Multiple linear regression

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

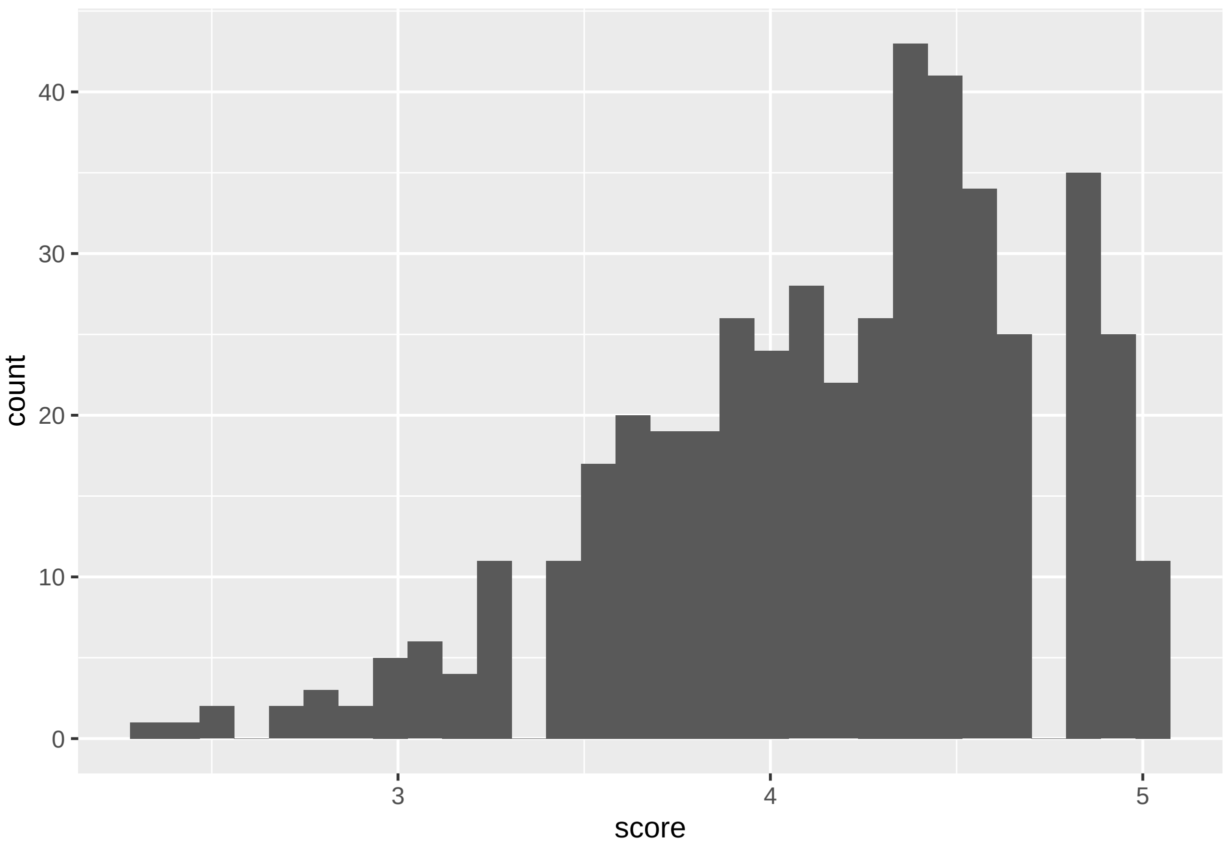
1. Is this an observational study or an experiment? The original research question posed in the paper is whether beauty leads directly to the differences in course evaluations. Given the study design, is it possible to answer this question as it is phrased? If not, rephrase the question.

- ANSWER: Observational studies are observing a variable naturally, without intervening (which would be an experiment). That is what's occurring here.
- ANSWER: Whether beauty leads "directly" to the difference in course evaluations (causation) cannot be *proven*. The probability of a connection between the two is what can be assessed. A better, rephrased question might be: Based on these data, does there appear to be an association between perceived beauty scores and course evaluation scores?

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?

- ANSWER: This distribution is negatively or left skewed, with the average of the ratings (roughly 4.2) higher than the average of the scale (2.5). Students tend to give positive ("above (scale) average") reviews. This is approximately what I expected due to survivorship bias: Teachers who manage to become and stay as professors tend to be good teachers.

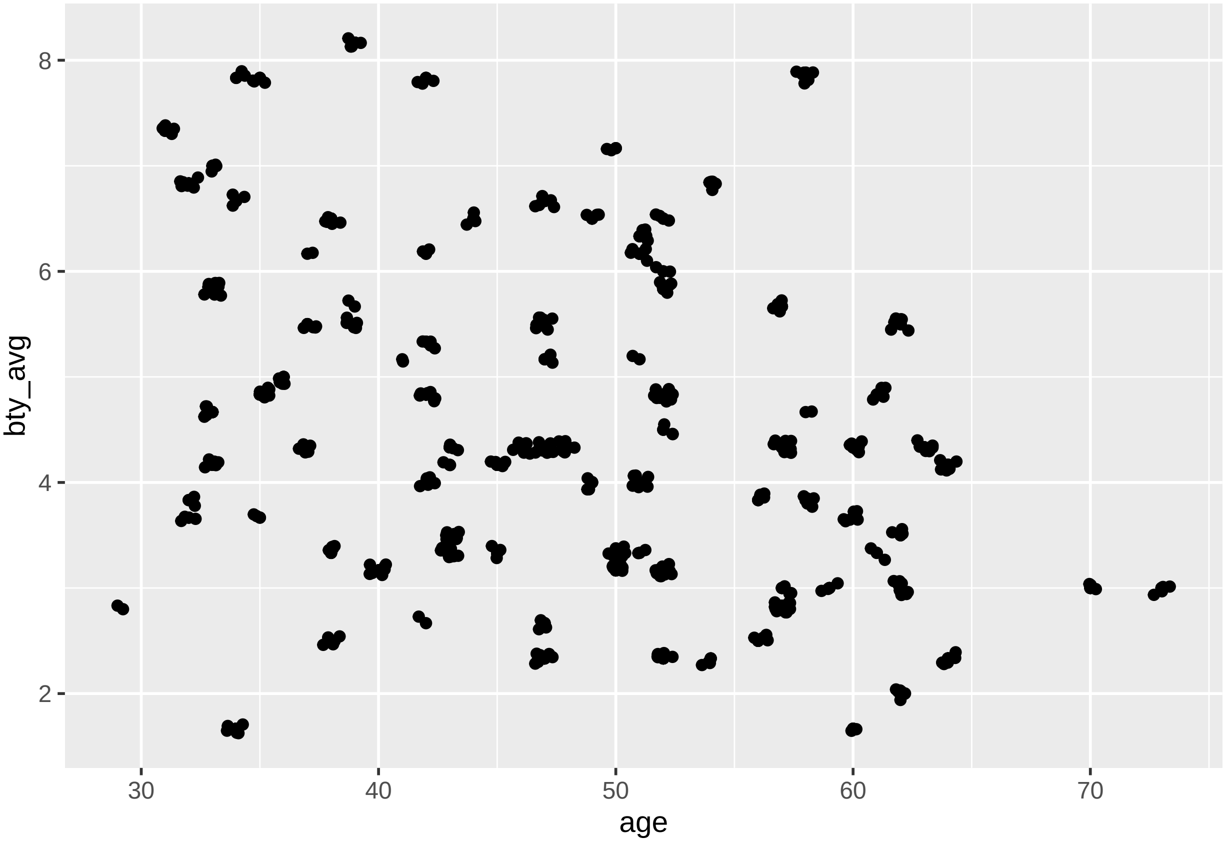
```
ggplot(evals,aes(x=score)) +
  geom_histogram()
```



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

- ANSWER: Below I look at the relationship between age and average beauty score. There's no obvious relationship between age and beauty score. Perhaps at the high end of "bty_avg" the "age" tends to be younger.

