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.
Table 5.1  Salary Survey Data

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

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, \( E_3 \), defined to be 1 or 0 depending on whether or not the \( i\)th 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

<table>
<thead>
<tr>
<th>Category</th>
<th>E</th>
<th>M</th>
<th>Regression Equation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>$S = \left( \beta_0 + \gamma_1 \right) + \beta_1 X + \varepsilon$</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>1</td>
<td>$S = \left( \beta_0 + \gamma_1 + \delta_1 \right) + \beta_1 X + \varepsilon$</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>0</td>
<td>$S = \left( \beta_0 + \gamma_2 \right) + \beta_1 X + \varepsilon$</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>1</td>
<td>$S = \left( \beta_0 + \gamma_2 + \delta_1 \right) + \beta_1 X + \varepsilon$</td>
</tr>
<tr>
<td>5</td>
<td>3</td>
<td>0</td>
<td>$S = \beta_0 + \delta_1 + \beta_1 X + \varepsilon$</td>
</tr>
<tr>
<td>6</td>
<td>3</td>
<td>1</td>
<td>$S = \left( \beta_0 + \delta_1 \right) + \beta_1 X + \varepsilon$</td>
</tr>
</tbody>
</table>

Table 5.3  Regression Analysis of Salary Survey Data

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient</th>
<th>s.e.</th>
<th>t-Test</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>11031.800</td>
<td>383.2</td>
<td>28.80</td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td>$X$</td>
<td>546.184</td>
<td>30.5</td>
<td>17.90</td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td>$E_1$</td>
<td>-2996.210</td>
<td>411.8</td>
<td>-7.28</td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td>$E_2$</td>
<td>147.825</td>
<td>387.7</td>
<td>0.38</td>
<td>0.7049</td>
</tr>
<tr>
<td>$M$</td>
<td>6883.530</td>
<td>313.9</td>
<td>21.90</td>
<td>&lt; 0.0001</td>
</tr>
</tbody>
</table>

$n = 46 \quad R^2 = 0.957 \quad R^2_a = 0.953 \quad \hat{\sigma} = 1027 \quad df = 41$

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

$$S = \beta_0 + \beta_1 X + \gamma_1 E_1 + \gamma_2 E_2 + \delta_1 M + \varepsilon.$$  (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, $\delta_1$, 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, $\gamma_1$ measures the salary differential for the H.S. category relative to the advanced degree category and $\gamma_2$ 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 categorial 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 $\delta_1$, 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

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient</th>
<th>s.e.</th>
<th>( t )-Test</th>
<th>( p )-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>11203.40</td>
<td>79.07</td>
<td>141.7</td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td>( X )</td>
<td>496.99</td>
<td>5.57</td>
<td>89.3</td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td>( E_1 )</td>
<td>-1730.75</td>
<td>105.30</td>
<td>-16.4</td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td>( E_2 )</td>
<td>-349.08</td>
<td>97.57</td>
<td>-3.6</td>
<td>0.0009</td>
</tr>
<tr>
<td>( M )</td>
<td>7047.41</td>
<td>102.60</td>
<td>68.7</td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td>( E_1 \cdot M )</td>
<td>-3066.04</td>
<td>149.30</td>
<td>-20.5</td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td>( E_2 \cdot M )</td>
<td>1836.49</td>
<td>131.20</td>
<td>14.0</td>
<td>&lt; 0.0001</td>
</tr>
</tbody>
</table>

\( n = 46 \)  \( R^2 = 0.999 \)  \( R^2_a = 0.999 \)  \( \hat{\sigma} = 173.8 \)  \( \text{df} = 39 \)

The expanded model is
\[
S = \beta_0 + \beta_1 X + \gamma_1 E_1 + \gamma_2 E_2 + \delta_1 M \\
+ \alpha_1 (E_1 \cdot M) + \alpha_2 (E_2 \cdot M) + \varepsilon. \tag{5.2}
\]

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)

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

$^a$ 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.
Let \( Y \) represent job performance and let \( X \) be the score on the preemployment test. We want to compare

Model 1 (Pooled): \[ y_{ij} = \beta_0 + \beta_1 x_{ij} + \varepsilon_{ij}, \quad j = 1, 2; \quad i = 1, 2, \ldots, n_j, \]

Model 2 (Minority): \[ y_{i1} = \beta_{01} + \beta_{11} x_{i1} + \varepsilon_{i1}, \quad (5.3) \]

Model 2 (White): \[ y_{i2} = \beta_{02} + \beta_{12} x_{i2} + \varepsilon_{i2}. \]

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

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

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 $H_0: \beta_{11} = \beta_{12}, \beta_{01} = \beta_{02}$ 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

Model 1: \[ y_{ij} = \beta_0 + \beta_1 x_{ij} + \varepsilon_{ij} \]

Model 3: \[ y_{ij} = \beta_0 + \beta_1 x_{ij} + \gamma z_{ij} + \delta (z_{ij} \cdot x_{ij}) + \varepsilon_{ij}. \] (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

\[ y_{i1} = \beta_0 + \beta_1 x_{i1} + \gamma + \delta x_{i1} + \varepsilon_{i1} \]
\[ = (\beta_0 + \gamma) + (\beta_1 + \delta) x_{i1} + \varepsilon_{i1} \]
\[ = \beta_{01} + \beta_{11} x_{i1} + \varepsilon_{i1}, \]

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

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient</th>
<th>s.e.</th>
<th>t-Test</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>1.03</td>
<td>0.87</td>
<td>1.19</td>
<td>0.2486</td>
</tr>
<tr>
<td>TEST (X)</td>
<td>2.36</td>
<td>0.54</td>
<td>4.39</td>
<td>0.0004</td>
</tr>
</tbody>
</table>

\[ n = 20 \quad R^2 = 0.52 \quad R_a^2 = 0.49 \quad \sigma = 1.59 \quad df = 18 \]

