Stat 230, Spring 2016
Homework 10: Model Selection, Cross Validation

Due Thursday 4/14/16 at 11:59pm on bcourses.
The file HW10.rda has a data frame named `data`. The first column is Y, the other 20 are variables named a-t. Read the whole lab before you get started, you'll be able to reuse earlier functions for later parts if you write them carefully.

1. Full model, R^2

   (a) Use OLS to fit the full model using all 20 variables for X.

   (b) Compute R^2 as decribed on page 51 of Freedman. (I don't care if you do it using fitted values or residuals.) Note that $R^2 = 1 - \frac{RSS}{\sum {Y_i}^2}$ so having a large R^2 i s equivalent to small RSS (residual sum of squares).

   (c) Use 10-fold cross validation to get a cross-validated R^2. In other words, do 10 OLS fits based on 90% of the data (training sets), and for each one use the coefficients generated to get fitted values for the 10% of the data left out (test sets). Once you do all 10 OLS fits, each point will be left out exactly once, and you'll get a fitted value for each point. Based on these fitted values, compute R^2. How do these two values compare to the Multiple R^2 and Adjusted R^2 from the summary of the results from lm()? If you want to read about what the Adjusted R^2 is really doing,
   `http://en.wikipedia.org/wiki/Coefficient_of_determination#Adjusted_R2`
   has a reasonably short explanation.

2. Model Selection (backward) The process below will generate a sequence of nested models, each is a reasonable candidate for best model of a given size. Backwards and forwards model selection are also described on pages 56-57 of HTF.

   (a) Leave out the variable with the smallest t statistic (in absolute value). Don't take out the intercept even if it has the smallest t value.

   (b) Rather than doing this based on R^2, we'll do the rest of the lab in terms of the training error and test error as described in HTF pages 220-221. For this lab, just get training error based on RSS for best model of each size. Get test error by using cross validation. This is similar but not exactly the same as what was done in the figure in HTF on page 220. You should be able to adjust your function for R^2 slightly to output test error $\mathrm{Err}_{\mathcal{T}}$ and training error $\overline{\mathrm{err}}$ for each model.

   (c) Repeat steps a) and b) until you are left with just the intercept term. Note that the next variable left out should be the one with the smallest t value based on the most recent fit, not based on the smallest remaining t value from the original fit. Record the order in which the variables are removed as well as the values of both prediction errors, you'll use them later.

3. Model Selection (forward) The process will generate a sequence of nested models, each is a reasonable candidate for best model of a given size.

   (a) Do the same as in backward selection, but now start with the intercept and add the single variable that improves the fit the most (based on smallest RSS). at every step. (So at the kth step you'll need to do 20-k+1 regressions to figure out which added variable will help the most.) Add one variable at a time until they are all included, and record the values of both prediction error values at every step.

   (b) Did forward selection and backward selection give the same sequence of models? Will that always be true? Explain briefly.

4. Plot the results
   Make a plot that has four lines with points (use `type="b"`). The horizontal axis will be for p. The vertical axis will be for the prediction error value. Use colors `col=` and line type `lty=` to distinguish the two different kinds of prediction error and the two methods of model selection. You might want to specify the colors using `rgb()` and the transparency argument `alpha=` or use `adjustcolor()`. Include appropriate colored vertical lines for the optimum model based on smallest prediction error for each of the 4 ways of doing this (backwards and forwards, two choices of prediction error for each). Comment briefly.