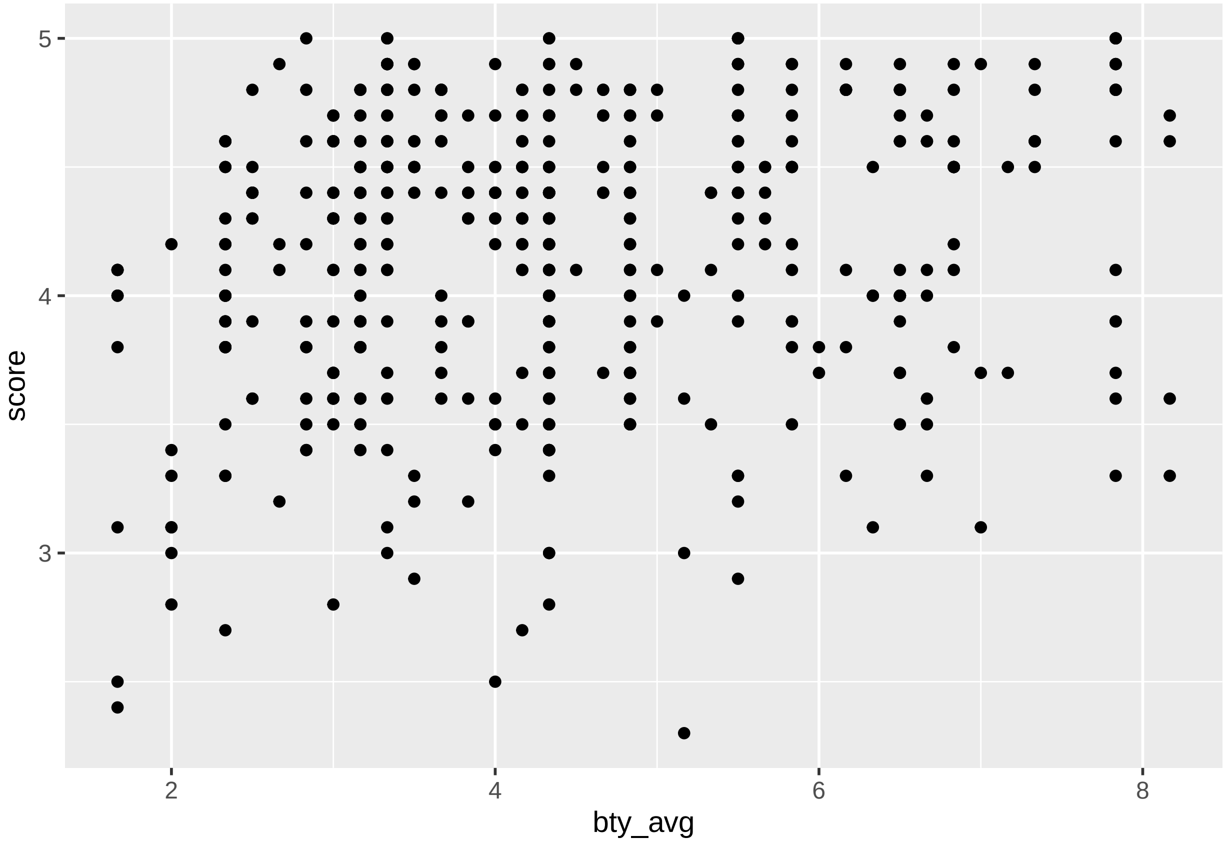
```
ggplot(data = evals, aes(x = age, y = bty_avg)) +
  geom_jitter() +
  geom_point()
```



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?

- The dataframe "evals" has 464 observations and there are appear to be fewer points on the scatterplot, perhaps due to missing "score" or "bty_avg" values.

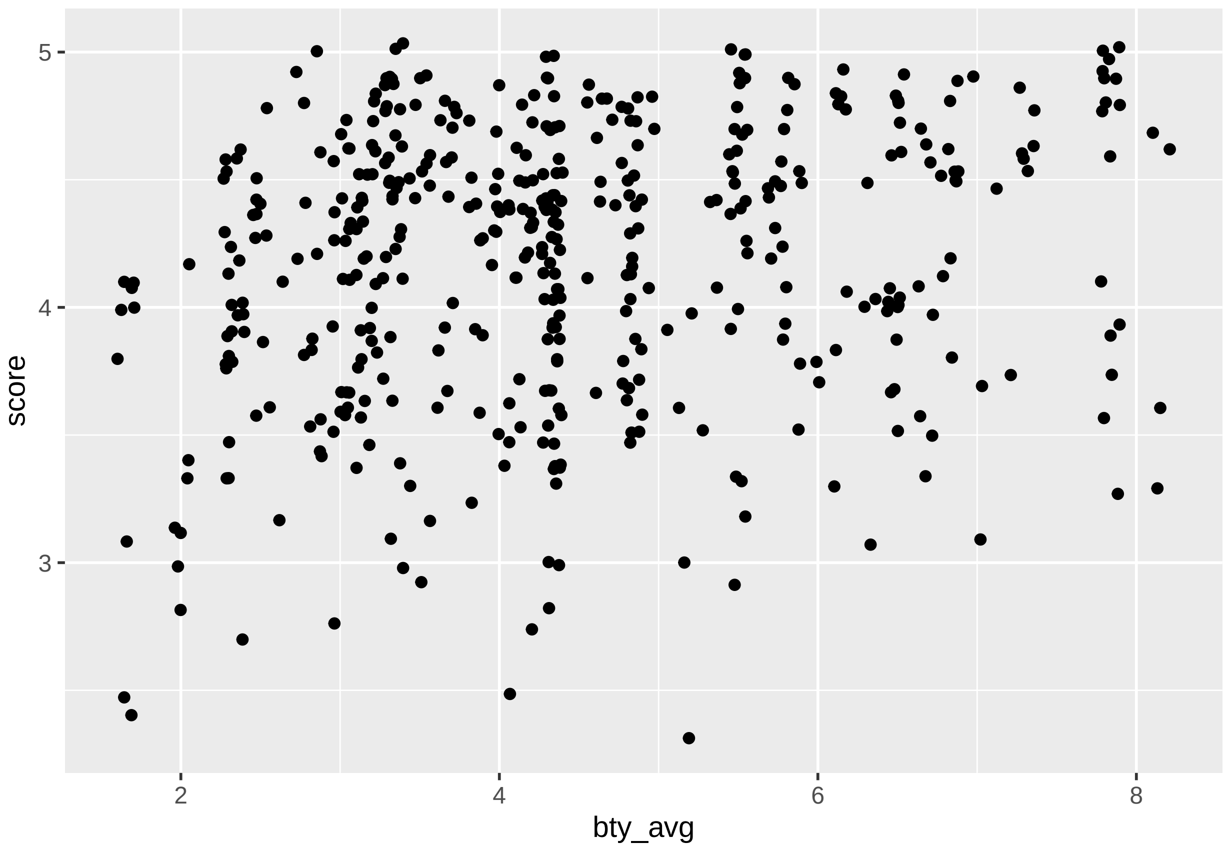
```
nrow(evals)
```

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

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

- ANSWER: The scatterplot had many observations with highly similar values, such that they couldn't be differentiated graphically on the scatterplot. The jitter fixes this.

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



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?

- ANSWER: y hat m_bty = 3.88034 (y intercept) + 0.06664*evals$bty_avg
- ANSWER: Interpretation of slope: Assuming a base of 3.88 score at 0 bty_avg, for every increase of 1 in bty_avg, score increases by 0.06664.
- ANSWER: Average beauty score appears to be a statistically significant, with a p value of 0.0000508. Assuming a standard significance threshold of less than 0.05 (likelihood of relationship occurring due to randomness), the average beauty score `bty_avg` appears to be a statistically significant predictor of the average professor evaluation score `score`.

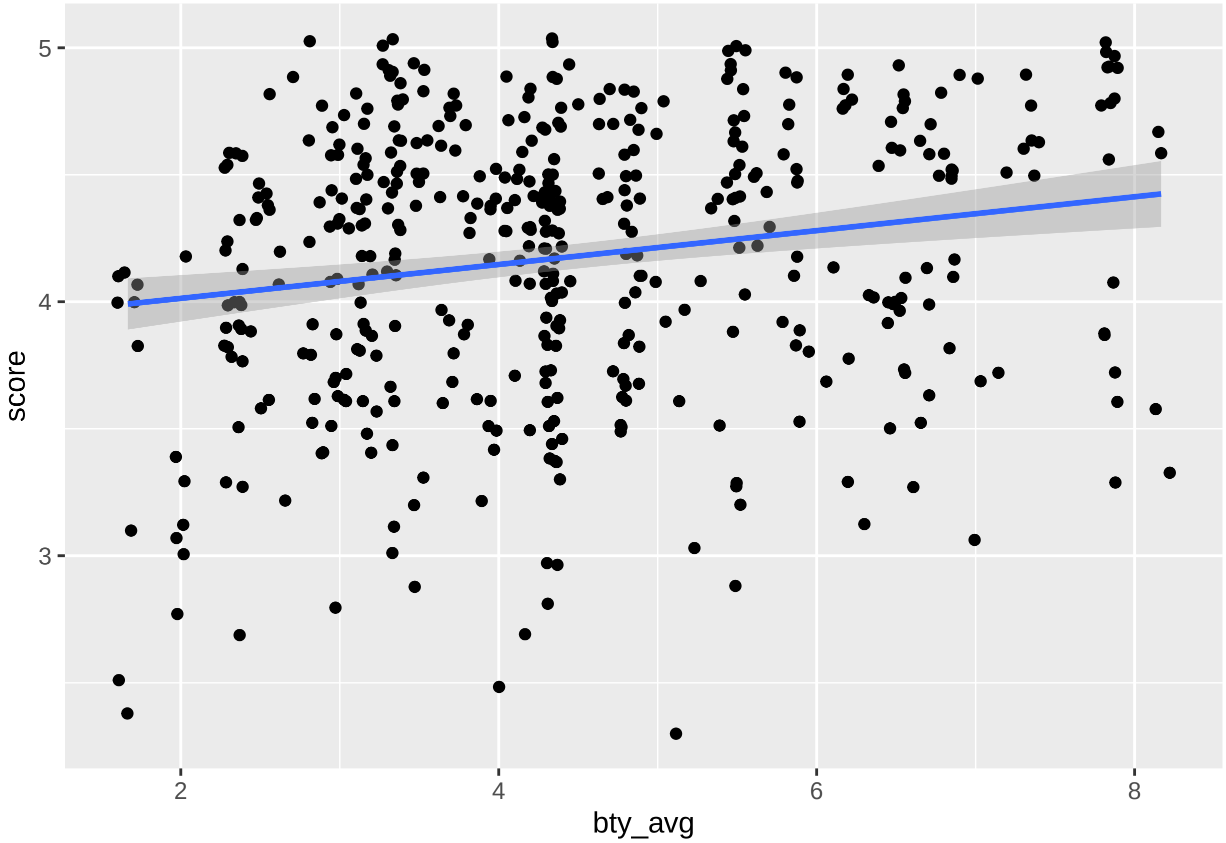
```
m_bty <- lm(score ~ bty_avg, data=evals)
summary(m_bty)
```

```
##
## 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
```

