Homework #4, ST518

Instructions: use statistical software to address the following questions.

1. Consider the built-in `mtcars` dataset in R (available on moodle as "mtcars.csv").
   (a) Obtain scatterplots of all pairs of variables involving `mpg`.
   (b) Obtain sample correlation coefficients for all pairs of variables involving `mpg`.
   (c) Obtain Mallows C_p and AIC for all subsets of the variables in additive multiple linear regressions: `cyl disp hp drat wt qsec vs am gear carb` Report the 10 models with the lowest values of C_p.
   (d) Which are the two best subset models that contain three variables? Report the error mean square and multiple coefficient of determination from these two models. (Note that they don't have any predictor variables in common!)
   (e) Add `disp` and its square, `disp*disp` to the model that is selected for having the lowest AIC in part (d). Use an F-test to compare the models with and without the pair of variables, `disp` and `disp*disp`. Write out the hypothesis and draw a conclusion using level of significance $\alpha = .05$.
   (f) Report $MS(E)$ and r^2 from the model selected in part (e) and compare them to those obtained from the model in part (d). Which model is better? In what sense(s)?
   (g) Use the partial p-values to retest each regression coefficient in the model using the liberal significance level of $\alpha = .10$. If a regression coefficient does not meet this threshold, drop it from the model. Call this the final model.
   (h) Using the final model, ...
      i. Plot the residuals versus the fitted values.
      ii. Obtain a qqnormal plot of the sorted residuals.
      iii. Plot the observed versus the fitted values. Report the observed correlation between the observed and fitted values. What is the square of this correlation called?
   (i) Construct a table with three rows, one for each of the two models identified in part (d) and one for the final model. Report the $MS(E)$ and r^2 for each.

2. Suppose a researcher from horticulture tells you he randomized 14 seedlings to two different fertilizer treatments ($n = 7$ each) and measured the heights of the plants after one week with the following results:

Treatment	mean	std. dev.
A	7.5	0.6
B	7.3	0.7

   (a) Conduct an equal-variances independent samples t-test for the effect of the fertilizer treatment. Is the observed difference of 0.2 significant from 0, using level of significance $\alpha = 0.05$?
   (b) Suppose he forgot to tell you that this was really a CRD with $t = 3$ treatments and a total of 21 plants. The mean for the last treatment was $\overline{y}_{3.} = 6.7$ with a standard deviation of $s_3 = 0.4$. Carry out an F-test for a treatment effect. Draw a conclusion using $\alpha = .05$. Do the three observed means differ significantly?