8.2 Interaction Regression Models

We have previously noted that regression models with cross-product interaction effects, such as regression model (6.15), are special cases of general linear regression model (6.7). We also encountered regression models with interaction effects briefly when we considered polynomial regression models, such as model (8.7). Now we consider in some detail regression models with interaction effects, including their interpretation and implementation.

Interaction Effects

A regression model with $p-1$ predictor variables contains additive effects if the response function can be written in the form:

$$E\{Y\} = f_1(X_1) + f_2(X_2) + \cdots + f_{p-1}(X_{p-1}) \tag{8.21}$$

where $f_1, f_2, \ldots, f_{p-1}$ can be any functions, not necessarily simple ones. For instance, the following response function with two predictor variables can be expressed in the form of (8.21):

$$E\{Y\} = \underbrace{\beta_0 + \beta_1 X_1 + \beta_2 X_1^2}_{f_1(X_1)} + \underbrace{\beta_3 X_2}_{f_2(X_2)}$$

We say here that the effects of X_1 and X_2 on Y are additive.

In contrast, the following regression function:

$$E\{Y\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2$$

cannot be expressed in the form (8.21). Hence, this latter regression model is not additive, or, equivalently, it contains an interaction effect.

A simple and commonly used means of modeling the interaction effect of two predictor variables on the response variable is by a cross-product term, such as $\beta_3 X_1 X_2$ in the above response function. The cross-product term is called an *interaction term*. More specifically, it is sometimes called a *linear-by-linear* or a *bilinear* interaction term. When there are three predictor variables whose effects on the response variable are linear, but the effects on Y of X_1 and X_2 and of X_1 and X_3 are interacting, the response function would be modeled as follows using cross-product terms:

$$E\{Y\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_1 X_2 + \beta_5 X_1 X_3$$

Interpretation of Interaction Regression Models with Linear Effects

We shall explain the influence of interaction effects on the shape of the response function and on the interpretation of the regression coefficients by first considering the simple case of two quantitative predictor variables where each has a linear effect on the response variable.

Interpretation of Regression Coefficients. The regression model for two quantitative predictor variables with linear effects on Y and interacting effects of X_1 and X_2 on Y represented by a cross-product term is as follows:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i1} X_{i2} + \varepsilon_i \tag{8.22}$$

The meaning of the regression coefficients β_1 and β_2 here is not the same as that given earlier because of the interaction term $\beta_3 X_{i1} X_{i2}$. The regression coefficients β_1 and β_2 no longer indicate the change in the mean response with a unit increase of the predictor variable, with the other predictor variable held constant at any given level. It can be shown that the change in the mean response with a unit increase in X_1 when X_2 is held constant is:

$$\beta_1 + \beta_3 X_2 \tag{8.23}$$

Similarly, the change in the mean response with a unit increase in X_2 when X_1 is held constant is:

$$\beta_2 + \beta_3 X_1 \tag{8.24}$$

Hence, in regression model (8.22) both the effect of X_1 for given level of X_2 and the effect of X_2 for given level of X_1 depend on the level of the other predictor variable.

We shall illustrate how the effect of one predictor variable depends on the level of the other predictor variable in regression model (8.22) by returning to the sales promotion response function shown in Figure 6.1 on page 215. The response function (6.3) for this example, relating locality sales (Y) to point-of-sale expenditures (X_1) and TV expenditures (X_2), is additive:

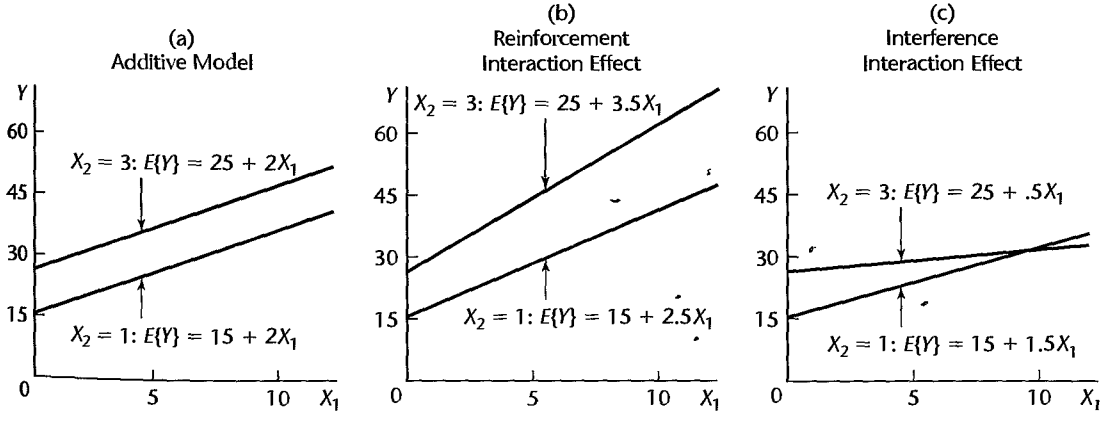
$$E\{Y\} = 10 + 2X_1 + 5X_2 \tag{8.25}$$

In Figure 8.7a, we show the response function $E\{Y\}$ as a function of X_1 when $X_2 = 1$ and when $X_2 = 3$. Note that the two response functions are parallel—that is, the mean sales response increases by the same amount $\beta_1 = 2$ with a unit increase of point-of-sale expenditures whether TV expenditures are $X_2 = 1$ or $X_2 = 3$. The plot in Figure 8.7a is called a *conditional effects plot* because it shows the effects of X_1 on the mean response conditional on different levels of the other predictor variable.

In Figure 8.7b, we consider the same response function but with the cross-product term $.5X_1X_2$ added for interaction effect of the two types of promotional expenditures on sales:

$$E\{Y\} = 10 + 2X_1 + 5X_2 + .5X_1X_2 \tag{8.26}$$

FIGURE 8.7 Illustration of Reinforcement and Interference Interaction Effects—Sales Promotion Example.



We again use a conditional effects plot to show the response function $E\{Y\}$ as a function of X_1 conditional on $X_2 = 1$ and on $X_2 = 3$. Note that the slopes of the response functions plotted against X_1 now differ for $X_2 = 1$ and $X_2 = 3$. The slope of the response function when $X_2 = 1$ is by (8.23):

$$\beta_1 + \beta_3 X_2 = 2 + .5(1) = 2.5$$

and when $X_2 = 3$, the slope is:

$$\beta_1 + \beta_3 X_2 = 2 + .5(3) = 3.5$$

Thus, a unit increase in point-of-sale expenditures has a larger effect on sales when TV expenditures are at a higher level than when they are at a lower level.

Hence, β_1 in regression model (8.22) containing a cross-product term for interaction effect no longer indicates the change in the mean response for a unit increase in X_1 for any given X_2 level. That effect in this model depends on the level of X_2. Although the mean response in regression model (8.22) when X_2 is constant is still a linear function of X_1, now both the intercept and the slope of the response function change as the level at which X_2 is held constant is varied. The same holds when the mean response is regarded as a function of X_2, with X_1 constant.

