Lab 3: Multiple linear regression

Due date: Thursday, Feb 27, 2025 submitted as Word document to Canvas ***Lab03*** link

This lab counts 9 % toward your total grade.

Objectives: In this lab, you will practice your skills in

a) Explore multiple regression
b) Multicollinearity
c) Model selection

Format of answer: Submit your answers as a **Word document** with graphs and verbal descriptions, properly labeled in the task sequence, with answers in red text and only relevant content included

Data Introduction - evals

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.

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

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 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.

Task 1: Exploring the data (4 pts)

a) 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 quick look at the relationship between one of these scores (bty_f1lower) and the average beauty score(bty_avg) using **plot()**.Please explain the relationship between these two variables. (0.5 pts)

b) Using **car::vif()** function to check the multicollinearity for independent variables (bty_f1lower,bty_f1upper,bty_f2upper, bty_m1lower, bty_m1upper, bty_m2upper, bty_avg) and dependent variable(score). Please illustrate the multicollinearity relationship among them and justify the reason that use the average beauty score (bty_avg) as a single representative of these variables. (1 pts)

c) The model you finalized in Task1.a is a bivariate model. 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 (gender) into the model. (0.5 pts)

d) p-values and parameter estimates should only be trusted if the conditions for the regression are reasonable. Verify that the conditions for Task1.c model are reasonable using diagnostic plots (plot the histogram of residual to assess the normality of the residual; using residual plot (residual vs fitted value) to assess the variance in residual. (1 pts)

e) 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.). 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.

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

$$\widehat{score} = \widehat{\beta_0} + \widehat{\beta_1} * bty_{ave} + \widehat{\beta_2} * 0$$

a) What is the equation corresponding to males? For two professors who received the same beauty rating, which gender tends to have the higher course evaluation score? Please explain the $\beta_0, \beta_1, \beta_2$. (0.5 pts)

f) 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. (0.5 pts)

Task 2: The search for the best model (5 pts)

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.

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)

a) 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.(0.5 pts)

b) Check your suspicions based on the summary from the full model. Include the model output in your response. (1 pts)

c) Interpret the coefficient associated with the ethnicity variable. (1 pts)

d) Using different model selection methods (best subset, backward stepwise, forward stepwise), determining the best model based on AIC as the selection criterion (1 pts)

b) Based on your final model (You can use any best model from the three

methods), describe the characteristics of a professor and course at University of Texas at Austin that would be associated with a high evaluation score. (0.5 pts)

e) Would you be comfortable generalizing your conclusions to apply to professors generally (at any university)? Why or why not? (1 pts)