Chapter

Analysis of Covariance

Analysis of covariance (ANCOVA) is a technique that combines features of analysis of variance and regression. It can be used for either observational studies or designed experiments. The basic idea is to augment the analysis of variance model containing the factor effects with one or more additional quantitative variables that are related to the response variable. This augmentation is intended to reduce the variance of the error terms in the model, i.e., to make the analysis more precise. We considered covariance models briefly in Chapter 8 on page 329, and noted there that they are linear models containing both qualitative and quantitative predictor variables. Thus, covariance models are just a special type of regression model.

In this chapter, we shall first consider how a covariance model can be more effective than an ordinary ANOVA model. Then we shall discuss how to use a single-factor covariance model for making inferences. We conclude by taking up analysis of covariance models for two-factor studies and some additional considerations for the use of covariance analysis.

22.1 Basic Ideas

How Covariance Analysis Reduces Error Variability

Covariance analysis may be helpful in reducing large error term variances that sometimes are present in analysis of variance models. Consider a study in which the effects of three different films promoting travel in a state are studied. A subject receives an initial questionnaire to elicit information about the subject’s attitudes toward the state. The subject is then shown one of the three five-minute films, and immediately afterwards is questioned about the film, about desire to travel in the state, and so on.

In this type of situation, covariance analysis can be utilized. To see why it might be highly effective, consider Figure 22.1a. Here are plotted the desire-to-travel scores, obtained after each of the three promotional films was shown to a different group of five subjects. Three different symbols are used to distinguish the different treatments. It is evident from Figure 22.1a that the error terms, as shown by the scatter around the estimated treatment means \( \bar{Y}_i \), are fairly large, indicating a large error term variance.

Suppose now that we were to utilize also the subjects’ initial attitude scores. We plot in Figure 22.1b the desire-to-travel score (obtained after exposure to the film) against the initial attitude score for each of the 15 subjects. Note that the three treatment regression relations
FIGURE 22.1 Illustration of Error Variability Reduction by Covariance Analysis.

(a) Error Variability with Single-factor Analysis of Variance Model

(b) Error Variability with Covariance Analysis Model

happen to be linear (this need not be so). Also note that the scatter around the treatment regression lines is much less than the scatter in Figure 22.1a around the treatment means $\bar{Y}_t$, as a result of the desire-to-travel scores being highly linearly related to the initial attitude scores. The relatively large scatter in Figure 22.1a reflects the large error term variability that would be encountered with an analysis of variance model for this single-factor study. The smaller scatter in Figure 22.1b reflects the smaller error term variability that would be involved in an analysis of covariance model.

Covariance analysis, it is thus seen, utilizes the relationship between the response variable (desire-to-travel score, in our example) and one or more quantitative variables for which observations are available (prestudy attitude score, in our example) in order to reduce the error term variability and make the study a more powerful one for comparing treatment effects.
Comitant Variables

In covariance analysis terminology, each quantitative variable added to the ANOVA model is called a concomitant variable. We already encountered concomitant variables in Chapter 9, though not by that name. We mentioned in Chapter 9 that supplemental or uncontrolled variables are sometimes used in regression models for controlled experiments to reduce the variance of the experimental error terms. We also noted in that chapter that control variables may be added to the regression model in confirmatory observational studies to reflect the effects of previously identified explanatory variables as the effects of the new, primary explanatory variables on the response variable are being tested. Both the supplemental or uncontrolled variables in a controlled experiment and the control variables in a confirmatory observational study are concomitant variables that are added to the model primarily to reduce the variance of the error terms. Concomitant variables are sometimes also called covariates.

Choice of Concomitant Variables. The choice of concomitant variables is an important one. If such variables have no relation to the response variable, nothing is to be gained by covariance analysis, and one might as well use a simpler analysis of variance model. Concomitant variables frequently used with human subjects include prestudy attitudes, age, socioeconomic status, and aptitude. When retail stores are used as study units, concomitant variables might be last period’s sales or number of employees.

Concomitant Variables Unaffected by Treatments. For a clear interpretation of the results, a concomitant variable should be observed before the study; or if observed during the study, it should not be influenced by the treatments in any way. A prestudy attitude score meets this requirement. Also, if a subject’s age is ascertained during the study, it would be reasonable in many instances to expect that the information about age provided by the subject will not be affected by the treatment. The reason for this requirement can be seen readily from the following example. A company was conducting a training school for engineers to teach them accounting and budgeting principles. Two teaching methods were used, and engineers were assigned at random to one of the two. At the end of the program, a score was obtained for each engineer reflecting the amount of learning. The analyst decided to use as a concomitant variable in covariance analysis the amount of time devoted to study (which the engineers were required to record). After conducting the analysis of covariance, the analyst found that training method had virtually no effect. The analyst was baffled by this finding until it was pointed out that the amount of study time probably was also affected by the treatments, and analysis indeed confirmed this. One of the training methods involved computer-assisted learning which appealed to the engineers so that they spent more time studying and also learned more. In other words, both the learning score and the amount of study time were influenced by the treatment in this case. As a result of the high correlation between the amount of study time and the learning score, the marginal treatment effect of the teaching methods on amount of learning was small and the test for treatment effects showed no significant difference between the two teaching methods.

Whenever a concomitant variable is affected by the treatments, covariance analysis will fail to show some (or much) of the effects that the treatments had on the response variable, so that an uncritical analysis may be badly misleading.

A symbolic scatter plot can provide evidence as to whether the concomitant variable is affected by the treatments. Figure 22.2 shows a scatter plot of learning score and amount of
study time for the engineer training example. Treatment 1 is the one using computer-assisted learning. Note that most persons with this treatment devoted large amounts of time to study. On the other hand, persons receiving treatment 2 tended to devote smaller amounts of time to study. As a result, the observations for the two treatments tend to be concentrated over different intervals on the X scale.

Contrast this situation with the one seen in Figure 22.1b for the study on promotional films. Figure 22.1b illustrates how the concomitant variable observations should be scattered in a randomized experiment if the treatments have no effect on the concomitant variable. Here, the distribution of subjects along the X scale by prestudy attitude scores is roughly similar for all treatments, subject only to chance variation.

Comment
Covariance analysis is concerned with quantitative concomitant variables. When qualitative concomitant variables need to be added (e.g., gender, geographic region), the model remains an analysis of variance model where some of the factors are of primary interest and the others represent concomitant variables that are included for the purpose of error variance reduction.

22.2 Single-Factor Covariance Model

The covariance models to be presented in this chapter are applicable to observational studies and to experimental studies based on a completely randomized design. In the earlier engineer training example, the 24 engineers participating in the study were randomly assigned to the two teaching methods, with 12 engineers assigned to each teaching method. Thus, this experimental study was based on a completely randomized design.

The covariance models to be taken up in this chapter are also applicable to observational studies, such as an investigation of the salary increases of a company's employees in the accounting department by gender, where age is utilized as a concomitant variable.
We shall employ the notation for single-factor analysis of variance. The number of cases for the $i$th factor level is denoted by $n_i$, the total number of cases by $n_T = \sum n_i$, and the $j$th observation on the response variable for the $i$th factor level is denoted by $Y_{ij}$. We shall initially consider a single-factor covariance model with only one concomitant variable. Later we shall take up models with more than one concomitant variable. We shall denote the value of the concomitant variable associated with the $j$th case for the $i$th factor level by $X_{ij}$.

**Development of Covariance Model**

The single-factor ANOVA model in terms of fixed factor effects was given in (16.62):

$$Y_{ij} = \mu_\cdot + \tau_i + \varepsilon_{ij} \tag{22.1}$$

The covariance model starts with this ANOVA model and adds another term (or several), reflecting the relationship between the response variable and the concomitant variable. Usually, a linear relation is utilized as a first approximation:

$$Y_{ij} = \mu_\cdot + \tau_i + \gamma X_{ij} + \varepsilon_{ij} \tag{22.2}$$

Here $\gamma$ is a regression coefficient for the relation between $Y$ and $X$. The constant $\mu_\cdot$ now is no longer an overall mean. We can, however, make this constant an overall mean, and incidentally simplify some computations, if we center the concomitant variable around the overall mean $\overline{X}_\cdot$. The resulting model is the usual covariance model for a single-factor study with fixed factor levels:

$$Y_{ij} = \mu_\cdot + \tau_i + \gamma(X_{ij} - \overline{X}_\cdot) + \varepsilon_{ij} \tag{22.3}$$

where:

- $\mu_\cdot$ is an overall mean
- $\tau_i$ are the fixed treatment effects subject to the restriction $\sum \tau_i = 0$
- $\gamma$ is a regression coefficient for the relation between $Y$ and $X$
- $X_{ij}$ are constants
- $\varepsilon_{ij}$ are independent $N(0, \sigma^2)$
- $i = 1, \ldots, r; j = 1, \ldots, n_i$

Covariance model (22.3) corresponds to ANOVA model (22.1) except for the term $\gamma(X_{ij} - \overline{X}_\cdot)$, which is added to reflect the relationship between $Y$ and $X$. Note that the concomitant observations $X_{ij}$ are assumed to be constants. Since $\varepsilon_{ij}$ is the only random variable on the right side of (22.3), it follows at once that:

$$E[Y_{ij}] = \mu_\cdot + \tau_i + \gamma(X_{ij} - \overline{X}_\cdot) \tag{22.4a}$$

$$\sigma^2[Y_{ij}] = \sigma^2 \tag{22.4b}$$
In view of the independence of the $\varepsilon_{ij}$, the $Y_{ij}$ are also independent. Hence, an alternative statement of covariance model (22.3) is:

$$Y_{ij} \text{ are independent } N(\mu_{ij}, \sigma^2)$$

(22.5)

where:

$$\mu_{ij} = \mu_\cdot + \tau_i + \gamma(X_{ij} - \bar{X}_{..})$$

$$\sum \tau_i = 0$$

**Properties of Covariance Model**

Some of the properties of covariance model (22.3) are identical to those of ANOVA model (22.1). For instance, the error terms $\varepsilon_{ij}$ are independent and have constant variance. There are also some new properties, and we discuss these now.

**Comparisons of Treatment Effects.** With the analysis of variance model, all observations for the $i$th treatment have the same mean response; i.e., $E\{Y_{ij}\} = \mu_i$ for all $j$. This is not so with the covariance model, since the mean response $E\{Y_{ij}\}$ here depends not only on the treatment but also on the value of the concomitant variable $X_{ij}$ for the study unit. Thus, the expected response for the $i$th treatment with covariance model (22.3) is given by a regression line:

$$\mu_{ij} = \mu_\cdot + \tau_i + \gamma(X_{ij} - \bar{X}_{..})$$

(22.6)

This regression line indicates, for any value of $X$, the mean response with treatment $i$. Figure 22.3 illustrates for a study with three treatments how these treatment regression lines might appear. Note that $\mu_\cdot + \tau_i$ is the ordinate of the line for the $i$th treatment when

**FIGURE 22.3**

Example of Treatment Regression Lines with Covariance Model (22.3).
$X - \bar{X}_.. = 0$, that is, when $X = \bar{X}_..$, and that $\gamma$ is the slope of each line. Since all treatment regression lines have the same slope, they are parallel.