Table 5.9  Regression Results, Preemployment Testing Data: Model 3

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient</th>
<th>s.e.</th>
<th>t-Test</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>2.01</td>
<td>1.05</td>
<td>1.91</td>
<td>0.0736</td>
</tr>
<tr>
<td>TEST (X)</td>
<td>1.31</td>
<td>0.67</td>
<td>1.96</td>
<td>0.0677</td>
</tr>
<tr>
<td>RACE (Z)</td>
<td>−1.91</td>
<td>1.54</td>
<td>−1.24</td>
<td>0.2321</td>
</tr>
<tr>
<td>RACE · TEST (X · Z)</td>
<td>2.00</td>
<td>0.95</td>
<td>2.09</td>
<td>0.0527</td>
</tr>
</tbody>
</table>

\[ n = 20 \quad R^2 = 0.664 \quad R_a^2 = 0.601 \quad \sigma = 1.41 \quad df = 16 \]

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

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

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

<table>
<thead>
<tr>
<th>Sample</th>
<th>$\hat{\beta}_0$</th>
<th>$\hat{\beta}_1$</th>
<th>$t_1$</th>
<th>$R^2$</th>
<th>$\hat{\sigma}$</th>
<th>df</th>
</tr>
</thead>
<tbody>
<tr>
<td>Minority</td>
<td>0.10</td>
<td>3.31</td>
<td>5.31</td>
<td>0.78</td>
<td>1.29</td>
<td>8</td>
</tr>
<tr>
<td>White</td>
<td>2.01</td>
<td>1.31</td>
<td>1.82</td>
<td>0.29</td>
<td>1.51</td>
<td>8</td>
</tr>
</tbody>
</table>

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)

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

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 $\sigma^2$. 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, $\beta_1$, 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

Model 1 (Pooled): \hspace{1cm} \begin{align*}
    y_{ij} &= \beta_0 + \beta_1 x_{ij} + \epsilon_{ij}, \quad j = 1, 2, \ldots, n_j, \\
end{align*}

Model 2 (Minority): \hspace{1cm} \begin{align*}
    y_{i1} &= \beta_{01} + \beta_1 x_{i1} + \epsilon_{i1}, \\
end{align*}

Model 2 (White): \hspace{1cm} \begin{align*}
    y_{i2} &= \beta_{02} + \beta_1 x_{i2} + \epsilon_{i2}. \\
end{align*}

\text{(5.5)}
Notice that the two models have the same value of the slope $\beta_1$ but different values of the intercepts $\beta_{01}$ and $\beta_{02}$. Using the indicator variable $Z$ defined earlier, we can write Model 2 as

$$\text{Model 2: } y_{ij} = \beta_0 + \beta_1 x_{ij} + \gamma z_{ij} + \varepsilon_{ij}. \quad (5.6)$$

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

\[
y_{i1} = \beta_0 + \beta_1 x_{i1} + \gamma + \varepsilon_{i1} = (\beta_0 + \gamma) + \beta_1 x_{i1} + \varepsilon_{i1} = \beta_{01} + \beta_1 x_{i1} + \varepsilon_{i1},
\]

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 $\beta_0$. Therefore, our null hypothesis implies a restriction on $\gamma$ in Model 3, namely, $H_0 : \gamma = 0$. 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, $\beta_0$, and we wish to test the hypothesis that the two groups also have the same slope,\(^5\)

\(^4\) 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

Model 1 (Pooled): \[ y_{ij} = \beta_0 + \beta_1 x_{ij} + \varepsilon_{ij}, \quad j = 1, 2; \quad i = 1, 2, \ldots, n_j, \]

Model 2 (Minority): \[ y_{i1} = \beta_0 + \beta_{11} x_{i1} + \varepsilon_{i1}, \quad (5.7) \]

Model 2 (White): \[ y_{i2} = \beta_0 + \beta_{12} x_{i2} + \varepsilon_{i2}. \]

Note that the two models have the same value of the intercept $\beta_0$ but different values of the slopes $\beta_{11}$ and $\beta_{12}$. Using the indicator variable $Z$ defined earlier, we can write Model 2 as

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

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

\[
y_{i1} = \beta_0 + \beta_1 x_{i1} + \delta x_{i1} + \varepsilon_{i1} \\
= \beta_0 + (\beta_1 + \delta) x_{i1} + \varepsilon_{i1} \\
= \beta_0 + \beta_{11} x_{i1} + \varepsilon_{i1},
\]

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 $\delta$ 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 \( j \)th 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 \( i \)th response in the \( j \)th sample. Then \( y_{ij} \) can be modeled as

\[
y_{ij} = \mu_0 + \mu_1 x_{i1} + \cdots + \mu_p x_{ip} + \varepsilon_{ij}.
\] (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

\[
y_{ij} = \mu_0 + \varepsilon_{ij}.
\] (5.10)

In both (5.9) and (5.10), \( \varepsilon_{ij} \) are random errors assumed to be independent normal variables with zero means and constant variance \( \sigma^2 \). The constant \( \mu_0 \) represents the mean of the control group and the regression coefficient \( \mu_j \) can be interpreted as the difference between the means of the control and \( j \)th groups. If \( \mu_j = 0 \), then the means of the control and \( j \)th groups are equal. The null hypothesis \( H_0 : \mu_1 = \cdots = \mu_p = 0 \) 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 \( \mu_j \)'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

