REGRESSION ANALYSIS BY EXAMPLE

Fifth Edition

Samprit Chatterjee
Ali S. Hadi

WILEY-INTERSCIENCE

A JOHN WILEY & SONS, INC., PUBLICATION

CHAPTER 5

QUALITATIVE VARIABLES AS PREDICTORS

5.1 INTRODUCTION

Qualitative or categorical variables can be very useful as predictor variables in regression analysis. Qualitative variables such as gender, marital status, or political affiliation can be represented by *indicator* or *dummy* variables. These variables take on only two values, usually 0 and 1. The two values signify that the observation belongs to one of two possible categories. The numerical values of indicator variables are not intended to reflect a quantitative ordering of the categories, but only serve to identify category or class membership. For example, an analysis of salaries earned by computer programmers may include variables such as education, years of experience, and gender as predictor variables. The gender variable could be quantified, say, as 1 for female and 0 for male. Indicator variables can also be used in a regression equation to distinguish among three or more groups as well as among classifications across various types of groups. For example, the regression described above may also include an indicator variable to distinguish whether the observation was for a systems or applications programmer. The four conditions determined by gender and type of programming can be represented by combining the two variables, as we shall see in this chapter.

Regression Analysis by Example, Fifth Edition. By Samprit Chatterjee and Ali S. Hadi

Table 5.1 Salary Survey Data

Row	S	X	E	M	Row	S	X	E	M
1	13876	1	1	1	24	22884	6	2	1
2	11608	1	3	0	25	16978	7	1	1
3	18701	1	3	1	26	14803	8	2	0
4	11283	1	2	0	27	17404	8	1	1
5	11767	1	3	0	28	22184	8	3	1
6	20872	2	2	1	29	13548	8	1	0
7	11772	2	2	0	30	14467	10	1	0
8	10535	2	1	0	31	15942	10	2	0
9	12195	2	3	0	32	23174	10	3	1
10	12313	3	2	0	33	23780	10	2	1
11	14975	3	1	1	34	25410	11	2	1
12	21371	3	2	1	35	14861	11	1	0
13	19800	3	3	1	36	16882	12	2	0
14	11417	4	1	0	37	24170	12	3	1
15	20263	4	3	1	38	15990	13	1	0
16	13231	4	3	0	39	26330	13	2	1
17	12884	4	2	0	40	17949	14	2	0
18	13245	5	2	0	41	25685	15	3	1
19	13677	5	3	0	42	27837	16	2	1
20	15965	5	1	1	43	18838	16	2	0
21	12336	6	1	0	44	17483	16	1	0
22	21352	6	3	1	45	19207	17	2	0
23	13839	6	2	0	46	19346	20	1	0

Indicator variables can be used in a variety of ways and may be considered whenever there are qualitative variables affecting a relationship. We shall illustrate some of the applications with examples and suggest some additional applications. It is hoped that the reader will recognize the general applicability of the technique from the examples. In the first example, we look at data on a salary survey, such as the one mentioned above, and use indicator variables to adjust for various categorical variables that affect the regression relationship. The second example uses indicator variables for analyzing and testing for equality of regression relationships in various subsets of a population.

We continue to assume that the response variable is a quantitative continuous variable, but the predictor variables can be quantitative and/or categorical. The case where the response variable is an indicator variable is dealt with in Chapter 12.

5.2 SALARY SURVEY DATA

The Salary Survey data set was developed from a salary survey of computer professionals in a large corporation. The objective of the survey was to identify and quantify those variables that determine salary differentials. In addition, the data

could be used to determine if the corporation's salary administration guidelines were being followed. The data appear in Table 5.1 and can be obtained from the book's Website.[1] The response variable is salary (S) and the predictors are: (1) experience (X), measured in years; (2) education (E), coded as 1 for completion of a high school (H.S.) diploma, 2 for completion of a bachelor degree (B.S.), and 3 for the completion of an advanced degree; and (3) management (M), which is coded as 1 for a person with management responsibility and 0 otherwise. We shall try to measure the effects of these three variables on salary using regression analysis.

A linear relationship will be used for salary and experience. We shall assume that each additional year of experience is worth a fixed salary increment. Education may also be treated in a linear fashion. If the education variable is used in the regression equation in raw form, we would be assuming that each step up in education is worth a fixed increment in salary. That is, with all other variables held constant, the relationship between salary and education is linear. That interpretation is possible but may be too restrictive. Instead, we shall view education as a categorical variable and define two indicator variables to represent the three categories. These two variables allow us to pick up the effect of education on salary whether or not it is linear. The management variable is also an indicator variable designating the two categories, 1 for management positions and 0 for regular staff positions.

Note that when using indicator variables to represent a set of categories, the number of these variables required is one less than the number of categories. For example, in the case of the education categories above, we create two indicator variables E_1 and E_2, where

$$E_{i1} = \begin{cases} 1, & \text{if the } i\text{th person is in the H.S. category,} \\ 0, & \text{otherwise,} \end{cases}$$

and

$$E_{i2} = \begin{cases} 1, & \text{if the } i\text{th person is in the B.S. category,} \\ 0, & \text{otherwise.} \end{cases}$$

As stated above, these two variables taken together uniquely represent the three groups. For H.S., $E_1 = 1, E_2 = 0$; for B.S., $E_1 = 0, E_2 = 1$; and for advanced degree, $E_1 = 0, E_2 = 0$. Furthermore, if there were a third variable, E_{i3}, defined to be 1 or 0 depending on whether or not the ith person is in the advanced degree category, then for each person we have $E_1 + E_2 + E_3 = 1$. Then $E_3 = 1 - E_1 - E_2$, showing clearly that one of the variables is superfluous.[2] Similarly, there is only one indicator variable required to distinguish the two management categories. The category that is not represented by an indicator variable is referred to as the *base category* or the *control group* because the regression coefficients of the indicator variables are interpreted relative to the control group.

[1] http://www.aucegypt.edu/faculty/hadi/RABE5

[2] Had E_1, E_2, and E_3 been used, there would have been a perfect linear relationship among the predictors, which is an extreme case of collinearity, a problem described in Chapters 9 and 10.

Table 5.2 Regression Equations for the Six Categories of Education and Management

Category	E	M	Regression Equation
1	1	0	$S = (\beta_0 + \gamma_1) + \beta_1 X + \varepsilon$
2	1	1	$S = (\beta_0 + \gamma_1 + \delta_1) + \beta_1 X + \varepsilon$
3	2	0	$S = (\beta_0 + \gamma_2) + \beta_1 X + \varepsilon$
4	2	1	$S = (\beta_0 + \gamma_2 + \delta_1) + \beta_1 X + \varepsilon$
5	3	0	$S = \beta_0 + \beta_1 X + \varepsilon$
6	3	1	$S = (\beta_0 + \delta_1) + \beta_1 X + \varepsilon$

Table 5.3 Regression Analysis of Salary Survey Data

Variable	Coefficient	s.e.	t-Test	p-value
Constant	11031.800	383.2	28.80	< 0.0001
X	546.184	30.5	17.90	< 0.0001
E_1	-2996.210	411.8	-7.28	< 0.0001
E_2	147.825	387.7	0.38	0.7049
M	6883.530	313.9	21.90	< 0.0001
$n = 46$	$R^2 = 0.957$	$R_a^2 = 0.953$	$\hat{\sigma} = 1027$	df = 41

In terms of the indicator variables described above, the regression model is

$$S = \beta_0 + \beta_1 X + \gamma_1 E_1 + \gamma_2 E_2 + \delta_1 M + \varepsilon. \tag{5.1}$$

By evaluating (5.1) for the different values of the indicator variables, it follows that there is a different regression equation for each of the six (three education and two management) categories as shown in Table 5.2. According to the proposed model, we may say that the indicator variables help to determine the base salary level as a function of education and management status after adjustment for years of experience.