While we no longer can speak of the mean response with the $i$th treatment since it varies with $X$, we can still measure the effect of any treatment compared with any other by a single number. In Figure 22.3, for instance, treatment 1 leads to a higher mean response than treatment 2 by an amount that is the same no matter what is the value of $X$. The difference between the two mean responses is the same for all values of $X$ because the slopes of the regression lines are equal. Hence, we can measure the difference at any convenient $X$, say, at $X = \bar{X}_..$:

$$\mu_.. + \tau_1 - (\mu_.. + \tau_2) = \tau_1 - \tau_2$$

(22.7)

Thus, $\tau_1 - \tau_2$ measures how much higher the mean response is with treatment 1 than with treatment 2 for any value of $X$. We can compare any other two treatments similarly. It follows directly from this discussion that when all treatments have the same mean responses for any $X$ (i.e., the treatments have no differential effects), the treatment regression lines must be identical; and hence, $\tau_1 - \tau_2 = 0$, $\tau_1 - \tau_3 = 0$, etc. Indeed, all $\tau_i$ equal zero in that case.

**Constancy of Slopes.** The assumption in covariance model (22.3) that all treatment regression lines have the same slope is a crucial one. Without it, the difference between the effects of two treatments cannot be summarized by a single number based on the main effects, such as $\tau_2 - \tau_1$. Figure 22.4 illustrates the case of nonparallel slopes for two treatments. Here, treatment 1 leads to higher mean responses than treatment 2 for smaller values of $X$, and the reverse holds for larger values of $X$. When the treatments interact with the concomitant variable $X$, resulting in nonparallel slopes, covariance analysis is not appropriate. Instead, separate treatment regression lines need to be estimated and then compared.

**Generalizations of Covariance Model**

Covariance model (22.3) for single-factor studies can be generalized in several respects. We mention briefly three ways in which this model can be generalized.

**Nonconstant $X$s.** Covariance model (22.3) assumes that the observations $X_{ij}$ on the concomitant variable are constants. At times, it might be more reasonable to consider the concomitant observations as random variables. In that case, if covariance model (22.3) can be interpreted as a conditional one, applying for any $X$ values that might be observed, the covariance analysis to be presented is still appropriate.
**Nonlinearity of Relation.** The linear relation between $Y$ and $X$ assumed in covariance model (22.3) is not essential to covariance analysis. Any other relation could be used. For instance, the model for a quadratic relation is as follows:

$$Y_{ij} = \mu_i + \tau_i + \gamma_1(X_{ij} - \bar{X}_{..}) + \gamma_2(X_{ij} - \bar{X}_{..})^2 + \epsilon_{ij} \tag{22.8}$$

Linearity of the relation leads to simpler analysis and is often a sufficiently good approximation to provide meaningful results. If a linear relation is not a good approximation, however, a more adequate description of the relation should be utilized in the covariance model. Covariance analysis does require, however, that the treatment response functions be parallel; in other words, there must not be any interaction effects between the treatment and concomitant variables.

**Several Concomitant Variables.** Covariance model (22.3) uses a single concomitant variable. This is often sufficient to reduce the error variability substantially. However, the model can be extended in a straightforward fashion to include two or more concomitant variables. The single-factor covariance model for two concomitant variables, $X_1$ and $X_2$, to the first order is as follows:

$$Y_{ij} = \mu_i + \tau_i + \gamma_1(X_{ij1} - \bar{X}_{..1}) + \gamma_2(X_{ij2} - \bar{X}_{..2}) + \epsilon_{ij} \tag{22.9}$$

**Regression Formulation of Covariance Model**

An easy way to estimate the parameters of covariance model (22.3) and make inferences is through the regression approach. Computational formulas for manual calculation were developed before the advent of computers, making use of the special structure of the $X$ matrix for covariance models. Today, however, covariance calculations can be carried out readily by means of standard regression packages.

As for the regression formulation of analysis of variance models, we shall employ $r - 1$ indicator variables taking on the values 1, $-1$, or 0 to represent the $r$ treatments in a covariance analysis model:

$$I_t = \begin{cases} 
1 & \text{if case from treatment 1} \\
-1 & \text{if case from treatment } r \\
0 & \text{otherwise} 
\end{cases} \tag{22.10}$$

$$I_{r-1} = \begin{cases} 
1 & \text{if case from treatment } r - 1 \\
-1 & \text{if case from treatment } r \\
0 & \text{otherwise} 
\end{cases}$$

Note that we now denote the indicator variables by the symbol $I$ to clearly distinguish the treatment effects from the concomitant variable $X$.

In expressing covariance model (22.3) in regression form, we shall, as in the regression chapters, denote the centered observations $X_{ij} - \bar{X}_{..}$ by $x_{ij}$. Covariance model (22.3) can then be expressed as follows:

$$Y_{ij} = \mu + \tau_1 I_{ij1} + \cdot + \tau_{r-1} I_{ij,r-1} + \gamma x_{ij} + \epsilon_{ij} \tag{22.11}$$

where:

$$x_{ij} = X_{ij} - \bar{X}_{..}$$
Here, \( I_{ij} \) is the value of indicator variable \( I_i \) for the \( j \)th case from treatment \( i \), and similarly for the other indicator variables. Note that the treatment effects \( \tau_1, \ldots, \tau_r-1 \) are the regression coefficients for the indicator variables.

Now that we have formulated covariance model (22.3) as a regression model, our discussion of regression analysis in previous chapters applies. We therefore consider only briefly how to examine the appropriateness of the covariance model and how to make relevant inferences before turning to an example to illustrate the procedures.

**propriateness of Covariance Model**

Some of the key issues concerning the appropriateness of covariance model (22.3) and the equivalent regression model (22.11) deal with:

1. Normality of error terms.
2. Equality of error variances for different treatments.
3. Equality of slopes of the different treatment regression lines.
4. Linearity of regression relation with concomitant variable.
5. Uncorrelatedness of error terms.

The third issue, concerning the equality of the slopes of the different treatment regression lines, is particularly important in evaluating the appropriateness of covariance model (22.3). The test in Section 8.7 to compare several regression lines is applicable for determining whether the condition of equal slopes in the covariance model is met. We shall illustrate this test in the example in Section 22.3.

**Inferences of Interest**

The key statistical inferences of interest in covariance analysis are the same as with analysis of variance models, namely, whether the treatments have any effects, and if so what these effects are. Testing for fixed treatment effects involves the same alternatives as for analysis of variance models:

\[
H_0: \tau_1 = \tau_2 = \cdots = \tau_r = 0 \\
H_a: \text{not all } \tau_i \text{ equal zero} \tag{22.12}
\]

As we can see by referring to the equivalent regression model (22.11), this test involves testing whether several regression coefficients equal zero. The appropriate test statistic therefore is (7.27).

If the treatment effects are found to differ, the next step usually is to investigate the nature of these effects. Pairwise comparisons of treatment effects \( \tau_i - \tau_j \) (the vertical distance between the two treatment regression lines) may be of interest, or more general contrasts of the \( \tau_i \) may be relevant. In either case, linear combinations of the regression coefficients \( \tau_1, \ldots, \tau_{r-1} \) are to be estimated.

Occasionally, the nature of the regression relationship between \( Y \) and \( X \) is of interest, but usually the concomitant variable \( X \) is only employed in ANCOVA models to help reduce the error variability.

**Comment**

In covariance analysis there is usually no concern with whether the regression coefficient \( \gamma \) is zero, that is, whether there is indeed a regression relation between \( Y \) and \( X \). If there is no relation, no bias
results in the covariance analysis. The error mean square would simply be the same as for the analysis of variance model (allowing for sampling variation), and one degree of freedom would be lost for the error mean square.

### 22.3 Example of Single-Factor Covariance Analysis

A company studied the effects of three different types of promotions on sales of its crackers:

- **Treatment 1**—Sampling of product by customers in store and regular shelf space
- **Treatment 2**—Additional shelf space in regular location
- **Treatment 3**—Special display shelves at ends of aisle in addition to regular shelf space

Fifteen stores were selected for the study, and a completely randomized experimental design was utilized. Each store was randomly assigned one of the promotion types, with five stores assigned to each type of promotion. Other relevant conditions under the control of the company, such as price and advertising, were kept the same for all stores in the study. Data on the number of cases of the product sold during the promotional period, denoted by \( Y \), are presented in Table 22.1, as are also data on the sales of the product in the preceding period, denoted by \( X \). Sales in the preceding period are to be used as the concomitant variable.

### Development of Model

Figure 22.5 presents the data of Table 22.1 in the form of a symbolic scatter plot. Linear regression and parallel slopes for the treatment regression lines appear to be reasonable. Therefore, the following regression model was tentatively selected:

\[
Y_{ij} = \mu_i + \tau_1 I_{ij1} + \tau_2 I_{ij2} + \gamma X_{ij} + \epsilon_{ij} \quad \text{Full model (22.13)}
\]

where:

\[
I_1 = \begin{cases} 
1 & \text{if store received treatment 1} \\
-1 & \text{if store received treatment 3} \\
0 & \text{otherwise}
\end{cases}
\]

\[
I_2 = \begin{cases} 
1 & \text{if store received treatment 2} \\
-1 & \text{if store received treatment 3} \\
0 & \text{otherwise}
\end{cases}
\]

\[
x_{ij} = X_{ij} - \bar{X}_i.
\]

| Table 22.1 Data—Cracker Promotion Example (number of cases sold) |
|-----------------------------------------------|----------------|----------------|
| Treatment \( i \) | 1 | 2 | 3 | 4 | 5 |
| 1 | \( Y_{i1} \) | \( X_{i1} \) | \( Y_{i2} \) | \( X_{i2} \) | \( Y_{i3} \) | \( X_{i3} \) | \( Y_{i4} \) | \( X_{i4} \) | \( Y_{i5} \) | \( X_{i5} \) |
| 2 | 38 | 21 | 39 | 26 | 36 | 22 | 45 | 28 | 33 | 19 |
| 2 | 43 | 34 | 38 | 26 | 38 | 29 | 27 | 18 | 34 | 25 |
| 3 | 24 | 23 | 32 | 29 | 31 | 30 | 21 | 16 | 28 | 29 |
Table 22.2 repeats a portion of the data on the responses $Y$ and the concomitant variable $X$ in columns 1 and 2. The centered concomitant variable $x$ is presented in column 3 and the indicator variables for the treatments in columns 4 and 5. Note that the centering of the concomitant variable is around the overall mean $\bar{x}_{..} = 25$. Regressing $Y$ in column 1 of Table 22.2 on $x$, $I_1$, and $I_2$ in columns 3–5 by a computer package led to the results summarized in Table 22.3.

