# Lab 11 - Math 58 / 58b: full linear regression analysis

*done during lab April 22 or 24, 2020*

*not due*

## Lab Goals

- a full analysis including multiple variables!
- including inference in the linear model

## 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" (Hamermesh and Parker, 2005) found that instructors who are viewed to be better looking receive higher instructional ratings. (Daniel S. Hamermesh, Amy Parker, Beauty in the classroom: instructors pulchritude and putative pedagogical productivity, *Economics of Education Review*, Volume 24, Issue 4, August 2005, Pages 369-376, ISSN 0272-7757, 10.1016/j.econedurev.2004.07.013. http://www.sciencedirect.com/science/article/pii/S0272775704001165.)

## The data

The data were gathered from end of semester student evaluations for a large sample of professors at the University of Texas, Austin. Additionally, six students rated the professors' physical appearance. (Ours is a slightly modified version of the original data set released as part of the replication data for *Data Analysis Using Regression and Multilevel/Hierarchical Models* (Gelman & Hill, 2007).) In the resulting data frame, each row represents a different course and columns represent variables about the courses and professors.

| variable | description |
| --- | --- |
| score | average professor evaluation score: (1) very unsatisfactory - (5) excellent. |
| rank | rank of professor: teaching, tenure track, tenured. |
| ethnicity | ethnicity of professor: not minority, minority. |
| gender | gender of professor: female, male. |
| language | language of school where professor received education: english or non-english. |
| age | age of professor. |
| cls_perc_eval | percent of students in class who completed evaluation. |
| cls_did_eval | number of students in class who completed evaluation. |
| cls_students | total number of students in class. |
| cls_level | class level: lower, upper. |
| cls_profs | number of professors teaching sections in course in sample: single, multiple. |
| cls_credits | number of credits of class: one credit (lab, PE, etc.), multi credit. |
| bty_f1lower | beauty rating of professor from lower level female: (1) lowest - (10) highest. |
| bty_f1upper | beauty rating of professor from upper level female: (1) lowest - (10) highest. |
| bty_f2upper | beauty rating of professor from second upper level female: (1) lowest - (10) highest. |
| bty_m1lower | beauty rating of professor from lower level male: (1) lowest - (10) highest. |
| bty_m1upper | beauty rating of professor from upper level male: (1) lowest - (10) highest. |
| bty_m2upper | beauty rating of professor from second upper level male: (1) lowest - (10) highest. |
| bty_avg | average beauty rating of professor. |
| pic_outfit | outfit of professor in picture: not formal, formal. |
| pic_color | color of professor's picture: color, black & white. |

```
evals <- read_csv("https://www.openintro.org/data/csv/evals.csv")
```

Note that the lab was taken in its entirety from the OpenIntro materials. The lab was written by Mine Çetinkaya-Rundel and Andrew Bray.

In this lab we will analyze the data from this study in order to learn what goes into a positive professor evaluation.

## 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.

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?

3. Excluding `score`, select two other variables and describe their relationship using an appropriate visualization (scatterplot, side-by-side boxplots, barplot, or histogram).

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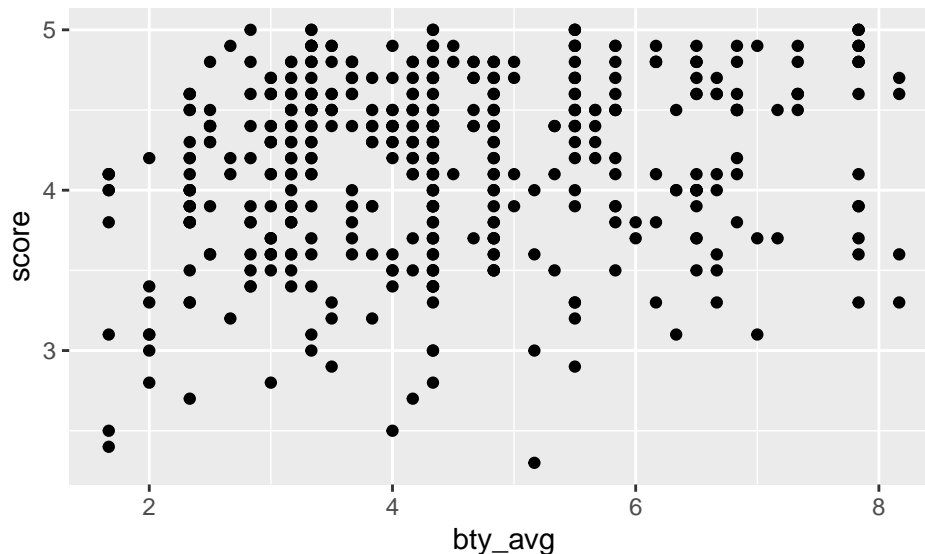


Before we draw conclusions about the trend, compare the number of observations in the data frame with the approximate number of points on the scatterplot. Is anything awry?