The results of the regression computations for the model given in (5.1) appear in Table 5.3. The proportion of salary variation accounted for by the model is quite high ($R^2 = 0.957$). At this point in the analysis we should investigate the pattern of residuals to check on model specification. We shall postpone that investigation for now and assume that the model is satisfactory so that we can discuss the interpretation of the regression results. Later we shall return to analyze the residuals and find that the model must be altered.

We see that the coefficient of X is 546.16. That is, each additional year of experience is estimated to be worth an annual salary increment of \$546. The other coefficients may be interpreted by looking into Table 5.2. The coefficient of the management indicator variable, δ_1, is estimated to be 6883.50. From Table

5.2 we interpret this amount to be the average incremental value in annual salary associated with a management position. For the education variables, γ_1 measures the salary differential for the H.S. category relative to the advanced degree category and γ_2 measures the differential for the B.S. category relative to the advanced degree category. The difference, $\gamma_2 - \gamma_1$, measures the differential salary for the H.S. category relative to the B.S. category. From the regression results, in terms of salary for computer professionals, we see that an advanced degree is worth \$2996 more than a high school diploma, a B.S. is worth \$148 more than an advanced degree (this differential is not statistically significant, $t = 0.38$), and a B.S. is worth about \$3144 more than a high school diploma. These salary differentials hold for every fixed level of experience.

5.3 INTERACTION VARIABLES

Returning now to the question of model specification, consider Figure 5.1, where the residuals are plotted against X. The plot suggests that there may be three or more specific levels of residuals. Possibly the indicator variables that have been defined are not adequate for explaining the effects of education and management status. Actually, each residual is identified with one of the six education-management combinations. To see this we plot the residuals against Category (a new categorical variable that takes a separate value for each of the six combinations). This graph is, in effect, a plot of residuals versus a potential predictor variable that has not yet been used in the equation. The graph is given in Figure 5.2. It can be seen from the graph that the residuals cluster by size according to their education-management category. The combinations of education and management have not been satisfactorily treated in the model. Within each of the six groups, the residuals are either almost totally positive or totally negative. This behavior implies that the model given in (5.1) does not adequately explain the relationship between salary and experience, education, and management variables. The graph points to some hidden structure in the data that has not been explored.

The graphs strongly suggest that the effects of education and management status on salary determination are not additive. Note that in the model in (5.1) and its further exposition in Table 5.2, the incremental effects of both variables are determined by additive constants. For example, the effect of a management position is measured as δ_1, independently of the level of educational attainment. The nonadditive effects of these variables can be evaluated by constructing additional variables that are used to measure what may be referred to as *multiplicative* or *interaction effects*. Interaction variables are defined as products of the existing indicator variables $(E_1 \cdot M)$ and $(E_2 \cdot M)$. The inclusion of these two variables on the right-hand side of (5.1) leads to a model that is no longer additive in education and management, but recognizes the multiplicative effect of these two variables.

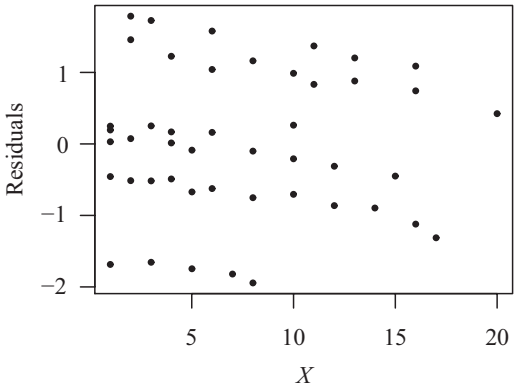

Figure 5.1 Standardized residuals versus years of experience (X).

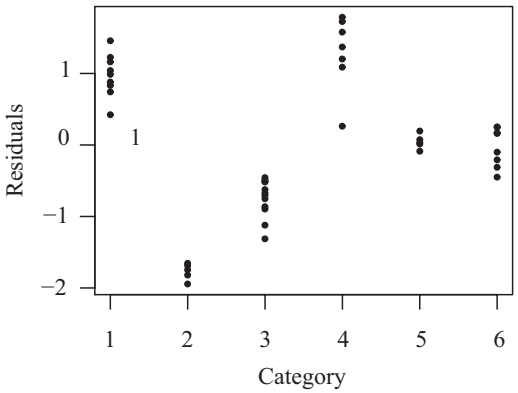

Figure 5.2 Standardized residuals versus education-management categorical variable.

Table 5.4 Regression Analysis of Salary Data: Expanded Model

Variable	Coefficient	s.e.	t-Test	p-value
Constant	11203.40	79.07	141.7	< 0.0001
X	496.99	5.57	89.3	< 0.0001
E_1	-1730.75	105.30	-16.4	< 0.0001
E_2	-349.08	97.57	-3.6	0.0009
M	7047.41	102.60	68.7	< 0.0001
$E_1 \cdot M$	-3066.04	149.30	-20.5	< 0.0001
$E_2 \cdot M$	1836.49	131.20	14.0	< 0.0001
$n = 46$	$R^2 = 0.999$	$R_a^2 = 0.999$	$\hat{\sigma} = 173.8$	df = 39

Table 5.5 Regression Analysis of Salary Data: Expanded Model, Observation 33 Deleted.

Variable	Coefficient	s.e.	t-Test	p-value
Constant	11199.70	30.54	367.0	< 0.0001
X	498.41	2.15	232.0	< 0.0001
E_1	-1741.28	40.69	-42.8	< 0.0001
E_2	-357.00	37.69	-9.5	< 0.0001
M	7040.49	39.63	178.0	< 0.0001
$E_1 \cdot M$	-3051.72	57.68	-52.9	< 0.0001
$E_2 \cdot M$	1997.62	51.79	38.6	< 0.0001
$n = 45$	$R^2 = 1.0$	$R_a^2 = 1.0$	$\hat{\sigma} = 67.13$	df = 38

The expanded model is

$$\begin{aligned} S &= \beta_0 + \beta_1 X + \gamma_1 E_1 + \gamma_2 E_2 + \delta_1 M \\ &\quad + \alpha_1 (E_1 \cdot M) + \alpha_2 (E_2 \cdot M) + \varepsilon. \end{aligned} \tag{5.2}$$

The regression results are given in Table 5.4. The residuals from the regression of the expanded model are plotted against X in Figure 5.3. Note that observation 33 is an outlier. Salary is overpredicted by the model. Checking this observation in the listing of the raw data, it appears that this particular person seems to have fallen behind by a couple of hundred dollars in annual salary as compared to other persons with similar characteristics. To be sure that this single observation is not overly affecting the regression estimates, it has been deleted and the regression rerun. The new results are given in Table 5.5.

