Assignment 06

Regression Assumptions

The file *beauty.csv* contains data collected from student evaluations of instructors' beauty and teaching quality for several courses at the University of Texas. The teaching evaluations were conducted at the end of the semester, and the beauty judgments were made later, by six students who had not attended the classes and were not aware of the course evaluations. The variables are:

- `prof`: Professor ID number
- `avgeval`: Average course rating
- `btystdave`: Measure of the professor's beauty composed of the average score on six standardized beauty ratings
- `tenured`: 0 = non-tenured; 1 = tenured
- `nonenglish`: 0 = native English speaker; 1 = non-native English speaker
- `age`: Professor's age (in years)
- `female`: 0 = male; 1 = female
- `students`: Number of students enrolled in the course
- `percentevaluating`: Percentage of enrolled students who completed an evaluation

These source of these data is Hamermesh, D. S. & Parker, A. M. (2005). Beauty in the classroom: Instructors' pulchritude and putative pedagogical productivity. *Economics of Education Review, 24*, 369–376. The data were made available by Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. New York: Cambridge University Press. Use these data to answer each of the following questions. Each question is worth one point unless otherwise noted. The entire assignment is worth 19 points.

Fitting Models

Fit the following regression model using R. You will use the output from the fitted model to answer the questions in the assignment.

- Model 1: `avgeval ~ btystdave + percentevaluating`

Preliminary Examination of Assumptions

1. Create the density plots for the outcome and each of the two predictors (three total). Size them so that they do not take up any more than one page altogether. **(2pts.)**
2. Do any of these distributions foreshadow problems for the normality assumption? Explain.
3. Create the scatterplot of the outcome vs. each of the two predictors (two total). Include the fitted simple regression line in each plot. Size them so that they do not take up any more than one-half a page altogether. **(2pts.)**
4. For each of the two scatterplots describe the structural relationship between the outcome and predictor (e.g., linear, non-linear, etc.). Do any of these relationships foreshadow problems for the linearity assumption? Explain.
5. Based on the scatterplots, are there any observations that you would identify as outliers? If so, identify the observation(s) by marking them in the scatterplot in which they are problematic, and identifying the observation number (row in the data frame).

Examination of the Standardized Residuals from Each Simple Regression Model

6. Create the residual plots using the standardized residuals for each of the simple regression models (two total). **(2pts.)**
7. Do any of these plots foreshadow problems for the linearity assumption? Explain.
8. Do any of these plots foreshadow problems for the homogeneity of variance assumption? Explain.
9. Do any of these plots point to any extreme observations? If so, identify the observation(s) by marking them in the scatterplot in which they are problematic, and identifying the observation number (row in the data frame).

Examination of the Standardized Residuals from the Multiple Regression Model

10. Create the residual plots using the standardized residuals for the two-predictor multiple regression model (one total). **(2pts.)**
11. Does this plot suggest problems about meeting the linearity assumption? Explain.
12. Does this plot suggest problems problems about meeting the homogeneity of variance assumption? Explain.
13. Does this plot point to any extreme observations? If so, identify the observation(s) by marking them in the scatterplot in which they are problematic, and identifying the observation number (row in the data frame).
14. Create the density plot of the marginal distribution of the standardized residuals. Add the confidence envelope for the normal distribution.
15. Does this plot suggest problems about meeting the assumption of normality? Explain.