Various residual plots were obtained to examine the appropriateness of regression model (22.13). Figure 22.6 contains two of these. Figure 22.6a contains aligned residual dot plots for the three treatments. These do not suggest any major differences in the variances of the error terms. Figure 22.6b contains a normal probability plot of the residuals, which shows some modest departure from linearity. However, the coefficient of correlation between the ordered residuals and their expected values under normality is .958, for which Table B.6 does not suggest any significant departure from normality. The analyst also conducted a test to confirm the equality of the slopes of the three treatment regression lines. This test will be described shortly. On the basis of these analyses, the analyst concluded that regression model (22.13) is appropriate here.
Table 22.3

(a) Regression Coefficients

\[ \hat{\mu} = 33.800 \quad \hat{\tau}_2 = 0.942 \]
\[ \hat{\tau}_1 = 6.017 \quad \hat{\gamma} = 0.899 \]

(b) Analysis of Variance

<table>
<thead>
<tr>
<th>Source of Variation</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>SSR = 607.829</td>
<td>3</td>
<td>MSR = 202.610</td>
</tr>
<tr>
<td>Error</td>
<td>SSE = 38.571</td>
<td>11</td>
<td>MSE = 3.506</td>
</tr>
<tr>
<td>Total</td>
<td>SSTO = 646.400</td>
<td>14</td>
<td></td>
</tr>
</tbody>
</table>

(c) Estimated Variance-Covariance Matrix of Regression Coefficients

\[
\begin{bmatrix}
\hat{\mu} & \hat{\tau}_1 & \hat{\tau}_2 & \hat{\gamma} \\
\hat{\tau}_1 & 0 & 0.5016 & 0 & 0.2603 & 0.4882 & 0.0189 & -0.0147 & 0.0105 \\
\hat{\tau}_2 & 0 & 0 & 0.4882 & 0.0105 & 0.4882 & 0.0105 & 0.0105 & 0.0105 & 0.0105 \\
\hat{\gamma} & 0 & 0.2603 & 0.4882 & 0.0105 & 0.4882 & 0.0105 & 0.0105 & 0.0105 & 0.0105 \\
\end{bmatrix}
\]

Figure 22.6 Diagnostic Residual Plots—Cracker Promotion Example.

(a) Residual Dot Plots

(b) Normal Probability Plot

Test for Treatment Effects

To test whether or not the three cracker promotions differ in effectiveness, we can either follow the general linear test approach of fitting full and reduced models and using statistic (2.70) or use extra sums of squares and test statistic (7.27). In either case,
TABLE 22.4 Regression ANOVA Results for Reduced Model (22.15)—Cracker Promotion Example.

<table>
<thead>
<tr>
<th>Source of Variation</th>
<th>SS</th>
<th>df</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>$SSR = 190.678$</td>
<td>1</td>
</tr>
<tr>
<td>Error</td>
<td>$SSE = 455.722$</td>
<td>13</td>
</tr>
<tr>
<td>Total</td>
<td>$SSTO = 646.400$</td>
<td>14</td>
</tr>
</tbody>
</table>

alternatives are:

$$H_0: \tau_1 = \tau_2 = 0$$

$$H_a: \text{not both } \tau_1 \text{ and } \tau_2 \text{ equal zero}$$ (22.14)

Note that $\tau_3 = -\tau_1 - \tau_2$ must equal zero when $\tau_1 = \tau_2 = 0$.

We shall conduct the test by means of the general linear test approach. First, we develop the reduced model under $H_0$:

$$Y_{ij} = \mu. + \gamma x_{ij} + \epsilon_{ij} \quad \text{Reduced model}$$ (22.15)

Model (22.15) is just a simple linear regression model where none of the parameters vary for the different treatments. When regressing $Y$ in column 1 of Table 22.2 on $x$ in column 3, we obtain the analysis of variance results in Table 22.4.

We see from Table 22.4 that $SSE(R) = 455.722$ and from Table 22.3b that $SSE(F) = 38.571$. Hence, test statistic (2.70) here is:

$$F^* = \frac{SSE(R) - SSE(F)}{(n_T - 2) - (n_T - (r + 1))} \div \frac{SSE(F)}{n_T - (r + 1)}$$

$$= \frac{455.722 - 38.571}{13 - 11} \div \frac{38.571}{11} = 59.5$$

The level of significance is to be controlled at $\alpha = .05$; hence, we need to obtain $F(.95; 2, 11) = 3.98$. The decision rule therefore is:

If $F^* \leq 3.98$, conclude $H_0$

If $F^* > 3.98$, conclude $H_a$

Since $F^* = 59.5 > 3.98$, we conclude $H_a$, that the three cracker promotions differ in sales effectiveness. The $P$-value of the test is 0+.

Comment

Occasionally, a test whether or not $\gamma = 0$ is of interest. This is simply the ordinary test whether or not a single regression coefficient equals zero. It can be conducted by means of the $t^*$ test statistic (7.25) or by means of the $F^*$ test statistic (7.24).
Estimation of Treatment Effects

Since treatment effects were found to be present in the cracker promotion study, the analyst next wished to investigate the nature of these effects. We noted earlier that a comparison of two treatments involves \( \tau_i - \tau_j \), the vertical distance between the two treatment regression lines. Using the fact that \( \tau_3 = -\tau_1 - \tau_2 \) and (A.30b) for the variance of a linear combination of two random variables, we see that the estimators of all pairwise comparisons and their variances are as follows:

\[
\begin{array}{ccc}
\text{Comparison} & \text{Estimator} & \text{Variance} \\
\tau_1 - \tau_2 & \hat{\tau}_1 - \hat{\tau}_2 & \sigma^2 \{ \hat{\tau}_1 \} + \sigma^2 \{ \hat{\tau}_2 \} - 2\sigma \{ \hat{\tau}_1, \hat{\tau}_2 \} \\
\tau_1 - \tau_3 = 2\tau_1 + \tau_2 & 2\hat{\tau}_1 + \hat{\tau}_2 & 4\sigma^2 \{ \hat{\tau}_1 \} + \sigma^2 \{ \hat{\tau}_2 \} + 4\sigma \{ \hat{\tau}_1, \hat{\tau}_2 \} \\
\tau_2 - \tau_3 = \tau_1 + 2\tau_2 & \hat{\tau}_1 + 2\hat{\tau}_2 & \sigma^2 \{ \hat{\tau}_1 \} + 4\sigma^2 \{ \hat{\tau}_2 \} + 4\sigma \{ \hat{\tau}_1, \hat{\tau}_2 \} \\
\end{array}
\]

Table 22.3a furnishes the needed estimated regression coefficients, and Table 22.3c provides their estimated variances and covariances. We obtain from there:

\[
\begin{array}{ccc}
\text{Comparison} & \text{Estimate} & \text{Variance} \\
\tau_1 - \tau_2 & 6.017 - .942 & .5016 + .4882 - 2(-.2603) \\
& = 5.075 & = 1.5104 \\
\tau_1 - \tau_3 & 2(6.017) + .942 & 4(.5016) + .4882 + 4(-.2603) \\
& = 12.976 & = 1.4534 \\
\tau_2 - \tau_3 & 6.017 + 2(.942) & .5016 + 4(.4882) + 4(-2.2603) \\
& = 7.901 & = 1.4132 \\
\end{array}
\]

When a single interval estimate is to be constructed, the \( t \) distribution with \( n_T - r - 1 \) degrees of freedom is used. (The degrees of freedom are those associated with \( MSE \) in the full covariance model.) Usually, however, a family of interval estimates is desired. In that case, the Scheffé multiple comparison procedure may be employed with the \( S \) multiple defined by:

\[
S^2 = (r - 1) F(1 - \alpha; r - 1, n_T - r - 1)
\]

or the Bonferroni method may be employed with the \( B \) multiple:

\[
B = t(1 - \alpha/2g; n_T - r - 1)
\]

where \( g \) is the number of statements in the family. The Tukey method is not appropriate for covariance analysis.

In the case at hand, the analyst wished to obtain all pairwise comparisons with a 95 percent family confidence coefficient. The analyst used the Scheffé procedure in anticipation that
some additional estimates of contrasts might be desired. We require therefore:

\[ S^2 = (3 - 1) F(0.95; 2, 11) = 2(3.98) = 7.96 \quad S = 2.82 \]

Using the results in (22.16a), the confidence intervals for all pairwise treatment comparisons with a 95 percent family confidence coefficient then are:

\[
1.61 = 5.075 - 2.82\sqrt{1.5104} \leq \tau_1 - \tau_2 \leq 5.075 + 2.82\sqrt{1.5104} = 8.54
\]

\[
9.58 = 12.976 - 2.82\sqrt{1.4534} \leq \tau_1 - \tau_3 \leq 12.976 + 2.82\sqrt{1.4534} = 16.38
\]

\[
4.55 = 7.901 - 2.82\sqrt{1.4132} \leq \tau_2 - \tau_3 \leq 7.901 + 2.82\sqrt{1.4132} = 11.25
\]

These results indicate clearly that sampling in the store (treatment 1) is significantly better for stimulating cracker sales than either of the two shelf promotions, and that increasing the regular shelf space (treatment 2) is superior to additional displays at the end of the aisle (treatment 3).

