Multiple linear regression

Kenan Sooklall

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

Exercise 1

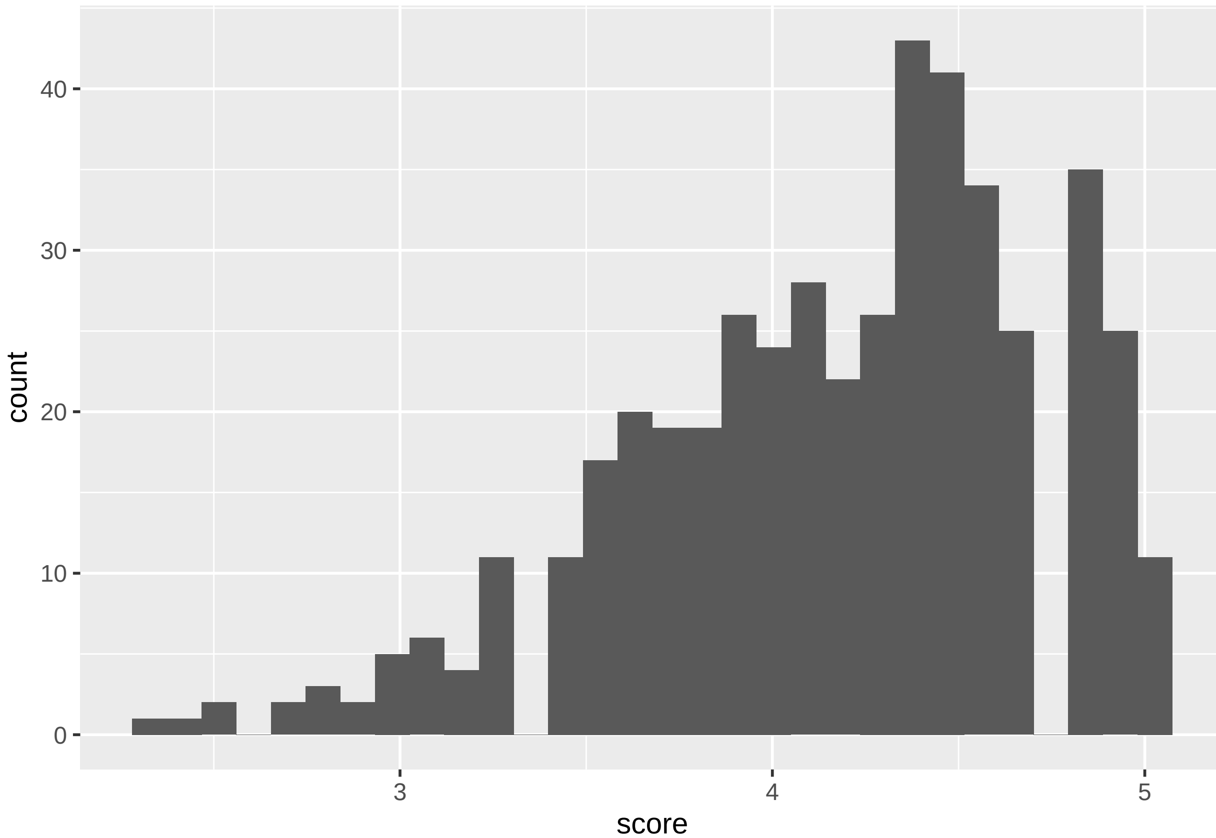
1. Is this an observational study or an experiment? The original research question posed in the paper is whether beauty leads directly to the differences in course evaluations. Given the study design, is it possible to answer this question as it is phrased? If not, rephrase the question.

This is an observational study since we aren't conducting any experiment.

Exercise 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?

```
evals %>% ggplot(aes(x=score)) + geom_histogram()
```

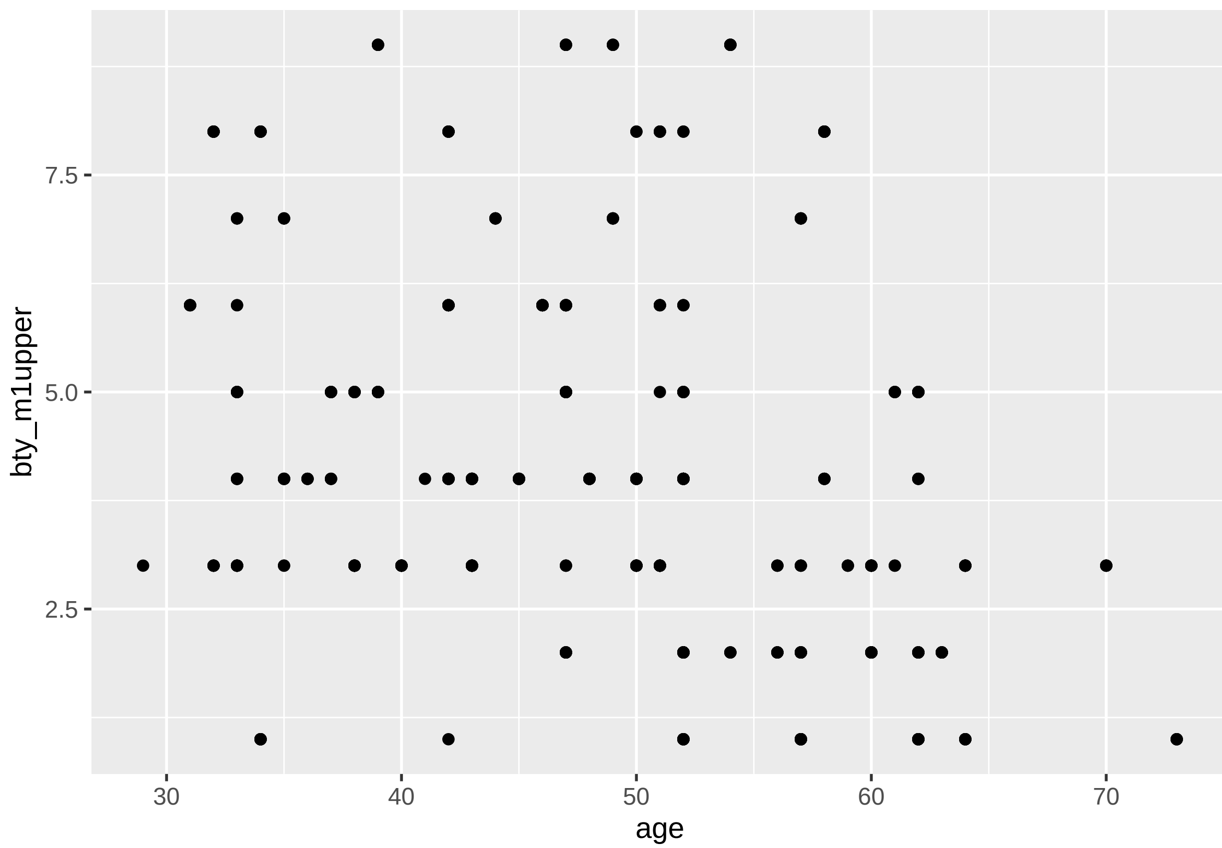


Scores are left skewed, meaning more student gave higher scores than lower ones. I would expect more neutral scores around 3, but I'm guess student's really like the course or there was another variable.