<table>
<thead>
<tr>
<th>Row</th>
<th>Date</th>
<th>Sales</th>
<th>PDI</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Q1/64</td>
<td>37.0</td>
<td>109</td>
</tr>
<tr>
<td>2</td>
<td>Q2/64</td>
<td>33.5</td>
<td>115</td>
</tr>
<tr>
<td>3</td>
<td>Q3/64</td>
<td>30.8</td>
<td>113</td>
</tr>
<tr>
<td>4</td>
<td>Q4/64</td>
<td>37.9</td>
<td>116</td>
</tr>
<tr>
<td>5</td>
<td>Q1/65</td>
<td>37.4</td>
<td>118</td>
</tr>
<tr>
<td>6</td>
<td>Q2/65</td>
<td>31.6</td>
<td>120</td>
</tr>
<tr>
<td>7</td>
<td>Q3/65</td>
<td>34.0</td>
<td>122</td>
</tr>
<tr>
<td>8</td>
<td>Q4/65</td>
<td>38.1</td>
<td>124</td>
</tr>
<tr>
<td>9</td>
<td>Q1/66</td>
<td>40.0</td>
<td>126</td>
</tr>
<tr>
<td>10</td>
<td>Q2/66</td>
<td>35.0</td>
<td>128</td>
</tr>
<tr>
<td>11</td>
<td>Q3/66</td>
<td>34.9</td>
<td>130</td>
</tr>
<tr>
<td>12</td>
<td>Q4/66</td>
<td>40.2</td>
<td>132</td>
</tr>
<tr>
<td>13</td>
<td>Q1/67</td>
<td>41.9</td>
<td>133</td>
</tr>
<tr>
<td>14</td>
<td>Q2/67</td>
<td>34.7</td>
<td>135</td>
</tr>
<tr>
<td>15</td>
<td>Q3/67</td>
<td>38.8</td>
<td>138</td>
</tr>
<tr>
<td>16</td>
<td>Q4/67</td>
<td>43.7</td>
<td>140</td>
</tr>
<tr>
<td>17</td>
<td>Q1/68</td>
<td>44.2</td>
<td>143</td>
</tr>
<tr>
<td>18</td>
<td>Q2/68</td>
<td>40.4</td>
<td>147</td>
</tr>
<tr>
<td>19</td>
<td>Q3/68</td>
<td>38.4</td>
<td>148</td>
</tr>
<tr>
<td>20</td>
<td>Q4/68</td>
<td>45.4</td>
<td>151</td>
</tr>
</tbody>
</table>

Aggregate measure of purchasing potential.

The model is an equation that relates $S_t$ to PDI, that is, $S_t = \beta_0 + \beta_1 PDI_t + \epsilon_t$, where $S_t$ is sales in millions in the $t$th 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

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

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

$$z_{t3} = \begin{cases} 1, & \text{if the } t\text{th period is a third quarter,} \\ 0, & \text{otherwise.} \end{cases}$$
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 = Northeast, 2 = North Central, 3 = South, 4 = 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, 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

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

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

* 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)