Note that as a result of the interaction effect in regression model (8.26), the increase in sales with a unit increase in point-of-sale expenditures is greater, the higher the level of TV expenditures, as shown by the larger slope of the response function when $X_2 = 3$ than when $X_2 = 1$. A similar increase in the slope occurs if the response function against X_2 is considered for higher levels of X_1. When the regression coefficients β_1 and β_2 are positive, we say that the interaction effect between the two quantitative variables is of a *reinforcement* or *synergistic* type when the slope of the response function against one of the predictor variables increases for higher levels of the other predictor variable (i.e., when β_3 is positive).

If the sign of β_3 in regression model (8.26) were negative:

$$E\{Y\} = 10 + 2X_1 + 5X_2 - .5X_1X_2 \qquad \textbf{(8.27)}$$

the result of the interaction effect of the two types of promotional expenditures on sales would be that the increase in sales with a unit increase in point-of-sale expenditures becomes smaller, the higher the level of TV expenditures. This effect is shown in the conditional effects plot in Figure 8.7c. The two response functions for $X_2 = 1$ and $X_2 = 3$ are again nonparallel, but now the slope of the response function is smaller for the higher level of TV expenditures. A similar decrease in the slope would occur if the response function against X_2 is considered for higher levels of X_1. When the regression coefficients β_1 and β_2 are positive, we say that the interaction effect between two quantitative variables is of an *interference* or *antagonistic* type when the slope of the response function against one of the predictor variables decreases for higher levels of the other predictor variable (i.e., when β_3 is negative).

Comments

1. When the signs of β_1 and β_2 in regression model (8.22) are negative, a negative β_3 is usually viewed as a reinforcement type of interaction effect and a positive β_3 as an interference type of effect.

2. To derive (8.23) and (8.24), we differentiate:

$$E\{Y\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2$$

with respect to X_1 and X_2, respectively:

$$\frac{\partial E\{Y\}}{\partial X_1} = \beta_1 + \beta_3 X_2 \qquad \frac{\partial E\{Y\}}{\partial X_2} = \beta_2 + \beta_3 X_1$$

■

Shape of Response Function. Figure 8.8 shows for the sales promotion example the impact of the interaction effect on the shape of the response function. Figure 8.8a presents the additive response function in (8.25), and Figures 8.8b and 8.8c present the response functions with the reinforcement interaction effect in (8.26) and with the interference interaction effect in (8.27), respectively. Note that the additive response function is a plane, but that the two response functions with interaction effects are not. Also note in Figures 8.8b and 8.8c that the mean response as a function of X_1, for any given level of X_2, is no longer parallel to the same function at a different level of X_2, for either type of interaction effect.

We can also illustrate the difference in the shape of the response function when the two predictor variables do and do not interact by representing the response surface by means of a contour diagram. As we noted previously, such a diagram shows for different response levels the various combinations of levels of the two predictor variables that yield the same level of response. Figure 8.8d shows a contour diagram for the additive response surface in Figure 8.8a when the two predictor variables do not interact. Note that the contour curves are straight lines and that the contour lines are parallel and hence equally spaced. Figures 8.8e and 8.8f show contour diagrams for the response surfaces in Figures 8.8b and 8.8c, respectively, where the two predictor variables interact. Note that the contour curves are no longer straight lines and that the contour curves are not parallel here. For instance, in Figure 8.8e the vertical distance between the contours for $E\{Y\} = 200$ and $E\{Y\} = 400$ at $X_1 = 10$ is much larger than at $X_1 = 50$.

In general, additive or noninteracting predictor variables lead to parallel contour curves, whereas interacting predictor variables lead to nonparallel contour curves.

Interpretation of Interaction Regression Models with Curvilinear Effects

When one or more of the predictor variables in a regression model have curvilinear effects on the response variable, the presence of interaction effects again leads to response functions whose contour curves are not parallel. Figure 8.9a shows the response surface for a study of the volume of a quick bread:

$$E\{Y\} = 65 + 3X_1 + 4X_2 - 10X_1^2 - 15X_2^2 + 35X_1X_2$$

Here, Y is the percentage increase in the volume of the quick bread from baking, X_1 is the amount of a leavening agent (coded), and X_2 is the oven temperature (coded). Figure 8.9b shows contour curves for this response function. Note the lack of parallelism in the contour curves, reflecting the interaction effect. Figure 8.10 presents a conditional effects plot to show in a simple fashion the nature of the interaction in the relation of oven temperature (X_2) to the mean volume when leavening agent amount (X_1) is held constant at different levels. Note that increasing oven temperature increases volume when leavening agent amount is high, and the opposite is true when leavening agent amount is low.

