CHLH 573 – \_Final Project

(300 points)

Due: 11:59 PM, May 5th

**Answer the following questions, justifying with the appropriate analyses. Do not just answer yes or no, you must justify your response and provide data to back it up.**

1. **Property Valuation: Scientific mass appraisal uses linear regression methods to assess property valuation. Twenty-four observations were obtained from a property listing for Erie PA. The problem is to find the best fitting regression model to predict sales price (Y) using the following independent variables: taxes in $1000s of dollars (X1), number of bathrooms (X2), lot size (X3), living space (X4), number of garage stalls (X5), number of rooms (X6), number of bedrooms (X7), age of the home in years (X8) and number of fireplaces (X9). The data have been uploaded to compass (FPQ1.dta).** **You can treat all predictors as continuous and no interactions will be considered**.
   1. A veteran real estate agent (but new to statistics) suggested to include all variables collected as predictors (X1-X9) in the regression model to predict sales price. A former student in CHLH 573 carefully reviewed the list of predictors (X1-X9) and shared some concern right away of including them all **without even accessing the data (the concern is not whether all predictors included are significant)**. What is the concern? Now use the data to see whether the concern is valid. (6%)
   2. Another agent has suggested that a model with taxes, the number of rooms, and the age of the house should adequately describe sales price. Do you agree? (6%)
   3. Another suggestion was that the selling price is determined by its desirability which is a function of the physical characteristics (X2-X9) of the building. The physical characteristics of the building are reflected in the local taxes paid on the building; thus **the best** predictor of sales price is local taxes. Do you agree? (6%)
   4. Because of the suggestion in c), it was also then suggested that the building characteristics (X2-X9) in an equation with local taxes would be redundant in describing sales price. Do you agree? (6%)
   5. Based on your answers in a) to d) present what you consider to be the most adequate model or models for predicting the sales price of homes in Erie PA. (6%)

**Now suppose we are only allowed to use physical characteristics of the building as predictors (X2- X9). Answer the following questions:**

* 1. Run stepwise regression with Pr=0.3 and Pe=0.25. (7%)
  2. Run forward regression with Pe=0.25. (7%)
  3. To make all regression approach practical, we first need to narrow down our predictor pool by running univariate regressions for each predictor respectively (e.g. Y~X2, Y~X3, etc.): We identify and keep the top three most significant predictors while removing the others. Then all regression approach can be implemented to find the best model containing all or some of the three predictors identified. Implement this procedure and identify the best model. (10%)
  4. Based on your answers in f) to h), present the best model and justify your choice. Use it to answer following questions: (6%)
  5. Interpret slope/slopes of the model in i) (6%)
  6. Check assumptions (normality, linearity and homogeneity) of the model in i) using appropriate graphical tools and/or statistical tests. (8%)
  7. Identify any potential high-leverage points, outliers, influential points of the model in i). (8%)
  8. Summarize your findings from a) to l) and comment on how you have helped to build the appraisal tool with and without taxes information. (4%)

1. **Data from a British government survey of household spending were collected to examine the relationship between household spending on tobacco products and alcoholic beverages. The data are average weekly alcohol spending (y), average weekly tobacco spending (x) and area in Great Britain (z), where area is Northern Ireland when z=1 and area is not Northern Ireland when z=0. The data have been uploaded to compass (FPQ2.dta).** 
   1. Fit a model to predict y using x, z and their interactions as predictors. Write down the model explicitly (You do not need to replace betas with estimates in this step). (4%) b) Write out the prediction formulas of y in Northern Ireland and not Northern Ireland. (4%)
   2. Write out the prediction formulas of y in Northern Ireland and not Northern Ireland. (4%)
   3. Test whether the interaction term is significant given x and z already in the model. State the null and alternative hypothesis in terms of beta (use the same notation as in a)). What can you conclude about the relationship of y to x when z equals to 1 versus z not equals to 1? (6%)