It is hard to count the points, but we can definitely see some patterns that seem to have more to do with the data collection than the effect. That is, `bty_avg` is only scored on a few different values (similar for `score`, but less so).

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

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 and add the line to your plot

using `geom_smooth(method = "lm")`. 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?

6. Use residual plots to evaluate whether the conditions of least squares regression are reasonable. Provide plots and comments for each one.

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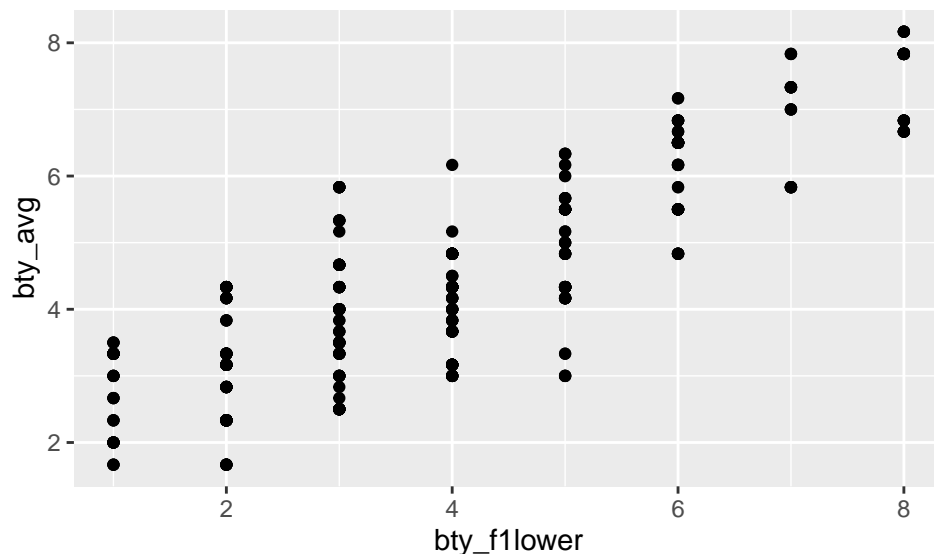


```
evals %>% select(bty_f1lower, bty_avg) %>%
  cor()
```

```
##             bty_f1lower   bty_avg
## bty_f1lower   1.0000000 0.8439112
## bty_avg       0.8439112 1.0000000
```

As expected the relationship is quite strong - after all, the average score is calculated using the individual scores. We can actually take a look at the relationships between all beauty variables (columns 13 through 19) using the following command:

```
library(GGally)
ggpairs(evals, columns = c(13:19))
```
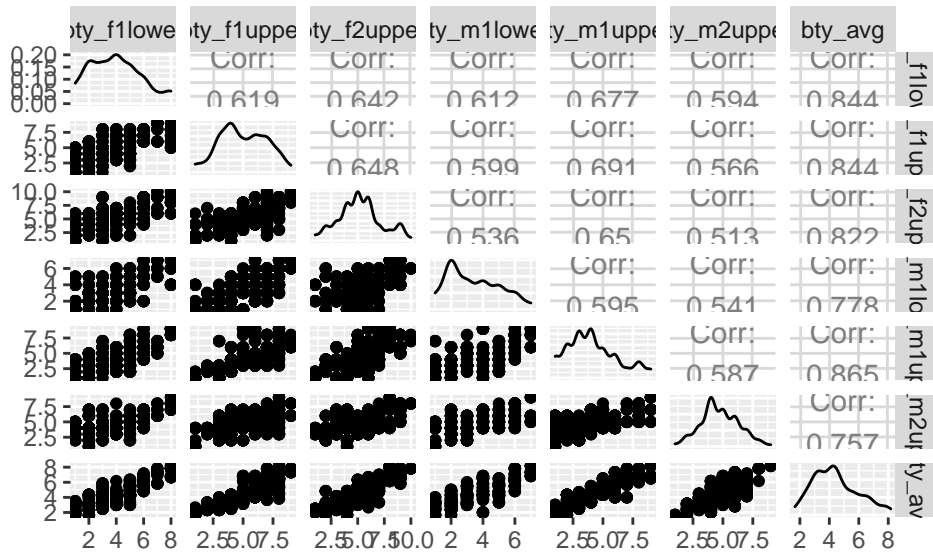
These variables are collinear (correlated), and adding more than one of these variables to the model would not add much value to the model. In this application and with these highly-correlated predictors, it is reasonable to use the average beauty score as the single representative of these variables.

