Stat 230, Spring 2013
Homework 10: Cross Validation
Due Wednesday 4/17/13 at 11:55pm on bspace.

The file HW10.rda has a data frame named data. The first column is Y, the other twenty are variables named a-t.

1. Use OLS to fit the full model using all 20 variables for X.

2. Compute R^2 as described on page 51 of the text. Then use 10-fold cross validation to compute R^2. This will involve 10 OLS fits, each based on 90% of the data, and for each fit you'll use the coefficients to get residuals for the points not used to fit the data. With these residuals, you can get R^2, then take the average of the 10 R^2 values to get the cross validation R^2. How do these values compare to the Multiple R-squared and Adjusted R-squared given by the summary function?

3. Model Selection

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

   (b) Get both "regular" R^2 and cross validation R^2 as in step 2.

   (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 remainding t value from the original fit.

4. Plot the number of variables on the horizontal axis and for each number of variables, both the R^2 values, color coded and with lines connecting the points. How many variables are used for the smallest R^2?