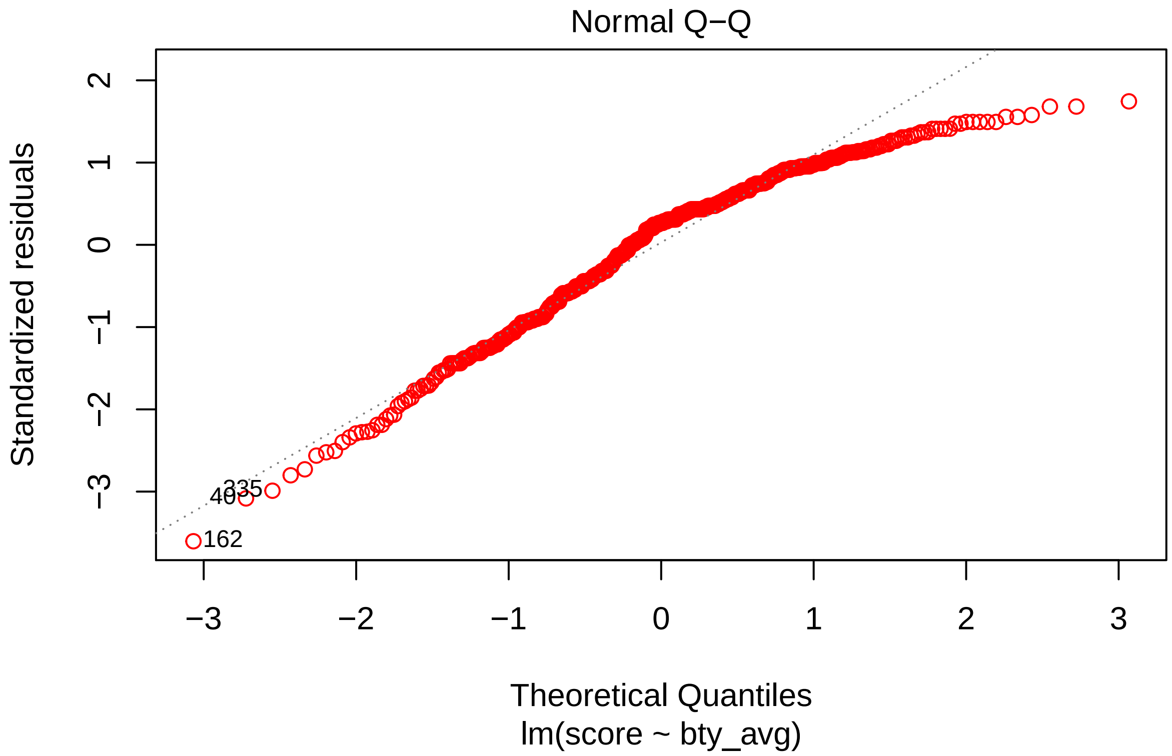
```
#summary(evals1_fit)
# let's add predicted and residual into data
evals2$predicted <- predict(evals1_fit)
evals2$residuals <- residuals(evals1_fit)
# Let's take a look at evals1_fit
#glimpse(evals1_fit)

ggplot (evals2, aes(x = bty_avg, y = score)) +
  geom_smooth(method = "lm", se = FALSE, color = "green") +
  geom_segment(aes(xend = bty_avg, yend = predicted), alpha = .2) +
  geom_point(aes(color = abs(residuals), size = abs(residuals))) +
  scale_color_continuous(low = "black", high = "red") +
  guides(color = FALSE, size = FALSE) +
  geom_point(aes(y = predicted), shape = 1) +
  theme_bw()
```

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



The Residuals Vs Fitted show points that are not about the same on each side of the x-axis, so there is no constant variablitiy . This time we wanted to highlight the residual for a better visualization. the red points show residuals points, green points show predicted points and black points show actual points. So, we can see that one of the conditions of least squares regression, linearity is not reasonable. Data does not show a linear trend. Let's evaluate normal residuals.

```r
plot(evals1_fit, which = 2, col = c("red")) #another way of plotting Q-Q plot
```