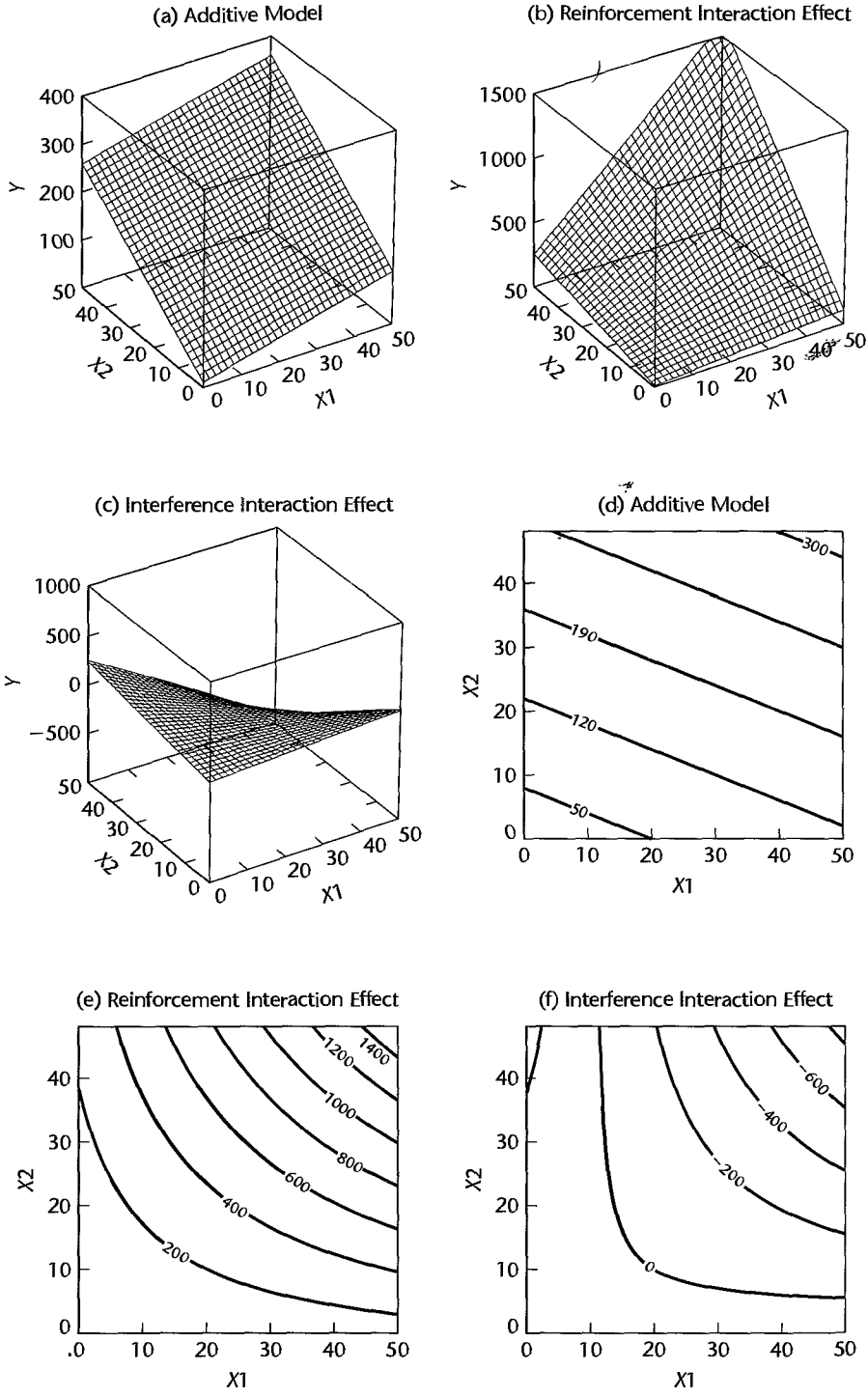


FIGURE 8.8 **Response Surfaces and Contour Plots for Additive and Interaction Regression Models—Sales Promotion Example.**

FIGURE 8.9 Response Surface and Contour Curves for Curvilinear Regression Model with Interaction Effect—Quick Bread Volume Example.

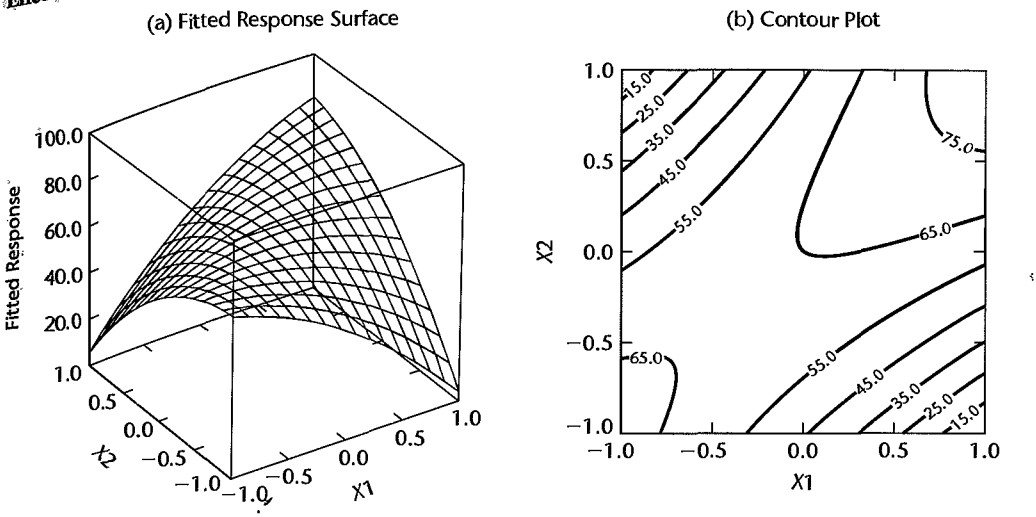
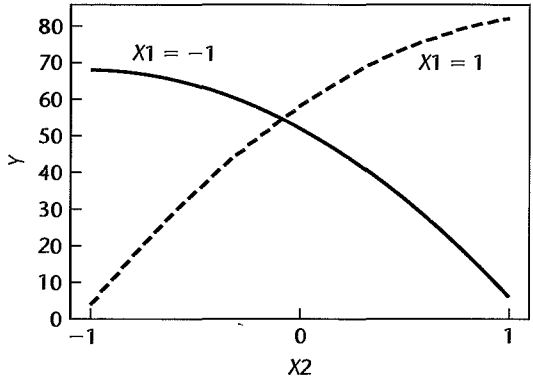


FIGURE 8.10 Conditional Effects Plot for Curvilinear Regression Model with Interaction Effect—Quick Bread Volume Example.



Implementation of Interaction Regression Models

The fitting of interaction regression models is routine, once the appropriate cross-product terms have been added to the data set. Two considerations need to be kept in mind when developing regression models with interaction effects.

1. When interaction terms are added to a regression model, high multicollinearities may exist between some of the predictor variables and some of the interaction terms, as well as among some of the interaction terms. A partial remedy to improve computational accuracy is to center the predictor variables; i.e., to use $x_{ik} = X_{ik} - \bar{X}_k$.

2. When the number of predictor variables in the regression model is large, the potential number of interaction terms can become very large. For example, if eight predictor

variables are present in the regression model in linear terms, there are potentially 28 pairwise interaction terms that could be added to the regression model. The data set would need to be quite large before 36 X variables could be used in the regression model.

It is therefore desirable to identify in advance, whenever possible, those interactions that are most likely to influence the response variable in important ways. In addition to utilizing *a priori* knowledge, one can plot the residuals for the additive regression model against the different interaction terms to determine which ones appear to be influential in affecting the response variable. When the number of predictor variables is large, these plots may need to be limited to interaction terms involving those predictor variables that appear to be the most important on the basis of the initial fit of the additive regression model.

Example

We wish to test formally in the body fat example of Table 7.1 whether interaction terms between the three predictor variables should be included in the regression model. We therefore need to consider the following regression model:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i1} X_{i2} + \beta_5 X_{i1} X_{i3} + \beta_6 X_{i2} X_{i3} + \varepsilon_i \quad \textbf{(8.28)}$$

This regression model requires that we obtain the new variables X_1X_2, X_1X_3, and X_2X_3 and add these X variables to the ones in Table 7.1. We find upon examining these X variables that some of the predictor variables are highly correlated with some of the interaction terms, and that there are also some high correlations among the interaction terms. For example, the correlation between X_1 and X_1X_2 is .989 and that between X_1X_3 and X_2X_3 is .998.

We shall therefore use centered variables in the regression model:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i1} x_{i2} + \beta_5 x_{i1} x_{i3} + \beta_6 x_{i2} x_{i3} + \varepsilon_i \quad \textbf{(8.29)}$$

where:

$$x_{i1} = X_{i1} - \bar{X}_1 = X_{i1} - 25.305$$
$$x_{i2} = X_{i2} - \bar{X}_2 = X_{i2} - 51.170$$
$$x_{i3} = X_{i3} - \bar{X}_3 = X_{i3} - 27.620$$

Upon obtaining the cross-product terms using the centered variables, we find that the intercorrelations involving the cross-product terms are now smaller. For example, the largest correlation, which was between X_1X_3 and X_2X_3, is reduced from .998 to .891. Other correlations are reduced in absolute magnitude even more.

