Using Y to represent per capita expenditure on schools, the model takes the form

$$\begin{aligned} Y &= \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \gamma_1 T_1 + \gamma_2 T_2 + \delta_1 T_1 \cdot X_1 \\ &\quad + \delta_2 T_1 \cdot X_2 + \delta_3 T_1 \cdot X_3 + \alpha_1 T_2 \cdot X_1 + \alpha_2 T_2 \cdot X_2 \\ &\quad + \alpha_3 T_2 \cdot X_3 + \varepsilon. \end{aligned}$$

From the definitions of T_1 and T_2, the above model is equivalent to

$$\begin{aligned} \text{For 1960:} \quad & Y = (\beta_0 + \gamma_1) + (\beta_1 + \delta_1) X_1 + (\beta_2 + \delta_2) X_2 \\ & \qquad + (\beta_3 + \delta_3) X_3 + \varepsilon, \\ \text{For 1970:} \quad & Y = (\beta_0 + \gamma_2) + (\beta_1 + \alpha_1) X_1 + (\beta_2 + \alpha_2) X_2 \\ & \qquad + (\beta_3 + \alpha_3) X_3 + \varepsilon, \\ \text{For 1975:} \quad & Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon. \end{aligned}$$

As noted earlier, this method of analysis necessarily implies that the variability about the regression function is assumed to be equal for all three years. One formal hypothesis of interest is

$$\mathrm{H}_0 : \gamma_1 = \gamma_2 = \delta_1 = \delta_2 = \delta_3 = \alpha_1 = \alpha_2 = \alpha_3 = 0,$$

which implies that the regression system has remained unchanged throughout the period of investigation (1960–1975).

The data for this example, which we refer to as the Education Expenditures data, appear in Tables 5.12, 5.13, and 5.14 and can be obtained from the book's Website. The reader is invited to perform the analysis described above as an exercise.

EXERCISES

5.1 Using the model defined in (5.6):

(a) Check to see if the usual least squares assumptions hold.

(b) Test $\mathrm{H}_0 : \gamma = 0$ using the F-Test.

(c) Test $\mathrm{H}_0 : \gamma = 0$ using the t-Test.

(d) Verify the equivalence of the two tests above.

5.2 Using the model defined in (5.8):

(a) Check to see if the usual least squares assumptions hold.

(b) Test $\mathrm{H}_0 : \delta = 0$ using the F-Test.

(c) Test $\mathrm{H}_0 : \delta = 0$ using the t-Test.

(d) Verify the equivalence of the two tests above.

5.3 Perform a thorough analysis of the Ski Sales data in Table 5.11 using the ideas presented in Section 5.6.

5.4 Perform a thorough analysis of the Education Expenditures data in Tables 5.12, 5.13, and 5.14 using the ideas presented in Section 5.7. **(Ignore this exercise)**

Table 5.15 Regression Output from the Regression of the Weekly Wages, Y, on X (Gender: 1 = Male, 0 = Female)

ANOVA Table				
Source	Sum of Squares	df	Mean Square	F-Test
Regression	98.8313	1	98.8313	14
Residual	338.449	48	7.05101	

Coefficients Table				
Variable	Coefficient	s.e.	t-Test	p-value
Constant	15.58	0.54	28.8	< 0.0001
X	-2.81	0.75	-3.74	0.0005

5.5 Table 5.15 shows a regression output obtained from fitting the model $Y = \beta_0 + \beta_1 X + \varepsilon$ to a set of data consisting of n workers in a given company, where Y is the weekly wages in \$100 and X is the gender. The Gender variable is coded as 1 for Males and 0 for Females.

(a) How many workers are there in this data set?

(b) Compute the variance of Y?

(c) Given that $\bar{X} = 0.52$, what is $\bar{Y}$?

(d) Given that $\bar{X} = 0.52$, how many women are there in this data set?

(e) What percentage of the variability in Y can be accounted for by X?

(f) Compute the correlation coefficient between Y and X?

