Note incidentally that even though the coefficient of multiple determination, $R^2 = .917$, is high, the prediction limits here are not fully satisfactory. This serves as another reminder that a high value of R^2 does not necessarily indicate that precise predictions can be made.

Cited Reference

6.1. Box, G. E. P., and P. W. Tidwell. "Transformations of the Independent Variables," *Technometrics* 4 (1962), pp. 531–50.

Problems

6.1. Set up the **X** matrix and $\boldsymbol{\beta}$ vector for each of the following regression models (assume $i = 1, \ldots, 4$):

a. $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i1} X_{i2} + \varepsilon_i$

b. $\log Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i$

6.2. Set up the **X** matrix and $\boldsymbol{\beta}$ vector for each of the following regression models (assume $i = 1, \ldots, 5$):

a. $Y_i = \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i1}^2 + \varepsilon_i$

b. $\sqrt{Y_i} = \beta_0 + \beta_1 X_{i1} + \beta_2 \log_{10} X_{i2} + \varepsilon_i$

6.3. A student stated: "Adding predictor variables to a regression model can never reduce R^2, so we should include all available predictor variables in the model." Comment.

6.4. Why is it not meaningful to attach a sign to the coefficient of multiple correlation R, although we do so for the coefficient of simple correlation r_{12}?

6.5. **Brand preference.** In a small-scale experimental study of the relation between degree of brand liking (Y) and moisture content (X_1) and sweetness (X_2) of the product, the following results were obtained from the experiment based on a completely randomized design (data are coded):

i:	1	2	3	...	14	15	16
X_{i1}:	4	4	4	...	10	10	10
X_{i2}:	2	4	2	..	4	2	4
Y_i:	64	73	61	...	95	94	100

a. Obtain the scatter plot matrix and the correlation matrix. What information do these diagnostic aids provide here?

b. Fit regression model (6.1) to the data. State the estimated regression function. How is b_1 interpreted here?

c. Obtain the residuals and prepare a box plot of the residuals. What information does this plot provide?

d. Plot the residuals against $\hat{Y}$, X_1, X_2, and $X_1 X_2$ on separate graphs. Also prepare a normal probability plot. Interpret the plots and summarize your findings.

e. Conduct the Breusch-Pagan test for constancy of the error variance, assuming $\log \sigma_i^2 = \gamma_0 + \gamma_1 X_{i1} + \gamma_2 X_{i2}$; use $\alpha = .01$. State the alternatives, decision rule, and conclusion.

f. Conduct a formal test for lack of fit of the first-order regression function; use $\alpha = .01$. State the alternatives, decision rule, and conclusion.

6.6. Refer to **Brand preference** Problem 6.5. Assume that regression model (6.1) with independent normal error terms is appropriate.

a. Test whether there is a regression relation, using $\alpha = .01$. State the alternatives, decision rule, and conclusion. What does your test imply about β_1 and β_2?

level (X_3, an index). The administrator randomly selected 46 patients and collected the data presented below, where larger values of Y, X_2, and X_3 are, respectively, associated with more satisfaction, increased severity of illness, and more anxiety.

i:	1	2	3	...	44	45	46
X_{i1}:	50	36	40	...	45	37	28
X_{i2}:	51	46	48	...	51	53	46
X_{i3}:	2.3	2.3	2.2	...	2.2	2.1	1.8
Y_i:	48	57	66	...	68	59	92

a. Prepare a stem-and-leaf plot for each of the predictor variables. Are any noteworthy features revealed by these plots?

b. Obtain the scatter plot matrix and the correlation matrix. Interpret these and state your principal findings.

c. Fit regression model (6.5) for three predictor variables to the data and state the estimated regression function. How is b_2 interpreted here?

d. Obtain the residuals and prepare a box plot of the residuals. Do there appear to be any outliers?

e. Plot the residuals against $\hat{Y}$, each of the predictor variables, and each two-factor interaction term on separate graphs. Also prepare a normal probability plot. Interpret your plots and summarize your findings.

f. Can you conduct a formal test for lack of fit here?

g. Conduct the Breusch-Pagan test for constancy of the error variance, assuming $\log \sigma_i^2 = \gamma_0 + \gamma_1 X_{i1} + \gamma_2 X_{i2} + \gamma_3 X_{i3}$; use $\alpha = .01$. State the alternatives, decision rule, and conclusion.

*6.16. Refer to **Patient satisfaction** Problem 6.15. Assume that regression model (6.5) for three predictor variables with independent normal error terms is appropriate.

a. Test whether there is a regression relation; use $\alpha = .10$. State the alternatives, decision rule, and conclusion. What does your test imply about β_1, β_2, and β_3? What is the P-value of the test?

b. Obtain joint interval estimates of β_1, β_2, and β_3, using a 90 percent family confidence coefficient. Interpret your results.

c. Calculate the coefficient of multiple determination. What does it indicate here?

*6.17. Refer to **Patient satisfaction** Problem 6.15. Assume that regression model (6.5) for three predictor variables with independent normal error terms is appropriate.

a. Obtain an interval estimate of the mean satisfaction when $X_{h1} = 35$, $X_{h2} = 45$, and $X_{h3} = 2.2$. Use a 90 percent confidence coefficient. Interpret your confidence interval.

b. Obtain a prediction interval for a new patient's satisfaction when $X_{h1} = 35$, $X_{h2} = 45$, and $X_{h3} = 2.2$. Use a 90 percent confidence coefficient. Interpret your prediction interval.

