PSTAT 126: Quiz #3 Solutions

Question 1: Consider the "mtcars" dataset "built-in" with R. With disp as the response variable and with the variables mpg, hp, and wt as predictors in a Stepwise Search, use standard Stepwise Regression in R, along with the AIC criterion, to identify a regression model in this setting with relatively high goodness-of-fit but for which model complexity is controlled. What model did the stepwise regression search procedure generate? That is, which of the three predictors should be included in the model and which excluded? Include (cut and paste into the answer field here) the algorithm output results describing the final Step iteration of the algorithm with your answer. (If for some reason you experience any difficulties cutting and pasting these algorithm output results into the answer field here, then instead please include as part of your answer here -- simply manually type into the answer field here -- the four numbers under the column heading "AIC" in the final Step iteration of the algorithm results. Please do not attach any files or screen shots here and also do not email them separately either. Also, do not include any R code with your answer, only code output results are being requested.)

Answer: The variables the Stepwise Regression identified for inclusion into an enhanced model according to the AIC criterion were wt and hp, and mpg was excluded. The final step iteration in the output results was

Step: AIC=249.73
disp ~ wt + hp

	Df	Sum of Sq	RSS	AIC
<none>			65012	249.73
+ mpg	1	19	64993	251.72
- hp	1	35697	100709	261.74
- wt	1	113271	178284	280.01

(More details and the relevant R code for producing these results can be found on course slides 93-94.)

Question 2: Recall the Akaike information Criterion (AIC) as well as the related Bayesian Information Criterion (BIC) from our study of Variable Selection and Model Building. Without heavy mathematical notation, explain how both the AIC and BIC emphasize goodness-of-fit of a regression model while also controlling model complexity. Why would we wish to do this (that is, to promote goodness-of-it while at the same time controlling model complexity)? Why is model complexity so undesirable?

Answer: Clearly the first term of both the AIC and BIC involving the logarithm of the sum of squared differences of the y-data samples and their associated fitted values is smaller the better the fit of the model (represented by the fitted values) is to the data samples. Moreover, one very simple measure of model complexity is the number predictor variables included in the model, and this is incorporated into the AIC or BIC measure in the term on the right-hand side of the definition of both the AIC and BIC.

Model complexity is undesirable because, for example, it can lead to overfitting (also called overdetermination), which can cause poor generalization/prediction performance of the model for new x-values (that is, new predictor variable samples) not used in the actual computation of the least-squares regression model coefficients (that is, not used in the regression model

training). Moreover, all models should as a rule be as simple as possible for a given level of performance. Excess model complexity without corresponding improvement in model performance leads to longer computational process times without any resulting benefit and can also make it harder to interpret model results.

Question 3: Does Stepwise Regression using, say, the AIC criterion necessarily always find an optimal choice of predictor variables -- optimal (lowest) according to the AIC -- among a given set of candidate predictor variables? Why or why not?

Answer: The answer is that it does NOT necessarily find an optimal choice. It is not in general an exhaustive search. It adds and subtracts predictor variables one at a time and does not check all possible combinations of the predictors to find the optimal one.

Question 4: Consider the “mtcars” dataset “built-in” with R. Perform standard Stepwise Regression in R using AIC as the search criterion, starting with the model mpg ~ 1 and searching up to mpg ~ wt + drat + disp + qsec. What is the subset of predictors (from among wt, drat, disp, and qsec) that the algorithm came up with? Now do the same thing using the BIC criterion instead. Which subset of these four predictors do you get now? For both of these two sub-questions in this question (Question 4) you only need to give the corresponding subset of predictors as your answer. You do not need to include any code, code output, or explanation with your answer for this question.

Answer:
Stepwise Regression with AIC gives: wt, qsec, drat.
Stepwise Regression with BIC gives: wt, qsec.

(The relevant R code to produce these results can be found on course slides 93-94.)