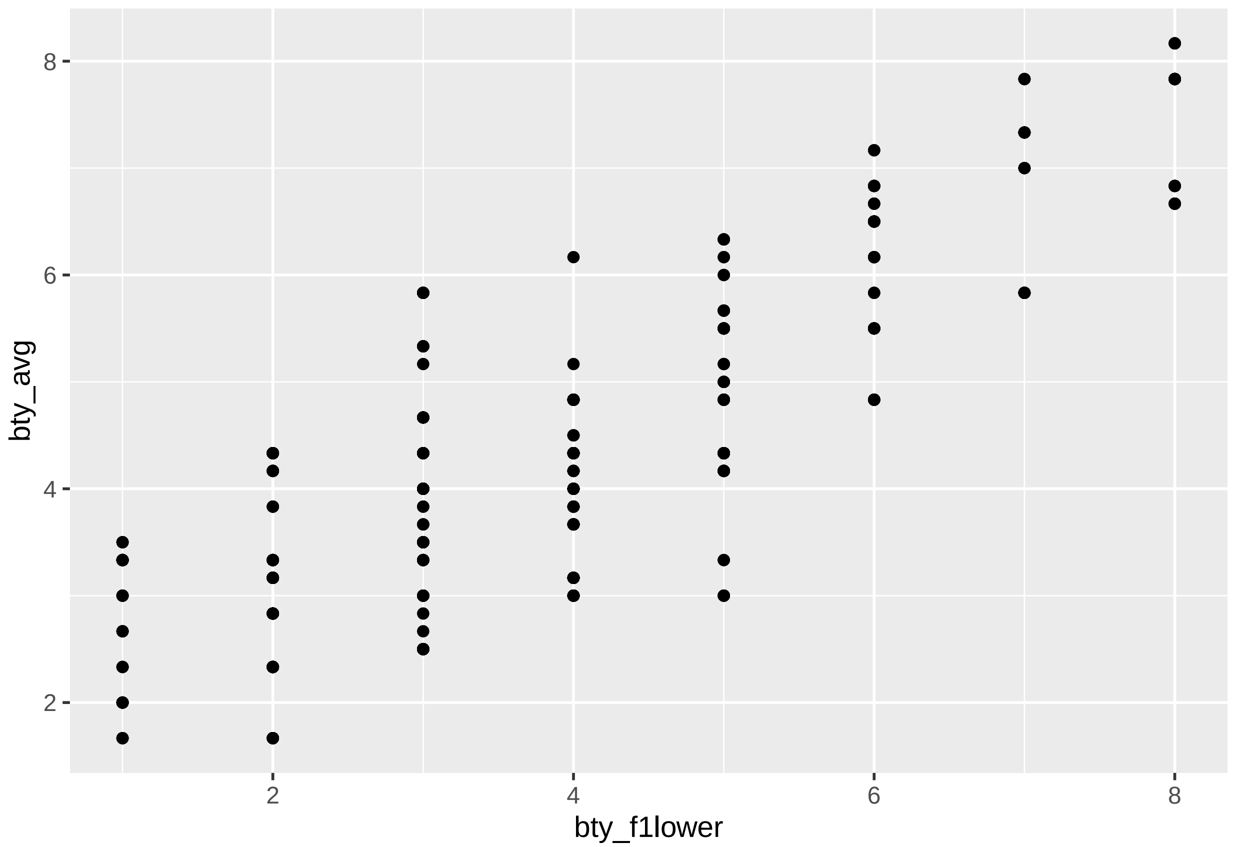
```
#qqnorm(evals1$residuals)
#qqline(evals1$residuals)
```

Based on the normal Q-Q plot , the residuals show a deviation from the line around the tails. This is an indicator that data is not normally distributed.

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.

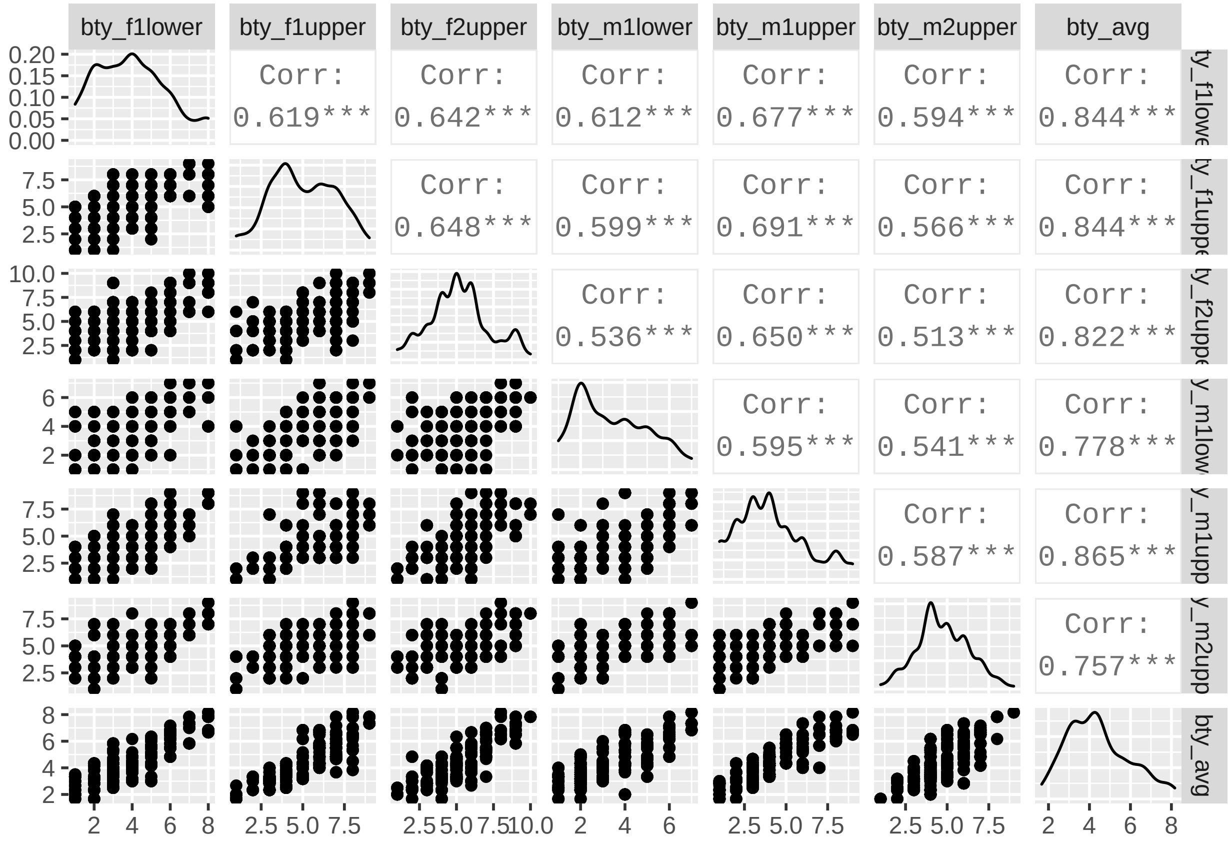
```
ggplot(data = evals, aes(x = bty_f1lower, y = bty_avg)) +
  geom_point()
```



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

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.