Fitting regression model (8.29) yields the following estimated regression function, mean square error, and extra sums of squares:

$$\hat{Y} = 20.53 + 3.438x_1 - 2.095x_2 - 1.616x_3 + .00888x_1x_2 - .08479x_1x_3 + .09042x_2x_3$$
$$MSE = 6.745$$

Variable	Extra Sum of Squares
X_1	$SSR(X_1) = 352.270$
X_2	$SSR(X_2 \mid X_1) = 33.169$
X_3	$SSR(X_3 \mid X_1, X_2) = 11.546$
X_1X_2	$SSR(X_1X_2 \mid X_1, X_2, X_3) = 1.496$
X_1X_3	$SSR(X_1X_3 \mid X_1, X_2, X_3, X_1X_2) = 2.704$
X_2X_3	$SSR(X_2X_3 \mid X_1, X_2, X_3, X_1X_2, X_1X_3) = 6.515$

We wish to test whether any interaction terms are needed:

$$H_0: \beta_4 = \beta_5 = \beta_6 = 0$$
$$H_a: \text{not all } \beta s \text{ in } H_0 \text{ equal zero}$$

The partial F test statistic (7.27) requires here the following extra sum of squares:

$$SSR(X_1X_2, X_1X_3, X_2X_3 \mid X_1, X_2, X_3) = 1.496 + 2.704 + 6.515 = 10.715$$

and the test statistic is:

$$F^* = \frac{SSR(X_1X_2, X_1X_3, X_2X_3 \mid X_1, X_2, X_3)}{3} \div MSE$$
$$= \frac{10.715}{3} \div 6.745 = .53$$

For level of significance $\alpha = .05$, we require $F(.95; 3, 13) = 3.41$. Since $F^* = .53 \leq 3.41$, we conclude H_0, that the interaction terms are not needed in the regression model. The P-value of this test is .67.

8.3 Qualitative Predictors

As mentioned in Chapter 6, qualitative, as well as quantitative, predictor variables can be used in regression models. Many predictor variables of interest in business, economics, and the social and biological sciences are qualitative. Examples of qualitative predictor variables are gender (male, female), purchase status (purchase, no purchase), and disability status (not disabled, partly disabled, fully disabled).

In a study of innovation in the insurance industry, an economist wished to relate the speed with which a particular insurance innovation is adopted (Y) to the size of the insurance firm (X_1) and the type of firm. The response variable is measured by the number of months elapsed between the time the first firm adopted the innovation and the time the given firm adopted the innovation. The first predictor variable, size of firm, is quantitative, and is measured by the amount of total assets of the firm. The second predictor variable, type of firm, is qualitative and is composed of two classes—stock companies and mutual companies. In order that such a qualitative variable can be used in a regression model, quantitative indicators for the classes of the qualitative variable must be employed.

Qualitative Predictor with Two Classes

There are many ways of quantitatively identifying the classes of a qualitative variable. We shall use indicator variables that take on the values 0 and 1. These indicator variables are easy to use and are widely employed, but they are by no means the only way to quantify a qualitative variable.

For the insurance innovation example, where the qualitative predictor variable has two classes, we might define two indicator variables X_2 and X_3 as follows:

$$X_2 = \begin{cases} 1 & \text{if stock company} \\ 0 & \text{otherwise} \end{cases}$$
$$X_3 = \begin{cases} 1 & \text{if mutual company} \\ 0 & \text{otherwise} \end{cases} \tag{8.30}$$

A first-order model then would be the following:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \varepsilon_i \tag{8.31}$$

This intuitive approach of setting up an indicator variable for each class of the qualitative predictor variable unfortunately leads to computational difficulties. To see why, suppose we have $n = 4$ observations, the first two being stock firms (for which $X_2 = 1$ and $X_3 = 0$), and the second two being mutual firms (for which $X_2 = 0$ and $X_3 = 1$). The $\mathbf{X}$ matrix would then be:

$$\mathbf{X} = \begin{array}{c} \\ 1 \\ 1 \\ 1 \\ 1 \end{array} \begin{array}{c} \begin{array}{cccc} X_1 & X_2 & X_3 \end{array} \\ \begin{bmatrix} 1 & 1 & 0 \\ X_{11} & 1 & 0 \\ X_{21} & 1 & 0 \\ X_{31} & 0 & 1 \\ X_{41} & 0 & 1 \end{bmatrix} \end{array}$$

Note that the first column is equal to the sum of the X_2 and X_3 columns, so that the columns are linearly dependent according to definition (5.20). This has a serious effect on the $\mathbf{X}'\mathbf{X}$ matrix:

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ X_{11} & X_{21} & X_{31} & X_{41} \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{bmatrix} \begin{bmatrix} 1 & X_{11} & 1 & 0 \\ 1 & X_{21} & 1 & 0 \\ 1 & X_{31} & 0 & 1 \\ 1 & X_{41} & 0 & 1 \end{bmatrix}$$

$$= \begin{bmatrix} 4 & \sum_{i=1}^{4} X_{i1} & 2 & 2 \\ \sum_{i=1}^{4} X_{i1} & \sum_{i=1}^{4} X_{i1}^2 & \sum_{i=1}^{2} X_{i1} & \sum_{i=3}^{4} X_{i1} \\ 2 & \sum_{i=1}^{2} X_{i1} & 2 & 0 \\ 2 & \sum_{i=3}^{4} X_{i1} & 0 & 2 \end{bmatrix}$$

We see that the first column of the $\mathbf{X}'\mathbf{X}$ matrix equals the sum of the last two columns, so that the columns are linearly dependent. Hence, the $\mathbf{X}'\mathbf{X}$ matrix does not have an inverse, and no unique estimators of the regression coefficients can be found.

A simple way out of this difficulty is to drop one of the indicator variables. In our example, we might drop X_3. Dropping one indicator variable is not the only way out of the difficulty, but it leads to simple interpretations of the parameters. In general, therefore, we shall follow the principle:

> A qualitative variable with c classes will be represented by $c - 1$ indicator variables, each taking on the values 0 and 1. **(8.32)**

Comment

Indicator variables are frequently also called *dummy variables* or *binary variables*. The latter term has reference to the binary number system containing only 0 and 1. ■

Interpretation of Regression Coefficients

Returning to the insurance innovation example, suppose that we drop the indicator variable X_3 from regression model (8.31) so that the model becomes:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i \qquad \textbf{(8.33)}$$

where:

X_{i1} = size of firm

$$X_{i2} = \begin{cases} 1 & \text{if stock company} \\ 0 & \text{if mutual company} \end{cases}$$

The response function for this regression model is:

$$E\{Y\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \qquad \textbf{(8.34)}$$

To understand the meaning of the regression coefficients in this model, consider first the case of a mutual firm. For such a firm, $X_2 = 0$ and response function (8.34) becomes:

$$E\{Y\} = \beta_0 + \beta_1 X_1 + \beta_2(0) = \beta_0 + \beta_1 X_1 \qquad \text{Mutual firms} \qquad \textbf{(8.34a)}$$

Thus, the response function for mutual firms is a straight line, with Y intercept β_0 and slope β_1. This response function is shown in Figure 8.11.

For a stock firm, $X_2 = 1$ and response function (8.34) becomes:

$$E\{Y\} = \beta_0 + \beta_1 X_1 + \beta_2(1) = (\beta_0 + \beta_2) + \beta_1 X_1 \qquad \text{Stock firms} \qquad \textbf{(8.34b)}$$

This also is a straight line, with the same slope β_1 but with Y intercept $\beta_0 + \beta_2$. This response function is also shown in Figure 8.11.

