Multiple linear regression

Gabriel Campos

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(DATA606)
library(GGally)
data(evals)
```

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)
```

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.

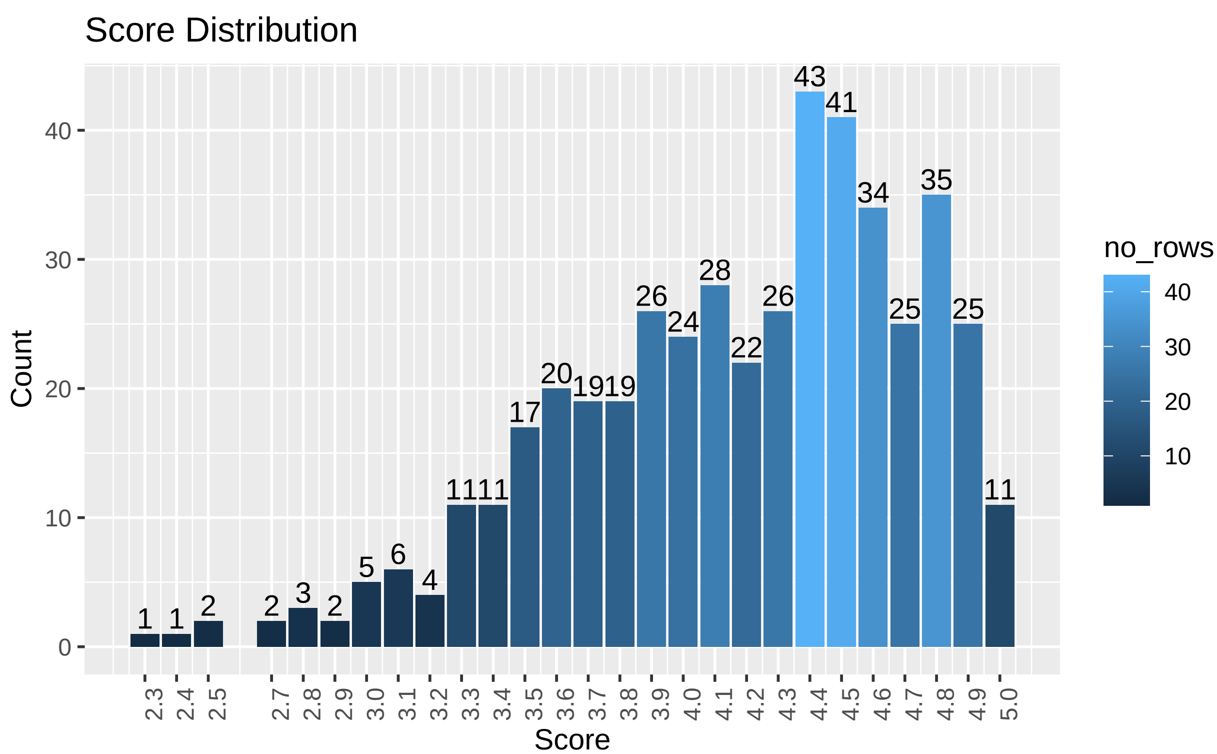
```
Is this an observational study or an experiment? The original research
question posed in the paper is whether beauty leads directly to the
differences in course evaluations. Given the study design, is it possible to
answer this question as it is phrased? If not, rephrase the question.
```

ANSWER:
This is an observational studies (no control groups, experimentation. No it is not phrased correctly, *'Does beauty lead to* `higher` *(or* `lower`*) course evaluations for professors.'*
are more appropriate potential questions.

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:
The data is left skewed, but for the most part not at the endpoints. This is expected, simply because scores below 3.3 and at or above 4.7 can be viewed as critical of the professor both positively and negatively, and hence would be less likely scores a student would pick in general. This also may imply higher scoring from students on average.