<table>
<thead>
<tr>
<th>Row</th>
<th>STATE</th>
<th>Y</th>
<th>X_1</th>
<th>X_2</th>
<th>X_3</th>
<th>Region</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>ME</td>
<td>61</td>
<td>1704</td>
<td>388</td>
<td>399</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>NH</td>
<td>68</td>
<td>1885</td>
<td>372</td>
<td>598</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>VT</td>
<td>72</td>
<td>1745</td>
<td>397</td>
<td>370</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>MA</td>
<td>72</td>
<td>2394</td>
<td>358</td>
<td>868</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>RI</td>
<td>62</td>
<td>1966</td>
<td>357</td>
<td>899</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>CT</td>
<td>91</td>
<td>2817</td>
<td>362</td>
<td>690</td>
<td>1</td>
</tr>
<tr>
<td>7</td>
<td>NY</td>
<td>104</td>
<td>2685</td>
<td>341</td>
<td>728</td>
<td>1</td>
</tr>
<tr>
<td>8</td>
<td>NJ</td>
<td>99</td>
<td>2521</td>
<td>353</td>
<td>826</td>
<td>1</td>
</tr>
<tr>
<td>9</td>
<td>PA</td>
<td>70</td>
<td>2127</td>
<td>352</td>
<td>656</td>
<td>1</td>
</tr>
<tr>
<td>10</td>
<td>OH</td>
<td>82</td>
<td>2184</td>
<td>387</td>
<td>674</td>
<td>2</td>
</tr>
<tr>
<td>11</td>
<td>IN</td>
<td>84</td>
<td>1990</td>
<td>392</td>
<td>568</td>
<td>2</td>
</tr>
<tr>
<td>12</td>
<td>IL</td>
<td>84</td>
<td>2435</td>
<td>366</td>
<td>759</td>
<td>2</td>
</tr>
<tr>
<td>13</td>
<td>MI</td>
<td>104</td>
<td>2099</td>
<td>403</td>
<td>650</td>
<td>2</td>
</tr>
<tr>
<td>14</td>
<td>WI</td>
<td>84</td>
<td>1936</td>
<td>393</td>
<td>621</td>
<td>2</td>
</tr>
<tr>
<td>15</td>
<td>MN</td>
<td>103</td>
<td>1916</td>
<td>402</td>
<td>610</td>
<td>2</td>
</tr>
<tr>
<td>16</td>
<td>IA</td>
<td>86</td>
<td>1863</td>
<td>385</td>
<td>522</td>
<td>2</td>
</tr>
<tr>
<td>17</td>
<td>MO</td>
<td>69</td>
<td>2037</td>
<td>364</td>
<td>613</td>
<td>2</td>
</tr>
<tr>
<td>18</td>
<td>ND</td>
<td>94</td>
<td>1697</td>
<td>429</td>
<td>351</td>
<td>2</td>
</tr>
<tr>
<td>19</td>
<td>SD</td>
<td>79</td>
<td>1644</td>
<td>411</td>
<td>390</td>
<td>2</td>
</tr>
<tr>
<td>20</td>
<td>NB</td>
<td>80</td>
<td>1894</td>
<td>379</td>
<td>520</td>
<td>2</td>
</tr>
<tr>
<td>21</td>
<td>KS</td>
<td>98</td>
<td>2001</td>
<td>380</td>
<td>564</td>
<td>2</td>
</tr>
<tr>
<td>22</td>
<td>DE</td>
<td>124</td>
<td>2760</td>
<td>388</td>
<td>326</td>
<td>3</td>
</tr>
<tr>
<td>23</td>
<td>MD</td>
<td>92</td>
<td>2221</td>
<td>393</td>
<td>562</td>
<td>3</td>
</tr>
<tr>
<td>24</td>
<td>VA</td>
<td>67</td>
<td>1674</td>
<td>402</td>
<td>487</td>
<td>3</td>
</tr>
<tr>
<td>25</td>
<td>WV</td>
<td>66</td>
<td>1509</td>
<td>405</td>
<td>358</td>
<td>3</td>
</tr>
<tr>
<td>26</td>
<td>NC</td>
<td>65</td>
<td>1384</td>
<td>423</td>
<td>362</td>
<td>3</td>
</tr>
<tr>
<td>27</td>
<td>SC</td>
<td>57</td>
<td>1218</td>
<td>453</td>
<td>343</td>
<td>3</td>
</tr>
<tr>
<td>28</td>
<td>GA</td>
<td>60</td>
<td>1487</td>
<td>420</td>
<td>498</td>
<td>3</td>
</tr>
<tr>
<td>29</td>
<td>FL</td>
<td>74</td>
<td>1876</td>
<td>334</td>
<td>628</td>
<td>3</td>
</tr>
<tr>
<td>30</td>
<td>KY</td>
<td>49</td>
<td>1397</td>
<td>594</td>
<td>377</td>
<td>3</td>
</tr>
<tr>
<td>31</td>
<td>TN</td>
<td>60</td>
<td>1439</td>
<td>346</td>
<td>457</td>
<td>3</td>
</tr>
<tr>
<td>32</td>
<td>AL</td>
<td>59</td>
<td>1359</td>
<td>637</td>
<td>517</td>
<td>3</td>
</tr>
<tr>
<td>33</td>
<td>MS</td>
<td>68</td>
<td>1053</td>
<td>448</td>
<td>362</td>
<td>3</td>
</tr>
<tr>
<td>34</td>
<td>AR</td>
<td>56</td>
<td>1225</td>
<td>403</td>
<td>416</td>
<td>3</td>
</tr>
<tr>
<td>35</td>
<td>LA</td>
<td>72</td>
<td>1576</td>
<td>433</td>
<td>562</td>
<td>3</td>
</tr>
<tr>
<td>36</td>
<td>OK</td>
<td>80</td>
<td>1740</td>
<td>378</td>
<td>610</td>
<td>3</td>
</tr>
<tr>
<td>37</td>
<td>TX</td>
<td>79</td>
<td>1814</td>
<td>409</td>
<td>727</td>
<td>3</td>
</tr>
<tr>
<td>38</td>
<td>MT</td>
<td>95</td>
<td>1920</td>
<td>412</td>
<td>463</td>
<td>4</td>
</tr>
<tr>
<td>39</td>
<td>ID</td>
<td>79</td>
<td>1701</td>
<td>418</td>
<td>414</td>
<td>4</td>
</tr>
<tr>
<td>40</td>
<td>WY</td>
<td>142</td>
<td>2088</td>
<td>415</td>
<td>568</td>
<td>4</td>
</tr>
<tr>
<td>41</td>
<td>CO</td>
<td>108</td>
<td>2047</td>
<td>399</td>
<td>621</td>
<td>4</td>
</tr>
<tr>
<td>42</td>
<td>NM</td>
<td>94</td>
<td>1838</td>
<td>458</td>
<td>618</td>
<td>4</td>
</tr>
<tr>
<td>43</td>
<td>AZ</td>
<td>107</td>
<td>1932</td>
<td>425</td>
<td>699</td>
<td>4</td>
</tr>
<tr>
<td>44</td>
<td>UT</td>
<td>109</td>
<td>1753</td>
<td>494</td>
<td>665</td>
<td>4</td>
</tr>
<tr>
<td>45</td>
<td>NV</td>
<td>114</td>
<td>2569</td>
<td>372</td>
<td>663</td>
<td>4</td>
</tr>
<tr>
<td>46</td>
<td>WA</td>
<td>112</td>
<td>2160</td>
<td>386</td>
<td>584</td>
<td>4</td>
</tr>
<tr>
<td>47</td>
<td>OR</td>
<td>105</td>
<td>2006</td>
<td>382</td>
<td>534</td>
<td>4</td>
</tr>
<tr>
<td>48</td>
<td>CA</td>
<td>129</td>
<td>2557</td>
<td>373</td>
<td>717</td>
<td>4</td>
</tr>
<tr>
<td>49</td>
<td>AK</td>
<td>107</td>
<td>1900</td>
<td>434</td>
<td>379</td>
<td>4</td>
</tr>
<tr>
<td>50</td>
<td>HI</td>
<td>77</td>
<td>1852</td>
<td>431</td>
<td>693</td>
<td>4</td>
</tr>
</tbody>
</table>
### Table 5.13  Education Expenditures Data (1970)