The regression coefficients are basically unchanged. However, the standard deviation of the residuals has been reduced to \$67.28 and the proportion of explained variation has reached 0.9998. The plot of residuals versus X (Figure 5.4) appears

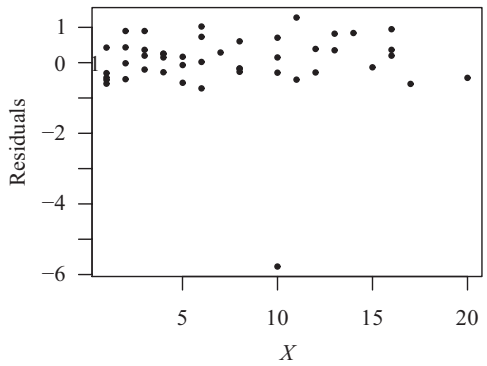

Figure 5.3 Standardized residuals versus years of experience: Expanded model.

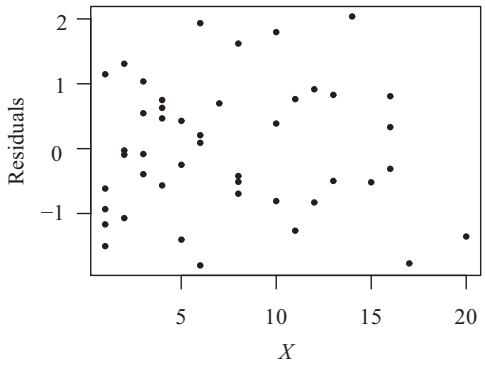

Figure 5.4 Standardized residuals versus years of experience: Expanded model, observation 33 deleted.

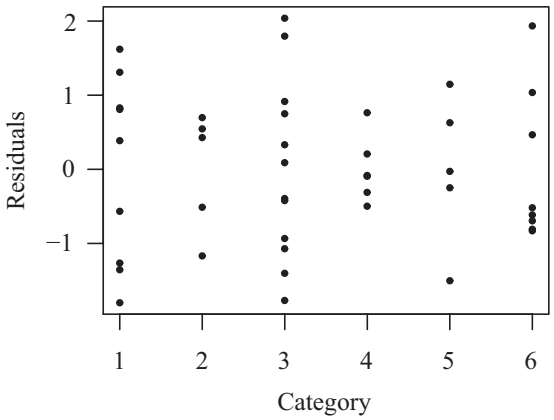

Figure 5.5 Standardized residuals versus education-management categorical variable: Expanded model, observation 33 deleted.

to be satisfactory compared with the similar residual plot for the additive model. In addition, the plot of residuals for each education-management category (Figure 5.5) shows that each of these groups has residuals that appear to be symmetrically distributed about zero. Therefore the introduction of the interaction terms has produced an accurate representation of salary variations. The relationships between salary and experience, education, and management status appear to be adequately described by the model given in (5.2).

With the standard error of the residuals estimated to be \$67.28, we can believe that we have uncovered the actual and very carefully administered salary formula. Using 95% confidence intervals, each year of experience is estimated to be worth between \$494.08 and \$502.72. These increments of approximately \$500 are added to a starting salary that is specified for each of the six education-management groups. Since the final regression model is not additive, it is rather difficult to directly interpret the coefficients of the indicator variables. To see how the qualitative variables affect salary differentials, we use the coefficients to form estimates of the base salary for each of the six categories. These results are presented in Table 5.6 along with standard errors and confidence intervals. The standard errors are computed using (A.12) in the Appendix to Chapter 3.

Using a regression model with indicator variables and interaction terms, it has been possible to account for almost all the variation in salaries of computer professionals selected for this survey. The level of accuracy with which the model explains the data is very rare! We can only conjecture that the methods of salary administration in this company are precisely defined and strictly applied.

In retrospect, we see that an equivalent model may be obtained with a different set of indicator variables and regression parameters. One could define five variables, each taking on the values of 1 or 0, corresponding to five of the six education-management categories. The numerical estimates of base salary and the standard errors of Table 5.6 would be the same. The advantage to proceeding as we have is that it allows us to separate the effects of the three sets of predictor variables, (1) education, (2) management, and (3) education-management interaction. Recall that interaction terms were included only after we found that an additive model did not satisfactorily explain salary variations. In general, we start with simple models and proceed sequentially to more complex models if necessary. We shall always hope to retain the simplest model that has an acceptable residual structure.

5.4 SYSTEMS OF REGRESSION EQUATIONS: COMPARING TWO GROUPS

A collection of data may consist of two or more distinct subsets, each of which may require a separate regression equation. Serious bias may be incurred if one regression relationship is used to represent the pooled data set. An analysis of this problem can be accomplished using indicator variables. An analysis of separate regression equations for subsets of the data may be applied to *cross-sectional* or *time*

Table 5.6 Estimates of Base Salary Using the Nonadditive Model in (5.2)

Category	E	M	Coefficients	Estimate of Base Salary[a]	s.e.[a]	95% Confidence Interval
1	1	0	$\beta_0 + \gamma_1$	9459	31	(9398, 9520)
2	1	1	$\beta_0 + \gamma_1 + \delta + \alpha_1$	13448	32	(13385, 13511)
3	2	0	$\beta_0 + \gamma_2$	10843	26	(10792, 10894)
4	2	1	$\beta_0 + \gamma_2 + \delta + \alpha_2$	19880	33	(19815, 19945)
5	3	0	β_0	11200	31	(11139, 11261)
6	3	1	$\beta_0 + \delta$	18240	29	(18183, 18297)

[a] Recorded to the nearest dollar.

series data. The example discussed below treats cross-sectional data. Applications to time series data are discussed in Section 5.5.

The model for the two groups can be different in all aspects or in only some aspects. In this section we discuss three distinct cases:

1. Each group has a separate regression model.

2. The models have the same intercept but different slopes.

3. The models have the same slope but different intercepts.

We illustrate these cases below when we have only one quantitative predictor variable. These ideas can be extended straightforwardly to the cases where there are more than one quantitative predictor variable.

5.4.1 Models with Different Slopes and Different Intercepts

We illustrate this case with an important problem concerning equal opportunity in employment. Many large corporations and government agencies administer a preemployment test in an attempt to screen job applicants. The test is supposed to measure an applicant's aptitude for the job and the results are used as part of the information for making a hiring decision. The federal government has ruled[3] that these tests (1) must measure abilities that are directly related to the job under consideration and (2) must not discriminate on the basis of race or national origin. Operational definitions of requirements (1) and (2) are rather elusive. We shall not try to resolve these operational problems. We shall take one approach involving race represented as two groups, white and minority. The hypothesis that there are separate regressions relating test scores to job performance for the two groups will be examined. The implications of this hypothesis for discrimination in hiring are discussed.

