STAT 516 - Spring 2001 - Homework 2

Due: Wednesday, February 14, 2001

1) *(3 points)* A continuation of Problem 3, on Homework 1. For both portions of this question, you may assume the assumptions of the regression hold.

a) One of the questions that Galton was interested in was whether or not there was a regression to the mean effect. That is, do the larger parent peas tend to produce offspring that are somewhat smaller than themselves, and do the smaller parent peas tend to produce offspring that are somewhat larger than themselves. This could be examined by testing the null hypothesis H_0: $\beta_1=1$ against the appropriate alternate hypothesis. What is the appropriate alternate hypothesis to test Galton's experimental hypothesis? Using the output contained in the "Parameter Estimates" box on the PROC INSIGHT output, construct this test of hypothesis by hand (use $\alpha=0.05$).

b) Using the simple linear regression model, what is the 95% confidence interval for predicting the average size of the offspring peas of a parent pea plant with peas of diameter 20.5.

2) *(3 points)* Page 108, problem 3.6

3) *(4 points)* The data set on the web 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
`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.)

a) Test whether the other five variables as a group are statistically significant predictors of the states average SAT scores. Check the assumptions, and state whether you feel comfortable trusting the results of the regression. What percentage of the variation in state average SAT score do they explain?

b) If you use `takers` alone to predict SAT score, you get a p-value of less than .0001 and an r^2 of 0.7358. Why does the multiple regression <u>seem</u> to say that `takers` isn't significant?

c) Which two states most "under-achieved" according to the model? "over-achieved"? Were there any states that were large enough outliers to substantially change the regression model?

d) Answer the following question with either true or false, and justify your answer. If we used `PROC GLM` to construct the "95% confidence interval for an observation" for each state, we would expect 95% of the students in that state to have scores in that interval.