Project (100 pts.)
due April 18

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. Don't forget to put all necessary information (see course policies) on the first page including names for each group member. Include the SAS input for all questions at the very end of your project; this could be important even though it won't be graded.

The project is composed of 2 parts. The first part includes the 8 steps below and is worth 90 points. Please turn in one Part 1 per group; this part will be graded by the grader. The second part is to be completed individually and is worth 10 points. Each group member is to rate the performance of the other group members. The grading rubric to use is at the end of this document. Please fill out one for each group member; circling the answer on the form and totaling the points and turn it in directly to me. Each person will get an average of the scores from her/his group mates. If you want to make any additional comments about the group participation of an individual person, please indicate that on the back of the form. Note that your group members will not see how you rate them. Note: If you do this individually, you will automatically lose the 10 points for group participation.

This project is concerning the complete analysis using multiple regression of the Real Estate Sales Data set described in Data Set C.7 (APPENC07.DAT). In brief, a city tax assessor is interested in what factors are affecting residential home sale prices. In this project, no interaction terms will be used (until 8.e).

1. (4 pts.) The project begins by determining if a multiple regression is appropriate. Remember, if there are qualitative variables with more than two options, you will need to make dummy variables before you run the regression. To perform this step, test the regression relation using all of the explanatory variables. State the hypotheses, test statistic and degrees of freedom, the p-value, the decision and the conclusion in words.
2. (7 pts.) The next step is to check the assumptions and look at the original variables. Use all of the "usual plots" (no partial residual plots). It is acceptable to just show the histograms for the explanatory variables. This step does not require any quantitative analysis. (You may use the automated plots generated in SAS.) Be sure to list all of the assumptions and whether they are appropriate or not using the graphs displayed in this step.
3. (4 pts.) The next step is to look at multicollinearity.
   (a) Do the results from the regression indicate that there might be a problem with multicollinearity? Explain your answer. You might need to include additional output from what you displayed in part 1.
   (b) Roughly determine which variables (if any) might cause a problem with multicollinearity by looking at the correlations between the explanatory variables both visually (scatterplot) and quantitatively (proc corr). The scatterplot should be repeated from question 2. Which variables do you think might be causing problems? Explain your answer.
   (c) Is your analysis consistent in parts a) and b)? Explain your answer.

4. (11.5 pts.) Before we run model selection, it is a good idea to make some predictions on which variables might be included in the final model.
   (a) From the regression analysis, which variables do you think might be important in the final model. Explain your answer.
   (b) Which variables do you think might be important by looking at the correlations between the sales price and each of the explanatory variables. Your SAS output should include both a visual representation (scatterplot) and quantitative data (proc corr). Which variables do you think should be included in the final model and which variables do you think might be dropped. Explain your answer.
   (c) Generate partial regression plots (it is extremely useful to have the best fit line on the plot). Which variables do you think should be included in the final model. Explain your answer. Please comment on each of the plots.
   (d) Compare the methods used in parts a), b) and c). Are the results the same? different? Which variables do you think should be included in the final model?
   (e) Does multicollinearity (question 3) affect your answer in part d)? Explain your answer.
5. (9.5 pts.) In this step, we will run the model selection. Remember, the best model has the least number of explanatory variables that can adequately predict the response variable.
   (a) Determine the three best regression models using the Cp criterion. Summarize your results (include the explanatory variables but not their values, and values of R^2, adjusted R^2 and Cp). Explain your answer.
   (b) Determine the best regression method using each of the automatic methods, forward stepwise regression, forward selection and backward elimination. Again just include the explanatory variables for each of them. Are these models the same or different from each other and of the models chosen in part a)?
   (c) Normally, what we would do is to perform the rest of the project on each of the best models and then determine which model is the best at the end. However, to save time, we will choose the best model first and then only perform diagnostics and remedial actions on that one model.
      i. Run a linear regression on each of the best methods from parts a) and b). If necessary, run a manual backwards elimination and repeat the calculation. (p-values of close to 0.05 are still acceptable.) Give the equation of the fitted regression line for your final model. Explain your choice.
      ii. Is your answer in part i) consistent with the results from questions 3 and 4? Do you expect a problem with multicollinearity?
6. (28 pts.) We are now going to check our assumptions on the best model obtained in Step 5. For ease in grading, I am splitting your answer into the type of information that you are checking. Each assumption will be addressed in at least one (normally more than one) of the parts. Please comment on all graphs shown and write at least one sentence of conclusion for each part. We have already looked at the original variables and the scatterplots between them so you do not need to repeat that here.
   (a) residual plots.
   (b) normality plots.
   (c) partial regression plots.
   (d) qualitative outlier and influential points (Note: this is considered a large data set). Be sure to include your results from looking at the interactive scatterplot. Do not consider non-outliers as being influential.