[3] Tower amendment to Title VII, Civil Rights Act of 1964.

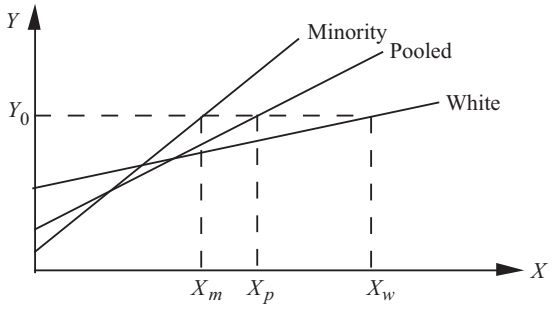

Figure 5.6 Requirements for employment on pretest.

Let Y represent job performance and let X be the score on the preemployment test. We want to compare

$$\begin{aligned}
&\text{Model 1 (Pooled):} && y_{ij} = \beta_0 \ + \beta_1 x_{ij} \ + \varepsilon_{ij}, \ \ j = 1, 2; \ \ i = 1, 2, \cdots, n_j, \\
&\text{Model 2 (Minority):} && y_{i1} = \beta_{01} + \beta_{11} x_{i1} + \varepsilon_{i1}, \\
&\text{Model 2 (White):} && y_{i2} = \beta_{02} + \beta_{12} x_{i2} + \varepsilon_{i2}.
\end{aligned} \tag{5.3}$$

Figure 5.6 depicts the two models. In model 1, race distinction is ignored, the data are pooled, and there is one regression line. In model 2 there is a separate regression relationship for the two subgroups, each with distinct regression coefficients. We shall assume that the variances of the residual terms are the same in each subgroup.

Before analyzing the data, let us briefly consider the types of errors that could be present in interpreting and applying the results. If Y_0, as seen on the graph, has been set as the minimum required level of performance, then using Model 1, an acceptable score on the test is one that exceeds X_p. However, if Model 2 is in fact correct, the appropriate test score for whites is X_w and for minorities is X_m. Using X_p in place of X_m and X_w represents a relaxation of the pretest requirement for whites and a tightening of that requirement for minorities. Since inequity can result in the selection procedure if the wrong model is used to set cutoff values, it is necessary to examine the data carefully. It must be determined whether there are two distinct relationships or whether the relationship is the same for both groups and a single equation estimated from the pooled data is adequate. Note that whether Model 1 or Model 2 is chosen, the values X_m, X_w, and X_p are estimates subject to sampling errors and should only be used in conjunction with appropriate confidence intervals. (Construction of confidence intervals is discussed in the following paragraphs.)

Data were collected for this analysis using a special employment program. Twenty applicants were hired on a trial basis for six weeks. One week was spent in a training class. The remaining five weeks were spent on the job. The participants were selected from a pool of applicants by a method that was not related to the preemployment test scores. A test was given at the end of the training period and a work performance evaluation was developed at the end of the six-week period.

Table 5.7 Data on Preemployment Testing Program

Row	TEST	RACE	JPERF	Row	TEST	RACE	JPERF
1	0.28	1	1.83	11	2.36	0	3.25
2	0.97	1	4.59	12	2.11	0	5.30
3	1.25	1	2.97	13	0.45	0	1.39
4	2.46	1	8.14	14	1.76	0	4.69
5	2.51	1	8.00	15	2.09	0	6.56
6	1.17	1	3.30	16	1.50	0	3.00
7	1.78	1	7.53	17	1.25	0	5.85
8	1.21	1	2.03	18	0.72	0	1.90
9	1.63	1	5.00	19	0.42	0	3.85
10	1.98	1	8.04	20	1.53	0	2.95

These two scores were combined to form an index of job performance. (Those employees with unsatisfactory performance at the end of the six-week period were dropped.) The data appear in Table 5.7 and can be obtained from the book's Website. We refer to this data set as the Preemployment Testing data.

Formally, we want to test the null hypothesis $\mathrm{H}_0 : \beta_{11} = \beta_{12}, \beta_{01} = \beta_{02}$ against the alternative that there are substantial differences in these parameters. The test can be performed using indicator variables. Let z_{ij} be defined to take the value 1 if $j = 1$ and to take the value 0 if $j = 2$. That is, Z is a new variable that has the value 1 for a minority applicant and the value 0 for a white applicant. We consider the two models

$$\begin{aligned} &\text{Model 1:} \quad y_{ij} = \beta_0 + \beta_1 x_{ij} + \varepsilon_{ij} \\ &\text{Model 3:} \quad y_{ij} = \beta_0 + \beta_1 x_{ij} + \gamma z_{ij} + \delta(z_{ij} \cdot x_{ij}) + \varepsilon_{ij}. \end{aligned} \tag{5.4}$$

The variable $(z_{ij} \cdot x_{ij})$ represents the interaction between the group (race) variable Z and the preemployment test X. Note that Model 3 is equivalent to Model 2. This can be seen if we observe that for the minority group, $x_{ij} = x_{i1}$ and $z_{ij} = 1$; hence Model 3 becomes

$$\begin{aligned} y_{i1} &= \beta_0 + \beta_1 x_{i1} + \gamma + \delta x_{i1} + \varepsilon_{i1} \\ &= (\beta_0 + \gamma) + (\beta_1 + \delta) x_{i1} + \varepsilon_{i1} \\ &= \beta_{01} + \beta_{11} x_{i1} + \varepsilon_{i1}, \end{aligned}$$

which is the same as Model 2 for minority with $\beta_{01} = \beta_0 + \gamma$ and $\beta_{11} = \beta_1 + \delta$. Similarly, for the white group, we have $x_{ij} = x_{i2}$, $z_{ij} = 0$, and Model 3 becomes

$$y_{i2} = \beta_0 + \beta_1 x_{i2} + \varepsilon_{i2},$$

which is the same as Model 2 for white with $\beta_{02} = \beta_0$ and $\beta_{12} = \beta_1$. Therefore, a comparison between Models 1 and 2 is equivalent to a comparison between Models 1 and 3. Note that Model 3 can be viewed as a full model (FM) and Model 1

Table 5.8 Regression Results, Preemployment Testing Data: Model 1

Variable	Coefficient	s.e.	t-Test	p-value
Constant	1.03	0.87	1.19	0.2486
TEST (X)	2.36	0.54	4.39	0.0004
$n = 20$	$R^2 = 0.52$	$R_a^2 = 0.49$	$\hat{\sigma} = 1.59$	df = 18

Table 5.9 Regression Results, Preemployment Testing Data: Model 3

Variable	Coefficient	s.e.	t-Test	p-value
Constant	2.01	1.05	1.91	0.0736
TEST (X)	1.31	0.67	1.96	0.0677
RACE (Z)	-1.91	1.54	-1.24	0.2321
RACE · TEST ($X \cdot Z$)	2.00	0.95	2.09	0.0527
$n = 20$	$R^2 = 0.664$	$R_a^2 = 0.601$	$\hat{\sigma} = 1.41$	df = 16

as a restricted model (RM) because Model 1 is obtained from Model 3 by setting $\gamma = \delta = 0$. Thus, our null hypothesis H_0 now becomes $H_0 : \gamma = \delta = 0$. The hypothesis is tested by constructing an F-Test for the comparison of two models as described in Chapter 3. In this case, the test statistics is

$$F = \frac{[\text{SSE(RM)} - \text{SSE(FM)}]/2}{\text{SSE(FM)}/16},$$

which has 2 and 16 degrees of freedom. (Why?) Proceeding with the analysis of the data, the regression results for Model 1 and Model 3 are given in Tables 5.8 and 5.9. The plots of residuals against the predictor variable (Figures 5.7 and 5.8) look acceptable in both cases. The one residual at the lower right in Model 1 may require further investigation.