Exercise 3

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

```
evals %>% ggplot(aes(x=age, y=bty_m1upper)) + geom_point()
```

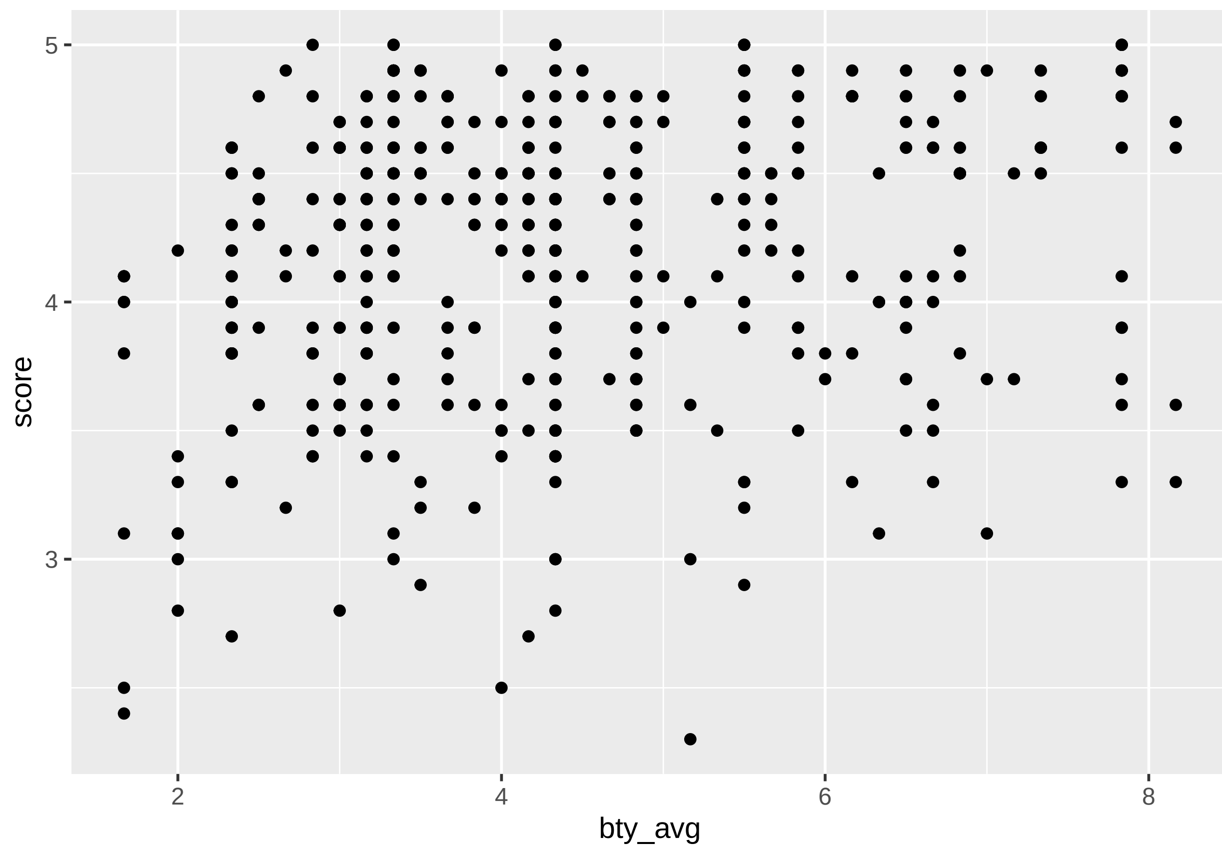


From the scatter plot there doesn't seem to be a trend on the relationship between beauty rating for upper level males and their age.

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?

```
evals %>% select(score, bty_avg) %>% filter(score == 4)
```

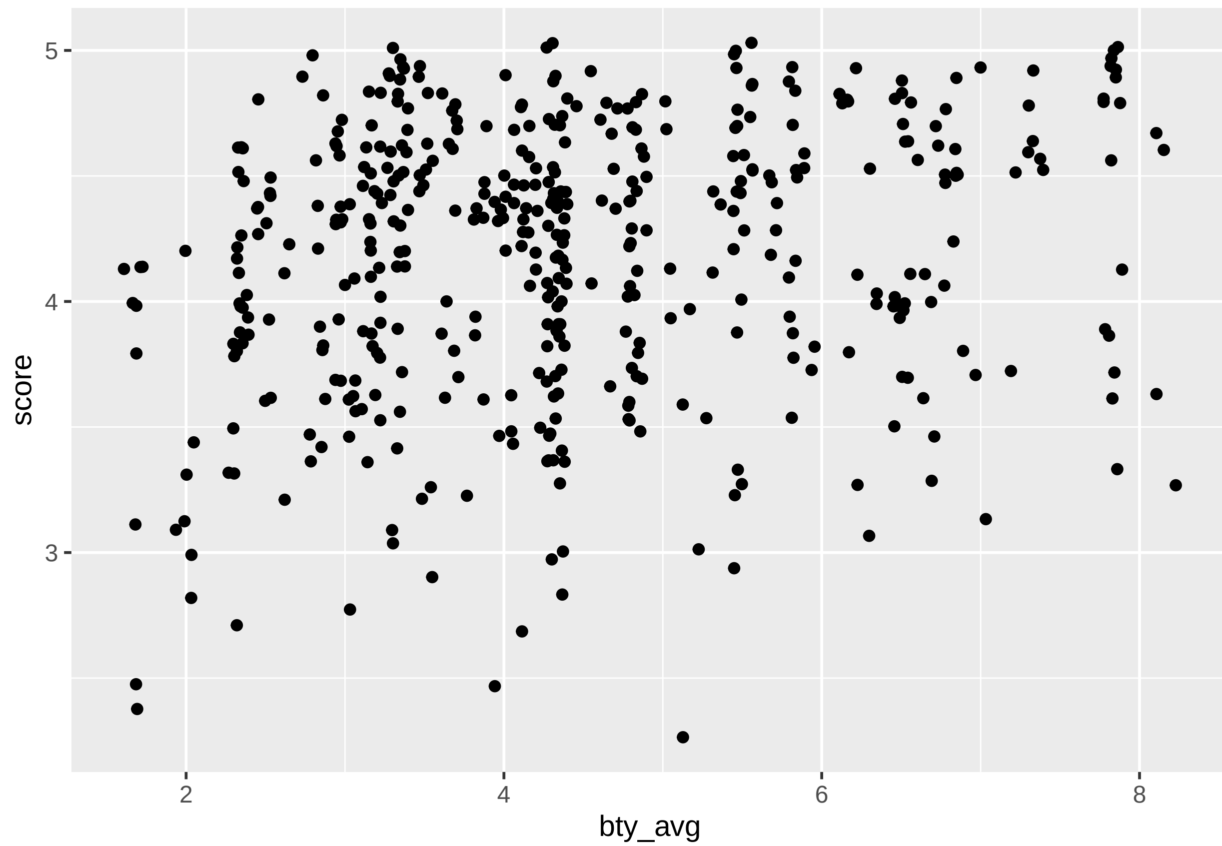
```
## # A tibble: 24 x 2
##    score bty_avg
##    <dbl>   <dbl>
##  1     4    5.5
##  2     4    4.83
##  3     4    4.83
##  4     4    4.33
##  5     4    4.33
##  6     4    4.33
##  7     4    4.33
##  8     4    6.5
##  9     4    6.5
## 10     4    2.33
## # ... with 14 more rows
```

It looks like there are a lot of duplicated data points.

