6.27. In a small-scale regression study, the following data were obtained:

i:	1	2	3	4	5	6
X_{i1}:	7	4	16	3	21	8
X_{i2}:	33	41	7	49	5	31
Y_i:	42	33	75	28	91	55

Assume that regression model (6.1) with independent normal error terms is appropriate. Using matrix methods, obtain (a) **b**; (b) **e**; (c) **H**; (d) *SSR*; (e) $\mathbf{s}^2\{\mathbf{b}\}$; (f) $\hat{Y}_h$ when $X_{h1} = 10$, $X_{h2} = 30$; (g) $s^2\{\hat{Y}_h\}$ when $X_{h1} = 10$, $X_{h2} = 30$.

Projects

6.28. 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_1), land area (X_2), and total personal income (X_3). Proposed model II includes as predictor variables population density (X_1, total population divided by land area), percent of population greater than 64 years old (X_2), and total personal income (X_3).

a. Prepare a stem-and-leaf 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 R^2 for each model. Is one model clearly preferable in terms of this measure?

e. For each model, obtain the residuals and plot them against $\hat{Y}$, 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?

6.29. Refer to the **CDI** data set in Appendix C.2.

a. For each geographic region, regress the number of serious crimes in a CDI (Y) against population density (X_1, total population divided by land area), per capita personal income (X_2), and percent high school graduates (X_3). Use first-order regression model (6.5) with three predictor variables. State the estimated regression functions.

b. Are the estimated regression functions similar for the four regions? Discuss.

c. Calculate *MSE* and R^2 for each region. Are these measures similar for the four regions? Discuss.

d. Obtain the residuals for each fitted model and prepare a box plot of the residuals for each fitted model. Interpret your plots and state your findings.

6.30. Refer to the **SENIC** data set in Appendix C.1. Two models have been proposed for predicting the average length of patient stay in a hospital (Y). Model I utilizes as predictor variables age (X_1), infection risk (X_2), and available facilities and services (X_3). Model II uses as predictor variables number of beds (X_1), infection risk (X_2), and available facilities and services (X_3).

a. Prepare a stem-and-leaf plot for each of the predictor variables. What information do these plots provide?

b. Obtain the scatter plot matrix and the correlation matrix for each proposed model. Interpret these and state your principal findings.

State the reduced models for testing whether or not: (1) $\beta_1 = \beta_3 = 0$, (2) $\beta_0 = 0$, (3) $\beta_3 = 5$, (4) $\beta_0 = 10$, (5) $\beta_1 = \beta_2$.

7.33. Show the equivalence of the expressions in (7.36) and (7.41) for $R^2_{Y2|1}$.

7.34. Refer to the work crew productivity example in Table 7.6.

a. For the variables transformed according to (7.44), obtain: (1) $\mathbf{X'X}$, (2) $\mathbf{X'Y}$, (3) $\mathbf{b}$, (4) $\mathbf{s}^2\{\mathbf{b}\}$.

b. Show that the standardized regression coefficients obtained in part (a3) are related to the regression coefficients for the regression model in the original variables according to (7.53).

7.35. Derive the relations between the β_k and β_k^* in (7.46a) for $p - 1 = 2$.

7.36. Derive the expression for $\mathbf{X'Y}$ in (7.51) for standardized regression model (7.30.) for $p - 1 = 2$.

Projects

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 (X_1) and total personal income (X_2) 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 X_1 and X_2 are included in the model: land area (X_3), percent of population 65 or older (X_4), number of hospital beds (X_5), and total serious crimes (X_6).

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 X_1 and X_2 are included in the model; use $\alpha = .01$. 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.

7.38. Refer to the **SENIC** data set in Appendix C.1. For predicting the average length of stay of patients in a hospital (Y), it has been decided to include age (X_1) and infection risk (X_2) 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 X_1 and X_2 are included in the model: routine culturing ratio (X_3), average daily census (X_4), number of nurses (X_5), and available facilities and services (X_6).

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 X_1 and X_2 are included in the model; use $\alpha = .05$. 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.