To evaluate the formal hypothesis we compute the F-ratio specified previously, which is equal to

$$F = \frac{(45.51 - 31.81)/2}{31.81/16} = 3.4$$

and is significant at a level slightly above 5%. Therefore, on the basis of this test we would conclude that the relationship is probably different for the two groups. Specifically, for minorities we have

$$Y_1 = 0.10 + 3.31X_1$$

and for whites we have

$$Y_2 = 2.01 + 1.32X_2.$$

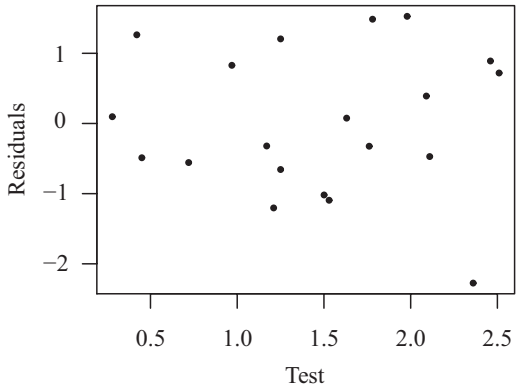

Figure 5.7 Standardized residuals versus test score: Model 1.

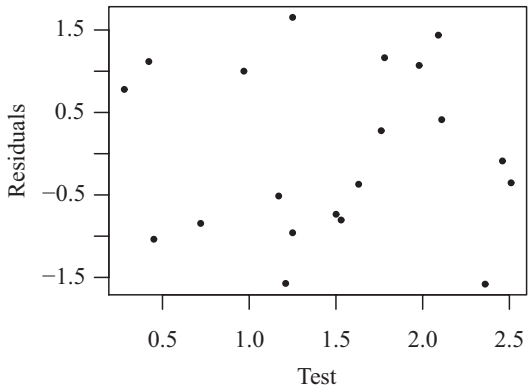

Figure 5.8 Standardized residuals versus test score: Model 3.

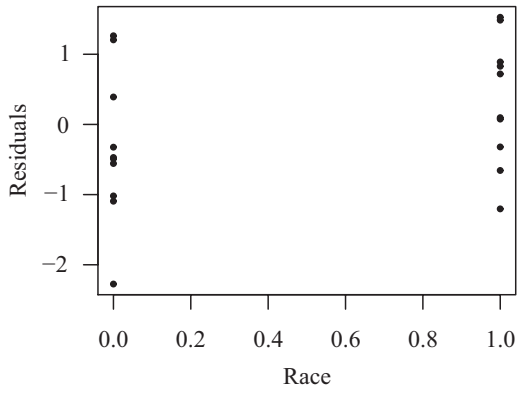

Figure 5.9 Standardized residuals versus race: Model 1.

Table 5.10 Separate Regression Results

Sample	$\hat{\beta}_0$	$\hat{\beta}_1$	t_1	R^2	$\hat{\sigma}$	df
Minority	0.10	3.31	5.31	0.78	1.29	8
White	2.01	1.31	1.82	0.29	1.51	8

The results are very similar to those that were described in Figure 5.5 when the problem of bias was discussed. The straight line representing the relationship for minorities has a larger slope and a smaller intercept than the line for whites. If a pooled model were used, the types of biases discussed in relation to Figure 5.6 would occur.

Although the formal procedure using indicator variables has led to the plausible conclusion that the relationships are different for the two groups, the data for the individual groups have not been looked at carefully. Recall that it was assumed that the variances were identical in the two groups. This assumption was required so that the only distinguishing characteristic between the two samples was the pair of regression coefficients. In Figure 5.9 a plot of residuals versus the indicator variable is presented. There does not appear to be a difference between the two sets of residuals. We shall now look more closely at each group. The regression coefficients for each sample taken separately are presented in Table 5.10. The residuals are shown in Figures 5.10 and 5.11. The regression coefficients are, of course, the values obtained from Model 3. The standard errors of the residuals are 1.29 and 1.51 for the minority and white samples, respectively. The residual plots against the test score are acceptable in both cases. An interesting observation that was not available in the earlier analysis is that the preemployment test accounts for a major portion of the variation in the minority sample, but the test is only marginally useful in the white sample.

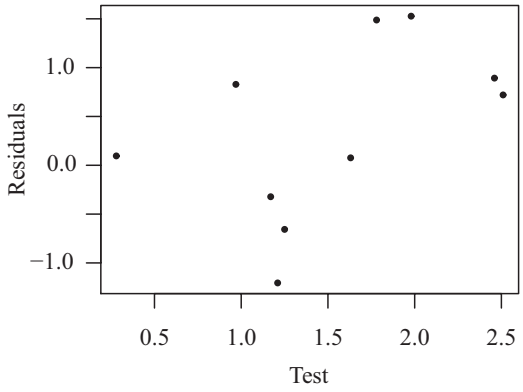

Figure 5.10 Standardized residuals versus test: Model 1, minority only.

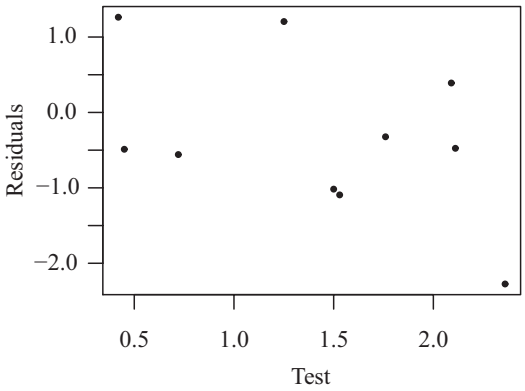

Figure 5.11 Standardized residuals versus test: Model 1, white only.

Our previous conclusion is still valid. The two regression equations are different. Not only are the regression coefficients different, but the residual mean squares also show slight differences. Of more importance, the values of R^2 are greatly different. For the white sample, $R^2 = 0.29$ is so small ($t = 1.82; 2.306$ is required for significance) that the preemployment test score is not deemed an adequate predictor of job success. This finding has bearing on our original objective since it should be a prerequisite for comparing regressions in two samples that the relationships be valid in each of the samples when taken alone. Concerning the validity of the preemployment test, we conclude that if applied as the law prescribes, with indifference to race, it will give biased results for both racial groups. Moreover, based on these findings, we may be justified in saying that the test is of no value for screening white applicants.

