Homework 6 (16 pts.)
due Feb. 28

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.

The following two problems are a continuation of Problem 3 in HW 4 and Homework 6. Consider the data set from Problem 6.18, p. 251 about "Commercial Properties" (CH06PR18.DAT) which describes a data set (n = 81) used to evaluate the relation because rental rates (Y) and the age of the unit (X_1), operating expenses and taxes (X_2), vacancy rates (X_3), total square footage (X_4).

1. (2.5 pts.) According to the scatter plot (HW5 1c) and results from `proc corr` (HW 1e), it looks like we might have a correlation problem in this system.
   (a) Refer to your answer (or the one in the answer key) in HW 5:1e and repeat it here. [Note: this is not worth any points, but it is useful to have it available.]
   (b) Run the regression to determine if any of the explanatory variables can be explained by the other explanatory variables. Summarize the results by giving the percent variation explained (R^2) by each of the following models. (Please do not include the SAS output for all of these models. Only the R^2 value is needed. Note that you can run `proc reg` with multiple model statements to save typing.)
      i. Predict age using operating expenses, vacancy and square footage.
      ii. Predict operating expenses using age, vacancy and square footage.
      iii. Predict vacancy using age, operating expenses and square footage.
      iv. Predict square footage using age, operating expenses and vacancy.
   (c) Do you think this multicollinearity is a problem in this example?
2. (5.5 pts.) From the scatterplot of age of the property (X_1) versus rental rate (Y) (HW5 1c), there seem to be some curvature.
   (a) To determine if there is a quadratic relationship between age and rental rate, we want to fit a polynomial regression with age (X_1), the age squared (X_1^2), operating expenses and taxes (X_2) and total square footage (X_4) and the three 2nd order interaction terms. Run the regression to determine if there are any obvious problems with the p-values. In addition, run `proc corr` between the explanatory variables. Should we center the age? Why or not?
   (b) If centering is required, center the variables and repeat part a) to be sure that this removes the problem. From this information, do you think that we should include any 2nd order terms? Why or why not?
   (c) Run the regression analysis and perform a general linear test to determine if the 2nd order terms should be included in the model. Provide the null and alternative hypotheses, the test statistic (with degrees of freedom), the p value, decision and a conclusion in words.

3. (6.5 pts.) This problem is related to the Assessed valuations, KNNL Problem 8.24 (CH08PR24.DAT). The response variable (Y) is the selling price for the one-family home. X_1 is the assess valuation of the home and $X_2 = 1$ if the home is on a corner lot and 0 if not.

   (a) Plot the data for the two populations on the same graph using different symbols (v =) and lines. Does the relationship between the selling price and the assessed valuation appear to be the same whether the house is on a corner or not?

   (b) Run the multiple linear regression with the two explanatory variables and the interaction term.

      i. Give the equation of the fitted regression line using all three explanatory variables.

      ii. Give the equation of the fitted regression line for houses not a corner lot.

      iii. Give the equation of the fitted regression line for house on a corner lot.

   (c) Examine the question of whether or not the two lines are the same. Perform a general linear test to determine if the two lines are equal. State the null and alternative hypothesis, the test statistic with degrees of freedom, the p-value and your conclusion in words.

   (d) Examine the question whether the slopes are the same or not. Provide a 95% confidence interval for the difference in the slopes. (Hint: what parameter represents the difference between the slopes?).

4. (1.5 pts.) This problem is a continuation of the Commercial Properties question from Homeworks 4 and 5 and above (Problem 6.18, CH06PR18.DAT).

   Using the first-order, second-order and interaction terms for the three good predictor variables (please remove vacancy, X_3), all centered around the mean, determine the three best regression models using the C_p criterion. Best means the C_p values that are closest to the number of predictors; however, if a C_p values is much smaller relative to the p, that should also be considered. Summarize your results (include the explanatory variables, and values of R^2, adjusted R^2 and C_p). Do the other criterion (R^2 and adjusted R^2) determine the same models?