(e) multicollinearity including visual (scatter plots), pairwise correlation and quantitative information.

(f) Please summarize your answer by listing each of the assumptions (and multicollinearity) and whether there is a problem or not.

7. (12.5 pts.) The next step is to perform remedial actions. Please do not correct for influential points.

   (a) Only perform one of the following possible actions. Please explain your choice and provide the results of your action.

      i. if there is a problem in normality/constant variance and linearity, perform a Y transformation.

      ii. if there is a problem in just constant variance, perform a weighted regression.

      iii. If there is a problem with multicollinearity, perform a ridge regression.

      iv. if there is a problem with linearity that is due to a curve in one of the explanatory variables, center the variable and include the quadratic term in the regression.

   (b) Did your remedial action, correct the problem? Explain your answer by displaying the appropriate plots or other SAS output.

   (c) Regenerate any of the other parts (a) – (e) in question 6 that are required to be sure that all of the assumptions are met. Again, comments are necessary on all plots. If any assumptions are still not met, what would you do next to correct the problem? Please explain your answer. YOU DO NOT NEED TO PERFORM ANY MORE REMEDIAL ACTIONS!

8. (13.5 pts.) Finally, we need to summarize our results using the specifics of the variables.

   (a) Give the equation of the fitted regression using the selected explanatory variables.

   (b) If there is a indicator variable left in the model (Air conditioning, Pool, Quality, Style, Adjacent to highway), give the equation with and without that variable and using the estimated regression coefficient explain how the variable predicts the final sales price. Does this make sense? Explain your answer.

   (c) For the continuous variables, explain how each variable predicts the final sales price. Does this make sense? Explain your answer.

   (d) To me (having recently bought a house), all of the predictor variables are important in the sales price. Can you think of a reason why not all of the predictors are in the final model. (Possible reasons: multicollinearity, all houses have it, very few houses have it, variety of preferences.) There should be one statement per predictor variable NOT in the final model.

   (e) Though we are not including any interaction terms in this model, can you think any of the possible interaction terms might have been important in this model? Why or why not? (You just need to include one possibility.) Rerun your chosen model given in part a) including your chosen interaction term (remember that both of the first order terms need to be included if they are not already there). Is this interaction term important? Please comment.

Name of Student Being Evaluated _____ **Evaluator's Name** _____

Total Score _____ **Evaluation Date** _____

	1 (0 Points)	**2** (0.5 Points)	**3** (1.5 Points - *unless otherwise indicated*)	**4** (2 Points - *unless otherwise indicated*)
Contribution to Group's Tasks *(2 Points)*	• Chooses not to participate • Shows no concern for goals	• Participates inconsistently in group • Shows sporadic concern for goals	• Participates in group most of the time • Shows concern for goals most of the time	• Participates actively • Shows concern about goals actively
Completion of Personal Tasks *(3 Points)*	• Impedes goal setting process • Impedes group from meeting goals • Does not complete assigned tasks	• Participates sporadically in goal setting • Participates sometimes in meeting goals • Completes assigned tasks some of the time.	• Participates in goal setting most of the time • Participates in meeting goals most of the time • Completes assigned tasks the majority of the time ***(2 pts.)***	• Helps direct the group in setting goals • Helps direct group in meeting goals • Thoroughly completes assigned tasks ***(3 pts.)***
Discussion Skills *(2 Points)*	• Discourages sharing • Does not participate in group discussions	• Shares ideas occasionally when encouraged • Allows sharing by most group members	• Shares ideas most of the time • Sometimes encourages groups	• Shares many ideas related to the goals • Encourages all group members to share their ideas
Active Listening *(1 Points)*	• Does not listen to others • Not considerate of others' feelings and ideas	• Listens to others sometimes • Considers other people's feelings and ideas sometimes	• Listens and takes other's feelings into consideration most of the time ***(0.7 pts)***	• Listens attentively to others • Empathetic to other people's feelings and ideas ***(1 pt.)***
Problem-solving *(2 Points)*	• Chooses not to participate in problem-solving	• Offers suggestions occasionally to solve problems • Demonstrates effort sometimes to help the group work together	• Offers suggestions to solve problems and sometimes encourages group participation	• Involves the whole group in problem-solving