(g) What is your interpretation of the estimated coefficient $\hat{\beta}_1$?

(h) What is the estimated weekly wages of a man chosen at random from the workers in the company?

(i) What is the estimated weekly wages of a woman chosen at random from the workers in the company?

(j) Construct a 95% confidence interval for β_1.

(k) Test the hypothesis that the average weekly wages of men is equal to that of women. [Specify (a) the null and alternative hypotheses, (b) the test statistics, (c) the critical value, and (d) your conclusion.]

5.6 The price of a car is thought to depend on the horsepower of the engine and the country where the car is made. The variable Country has four categories: USA, Japan, Germany, and Others. To include the variable Country in a regression equation, three indicator variables are created, one for USA, another for Japan, and the third for Germany. In addition, there are three interaction variables between the horsepower and each of the three Country categories (HP*USA, HP*Japan, and HP*Germany). Some regression outputs when

Table 5.16 Some Regression Outputs When Fitting Three Models to the Car Data

Model 1

Source	Sum of Squares	df	Mean Square	F-Test
Regression	4604.7	1	4604.7	253
Residual	1604.44	88	18.2323	

Variable	Coefficient	s.e.	t-Test	p-value
Constant	-6.107	1.487	-4.11	0.0001
Horsepower	0.169	0.011	15.9	0.0001

Model 2

Source	Sum of Squares	df	Mean Square	F-Test
Regression	4818.84	4	1204.71	73.7
Residual	1390.31	85	16.3566	

Variable	Coefficient	s.e.	t-Test	p-value
Constant	-4.117	1.582	-2.6	0.0109
Horsepower	0.174	0.011	16.6	0.0001
USA	-3.162	1.351	-2.34	0.0216
Japan	-3.818	1.357	-2.81	0.0061
Germany	0.311	1.871	0.166	0.8682

Model 3

Source	Sum of Squares	df	Mean Square	F-Test
Regression	4889.3	7	698.471	43.4
Residual	1319.85	82	16.0957	

Variable	Coefficient	s.e.	t-Test	p-value
Constant	-10.882	4.216	-2.58	0.0116
Horsepower	0.237	0.038	6.21	0.0001
USA	2.076	4.916	0.42	0.6740
Japan	4.755	4.685	1.01	0.3131
Germany	11.774	9.235	1.28	0.2059
HP*USA	-0.052	0.042	-1.23	0.2204
HP*Japan	-0.077	0.041	-1.88	0.0631
HP*Germany	-0.095	0.066	-1.43	0.1560

fitting three models to the data is shown in Table 5.16. The usual regression assumptions hold.

(a) Compute the correlation coefficient between the price and the horsepower.

(b) What is the least squares estimated price of an American car with a 100 horsepower engine?

(c) Holding the horsepower fixed, which country has the least expensive car? Why?

(d) Test whether there is an interaction between Country and horsepower. Specify the null and alternative hypotheses, test statistics, and conclusions.

(e) Given the horsepower of the car, test whether the Country is an important predictor of the price of a car. Specify the null and alternative hypotheses, test statistics, and conclusions.

(f) Would you recommend that the number of categories of Country be reduced? If so, which categories can be joined together to form one category?

(g) Holding the horsepower fixed, write down the formula for the test statistic for testing the equality of the price of American and Japanese cars?

5.7 Three types of fertilizer are to be tested to see which one yields more corn crop. Forty similar plots of land were available for testing purposes. The 40 plots are divided at random into four groups, 10 plots in each group. Fertilizer 1 was applied to each of the 10 corn plots in Group 1. Similarly, Fertilizers 2 and 3 were applied to the plots in Groups 2 and 3, respectively. The corn plants in Group 4 were not given any fertilizer; it will serve as the control group. Table 5.17 gives the corn yield y_{ij} for each of the 40 plots.

(a) Create three indicator variables F_1, F_2, F_3, one for each of the three fertilizer groups.

