Department of Statistics

COURSE STATS 762

Assignment 6, 2007

Instructions: Hand in your completed assignment to me by 4pm on Thursday 18 October.

In this assignment you are requited to investigate the performance of the stepwise regression algorithm. You will be using the car data first encountered in Tutorial 3 and also used in Tutorial 4.

In Tutorial 3, we fitted a model of the form

1/CITY ~ PRICE + WEIGHT + DISP + COMP + HP + TORQ + TRANS + CYL

to the data with point 47 omitted (i.e. to a data set with 137 observations). In tutorial 5, we selected a submodel using stepwise regression. The selected submodel is

1/CITY ~ PRICE + WEIGHT + DISP + HP

In this assignment, we will assume that this 4-variable model is in fact the true model.

1. Fit the 4-variable model to the 137 observations, and calculate the fitted values and the estimated error standard deviation. Show the R code, the fitted values and the value of the standard deviation. [10 marks]

2. Assuming that this is the true model (i.e. that the estimated regression coefficients are in fact the true coefficients, and that the errors are normally distributed with error variance equal to the estimated variance), generate a new set of responses by generating a vector of 137 random normal values, and adding them to the vector of fitted values from Question 1. Select a subset of variables using stepwise regression. Does the selected model match the true model? (which we know in this case as we are simulating). Show the R code. [10 marks]

3. Repeat question 2 1000 times. Make a table of how many times different models are selected. How often is the true model selected? Which model is selected most often? What does this tell you about stepwise regression? Show the R code used.
   [20 marks]