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

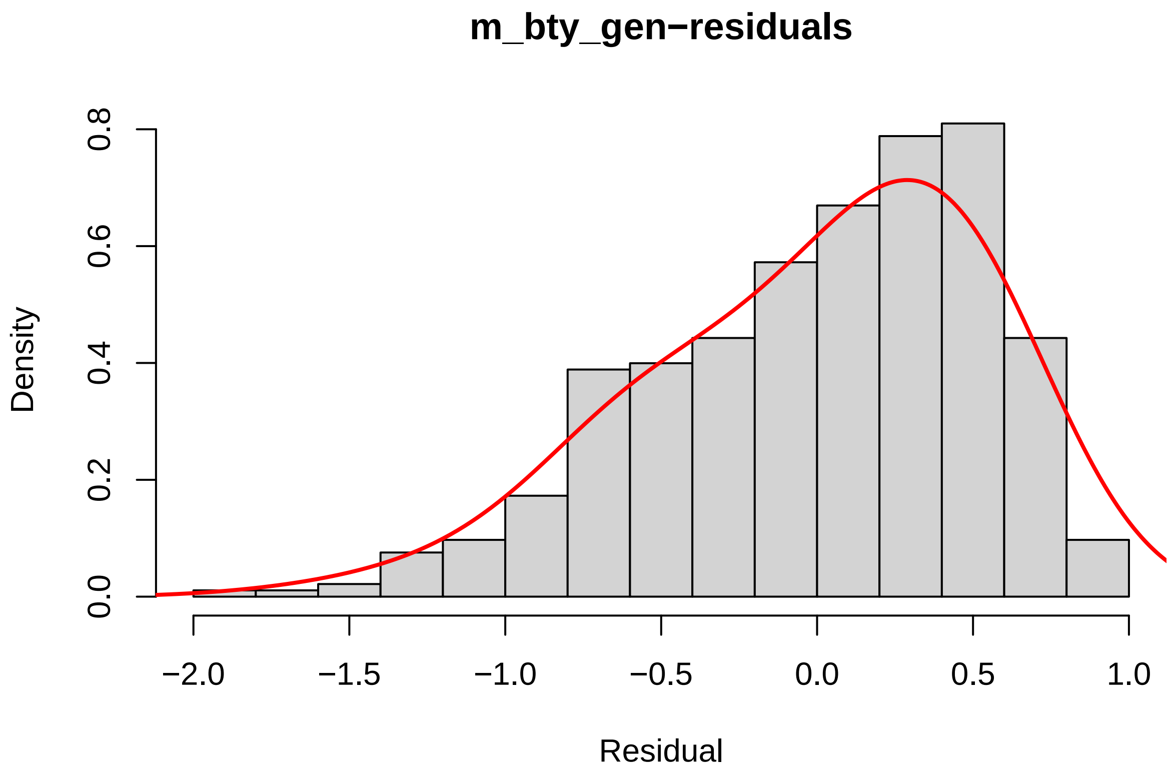
```
##
## Call:
## lm(formula = score ~ bty_avg + gender, data = evals)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.8305 -0.3625  0.1055  0.4213  0.9314
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.74734    0.08466  44.266  < 2e-16 ***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