We close the discussion with a note about determining the appropriate cutoff test score if the test were used. Consider the results for the minority sample. If Y_m is designated as the minimum acceptable job performance value to be considered successful, then from the regression equation (also see Figure 5.6)

$$X_m = \frac{Y_m - \hat{\beta}_0}{\hat{\beta}_1} ,$$

where $\hat{\beta}_0$ and $\hat{\beta}_1$ are the estimated regression coefficients. X_m is an estimate of the minimum acceptable test score required to attain Y_m. Since X_m is defined in terms of quantities with sampling variation, X_m is also subject to sampling variation. The variation is most easily summarized by constructing a confidence interval for X_m. An approximate 95% level confidence interval takes the form (Scheffé, 1959, p. 52)

$$X_m \pm \frac{t_{(n-2,\alpha/2)}(\hat{\sigma}/n)}{\hat{\beta}_1} ,$$

where $t_{(n-2,\alpha/2)}$ is the appropriate percentile point of the t-distribution and $\hat{\sigma}^2$ is the least squares estimate of σ^2. If Y_m is set at 4, then $X_m = (4 - 0.10)/3.31 = 1.18$ and a 95% confidence interval for the test cutoff score is (1.09, 1.27).

5.4.2 Models with Same Slope and Different Intercepts

In the previous subsection we dealt with the case where the two groups have distinct models with different sets of coefficients as given by Models 1 and 2 in (5.3) and as depicted in Figure 5.6. Suppose now that there is a reason to believe that the two groups have the same slope, β_1, and we wish to test the hypothesis that the two groups also have the same intercept, that is, $H_0 : \beta_{01} = \beta_{02}$. In this case we compare

$$\begin{aligned}
&\text{Model 1 (Pooled):} && y_{ij} = \beta_0 + \beta_1 x_{ij} + \varepsilon_{ij}, \quad j = 1, 2; \quad i = 1, 2, \cdots, n_j, \\
&\text{Model 2 (Minority):} && y_{i1} = \beta_{01} + \beta_1 x_{i1} + \varepsilon_{i1}, \\
&\text{Model 2 (White):} && y_{i2} = \beta_{02} + \beta_1 x_{i2} + \varepsilon_{i2}.
\end{aligned} \tag{5.5}$$

Notice that the two models have the same value of the slope β_1 but different values of the intercepts β_{01} and β_{02}. Using the indicator variable Z defined earlier, we can write Model 2 as

$$\text{Model 3:} \quad y_{ij} = \beta_0 + \beta_1 x_{ij} + \gamma z_{ij} + \varepsilon_{ij}. \tag{5.6}$$

Note the absence of the interaction variable $(z_{ij} \cdot x_{ij})$ from Model 3 in (5.6). If it is present, as it is in (5.4), the two groups would have two models with different slopes and different intercepts.

The equivalence of Models 2 and 3 can be seen by noting that for the minority group, where $x_{ij} = x_{i1}$ and $z_{ij} = 1$, Model 3 becomes

$$\begin{aligned} y_{i1} &= \beta_0 + \beta_1 x_{i1} + \gamma + \varepsilon_{i1} \\ &= (\beta_0 + \gamma) + \beta_1 x_{i1} + \varepsilon_{i1} \\ &= \beta_{01} + \beta_1 x_{i1} + \varepsilon_{i1}, \end{aligned}$$

which is the same as Model 2 for minority with $\beta_{01} = \beta_0 + \gamma$. Similarly, Model 3 for the white group becomes

$$y_{i2} = \beta_0 + \beta_1 x_{i2} + \varepsilon_{i2}.$$

Thus, Model 2 (or equivalently, Model 3) represents two parallel lines[4] (same slope) with intercepts $\beta_0 + \gamma$ and β_0. Therefore, our null hypothesis implies a restriction on γ in Model 3, namely, $\text{H}_0 : \gamma = 0$. To test this hypothesis, we use the F-Test

$$F = \frac{[\text{SSE(RM)} - \text{SSE(FM)}]/1}{\text{SSE(FM)}/17},$$

which has 1 and 17 degrees of freedom. Equivalently, we can use the t-Test for testing $\gamma = 0$ in Model 3, which is

$$t = \frac{\hat{\gamma}}{\text{s.e.}(\hat{\gamma})},$$

which has 17 degrees of freedom. Again, the validation of the assumptions of Model 3 should be done before any conclusions are drawn from these tests. For the current example, we leave the computations of the above tests and the conclusions based on them, as an exercise for the reader.

5.4.3 Models with Same Intercept and Different Slopes

Now we deal with the third case where the two groups have the same intercept, β_0, and we wish to test the hypothesis that the two groups also have the same slope,

[4] In the general case where the model contains $X_1, X_2, \cdots, X_p$ plus one indicator variable Z, Model 3 represents two parallel (hyper-) planes that differ only in the intercept.

that is, $H_0 : \beta_{11} = \beta_{12}$. In this case we compare

$$\begin{aligned} &\text{Model 1 (Pooled):} && y_{ij} = \beta_0 + \beta_1\ x_{ij} + \varepsilon_{ij},\ \ j = 1, 2;\ \ i = 1, 2, \cdots, n_j, \\ &\text{Model 2 (Minority):} && y_{i1} = \beta_0 + \beta_{11} x_{i1} + \varepsilon_{i1}, \\ &\text{Model 2 (White):} && y_{i2} = \beta_0 + \beta_{12} x_{i2} + \varepsilon_{i2}. \end{aligned} \tag{5.7}$$

Note that the two models have the same value of the intercept β_0 but different values of the slopes β_{11} and β_{12}. Using the indicator variable Z defined earlier, we can write Model 2 as

$$\text{Model 3:} \quad y_{ij} = \beta_0 + \beta_1 x_{ij} + \delta(z_{ij} \cdot x_{ij}) + \varepsilon_{ij}. \tag{5.8}$$

Observe the presence of the interaction variable $(z_{ij} \cdot x_{ij})$ but the absence of the individual contribution of the variable Z. The equivalence of Models 2 and 3 can be seen by observing that for the minority group, where $x_{ij} = x_{i1}$ and $z_{ij} = 1$, Model 3 becomes

$$\begin{aligned} y_{i1} &= \beta_0 + \beta_1 x_{i1} + \delta x_{i1} + \varepsilon_{i1} \\ &= \beta_0 + (\beta_1 + \delta) x_{i1} + \varepsilon_{i1} \\ &= \beta_0 + \beta_{11} x_{i1} + \varepsilon_{i1}, \end{aligned}$$

which is the same as Model 2 for minority with $\beta_{11} = \beta_1 + \delta$. Similarly, Model 3 for the white group becomes

$$y_{i2} = \beta_0 + \beta_{12} x_{i2} + \varepsilon_{i2}.$$

Therefore, our null hypothesis implies a restriction on δ in Model 3, namely, $H_0 : \delta = 0$. To test this hypothesis, we use the F-Test

$$F = \frac{[\text{SSE(RM)} - \text{SSE(FM)}]/1}{\text{SSE(FM)}/17},$$

which has 1 and 17 degrees of freedom. Equivalently, we can use the t-Test for testing $\delta = 0$ in Model 3, which is

$$t = \frac{\hat{\delta}}{\text{s.e.}(\hat{\delta})},$$

which has 17 degrees of freedom. Validation of the assumptions of Model 3, the computations of the above tests, and the conclusions based on them are left as an exercise for the reader.