<table>
<thead>
<tr>
<th>Row</th>
<th>STATE</th>
<th>Y</th>
<th>X₁</th>
<th>X₂</th>
<th>X₃</th>
<th>Region</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>ME</td>
<td>189</td>
<td>2828</td>
<td>351</td>
<td>508</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>NH</td>
<td>169</td>
<td>3259</td>
<td>346</td>
<td>564</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>VT</td>
<td>230</td>
<td>3072</td>
<td>348</td>
<td>322</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>MA</td>
<td>168</td>
<td>3835</td>
<td>335</td>
<td>846</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>RI</td>
<td>180</td>
<td>3549</td>
<td>327</td>
<td>871</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>CT</td>
<td>193</td>
<td>4256</td>
<td>341</td>
<td>774</td>
<td>1</td>
</tr>
<tr>
<td>7</td>
<td>NY</td>
<td>261</td>
<td>4151</td>
<td>326</td>
<td>856</td>
<td>1</td>
</tr>
<tr>
<td>8</td>
<td>NJ</td>
<td>214</td>
<td>3954</td>
<td>333</td>
<td>889</td>
<td>1</td>
</tr>
<tr>
<td>9</td>
<td>PA</td>
<td>201</td>
<td>3419</td>
<td>326</td>
<td>715</td>
<td>1</td>
</tr>
<tr>
<td>10</td>
<td>OH</td>
<td>172</td>
<td>3509</td>
<td>354</td>
<td>753</td>
<td>2</td>
</tr>
<tr>
<td>11</td>
<td>IN</td>
<td>194</td>
<td>3412</td>
<td>359</td>
<td>649</td>
<td>2</td>
</tr>
<tr>
<td>12</td>
<td>IL</td>
<td>189</td>
<td>3981</td>
<td>349</td>
<td>830</td>
<td>2</td>
</tr>
<tr>
<td>13</td>
<td>MI</td>
<td>233</td>
<td>3675</td>
<td>369</td>
<td>738</td>
<td>2</td>
</tr>
<tr>
<td>14</td>
<td>WI</td>
<td>209</td>
<td>3363</td>
<td>361</td>
<td>659</td>
<td>2</td>
</tr>
<tr>
<td>15</td>
<td>MN</td>
<td>262</td>
<td>3341</td>
<td>365</td>
<td>664</td>
<td>2</td>
</tr>
<tr>
<td>16</td>
<td>IA</td>
<td>234</td>
<td>3265</td>
<td>344</td>
<td>572</td>
<td>2</td>
</tr>
<tr>
<td>17</td>
<td>MO</td>
<td>177</td>
<td>3257</td>
<td>336</td>
<td>701</td>
<td>2</td>
</tr>
<tr>
<td>18</td>
<td>ND</td>
<td>177</td>
<td>2730</td>
<td>369</td>
<td>443</td>
<td>2</td>
</tr>
<tr>
<td>19</td>
<td>SD</td>
<td>187</td>
<td>2876</td>
<td>369</td>
<td>446</td>
<td>2</td>
</tr>
<tr>
<td>20</td>
<td>NB</td>
<td>148</td>
<td>3239</td>
<td>350</td>
<td>615</td>
<td>2</td>
</tr>
<tr>
<td>21</td>
<td>KS</td>
<td>196</td>
<td>3303</td>
<td>340</td>
<td>661</td>
<td>2</td>
</tr>
<tr>
<td>22</td>
<td>DE</td>
<td>248</td>
<td>3795</td>
<td>376</td>
<td>722</td>
<td>3</td>
</tr>
<tr>
<td>23</td>
<td>MD</td>
<td>247</td>
<td>3742</td>
<td>364</td>
<td>766</td>
<td>3</td>
</tr>
<tr>
<td>24</td>
<td>VA</td>
<td>180</td>
<td>3068</td>
<td>353</td>
<td>631</td>
<td>3</td>
</tr>
<tr>
<td>25</td>
<td>WV</td>
<td>149</td>
<td>2470</td>
<td>329</td>
<td>390</td>
<td>3</td>
</tr>
<tr>
<td>26</td>
<td>NC</td>
<td>155</td>
<td>2664</td>
<td>354</td>
<td>450</td>
<td>3</td>
</tr>
<tr>
<td>27</td>
<td>SC</td>
<td>149</td>
<td>2380</td>
<td>377</td>
<td>476</td>
<td>3</td>
</tr>
<tr>
<td>28</td>
<td>GA</td>
<td>156</td>
<td>2781</td>
<td>371</td>
<td>603</td>
<td>3</td>
</tr>
<tr>
<td>29</td>
<td>FL</td>
<td>191</td>
<td>3191</td>
<td>336</td>
<td>805</td>
<td>3</td>
</tr>
<tr>
<td>30</td>
<td>KY</td>
<td>140</td>
<td>2645</td>
<td>349</td>
<td>523</td>
<td>3</td>
</tr>
<tr>
<td>31</td>
<td>TN</td>
<td>137</td>
<td>2579</td>
<td>343</td>
<td>588</td>
<td>3</td>
</tr>
<tr>
<td>32</td>
<td>AL</td>
<td>112</td>
<td>2337</td>
<td>362</td>
<td>584</td>
<td>3</td>
</tr>
<tr>
<td>33</td>
<td>MS</td>
<td>130</td>
<td>2081</td>
<td>385</td>
<td>445</td>
<td>3</td>
</tr>
<tr>
<td>34</td>
<td>AR</td>