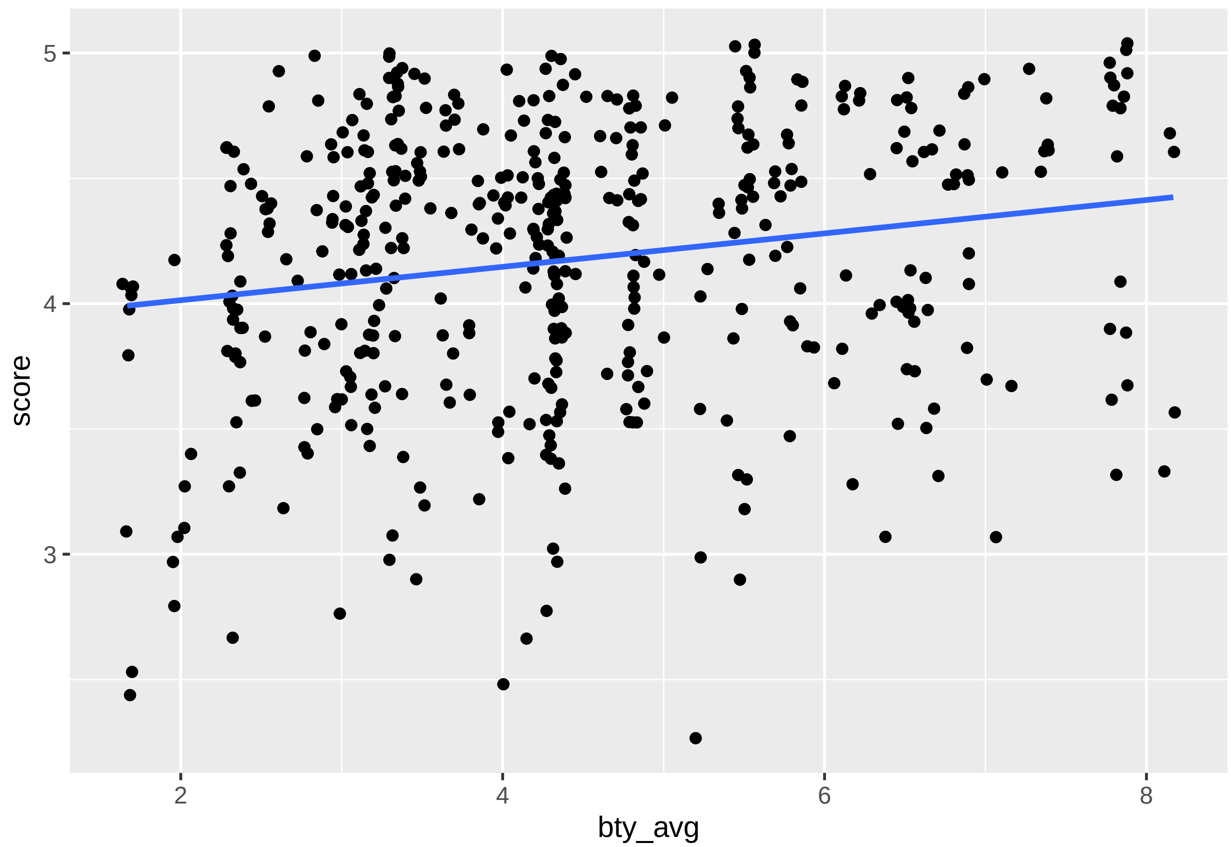
Add the line of the best fit model to your plot using the following:

```
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)
```

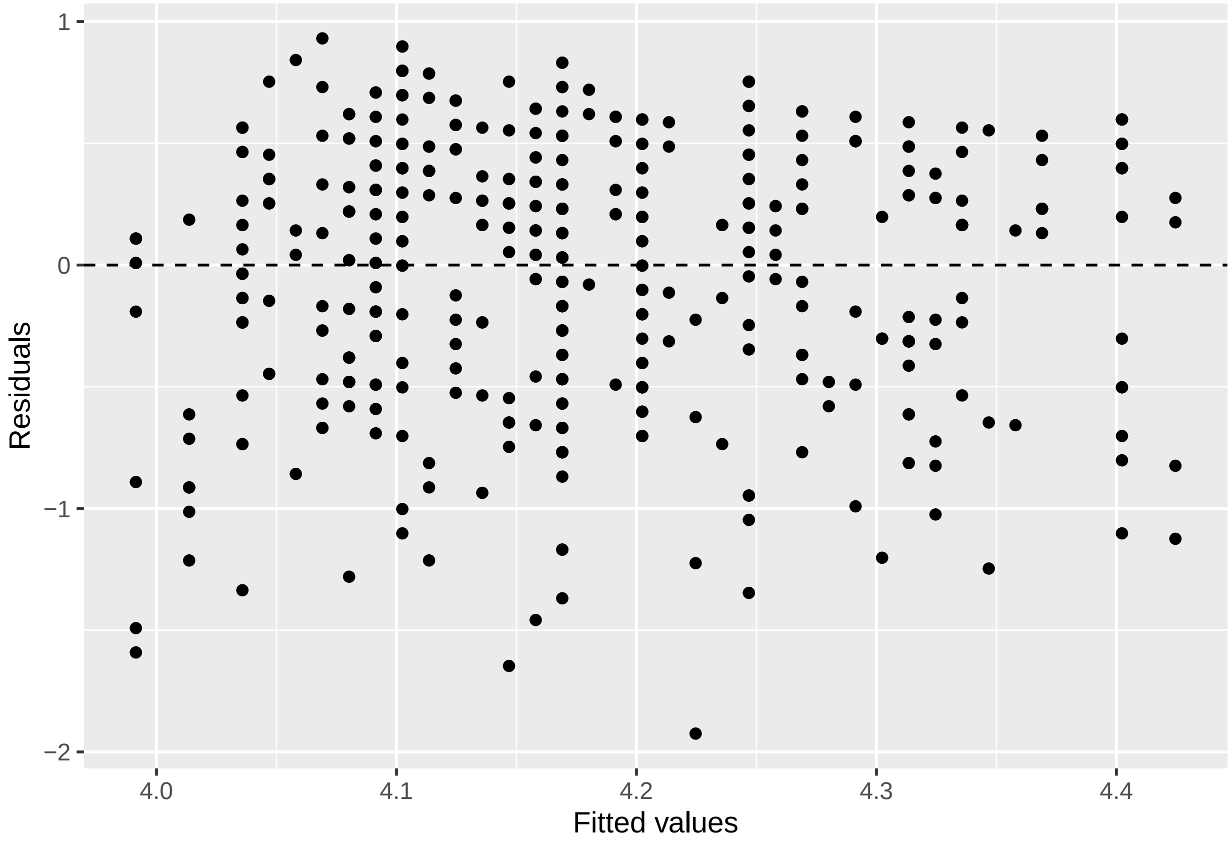


6. Use residual plots to evaluate whether the conditions of least squares regression are reasonable. Provide plots and comments for each one (see the Simple Regression Lab for a reminder of how to make these).