Exercise 4

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

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



The initial scatterplot was more structured due to the duplicates.

Exercise 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?

```r
model = lm(bty_avg ~ score, data=evals)
summary(model)
```

```
##
## Call:
## lm(formula = bty_avg ~ score, data = evals)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.7116 -1.2116 -0.2032  0.9328  4.2089
##
## Coefficients:
```

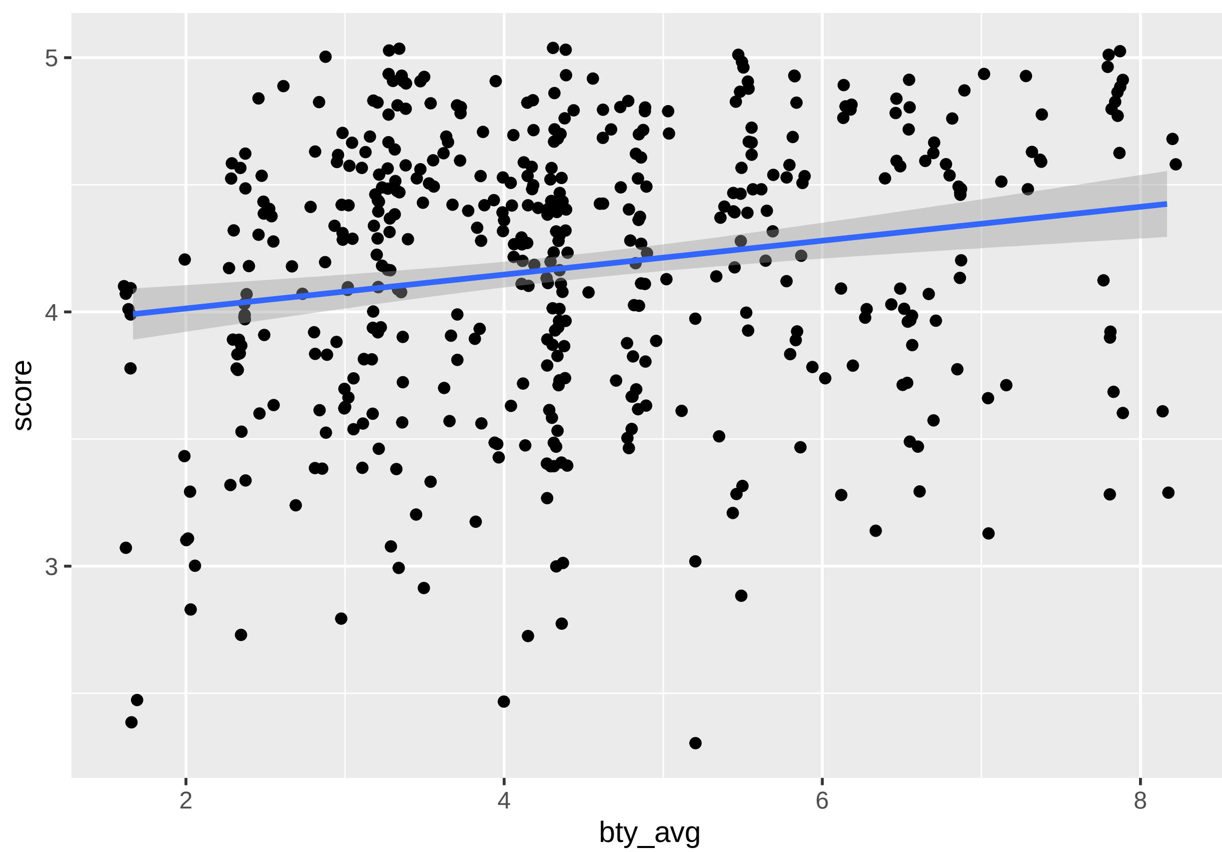
```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.2237     0.5409   4.111 4.66e-05 ***
## score         0.5256     0.1285   4.090 5.08e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.502 on 461 degrees of freedom
## Multiple R-squared:  0.03502,    Adjusted R-squared:  0.03293
## F-statistic: 16.73 on 1 and 461 DF,  p-value: 5.083e-05
```

Equation $\hat{y} = 2.2237 + 0.5256 * btyAvg$

The p-value is less than 0.05 making this model a statistically significant predictor ;however, the r^2 is only 0.033 making this model not a good predictor.