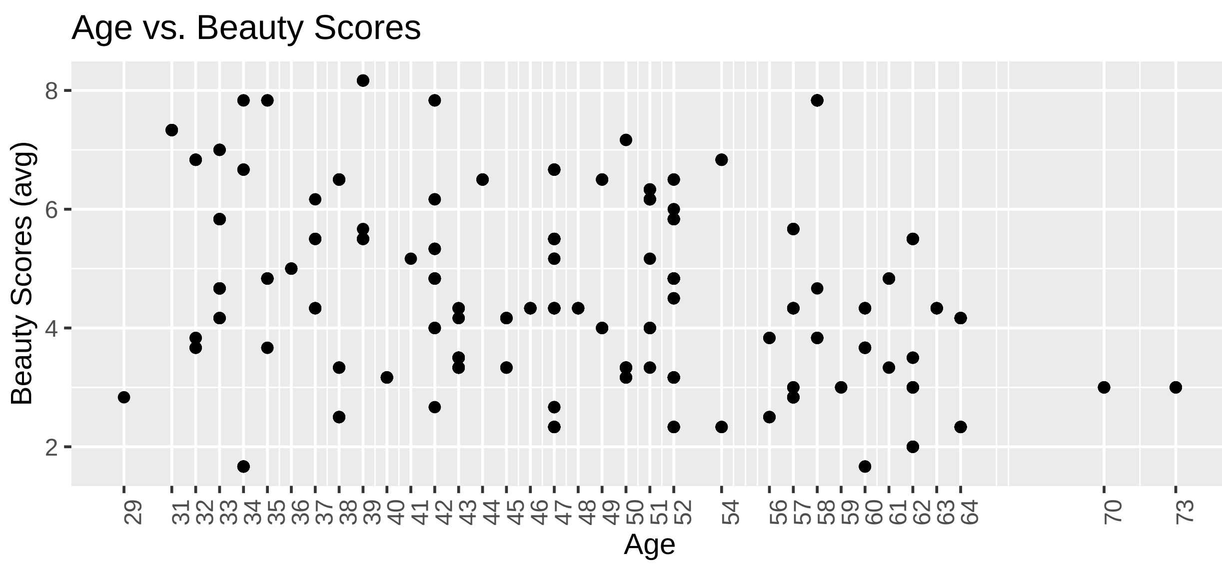


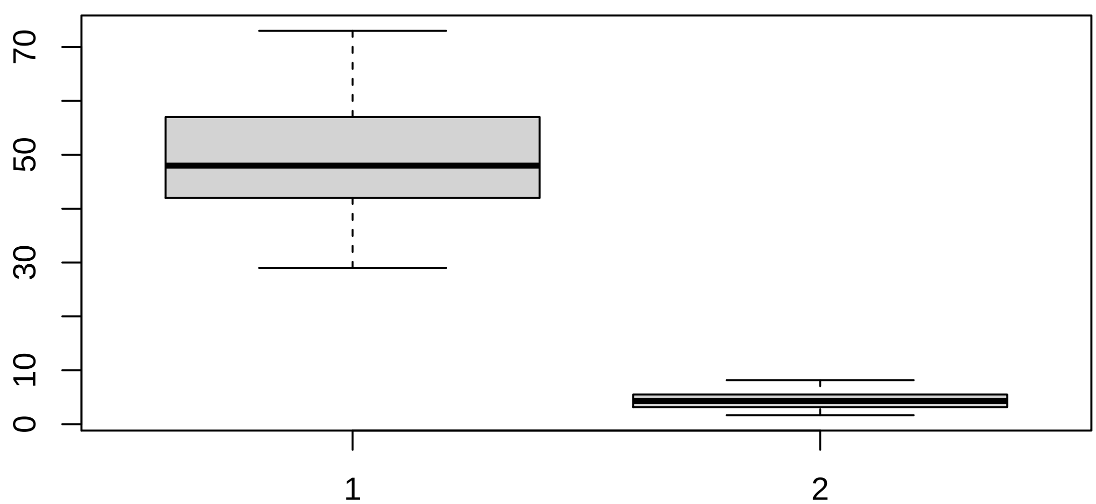
3.

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

ANSWER:
I chose Age and `bty_avg` and `age` which is defined by Open Intro ‘eval’ data set as:

- age = Age of professor
- bty_avg = Average beauty rating of professor.

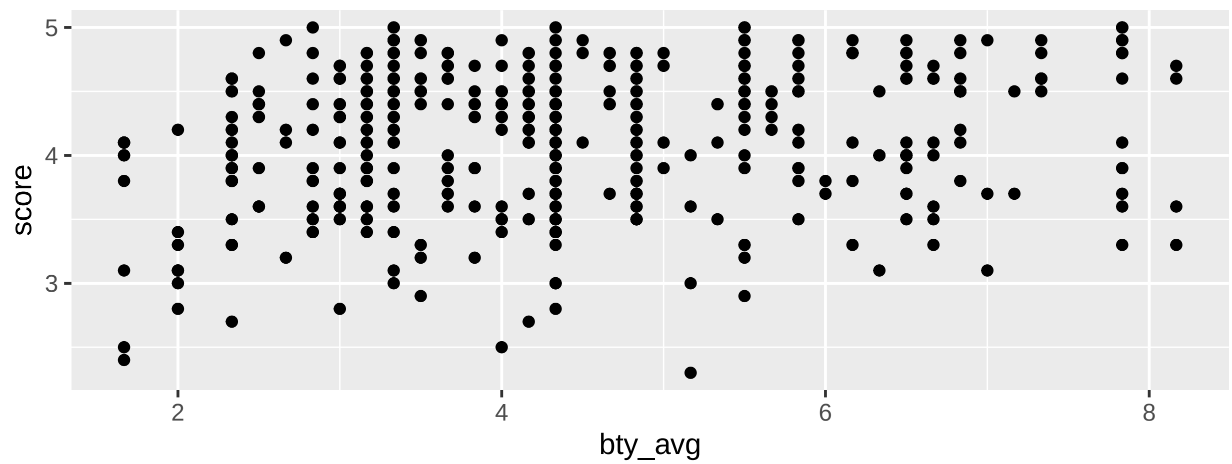




Simple linear regression

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:

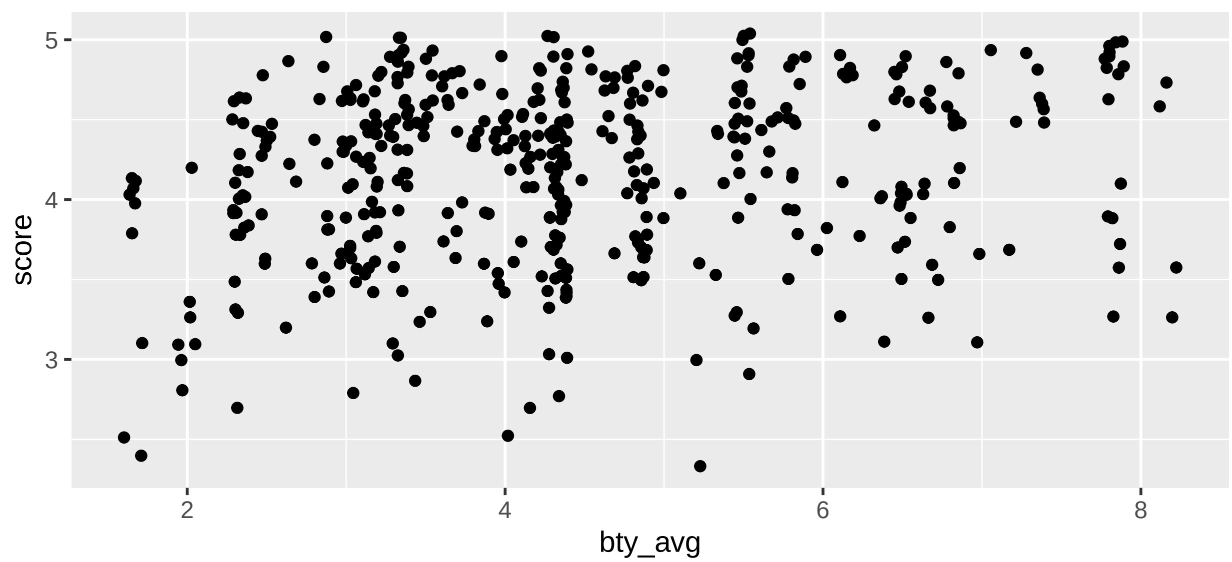
```
ggplot(data = evals, aes(x = bty_avg, y = score)) +
  geom_point()
```



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?**I think we're missing data**

4.

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

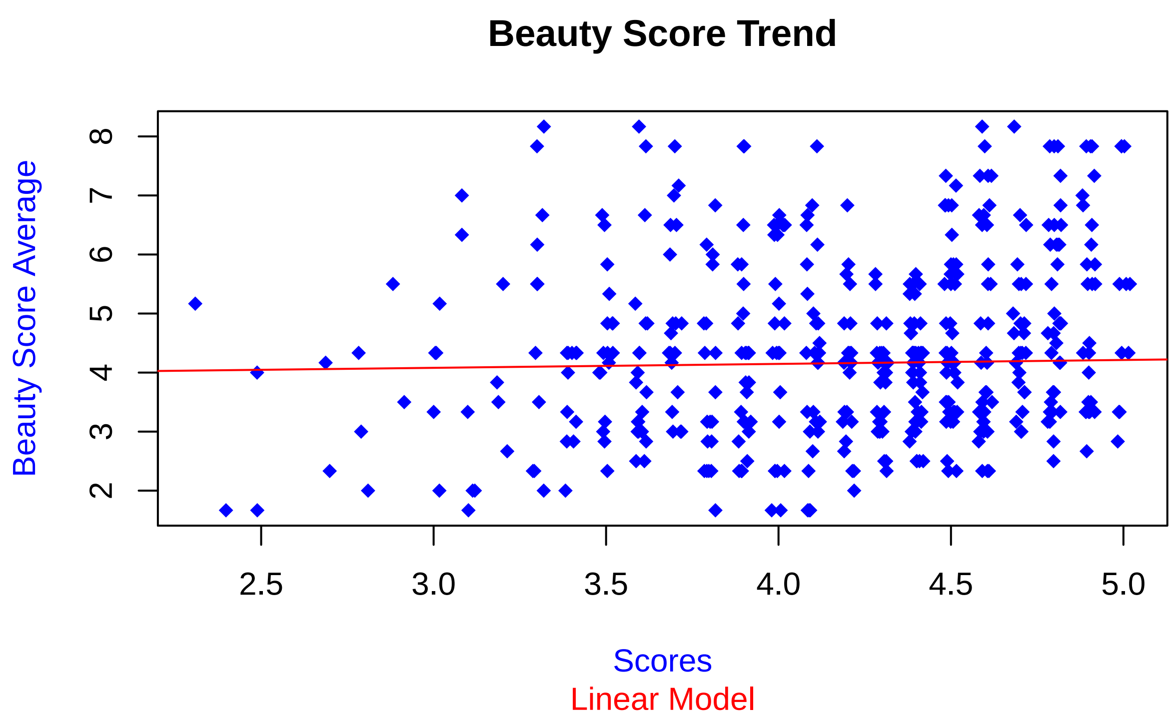


ANSWER:
We definitely were not showing all the data, some rows were missing without use of geom_jitter()

5.

