STAT 516 - Spring 2003 - Homework 3

Due: Monday, February 16th

The data set on the web at `http://www.stat.sc.edu/~habing/courses/data/sat.txt` is from the first year that SAT scores were published on a state-by-state basis in the U.S. It was originally published in the *Harvard Educational Review* in 1984, and is also reported in Ramsey and Schafer, 1997. The variables included are:

`sat` = average total SAT score for the state
`takers` = percent of eligible students in the state who took the exam
`income` = the median family income of students in the state who took the exam
`years` = the average number of years that the test-takers had for studies in the core subjects\
`public` = percentage of test takers attending public secondary schools
`expend` = the states expenditures on education in hundreds of dollars per student
`rank` = the median percentile ranking of the test-takers in their high-school class

Perform a multiple regression to predict the average SAT score in the state from the other variables, and answer the following questions. Present copies of the relevant portions of the SAS output, and for each question indicate which portion of the output you used and how you used it.

a) Test whether the other six variables as a group are statistically significant predictors of the states average SAT scores. (Report the p-value and your conclusion)

b) Check the assumptions, and state whether you feel comfortable trusting the results of the regression.

c) Report the p-values for the Type I and Type III tests for takers and interpret the results at an α=0.05 level.

d) What percentage of the variation in state average SAT score do the six variables explain?

e) For which, if any, of the predictor variables is multicollinearity a concern?

f) Which two states' independent variables are most different from those of the other states? Most similar?

g) Which two states have the greatest effect on the estimated regression line? How would you classify their effect?

h) What subset of independent variables gives the "best" model for this problem?

i) How would you respond to someone who claimed that the variables not included in the model in part h did not affect the average SAT score of the states?