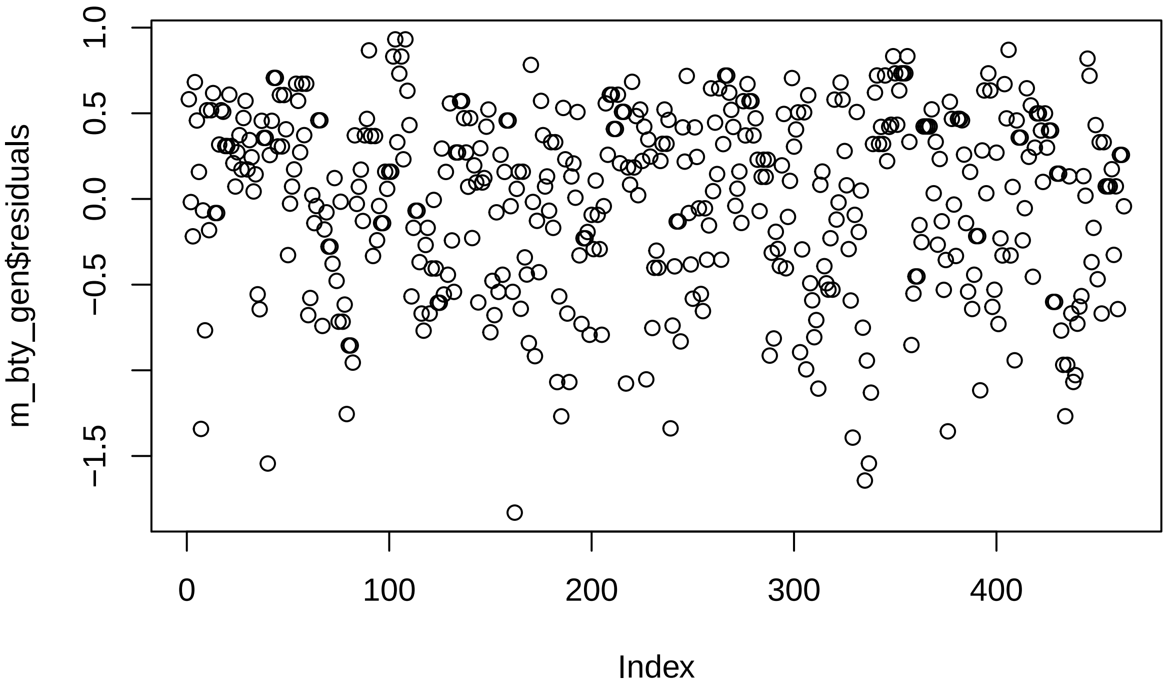
Exercice 7

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.

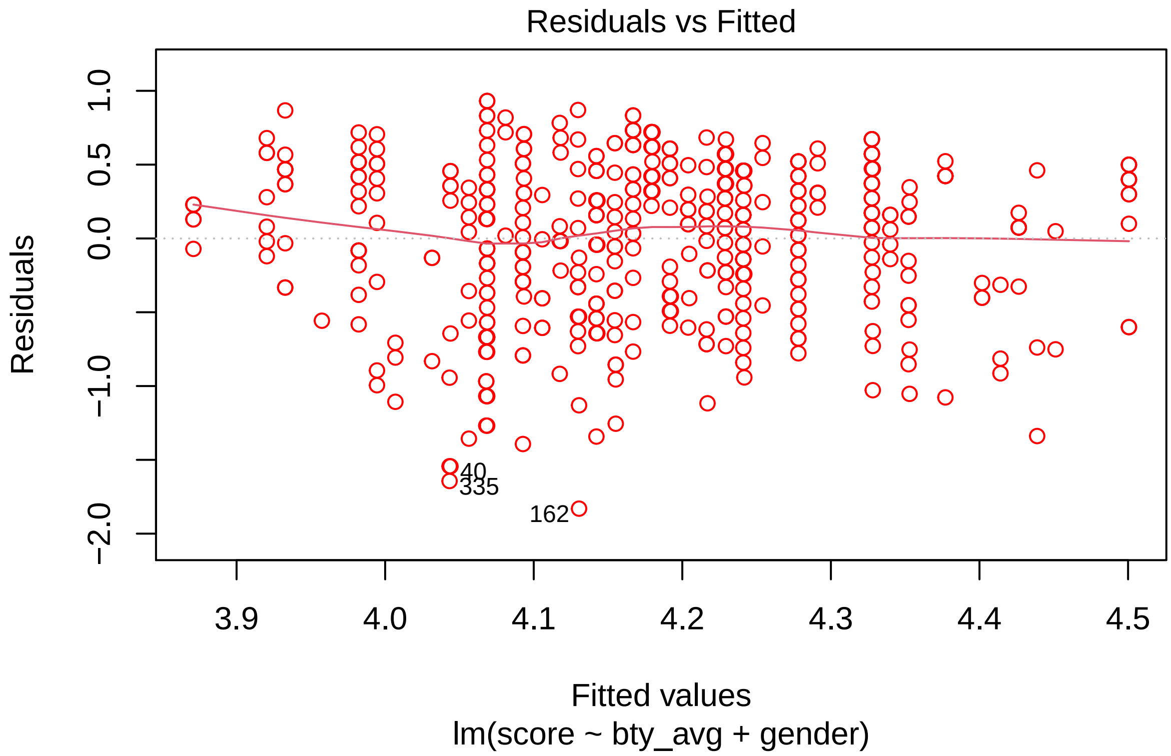
```
hist(m_bty_gen$residuals, prob = TRUE, breaks = 11,  main = "m_bty_gen-residuals", xlab = "Residual")
lines(density(m_bty_gen$residuals, adjust = 1.8), col = "Red", lwd = 2)
```



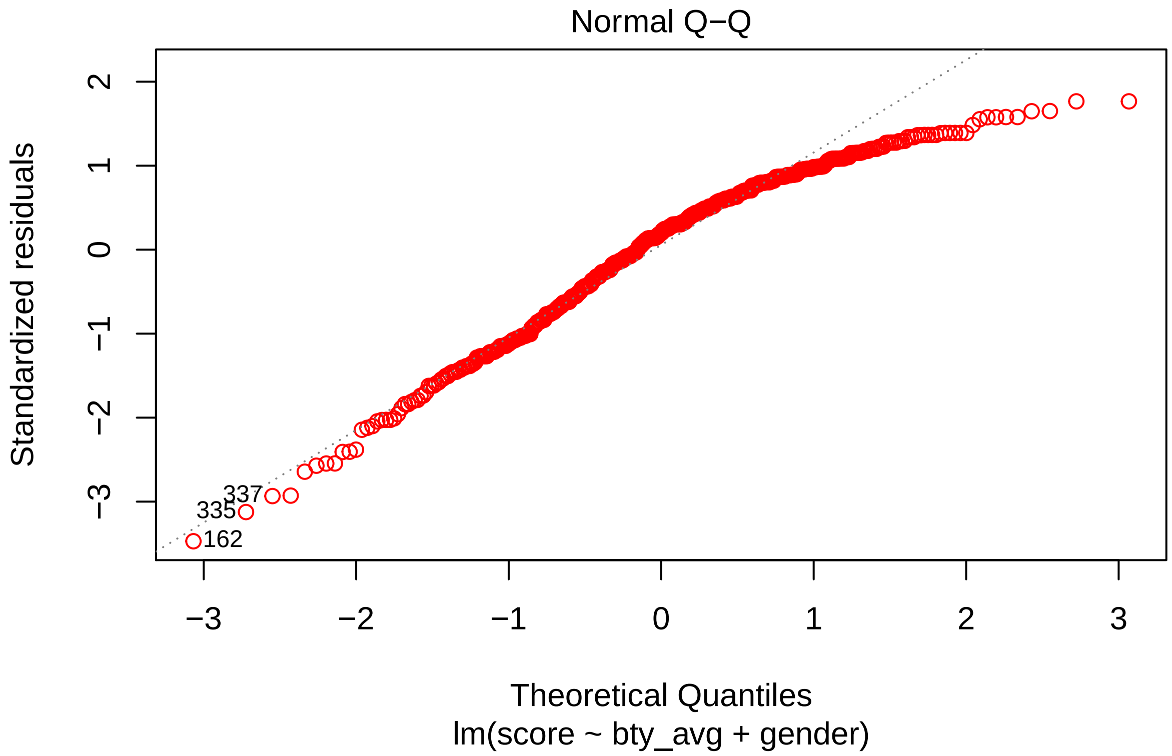
```
plot(m_bty_gen$residuals)
```