Let us consider now the meaning of the regression coefficients in response function (8.34) with specific reference to the insurance innovation example. We see that the mean time elapsed before the innovation is adopted, $E\{Y\}$, is a linear function of size of firm (X_1), with the same slope β_1 for both types of firms. β_2 indicates how much higher (lower) the response function for stock firms is than the one for mutual firms, for any given size of firm. Thus, β_2 measures the differential effect of type of firm. In general, β_2 shows how much higher (lower) the mean response line is for the class coded 1 than the line for the class coded 0, for any given level of X_1.

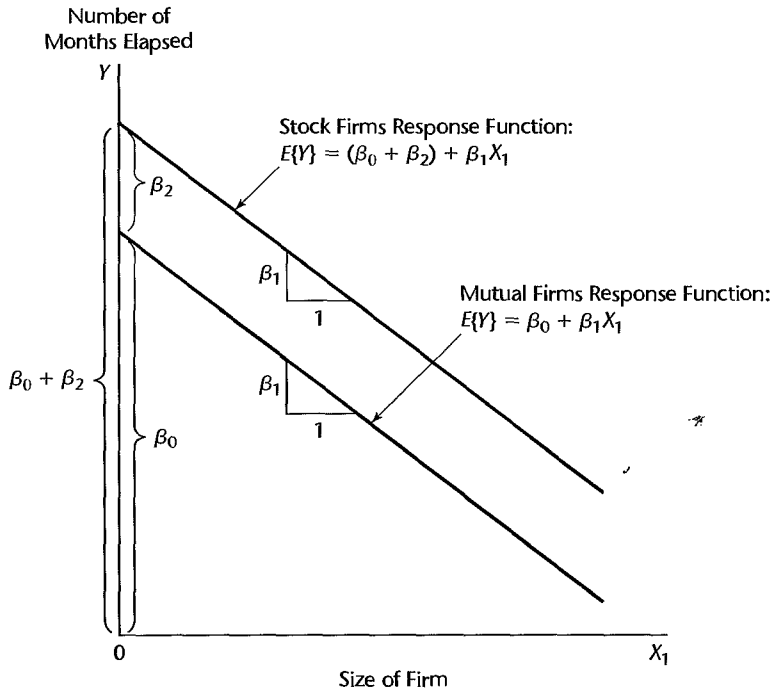


FIGURE 8.11 **Illustration of Meaning of Regression Coefficients for Regression Model (8.33) with Indicator Variable X_2—Insurance Innovation Example.**

Example

In the insurance innovation example, the economist studied 10 mutual firms and 10 stock firms. The basic data are shown in Table 8.2, columns 1–3. The indicator coding for type of firm is shown in column 4. Note that $X_2 = 1$ for each stock firm and $X_2 = 0$ for each mutual firm.

The fitting of regression model (8.33) is now straightforward. Table 8.3 presents the key results from a computer run regressing Y on X_1 and X_2. The fitted response function is:

$$\hat{Y} = 33.87407 - .10174X_1 + 8.05547X_2$$

Figure 8.12 contains the fitted response function for each type of firm, together with the actual observations.

The economist was most interested in the effect of type of firm (X_2) on the elapsed time for the innovation to be adopted and wished to obtain a 95 percent confidence interval for β_2. We require $t(.975; 17) = 2.110$ and obtain from the results in Table 8.3 the confidence limits $8.05547 \pm 2.110(1.45911)$. The confidence interval for β_2 therefore is:

$$4.98 \leq \beta_2 \leq 11.13$$

Thus, with 95 percent confidence, we conclude that stock companies tend to adopt the innovation somewhere between 5 and 11 months later, on the average, than mutual companies for any given size of firm.

A formal test of:

$$H_0: \beta_2 = 0$$
$$H_a: \beta_2 \neq 0$$

TABLE 8.2 Data and Indicator Coding—Insurance Innovation Example.

Firm i	(1) Number of Months Elapsed Y_i	(2) Size of Firm (million dollars) X_{i1}	(3) Type of Firm	(4) Indicator Code X_{i2}	(5) $X_{i1}X_{i2}$
1	17	151	Mutual	0	0
2	26	92	Mutual	0	0
3	21	175	Mutual	0	0
4	30	31	Mutual	0	0
5	22	104	Mutual	0	0
6	0	277	Mutual	0	0
7	12	210	Mutual	0	0
8	19	120	Mutual	0	0
9	4	290	Mutual	0	0
10	16	238	Mutual	0	0
11	28	164	Stock	1	164
12	15	272	Stock	1	272
13	11	295	Stock	1	295
14	38	68	Stock	1	68
15	31	85	Stock	1	85
16	21	224	Stock	1	224
17	20	166	Stock	1	166
18	13	305	Stock	1	305
19	30	124	Stock	1	124
20	14	246	Stock	1	246

TABLE 8.3 Regression Results for Fit of Regression Model (8.33)—Insurance Innovation Example.

(a) Regression Coefficients

Regression Coefficient	Estimated Regression Coefficient	Estimated Standard Deviation	t^*
β_0	33.87407	1.81386	18.68
β_1	−.10174	.00889	−11.44
β_2	8.05547	1.45911	5.52

(b) Analysis of Variance

Source of Variation	SS	df	MS
Regression	1,504.41	2	752.20
Error	176.39	17	10.38
Total	1,680.80	19	

with level of significance .05 would lead to H_a, that type of firm has an effect, since the 95 percent confidence interval for β_2 does not include zero.

The economist also carried out other analyses, some of which will be described shortly.

Comment

The reader may wonder why we did not simply fit separate regressions for stock firms and mutual firms in our example, and instead adopted the approach of fitting one regression with an indicator

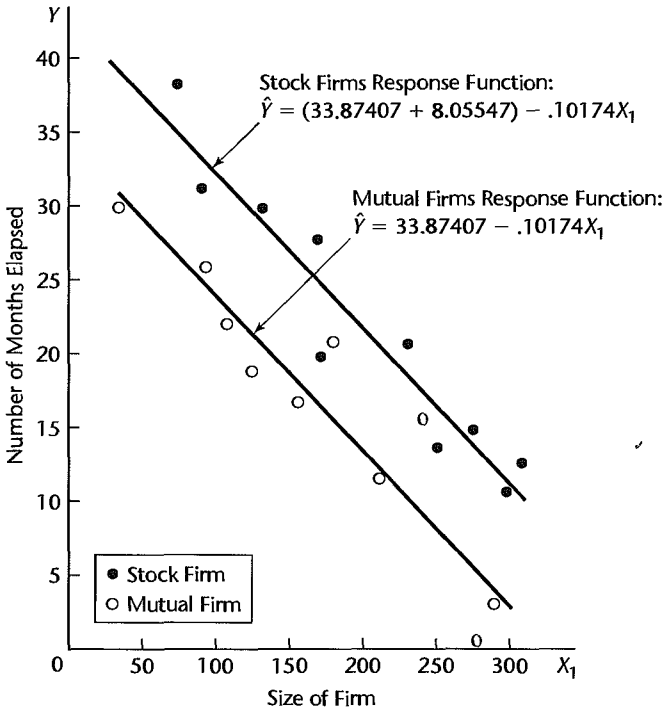


FIGURE 8.12 **Fitted Regression Functions for Regression Model (8.33)—Insurance Innovation Example.**

variable. There are two reasons for this. Since the model assumes equal slopes and the same constant error term variance for each type of firm, the common slope β_1 can best be estimated by pooling the two types of firms. Also, other inferences, such as for β_0 and β_2, can be made more precisely by working with one regression model containing an indicator variable since more degrees of freedom will then be associated with *MSE*. ■

