Problem 1

The publisher would like to develop a model that will help it in determining the cost of books it publishes. The publisher obtains a sample of 207 books that have been recently published by a publisher. Of the 207 books in the sample, 83 are hardcover and 124 are softcover. Hardcover books are obviously priced at a premium, so some adjustment for this will need to be made. The variables include the cost of producing the book (COST), the number of pages in the book (PAGES), and an indicator variables SOFTCOVER which is equal to 1 if a book is softcover and 0 if a book is hardcover.

(a) Plot the scatterplot of COST versus PAGES, by SOFTCOVER (hardcover and softcover books have to use different plotting symbols). Comment on the scatterplot about the similarities and differences between the two groups, softcover and hardcover books.

(b) Fit an MLR of COST on PAGES and SOFTCOVER. Comment on the overall significance of the parameter estimates and the model.

(c) For the model in part (b), write down the regression equation and interpret the slope parameter estimates associated with PAGES and SOFTCOVER in the context of the problem.

(d) For the model in part (b), check the linear model assumption using lack-of-fit test. Check the constant variance assumption. Also check the normality assumption using Anderson-Darling normality test procedure.

(e) Do you think the model in (b) is adequate to capture the type of relationship between COST and PAGES, as seen from the scatterplot in (a)? Why or why not?

(f) Now create an interaction variable, called PAGES-SOFT by multiplying PAGES ad SOFTCOVER. Fit an MLR of COST on PAGES, SOFTCOVER and PAGES-SOFT. Comment on the overall significance of the parameter estimates and the model. Is there a potential multicollinearity problem? Why or why not?

(g) For the model in part (f), write down the regression equation and interpret the slope parameter estimates associated with PAGES and SOFTCOVER in the context of the problem.

(h) For the model in part (f), check the linear model assumption using lack-of-fit test. Check the constant variance assumption. Also check the normality assumption using Anderson-Darling normality test procedure.

(i) Compare the two models, (b) and (f). Which one will you prefer to use? You have to provide evidence to support your conclusion.

Note that you must include relevant outputs from the designated software package in order to earn full credit.

Problem 2

"LaQuinta.csv" has financial information of La Quinta Inns. The data set consists of a random sample of 100 inns belonging to La Quinta. The response variable of interest is MARGIN, which is the operating margin (in percent), as the sum of profit, depreciation and interest expenses divided by total revenue. In an effort to determine which factors (variables) could impact operating margin, the company also collected a number of variables, such as the total number of hotel and motel rooms available within three miles of each La Quinta Inn (ROOMS), the number of miles to the nearest competitor (NEAREST), the total office space (in 1,000 ft) in surrounding community (OFFICE), college and university enrollment in thousands (COLLEGE), the median household income in $1,000's (INCOME), and the distance to downtown (DISTTWN).

A) Use all possible regressions approach to select the predictor variables that will be included in the regression modeling and analysis. You have to explain why and how you reach your conclusions.
B) Use backward elimination approach, using α=.10, to select the predictor variables that will be included in the regression modeling and analysis. You have to explain why and how you reach your conclusions.
C) Use forward selection approach, using α=.25, to select the predictor variables that will be included in the regression modeling and analysis. You have to explain why and how you reach your conclusions.
D) Use stepwise approach, using "α-to-remove"="α-to-enter"=.25, to select the predictor variables that will be included in the regression modeling and analysis. You have to explain why and how you reach your conclusions.
E) Based on the conclusion in parts (a), (b), (c) and (d), what will be your final choice of variables to be included in the regression modeling and analysis? Explain why and how you reach your conclusion.