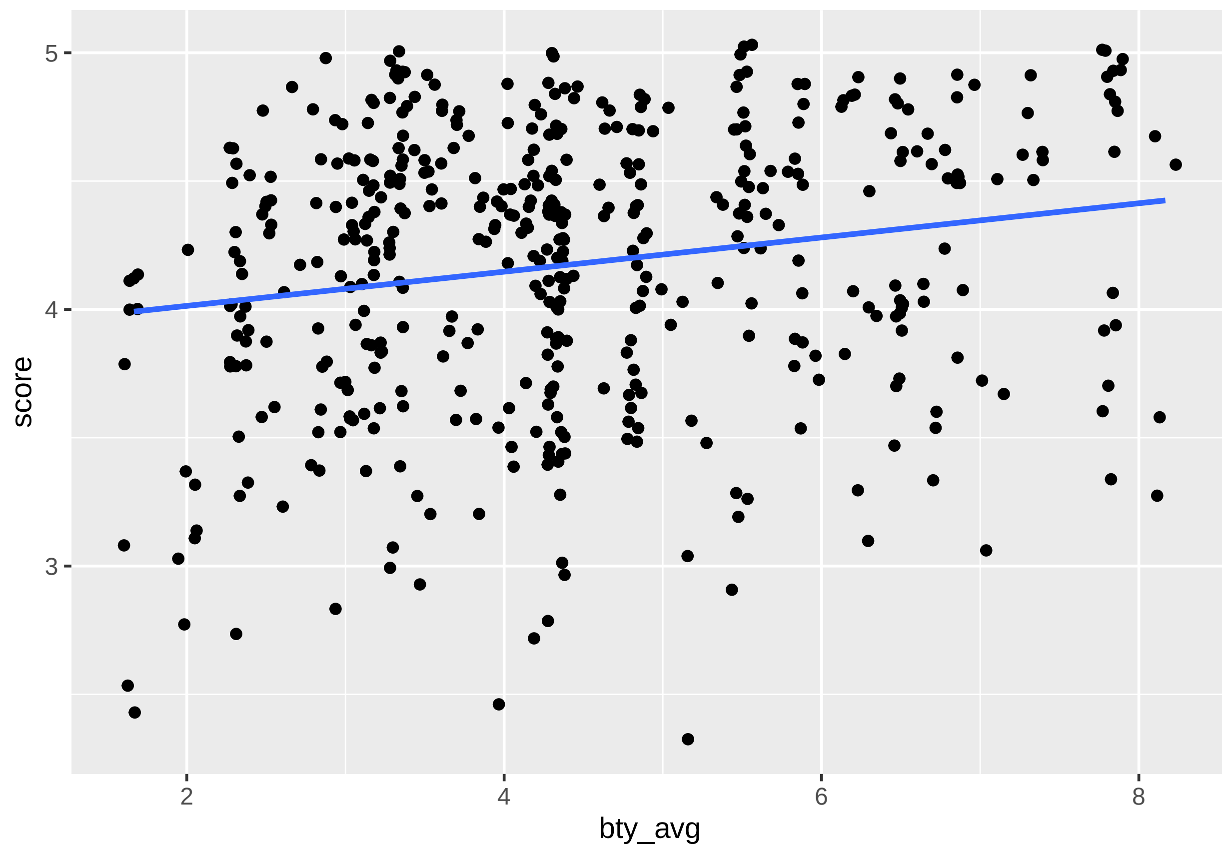
Add the line of the bet fit model to your plot using the following:

```
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`.

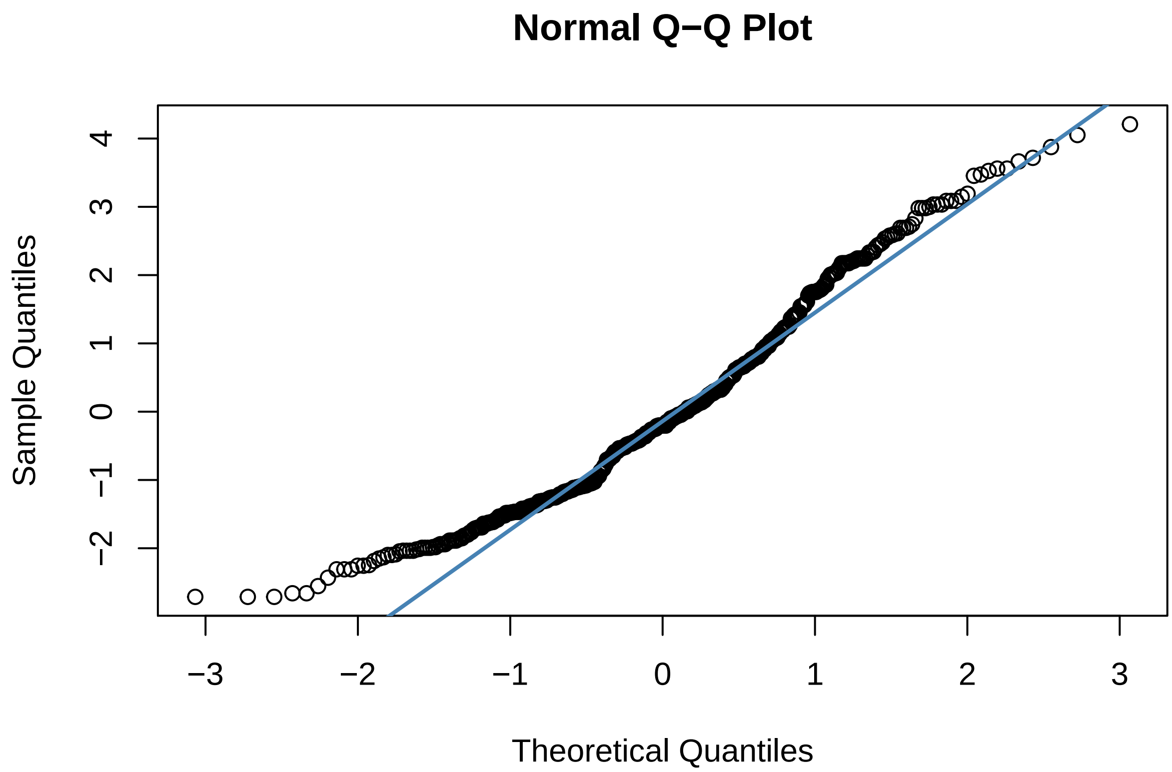
```
ggplot(data = evals, aes(x = bty_avg, y = score)) +
  geom_jitter() +
  geom_smooth(method = "lm", se = FALSE)
```



Exercise 6

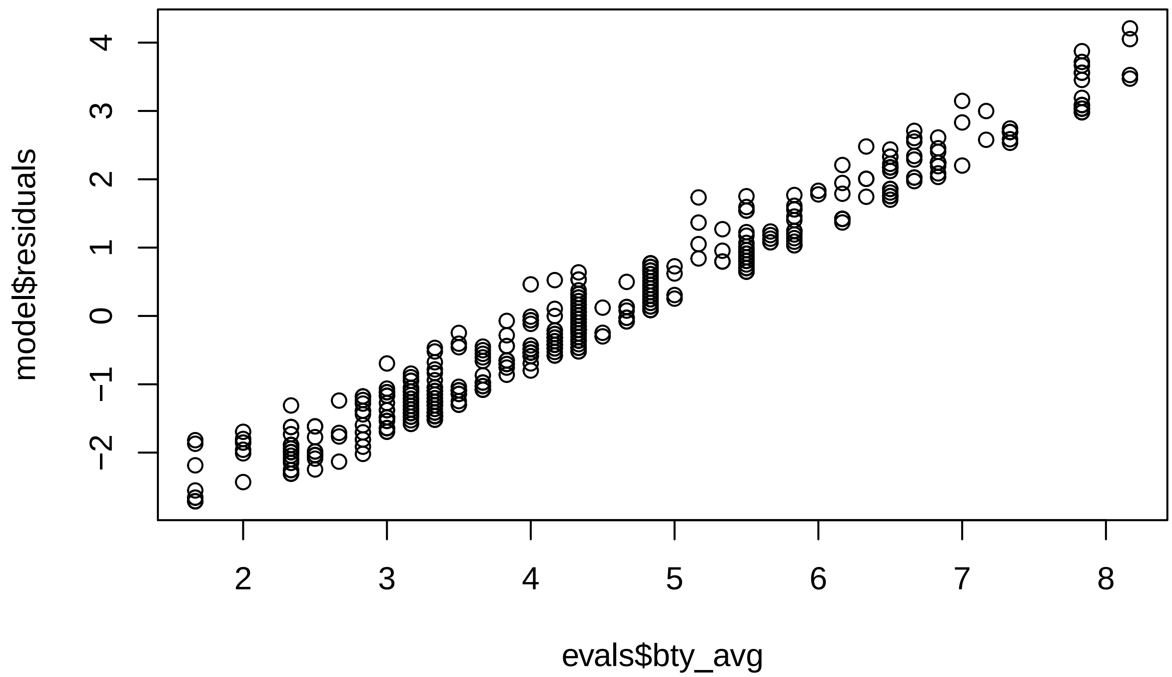
6. Use residual plots to evaluate whether the conditions of least squares regression are reasonable. Provide plots and comments for each one (see the Simple Regression Lab for a reminder of how to make these).

