ST505: Fall 2012 Homework 8.

Problem 1 is due Friday, Nov. 30 at beginning of class.
See notes on Problem 2.

1. This is the second suggested problem that I had on hwk. 7, with a few modest changes.

   (a) .
   This problem uses the setting of problem 8.15. I've posted a data file that combines the data from 1.20 combined with that from 8.15. The data is in a form where the first variable is service time (Y), the second variables is number of copies serviced (X_1) and the third variable is type (X_2) which is 1 if the type of copier is small and is 0 if it is large.

      i. First do problem 8.15 which uses the model in (8.33). Note that X_{2i} is what we defined as Z_{1i} in the notes.
      - In doing b), fit the model by regressing Y on X_1 and X_2 in the usual way. Explicitly interpret each of the coefficients that are fitted.
      Note that part c) is asking for an estimate of the difference in $E(Y)$ under small and large copiers at a common X_1

      * For the remainder of the problem we'll use a linear regression model **allowing separate coefficients for each type**. That is: if type = S then $E(Y|X_1) = \beta_{S0} + \beta_{S1}X_1$; if type = L then $E(Y|X_1) = \beta_{L0} + \beta_{L1}X_1$.
      We'll assume the variance is constant throughout for now.

      ii. Define $Z1$ = type and define $Z2 = 1-$ type (so $Z2 = 1$ when type is large and $Z2 = 0$ when type=small). Using these run **ONE!!** proc reg such that it gives you four estimated coefficients in the output and those are estimates of β_{S0}, β_{S1}, β_{L0} and β_{L1}.

      iii. Regress Y on X_1, X_2 and $X_1 * X_2$ including use of the clb option in SAS or use confint in R. This is fitting the model

      $$E(Y|X_1, X_2) = \beta_0 + \beta_1 * X_1 + \beta_2 * X_2 + \beta_3(X_1 * X_2) \tag{1}$$

      - Show the relationship between the β_0, β_1, β_2 and β_3 in (1) and β_{S0}, β_{S1},β_{L0} and β_{L1} defined earlier.
      - Identify how the β's in (1) relate to the differences between intercepts and slopes for the two types.
      - Use the output from the regression run to obtain i) Separate 95% confidence intervals for $\beta_{S0} - \beta_{L0}$ and $\beta_{S1} - \beta_{L1}$ and ii) separate t-tests for $H_0 : \beta_{S0} = \beta_{L0}$ and $H_0 : \beta_{S1} = \beta_{L1}$. State your conclusions.

      iv. Carry out a test that the two regression lines are equal meaning that they have both equal slopes AND equal intercepts. Do this under the assumption of equal variance throughout. Get the p-value associated with the test.

2. This problem involves variable selection.

   ST697R students This will be handed in along with any additional problems (homework 9), due last day of class. Use the SMSA data. Eliminate noxpot and hcpot (see homework 7 solution).

   (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.

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

   (c) Summarize your findings.

ST505 students you will carry out variable selection using your individual IE project data and the results will be part of your IE write. All students should do this piece individually and then compare results since this is something you should know how to do. Note that variable selection may not be essential for the utlimate objective in all of the the projects. But, I want you to understand how it works without you having to work on another data set. So, I'll have you work on your data in some cases dropping some variables for illustration. **But any suggestion for eliminating certain variables below is for this variable selection piece!! not for the whole project.** You should also look at the solution using the SMSA data (which the grad students are working on) when it comes out.

Proceed as described above for the 697R students. Here are some things for the different IE projects.

- MCAS: All quantitative variables, so use them all, with $Y=$ log(proppass/(1-proppass)).
- Baseball: You can do this by ignoring the qualitative variables we kept and use the numeric ones.
- Insurance and discrimination. You can drop the Zip code for doing this.
- Hospital infections. You can use Med school affiliation as is for one of the variables since it is 0/1. For region create three dummy variables (for NE, NC and S) and include them the list of potential variables.
- Sex discrimination. You can have sex in the list of variables as a single dummy variable.
- Wages and gender discrimination with CPS data. This has a number of qualitative variables. For this variable selection piece, consider building for wage using age, education, experience (three quantitative variables) then single numerical dummy variables for south, sex, union and married and two dummy variables for Race (hispanic and white). Ignore the others for this part.