Statistics 135 – Assignment 4
Due: Friday, December 16, 2005

For this assignment, please submit your answers typeset in some package. All calculations are to be done in **SAS**. Please include your **SAS** code as part of your answers.

1. The computer science department at a large university was interested in why a large proportion of their first-year students failed to graduate as computer science majors. An examination of records from the registrar indicated that most of the attrition occurred during the first three semesters. Therefore, they decided to study all first-year students entering their program in a particular year and to follow their progress for the first three semesters.

   Data on 224 students who began study in computer science were collected. The variables collected include the GPA after 3 semesters (`gpa`), sex (`sex` - 1 for men, 2 for women), high school math, science, and English grades (hsm, hss, hse), math and verbal SAT scores (`satm` and `satv`), and their major at the end of three semesters (`major` - 1 for computer science, 2 for engineering and other sciences, and 3 for other). The data for this study is available in the file `csdata.dat` available on the course web site.

   (a) Read the data into **SAS** and print out the data on `gpa`, `satm`, `satv` and `sex` for the first 10 observations. When printing the `sex` variable, use `Male` and `Female` instead of their respective codes 1 and 2. When doing this, just print the 10 required observations. Don't print all of them and cut and paste the 10 you want.

   (b) Generate a contingency table showing the breakdown of the students by `sex` and `major`. Is their any relationship between the two variables or do they appear to be independent.

   (c) Generate a histogram of `gpa`. To any patterns appear in the figure.

   (d) Generate a line printer version of a scatter plot (using `PROC PLOT`) of `gpa` versus `major`. Also generate high resolution scatter plots (using `PROC GPLOT`) of `gpa` versus each of `hsm`, `hss`, and `hse`. Describe the important features, if any, in these plots.

   (e) Using all of the variables (expect `obs`) find the 'best' model by stepwise regression and the 'best' two models by Mallow's C_p.

   (f) For the 'best' model selected by stepwise regression, refit this model and find the fitted `gpa` for a man who stays in computer science with `satm` = 650, `satv` = 540, `hsm` = 8, `hss` = 7, `hse` = 5. Note the your fitted value may not involve all of these predictor levels.

   (g) Again for the 'best' stepwise regression model, generate a normal score plot of the residuals and a scatter plot of the residuals versus the fits. Are there any indications the standard regression assumptions are violated for this model.

   (h) Are any of the predictor variables useful in indicating that a student might have a low gpa after three semesters?