Assignment #3

Due: November 17, on-line

This assignment uses the **Boston housing data** from in Module 3. The data is described further in that module; here is a brief description of the variables.

Value	The response: median value of owner-occupied homes in USD 1000's
NOx	Nitric oxides concentration (parts per 10 million)
Distance	Weighted distances to five Boston employment centres
Lower class	Percentage of lower status of the population
Pupil/teacher	Pupil-teacher ratio by town
Zoning	Proportion of residential land zoned for lots over 25,000 sq.ft.
Crime rate	Per capita crime rate by tract
Industrial	Proportion of non-retail business acres per town
Rooms	Average number of rooms per dwelling
Age	Proportion of owner-occupied units built prior to 1940
Highway	Index of accessibility to radial highways
Tax Rate	Full-value property-tax rate per USD 10,000
Minority	$1000(B - 0.63)^2$ where B is the proportion of blacks by tract
Charles River	Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)

These data are from the 1970's, so home prices are much lower than today. Use all 500 cases. The first row in the data is there for prediction and omits the response, so your model will have $n = 500$. The models that illustrate interactions in Module 3 are flawed and not calibrated. Here's your chance to do better! The questions include several check-points to get you in the habit of including important steps when building a model.

1. As a first step, inspect histograms of each variable (*Value* through *Charles River*). Does the marginal distribution of any variable indicate issues for building a regression model? (Do this briefly, with no more than a sentence for each variable. Don't show all the histograms, just the interesting ones.)
2. With so many observations relative to the number of explanatory variables ($n = 500$, $k = 13$), it is useful to begin with a regression that includes all 13 explanatory variables, *NOx* through *Charles River*. Do the diagnostic plots for this model appear satisfactory? (Use the script "13-predictor Model" to fit this model. Show *only* plots that are relevant to your answer.
3. Consider a tract for which the fitted 13-variable model predicts the value of housing to be $25,000. Do you think this prediction is reasonable? Explain your answer briefly. If not $25,000, indicate your predicted value.
4. Build a 95% prediction interval based on the predicted value in Q3 as
   (prediction from Q3) ± 2 (appropriate scale)
   Do you think this will be a 95% prediction interval?
5. Add the following 4 interactions to the 13-predictor regression (be sure that centering is turned on in the dialog that builds the model)
   Rooms with itself, *NOx* with itself, and *Lower class* with itself
   Charles River with *Distance*
   Does the addition of this collection of interactions produce a statistically significant improvement to the fit of the 13-predictor model? (We will soon see how to discover interactions like these.)
6. What's the effect of outliers on slopes in the model? Answer this question based on inspecting the relevant plots, *without* removing any cases.
7. Interpret the significance and implications of the interaction of *Rooms* with itself.
   (a) Is this effect statistically significant?
   (b) Interpret the effect of the interaction.
8. Interpret the significance and implications of the interaction between *Charles River* and *Distance*.
   (a) Is this effect statistically significant?
   (b) Interpret the effect of the interaction.

9. Does the model fit in Q5 require some improvement? If so, do it. (Don't get carried away. Don't try looking for transformations like logs. You'll spend too much time. As for adding things, the model leaves out a big interaction described in Module 3.)
10. Is your final model calibrated?