```
qqnorm(model$residuals)
qqline(model$residuals, col = "steelblue", lwd = 2)
```



The QQ plot shows a lot of values within the line with the smaller values away from the line

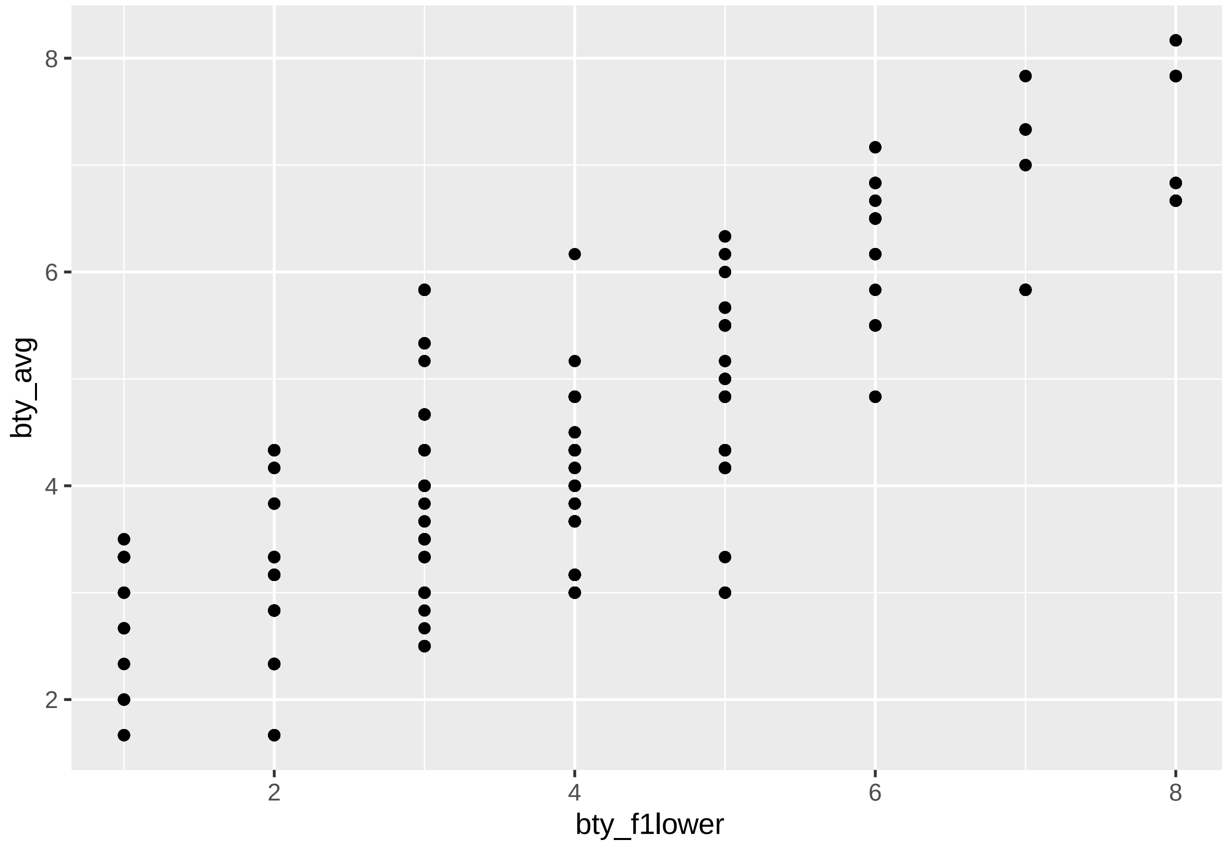
```
plot(model$residuals ~ evals$bty_avg)
```



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()
```

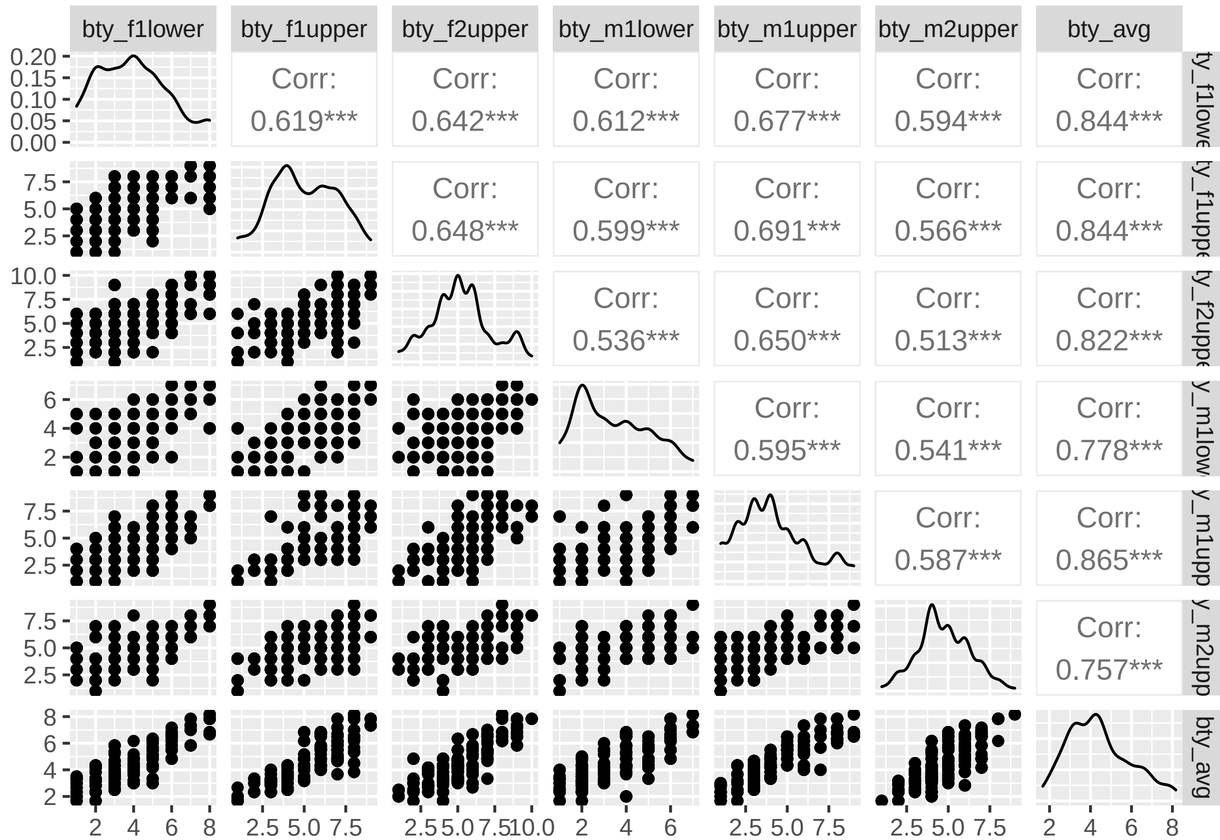



```
evals %>%
  summarise(cor(bty_avg, bty_f1lower))
```

```
## # 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)
```