```
plot(m_bty_gen, which = 1, col = c("red"))
```



```r
plot(m_bty_gen, which = 2, col = c("red"))
```



The histogram of the residual shows a left skewed distributed data. this is not a normal distributed data. the residuals show a deviation from the line around the tails. This is an indicator that data is not normally distributed. The We are bit confused…

Exercice 8

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`?

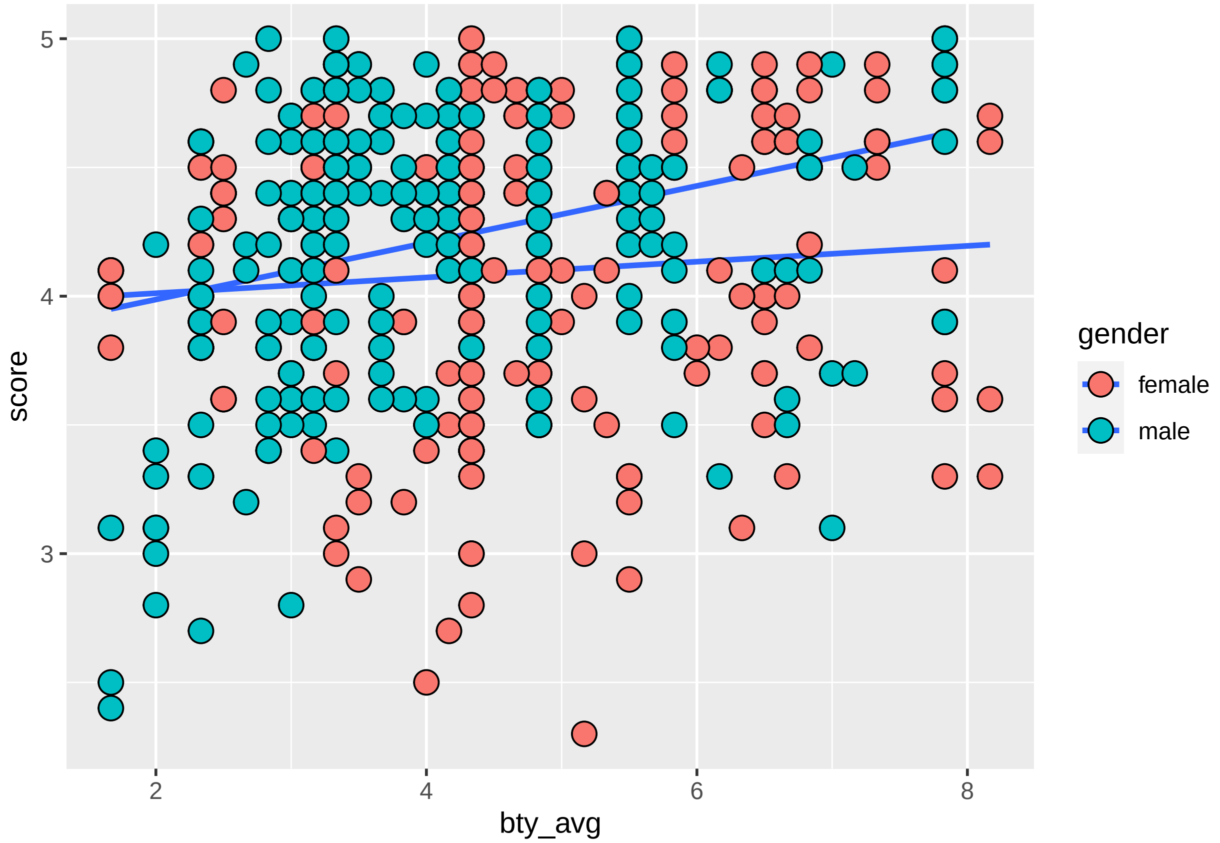
We are bit confused. Adding gender to the model decreased the p-value from 5.083e-05 to 8.177e-07

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
byt_df <- select(evals, gender, bty_avg, score)

ggplot(byt_df, aes(x = bty_avg, y = score, fill = gender)) +
  geom_smooth(method = "lm" , formula = y ~ x , se = FALSE) +
  geom_point(size = 4, shape = 21)
```



Exercice 9

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? ### Answer:

$$
\begin{aligned}
\widehat{score} &= \hat{\beta}_0 + \hat{\beta}_1 \times bty\_avg + \hat{\beta}_2 \times (x2) \\
&= 3.74734 + 0.07416 \times bty\_avg + 0.17239 \times (x2)
\end{aligned}
$$

### For two professors who received the same beauty rating,male professor tends to have the higher course evaluation score (strange from experience).

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.)

Exercice 10

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.