Qualitative Predictor with More than Two Classes

If a qualitative predictor variable has more than two classes, we require additional indicator variables in the regression model. Consider the regression of tool wear (Y) on tool speed (X_1) and tool model, where the latter is a qualitative variable with four classes (M1, M2, M3, M4). We therefore require three indicator variables. Let us define them as follows:

$$X_2 = \begin{cases} 1 & \text{if tool model M1} \\ 0 & \text{otherwise} \end{cases}$$

$$X_3 = \begin{cases} 1 & \text{if tool model M2} \\ 0 & \text{otherwise} \end{cases} \qquad (8.35)$$

$$X_4 = \begin{cases} 1 & \text{if tool model M3} \\ 0 & \text{otherwise} \end{cases}$$

First-Order Model. A first-order regression model is:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i4} + \varepsilon_i \qquad (8.36)$$

For this model, the data input for the X variables would be as follows:

Tool Model	X_1	X_2	X_3	X_4
M1	X_{i1}	1	0	0
M2	X_{i1}	0	1	0
M3	X_{i1}	0	0	1
M4	X_{i1}	0	0	0

The response function for regression model (8.36) is:

$$E\{Y\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 \qquad \textbf{(8.37)}$$

To understand the meaning of the regression coefficients, consider first what response function (8.37) becomes for tool models M4 for which $X_2 = 0$, $X_3 = 0$, and $X_4 = 0$:

$$E\{Y\} = \beta_0 + \beta_1 X_1 \qquad \text{Tool models M4} \qquad \textbf{(8.37a)}$$

For tool models M1, $X_2 = 1$, $X_3 = 0$, and $X_4 = 0$, and response function (8.37) becomes:

$$E\{Y\} = (\beta_0 + \beta_2) + \beta_1 X_1 \qquad \text{Tool models M1} \qquad \textbf{(8.37b)}$$

Similarly, response functions (8.37) becomes for tool models M2 and M3:

$$E\{Y\} = (\beta_0 + \beta_3) + \beta_1 X_1 \qquad \text{Tool models M2} \qquad \textbf{(8.37c)}$$

$$E\{Y\} = (\beta_0 + \beta_4) + \beta_1 X_1 \qquad \text{Tool models M3} \qquad \textbf{(8.37d)}$$

Thus, response function (8.37) implies that the regression of tool wear on tool speed is linear, with the same slope for all four tool models. The coefficients β_2, β_3, and β_4 indicate, respectively, how much higher (lower) the response functions for tool models M1, M2, and M3 are than the one for tool models M4, for any given level of tool speed. Thus, β_2, β_3, and β_4 measure the differential effects of the qualitative variable classes on the height of the response function for any given level of X_1, always compared with the class for which $X_2 = X_3 = X_4 = 0$. Figure 8.13 illustrates a possible arrangement of the response functions.

When using regression model (8.36), we may wish to estimate differential effects other than against tool models M4. This can be done by estimating differences between regression coefficients. For instance, $\beta_4 - \beta_3$ measures how much higher (lower) the response function for tool models M3 is than the response function for tool models M2 for any given level of tool speed, as may be seen by comparing (8.37c) and (8.37d). The point estimator of this quantity is, of course, $b_4 - b_3$, and the estimated variance of this estimator is:

$$s^2\{b_4 - b_3\} = s^2\{b_4\} + s^2\{b_3\} - 2s\{b_4, b_3\} \qquad \textbf{(8.38)}$$

The needed variances and covariance can be readily obtained from the estimated variance-covariance matrix of the regression coefficients.

Time Series Applications

Economists and business analysts frequently use time series data in regression analysis. Indicator variables often are useful for time series regression models. For instance, savings (Y) may be regressed on income (X), where both the savings and income data are annual

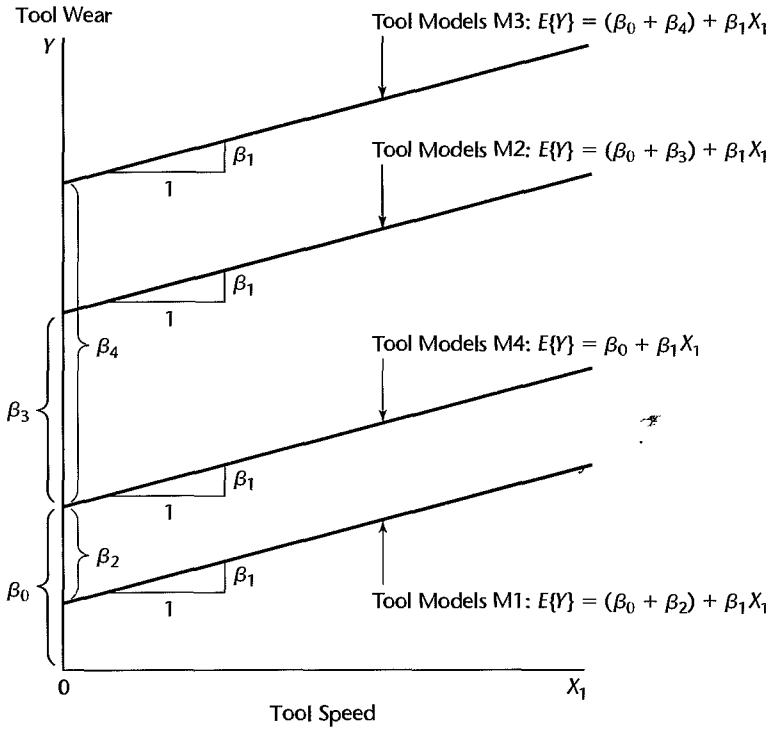


FIGURE 8.13 **Illustration of Regression Model (8.36)—Tool Wear Example.**

data for a number of years. The model employed might be:

$$Y_t = \beta_0 + \beta_1 X_t + \varepsilon_t \qquad t = 1, \ldots, n \tag{8.39}$$

where Y_t and X_t are savings and income, respectively, for time period t. Suppose that the period covered includes both peacetime and wartime years, and that this factor should be recognized since savings in wartime years tend to be higher. The following model might then be appropriate:

$$Y_t = \beta_0 + \beta_1 X_{t1} + \beta_2 X_{t2} + \varepsilon_t \tag{8.40}$$

where:

$X_{t1} =$ income

$$X_{t2} = \begin{cases} 1 & \text{if period } t \text{ peacetime} \\ 0 & \text{otherwise} \end{cases}$$

Note that regression model (8.40) assumes that the marginal propensity to save (β_1) is constant in both peacetime and wartime years, and that only the height of the response function is affected by this qualitative variable.