- ANSWER: Conditions of least squares regression: 1) Linearity; 2) Nearly normal residuals; 3) Constant variability. Each is evaluated below:

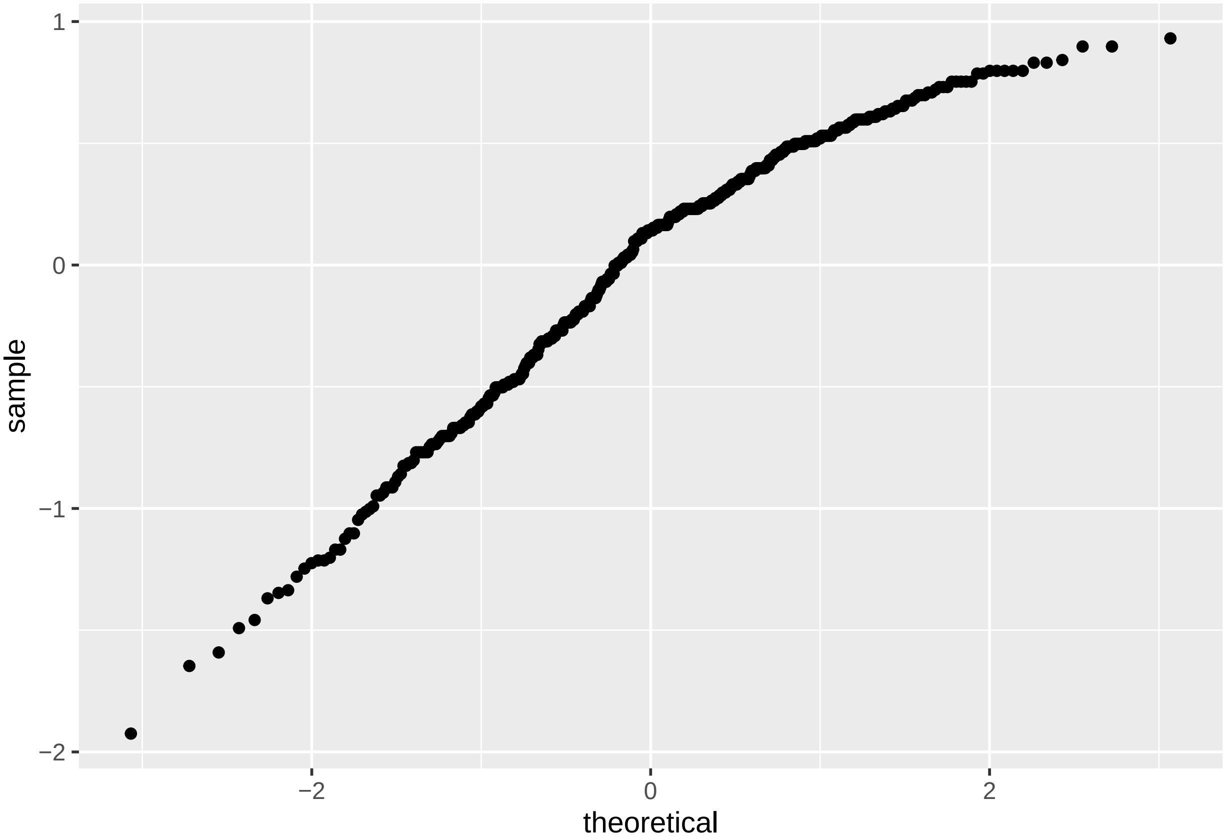
Linearity: Relationship appears linear. The variability of the residuals is approximately constant across the distribution, without curvature or indication of non-normality. There is no apparent pattern in the residuals plot, which indicates that the relationship is linear. The linear model is approximating the data points without favoring certain inputs.

```
ggplot(data = m_bty, aes(x = .fitted, y = .resid)) +
  geom_point() +
  geom_hline(yintercept = 0, linetype = "dashed") +
  xlab("Fitted values") +
  ylab("Residuals")
```



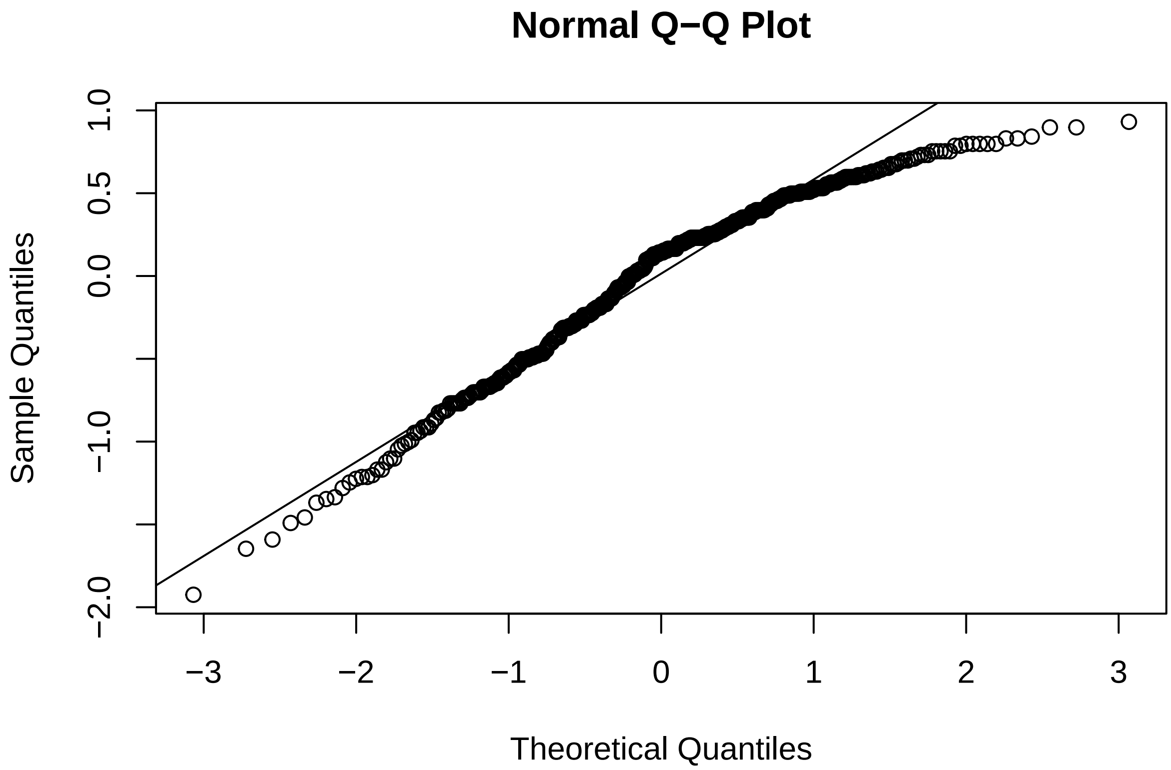
Nearly normal residuals: Normal probability plot of the residuals below. The relationship between "theoretical" and "sample" appears generally linear. There is a skew at higher values, however not enough to describe the relationship as non-linear.

```
ggplot(data = m_bty, aes(sample = .resid)) +
  stat_qq()
```



Constant variability: The constant variability condition states that the variability of points around the least squares line should be roughly constant. When I plotted using qqnorm and qqline below, there was a roughly constant relationship between the two axes. Therefore, yes, the constant variability condition appears to be met.

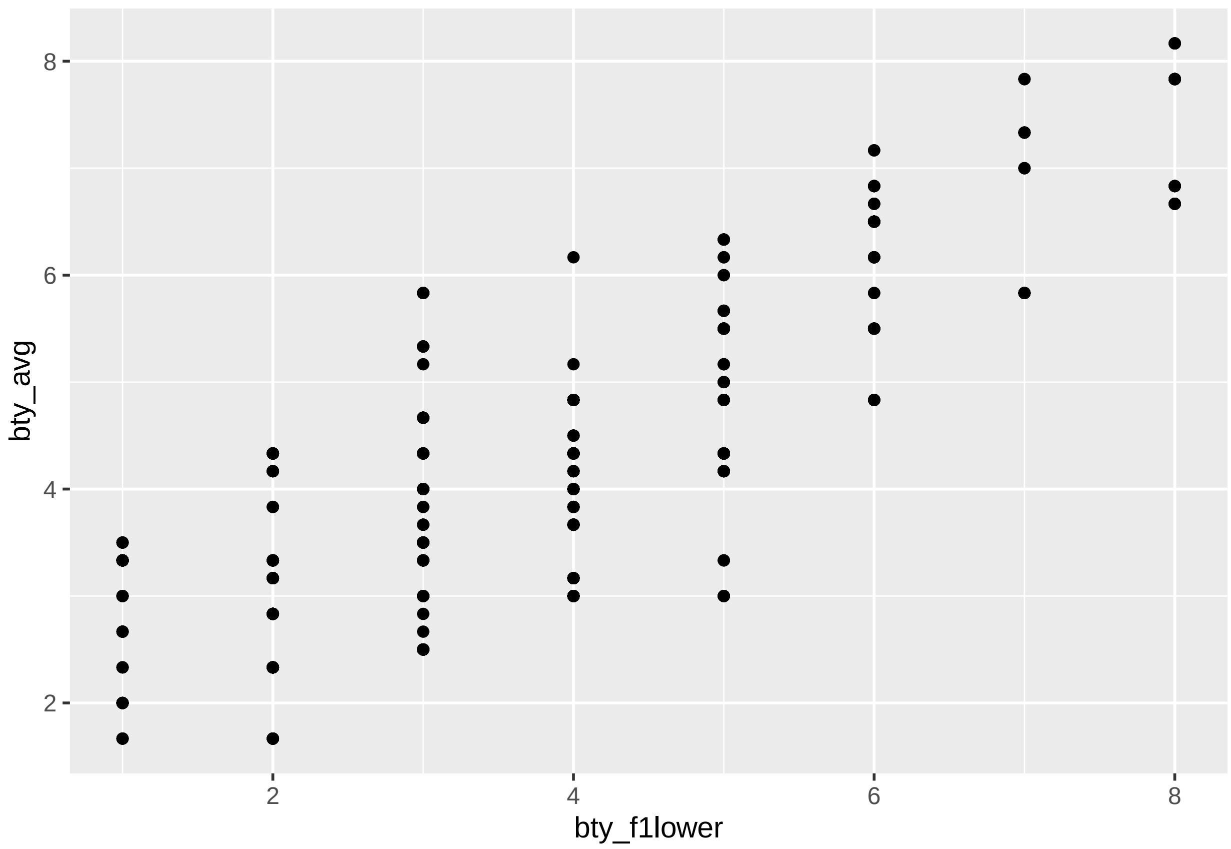
```
qqnorm(m_bty$residuals)
qqline(m_bty$residuals)
```