(b) Fit the model $y_{ij} = \mu_0 + \mu_1 F_{i1} + \mu_2 F_{i2} + \mu_3 F_{i3} + \varepsilon_{ij}$.

(c) Test the hypothesis that, on the average, none of the three types of fertilizer has an effect on corn crops. Specify the hypothesis to be tested, the test used, and your conclusions at the 5% significance level.

(d) Test the hypothesis that, on the average, the three types of fertilizer have equal effects on corn crop but different from that of the control group. Specify the hypothesis to be tested, the test used, and your conclusions at the 5% significance level.

(e) Which of the three fertilizers has the greatest effects on corn yield?

5.8 In a statistics course personal information was collected on all the students for class analysis. Data on age (in years), height (in inches), and weight (in pounds) of the students are given in Table 5.18 and can be obtained from the book's Website. The gender of each student is also noted and coded as 1 for women and 0 for men. We want to study the relationship between the height

Table 5.17 Corn Yields by Fertilizer Group

Fertilizer 1	Fertilizer 2	Fertilizer 3	Control Group
31	27	36	33
34	27	37	27
34	25	37	35
34	34	34	25
43	21	37	29
35	36	28	20
38	34	33	25
36	30	29	40
36	32	36	35
45	33	42	29

and weight of students. Weight is taken as the response variable, and the height as the predictor variable.

(a) Do you agree or do you think the roles of the variables should be reversed?

(b) Is a single equation adequate to describe the relationship between height and weight for the two groups of students? Examine the standardized residual plot from the model fitted to the pooled data, distinguishing between the male and female students.

(c) Find the best model that describes the relationship between the weight and the height of students. Use interaction variables and the methodology described in this chapter.

(d) Do you think we should include age as a variable to predict weight? Give an intuitive justification for your answer.

5.9 Presidential Election Data (1916–1996): The data in Table 5.19 were kindly provided by Professor Ray Fair of Yale University, who has found that the proportion of votes obtained by a presidential candidate in a U.S.A. presidential election can be predicted accurately by three macroeconomic variables, incumbency, and a variable which indicates whether the election was held during or just after a war. The variables considered are given in Table 5.20. All growth rates are annual rates in percentage points. Consider fitting the following initial model to the data:

$$\begin{aligned} V &= \beta_0 + \beta_1 I + \beta_2 D + \beta_3 W + \beta_4 (G \cdot I) \\ &\quad + \beta_5 P + \beta_6 N + \varepsilon. \end{aligned} \tag{5.11}$$

(a) Write the regression model corresponding to each of the three possible values of D in (5.11) and interpret the regression coefficient of D (β_2).

(b) Do we need to keep the variable I in the above model?

Table 5.18 Class Data on Age (in Years), Height (in Inches), Weight (in Pounds), and Gender (1 = Female, 0 = Male)

Age	Height	Weight	Gender	Age	Height	Weight	Gender
19	61	180	0	19	65	135	1
19	70	160	0	19	70	120	0
19	70	135	0	21	69	142	0
19	71	195	0	20	63	108	1
19	64	130	1	19	63	118	1
19	64	120	1	20	72	135	0
21	69	135	1	19	73	169	0
19	67	125	0	19	69	145	0
19	62	120	1	27	69	130	1
20	66	145	0	18	64	135	0
19	65	155	0	20	61	115	1
19	69	135	1	19	68	140	0
19	66	140	0	21	70	152	0
19	63	120	1	19	64	118	1
19	69	140	0	19	62	112	1
18	66	113	1	19	64	100	1
18	68	180	0	20	67	135	1
19	72	175	0	20	63	110	1
19	70	169	0	20	68	135	0
19	74	210	0	18	63	115	1
20	66	104	1	19	68	145	0
20	64	105	1	19	65	115	1
20	65	125	1	19	63	128	1
20	71	120	1	20	68	140	1
19	69	119	1	19	69	130	0
20	64	140	1	19	69	165	0
20	67	185	1	19	69	130	0
19	60	110	1	20	70	180	0
20	66	120	1	28	65	110	1
19	71	175	0	19	55	155	0

(c) Do we need to keep the interaction variable $(G \cdot I)$ in the above model?

