**Statistics 512: Homework#6**

Due March 7, 2014 BEFORE CLASS

**For the following 3 problems use the computer science data that we have been discussing in class. You can get a copy of the data set `csdata.dat` from the class website. The variables are: `id`, a numerical identifier for each student; `GPA`, the grade point average after three semesters; `HSM`; `HSS`; `HSE`; `SATM`; `SATV`, which were all explained in class; and `GENDER`, coded as 1 for men and 2 for women.**

1. In a `data` step, create a new variable `GENDERW` that has values 1 for women and 0 for men (use arithmetic on the original variable `GENDER`). Run a regression to predict `GPA` using the explanatory variables `HSM`, `HSS`, `HSE`, `SATM`, `SATV`, and `GENDERW`. (Do not include any interaction terms.)

   (a) Give the equation of the fitted regression line using all six explanatory variables.

   (b) Give the fitted regression line for women (use part a).

   (c) Give the fitted regression line for men (use part a).

   DO NOT attempt to run `proc reg` on a subset of the data to answer this question.

2. Use the $C_p$ criterion to select the best subset of variables for this problem (i.e. use the options “ `/ selection = cp b;`”) . Use only the original six explanatory variables, not `HS` or `SAT`, and use either `GENDER` or `GENDERW`, not both. Summarize the results and explain your choice of the best model.

3. Check the assumptions of this “best” model using all the usual plots (you know what they are by now). Explain in detail whether or not each assumption appears to be substantially violated.

**A testing laboratory with equipemtn that simulates highway driving studies for two makes $(A, B)$ of a certain type of truck tire the relation between operating cost per mile $(Y)$ and cruising speed $(X_1)$. The observations are given in `CH11PR15.DAT`, where the columns are ordered $(Y_i, X_{i,1}, X_{i,2})$ where $X_{i,2} = 1$ for Make A and $X_{i,2} = 0$ for Make B. An engineer now wishes to decide whether or not the regression of operating cost on cruising speed is the same for the two makes of tires. Assume the error variances for the two makes are the same and that an interaction-effect regression model is appropriate.**

4. Plot the data for the two populations on the same graph, using different symbols (`v=`) and lines. Does the relationship between speed and operating cost appear to be the same for the two makes of tire?

5. Examine the question of whether or not the two lines are the same. Write a model that allows the two makes of tires to have different intercepts and slopes. Then, perform the general linear test to determine whether the two lines are equal. State the null and alternative hypothesis, the test statistic with degrees of freedom, the $p$-value and your conclusion.

6. Using the model that fits two different lines, give a 95% confidence interval for the difference in slopes. (Hint: what parameter represents the difference between the slopes?)