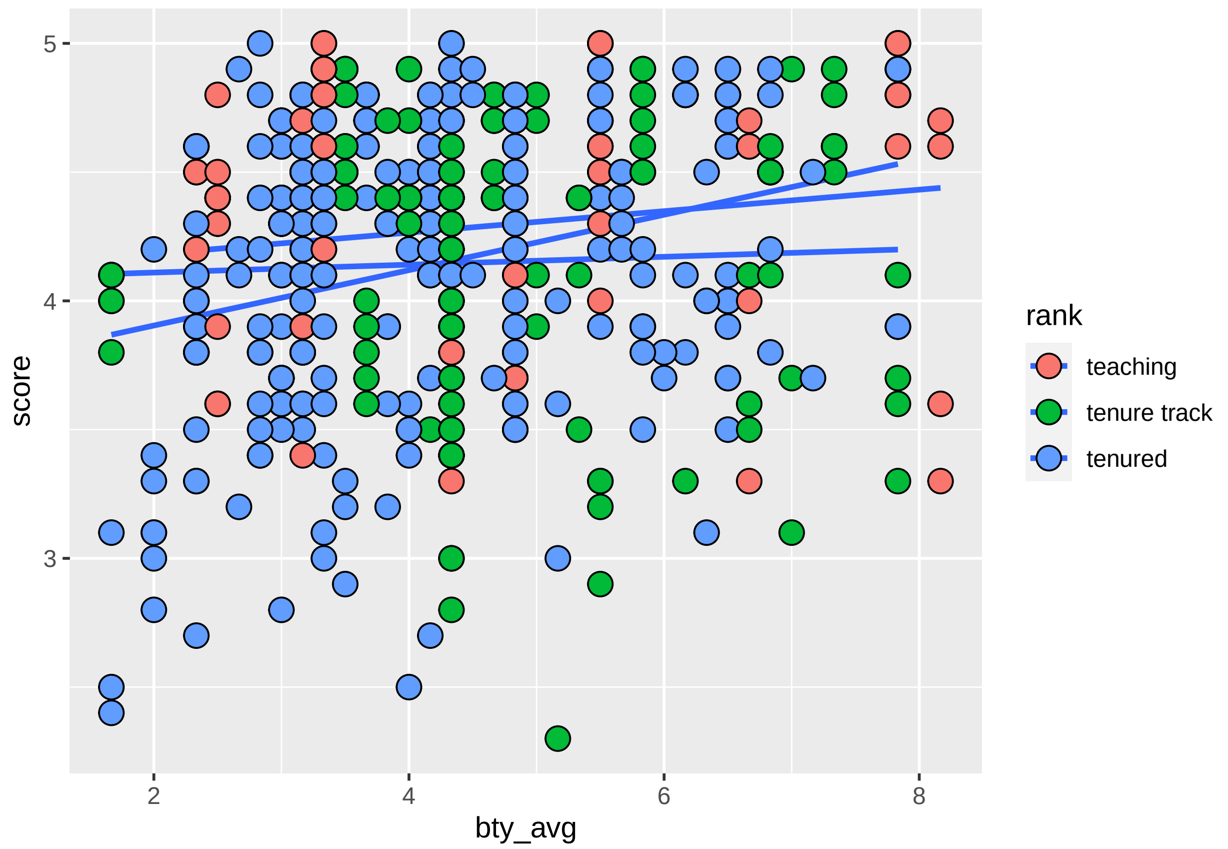
```
m_bty_rank <- lm(score~bty_avg + rank, data = evals)
summary(m_bty_rank)
```

```
##
## 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
```

```
head(evals$rank)
```

```
## [1] tenure track tenure track tenure track tenure track tenured
## [6] tenured
## Levels: teaching tenure track tenured
```

```
byt_df1 <- select(evals, rank, bty_avg, score)

ggplot(byt_df1, aes(x = bty_avg, y = score, fill = rank)) +
  geom_smooth(method = "lm" , formula = y ~ x , se = FALSE) +
  geom_point(size = 4, shape = 21)
```



R appear to handle categorical variable that have more than two levels well. In this case rank as 03 levels and teaching parameter estimate is multiplied by 0 while tenured and tenure track are multiplied by 1 respectively. We get the following equation of the line.

$$
\begin{aligned}
\widehat{score} &= \hat{\beta}_0 + \hat{\beta}_1 \times bty\_avg + \hat{\beta}_2 \times (x2) + \hat{\beta}_3 \times (x3) \\
&= 3.98155 + 0.06783 \times bty\_avg - 0.16070 \times (x2) - 0.12623 \times (x3)
\end{aligned}
$$

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.

Exercice 11

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. ### Answer: We would expect the cls_level (class level:lower, upper). We think class level has no association

with the professor score because professor score is independent of class level. We think each student regardless of his/her level care more about how good the professor is. The only apprehension on this choice is the judgement a student might have. Maybe students from upper level are more objective than those from lower level. This sounds more like a bias assumption since there is no fact to prove it.

Let's run the model…

```r
#?evals
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
```

Exercice 12

12. Check your suspicions from the previous exercise. Include the model output in your response. ### Answer: My suspicion was very close. cls_levelupper 0.0605140 0.0575617 1.051 0.29369

Exercice 13

13. Interpret the coefficient associated with the ethnicity variable. ### Answer: ethnicitynot minority coefficient is 0.1234929. This parameter estimate is multiplied by 1 in the equation of the line.