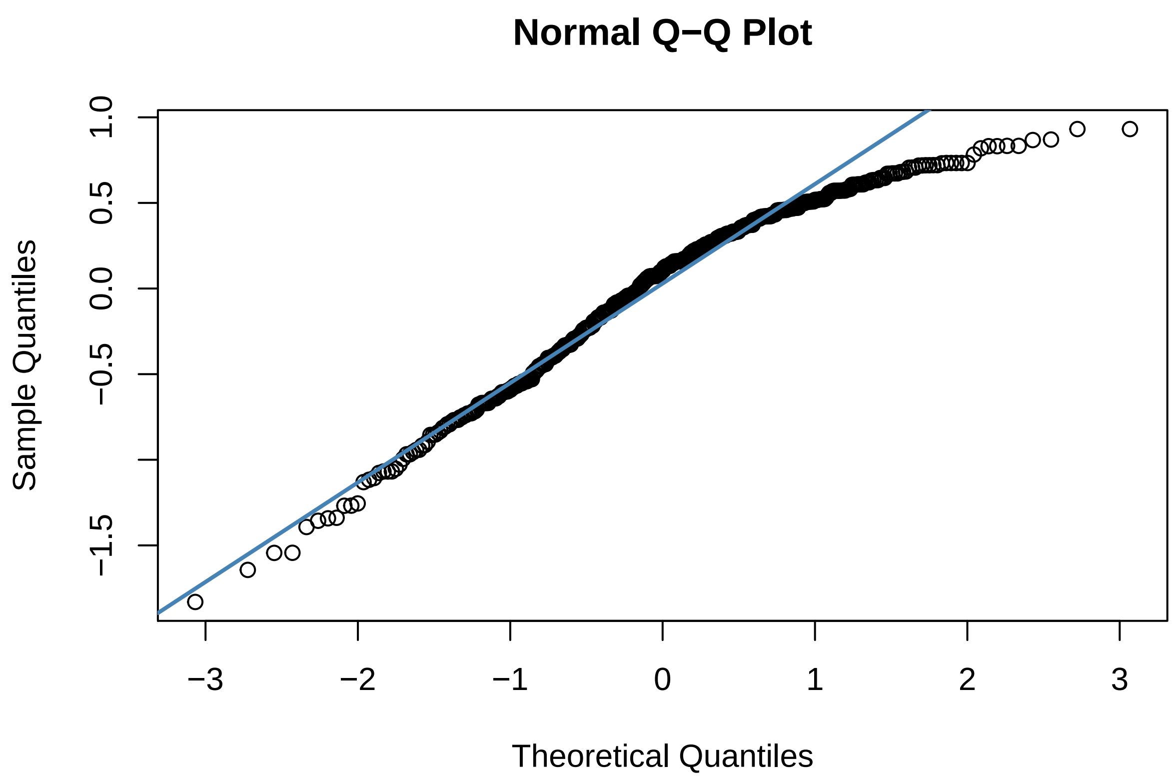
```
##
## Call:
## lm(formula = score ~ bty_avg + gender, data = evals)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.8305 -0.3625  0.1055  0.4213  0.9314
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.74734    0.08466  44.266  < 2e-16 ***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

Exercise 7

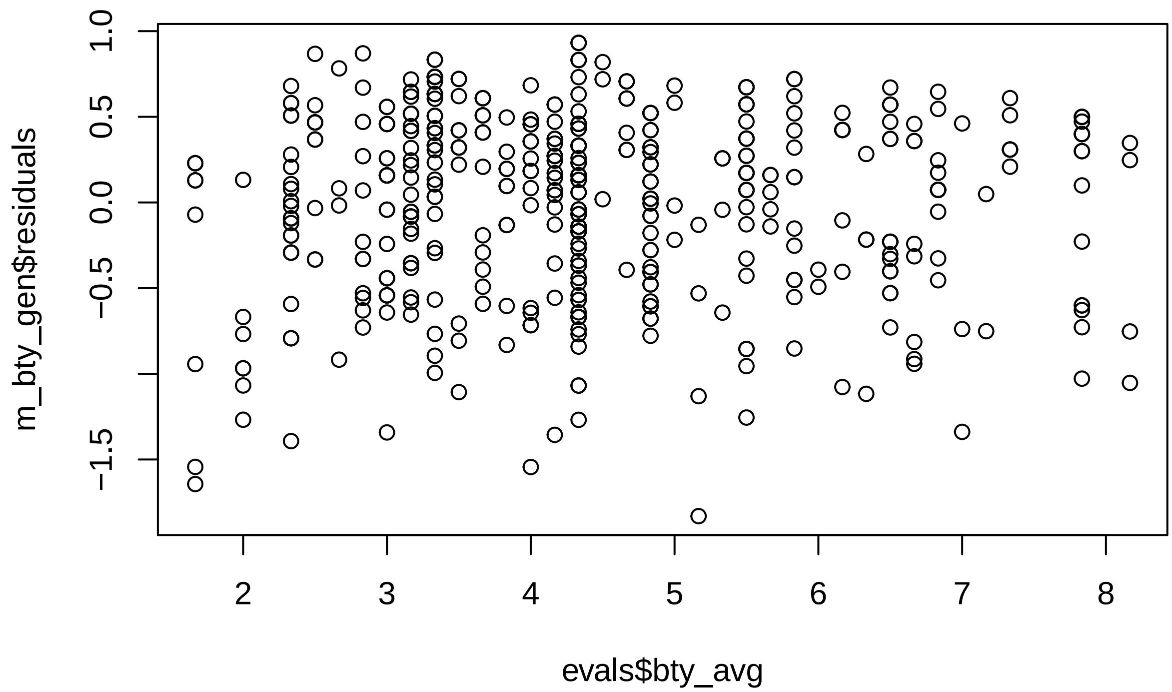
7. P-values and parameter estimates should only be trusted if the conditions for the regression are reasonable. Verify that the conditions for this model are reasonable using diagnostic plots.