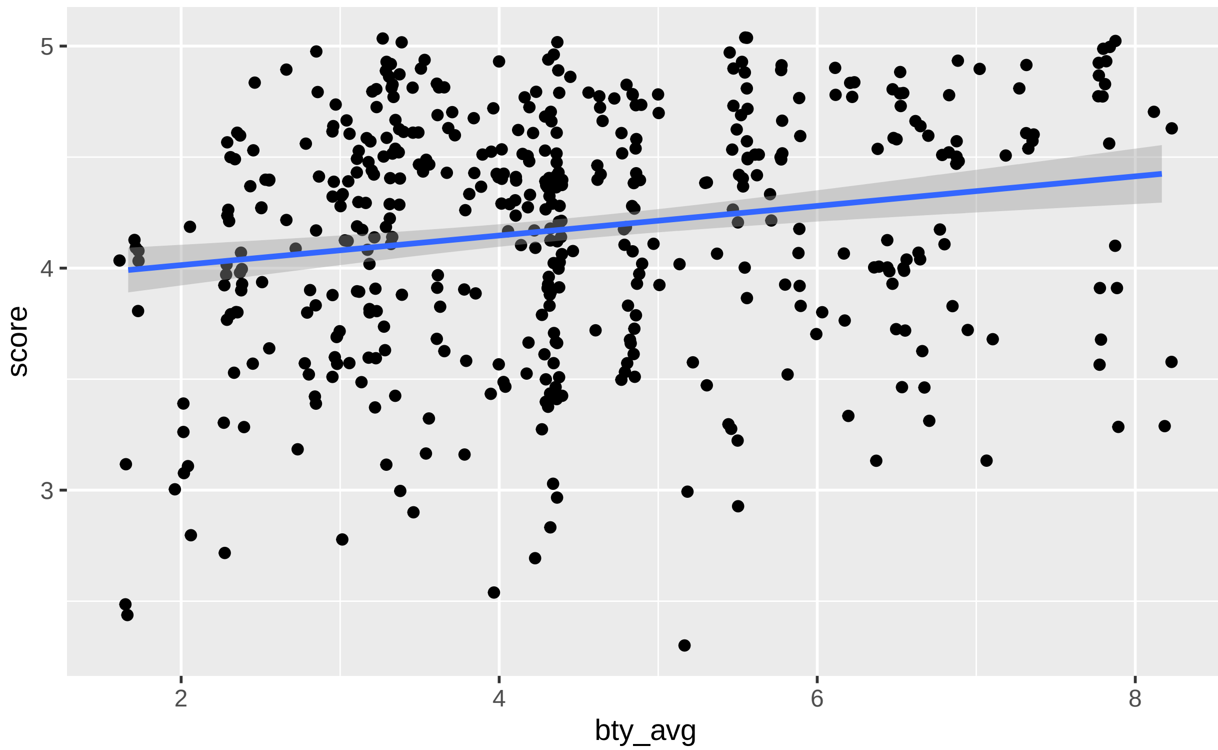
```
Let's see if the apparent trend in the plot is something more than
natural variation. Fit a linear model called 'm_bty' to predict average
professor score by average beauty rating. Write out the equation for the linear
model and interpret the slope. Is average beauty score a statistically significant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

```r
m_bty <- lm(evals$score~evals$bty_avg)
m_bty
plot(evals$bty_avg~jitter(evals$score),main = 'Beauty Score Trend',sub = 'Linear Model', xlab = 'Scores
abline(m_bty, col='red')
```



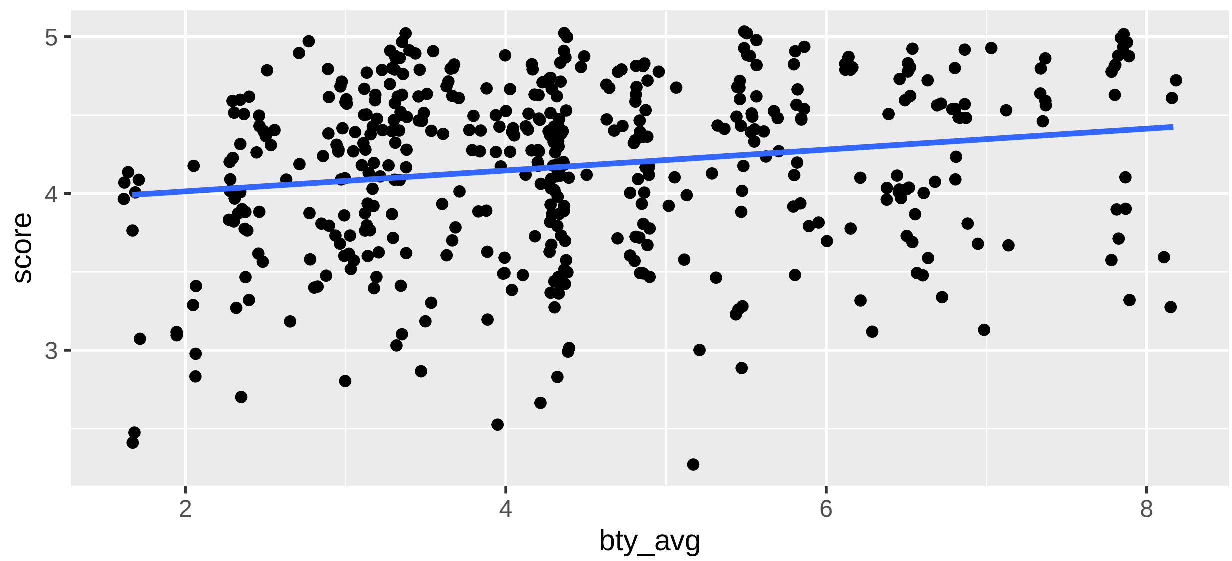
Add the line of the bet fit model to your plot using the following:

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



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

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



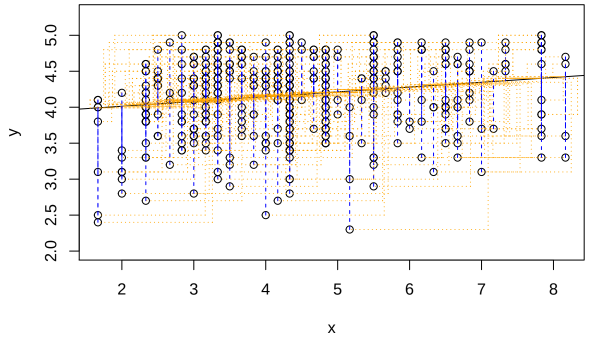
6.

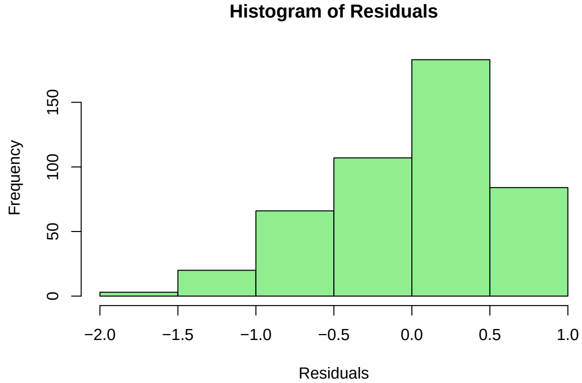
Use residual plots to evaluate whether the conditions of least squares regression are reasonable. Provide plots and comments for each one (see the Simple Regression Lab for a reminder of how to make these).

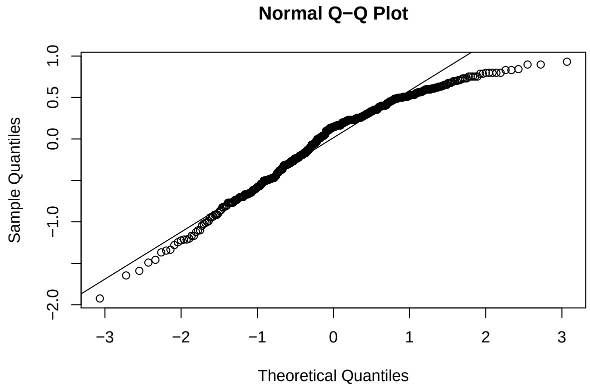
```
plot_ss(x = evals$bty_avg, y = evals$score, showSquares = TRUE)
```

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