**Comments**

1. Occasionally, more general contrasts among treatment effects than pairwise comparisons are desired. No new problems arise either in the use of the \( t \) distribution for a single contrast or in the use of the Scheffé or Bonferroni procedures for multiple comparisons. For instance, if the analyst desired in the cracker promotion example to compare the treatment effect for sampling in the store (treatment 1) with the two treatments involving shelf displays (treatments 2 and 3), the following contrast would be of interest:

\[
L = \tau_1 - \frac{\tau_2 + \tau_3}{2}
\]

The appropriate estimator is:

\[
\hat{L} = \hat{\tau}_1 - \frac{\hat{\tau}_2 + (\hat{\tau}_1 - \hat{\tau}_2)}{2} = \frac{3}{2}\hat{\tau}_1
\]

The variance of this estimator is by (A.16b):

\[
\sigma^2(\hat{L}) = \frac{9}{4}\sigma^2(\hat{\tau}_1)
\]

2. Sometimes there is interest in estimating the mean response with the \( i \)th treatment for a “typical” value of \( X \). Frequently \( X = \bar{X}_{..} \) is considered to be a “typical” value. We know from Figure 22.3 that at \( X = \bar{X}_{..} \), the mean response for the \( i \)th treatment is the intercept of the treatment regression line, \( \mu_0 + \tau_i \). An estimator of \( \mu_0 + \tau_i \) can be readily developed. For the cracker promotion example, we obtain the following estimators and their variances:

<table>
<thead>
<tr>
<th>Mean Response</th>
<th>Estimator</th>
<th>Variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>at ( X = \bar{X}_{..} )</td>
<td>( \mu_0 + \hat{\tau}_i )</td>
<td>( \sigma^2(\hat{\mu}_0) + \sigma^2(\hat{\tau}_i) + 2\sigma(\hat{\mu}_0, \hat{\tau}_i) )</td>
</tr>
<tr>
<td>( \mu_0 + \tau_1 )</td>
<td>( \hat{\mu}_0 + \hat{\tau}_1 )</td>
<td>( \sigma^2(\hat{\mu}_0) + \sigma^2(\hat{\tau}_1) + 2\sigma(\hat{\mu}_0, \hat{\tau}_1) )</td>
</tr>
<tr>
<td>( \mu_0 + \tau_2 )</td>
<td>( \hat{\mu}_0 + \hat{\tau}_2 )</td>
<td>( \sigma^2(\hat{\mu}_0) + \sigma^2(\hat{\tau}_2) + 2\sigma(\hat{\mu}_0, \hat{\tau}_2) )</td>
</tr>
<tr>
<td>( \mu_0 + \tau_3 )</td>
<td>( \hat{\mu}_0 - \hat{\tau}_1 - \hat{\tau}_2 )</td>
<td>( \sigma^2(\hat{\mu}_0) + \sigma^2(\hat{\tau}_1) + \sigma^2(\hat{\tau}_2) - 2\sigma(\hat{\mu}_0, \hat{\tau}_1) - 2\sigma(\hat{\mu}_0, \hat{\tau}_2) )</td>
</tr>
</tbody>
</table>

(22.22)
Use of the results in Table 22.3 leads to the following estimates:

<table>
<thead>
<tr>
<th>Treatment</th>
<th>Estimated Mean Response at $\bar{X}_i$</th>
<th>Estimated Variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$33.800 + 6.017 = 39.817$</td>
<td>$.2338 + .5016 + 2(0) = .7354$</td>
</tr>
<tr>
<td>2</td>
<td>$33.800 + .942 = 34.742$</td>
<td>$.2338 + .4882 + 2(0) = .7220$</td>
</tr>
<tr>
<td>3</td>
<td>$33.800 - 6.017 - .942 = 26.841$</td>
<td>$.2338 + .5016 + .4882 - 2(0) - 2(0)$ + $2(-.2603) = .7030$</td>
</tr>
</tbody>
</table>

The estimated mean response for treatment $i$ at $X = \bar{X}_i$ is often called the adjusted estimated treatment mean. It is said to be "adjusted" because it takes into account the effect of the concomitant variable. A comparison of the adjusted treatment means leads, of course, to the same pairwise comparisons of treatment effects as before; for instance, $39.817 - 34.742 = 5.075 = T_1 - T_2$.

**Test for Parallel Slopes**

An important assumption in covariance analysis is that all treatment regression lines have the same slope $\gamma$. The analyst who conducted the cracker promotion study indeed tested this assumption before proceeding with the analysis discussed earlier. We know from Chapter 8 that regression model (22.13) can be generalized to allow for different slopes for the treatments by introducing cross-product interaction terms. Specifically, interaction variables $I_1x_i$ and $I_2x_i$ will be required here. We shall denote the corresponding regression coefficients by $\beta_1$ and $\beta_2$. Thus, the generalized model is:

$$Y_{ij} = \mu + \tau_1I_{ij1} + \tau_2I_{ij2} + \gamma x_{ij} + \beta_1 I_{ij1}x_{ij} + \beta_2 I_{ij2}x_{ij} + \epsilon_{ij}$$

**Generalized model (22.23)**

Table 22.2 contains in columns 6 and 7 the interaction variables for this model for the cracker promotion example. Regressing the response variable $Y$ in column 1 of Table 22.2 on $x$, $I_1$, $I_2$, $I_1x$, $I_2x$ in columns 3–7 by means of a computer multiple regression package yielded the ANOVA results in Table 22.5. The error sum of squares $SSE$ obtained by fitting generalized model (22.23) is the equivalent of fitting separate regression lines for each treatment and summing these error sums of squares.

**TABLE 22.5 Regression ANOVA Results for Generalized Model (22.23)—Cracker Promotion Example.**

<table>
<thead>
<tr>
<th>Source of Variation</th>
<th>$SS$</th>
<th>$df$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>$SSR = 614.879$</td>
<td>5</td>
</tr>
<tr>
<td>Error</td>
<td>$SSE = 31.521$</td>
<td>9</td>
</tr>
<tr>
<td>Total</td>
<td>$SSTO = 646.400$</td>
<td>14</td>
</tr>
</tbody>
</table>
The test for parallel slopes is equivalent to testing for no interactions in generalized model (22.23):

\[ H_0: \beta_1 = \beta_2 = 0 \]
\[ H_a: \text{not both } \beta_1 \text{ and } \beta_2 \text{ equal zero} \quad (22.24) \]

We need to recognize that generalized model (22.23) now is the “full” model and covariance model (22.13) is the “reduced” model. Hence, we have from Tables 22.3b and 22.5:

\[ SSE(F) = 31.521 \quad SSE(R) = 38.571 \]

Thus, test statistic (2.70) becomes here:

\[ F^* = \frac{38.571 - 31.521}{11 - 9} \div \frac{31.521}{9} = 1.01 \]

For level of significance \( \alpha = 0.05 \), we require \( F(.95; 2, 9) = 4.26 \). Since \( F^* = 1.01 \leq 4.26 \), we conclude \( H_0 \), that the three treatment regression lines have the same slope. The \( P \)-value of the test is .40. Hence, the requirement of equal treatment slopes in analysis of covariance model (22.13) is met in the cracker promotion example.

**Comments**

1. An indication of the effectiveness of the analysis of covariance in reducing error variability can be obtained by comparing \( MSE \) for covariance analysis with \( MSE \) for regular analysis of variance. For the cracker promotion example, we know from Table 22.3 that \( MSE \) for the covariance analysis is 3.51. It can be shown that the error mean square for regular analysis of variance would have been 26.63. Hence, in this case, covariance analysis was able to reduce the residual variability by about 87 percent, a substantial reduction.

2. Covariance analysis and analysis of variance need not lead to the same conclusions about the treatment effects. For instance, analysis of variance might not indicate any treatment effects, whereas covariance analysis with a smaller error variance could show significant treatment effects. Ordinarily, of course, one should decide in advance which of the two analyses is to be used.

**22.4 Two-Factor Covariance Analysis**

We have until now considered covariance analysis for single-factor studies with \( r \) treatments. Covariance analysis can also be employed with two-factor and multifactor studies. We illustrate now the use of covariance analysis for two-factor studies with one concomitant variable. For notational simplicity, we consider the case where the treatment sample size is the same for all treatments. However, the regression approach to covariance analysis is general and applies directly when the study is unbalanced, with unequal treatment sample sizes.

**Covariance Model for Two-Factor Studies**

The fixed effects ANOVA model for a two-factor balanced study was given in (19.23):

\[ Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha \beta)_{ij} + \varepsilon_{ijk} \quad i = 1, \ldots, a; \ j = 1, \ldots, b; \ k = 1, \ldots, n \quad (22.25) \]
Regression Approach

We illustrate the regression approach to covariance analysis for a balanced two-factor study with one concomitant variable when both factors A and B are at two levels, i.e., when \( a = b = 2 \). The regression model counterpart to covariance model (22.26) then is:

\[
y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha \beta)_{ij} + \gamma (x_{ijk} - \bar{x}) \ldots + e_{ijk}
\]

\[
i = 1, \ldots, a; j = 1, \ldots, b; k = 1, \ldots, n
\]

(22.27)

where:

\[
I_1 = \begin{cases} 
1 & \text{if case from level 1 for factor A} \\
-1 & \text{if case from level 2 for factor A}
\end{cases}
\]

\[
I_2 = \begin{cases} 
1 & \text{if case from level 1 for factor B} \\
-1 & \text{if case from level 2 for factor B}
\end{cases}
\]

\[
x_{ijk} = X_{ijk} - \bar{X} \ldots
\]

Note that the regression coefficients in (22.27) are the analysis of variance factor effects \( \alpha_i \), \( \beta_1 \), and \( (\alpha \beta)_{11} \) and the concomitant variable coefficient \( \gamma \).

Testing for factor A main effects requires that \( \alpha_1 = 0 \) in the reduced model. Correspondingly, \( \beta_1 = 0 \) is required in the reduced model when testing for factor B main effects, and \( (\alpha \beta)_{11} = 0 \) is required in the reduced model when testing for AB interactions.

Estimation of factor A and factor B main effects can easily be done in terms of comparisons among the regression coefficients. The use of the Scheffé and Bonferroni multiple comparison procedures presents no new issues. For instance, the \( S \) multiple for multiple comparisons among the factor A level means is defined as follows:

\[
S^2 = (a - 1)F(1 - \alpha; a - 1, nab - ab - 1)
\]

(22.28)

and the B multiple is the same as in (22.18), with \( n_T = nab \) and \( r = ab \).

Example

A horticulturist conducted an experiment to study the effects of flower variety (factor A: varieties LP, WB) and moisture level (factor B: low, high) on yield of salable flowers \( (Y) \). Because the plots were not of the same size, the horticulturist wished to use plot size \( (X) \) as the concomitant variable. Six replications were made for each treatment. A portion of the data are presented in Table 22.6. Figure 22.7 contains a symbolic scatter plot of the data. The model assumptions of linear relations between \( Y \) and the concomitant variable \( X \), as well as of parallel slopes for the four treatments, appear to be reasonable here.

A fit of regression model (22.27) to the data by a computer regression package yielded the fitted regression function in Table 22.7a. The analyst plotted the data together with the fitted regression lines and made a variety of residual plots and tests (not shown). On the
basis of these diagnostics, the analyst was satisfied that regression model (22.27), which assumes parallel linear regression functions and constant error variances, is suitable here.

