Assignment #3: Multiple Regression Model (100 points)

Data Directory: Data can be accessed on the SAS OnDemand server using this libname statement.

```
libname mydata        '/courses/u_northwestern.edu1/i_833463/c_3505/SAS_Data/' access=readonly;
```

Data Set: mydata.building_prices

Data Description: See the data dictionary or pp. 328-329 of *Regression Analysis By Example*.

Assignment Instructions:

Part 1: For this assignment we will fit a multiple regression (a regression model with one or more predictor variables) to the building_prices data set. First, fit the simple linear regression model from Assignment #2 to use as a reference model. Second, find the optimal regression models using the automated variable selection procedures using the *selection* option in PROC REG equal to *forward*, *backward*, and *stepwise* using the default values for *SLENTRY* and *SLSTAY* (see Chapter 9 in *SAS Statistics By Example*). Did each procedure select the same model? Why could each procedure select a different model? Are these models more predictive than the simple linear regression model that you obtained from using *selection=rsquare* in Assignment #2? Which metric should we use to compare the models? (Hint: it is not R-squared.)

Part 2: For the regression model that you choose as the optimal multiple regression model in Part 1, fit the model and set the parameters in the PROC REG statement to produce the default diagnostic plots and the Variance Inflation Factors for the model. Perform an assessment of the model adequacy by commenting on the diagnostic and residual plots from the *plots = (diagnostics residualplot)* in PROC REG (see Section 9.11 in *The Little SAS Book (5th Edition)* and Chapter 9 in *SAS Statistics By Example*). (Note that the *fitplot* is not a valid *plots* option for a multiple regression model.) Note that the default SAS check of model adequacy includes: (1) plot the fitted regression model over the scatterplot, (2) an assessment of the normality of the residuals using a Quantile-Quantile plot (QQ plot), and (3) an assessment of the specification of the predictor variable by plotting the predictor variable against the residuals, and (4) a check for potential outliers using Cook's Distance. Comment on each of these diagnostics of model adequacy. In addition, output the Variance Inflation Factors using the VIF option in PROC REG and discuss any multicollinearity issues as step (5) of the assessment of model adequacy.

Part 3: Consider a model that contains only taxes in thousands of dollars (X1) and number of bathrooms (X2). Fit this model. Do you notice anything peculiar about X2? Should we treat X2 as a continuous or discrete predictor variable? Now code X2 into a dummy variable (see pp. 153-155 in *SAS Statistics By Example*) using an *if-then/else* statement (see pp. 82-85 in *The Little SAS Book (5th Edition)*) as follows.

```
data temp;
      set mydata.building_prices;
      if (X2=1.5) then bath_dummy=1;
      else bath_dummy=0;
run;
```

Which model fits better, the model that treats X2 as a continuous predictor variable or the model that treats X2 through the use of a dummy variable? Could treating X2 as a continuous predictor variable violate any of the OLS model assumptions? Fit a model with only X2 and examine the diagnostic plots produced by SAS to see if this simple linear regression model violates any of the OLS assumptions.

Assignment Document:

All assignment reports should conform to the standards and style of the report template provided to you. Results should be presented and discussed in an organized manner with the discussion in close proximity of the results. The report should not contain unnecessary results or information.

Treat each Part of the assignment as a separate subsection in your report. Part 1 will contain the parameter estimates of four regression models and the summary table from each of the variable selection procedures. Part 2 will contain the parameter estimates for one regression model and a full discussion of the model adequacy based on the results of the diagnostic plots. Part 3 will contain the parameter estimates and diagnostic plots for two regression models. Each section should also contain a discussion of any questions asked in this assignment. The document should be submitted in pdf format.