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()
```

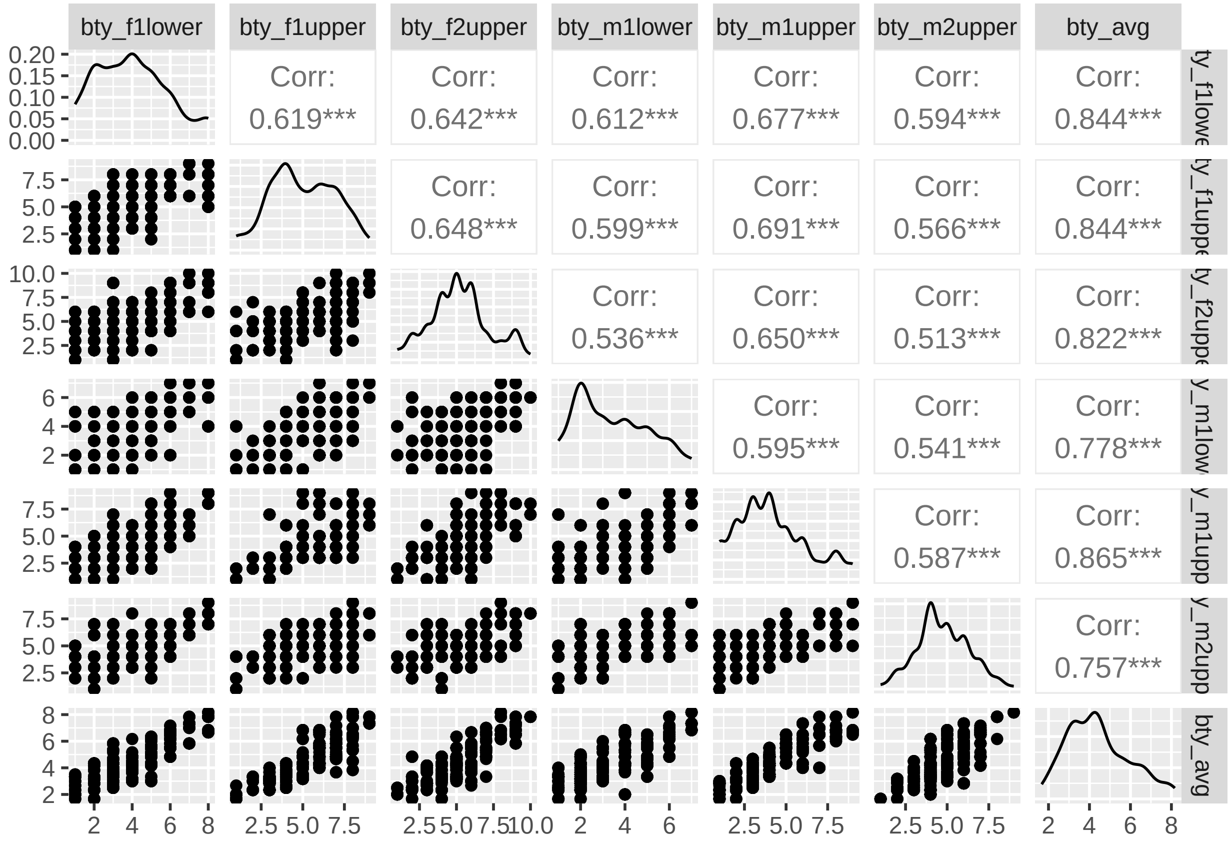


```r
evals %>%
  summarise(cor(bty_avg, bty_f1lower))
```

```
## # 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:

```r
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)
```

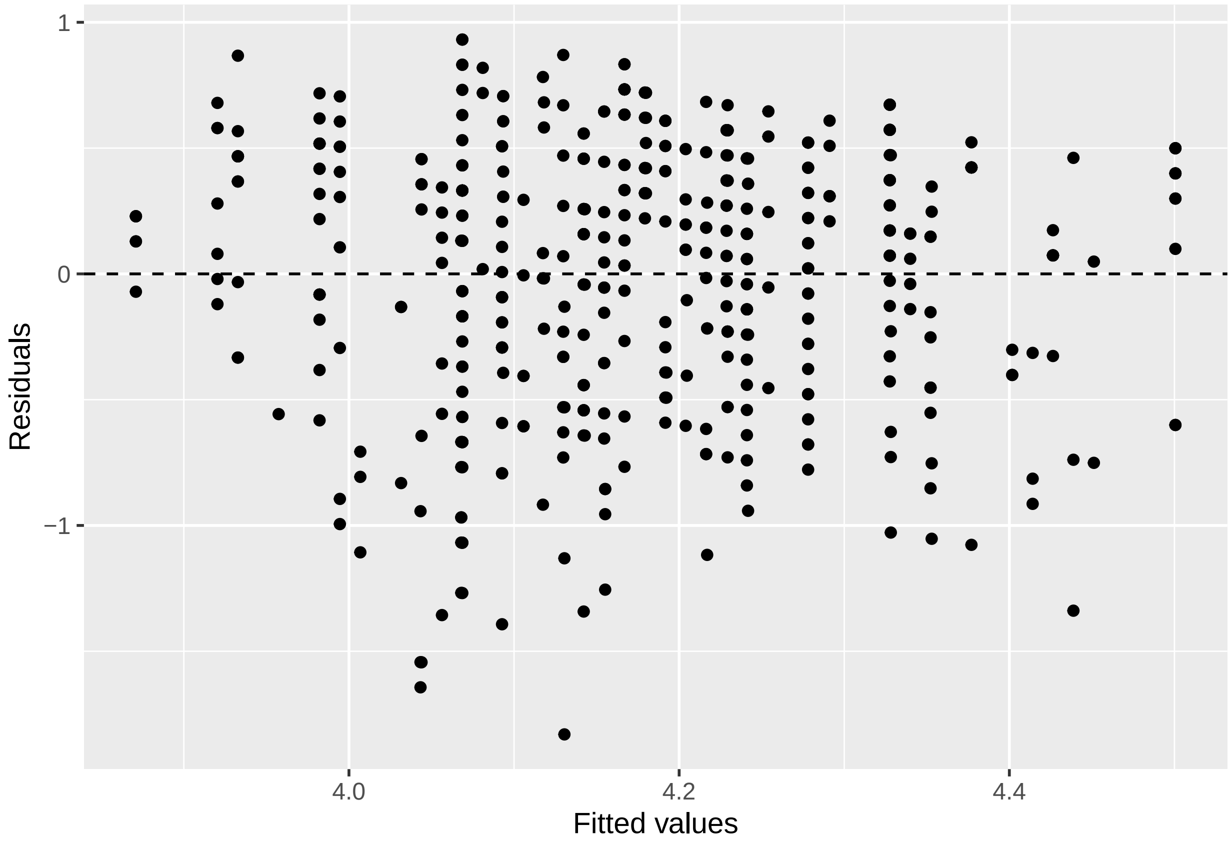
```
##
## Call:
## lm(formula = score ~ bty_avg + gender, data = evals)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.8305 -0.3625  0.1055  0.4213  0.9314
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.74734    0.08466  44.266  < 2e-16 ***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

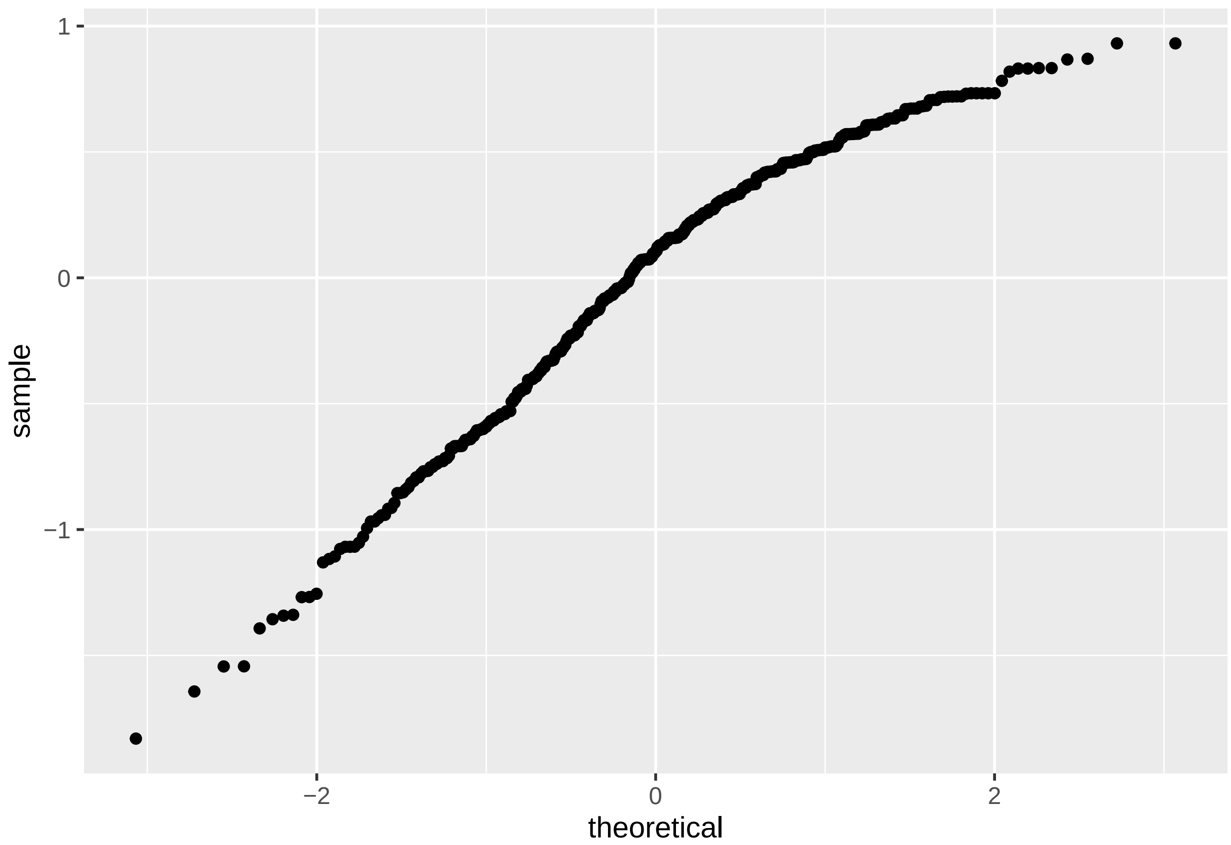
Linearity evaluation: Even spread and no pattern of values in the plot, with constant (flat) variability. Relationship appears linear.