To examine the nature of the factor effects, we show in Figure 22.8 the estimated treatment means plot for the two moisture levels $B_1$ and $B_2$. These estimated means all correspond to plot size $X = 8.25$ or $x = 0$. Any other plot size would yield exactly the same relationships as those in Figure 22.8. It appears from Figure 22.8 that there are no important interactions between flower variety and moisture level, and that there may be main effects for both factors, particularly for moisture level.

To study formally the factor effects, reduced models were formed by deleting from regression model (22.27) one predictor variable at a time (recall that both factors have only two levels), and the reduced models were then fitted. The extra sums of squares so obtained, as well as the error sum of squares for the full model, are presented in Table 22.7b, together with the degrees of freedom and mean squares. No total sum of squares is shown because the factor effect components are not orthogonal.
TABLE 22.7
Computer Output for Fit of Regression Model (22.27)—Salable Flowers Example.

<table>
<thead>
<tr>
<th>Regression Coefficient</th>
<th>Estimated Regression Coefficient</th>
<th>Estimated Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\alpha_1$</td>
<td>2.04234</td>
<td>.52108</td>
</tr>
<tr>
<td>$\beta_1$</td>
<td>3.68078</td>
<td>.51291</td>
</tr>
<tr>
<td>$(\alpha\beta)_{11}$</td>
<td>.81922</td>
<td>.51291</td>
</tr>
<tr>
<td>$\gamma$</td>
<td>3.27688</td>
<td>.13002</td>
</tr>
</tbody>
</table>

(b) Extra Sums of Squares

<table>
<thead>
<tr>
<th>Effect</th>
<th>Source of Variation</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Concomitant variable</td>
<td>$x_1l_1, l_2, l_1l_2$</td>
<td>3,994.52</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>$A$</td>
<td>$l_1x, l_2, l_1l_2$</td>
<td>96.60</td>
<td>1</td>
<td>96.60</td>
</tr>
<tr>
<td>$B$</td>
<td>$l_2x, l_1, l_1l_2$</td>
<td>323.85</td>
<td>1</td>
<td>323.85</td>
</tr>
<tr>
<td>$AB$</td>
<td>$l_1l_2x, l_1, l_2$</td>
<td>16.04</td>
<td>1</td>
<td>16.04</td>
</tr>
<tr>
<td>Error</td>
<td></td>
<td>119.48</td>
<td>19</td>
<td>6.2884</td>
</tr>
</tbody>
</table>

FIGURE 22.8
Estimated Treatment Means Plot—Salable Flowers Example.

We test first for the presence of interactions by means of the usual general linear statistic $F^*$, using the results in Table 22.7b:

$$F^* = \frac{SSR(l_1l_2x, l_1, l_2)}{MSE} = \frac{16.04}{6.2884} = 2.55$$

For $\alpha = .01$, we require $F(.99; 1, 19) = 8.18$. Since $F^* = 2.55 \leq 8.18$, we conclude no interactions are present. The $P$-value of the test is .13.
We now wish to compare both the factor $A$ main effects and the factor $B$ main effects by means of confidence intervals, with a 95 percent family confidence coefficient. Since $\alpha_2 = -\alpha_1$, we have for our example:

$$L_1 = \alpha_1 - \alpha_2 = \alpha_1 - (-\alpha_1) = 2\alpha_1$$

Similarly, we obtain for the comparison of factor $B$ main effects:

$$L_2 = 2\beta_1$$

Point estimates are readily obtained from the results in Table 22.7a:

$$\hat{L}_1 = 2\hat{\alpha}_1 = 2(2.04234) = 4.08$$

$$\hat{L}_2 = 2\hat{\beta}_1 = 2(3.68078) = 7.36$$

The estimated standard deviations also follow easily, using (A.16b):

$$s(\hat{L}_1) = 2s(\hat{\alpha}_1) = 2(.52108) = 1.042$$

$$s(\hat{L}_2) = 2s(\hat{\beta}_1) = 2(.51291) = 1.026$$

We utilize the Bonferroni simultaneous estimation procedure for $g = 2$ comparisons. For a 95 percent family confidence coefficient, we require $t[1 - .05/2; 19] = t(.9875; 19) = 2.433$. The two desired confidence intervals therefore are:

$$1.5 = 4.08 - 2.433(1.042) \leq \alpha_1 - \alpha_2 \leq 4.08 + 2.433(1.042) = 6.6$$

$$4.9 = 7.36 - 2.433(1.026) \leq \beta_1 - \beta_2 \leq 7.36 + 2.433(1.026) = 9.9$$

With family confidence coefficient .95, we conclude that variety LP yields, on the average, between 1.5 and 6.6 more salable flowers for any given plot size than variety WB. Also, for any given plot size, the mean number of salable flowers is between 4.9 and 9.9 flowers greater for the low moisture level than for the high one, thus indicating a substantial effect of moisture level on yield.

If interactions had been present, we could have studied the nature of the interaction effects by, for instance, comparing the effect of the moisture level for each of the two flower varieties. It can be shown that this comparison is given by $(\alpha\beta)_{12} = -(\alpha\beta)_{11}$. Hence, we could estimate the desired interaction effect by using the estimated regression coefficient $(\alpha\beta)_{11}$ and its estimated standard deviation in Table 22.7a.

### Covariance Analysis for Randomized Complete Block Designs

Covariance analysis can be employed to further reduce the experimental error variability in a randomized complete block design. The extension is a straightforward one from covariance analysis for a completely randomized design.

**Covariance Model.** The usual randomized block design model was given in (21.1). The covariance model for a randomized block design with one concomitant variable is obtained by simply adding a term (or several terms) for the relation between the response variable $Y$ and the concomitant variable $X$. Assuming this relation can be described by a linear function, we obtain:

$$Y_{ij} = \mu + \rho_i + \tau_j + \gamma(X_{ij} - \overline{X}.\, \pm \varepsilon_{ij} \quad i = 1, \ldots, n_b; j = 1, \ldots, r$$

(22.29)
Regression Approach. The regression approach to covariance model (22.29) involves no new principles. We shall denote the centered variable \( X_i - \bar{X} \) in covariance model (22.29) by \( x_i \). Further, we shall again use 1, -1, 0 indicator variables for the block and treatment effects. To illustrate an equivalent regression model, consider a randomized complete block design study with \( n_b = 4 \) blocks and \( r = 3 \) treatments. The regression model counterpart to covariance model (22.29) then is:

\[
Y_{ij} = \mu + \rho_1 I_{ij1} + \rho_2 I_{ij2} + \rho_3 I_{ij3} + \tau_1 I_{ij4} + \tau_2 I_{ij5} + \gamma x_{ij} + \varepsilon_{ij} \quad \text{Full model} \tag{22.30}
\]

where:

\[
I_1 = \begin{cases} 
1 & \text{if experimental unit from block 1} \\
-1 & \text{if experimental unit from block 4} \\
0 & \text{otherwise}
\end{cases}
\]

\( I_2, I_3 \) are defined similarly.

\[
I_4 = \begin{cases} 
1 & \text{if experimental unit received treatment 1} \\
-1 & \text{if experimental unit received treatment 3} \\
0 & \text{otherwise}
\end{cases}
\]

\[
I_5 = \begin{cases} 
1 & \text{if experimental unit received treatment 2} \\
-1 & \text{if experimental unit received treatment 3} \\
0 & \text{otherwise}
\end{cases}
\]

\( x_{ij} = X_{ij} - \bar{X} \).

To test for treatment effects:

\[
H_0: \tau_1 = \tau_2 = \tau_3 = 0 \quad \text{(22.31)}
\]

\[
H_a: \text{not all } \tau_j \text{ equal zero}
\]

we would either need to fit the reduced model under \( H_0 \):

\[
Y_{ij} = \mu + \rho_1 I_{ij1} + \rho_2 I_{ij2} + \rho_3 I_{ij3} + \gamma x_{ij} + \varepsilon_{ij} \quad \text{Reduced model} \tag{22.32}
\]

or else use the appropriate extra sum of squares. The test for treatment effects is then conducted in the usual way.

Comparisons of two treatment effects by the regression approach are straightforward. For estimating \( \tau_1 - \tau_2 \), for instance, we use the unbiased estimator \( \hat{\tau}_1 - \hat{\tau}_2 \) based on the estimated regression coefficients obtained when fitting the full model (22.30). The estimated variance of this estimator is:

\[
s^2(\hat{\tau}_1 - \hat{\tau}_2) = s^2(\hat{\tau}_1) + s^2(\hat{\tau}_2) - 2s(\hat{\tau}_1, \hat{\tau}_2) \tag{22.33}
\]

The estimated variance-covariance matrix of the regression coefficients, available in many regression package printouts, can then be used to obtain the required estimated variances and covariances.
Comment
Some computer packages for covariance analysis produce analyses that are only valid when all treatment sample sizes are equal. Computer packages should therefore be used with great care when the treatment sample sizes are unequal, to make sure that the package conducts the tests of interest.

5 Additional Considerations for the Use of Covariance Analysis

Covariance Analysis as Alternative to Blocking
At times, a choice exists between: (1) a completely randomized design, with covariance analysis used to reduce the experimental errors and (2) a randomized block design, with the blocks formed by means of the concomitant variable. Generally, the latter alternative is preferred. There are several reasons for this:

1. If the regression between the response variable and the concomitant (blocking) variable is linear, a randomized block design and covariance analysis are about equally efficient. If the regression is not linear but covariance analysis with a linear relationship is utilized, covariance analysis with a completely randomized design will tend to be not as effective as a randomized block design.

2. Randomized block designs are essentially free of assumptions about the nature of the relationship between the blocking variable and the response variable, while covariance analysis assumes a definite form of relationship.

3. Randomized block designs have somewhat fewer degrees of freedom available for experimental error than with covariance analysis for a completely randomized design. However, in all but small-scale experiments, this difference in degrees of freedom has little effect on the precision of the estimates.

Use of Differences
In a variety of studies, a prestudy observation X and a poststudy observation Y on the same variable are available for each unit. For instance, X may be the score for a subject's attitude toward a company prior to reading its annual report, and Y may be the score after reading the report. In this situation, an obvious alternative to covariance analysis is to do an analysis of variance on the differences Y - X. Sometimes, Y - X is called an index of response because it makes one observation out of two.

If the slope of the treatment regression lines is \( \gamma = 1 \), analysis of covariance and analysis of variance on \( Y - X \) are essentially equivalent. When \( \gamma = 1 \), covariance model (22.2) becomes:

\[
Y_{ij} = \mu + \tau_i + X_{ij} + \epsilon_{ij}
\]  \hspace{1cm} (22.34)

which can be written as a regular analysis of variance model:

\[
Y_{ij} - X_{ij} = \mu + \tau_i + \epsilon_{ij}
\]  \hspace{1cm} (22.34a)

Thus, if a unit change in X leads to about the same change in Y, it makes sense to perform an analysis of variance on \( Y - X \) rather than to use covariance analysis, because
the analysis of variance model is a simpler model. If the regression slope is not near 1, however, covariance analysis may be substantially more effective than use of the differences \( Y - X \).