```
hist(m_bty$residuals,xlim = c(-2,1),main = "Histogram of Residuals", col = 'light green',xlab = "Residu
```







ANSWER:

- **Linearity -** Positive & Slightly Linear
- **Nearly Normal residual -** Condition not met histogram left skewed
- **Constant Variability -** Condition met based of residual plot
- **Independent Observations -** Conditions assumed to be met

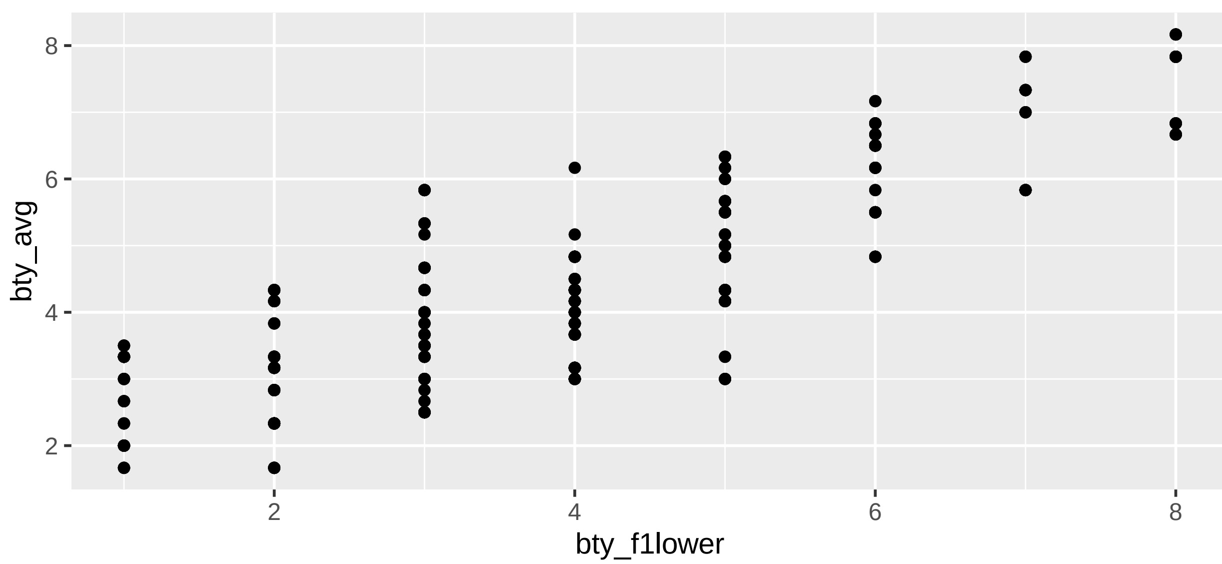
Not all conditions met, regression method will be rejected.

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.

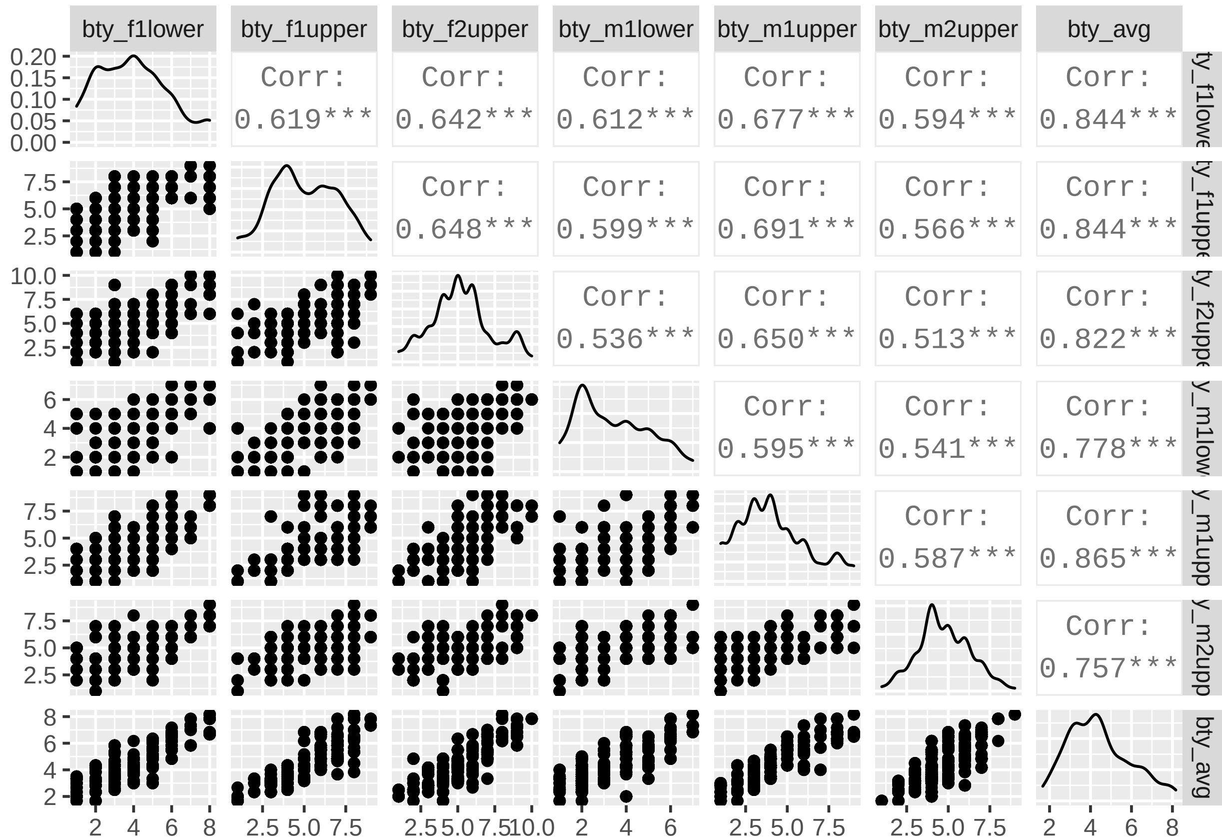
```
ggplot(data = evals, aes(x = bty_f1lower, y = bty_avg)) +
  geom_point()