```r
qqnorm(m_bty_gen$residuals)
qqline(m_bty_gen$residuals, col = "steelblue", lwd = 2)
```



The QQ plot shows a lot of values within the line with those to the right moving away

```r
plot(m_bty_gen$residuals ~ evals$bty_avg)
```



The residuals look all ovet the place

Exercise 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`?

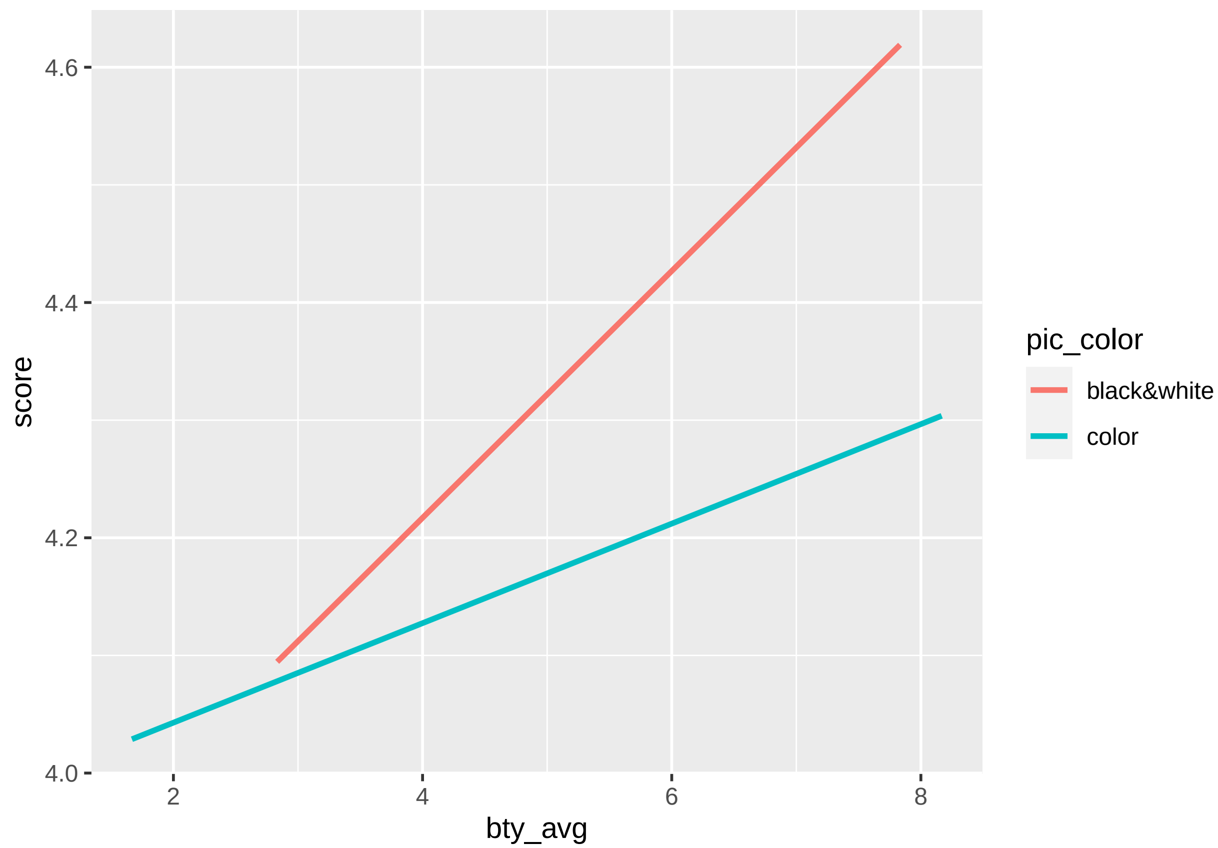
The adjusted R^2 for bty_avg along was 0.03, the addition of gender increased it to 0.06 thus making the model that accounts for gender better but not much.

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)
```



Exercise 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?

\$ y = 3.74734 + 0.17239 \$

If two professors both receive the same beauty rating, the male professor will have the higher course evaluation score because of the positive slope of ~0.17.

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.)

Exercise 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
```

It looks like R dropped the teaching level

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.

Exercise 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.

I think cls_credits and cls_profs would have high pvalues because a class that is single or multiple credits doesn't reflect a professor score. cls_profs is a strange variable because i never had a class taught with multiple professors.

Let's run the model…

```r
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
```

Exercise 12

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

```r
data.frame(summary(m_full)[4]$coefficients) %>% arrange('Pr...t..')
```

```
##                           Estimate  Std..Error    t.value     Pr...t..
## (Intercept)           4.0952140795 0.290527656 14.0957805 1.319326e-37
## ranktenure track     -0.1475932457 0.082067087 -1.7984462 7.277935e-02
## ranktenured          -0.0973377624 0.066329583 -1.4674864 1.429455e-01
## gendermale            0.2109481296 0.051822962  4.0705533 5.544372e-05
## ethnicitynot minority  0.1234929213 0.078627324  1.5706108 1.169791e-01
## languagenon-english  -0.2298111901 0.111375416 -2.0633924 3.965088e-02
## age                  -0.0090071896 0.003135911 -2.8722720 4.268765e-03
## cls_perc_eval         0.0053272412 0.001539323  3.4607683 5.902546e-04
## cls_students          0.0004546339 0.000377388  1.2046856 2.289607e-01
## cls_levelupper        0.0605139602 0.057561665  1.0512893 2.936925e-01
## cls_profssingle      -0.0146619208 0.051988497 -0.2820224 7.780566e-01
## cls_creditsone credit  0.5020431770 0.115938766  4.3302443 1.839347e-05
## bty_avg               0.0400333017 0.017506416  2.2867788 2.267440e-02
## pic_outfitnot formal -0.1126816871 0.073880036 -1.5251980 1.279153e-01
```

```
## pic_colorcolor        -0.2172629964 0.071502140 -3.0385523 2.516206e-03
```

It looks like credits, and cls_prof are among the highest pvalues.

Exercise 13

13. Interpret the coefficient associated with the ethnicity variable.

A unit increase in Ethnicity corresponds to 0.1234929 increase in score.

Exercise 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?

```
vals <- data.frame(summary(m_full)[4]$coefficients) %>% mutate(pval = Pr...t..) %>% filter(pval < 0.05)

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

```
##
## Call:
## lm(formula = score ~ cls_credits + gender + cls_perc_eval + age +
##     bty_avg + language, data = evals)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.87317 -0.31091  0.08655  0.38395  1.05632
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)           3.636831   0.192370  18.905  < 2e-16 ***
## cls_creditsone credit 0.440166   0.101955   4.317 1.94e-05 ***
## gendermale            0.198985   0.049972   3.982 7.95e-05 ***
## cls_perc_eval         0.004716   0.001441   3.272 0.001149 **
## age                  -0.004617   0.002621  -1.762 0.078786 .
## bty_avg               0.064377   0.016442   3.915 0.000104 ***
## languagenon-english  -0.249029   0.098839  -2.520 0.012091 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5066 on 456 degrees of freedom
## Multiple R-squared:  0.1437, Adjusted R-squared:  0.1324
## F-statistic: 12.75 on 6 and 456 DF,  p-value: 2.524e-13
```

Every variable estimated slope changed and the intercept fell.