5.5 OTHER APPLICATIONS OF INDICATOR VARIABLES

Applications of indicator variables such as those described in Section 5.4 can be extended to cover a variety of problems [see, e.g., Fox (1984), and Kmenta (1986)

for a variety of applications]. Suppose, for example, that we wish to compare the means of $k \geq 2$ populations or groups. The techniques commonly used here is known as the *analysis of variance* (ANOVA). A random sample of size n_j is taken from the jth population, $j = 1, \cdots, k$. We have a total of $n = n_1 + \cdots + n_k$ observations on the response variable. Let y_{ij} be the ith response in the jth sample. Then y_{ij} can be modeled as

$$y_{ij} = \mu_0 + \mu_1 x_{i1} + \cdots + \mu_p x_{ip} + \varepsilon_{ij}. \tag{5.9}$$

In this model there are $p = k - 1$ indicator predictor variables $x_{i1}, \cdots, x_{ip}$. Each variable x_{ij} is 1 if the corresponding response is from population j, and zero otherwise. The population that is left out is usually known as the *control* group. All indicator variables for the control group are equal to zero. Thus, for the control group, (5.9) becomes

$$y_{ij} = \mu_0 + \varepsilon_{ij}. \tag{5.10}$$

In both (5.9) and (5.10), ε_{ij} are random errors assumed to be independent normal variables with zero means and constant variance σ^2. The constant μ_0 represents the mean of the control group and the regression coefficient μ_j can be interpreted as the difference between the means of the control and jth groups. If $\mu_j = 0$, then the means of the control and jth groups are equal. The null hypothesis $\mathrm{H}_0 : \mu_1 = \cdots = \mu_p = 0$ that all groups have the same mean can be represented by the model in (5.10). The alternate hypothesis that at least one of the μ_j's is different from zero can be represented by the model in (5.9). The models in (5.9) and (5.10) can be viewed as full and reduced models, respectively. Hence H_0 can be tested using the F-Test given in (3.45). Thus, the use of indicator variables allowed us to express ANOVA techniques as a special case of regression analysis. Both the number of quantitative predictor variables and the number of distinct groups represented in the data by indicator variables may be increased.

Note that the examples discussed above are based on cross-sectional data. Indicator variables can also be utilized with time series data. In addition, there are some models of growth processes where an indicator variable is used as the dependent variable. These models, known as *logistic regression models*, are discussed in Chapter 12.

In Sections 5.6 and 5.7 we discuss the use of indicator variables with time series data. In particular, notions of *seasonality* and *stability* of parameters over time are discussed. These problems are formulated and the data are provided. The analyses are left to the reader.

5.6 SEASONALITY

The data set we use as an example here, referred to as the Ski Sales data, is shown in Table 5.11 and can be obtained from the book's Website. The data consist of two variables: the sales, S, in millions for a firm that manufactures skis and related

Table 5.11 Disposable Income and Ski Sales for Years 1964–1973

Row	Date	Sales	PDI	Row	Date	Sales	PDI
1	Q1/64	37.0	109	21	Q1/69	44.9	153
2	Q2/64	33.5	115	22	Q2/69	41.6	156
3	Q3/64	30.8	113	23	Q3/69	44.0	160
4	Q4/64	37.9	116	24	Q4/69	48.1	163
5	Q1/65	37.4	118	25	Q1/70	49.7	166
6	Q2/65	31.6	120	26	Q2/70	43.9	171
7	Q3/65	34.0	122	27	Q3/70	41.6	174
8	Q4/65	38.1	124	28	Q4/70	51.0	175
9	Q1/66	40.0	126	29	Q1/71	52.0	180
10	Q2/66	35.0	128	30	Q2/71	46.2	184
11	Q3/66	34.9	130	31	Q3/71	47.1	187
12	Q4/66	40.2	132	32	Q4/71	52.7	189
13	Q1/67	41.9	133	33	Q1/72	52.2	191
14	Q2/67	34.7	135	34	Q2/72	47.0	193
15	Q3/67	38.8	138	35	Q3/72	47.8	194
16	Q4/67	43.7	140	36	Q4/72	52.8	196
17	Q1/68	44.2	143	37	Q1/73	54.1	199
18	Q2/68	40.4	147	38	Q2/73	49.5	201
19	Q3/68	38.4	148	39	Q3/73	49.5	202
20	Q4/68	45.4	151	40	Q4/73	54.3	204

equipment for the years 1964–1973, and personal disposable income, PDI.[5] Each of these variables is measured quarterly. We use these data in Chapter 8 to illustrate the problem of *correlated errors*.

The model is an equation that relates S to PDI, that is, $S_t = \beta_0 + \beta_1 \text{PDI}_t + \varepsilon_t$, where S_t is sales in millions in the tth period and PDI_t is the corresponding personal disposable income. Our approach here is to assume the existence of a seasonal effect on sales that is determined on a quarterly basis. To measure this effect we may define indicator variables to characterize the seasonality. Since we have four quarters, we define three indicator variables, Z_1, Z_2, and Z_3, where

$$z_{t1} = \begin{cases} 1, & \text{if the } t\text{th period is a first quarter,} \\ 0, & \text{otherwise,} \end{cases}$$

$$z_{t2} = \begin{cases} 1, & \text{if the } t\text{th period is a second quarter,} \\ 0, & \text{otherwise,} \end{cases}$$

$$z_{t3} = \begin{cases} 1, & \text{if the } t\text{th period is a third quarter,} \\ 0, & \text{otherwise.} \end{cases}$$

[5] Aggregate measure of purchasing potential.

The analysis and interpretation of this data set are left to the reader. The authors have analyzed these data and found that there are actually only two seasons. (See the discussion of these sales data in Chapter 8 for an analysis using only one indicator variable, two seasons.) See Kmenta (1986) for further discussion on using indicator variables for analyzing seasonality.

5.7 STABILITY OF REGRESSION PARAMETERS OVER TIME

Indicator variables may also be used to analyze the stability of regression coefficients over time or to test for structural change. We consider an extension of the system of regressions problem when data are available on a cross-section of observations and over time. Our objective is to analyze the constancy of the relationships over time. The methods described here are suitable for intertemporal and interspatial comparisons. To outline the method we use the Education Expenditure data shown in Tables 5.12–5.14. The measured variables for the 50 states are:

Y	Per capita expenditure on public education
X_1	Per capita personal income
X_2	Number of residents per thousand under 18 years of age
X_3	Number of people per thousand residing in urban areas

The variable Region is a categorical variable representing geographical regions:

$$1 = \text{Northeast}, \quad 2 = \text{North Central}, \quad 3 = \text{South}, \quad 4 = \text{West}.$$

This data set is used in Chapter 7 to demonstrate methods of dealing with heteroscedasticity in multiple regression and to analyze the effects of regional characteristics on the regression relationships. Here we focus on the stability of the expenditure relationship with respect to time.