<td>134</td>
<td>2322</td>
<td>352</td>
<td>500</td>
<td>3</td>
</tr>
<tr>
<td>35</td>
<td>LA</td>
<td>162</td>
<td>2634</td>
<td>390</td>
<td>661</td>
<td>3</td>
</tr>
<tr>
<td>36</td>
<td>OK</td>
<td>135</td>
<td>2880</td>
<td>330</td>
<td>680</td>
<td>3</td>
</tr>
<tr>
<td>37</td>
<td>TX</td>
<td>155</td>
<td>3029</td>
<td>369</td>
<td>797</td>
<td>3</td>
</tr>
<tr>
<td>38</td>
<td>MT</td>
<td>238</td>
<td>2942</td>
<td>369</td>
<td>534</td>
<td>4</td>
</tr>
<tr>
<td>39</td>
<td>ID</td>
<td>170</td>
<td>2668</td>
<td>368</td>
<td>541</td>
<td>4</td>
</tr>
<tr>
<td>40</td>
<td>WY</td>
<td>238</td>
<td>3190</td>
<td>366</td>
<td>605</td>
<td>4</td>
</tr>
<tr>
<td>41</td>
<td>CO</td>
<td>192</td>
<td>3340</td>
<td>358</td>
<td>785</td>
<td>4</td>
</tr>
<tr>
<td>42</td>
<td>NM</td>
<td>227</td>
<td>2651</td>
<td>421</td>
<td>698</td>
<td>4</td>
</tr>
<tr>
<td>43</td>
<td>AZ</td>
<td>207</td>
<td>3027</td>
<td>387</td>
<td>796</td>
<td>4</td>
</tr>
<tr>
<td>44</td>
<td>UT</td>
<td>201</td>
<td>2790</td>
<td>412</td>
<td>804</td>
<td>4</td>
</tr>
<tr>
<td>45</td>
<td>NV</td>
<td>225</td>
<td>3957</td>
<td>385</td>
<td>809</td>
<td>4</td>
</tr>
<tr>
<td>46</td>
<td>WA</td>
<td>215</td>
<td>3688</td>
<td>342</td>
<td>726</td>
<td>4</td>
</tr>
<tr>
<td>47</td>
<td>OR</td>
<td>233</td>
<td>3317</td>
<td>333</td>
<td>671</td>
<td>4</td>
</tr>
<tr>
<td>48</td>
<td>CA</td>
<td>273</td>
<td>3968</td>
<td>348</td>
<td>909</td>
<td>4</td>
</tr>
<tr>
<td>49</td>
<td>AK</td>
<td>372</td>
<td>4146</td>
<td>440</td>
<td>484</td>
<td>4</td>
</tr>
<tr>
<td>50</td>
<td>HI</td>
<td>212</td>
<td>3513</td>
<td>383</td>
<td>831</td>
<td>4</td>
</tr>
<tr>
<td>Row</td>
<td>STATE</td>
<td>Y</td>
<td>X₁</td>
<td>X₂</td>
<td>X₃</td>
<td>Region</td>
</tr>
<tr>
<td>-----</td>
<td>-------</td>
<td>-----</td>
<td>------</td>
<td>------</td>
<td>------</td>
<td>--------</td>
</tr>
<tr>
<td>1</td>
<td>ME</td>
<td>235</td>
<td>3944</td>
<td>325</td>
<td>508</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>NH</td>
<td>231</td>
<td>4578</td>
<td>323</td>
<td>564</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>VT</td>
<td>270</td>
<td>4011</td>
<td>328</td>
<td>322</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>MA</td>
<td>261</td>
<td>5233</td>
<td>305</td>
<td>846</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>RI</td>
<td>300</td>
<td>4780</td>
<td>303</td>
<td>871</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>CT</td>
<td>317</td>
<td>5889</td>
<td>307</td>
<td>774</td>
<td>1</td>
</tr>
<tr>
<td>7</td>
<td>NY</td>
<td>387</td>
<td>5663</td>
<td>301</td>
<td>856</td>
<td>1</td>
</tr>
<tr>
<td>8</td>
<td>NJ</td>
<td>285</td>
<td>5759</td>
<td>310</td>
<td>889</td>
<td>1</td>
</tr>
<tr>
<td>9</td>
<td>PA</td>
<td>300</td>
<td>4894</td>
<td>300</td>
<td>715</td>
<td>1</td>
</tr>
<tr>
<td>10</td>
<td>OH</td>
<td>221</td>
<td>5012</td>
<td>324</td>
<td>753</td>
<td>2</td>
</tr>
<tr>
<td>11</td>
<td>IN</td>
<td>264</td>
<td>4908</td>
<td>329</td>
<td>649</td>
<td>2</td>
</tr>
<tr>
<td>12</td>
<td>IL</td>
<td>308</td>
<td>5753</td>
<td>320</td>
<td>830</td>
<td>2</td>
</tr>
<tr>
<td>13</td>
<td>MI</td>
<td>379</td>
<td>5439</td>
<td>337</td>
<td>738</td>
<td>2</td>
</tr>
<tr>
<td>14</td>
<td>WI</td>
<td>342</td>
<td>4634</td>
<td>328</td>
<td>659</td>
<td>2</td>
</tr>
<tr>
<td>15</td>
<td>MN</td>
<td>378</td>
<td>4921</td>
<td>330</td>
<td>664</td>
<td>2</td>
</tr>
<tr>
<td>16</td>
<td>IA</td>
<td>232</td>
<td>4869</td>
<td>318</td>