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.

```
lm_final <- lm(score ~ cls_credits + gender + cls_perc_eval + age, data = evals)
summary(lm_final)
```

```
##
## Call:
## lm(formula = score ~ cls_credits + gender + cls_perc_eval + age,
##     data = evals)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.84314 -0.30804  0.05629  0.36964  1.01798
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)            3.999749   0.170911  23.403  < 2e-16 ***
## cls_creditsone credit  0.404899   0.103482   3.913 0.000105 ***
## gendermale             0.191409   0.050987   3.754 0.000196 ***
## cls_perc_eval          0.005385   0.001458   3.695 0.000247 ***
## age                   -0.007448   0.002572  -2.896 0.003962 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5173 on 458 degrees of freedom
## Multiple R-squared:  0.1033, Adjusted R-squared:  0.09543
## F-statistic: 13.19 on 4 and 458 DF,  p-value: 3.565e-10
```

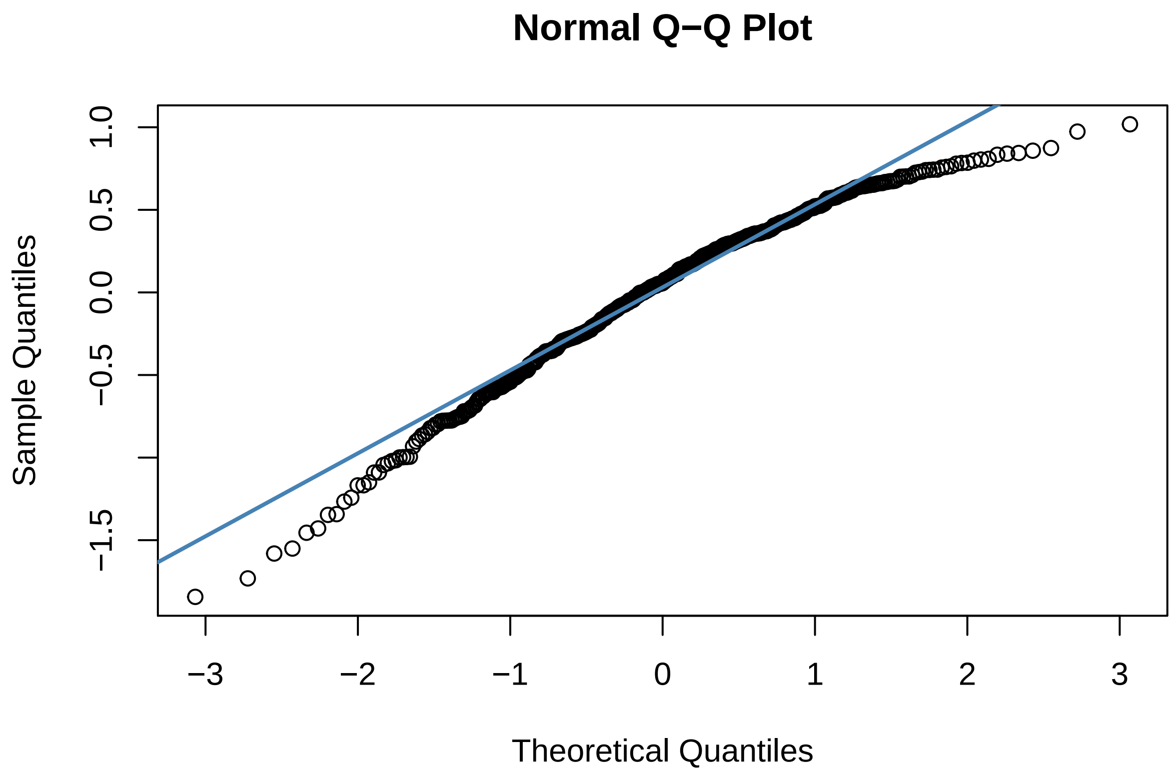
$ score = 3.99 + 0.4 * clsCredits + 0.19 * gender + 0.005 * clsPercEval - 0.007 * age $

Exercise 16

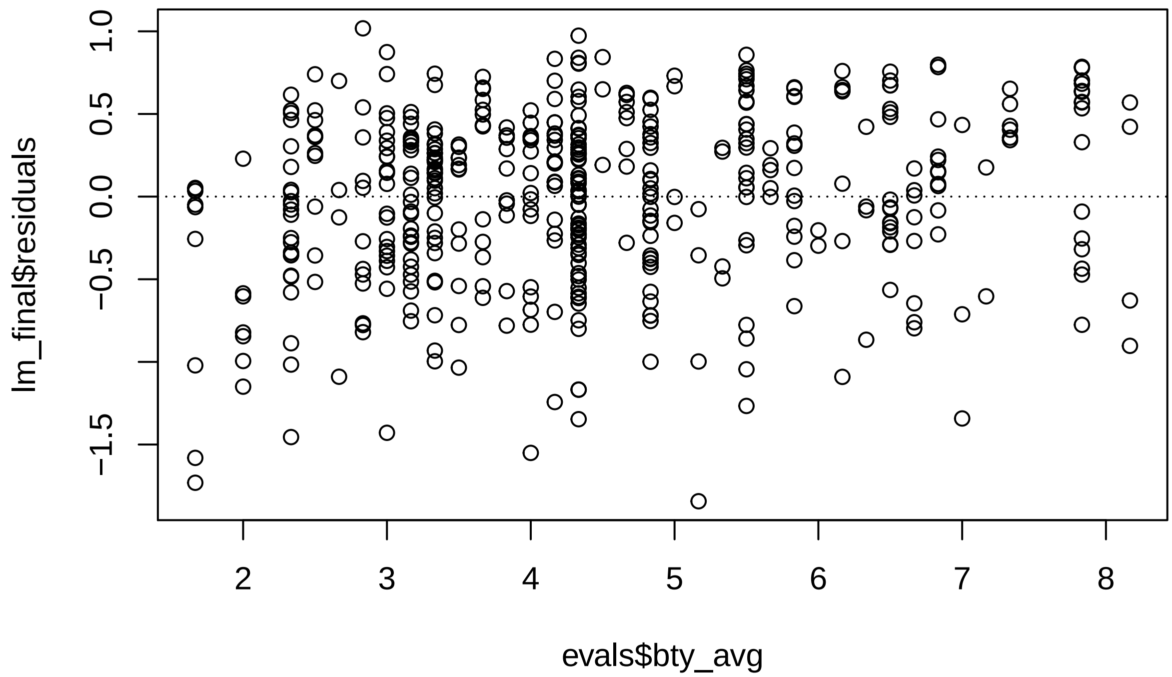
16. Verify that the conditions for this model are reasonable using diagnostic plots.

Using the same analysis done in Exercise 7 the plots look fine

```
qqnorm(lm_final$residuals)
qqline(lm_final$residuals, col = "steelblue", lwd = 2)
```




```r
plot(lm_final$residuals ~ evals$bty_avg)
abline(h = 0, lty = 3)
```



Exercise 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?

If each row is a course there will be a lot of overlap in student reviews thus causing some skewness.

Exercise 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.

Based on my final model the desired characteristics of a professor would be: male, young, high evaluations and high amount of credit.

Exercise 19

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

I think we can only generalize these conclusions for universities that carry the same values as University of Texas at Austin.