ST697R: Fall 2012 Homework 9. Due right at start of class Friday, Dec. 7.

This focuses mostly on applications/computing for variable selection and additional diagnostics. Be sure to also understand the basic principles involved.

1. Use the SMSA data. Eliminate noxpot and hcpot.

   (a) First choose models for fitting mortality using i) adjusted R^2, ii) Mallows C_p iii) AIC. For each criterion give what you think are the three best models (Note that here I'm asking about the three best overall).

   (b) Run forward, backward and stepwise selection. Give the final model in each case.

   (c) Summarize your findings.

2. Continue to use the SMSA data and consider fitting mortality as a function of all of the variables (without noxpot and hcpot as before).

   (a) Get the studentized residuals, the leverage values and the Cook's distances.

   (b) Plot the studentized residuals versus the regular residuals. Does it look like it matters which we use in our plots used to assess the model and variance assumption?

   (c) Do there appear to be any outliers in the X space? Use the rule on page 131 of the notes.

   (d) Using Cook's distance are there any influential observations. Use the rule on page 132 of notes. Note in the notes I referred to an observation being an outlier if Cook's distance was greater than $F(.5, p, n-p)$ but better to say that case was influential.

   (e) Using the studentized residual do any cases look like outliers? You can use the rule on page 132 first given in terms of studentized deleted residuals but apply it to the studentized residuals.

   Note that these diagnostics above rely on assuming the model is right, including having constant variance. If constant variance is not reasonable then some weighting should be done beforehand. (I'm not asking you to do that).

3. Consider the sex discrimination data and models we described in class with outcome being wage and predictors age, education and sex (qualitative = 0 or 1).

   (a) For sex = 0 and sex = 1, what does the model look like if we fit"

   - wage on age, educ, sex, educ*sex and age*sex
   - wage on age, educ, sex, educ*sex, age*sex and age*educ*sex

   (b) If we considered separate models with an intercept, linear and a product term for each sex, then we would have a total of eight different parameters. Why? Do any of the models above fit this? If not, write a model that would.