```
ggplot(data = m_bty_gen, aes(x = .fitted, y = .resid)) +
  geom_point() +
  geom_hline(yintercept = 0, linetype = "dashed") +
  xlab("Fitted values") +
  ylab("Residuals")
```



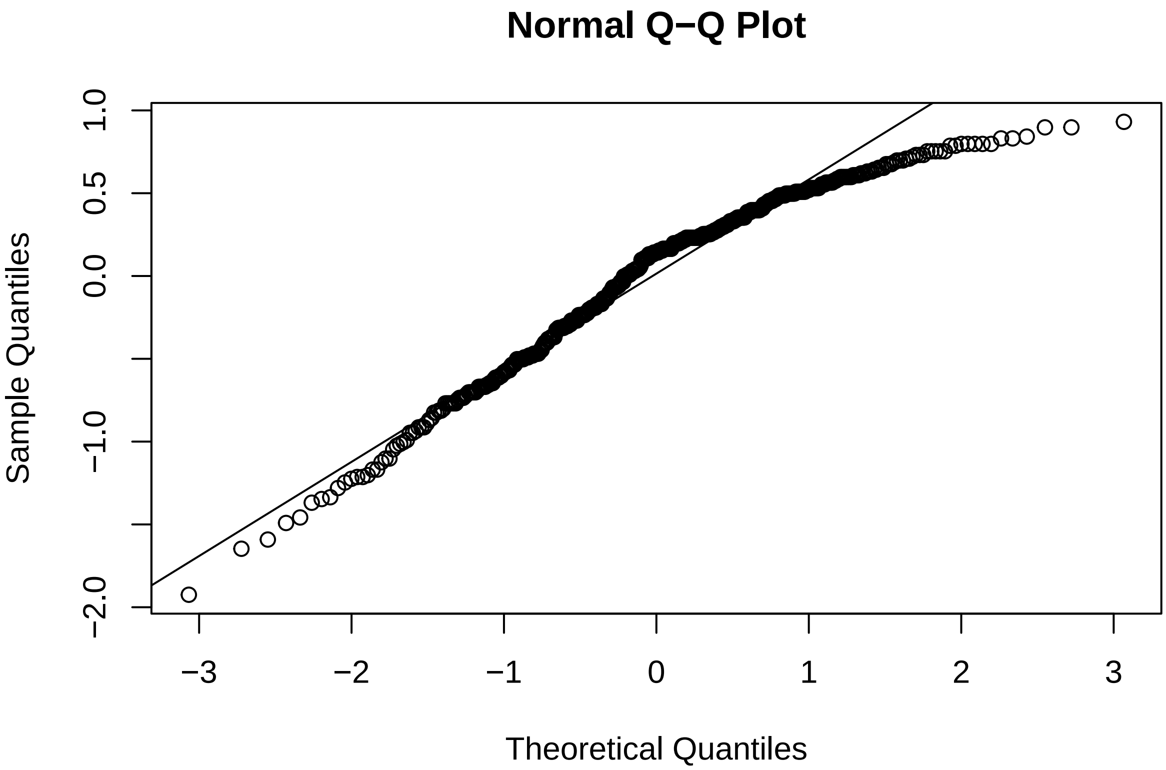
Nearly normal residuals evaluation: As with m_bty model, there is a skew at higher values of theoretical and sample, but the relationship is still linear.

```
ggplot(data = m_bty_gen, aes(sample = .resid)) +
  stat_qq()
```



Constant variability evaluation: Variability of points around the least squares line is roughly constant. Therefore the constant variability condition has been met.

```
qqnorm(m_bty$residuals)
qqline(m_bty$residuals)
```



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`?

- ANSWER: Re-running m__bty__gen model summary below. bty__avg remains a statistically significant predictor of `score` with a p value of 0.00000648.
- ANSWER: Before adding gender to the model, the equation for score ~ beauty model was y hat m__bty = 3.88034 (y intercept) + 0.06664(bty__avg). After adding gender to the model, the equation for the score ~ beauty + gender is y hat m__bty__gen = 3.74734 + 0.07416(bty__avg) + 0.17239(gendermale).
- ANSWER: Interpretation of slope: Assuming a base of 3.74734 score at 0 bty__avg, for every increase of 1 in bty__avg, score increases by 0.07416 and for every increase of 1 in gendermale (which is a binary variable), score increases by 0.17239. It's not only more attractive professors who get higher average course evaluation scores but also male professors.

```
summary(m_bty_gen)
```

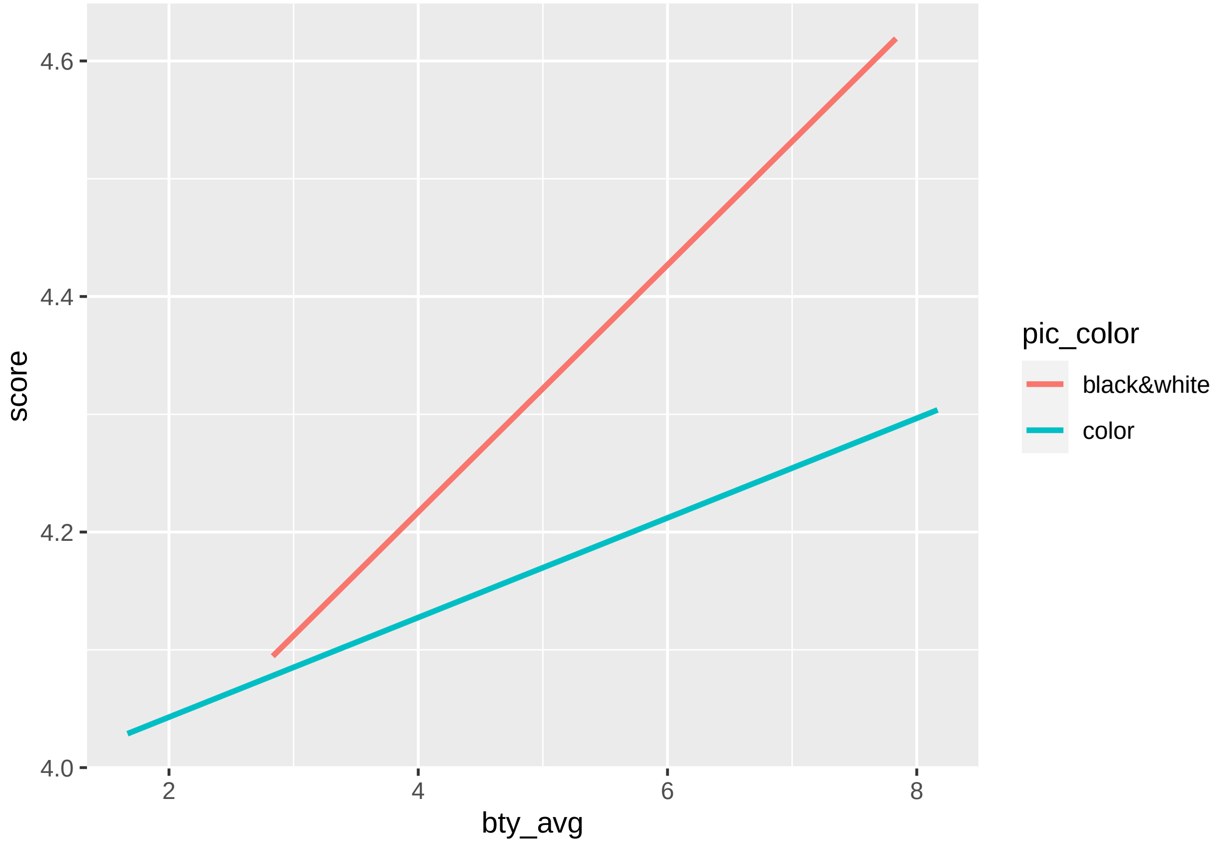
```
##
## Call:
## lm(formula = score ~ bty_avg + gender, data = evals)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.8305 -0.3625  0.1055  0.4213  0.9314
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.74734    0.08466  44.266  < 2e-16 ***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

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.

