Multiple linear regression

Your reproducible lab report: Before you get started, download the R Markdown template for this lab. Remember all of your code and answers go in this document:

```
download.file("https://dyurovsky.github.io/85309/post/rmd/lab10.Rmd",
              destfile = "lab10.Rmd")
```

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.

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

Getting Started

As usual, we’re going to load the `tidyverse` package for data manipulation. We’ll also be using the `GGally` package for the `ggpairs` function that generates pairwise correlation plots. If you don’t have `GGally` installed yet, you can install it by typing `install.packages('GGally')`.

We’ll also be reading in a dataset to work with just like we usually do.

```
library(tidyverse)
library(GGally)

evals <- read_csv("https://dyurovsky.github.io/85309/data/lab10/evals.csv")
```

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.

We have observations on 21 different variables, some categorical and some numerical. The meaning of each variable can be found in the codebook below:

Codebook

- `course_id` : Variable identifying the course (out of 463 courses).
- `score` : Average professor evaluation score: (1) very unsatisfactory - (5) excellent.

Professor Information:

- `prof_id` : Variable identifying the professor who taught the course (out of 94 professors).
- `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.
- `pic_outfit` : Outfit of professor in picture: not formal, formal.
- `pic_color` : Color of professor's picture: color, black & white.

Course Information:

- `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.

Beauty Information:

- `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 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.

Exploring the data

Exercise 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.

Exercise 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?

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

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

Exercise 3

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

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

Exercise 4

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

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

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

The blue line is the linear model (`lm`), and the `se` parameter being set to false tells *R* not to plot the estimated standard errors from the model.

Exercise 5 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).

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 %>%
  summarise(cor = cor(bty_avg, bty_f1lower)) %>%
  pull()
```

As expected the relationship is quite strong—after all, the average score is calculated using the individual scores. We 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 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)
summary(m_bty_gen)
```

Exercise 6 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.

Exercise 7 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.

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

Exercise 8

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. (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 9

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`.

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.

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)

summary(m_full)
```

Exercise 10 Interpret the coefficient associated with the ethnicity variable.

More Practice

Exercise 11 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?

Exercise 12 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.

Exercise 13 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.