• The project has three parts:

**Part I: Multiple linear regression I.**

This part consists of Project 6.28 in the book, with the following additional part:

628. Refer to the CDI data set in Appendix C.2. You have been asked to evaluate two alternative models for predicting the number of active physicians (Y) in a CDI. Proposed model I includes as predictor variables total population (X I), land area (X2), and total personal income (X3). Proposed model II includes as predictor variables population density (X I, total population divided by land area), percent of population greater than 64 years old (X2), and total personal income (X3).

a. Prepare a stem-and-Ieaf plot for each of the predictor variables. What noteworthy information is provided by your plots?

b. Obtain the scatter plot matrix and the correlation matrix for each proposed model. Summarize the information provided.

c. For each proposed model, fit the first-order regression model (6.5) with three predictor variables.

d. Calculate R2 for each model. Is one model clearly preferable in terms of this measure?

e. For each model, obtain the residuals and plot them against Y, each of the three predictor variables, and each of the two-factor interaction terms. Also prepare a normal probability plot for each of the two fitted models. Interpret your plots and state your findings. Is one model clearly preferable in terms of appropriateness?

f. Now expand both models proposed above by adding all possible two-factor interactions. Note that, for a model with X1, X2, X3 as the predictors, the two-factor interactions are X1X2, X1X3, X2X3. Repeat part d for the two expanded models.

**Part II: Multiple linear regression II.**

7.37. Refer to the CDI data set in Appendix C.2. For predicting the number of active physicians (Y) in a county, it has been decided to include total population (X1) and total personal income (X2) as predictor variables. The question now is whether an additional predictor variable would be helpful in the model and, if so, which variable would be most helpful. Assume that a first-order multiple regression model is appropriate.

a. For each of the following variables, calculate the coefficient of partial determination given that X1 and X2 are included in the model: land area (X3), percent of population 65 or older (X4), number of hospital beds (X5), and total serious crimes (X6).

b. On the basis of the results in part (a), which of the four additional predictor variables is best? Is the extra sum of squares associated with this variable larger than those for the other three variables?

c. Using the F\* test statistic, test whether or not the variable determined to be best in part (b) is helpful in the regression model when XI and X2 are included in the model; use alpha = 0.01. State the alternatives, decision rule, and conclusion. Would the F\* test statistics for the other three potential predictor variables be as large as the one here? Discuss. 1. Take out variable X6, that is, you will not consider the total serious crimes (X6) as a predictor in this project. Make the corresponding changes in parts a–c.

**2. Add the following part:**

d. Compute three additional coefficients of partial determination: R2 Y,X3,X4|X1,X2 , R2 Y,X3,X5|X1,X2 , and R2 Y,X4,X5|X1,X2 . Which pair of predictors is relatively more important than other pairs? Use the F test to find out whether adding the best pair to the model is helpful given that X1, X2 are already included.

**Part III: Discussion.**

Discuss about your results from a practical standpoint. What particular parts of the course material do you find most relevant to your analysis in this project (try to be as specific as possible)? Any suggestions on how to improve the linear regression models?