Regression Analysis
Assignment 4

Problem 5.1 on pages 234-235: This exercise uses the COLLGPA dataset and covers influential points (see pages 189-199). First, you'll fit a basic model with no transformations in part (a), discover that this model is unsatisfactory, and then improve the model by adding an interaction term and two quadratic terms; you'll use this improved model for all remaining parts of the exercise. You'll then investigate whether there are any outliers or points with high leverage for this model. Finally, you'll use Cook's distance to give you some additional insight into how influential the most influential points are.

Problem 5.1: This problem is adapted from one in McClave et al. (2005). The **COLLGPA** data file contains data that can be used to determine whether college grade point average (GPA) for 40 students (*Gpa*) can be predicted from:

Verb = verbal score on a college entrance examination (percentile),
Math = mathematics score on a college entrance examination (percentile).

Admission decisions are often based on college entrance examinations (among other things), so this is a common use of regression modeling.

a. Use statistical software to fit the model

$$E(Gpa) = b_0 + b_1 Verb + b_2 Math.$$

Save the studentized residuals and draw a scatterplot with these studentized residuals on the vertical axis and *Verb* on the horizontal axis. Repeat with *Math* on the horizontal axis. Add a "loess" fitted line to each of your plots to help you assess the zero mean regression errors assumption [computer help #35, #15, and #36]. Do either of the plots suggest that this assumption is violated for this model?

b. Use statistical software to create the interaction *VerbMath*, and the transformations $Verb^2$ and $Math^2$, and fit the full quadratic model:

$$E(Gpa) = b_0 + b_1 Verb + b_2 Math + b_3 VerbMath + b_4 Verb^2 + b_5 Math^2 .$$

Again save the studentized residuals, and in addition save leverages and Cook's distances [computer help #35, #37, and #38]. Draw a scatterplot of the studentized residuals on the vertical axis versus the (standardized) predicted values on the horizontal axis [computer help #15, #36, and #39]. This residual plot suggests that the zero mean, constant variance, and independence regression error assumptions are satisfied for this quadratic model. Briefly describe what you see (or fail to see) in this residual plot that leads you to this conclusion.

Note: Standardized predicted values are calculated by subtracting their sample mean and dividing by their sample standard deviation (so that they have a mean of 0 and standard deviation of 1). Some software packages can draw residual plots with standardized predicted values on the horizontal axis automatically. Such plots are essentially identical in appearance to residual plots with ordinary (unstandardized) predicted values on the horizontal axis (the only difference is the scale of the horizontal axis of the resulting plot). Plot whatever is easier to do in your particular software. In a real-life regression analysis you would also go on to check the zero mean, constant variance, and independent regression error assumptions in residual plots with Verb on the horizontal axis and also with Math on the horizontal axis. If you were to do that here, you would find that the assumptions are also satisfied in these residual plots.

c. Draw a histogram and QQ-plot of the studentized residuals for the quadratic model you fit in part (b) [computer help #14 and #22]. What do they suggest about the normality regression error assumption?
d. Are there any outliers for the quadratic model you fit in part (b)? (Remember to justify your answer).
e. Draw a scatterplot with the leverages from the quadratic model you fit in part (b) on the vertical axis and ID on the horizontal axis [computer help #15]. Which student has the highest leverage, and why? *If* you think you should investigate further, do so by seeing what happens if we exclude this student from the analysis [computer help #19].

   Hint: Look at this student's predictor values to see why his/her leverage is high, and consult Section 5.1.2 to see if you should investigate further (if so, delete this student, refit the model, and see how much regression parameter estimates change).

f. Draw a scatterplot with Cook's distances from the quadratic model in part (b) on the vertical axis and ID on the horizontal axis [computer help #15]. Which student has the highest Cook's distance, and why? *If* you think you should investigate further, do so by seeing what happens if we exclude this student from the analysis.

   Hint: Look at this student's studentized residual and leverage to see why his/her Cook's distance is high, and consult Section 5.1.3 to see if you should investigate further.

[20 points]

Problem 5.7 on pages 241-242: This exercise is an open-ended challenge to fit a multiple linear regression model to some data on restaurants. Try to follow the model building guidelines in Section 5.3 as best you can, and strive to come up with a "good" model (for this application, a "good" model should have an R-squared value of approximately 0.94 and a regression standard error, *s*, of approximately 10). You could potentially spend many hours on this exercise, but it should be possible to come up with a decent model within an hour or so; if you find yourself spending much more time than this, chances are you're on the wrong track or you're working too hard!

Problem 5.7: The following problem provides a challenging dataset that you can use to practice multiple linear regression model building. You've been asked to find out how profits for 120 restaurants in a particular restaurant chain are affected by certain characteristics of the restaurants. You would like to build a regression model for predicting *Profit* = annual profits (in thousands of dollars) from five potential predictor variables:

Cov = number of covers or customers served (in thousands)

Fco = food costs (in thousands of dollars)

Oco = overhead costs (in thousands of dollars)

Lco = labor costs (in thousands of dollars)

Region = geographical location (Mountain, Southwest, or Northwest)

Note that region is a qualitative (categorical) variable with three levels; the **RESTAURANT** data file contains two indicator variables to code the information in region: D_{Sw} = 1 for Southwest, 0 otherwise, and D_{Nw} = 1 for Northwest, 0 otherwise. Thus, the Mountain region is the reference level with 0 for both D_{Sw} and D_{Nw}. Build a suitable regression model and investigate the role of each of the predictors in the model through the use of predictor effect plots. You may want to consider the following topics in doing so:

- models with both quantitative and qualitative variables;
- polynomial transformations;
- interactions;

- comparing nested models.

You may use the following for terms in your model:

- *Cov*, *Fco*, *Oco*, *Lco*, D_{Sw}, D_{Nw};
- interactions between each of the quantitative predictors and the indicator variables, such as $D_{Sw}Cov$, $D_{Nw}Cov$, etc.;
- quadratic terms, such as Cov^2 (do not use terms like D_{Sw}^2, however!);
- use *Profit* as the response variable [i.e., do not use $\log_e(Profit)$ or any other transformation].

[20 points]