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.

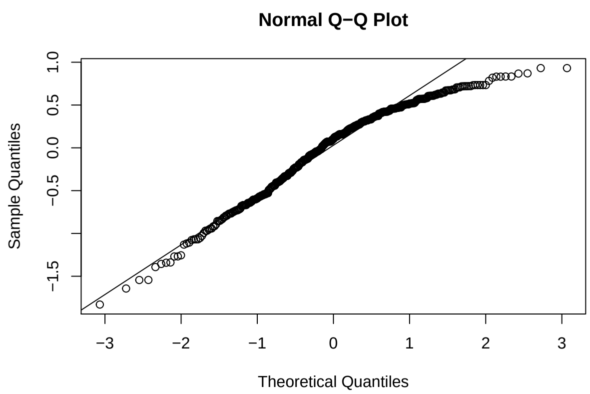
In order to see if beauty is still a significant predictor of professor score after you've accounted for the professor's gender, you can add the gender term into the model.

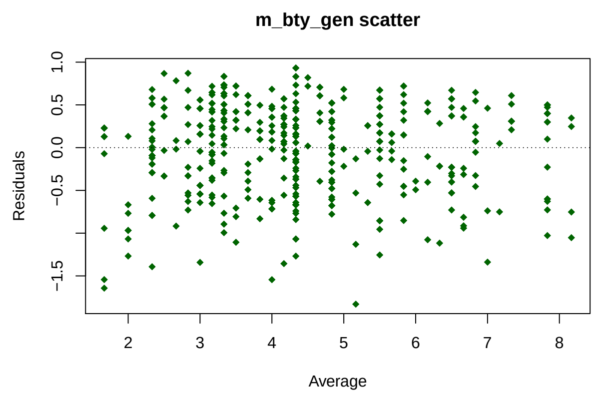
```
m_bty_gen <- lm(score ~ bty_avg + gender, data = evals)
summary(m_bty_gen)
```

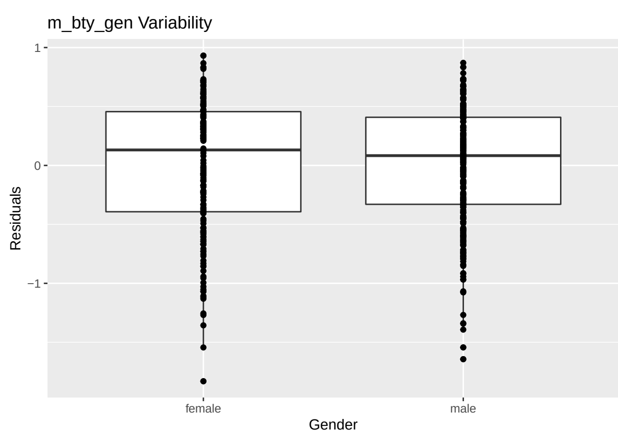
7.