Another use of indicator variables in time series applications occurs when monthly or quarterly data are used. Suppose that quarterly sales (Y) are regressed on quarterly advertising expenditures (X_1) and quarterly disposable personal income (X_2). If seasonal effects also have an influence on quarterly sales, a first-order regression model incorporating

seasonal effects would be:

$$Y_t = \beta_0 + \beta_1 X_{t1} + \beta_2 X_{t2} + \beta_3 X_{t3} + \beta_4 X_{t4} + \beta_5 X_{t5} + \varepsilon_t \quad (8.41)$$

where:

X_{t1} = quarterly advertising expenditures
X_{t2} = quarterly disposable personal income
$X_{t3} = \begin{cases} 1 & \text{if first quarter} \\ 0 & \text{otherwise} \end{cases}$
$X_{t4} = \begin{cases} 1 & \text{if second quarter} \\ 0 & \text{otherwise} \end{cases}$
$X_{t5} = \begin{cases} 1 & \text{if third quarter} \\ 0 & \text{otherwise} \end{cases}$

Regression models for time series data are susceptible to correlated error terms. It is particularly important in these cases to examine whether the modeling of the time series components of the data is adequate to make the error terms uncorrelated. We discuss in Chapter 12 a test for correlated error terms and a regression model that is often useful when the error terms are correlated.

8.4 Some Considerations in Using Indicator Variables

Indicator Variables versus Allocated Codes

An alternative to the use of indicator variables for a qualitative predictor variable is to employ *allocated codes*. Consider, for instance, the predictor variable "frequency of product use" which has three classes: frequent user, occasional user, nonuser. With the allocated codes approach, a single X variable is employed and values are assigned to the classes; for instance:

Class	X_1
Frequent user	3
Occasional user	2
Nonuser	1

The allocated codes are, of course, arbitrary and could be other sets of numbers. The first-order model with allocated codes for our example, assuming no other predictor variables, would be:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \varepsilon_i \quad (8.42)$$

The basic difficulty with allocated codes is that they define a metric for the classes of the qualitative variable that may not be reasonable. To see this concretely, consider the mean

responses with regression model (8.42) for the three classes of the qualitative variable:

Class	$E\{Y\}$
Frequent user	$E\{Y\} = \beta_0 + \beta_1(3) = \beta_0 + 3\beta_1$
Occasional user	$E\{Y\} = \beta_0 + \beta_1(2) = \beta_0 + 2\beta_1$
Nonuser	$E\{Y\} = \beta_0 + \beta_1(1) = \beta_0 + \beta_1$

Note the key implication:

$$E\{Y|\text{frequent user}\} - E\{Y|\text{occasional user}\} = E\{Y|\text{occasional user}\} - E\{Y|\text{nonuser}\} = \beta_1$$

Thus, the coding 1, 2, 3 implies that the mean response changes by the same amount when going from a nonuser to an occasional user as when going from an occasional user to a frequent user. This may not be in accord with reality and is the result of the coding 1, 2, 3, which assigns equal distances between the three user classes. Other allocated codes may, of course, imply different spacings of the classes of the qualitative variable, but these would ordinarily still be arbitrary.

Indicator variables, in contrast, make no assumptions about the spacing of the classes and rely on the data to show the differential effects that occur. If, for the same example, two indicator variables, say, X_1 and X_2, are employed to represent the qualitative variable, as follows:

Class	X_1	X_2
Frequent user	1	0
Occasional user	0	1
Nonuser	0	0

the first-order regression model would be:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i \tag{8.43}$$

Here, β_1 measures the differential effect:

$$E\{Y|\text{frequent user}\} - E\{Y|\text{nonuser}\}$$

and β_2 measures the differential effect:

$$E\{Y|\text{occasional user}\} - E\{Y|\text{nonuser}\}$$

Thus, β_2 measures the differential effect between occasional user and nonuser, and $\beta_1 - \beta_2$ measures the differential effect between frequent user and occasional user. Notice that there are no arbitrary restrictions to be satisfied by these two differential effects. Also note that if $\beta_1 = 2\beta_2$, then equal spacing between the three classes would exist.

Indicator Variables versus Quantitative Variables

Indicator variables can be used even if the predictor variable is quantitative. For instance, the quantitative variable age may be transformed by grouping ages into classes such as under

21, 21–34, 35–49, etc. Indicator variables are then used for the classes of this new predictor variable. At first sight, this may seem to be a questionable approach because information about the actual ages is thrown away. Furthermore, additional parameters are placed into the model, which leads to a reduction of the degrees of freedom associated with *MSE*.

Nevertheless, there are occasions when replacement of a quantitative variable by indicator variables may be appropriate. Consider a large-scale survey in which the relation between liquid assets (Y) and age (X) of head of household is to be studied. Two thousand households were included in the study, so that the loss of 10 or 20 degrees of freedom is immaterial. The analyst is very much in doubt about the shape of the regression function, which could be highly complex, and hence may utilize the indicator variable approach in order to obtain information about the shape of the response function without making any assumptions about its functional form.

Thus, for large data sets use of indicator variables can serve as an alternative to lowess and other nonparametric fits of the response function.

Other Codings for Indicator Variables

As stated earlier, many different codings of indicator variables are possible. We now describe two alternatives to our 0, 1 coding for $c - 1$ indicator variables for a qualitative variable with c classes. We illustrate these alternative codings for the insurance innovation example, where Y is time to adopt an innovation, X_1 is size of insurance firm, and the second predictor variable is type of company (stock, mutual).

The first alternative coding is:

$$X_2 = \begin{cases} 1 & \text{if stock company} \\ -1 & \text{if mutual company} \end{cases} \tag{8.44}$$

For this coding, the first-order linear regression model:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i \tag{8.45}$$

has the response function:

$$E\{Y\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \tag{8.46}$$

This response function becomes for the two types of companies:

$$E\{Y\} = (\beta_0 + \beta_2) + \beta_1 X_1 \qquad \text{Stock firms} \tag{8.46a}$$

$$E\{Y\} = (\beta_0 - \beta_2) + \beta_1 X_1 \qquad \text{Mutual firms} \tag{8.46b}$$

Thus, β_0 here may be viewed as an "average" intercept of the regression line, from which the stock company and mutual company intercepts differ by β_2 in opposite directions. A test whether the regression lines are the same for both types of companies involves H_0: $\beta_2 = 0$, H_a: $\beta_2 \neq 0$.

A second alternative coding scheme is to use a 0, 1 indicator variable for each of the c classes of the qualitative variable and to drop the intercept term in the regression model. For the insurance innovation example, the model would be:

$$Y_i = \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \varepsilon_i \tag{8.47}$$

where:

$$X_{i1} = \text{size of firm}$$

$$X_{i2} = \begin{cases} 1 & \text{if stock company} \\ 0 & \text{otherwise} \end{cases}$$

$$X_{i3} = \begin{cases} 1 & \text{if mutual company} \\ 0 & \text{otherwise} \end{cases}$$

Here, the two response functions are:

$$E\{Y\} = \beta_2 + \beta_1 X_1 \qquad \text{Stock firms} \qquad \textbf{(8.48a)}$$

$$E\{Y\} = \beta_3 + \beta_1 X_1 \qquad \text{Mutual firms} \qquad \textbf{(8.48b)}$$

A test of whether or not the two regression lines are the same would involve the alternatives H_0: $\beta_2 = \beta_3$, H_a: $\beta_2 \neq \beta_3$. This type of test, discussed in Section 7.3, cannot be conducted by using extra sums of squares and requires the fitting of both the full and reduced models.

