Homework 8 (20 pts. + 1.5 pt. Bonus)
due March 21

A reminder – Please do not hand in any unlabeled or unedited SAS output. Include in your write-up only those results that are necessary to present a complete solution (what you want the grader to grade). In particular, questions must be answered in order (including graphs), and all graphs must be fully labeled (main title should include the question number, and all axes should be labeled). Don't forget to put all necessary information (see course policies) on the first page. Include the SAS input for all questions at the very end of your homework; this could be important even though it won't be graded. You will often be asked to continue problems on successive homework assignments so save all your SAS code.

1. (9 pts.) This question uses the situation presented in Cosmetic sales, Problem 10.13 in the book (CH010PR13.txt: Note: this data set has been modified from the one in the textbook). For parts c, d, e, f, g ii, do not include in your answer the values for all cases. Use plots and verbal summaries instead. You may include values for a few selected cases if you wish.
    (a) From the previous problem set, write down the X and Y outliers that you have determined (HW7 2d and 2e). You will get no credit for this, but it is necessary to have this information to finish the problem.
    (b) Are there any influential points? (Hint: DFFITS, Cook's distance, DFBETAS), covariance ratio)
    (c) Is there a serious multicollinearity problem?
        i. Include an appropriate scatterplot and correlation values between the explanatory variables.
        ii. Using the new information in this section, do you think there is a problem with multicollinearity? (Hint: VIP or tolerance)
        iii. Compare your answers in parts i and ii. Are your conclusions the same or different? Please explain your answer.
    (d) Instead of removing variables, we are going to use the Ridge Regression to determine the parameter values.
        i. Make a ridge trace plot (Hint: the c value is higher here than in the example in class, please test values up to 0.2.) What value of c do you believe is best? Explain your choice.
        ii. Using the VIF factors, what value of c do you believe should be used? (Hint: Look at both the graph and the printed numbers.) Explain your choice.
        iii. Using the best choice for c, what are the standard regression parameters in this situation?
        iv. Compare your answer with d iii) with the parameters given in c ii). Are they similar? Which values are 'better'? Explain your answer.

2. (4.5 pts.) This question uses the situation presented in Machine speed, Problem 11.7 in the book (CH11PR07.DAT).

   (a) Fit a linear regression function by ordinary least squares; obtain the residuals and plot the residuals against X. What does the residual plot suggest?

   (b) Plot the absolute value of the residuals and the squared residuals vs. X. Which plot has a better line?

   (c) Perform a weighted least square using the squared residuals to compute the weights. Obtain the weighted least squares estimates for the estimated parameters and their standard errors. Are these values similar to the ones produced in a)? Which results are better, the ones generated in a) or c)? Please explain your answer.

   (d) Re-calculate the residuals for the weighted least squares and make a residual plot vs. X. Did this correct the problem that was seen in a).

   (e) **BONUS (1.5 pt.)**: Repeat parts c) and d) but plot the absolute value of the residuals vs. X to compute the weights. Again obtain the weighted least squares estimates for the estimated parameters and their standard errors and generate the new residual plot. Which method produces better results: using the squared residuals or the absolute value of the residuals. Explain your answer.?

3. (6.5 pts.) This question uses the situation presented in Filling machines, Problem 16.11 in the book (CH16PR11.DAT). (This problem will be continued in Homework 9.)

   (a) Give a table of sample sizes, means, and standard deviations for the six different filling machines.

      i. Based on this table, does the constant variance assumption appear to hold?
      ii. Based on this table, do the six machines appear to give the same average amount of detergent?

   (b) Make a plot of the means (with i=join) overlaid on a plot of the individual observations (with i=none) versus machine number. (Sample code is in nknw677a.sas)

      i. Based on this plot, does the constant variance assumption appear to hold?
      ii. Based on this plot, do the six machines appear to give the same average amount of detergent?

   (c) Examine the questions of whether or not the six machines have the same average amount of detergent.

      i. Write the cell means model for this analysis. State the null and alternative hypotheses in terms of the cell means model parameters, give the test statistic with degrees of freedom, the p-value, decision and your conclusion.
      ii. Write the factor effects model for this analysis. State the null and alternative hypotheses in terms of the factor effects model parameters, give the test statistic with degrees of freedom, the p-value, decision and your conclusion. [Note: some of this part, we will cover on Tuesday March 19.]