<td>572</td>
<td>2</td>
</tr>
<tr>
<td>17</td>
<td>MO</td>
<td>231</td>
<td>4672</td>
<td>309</td>
<td>701</td>
<td>2</td>
</tr>
<tr>
<td>18</td>
<td>ND</td>
<td>246</td>
<td>4782</td>
<td>333</td>
<td>443</td>
<td>2</td>
</tr>
<tr>
<td>19</td>
<td>SD</td>
<td>230</td>
<td>4296</td>
<td>330</td>
<td>446</td>
<td>2</td>
</tr>
<tr>
<td>20</td>
<td>NB</td>
<td>268</td>
<td>4827</td>
<td>318</td>
<td>615</td>
<td>2</td>
</tr>
<tr>
<td>21</td>
<td>KS</td>
<td>337</td>
<td>5057</td>
<td>304</td>
<td>661</td>
<td>2</td>
</tr>
<tr>
<td>22</td>
<td>DE</td>
<td>344</td>
<td>5540</td>
<td>328</td>
<td>722</td>
<td>3</td>
</tr>
<tr>
<td>23</td>
<td>MD</td>
<td>330</td>
<td>5331</td>
<td>323</td>
<td>766</td>
<td>3</td>
</tr>
<tr>
<td>24</td>
<td>VA</td>
<td>261</td>
<td>4715</td>
<td>317</td>
<td>631</td>
<td>3</td>
</tr>
<tr>
<td>25</td>
<td>WV</td>
<td>214</td>
<td>3828</td>
<td>310</td>
<td>390</td>
<td>3</td>
</tr>
<tr>
<td>26</td>
<td>NC</td>
<td>245</td>
<td>4120</td>
<td>321</td>
<td>450</td>
<td>3</td>
</tr>
<tr>
<td>27</td>
<td>SC</td>
<td>233</td>
<td>3817</td>
<td>342</td>
<td>476</td>
<td>3</td>
</tr>
<tr>
<td>28</td>
<td>GA</td>
<td>250</td>
<td>4243</td>
<td>339</td>
<td>603</td>
<td>3</td>
</tr>
<tr>
<td>29</td>
<td>FL</td>
<td>243</td>
<td>4647</td>
<td>287</td>
<td>805</td>
<td>3</td>
</tr>
<tr>
<td>30</td>
<td>KY</td>
<td>216</td>
<td>3967</td>
<td>325</td>
<td>523</td>
<td>3</td>
</tr>
<tr>
<td>31</td>
<td>TN</td>
<td>212</td>
<td>3946</td>
<td>315</td>
<td>588</td>
<td>3</td>
</tr>
<tr>
<td>32</td>
<td>AL</td>
<td>208</td>
<td>3724</td>
<td>332</td>
<td>584</td>
<td>3</td>
</tr>
<tr>
<td>33</td>
<td>MS</td>
<td>215</td>
<td>3448</td>
<td>358</td>
<td>445</td>
<td>3</td>
</tr>
<tr>
<td>34</td>
<td>AR</td>
<td>221</td>
<td>3680</td>
<td>320</td>
<td>500</td>
<td>3</td>
</tr>
<tr>
<td>35</td>
<td>LA</td>
<td>244</td>
<td>3825</td>
<td>355</td>
<td>661</td>
<td>3</td>
</tr>
<tr>
<td>36</td>
<td>OK</td>
<td>234</td>
<td>4189</td>
<td>306</td>
<td>680</td>
<td>3</td>
</tr>
<tr>
<td>37</td>
<td>TX</td>
<td>269</td>
<td>4336</td>
<td>335</td>
<td>797</td>
<td>3</td>
</tr>
<tr>
<td>38</td>
<td>MT</td>
<td>302</td>
<td>4418</td>
<td>335</td>
<td>534</td>
<td>4</td>
</tr>
<tr>
<td>39</td>
<td>ID</td>
<td>268</td>
<td>4323</td>
<td>344</td>
<td>541</td>
<td>4</td>
</tr>
<tr>
<td>40</td>
<td>WY</td>
<td>323</td>
<td>4813</td>
<td>331</td>
<td>605</td>
<td>4</td>
</tr>
<tr>
<td>41</td>
<td>CO</td>
<td>304</td>
<td>5046</td>
<td>324</td>
<td>785</td>
<td>4</td>
</tr>
<tr>
<td>42</td>
<td>NM</td>
<td>317</td>
<td>3764</td>
<td>366</td>
<td>698</td>
<td>4</td>
</tr>
<tr>
<td>43</td>
<td>AZ</td>
<td>332</td>
<td>4504</td>
<td>340</td>
<td>796</td>
<td>4</td>
</tr>
<tr>
<td>44</td>
<td>UT</td>
<td>315</td>
<td>4005</td>
<td>378</td>
<td>804</td>
<td>4</td>
</tr>
<tr>
<td>45</td>
<td>NV</td>
<td>291</td>
<td>5560</td>
<td>330</td>
<td>809</td>
<td>4</td>
</tr>
<tr>
<td>46</td>
<td>WA</td>
<td>312</td>
<td>4989</td>
<td>313</td>
<td>726</td>
<td>4</td>
</tr>
<tr>
<td>47</td>
<td>OR</td>
<td>316</td>
<td>4697</td>
<td>305</td>
<td>671</td>
<td>4</td>
</tr>
<tr>
<td>48</td>
<td>CA</td>
<td>332</td>
<td>5438</td>
<td>307</td>
<td>909</td>
<td>4</td>
</tr>
<tr>
<td>49</td>
<td>AK</td>
<td>546</td>
<td>5613</td>
<td>386</td>
<td>484</td>
<td>4</td>
</tr>
<tr>
<td>50</td>
<td>HI</td>
<td>311</td>
<td>5309</td>
<td>333</td>
<td>831</td>
<td>4</td>
</tr>
</tbody>
</table>
Using $Y$ to represent per capita expenditure on schools, the model takes the form

