Statistics 512: Homework 6

Divisions 1, 2, 3, and 4: Due Wednesday, April 15, 2015

Division 5: Due Tuesday, April 14, 2015

Important Note – *Every graph or plot you create should have your name printed as a subtitle. Consequently, any graph with no name will result in a* **20% points off** *on that question. Also, please attach your code at the end; any homework with no code provided will result in a* **50% points off** *on the entire assignment, NO EXCEPTIONS.*

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.

Use the data from problem 8.24 on page 339 (white text) or page 345 (blue text). (`CH08PR24.DAT`)

4. Plot the data for the two populations as a symbolic scatter plot on the same graph, using different symbols (`v=`) and lines. Does the relationship between valuation and selling price appear to be the same for the two lot locations?

5. Examine the question of whether or not the two lines are the same. Write a model that allows the two lot locations 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 94% confidence interval for the difference in slopes. (Hint: what parameter represents the difference between the slopes?)