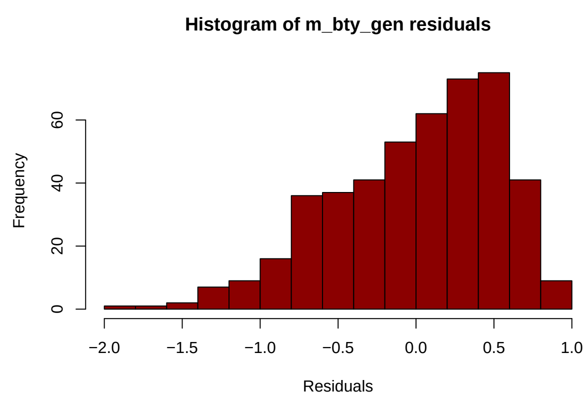
```
P-values and parameter estimates should only be trusted if the
conditions for the regression are reasonable. Verify that the conditions
for this model are reasonable using diagnostic plots.
```

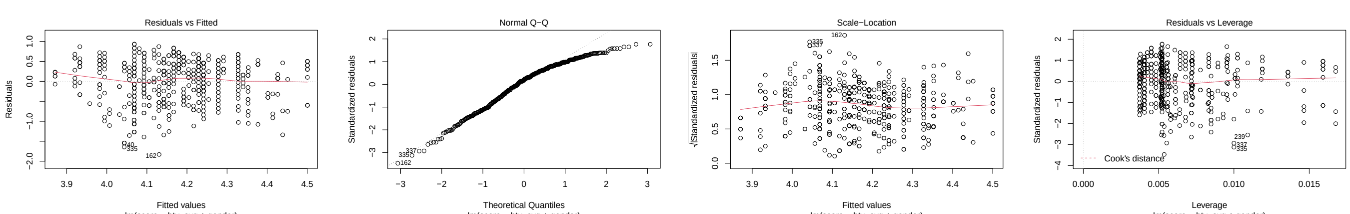
- **Linearity** - Condition met.
- **Nearly Normal residual** - Condition met based on `Histogram of m_bty_gen residuals` albiet slighly left skewed
- **Constant Variability** - Condition met based on `m_bty_gen scatter` & `m_bty_gen Variability` Where it is also noted the median are similar and each variable is related to the outcome
- **Independent Observations** - Conditions independence established based on how data was collected











8.

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

ANSWER:
Basically, `gender` enhances `bty_avg` more relevant.
P-value for `m_bty$bty_avg` vs `m_bty_gen$bty_avg` is 5.083e-05 > 6.48e-06 supporting this fact.

```
summary(m_bty)
summary(m_bty_gen)
```

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.

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

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:
As demonstrated below males tend to have a higher rating of at least 0.17239

	Estimate
(Intercept)	3.74734
bty_avg	0.0741
gendermale	0.17239

$\therefore B_0 = 3.74734$ and $B_1 = 0.07416$ and noting the details regarding males

- $scor\hat{e}_{male}$ = 3.74734+0.07416 × `bty_avg`+0.17239 × 1 = **3.74734+0.07416** × **`bty_avg`+0.17239**
- $scor\hat{e}_{female}$= 3.74734+0.07416 × `bty_avg`+0.17239 × 0 = **3.74734+0.07416** × **`bty_avg`**

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: R handles categorical variables in this case by creating 2 more levels:

ranktenure track	-0.16070 0.07395 -2.173 0.0303
ranktenured	-0.12623 0.06266 -2.014 0.0445

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

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:
Based on Open Intro ‘eval’ data set which states:

- **cls_profs** - Number of professors teaching sections in course in sample: single, multiple.

I’d assume this is the variable that would have the highest `p-value` since its least associated with the professor score.

Let’s run the model…

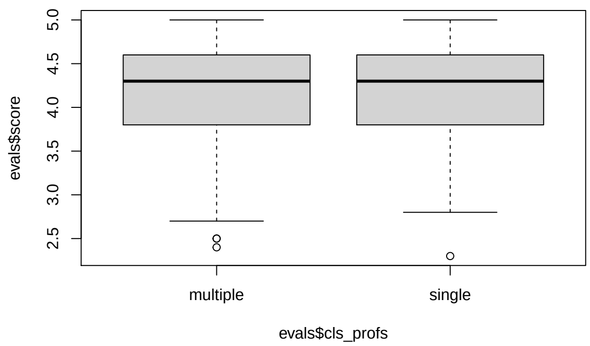
```
m_full <- lm(score ~ rank + gender + ethnicity + language + age + cls_perc_eval
             + cls_students + cls_level + cls_profs + cls_credits + bty_avg
             + pic_outfit + pic_color, data = evals)
summary(m_full)
```

12.