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \gamma_1 T_1 + \gamma_2 T_2 + \delta_1 T_1 \cdot X_1 + \delta_2 T_1 \cdot X_2 + \delta_3 T_1 \cdot X_3 + \alpha_1 T_2 \cdot X_1 + \alpha_2 T_2 \cdot X_2 + \alpha_3 T_2 \cdot X_3 + \varepsilon.$$

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

For 1960: $Y = (\beta_0 + \gamma_1) + (\beta_1 + \delta_1) X_1 + (\beta_2 + \delta_2) X_2 + (\beta_3 + \delta_3) X_3 + \varepsilon,$

For 1970: $Y = (\beta_0 + \gamma_2) + (\beta_1 + \alpha_1) X_1 + (\beta_2 + \alpha_2) X_2 + (\beta_3 + \alpha_3) X_3 + \varepsilon,$

For 1975: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon.$

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

$$H_0 : \gamma_1 = \gamma_2 = \delta_1 = \delta_2 = \delta_3 = \alpha_1 = \alpha_2 = \alpha_3 = 0,$$

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 $H_0 : \gamma = 0$ using the $F$-Test.

(c) Test $H_0 : \gamma = 0$ 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 $H_0 : \delta = 0$ using the $F$-Test.

(c) Test $H_0 : \delta = 0$ 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.
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 \( X = 0.52 \), what is \( \hat{Y} \)?
(d) Given that \( X = 0.52 \), 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 \( \beta_1 \).
(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

<table>
<thead>
<tr>
<th>Source</th>
<th>Sum of Squares</th>
<th>df</th>
<th>Mean Square</th>
<th>$F$-Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>4604.7</td>
<td>1</td>
<td>4604.7</td>
<td>253</td>
</tr>
<tr>
<td>Residual</td>
<td>1604.44</td>
<td>88</td>
<td>18.2323</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient</th>
<th>s.e.</th>
<th>$t$-Test</th>
<th>$p$-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>–6.107</td>
<td>1.487</td>
<td>–4.11</td>
<td>0.0001</td>
</tr>
<tr>
<td>Horsepower</td>
<td>0.169</td>
<td>0.011</td>
<td>15.9</td>
<td>0.0001</td>
</tr>
</tbody>
</table>

### Model 2

<table>
<thead>
<tr>
<th>Source</th>
<th>Sum of Squares</th>
<th>df</th>
<th>Mean Square</th>
<th>$F$-Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>4818.84</td>
<td>4</td>
<td>1204.71</td>
<td>73.7</td>
</tr>
<tr>
<td>Residual</td>
<td>1390.31</td>
<td>85</td>
<td>16.3566</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient</th>
<th>s.e.</th>
<th>$t$-Test</th>
<th>$p$-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>–4.117</td>
<td>1.582</td>
<td>–2.6</td>
<td>0.0109</td>
</tr>
<tr>
<td>Horsepower</td>
<td>0.174</td>
<td>0.011</td>
<td>16.6</td>
<td>0.0001</td>
</tr>
<tr>
<td>USA</td>
<td>–3.162</td>
<td>1.351</td>
<td>–2.34</td>
<td>0.0216</td>
</tr>
<tr>
<td>Japan</td>
<td>–3.818</td>
<td>1.357</td>
<td>–2.81</td>
<td>0.0061</td>
</tr>
<tr>
<td>Germany</td>
<td>0.311</td>
<td>1.871</td>
<td>0.166</td>
<td>0.8682</td>
</tr>
</tbody>
</table>

### Model 3

<table>
<thead>
<tr>
<th>Source</th>
<th>Sum of Squares</th>
<th>df</th>
<th>Mean Square</th>
<th>$F$-Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>4889.3</td>
<td>7</td>
<td>698.471</td>
<td>43.4</td>
</tr>
<tr>
<td>Residual</td>
<td>1319.85</td>
<td>82</td>
<td>16.0957</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient</th>
<th>s.e.</th>
<th>$t$-Test</th>
<th>$p$-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>–10.882</td>
<td>4.216</td>
<td>–2.58</td>
<td>0.0116</td>
</tr>
<tr>
<td>Horsepower</td>
<td>0.237</td>
<td>0.038</td>
<td>6.21</td>
<td>0.0001</td>
</tr>
<tr>
<td>USA</td>
<td>2.076</td>
<td>4.916</td>
<td>0.42</td>
<td>0.6740</td>
</tr>
<tr>
<td>Japan</td>
<td>4.755</td>
<td>4.685</td>
<td>1.01</td>
<td>0.3131</td>
</tr>
<tr>
<td>Germany</td>
<td>11.774</td>
<td>9.235</td>
<td>1.28</td>
<td>0.2059</td>
</tr>
<tr>
<td>HP*USA</td>
<td>–0.052</td>
<td>0.042</td>
<td>–1.23</td>
<td>0.2204</td>
</tr>
<tr>
<td>HP*Japan</td>
<td>–0.077</td>
<td>0.041</td>
<td>–1.88</td>
<td>0.0631</td>
</tr>
<tr>
<td>HP*Germany</td>
<td>–0.095</td>
<td>0.066</td>
<td>–1.43</td>
<td>0.1560</td>
</tr>
</tbody>
</table>
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 $y_{ij} = \mu + \mu_1 F_1 + \mu_2 F_2 + \mu_3 F_3 + \epsilon_{ij}$.
(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

<table>
<thead>
<tr>
<th>Fertilizer 1</th>
<th>Fertilizer 2</th>
<th>Fertilizer 3</th>
<th>Control Group</th>
</tr>
</thead>
<tbody>
<tr>
<td>31</td>
<td>27</td>
<td>36</td>
<td>33</td>
</tr>
<tr>
<td>34</td>
<td>27</td>
<td>37</td>
<td>27</td>
</tr>
<tr>
<td>34</td>
<td>25</td>
<td>37</td>
<td>35</td>
</tr>
<tr>
<td>34</td>
<td>34</td>
<td>34</td>
<td>25</td>
</tr>
<tr>
<td>43</td>
<td>21</td>
<td>37</td>
<td>29</td>
</tr>
<tr>
<td>35</td>
<td>36</td>
<td>28</td>
<td>20</td>
</tr>
<tr>
<td>38</td>
<td>34</td>
<td>33</td>
<td>25</td>
</tr>
<tr>
<td>36</td>
<td>30</td>
<td>29</td>
<td>40</td>
</tr>
<tr>
<td>36</td>
<td>32</td>
<td>36</td>
<td>35</td>
</tr>
<tr>
<td>45</td>
<td>33</td>
<td>42</td>
<td>29</td>
</tr>
</tbody>
</table>

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:

\[ V = \beta_0 + \beta_1 I + \beta_2 D + \beta_3 W + \beta_4 (G \cdot I) + \beta_5 P + \beta_6 N + \varepsilon. \]  

(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 \( D (\beta_2) \).

(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)

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

(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:

\[D_1 = 1 \text{ if } D = 1 \text{ (Democratic incumbent is running) and } 0 \text{ otherwise, and}
\]

\[D_2 = 1 \text{ if } D = -1 \text{ (Republican incumbent is running) and } 0 \text{ otherwise.}\]
Table 5.19  Presidential Election Data (1916–1996)

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

Then an alternative to the model in (5.11) is

$$V = \beta_0 + \beta_1 I + \alpha_1 D_1 + \alpha_2 D_2 + \beta_3 W + \beta_4 (G \cdot I)$$

$$+ \beta_5 P + \beta_6 N + \varepsilon.$$  \hspace{1cm} (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

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