6.18. **Commercial properties.** A commercial real estate company evaluates vacancy rates, square footage, rental rates, and operating expenses for commercial properties in a large metropolitan area in order to provide clients with quantitative information upon which to make rental decisions. The data below are taken from 81 suburban commercial properties that are the newest, best located, most attractive, and expensive for five specific geographic areas. Shown here are the age (X_1), operating expenses and taxes (X_2), vacancy rates (X_3), total square footage (X_4), and rental rates (Y).

i:	1	2	3	...	79	80	81
X_{i1}:	1	14	16	...	15	11	14
X_{i2}:	5.02	8.19	3.00	...	11.97	11.27	12.68
X_{i3}:	0.14	0.27	0	...	0.14	0.03	0.03
X_{i4}:	123,000	104,079	39,998	...	254,700	434,746	201,930
Y_i:	13.50	12.00	10.50	...	15.00	15.25	14.50

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

b. Obtain the scatter plot matrix and the correlation matrix. Interpret these and state your principal findings.

c. Fit regression model (6.5) for four predictor variables to the data. State the estimated regression function.

d. Obtain the residuals and prepare a box plot of the residuals. Does the distribution appear to be fairly symmetrical?

e. Plot the residuals against $\hat{Y}$, each predictor variable, and each two-factor interaction term on separate graphs. Also prepare a normal probability plot. Analyze your plots and summarize your findings.

f. Can you conduct a formal test for lack of fit here?

g. Divide the 81 cases into two groups, placing the 40 cases with the smallest fitted values $\hat{Y}_i$ into group 1 and the remaining cases into group 2. Conduct the Brown-Forsythe test for constancy of the error variance, using $\alpha = .05$. State the decision rule and conclusion.

6.19. Refer to **Commercial properties** Problem 6.18. Assume that regression model (6.5) for four predictor variables with independent normal error terms is appropriate.

a. Test whether there is a regression relation; use $\alpha = .05$. State the alternatives, decision rule, and conclusion. What does your test imply about β_1, β_2, β_3, and β_4? What is the P-value of the test?

b. Estimate β_1, β_2, β_3, and β_4 jointly by the Bonferroni procedure, using a 95 percent family confidence coefficient. Interpret your results.

c. Calculate R^2 and interpret this measure.

6.20. Refer to **Commercial properties** Problem 6.18. Assume that regression model (6.5) for four predictor variables with independent normal error terms is appropriate. The researcher wishes to obtain simultaneous interval estimates of the mean rental rates for four typical properties specified as follows:

	1	2	3	4
X_1:	5.0	6.0	14.0	12.0
X_2:	8.25	8.50	11.50	10.25
X_3:	0	0.23	0.11	0
X_4:	250,000	270,000	300,000	310,000

Obtain the family of estimates using a 95 percent family confidence coefficient. Employ the most efficient procedure.

6.21. Refer to **Commercial properties** Problem 6.18. Assume that regression model (6.5) for four predictor variables with independent normal error terms is appropriate. Three properties with the following characteristics did not have any rental information available.

	1	2	3
X_1:	4.0	6.0	12.0
X_2:	10.0	11.5	12.5
X_3:	0.10	0	0.32
X_4:	80,000	120,000	340,000

Develop separate prediction intervals for the rental rates of these properties, using a 95 percent statement confidence coefficient in each case. Can the rental rates of these three properties be predicted fairly precisely? What is the family confidence level for the set of three predictions?

Exercises

6.22. For each of the following regression models, indicate whether it is a general linear regression model. If it is not, state whether it can be expressed in the form of (6.7) by a suitable transformation:

a. $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 \log_{10} X_{i2} + \beta_3 X_{i1}^2 + \varepsilon_i$

b. $Y_i = \varepsilon_i \exp(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2}^2)$

c. $Y_i = \log_{10}(\beta_1 X_{i1}) + \beta_2 X_{i2} + \varepsilon_i$

d. $Y_i = \beta_0 \exp(\beta_1 X_{i1}) + \varepsilon_i$

e. $Y_i = [1 + \exp(\beta_0 + \beta_1 X_{i1} + \varepsilon_i)]^{-1}$

6.23. (Calculus needed.) Consider the multiple regression model:

$$Y_i = \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i \qquad i = 1, \ldots, n$$

where the ε_i are uncorrelated, with $E\{\varepsilon_i\} = 0$ and $\sigma^2\{\varepsilon_i\} = \sigma^2$.

a. State the least squares criterion and derive the least squares estimators of β_1 and β_2.

b. Assuming that the ε_i are independent normal random variables, state the likelihood function and obtain the maximum likelihood estimators of β_1 and β_2. Are these the same as the least squares estimators?

6.24. (Calculus needed.) Consider the multiple regression model:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i1}^2 + \beta_3 X_{i2} + \varepsilon_i \qquad i = 1, \ldots, n$$

where the ε_i are independent $N(0, \sigma^2)$.

a. State the least squares criterion and derive the least squares normal equations.

b. State the likelihood function and explain why the maximum likelihood estimators will be the same as the least squares estimators.

6.25. An analyst wanted to fit the regression model $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \varepsilon_i$, $i = 1, \ldots, n$, by the method of least squares when it is known that $\beta_2 = 4$. How can the analyst obtain the desired fit by using a multiple regression computer program?

6.26. For regression model (6.1), show that the coefficient of simple determination between Y_i and $\hat{Y}_i$ equals the coefficient of multiple determination R^2.

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.