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

ANSWER:
With a `p-value` of 0.5847 I believe my suspicions were correct.

```
m_a<-lm(score~cls_profs,data=evals)
summary(m_a)
plot(evals$score ~ evals$cls_profs)
```

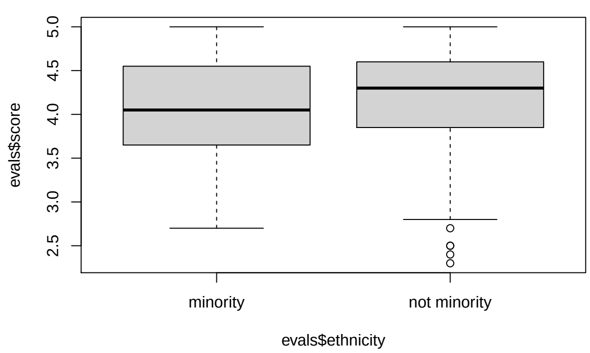


13.

```
Interpret the coefficient associated with the ethnicity variable.
```

ANSWER:
Ratings for professors that qualify as "not minority" is higher by 0.1032.

```
m_c<-lm(score~ethnicity,data=evals)
summary(m_c)
plot(evals$score ~ evals$ethnicity)
```



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:
Slight change in coefficients and significance, to where they're all slightly lower, implying more significance.

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

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:
$\hat{score} = \hat{\beta}_0 + \hat{\beta}_1 \times ethnicity\ not\ minority + \hat{\beta}_2 \times gender\ male + \hat{\beta}_3 \times language\ non - english + \hat{\beta}_4 \times age + \hat{\beta}_5 \times class\ percent\ evaluations + \hat{\beta}_6 \times one\ class\ credit + \hat{\beta}_7 \times beauty\ average + \hat{\beta}_8 \times colored\ picture$

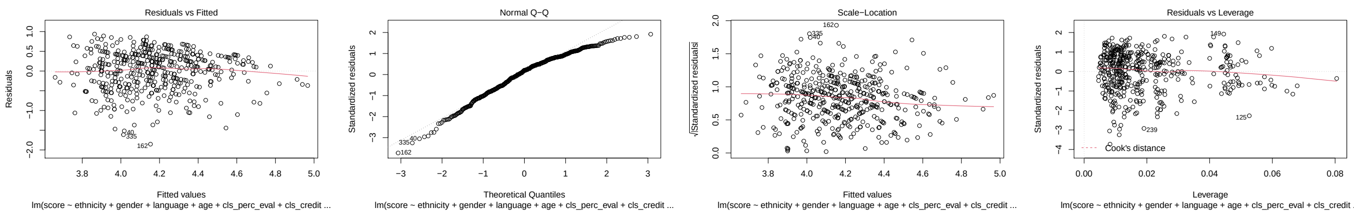
```
m_e <- lm(score ~ ethnicity + gender + language + age + cls_perc_eval
              + cls_credits + bty_avg
              + pic_color, data = evals)
summary(m_e)
```

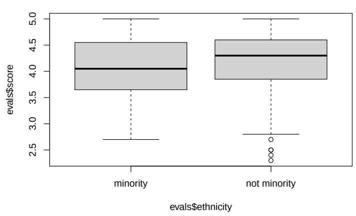
16.

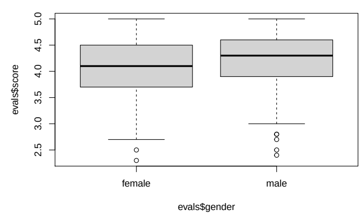
```
Verify that the conditions for this model are reasonable using diagnostic
plots.
```

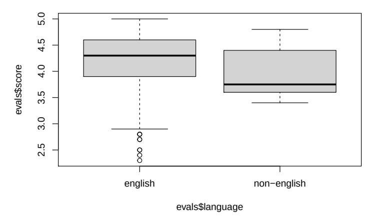
Answer:

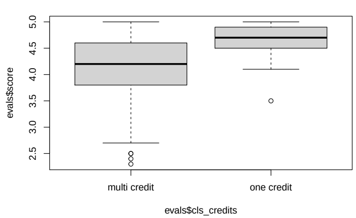
- Linearity ✓
- Nearly Normal residual ✓
- Constant Variability ✓
- Independent Observations ✓

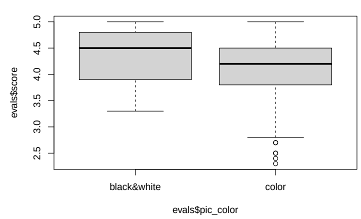


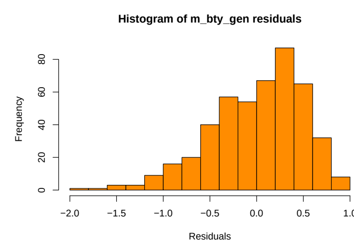












17.

```
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:
I imagine some may question if independence was violated, but I believe the scores given would still be considered independent, so no I do not believe it will make a difference.

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:
A professor with a high percentage for student evaluation in a one-credit class, who is a non-minority, english speaking teacher would get the highest evaluations based on our data.

19.

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

ANSWER:
I wouldn't simply because this is an observation in a specific region with it's own specific demographics. I would be comfortable with a larger, more diversified data set conducted as an experiment.

References