In the earlier cracker promotion example, use of \( Y - X \) would have been effective. It would have yielded the error mean square \( MSE = 3.500 \), which is practically the same as the error mean square for covariance analysis, \( MSE = 3.506 \) (see Table 22.3b). Recall that the regression slope in our example is close to 1 \( (\hat{\beta} = .899) \). hence, the approximate equivalence of the two procedures.

**Correction for Bias**

The suggestion is sometimes made that analysis of covariance can be helpful in correcting for bias with observational data. With such data, the groups under study may differ substantially with respect to a concomitant variable, and this may bias the comparisons of the groups. Consider, for instance, a study in which attitudes toward no-fault automobile insurance were compared for persons who are risk averse and persons who are risk seeking. It was found that many persons in the risk-averse group tended to be older (50 to 70 years old), while many persons in the risk-seeking group tended to be younger (20 to 40 years old).

In this type of situation, some researchers would advise that covariance analysis, with age as the concomitant variable, be employed to help remove any bias in the analysis of the observational data on attitudes toward no-fault insurance because the two age groups differ so much.

Even though there is great appeal in the idea of removing bias in observational data, covariance analysis should be used with caution for this purpose. In the first place, comparisons of means at a common value of \( X \) may require substantial extrapolation of the regression lines to a region where there are no or only few data points (in our example, to near 45 years). It may well be that the regression relationship used in the covariance analysis is not appropriate for substantial extrapolation. In the second place, the treatment variable may depend on the concomitant variable (or vice versa), which could affect the proper conclusions to be drawn.

**Interest in Nature of Treatment Effects**

Covariance analysis is sometimes employed for the principal purpose of shedding more light on the nature of the treatment effects, rather than merely for increasing the precision of the analysis. For instance, a market researcher in a study of the effects of three different advertisements on the maximum price consumers are willing to pay for a new type of home siding may use covariance analysis, with value of the consumer's home as the concomitant variable. The reason is because the researcher is truly interested in the relation for each advertisement between home value and maximum price. Reduction of error variance in this instance may be a secondary consideration.

As in all regression analyses, care must be used in drawing inferences about the causal nature of the relation between the concomitant variable and the response. In the advertising example, it might well be that value of a consumer's home is largely influenced by income. If this were so, the relation between value of the consumer's home and maximum price the consumer is willing to pay may actually be largely a reflection of the underlying relation between income and maximum price.
22.1. A student's reaction to the instructor's statement that covariance analysis is inappropriate when the treatment regression lines do not have the same slope was as follows: "It seems to me that this is ducking a real-world problem. If the treatment slopes are different, just use a covariance model that allows for different treatment slopes." Evaluate this reaction.

22.2. A survey analyst remarked: "When covariance analysis is used with survey data, there is a danger that the treatments may be related to the concomitant variable." What is the nature of the problem? Does this same problem exist when the treatments are randomly assigned to the experimental units?

22.3. Portray, analogously to the format of Figure 1.6 on page 11 for a regression model, the nature of covariance model (22.3) when there are three treatments and the parameter values are: \( \mu = 150, \tau_1 = 15, \tau_2 = -5, \tau_3 = -10, \gamma = 6, \bar{X} = 70, \sigma = 5. \) Show several distributions of \( Y \) for each treatment.

22.4. Refer to the cracker promotion example on page 926. A student stated, in discussing this case: "Strictly speaking, you cannot conclude anything about whether the three promotions differ in effectiveness because there was no control. The preceding period does not qualify as a control because it might have differed from the promotion period due to seasonal factors or other unique circumstances." Comment.

22.5. Refer to the cracker promotion example on pages 930 and 931, where three pairwise comparisons of treatment effects were made by the Scheffé procedure.

a. What would be the value of the Bonferroni multiple here for estimating the three comparisons?

b. Did the analyst obtain substantially less precise interval estimates using the Scheffé procedure, which permits making additional estimates without modifying the present ones?

22.6. State the analysis of covariance model for a single-factor study with four treatments when there are two concomitant variables, each with linear and quadratic terms in the model.

22.7. Refer to Productivity improvement Problem 16.7. The economist also has information on annual productivity improvement in the prior year and wishes to use this information as a concomitant variable. The data on the prior year's productivity improvement \((X_{ij})\) follow.

<table>
<thead>
<tr>
<th>(i)</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>8.2</td>
<td>7.9</td>
<td>7.0</td>
<td>5.7</td>
<td>7.2</td>
<td>7.0</td>
<td>6.5</td>
<td>7.9</td>
<td>6.3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>8.8</td>
<td>10.0</td>
<td>10.7</td>
<td>10.0</td>
<td>9.7</td>
<td>9.4</td>
<td>10.6</td>
<td>9.8</td>
<td>10.0</td>
<td>10.3</td>
<td>8.9</td>
<td>10.0</td>
</tr>
<tr>
<td>3</td>
<td>11.5</td>
<td>12.2</td>
<td>12.8</td>
<td>11.0</td>
<td>12.3</td>
<td>12.1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

a. Obtain the residuals for covariance model (22.3).

b. For each treatment, plot the residuals against the fitted values. Also prepare a normal probability plot of the residuals and calculate the coefficient of correlation between the ordered residuals and their expected values under normality. What do you conclude from your analysis?

c. State the generalized regression model to be employed for testing whether or not the treatment regression lines have the same slope. Conduct this test using \( \alpha = .01. \) State the alternatives, decision rule, and conclusion. What is the \( P \)-value of the test?

d. Could you conduct a formal test here as to whether the regression functions are linear? If so, how many degrees of freedom are there for the denominator mean square in the test statistic?
22.8. Refer to Productivity improvement Problems 16.7 and 22.7. Assume that covariance model (22.3) is appropriate.

a. Prepare a symbolic scatter plot of the data. Does it appear that there are effects of the level of research and development expenditures on mean productivity improvement? Discuss.

b. State the regression model equivalent to covariance model (22.3) for this case; use 1, -1, 0 indicator variables. Also state the reduced regression model for testing for treatment effects.

c. Fit the full and reduced regression models and test for treatment effects; use $\alpha = .05$. State the alternatives, decision rule, and conclusion. What is the $P$-value of the test?

d. Is $MSE(F)$ for the covariance model substantially smaller than $MSE$ for the analysis of variance model in Problem 16.7d? Does this affect the conclusion reached about treatment effects? Does it affect the $P$-value?

e. Estimate the mean productivity improvement for firms with moderate research and development expenditures that had a prior productivity improvement of $X = 9.0$; use a 95 percent confidence interval.

f. Make all pairwise comparisons between the treatment effects; use either the Bonferroni or the Scheffé procedure with a 90 percent family confidence coefficient, whichever is more efficient. State your findings.

22.9. Refer to Questionnaire color Problem 16.8. It has been suggested to the investigator that size of parking lot might be a useful concomitant variable. The number of spaces ($X_{ij}$) in each parking lot utilized in the study follow:

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>300</td>
<td>381</td>
<td>226</td>
<td>350</td>
<td>100</td>
</tr>
<tr>
<td>2</td>
<td>153</td>
<td>334</td>
<td>473</td>
<td>264</td>
<td>325</td>
</tr>
<tr>
<td>3</td>
<td>144</td>
<td>359</td>
<td>296</td>
<td>243</td>
<td>252</td>
</tr>
</tbody>
</table>

a. Obtain the residuals for covariance model (22.3).

b. For each treatment, plot the residuals against the fitted values. Also prepare a normal probability plot of the residuals and calculate the coefficient of correlation between the ordered residuals and their expected values under normality. What do you conclude from your analysis?

c. State the generalized regression model to be employed for testing whether or not the treatment regression lines have the same slope. Conduct this test using $\alpha = .005$. State the alternatives, decision rule, and conclusion. What is the $P$-value of the test?

d. Could you conduct a formal test here as to whether the regression functions are linear? Explain.

22.10. Refer to Questionnaire color Problems 16.8 and 22.9. Assume that covariance model (22.3) is applicable.

a. Prepare a symbolic scatter plot of the data. Does it appear that there are color effects on the mean response rate? Discuss.

b. State the regression model equivalent to covariance model (22.3) for this case; use 1, -1, 0 indicator variables. Also state the reduced regression model for testing for treatment effects.

c. Fit the full and reduced regression models and test for treatment effects; use $\alpha = .10$. State the alternatives, decision rule, and conclusion. What is the $P$-value of the test?
d. Is \( MSE(F) \) for the covariance model substantially smaller than \( MSE \) for the analysis of variance model in Problem 16.8d? How does this affect the conclusion reached about treatment effects?

e. Estimate the mean response rate for blue questionnaires in parking lots of size \( X = 280 \); use a 90 percent confidence interval.

f. Make all pairwise comparisons between the treatment effects; use either the Bonferroni or the Scheffé procedure with a 90 percent family confidence coefficient, whichever is more efficient. State your findings.

22.11. Refer to Rehabilitation therapy Problem 16.9. The rehabilitation researcher wishes to use age of patient as a concomitant variable. The ages \( \{X_{ij}\} \) of patients in the study follow.

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>18.3</td>
<td>30.0</td>
<td>26.5</td>
<td>28.1</td>
<td>29.7</td>
<td>27.8</td>
<td>19.8</td>
<td>29.3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>20.8</td>
<td>25.2</td>
<td>29.2</td>
<td>20.0</td>
<td>21.5</td>
<td>22.1</td>
<td>19.7</td>
<td>24.7</td>
<td>20.2</td>
<td>22.9</td>
</tr>
<tr>
<td>3</td>
<td>22.7</td>
<td>28.7</td>
<td>18.9</td>
<td>18.0</td>
<td>21.7</td>
<td>20.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

a. Obtain the residuals for covariance model (22.3).

b. For each treatment, plot the residuals against the fitted values. Also prepare a normal probability plot of the residuals and calculate the coefficient of correlation between the ordered residuals and their expected values under normality. What do you conclude from your analysis?

c. State the generalized regression model to be employed for testing whether or not the treatment regression lines have the same slope. Conduct this test using \( \alpha = .05 \). State the alternatives, decision rule, and conclusion. What is the \( P \)-value of the test?

d. Could you conduct a formal test here as to whether the regression functions are linear? Explain.

