Statistical Computing (140.776) Homework 3

Due Tuesday, Oct 18

1. Debug

Download debugtest.R from the course website. The program tries to compute a number *zsum*. When you run the program, you may encounter some problem. Use debug functions to find out where the problem is. Fix the problem and report the value of *zsum*.

2. Merge sort

Write an R function that implements the merge sort algorithm. Your function should be able to sort a numeric vector ***x*** of any length in ascending order. The function should have an optional argument *decreasing*. By default, *decreasing*=FALSE. If *decreasing*=TRUE, then the function will return a sorted vector in descending order. You are not allowed to use the sorting functions provided by R.

3. Variable selection

Download varsel.txt from the course website. The data contains 11 columns corresponding to y, $x_1, \ldots, x_{10}$. Among the 10 columns of x, only three are predictors of y. Let x_i, x_j, x_k denote these predictors. The relationship between y and the predictors can be described by a linear regression

$$y = a_0 + a_1 x_i + a_2 x_j + a_3 x_k + e$$

where $e \sim N(0,\sigma^2)$. Unfortunately, you don't know which of the 10 columns are predictors. In order to find the three predictors of y, you want to check all possible (i, j, k) combinations and find the optimal (i, j, k) that minimizes the following objective function:

$$g(x_i, x_j, x_k) = n \log (\text{MSE}) + p \log (n)$$

Here n is the sample size (i.e. number of rows in the table), p is the number of regression coefficients (i.e. 4 in this case), and MSE is the residual mean square (i.e., [sum of squared residuals / degrees of freedom], which is the estimate of σ^2).

(1) Write a program to find the optimal (i, j, k).

(2) Using the predictors you found, fit the linear regression and report regression results.