Data have been developed on the four variables described above for each state in 1960, 1970, and 1975. Assuming that the relationship can be identically specified in each of the three years,[6] the analysis of stability can be carried out by evaluating the variation in the estimated regression coefficients over time. Working with the pooled data set of 150 observations (50 states each in 3 years) we define two indicator variables, T_1 and T_2, where

$$T_{i1} = \begin{cases} 1, & \text{if the } i\text{th observation was from 1960,} \\ 0, & \text{otherwise,} \end{cases}$$

$$T_{i2} = \begin{cases} 1, & \text{if the } i\text{th observation was from 1970,} \\ 0, & \text{otherwise.} \end{cases}$$

[6] *Specification* as used here means that the same variables appear in each equation. Any transformations that are used apply to each equation. The assumption concerning identical specification should be empirically validated.

Table 5.12 Education Expenditures Data (1960)

Row	STATE	Y	X_1	X_2	X_3	Region
1	ME	61	1704	388	399	1
2	NH	68	1885	372	598	1
3	VT	72	1745	397	370	1
4	MA	72	2394	358	868	1
5	RI	62	1966	357	899	1
6	CT	91	2817	362	690	1
7	NY	104	2685	341	728	1
8	NJ	99	2521	353	826	1
9	PA	70	2127	352	656	1
10	OH	82	2184	387	674	2
11	IN	84	1990	392	568	2
12	IL	84	2435	366	759	2
13	MI	104	2099	403	650	2
14	WI	84	1936	393	621	2
15	MN	103	1916	402	610	2
16	IA	86	1863	385	522	2
17	MO	69	2037	364	613	2
18	ND	94	1697	429	351	2
19	SD	79	1644	411	390	2
20	NB	80	1894	379	520	2
21	KS	98	2001	380	564	2
22	DE	124	2760	388	326	3
23	MD	92	2221	393	562	3
24	VA	67	1674	402	487	3
25	WV	66	1509	405	358	3
26	NC	65	1384	423	362	3
27	SC	57	1218	453	343	3
28	GA	60	1487	420	498	3
29	FL	74	1876	334	628	3
30	KY	49	1397	594	377	3
31	TN	60	1439	346	457	3
32	AL	59	1359	637	517	3
33	MS	68	1053	448	362	3
34	AR	56	1225	403	416	3
35	LA	72	1576	433	562	3
36	OK	80	1740	378	610	3
37	TX	79	1814	409	727	3
38	MT	95	1920	412	463	4
39	ID	79	1701	418	414	4
40	WY	142	2088	415	568	4
41	CO	108	2047	399	621	4
42	NM	94	1838	458	618	4
43	AZ	107	1932	425	699	4
44	UT	109	1753	494	665	4
45	NV	114	2569	372	663	4
46	WA	112	2160	386	584	4
47	OR	105	2006	382	534	4
48	CA	129	2557	373	717	4
49	AK	107	1900	434	379	4
50	HI	77	1852	431	693	4

Table 5.13 Education Expenditures Data (1970)

Row	STATE	Y	X_1	X_2	X_3	Region
1	ME	189	2828	351	508	1
2	NH	169	3259	346	564	1
3	VT	230	3072	348	322	1
4	MA	168	3835	335	846	1
5	RI	180	3549	327	871	1
6	CT	193	4256	341	774	1
7	NY	261	4151	326	856	1
8	NJ	214	3954	333	889	1
9	PA	201	3419	326	715	1
10	OH	172	3509	354	753	2
11	IN	194	3412	359	649	2
12	IL	189	3981	349	830	2
13	MI	233	3675	369	738	2
14	WI	209	3363	361	659	2
15	MN	262	3341	365	664	2
16	IA	234	3265	344	572	2
17	MO	177	3257	336	701	2
18	ND	177	2730	369	443	2
19	SD	187	2876	369	446	2
20	NB	148	3239	350	615	2
21	KS	196	3303	340	661	2
22	DE	248	3795	376	722	3
23	MD	247	3742	364	766	3
24	VA	180	3068	353	631	3
25	WV	149	2470	329	390	3
26	NC	155	2664	354	450	3
27	SC	149	2380	377	476	3
28	GA	156	2781	371	603	3
29	FL	191	3191	336	805	3
30	KY	140	2645	349	523	3
31	TN	137	2579	343	588	3
32	AL	112	2337	362	584	3
33	MS	130	2081	385	445	3
34	AR	134	2322	352	500	3
35	LA	162	2634	390	661	3
36	OK	135	2880	330	680	3
37	TX	155	3029	369	797	3
38	MT	238	2942	369	534	4
39	ID	170	2668	368	541	4
40	WY	238	3190	366	605	4
41	CO	192	3340	358	785	4
42	NM	227	2651	421	698	4
43	AZ	207	3027	387	796	4
44	UT	201	2790	412	804	4
45	NV	225	3957	385	809	4
46	WA	215	3688	342	726	4
47	OR	233	3317	333	671	4
48	CA	273	3968	348	909	4
49	AK	372	4146	440	484	4
50	HI	212	3513	383	831	4

Table 5.14 Education Expenditures Data (1975)

Row	STATE	Y	X_1	X_2	X_3	Region
1	ME	235	3944	325	508	1
2	NH	231	4578	323	564	1
3	VT	270	4011	328	322	1
4	MA	261	5233	305	846	1
5	RI	300	4780	303	871	1
6	CT	317	5889	307	774	1
7	NY	387	5663	301	856	1
8	NJ	285	5759	310	889	1
9	PA	300	4894	300	715	1
10	OH	221	5012	324	753	2
11	IN	264	4908	329	649	2
12	IL	308	5753	320	830	2
13	MI	379	5439	337	738	2
14	WI	342	4634	328	659	2
15	MN	378	4921	330	664	2
16	IA	232	4869	318	572	2
17	MO	231	4672	309	701	2
18	ND	246	4782	333	443	2
19	SD	230	4296	330	446	2
20	NB	268	4827	318	615	2
21	KS	337	5057	304	661	2
22	DE	344	5540	328	722	3
23	MD	330	5331	323	766	3
24	VA	261	4715	317	631	3
25	WV	214	3828	310	390	3
26	NC	245	4120	321	450	3
27	SC	233	3817	342	476	3
28	GA	250	4243	339	603	3
29	FL	243	4647	287	805	3
30	KY	216	3967	325	523	3
31	TN	212	3946	315	588	3
32	AL	208	3724	332	584	3
33	MS	215	3448	358	445	3
34	AR	221	3680	320	500	3
35	LA	244	3825	355	661	3
36	OK	234	4189	306	680	3
37	TX	269	4336	335	797	3
38	MT	302	4418	335	534	4
39	ID	268	4323	344	541	4
40	WY	323	4813	331	605	4
41	CO	304	5046	324	785	4
42	NM	317	3764	366	698	4
43	AZ	332	4504	340	796	4
44	UT	315	4005	378	804	4
45	NV	291	5560	330	809	4
46	WA	312	4989	313	726	4
47	OR	316	4697	305	671	4
48	CA	332	5438	307	909	4
49	AK	546	5613	386	484	4
50	HI	311	5309	333	831	4

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.

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%