22.12. Refer to Rehabilitation therapy Problems 16.9 and 22.11. Assume that covariance model (22.3) is applicable.

a. Prepare a symbolic scatter plot of the data. Does it appear that there are effects of physical fitness status on the mean number of days required for therapy? Discuss.

b. State the regression model equivalent to covariance model (22.3) for this case; use 1, \(-1\), 0 indicator variables. Also state the reduced regression model for testing for treatment effects.

c. Fit the full and reduced regression models and test for treatment effects; use \( \alpha = .01 \). State the alternatives, decision rule, and conclusion. What is the \( P \)-value of the test?

d. Is \( MSE(F) \) for the covariance model substantially smaller than \( MSE \) for the analysis of variance model in Problem 16.9d? Does this affect the conclusion reached about treatment effects? Does it affect the \( P \)-value?

e. Estimate the mean number of days required for therapy for patients of average physical fitness and age 24 years; use a 99 percent confidence interval.

f. Make all pairwise comparisons between the treatment effects; use either the Bonferroni or the Scheffé procedure with a 95 percent family confidence coefficient, whichever is more efficient. State your findings.

22.13. Product display. A manufacturer of felt-tip markers investigated by an experiment whether a proposed new display, featuring a picture of a physician, is more effective in drugstores
than the present counter display, featuring a picture of an athlete and designed to be located in the stationery area. Fifteen drugstores of similar characteristics were chosen for the study. They were assigned at random in equal numbers to one of the following three treatments: (1) present counter display in stationery area, (2) new display in stationery area, (3) new display in checkout area. Sales with the present display \( X_{ij} \) were recorded in all 15 stores for a three-week period. Then the new display was set up in the 10 stores receiving it, and sales for the next three-week period \( Y_{ij} \) were recorded in all 15 stores. The data on sales (in dollars) follow.

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Treatment 1</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>First 3 weeks</td>
<td>92</td>
<td>68</td>
<td>74</td>
<td>52</td>
<td>65</td>
</tr>
<tr>
<td>Second 3 weeks</td>
<td>69</td>
<td>44</td>
<td>58</td>
<td>38</td>
<td>54</td>
</tr>
<tr>
<td><strong>Treatment 2</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>First 3 weeks</td>
<td>77</td>
<td>80</td>
<td>70</td>
<td>73</td>
<td>79</td>
</tr>
<tr>
<td>Second 3 weeks</td>
<td>74</td>
<td>75</td>
<td>73</td>
<td>78</td>
<td>82</td>
</tr>
<tr>
<td><strong>Treatment 3</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>First 3 weeks</td>
<td>64</td>
<td>43</td>
<td>81</td>
<td>68</td>
<td>71</td>
</tr>
<tr>
<td>Second 3 weeks</td>
<td>66</td>
<td>49</td>
<td>84</td>
<td>75</td>
<td>77</td>
</tr>
</tbody>
</table>

The analyst wishes to analyze the effects of the three different display treatments by means of covariance analysis.

a. Obtain the residuals for covariance model (22.3).

b. For each treatment, plot the residuals against the fitted values. Also prepare a normal probability plot of the residuals and calculate the coefficient of correlation between the ordered residuals and their expected values under normality. What do you conclude from your analysis?

c. State the generalized regression model to be employed for testing whether or not the treatment regression lines have the same slope. Conduct this test using \( \alpha = .05 \). State the alternatives, decision rule, and conclusion. What is the \( P \)-value of the test?

d. Could you conduct a formal test here as to whether the regression functions are linear? Explain.

22.14. Refer to **Product display** Problem 22.13. Assume that covariance model (22.3) is applicable.

a. Prepare a symbolic scatter plot of the data. Does it appear that there are display effects on mean sales? Discuss.

b. State the regression model equivalent to covariance model (22.3) for this case; use \( l, -1, 0 \) indicator variables. Also state the reduced regression model for testing for treatment effects.

c. Fit the full and reduced regression models and test for treatment effects; use \( \alpha = .05 \). State the alternatives, decision rule, and conclusion. What is the \( P \)-value of the test?

d. Is \( MSE(F) \) for the covariance model substantially smaller than the mean square error if analysis of variance model (16.2) had been employed?

e. Estimate the mean sales with display treatment 2 for stores whose sales in the preceding three-week period were $75; use a 95 percent confidence interval.

f. Make all pairwise comparisons between the treatment effects: use either the Bonferroni or the Scheffé procedure with a 90 percent family confidence coefficient, whichever is more efficient. State your findings.
22.15. Refer to Cash offers Problem 19.10. An analyst wishes to use each dealer's sales volume as a concomitant variable. The sales data ($X_{ijk}$, in hundred thousand dollars) follow.

\[
\begin{array}{ccc|ccc}
   & i=1 & & i=2 & & i=3 \\
   & j=1 & j=2 & j=1 & j=2 & j=1 & j=2 \\
3.0 & 3.5 & 6.5 & 2.2 & 5.0 & 4.0 \\
5.1 & 4.2 & 4.1 & 5.4 & 3.1 & .8 \\
... & ... & ... & ... & ... & ... \\
4.9 & 6.6 & 3.0 & 5.0 & 2.9 & 1.9 \\
\end{array}
\]

a. Obtain the residuals for covariance model (22.26).
b. For each treatment, plot the residuals against the fitted values. Also prepare a normal probability plot of the residuals and calculate the coefficient of correlation between the ordered residuals and their expected values under normality. What do you conclude from your analysis?
c. State the generalized regression model to be employed for testing whether or not the treatment regression lines have the same slope. Conduct this test using $\alpha = .01$. State the alternatives, decision rule, and conclusion. What is the $P$-value of the test?

22.16. Refer to Cash offers Problems 19.10 and 22.15. Assume that covariance model (22.26) is applicable.

a. State the regression model equivalent to covariance model (22.26) for this case; use 1, $-1$, 0 indicator variables. Fit this full model.
b. State the reduced regression models for testing for interaction and factor A and factor B main effects, respectively. Fit these reduced regression models.
c. Test for interaction effects; use $\alpha = .05$. State the alternatives, decision rule, and conclusion. What is the $P$-value of the test?
d. Test for factor A main effects; use $\alpha = .05$. State the alternatives, decision rule, and conclusion. What is the $P$-value of the test?
e. Test for factor B main effects; use $\alpha = .05$. State the alternatives, decision rule, and conclusion. What is the $P$-value of the test?
f. For each factor, make all pairwise comparisons between the factor level main effects. Use the Bonferroni procedure with a 90 percent family confidence coefficient. State your findings.

22.17. Refer to Eye contact effect Problem 19.12. Age of personnel officer is to be used as a concomitant variable. The ages ($X_{ijk}$) of the personnel officers follow.

\[
\begin{array}{ccc|ccc|ccc}
   & i=1 & & i=2 & & & i=2 & & & i=2 \\
   & j=1 & j=2 & j=1 & j=2 & j=1 & j=2 & j=1 & j=2 \\
42 & 51 & 43 & 42 \\
30 & 35 & 53 & 47 \\
... & ... & ... & ... \\
35 & 49 & 49 & 56 \\
\end{array}
\]

a. Obtain the residuals for covariance model (22.26).
b. For each treatment, plot the residuals against the fitted values. Also prepare a normal probability plot of the residuals and calculate the coefficient of correlation between the ordered
residuals, and their expected values under normality. What do you conclude from your analysis?

c. State the generalized regression model to be employed for testing whether or not the treatment regression lines have the same slope. Conduct this test using \( \alpha = .005 \). State the alternatives, decision rule, and conclusion. What is the \( P \)-value of the test?

22.18. Refer to Eye contact effect Problems 19.12 and 22.17. Assume that covariance model (22.26) is applicable.

a. State the regression model equivalent to covariance model (22.26) for this case; use 1, -1, 0 indicator variables. Fit this full model.

b. State the reduced regression models for testing for interaction and factor A and factor B main effects, respectively. Fit these reduced regression models.

c. Test for interaction effects; use \( \alpha = .01 \). State the alternatives, decision rule, and conclusion. What is the \( P \)-value of the test?

d. Test for factor A main effects; use \( \alpha = .01 \). State the alternatives, decision rule, and conclusion. What is the \( P \)-value of the test?

e. Test for factor B main effects; use \( \alpha = .01 \). State the alternatives, decision rule, and conclusion. What is the \( P \)-value of the test?

f. Compare the gender main effects by means of a 99 percent confidence interval. Interpret your interval estimate.

g. Estimate the mean success rating by female personnel officers aged 40 when eye contact is present; use a 99 percent confidence interval.

**22.19.** Refer to Auditor training Problem 21.5. The analyst wishes to examine whether use of pretraining statistical proficiency scores as a concomitant variable would help to reduce the experimental error variability significantly. The pretraining statistical proficiency scores for the auditors are as follows:

<table>
<thead>
<tr>
<th>Block</th>
<th>Training Method (j)</th>
<th>Block</th>
<th>Training Method (j)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>1</td>
<td>93</td>
<td>98</td>
<td>91</td>
</tr>
<tr>
<td>2</td>
<td>94</td>
<td>93</td>
<td>94</td>
</tr>
<tr>
<td>3</td>
<td>89</td>
<td>91</td>
<td>92</td>
</tr>
<tr>
<td>4</td>
<td>86</td>
<td>84</td>
<td>90</td>
</tr>
<tr>
<td>5</td>
<td>78</td>
<td>76</td>
<td>84</td>
</tr>
</tbody>
</table>

a. Would you expect the auditor’s age to have been a better concomitant variable here than the pretraining statistical proficiency score? Discuss.

b. State the regression model equivalent to covariance model (22.29); use 1, -1, 0 indicator variables.

c. Fit the full regression model.

d. State the reduced regression model for testing treatment effects. Fit the reduced model.

e. Test whether or not the training methods differ in mean effectiveness; use \( \alpha = .05 \). State the alternatives, decision rule, and conclusion. What is the \( P \)-value of the test?

f. Obtain a 95 percent confidence interval for \( \tau_1 - \tau_2 \). Interpret your interval estimate.

g. Has the error variance been reduced substantially by adding the concomitant variable? Explain.
22.20. Refer to Fat in diets Problem 21.7. The researcher wishes to examine whether each subject's body weight expressed as a percent of the ideal weight for that person would be a useful concomitant variable. The body weights as percents of the ideal weights for the 15 subjects are as follows:

<table>
<thead>
<tr>
<th>Block</th>
<th>Fat Content of Diet</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$j = 1$</td>
</tr>
<tr>
<td>1</td>
<td>94</td>
</tr>
<tr>
<td>2</td>
<td>97</td>
</tr>
<tr>
<td>3</td>
<td>105</td>
</tr>
<tr>
<td>4</td>
<td>108</td>
</tr>
<tr>
<td>5</td>
<td>118</td>
</tr>
</tbody>
</table>

a. State the regression model equivalent to covariance model (22.29); use $1, -1, 0$ indicator variables.
b. Fit the full regression model.
c. State the reduced regression model for testing treatment effects. Fit the reduced model.
d. Test whether or not the mean reductions in lipid level differ for the three diets; use $\alpha = .05$. State the alternatives, decision rule, and conclusion. What is the $P$-value of the test?
e. Obtain confidence intervals for $L_1 = \tau_1 - \tau_2$ and $L_2 = \tau_2 - \tau_3$, using the Bonferroni procedure with a 95 percent family confidence coefficient. Interpret your interval estimates.
f. Has the error variance been reduced substantially by adding the concomitant variable? Explain.

