Homework 5 (30 pts.)
due Feb. 7

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 problems are a continuation of Problem 3 in HW 4. 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. (21 pts.) This problem relates to material described in Chapter 6 about multiple regression.
   (a) Run the multiple linear regression with age, operating expenses, vacancy rates and total square footage as the explanatory variables and rental rate as the response variable. From the data from the last problem set (please include the relevant output again), please continue to summarize regression results by giving the estimated value of σ^2, the R^2 value and the adjusted R^2 value.

   (b) Now, perform the significance test to determine if there is some sort of association between at least one of the explanatory variables and the response variable (that is, the null hypothesis is that all four of the coefficients for the explanatory variables are zero). Please provide the null and alternative hypotheses, the test statistic with the degrees of freedom, p-value, decision and the conclusion in words.

   (c) Interactively generate a scatter plot for the response variable with the four explanatory variables. From this information, do you think that a linear model is appropriate? Do you think that there are any correlations between the explanatory variables?

   (d) Is the response variable pairwise correlated with any of the explanatory variables (using proc corr)? From this information, do you think that a linear model is appropriate?

   (e) Are any of the explanatory variables correlated with each other (using proc corr)? Do you have any concerns about this information? (We will address this issue in question 4 below.)

   (f) Give the 95% confidence intervals (do not use the Bonferroni correction) for the regression coefficients of age, operating expenses, vacancy rates and total square footage based on the multiple regression. Describe the results of the hypotheses tests for the individual regression coefficients. Please provide the null and alternative hypotheses, the test statistic with the degrees of freedom, p-value and the decision for all of them. Only provide the conclusion in words for age and vacancy. What is the relationship between these results and the confidence intervals?

   (g) Plot the residuals versus predicted rental rate and each of the explanatory variables (i.e., 4 residual plots for the explanatory variables and one for the predicted response variable). Are there any unusual patterns? Be sure to comment on each plot.

   (h) Examine the assumption of normality for the residuals using a qqplot and a histogram. State your conclusions.

(i) Predict the rental rate for a property with age equal to 15, operating expenses and taxes equal to 9.2, vacancy rate equal to 0.03 and total square footage equal to 90,000. Provide a 95% prediction interval in addition to the point estimate. Interpret your results.

2. (6 pts.) The following 2 problems are exploring some of the ideas described in Chapter 7 relating to extra sums of squares.

   (a) Run the following two regressions:

      i. Predict rental rate using the age , expense and square footage.
      ii. Predict rental rate using age, expense, vacancy and square footage.

      Calculate the extra sum of squares for the comparison of the two analysis. Use it to construct the F-statistic (the general linear test statistic) for testing the null hypothesis that the coefficient of the vacancy variable is zero in the model of the three predictors. What are the degrees of freedom for this test statistic?

   (b) Use the `test` statement in `proc reg` to obtain the same test statistic. Give the statistic, degrees of freedom, p-value, decision and conclusion in words.

   (c) Compare the test statistics and p-value from the test statement with the individual t-test for the coefficient of the vacancy variable in the full model. Explain the relationship.

   (d) Is it appropriate to remove the variable vacancy from the model? Why or why not?

   (e) Is it possible to remove both vacancy and the expenses from the model? Why or why not? (Hint: perform the appropriate `test`.)

3. (3 pts.) Run the regression to predict rental rate using age, operating expenses, vacancy rate and square footage. Put the variables in the order given above in the model statement. Calculate both the Type I SS (SS1) and the Type II SS (SS2).

   (a) Add the Type I sums of squares for the 4 predictor variables. Do the same for the Type II sums of squares. Do either of these sum to the model sum of squares? Are there any predictors for which the two sums of squares (Type I and Type II) are the same? Explain why.

   (b) Verify (by running additional regressions and doing some arithmetic with the results) that the Type I sum of squares for the variable operating expenses is the difference in the model sum of squares (or error sum of squares) for the following two analyses:

      i. Predict rental rage using age and operating expenses.
      ii. Predict rental rate using age.