Exercice 14

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? ### Answer: the highest p-value is cls_profssingle -0.0146619 0.0519885 -0.282 0.77806

```
m_full <- 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)
```

```
##
## 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
```

After dropping cls_profs, the coefficients and significance of the other explanatory variables change a little bit cls_levelupper 0.0606374 0.0575010 1.055 0.292200 compared to previously..cls_levelupper 0.0605140 0.0575617 1.051 0.29369

Exercise 15

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 <- lm(score ~ gender + ethnicity + language + age + cls_perc_eval + cls_credits + bty_avg + pic_c

summary(m_full)
```

```
##
## 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
```

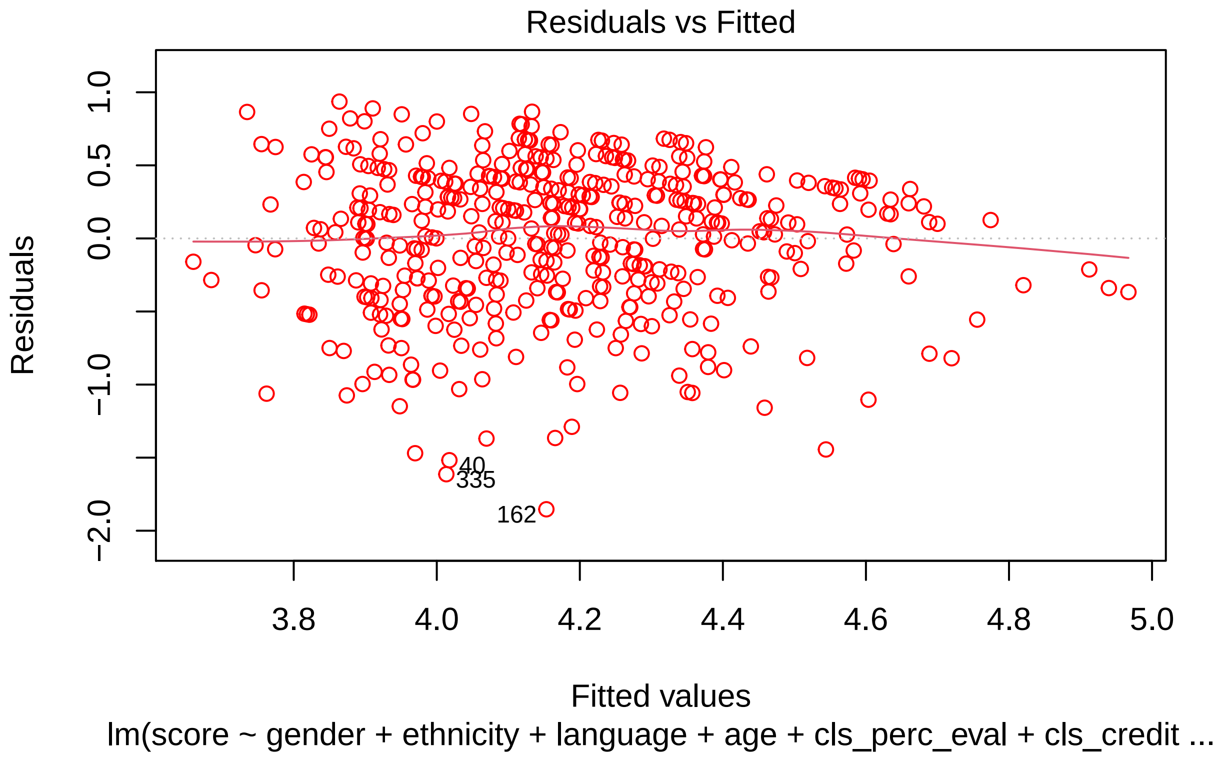
```
#?evals
```

$$\widehat{score} = 3.771922 + 0.051069 \times bty\_avg + 0.207112 \times (gendermale) + 0.167872 \times (ethnicitynotminority) - 0.206178 \times (lan$$

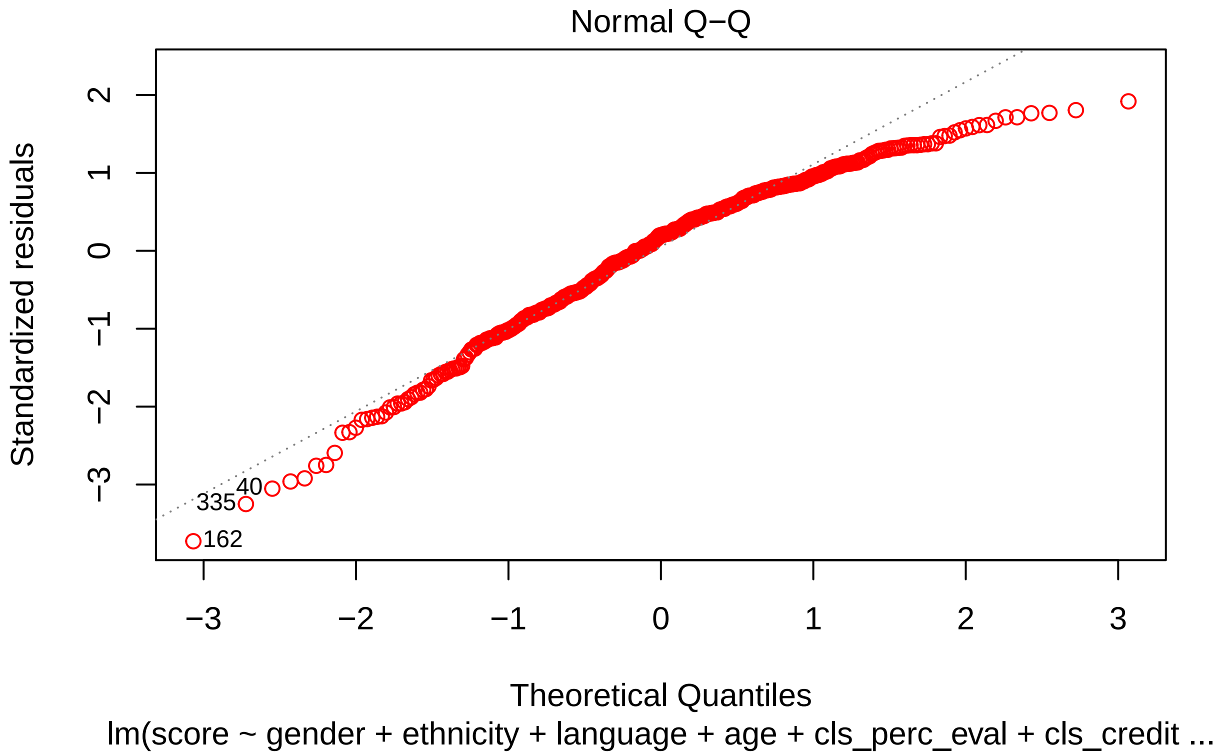
Exercise 16

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

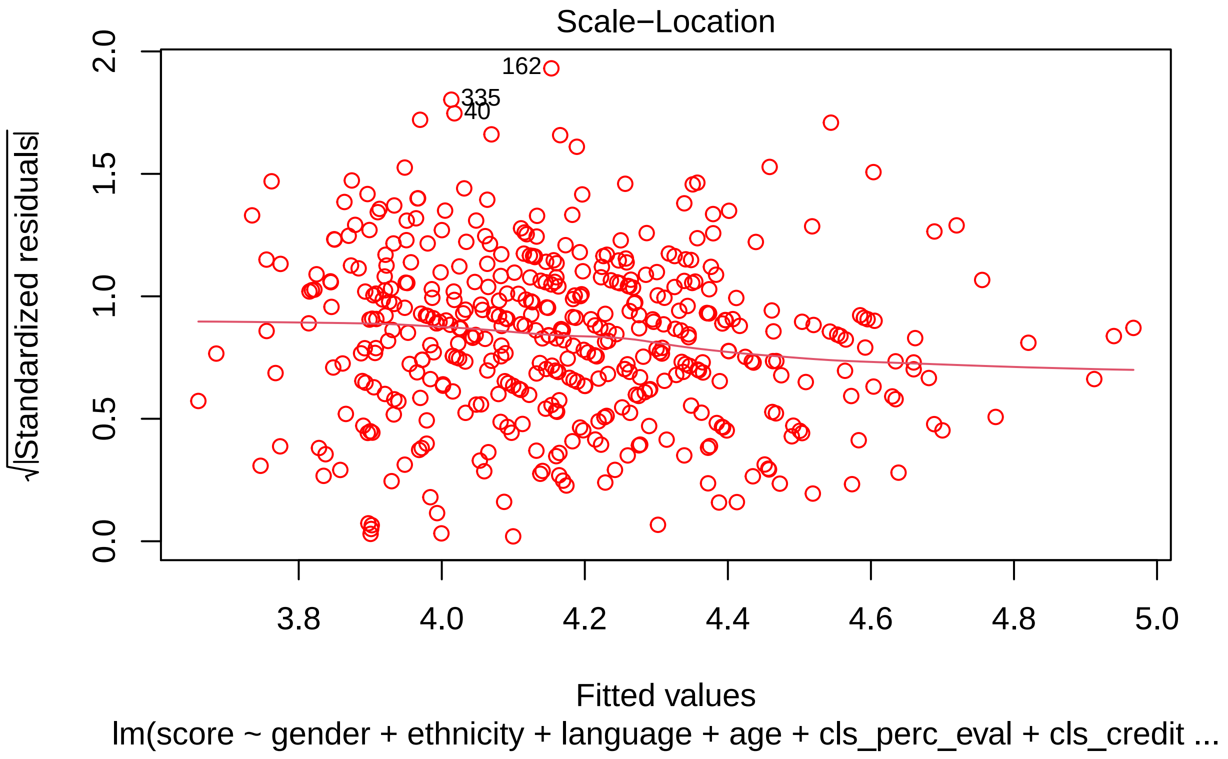
```
#hist(m_full$residuals, prob = TRUE, breaks = 11,  main = "m_full-residuals", xlab = "Residual")
#lines(density(m_full$residuals, adjust = 1.8), col = "Red", lwd = 2)