22.21. Refer to Productivity improvement Problems 22.7 and 22.8. The analyst is considering the use of the difference between the productivity improvements in the two years $(Y_{ij} - X_{ij})$ as the response variable with the regular analysis of variance model (22.29a).

a. Obtain the analysis of variance table.
b. How effective here is the use of differences with the regular ANOVA model compared to the use of covariance model (22.3)? Discuss.

22.22. Refer to Product display Problems 22.13 and 22.14. The analyst is considering the use of the difference in sales between the two periods $(Y_{ij} - X_{ij})$ as the response variable with the regular analysis of variance model (22.29a).

a. Obtain the analysis of variance table.
b. How effective here is the use of differences with the regular ANOVA model compared to the use of covariance model (22.3)? Discuss.

Exercise 22.23. (Calculus needed.) Denote $\mu_i + \tau_i$ in covariance model (22.3) by $\Delta_i$. Derive the least squares estimators for $\Delta_i$ and $\gamma$ in covariance model (22.3).

Projects 22.24. Refer to the SENIC data set in Appendix C.1. The following hospitals are to be considered in a study of the effects of region (variable 9) on the mean length of hospital stay of patients (variable 2), with available facilities and services (variable 12) as a concomitant variable:
a. Obtain the residuals for covariance model (22.3).

b. For each region, plot the residuals against the fitted values. Also prepare a normal probability plot of the residuals and calculate the coefficient of correlation between the ordered residuals and their expected values under normality. What do you conclude from your analysis?

c. State the generalized regression model to be employed for testing whether or not the treatment regression lines have the same slope. Conduct this test using $\alpha = .005$. State the alternatives, decision rule, and conclusion. What is the $P$-value of the test?

22.25. Refer to the SENIC data set in Appendix C.1 and Project 22.24. Assume that covariance model (22.3) is applicable.

a. Prepare a symbolic scatter plot of the data. Does it appear that there are region effects on the mean length of hospital stay? Discuss.

b. State the regression model equivalent to covariance model (22.3) for this case; use 1, $-1$, and 0 indicator variables. Also state the reduced regression model for testing for treatment effects.

c. Fit the full and reduced regression models and test for treatment effects; use $\alpha = .05$. State the alternatives, decision rule, and conclusion. What is the $P$-value of the test?

d. Make all pairwise comparisons between the region effects; use either the Bonferroni or the Scheffé procedure with a 90 percent family confidence coefficient, whichever is more efficient. State your findings.

22.26. Refer to the Market share data set in Appendix C.3 and Project 16.45. Use price (variable 3) as a concomitant variable.

a. Obtain the residuals for covariance model (22.3).

b. For each treatment, plot the residuals against the fitted values. Also prepare a normal probability plot of the residuals and calculate the coefficient of correlation between the ordered residuals and their expected values under normality. What do you conclude from your analysis?

c. State the generalized regression model to be employed for testing whether or not the treatment regression lines have the same slope. Conduct this test using $\alpha = .05$. State the alternatives, decision rule, and conclusion. What is the $P$-value of the test?

d. Could you conduct a formal test here as to whether the regression functions are linear? Explain.

22.27. Refer to the Market share data set in Appendix C.3 and Project 22.26.

a. Prepare a symbolic scatter plot of the data. Does it appear that mean monthly market share changes with the discount price and package promotion factor-level combinations? Discuss.

b. State the regression model equivalent to covariance model (22.3) for this case; use 1, $-1$, and 0 indicator variables. Also state the reduced regression model for testing for treatment effects.

c. Fit the full and reduced regression models and test for treatment effects; use $\alpha = .01$. State the alternatives, decision rule, and conclusion. What is the $P$-value of the test?

d. Is $MSE (F)$ for the covariance model substantially smaller than $MSE$ for the analysis of variance model in Project 16.45? Does this affect the conclusion reached about treatment effects? Does it affect the $P$-value?

e. Estimate the average monthly market share for product with discount price present, package promotion absent, and average monthly price of product 2.5; use a 99 percent confidence interval.
f. Make all pairwise comparisons between the treatment effects; use either the Bonferroni or the Scheffé procedure with a 95 percent family confidence coefficient, whichever is more efficient. State your findings.

22.28. Refer to the CDI data set in Appendix C.2 and Project 19.53. The metropolitan areas identified in Project 19.53 are to be considered in a study of the effects of region (factor A: variable 17) and percent below poverty level (factor B: variable 13) on crime rate (variable 10 ÷ variable 5), with percent of population 65 or older (variable 7) as a concomitant variable. For purposes of this analysis of covariance study, percent below poverty level is to be classified into two categories: less than 8.0 percent, and 8.0 percent or more.

a. Obtain the residuals for covariance model (22.26).

b. For each treatment, plot the residuals against the fitted values. Also prepare a normal probability plot of the residuals and calculate the coefficient of correlation between the ordered residuals and their expected values under normality. What do you conclude from your analysis?

c. State the generalized regression model to be employed for testing whether or not the treatment regression lines have the same slope. Conduct this test using \( \alpha = .001 \). State the alternatives, decision rule, and conclusion. What is the \( P \)-value of the test?

22.29. Refer to the CDI data set in Appendix C.2 and Project 22.28. Assume that covariance model (22.26) is applicable.

a. State the regression model equivalent to covariance model (22.26) for this case; use 1, -1, 0 indicator variables. Fit this full model.

b. State the reduced regression models for testing for interaction and factor A and factor B main effects, respectively. Fit these reduced regression models.

c. Test for interaction effects; use \( \alpha = .01 \). State the alternatives, decision rule, and conclusion. What is the \( P \)-value of the test?

d. Test for factor A main effects; use \( \alpha = .01 \). State the alternatives, decision rule, and conclusion. What is the \( P \)-value of the test?

e. Test for factor B main effects; use \( \alpha = .01 \). State the alternatives, decision rule, and conclusion. What is the \( P \)-value of the test?

22.30. Refer to the Market share data set in Appendix C.3 and Project 19.55. Use price (variable 3) as a concomitant variable.

a. Obtain the residuals for covariance model (22.26).

b. For each treatment, plot the residuals against the fitted values. Also prepare a normal probability plot of the residuals and calculate the coefficient of correlation between the ordered residuals and their expected values under normality. What do you conclude from your analysis?

c. State the generalized regression model to be employed for testing whether or not the treatment regression lines have the same slope. Conduct this test using \( \alpha = .05 \). State the alternatives, decision rule, and conclusion. What is the \( P \)-value of the test?

22.31. Refer to the Market share data set in Appendix C.3 and Project 22.30.

a. State the regression model equivalent to covariance model (22.26) for this case; use 1, -1, 0 indicator variables. Fit this full model.

b. State the reduced regression models for testing for interaction and factor A and factor B main effects, respectively. Fit these reduced regression models.

c. Test for interaction effects; use \( \alpha = .01 \). State the alternatives, decision rule, and conclusion. What is the \( P \)-value of the test?
d. Test for factor A main effects; use $\alpha = .01$. State the alternatives, decision rule, and conclusion. What is the $P$-value of the test?

e. Test for factor B main effects; use $\alpha = .01$. State the alternatives, decision rule, and conclusion. What is the $P$-value of the test?

Case Studies

22.32. Refer to the Prostate cancer data set in Appendix C.5 and Case Study 16.49. Carry out a one-way analysis of covariance of this data set, where the response of interest is PSA level (variable 2), the single factor is Gleason score (variable 9), and the possible covariates are cancer volume (variable 3) and weight (variable 4). The analysis should consider transformations of the response variable and the covariates. Document steps taken in your analysis, and justify your conclusions.

22.33. Refer to the Real estate sales data set in Appendix C.7 and Case Study 16.50. Carry out a one-way analysis of covariance of this data set, where the response of interest is sales price (variable 2), the single factor is number of bedrooms (variable 4), and the possible covariates are finished square feet (variable 3) and lot size (variable 12). Recode the number of bedrooms into four categories: 0–2, 3, 4, and greater than or equal to 5. The analysis should consider transformations of the response variable and the covariates. Document steps taken in your analysis, and justify your conclusions.

22.34. Refer to the Ischemic heart disease data set in Appendix C.9 and Case Study 16.51. Carry out a one-way analysis of covariance of this data set, where the response of interest is total cost (variable 2). the single factor is total number of interventions (variable 5), and the possible covariates are duration (variable 10) and age (variable 3). Recode the number of interventions into six categories: 0, 1, 2, 3–4, 5–7, and greater than or equal to 8. The analysis should consider transformations of the response variable and the covariates. Document steps taken in your analysis, and justify your conclusions.

22.35. Refer to the Real estate sales data set in Appendix C.7 and Case Study 19.59. Carry out a balanced two-way analysis of covariance of this data set where the response of interest is sales price (variable 2), the two crossed factors are quality (variable 10) and style (variable 11), and the possible covariates are finished square feet (variable 3) and lot size (variable 12). Style is recoded as either 1 or not 1. Order the observations in the six factor-level-combination cells from smallest to largest observation number and retain the first 25 observations in each cell for a total of 150 observations. The analysis should consider transformations of the response variable and the covariates. Document the steps taken in your analysis and justify your conclusions.

22.36. Refer to the Ischemic heart disease data set in Appendix C.9 and Case Study 16.60. Carry out a balanced two-way analysis of covariance of this data set where the response of interest is total cost (variable 2), the two crossed factors are number of interventions (variable 5) and number of comorbidities (variable 9), and the possible covariates are duration (variable 10) and age (variable 3). Recode the number of interventions into six categories: 0, 1, 2, 3–4, 5–7, and greater than or equal to 8. Recode the number of comorbidities into two categories: 0–1, and greater than or equal to 2. Order the observations in the twelve factor-level-combination cells from smallest to largest observation number and retain the first 43 observations in each cell for a total of 516 observations. The analysis should consider transformations of the response variable and the covariates. Document the steps taken in your analysis and justify your conclusions.