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

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

```
## # A tibble: 3 x 5
##   term        estimate std.error statistic   p.value
##   <chr>          <dbl>     <dbl>     <dbl>     <dbl>
## 1 (Intercept)   3.75     0.0847      44.3  6.23e-168
## 2 bty_avg       0.0742   0.0163       4.56 6.48e-  6
## 3 gendermale    0.172    0.0502       3.43 6.52e-  4
```

Let's visualize the two different models on the plot. Clearly the best fit linear model has a different intercept and a different slope across the two genders!

```
evals %>%
  ggplot(aes(x = bty_avg, y = score, color = gender)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE)
```
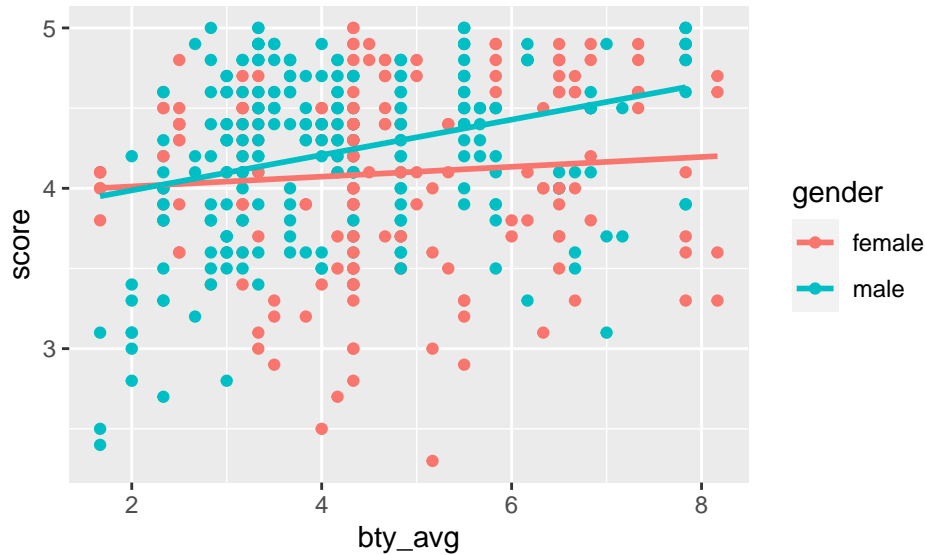
7. P-values and parameter estimates should only be trusted if the conditions for the regression are reasonable. Verify that the conditions for this model are reasonable using a residual plot.

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`?

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 `female` and `male` to being an indicator variable called `gendermale` that takes a value of 0 for females and a value of 1 for males. (Such variables are often referred to as "dummy" variables.)

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

$$
\begin{aligned}
\widehat{score}_F &= b_0 + b_1 \times bty\_avg + b_2 \times (0) \\
&= b_0 + b_1 \times bty\_avg
\end{aligned}
$$

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

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. (Advanced R topic: 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`.

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

m_bty_rank %>% tidy()

## # A tibble: 4 x 5
##   term          estimate std.error statistic  p.value
##   <chr>            <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)       3.98    0.0908      43.9 2.92e-166
```

```
## 2 bty_avg              0.0678      0.0165       4.10 4.92e- 5
## 3 ranktenure track  -0.161       0.0740      -2.17 3.03e- 2
## 4 ranktenured       -0.126       0.0627      -2.01 4.45e- 2
```

There are now **three** models, one for each rank of the professor!

$$
\begin{aligned}
\widehat{score}_{teach} &= b_0 + b_1 \times bty\_avg + b_2 \times (0) + b_3 \times (0) \\
&= b_0 + b_1 \times bty\_avg \\
\widehat{score}_{tt} &= b_0 + b_1 \times bty\_avg + b_2 \times (1) + b_3 \times (0) \\
&= (b_0 + b_2) + b_1 \times bty\_avg \\
\widehat{score}_{tenured} &= b_0 + b_1 \times bty\_avg + b_2 \times (0) + b_3 \times (1) \\
&= (b_0 + b_3) + b_1 \times bty\_avg
\end{aligned}
$$

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, ethnicity, gender, 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.

Let's run the model...

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

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

13. Interpret the coefficient associated with the ethnicity variable.

**Given that everything else stays the same**, a change from "ethnic" to "not ethnic" (i.e., a one unit increase in the varaible `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?

15. Verify that the conditions for this model are reasonable using a residual plots.

16. 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?

17. 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.

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