#plot(m_full)
plot(m_full, which = 1, col = c("red"))
```



```
plot(m_full, which = 2, col = c("red"))
```



```r
plot(m_full, which = 3, col = c("red"))
```



Answer: Based on the Normal Q-Q plots the residual of the model appears nearly normal (slightly left skewed). We really don't see outliers , however, the tails are off line Variability of the residuals is nearly constant: Based on the scale-location plot, we see that the residuals are equaly spread along the ranges of predictors. Each variable is linearly related to the outcome: based on the Residuals vs Fitted plot, there is no distinct patterns, so we are tempted to say we don't have non-linear relationship, however, the spread in the residuals is not equally distributed around the horizontal line. It looks like we have a "V" form.

Exercice 17

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? #### Answer: Yes and no, if each row represents a course, then we would like to know whether there are students who take one course or multiple courses and whether there are professors who teach more than one class. The assumption is that a student taking only one course might have a different opinio than the same student taking 02 courses with the same professor. It is hard to have the same goodness on two different topics.

Exercice 18

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. #### Answer: Students tend to

like mostly male teachers from not minority, with speaking english, teaching one credit and have color picture.

Exercice 19

19. Would you be comfortable generalizing your conclusions to apply to professors generally (at any university)? Why or why not? #### Answer: No, we Wouldn't be comfortable generalizing our conclusions to apply to professors generally (at any university). there is a bias in the assumption. Generalizing these conclusions would mean students are about the same all region. Whereas, how students appreciate a teacher is more a sociology/psychology aspect than it is a scientific law. For example, Students who go to school in the region where graduation and retention rate are very high would have a different feedback than those who go to school in the region where graduation and retention rate are very low.

References

---

https://fall2020.data606.net/assignments/labs/

https://www.statisticshowto.com/least-squares-regression-line/