(d) Examine different models to produce the model or models that might be expected to perform best in predicting future presidential elections. Include interaction terms if needed.

5.10 Refer to the Presidential Election Data in Exercise 5.9, where the variable D is a categorical variable with three categories. Now, if we replace D by two indicator variables such as:

$D_1 = 1$ if $D = 1$ (Democratic incumbent is running) and 0 otherwise, and
$D_2 = 1$ if $D = -1$ (Republican incumbent is running) and 0 otherwise.

Table 5.19 Presidential Election Data (1916–1996)

Year	V	I	D	W	G	P	N
1916	0.5168	1	1	0	2.229	4.252	3
1920	0.3612	1	0	1	−11.463	16.535	5
1924	0.4176	−1	−1	0	−3.872	5.161	10
1928	0.4118	−1	0	0	4.623	0.183	7
1932	0.5916	−1	−1	0	−14.901	7.069	4
1936	0.6246	1	1	0	11.921	2.362	9
1940	0.5500	1	1	0	3.708	0.028	8
1944	0.5377	1	1	1	4.119	5.678	14
1948	0.5237	1	1	1	1.849	8.722	5
1952	0.4460	1	0	0	0.627	2.288	6
1956	0.4224	−1	−1	0	−1.527	1.936	5
1960	0.5009	−1	0	0	0.114	1.932	5
1964	0.6134	1	1	0	5.054	1.247	10
1968	0.4960	1	0	0	4.836	3.215	7
1972	0.3821	−1	−1	0	6.278	4.766	4
1976	0.5105	−1	0	0	3.663	7.657	4
1980	0.4470	1	1	0	−3.789	8.093	5
1984	0.4083	−1	−1	0	5.387	5.403	7
1988	0.4610	−1	0	0	2.068	3.272	6
1992	0.5345	−1	−1	0	2.293	3.692	1
1996	0.5474	1	1	0	2.918	2.268	3

Then an alternative to the model in (5.11) is

$$\begin{aligned} V &= \beta_0 + \beta_1 I + \alpha_1 D_1 + \alpha_2 D_2 + \beta_3 W + \beta_4 (G \cdot I) \\ &\quad + \beta_5 P + \beta_6 N + \varepsilon. \end{aligned} \tag{5.12}$$

(a) Write the regression model corresponding to each of the three possible values of D in (5.12) and interpret the regression coefficients of D_1 and D_2.

(b) Show that the model in (5.11) can be obtained as a special case of the model in (5.11) by assuming that $\alpha_1 = -\alpha_2$.

(c) Do the data in Table 5.19 support the assumption that $\alpha_1 = -\alpha_2$?

5.11 Use the data given in Table 1.10 (A description of the data is found in Section 1.3.6).

(a) Examine the relationship between polishing times, the diameters, and the product type. Does the relationship vary between the categories?

(b) Polishing time plays an important part in the cost. Construct a regression model which connects price with the product types, polishing time, and diameter.

Table 5.20 Variables for the Presidential Election Data (1916–1996) in Table 5.19

Variable	Definition
YEAR	Election year
V	Democratic share of the two-party presidential vote
I	Indicator variable (1 if there is a Democratic incumbent at the time of the election and -1 if there is a Republican incumbent)
D	Categorical variable (1 if a Democratic incumbent is running for election, -1 if a Republican incumbent is running for election, and 0 otherwise)
W	Indicator variable (1 for the elections of 1920, 1944, and 1948, and 0 otherwise)
G	Growth rate of real per capita GDP in the first three quarters of the election year
P	Absolute value of the growth rate of the GDP deflator in the first 15 quarters of the administration
N	Number of quarters in the first 15 quarters of the administration in which the growth rate of real per capita GDP is greater than 3.2%