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

```
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?

- ANSWER: The equation of the line with color pictures is the line of best fit for score ~ bty_avg for the two different pic colors. In numeric terms it is y hat = 4.06318 (y intercept) + 0.05548(bty_avg) + -0.16059(pic_colorcolor) where if a given professor's photo is in color the pic_colorcolor value is 1. -ANSWER: For two professors who receive the same beauty rating, black and white photos tend to have a higher course evaluation score.

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

ANSWER: - If there are categorical variables with more than 2 levels, R will have one act as the base value (e.g. teaching = 0), with "ranktenure track" and "ranktenured" broken out as their own variables which can be 0 or 1.

```
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
```

```
table(evals$rank)
```

```
##
##     teaching tenure track      tenured
##          102          108          253
```

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.

- ANSWER: The variable with the highest p-value which I'd therefore not expect to have much association with the professor's score is cls_students because students would likely account for the professor's teaching quality within the limitations of class size – something out of their control.

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.

- ANSWER: I wasn't correct as cls_students had the third highest p-value of 0.22896, behind cls_profssingle and cls_levelupper which had 0.77806 and 0.29369 respectively.

13. Interpret the coefficient associated with the ethnicity variable.

- ANSWER: For every increase in 1 of the value of ethnicitynot minority (which is a binary variable), the average professor evaluation score `score` increased by 0.1234929. However this was not a statistically significant finding, therefore it should be dropped from the model.

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: Collinearity will inflate the variance, standard error, and p-value of coefficient estimates. When I droped cls_profs because it had the highest p-value, it did slightly decrease the p-value of most other predictors. therefore cls_profs had some collinearity with other predictors. When I dropped cls_profs, in addition to the signifiance levels changing, the coefficients also changed slightly.

```
m_no_cls_profs <- 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_no_cls_profs)
```

```
##
## 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
```

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.

- ANSWER: This model has the highest R-squared while also having statistical significance for all the predictors.

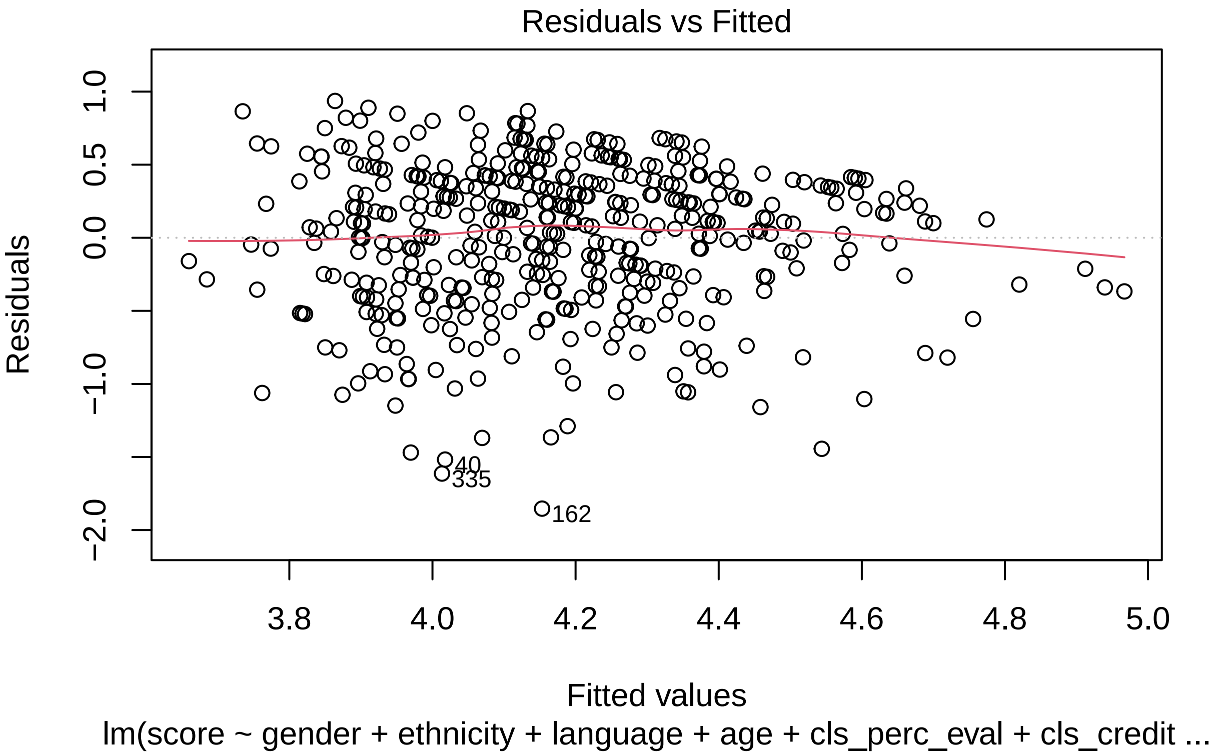
```
m_best <- lm(score ~ gender + ethnicity + language + age + cls_perc_eval +
              cls_credits + bty_avg + pic_color, data = evals)
summary(m_best)
```

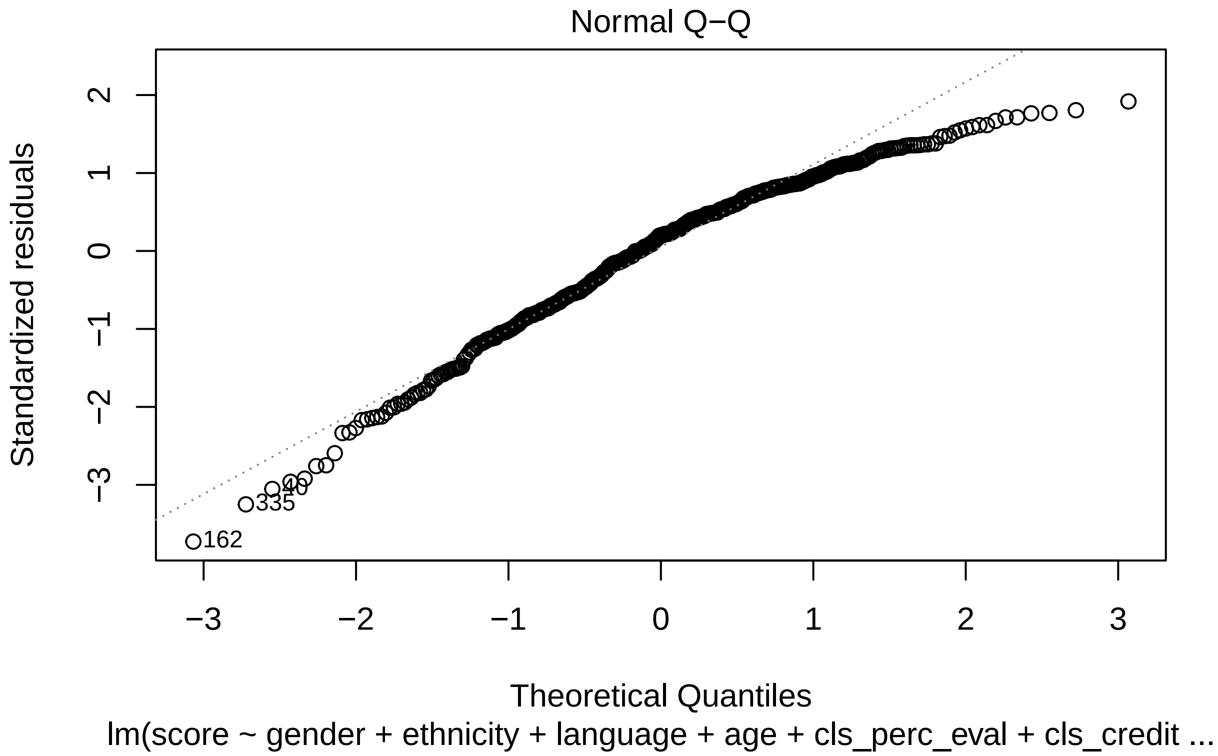
```
##
## Call:
## lm(formula = score ~ gender + ethnicity + language + age + cls_perc_eval +
##     cls_credits + bty_avg + pic_color, data = evals)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.85320 -0.32394  0.09984  0.37930  0.93610
##
## Coefficients:
##               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

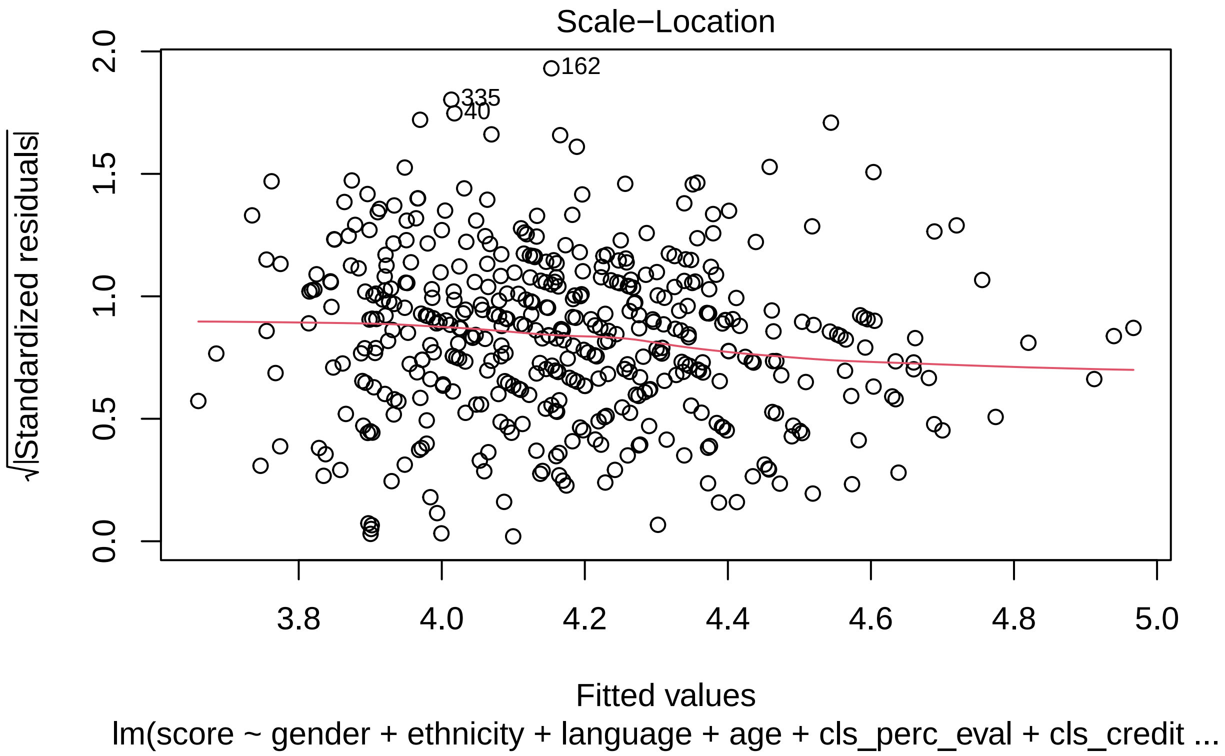
16. Verify that the conditions for this model are reasonable using diagnostic plots.

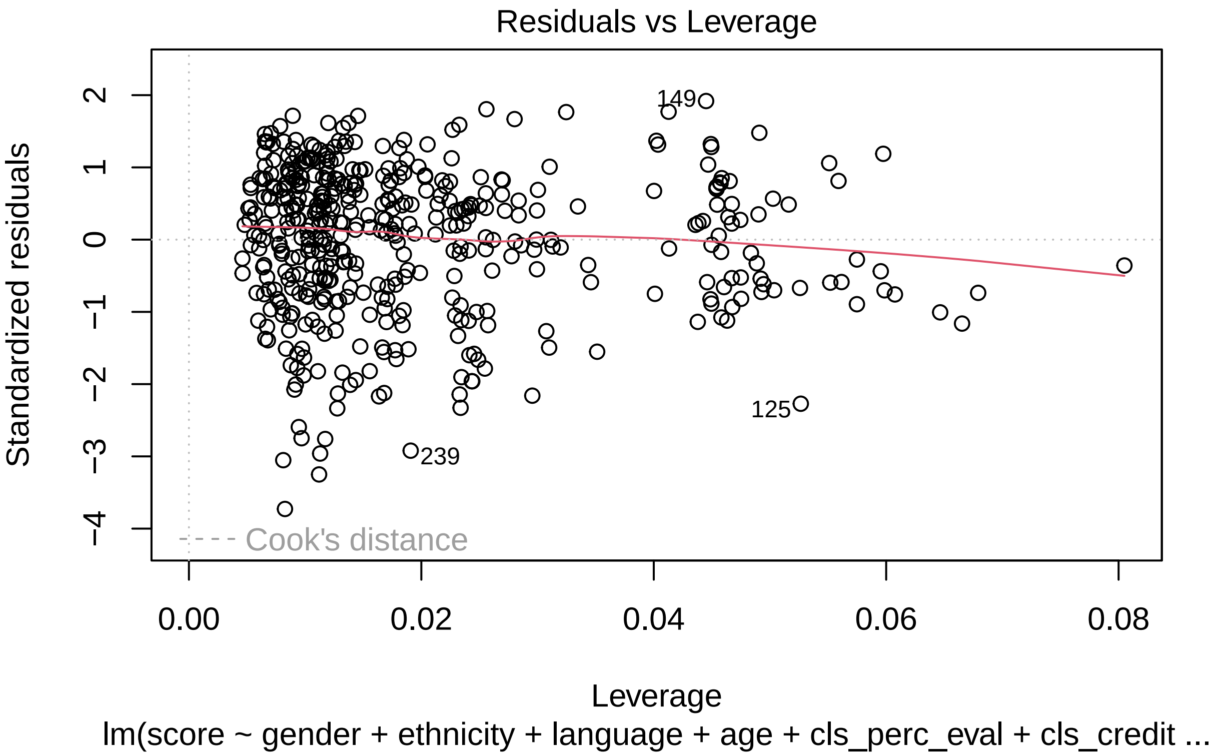
- ANSWER: (Graphed below via plot() function) Residuals vs. Fitted plot shows linearity via even spread and constant variability. Nearly normal residuals plot has skew at lowest and highest values but is generally linear; condition is met.

```
plot(m_best)
```


Residuals vs Fitted
Residuals
1.0
0.5
0.0
−1.0
−2.0
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 ...
40
335
162


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


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 ~ gender + ethnicity + language + age + cls_perc_eval + cls_credit ...



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, this is an issue because professors likely receive similar ratings across their various courses. If certain professors teach more courses than others, their data are over represented in the model. The below df prof shows that 7 professors taught 10 or more courses, while 7 professors only taught 1 course. In addition, if certain students have the same professor for multiple courses, their data could be especially over represented.

```
library(dplyr)
prof <- table(evals$prof_id)
prof %>% as.data.frame() %>% arrange(desc(Freq))
```

```
##    Var1 Freq
## 1    34   13
## 2    49   13
## 3    82   11
## 4    10   10
## 5    20   10
## 6    58   10
## 7    71   10
## 8    19    9
## 9    70    9
## 10    4    8
## 11   18    8
## 12   23    8
## 13   37    8
## 14   38    8
## 15   85    8
## 16    6    7
## 17    8    7
## 18    9    7
## 19   13    7
## 20   24    7
## 21   27    7
## 22   31    7
## 23   48    7
## 24   52    7
## 25   65    7
## 26   88    7
## 27   92    7
## 28    5    6
## 29   16    6
## 30   21    6
## 31   50    6
## 32   53    6
## 33   66    6
## 34   72    6
## 35   77    6
## 36   93    6
## 37    7    5
## 38   12    5
## 39   17    5
## 40   26    5
## 41   33    5
## 42   40    5
## 43   73    5
## 44   83    5
## 45   84    5
## 46    1    4
## 47   14    4
## 48   15    4
## 49   28    4
## 50   29    4
## 51   36    4
## 52   41    4
## 53   42    4
## 54   44    4
## 55   51    4
## 56   74    4
## 57   78    4
## 58   80    4
## 59   81    4
## 60   94    4
## 61    2    3
## 62   11    3
## 63   32    3
## 64   35    3
## 65   43    3
## 66   45    3
## 67   47    3
## 68   54    3
## 69   55    3
## 70   60    3
## 71   64    3
## 72   68    3
## 73   79    3
## 74   86    3
## 75   89    3
## 76   91    3
## 77    3    2
## 78   25    2
## 79   56    2
## 80   57    2
## 81   59    2
## 82   63    2
## 83   67    2
## 84   75    2
## 85   76    2
## 86   87    2
## 87   90    2
## 88   22    1
## 89   30    1
## 90   39    1
## 91   46    1
## 92   61    1
## 93   62    1
## 94   69    1
```

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: The characteristics most associated with receiving a high evaluation score are: Being male gender; Not being an ethnic minority; Having received an education at a predominantly English-language institution; Being young; Having a high beauty score; Having a high percentage of your class' students complete an evaluation; Having your class be one credit.

```
summary(m_best)
```

```
##
## 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
```

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

- ANSWER: I would need more data to understand whether 1) The University of Texas is similar to other universities (racially, socioeconomically, gender ratio, etc.) and 2) Whether this study was taken from a random sample of UT students. If these two conditions were met, then I'd generally expect results to be similar at other universities. As a demonstration of how this survey could be limited if condition 1 was not met: A non-ethnically diverse school of majority men might on average have different world views than a different sort of school, leading to different course evaluations. While identity politics have limitations, demographic characteristics can be used to predict such things as voting patterns, therefore they could have statistically significant impacts on professor evaluations too. Regarding condition 2, random samples are essential to representing the entire population. Otherwise, this sample might not even represent the UT, let alone be generalizable to other universities.