8.5 Modeling Interactions between Quantitative and Qualitative Predictors

In the insurance innovation example, the economist actually did not begin the analysis with regression model (8.33) because of the possibility of interaction effects between size of firm and type of firm on the response variable. Even though one of the predictor variables in the regression model here is qualitative, interaction effects can still be introduced into the model in the usual manner, by including cross-product terms. A first-order regression model with an added interaction term for the insurance innovation example is:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i1} X_{i2} + \varepsilon_i \qquad \textbf{(8.49)}$$

where:

$$X_{i1} = \text{size of firm}$$

$$X_{i2} = \begin{cases} 1 & \text{if stock company} \\ 0 & \text{otherwise} \end{cases}$$

The response function for this regression model is:

$$E\{Y\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 \qquad \textbf{(8.50)}$$

Meaning of Regression Coefficients

The meaning of the regression coefficients in response function (8.50) can best be understood by examining the nature of this function for each type of firm. For a mutual firm, $X_2 = 0$ and hence $X_1 X_2 = 0$. Response function (8.50) therefore becomes for mutual firms:

$$E\{Y\} = \beta_0 + \beta_1 X_1 + \beta_2(0) + \beta_3(0) = \beta_0 + \beta_1 X_1 \qquad \text{Mutual firms} \qquad \textbf{(8.50a)}$$

This response function is shown in Figure 8.14. Note that the Y intercept is β_0 and the slope is β_1 for the response function for mutual firms.

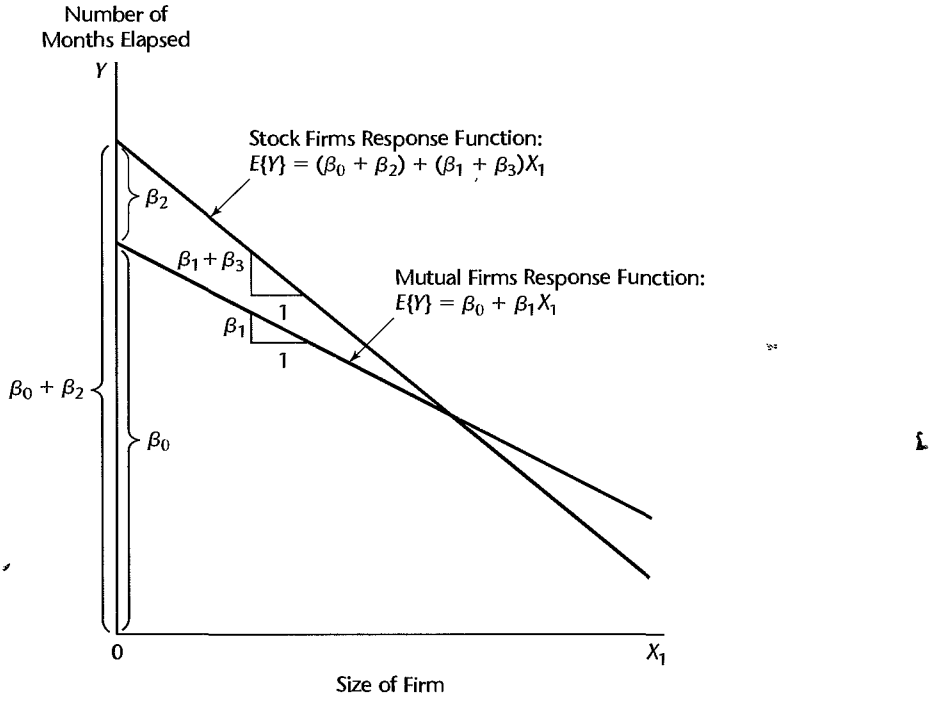


FIGURE 8.14 Illustration of Meaning of Regression Coefficients for Regression Model (8.49) with Indicator Variable X_2 and Interaction Term—Insurance Innovation Example.

For stock firms, $X_2 = 1$ and hence $X_1X_2 = X_1$. Response function (8.50) therefore becomes for stock firms:

$$E\{Y\} = \beta_0 + \beta_1 X_1 + \beta_2(1) + \beta_3 X_1$$

or:

$$E\{Y\} = (\beta_0 + \beta_2) + (\beta_1 + \beta_3)X_1 \qquad \text{Stock firms} \qquad \textbf{(8.50b)}$$

This response function is also shown in Figure 8.14. Note that the response function for stock firms has Y intercept $\beta_0 + \beta_2$ and slope $\beta_1 + \beta_3$.

We see that β_2 here indicates how much greater (smaller) is the Y intercept of the response function for the class coded 1 than that for the class coded 0. Similarly, β_3 indicates how much greater (smaller) is the slope of the response function for the class coded 1 than that for the class coded 0. Because both the intercept and the slope differ for the two classes in regression model (8.49), it is no longer true that β_2 indicates how much higher (lower) one response function is than the other for any given level of X_1. Figure 8.14 shows that the effect of type of firm with regression model (8.49) depends on X_1, the size of the firm. For smaller firms, according to Figure 8.14, mutual firms tend to innovate more quickly, but for larger firms stock firms tend to innovate more quickly. Thus, when interaction effects are present, the effect of the qualitative predictor variable can be studied only by comparing the regression functions within the scope of the model for the different classes of the qualitative variable.

Figure 8.15 illustrates another possible interaction pattern for the insurance innovation example. Here, mutual firms tend to introduce the innovation more quickly than stock firms

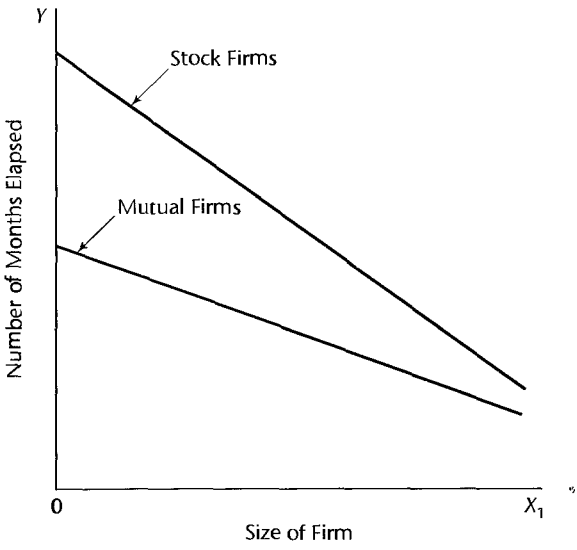


FIGURE 8.15 Another Illustration of Regression Model (8.49) with Indicator Variable X_2 and Interaction Term—Insurance Innovation Example.

for all sizes of firms in the scope of the model. but the differential effect is much smaller for large firms than for small ones.

The interactions portrayed in Figures 8.14 and 8.15 can no longer be viewed as reinforcing or interfering types of interactions because one of the predictor variables here is qualitative. When one of the predictor variables is qualitative and the other quantitative, nonparallel response functions that do not intersect within the scope of the model (as in Figure 8.15) are sometimes said to represent an *ordinal interaction*. When the response functions intersect within the scope of the model (as in Figure 8.14), the interaction is then said to be a *disordinal interaction*.

Example

Since the economist was concerned that interaction effects between size and type of firm may be present, the initial regression model fitted was model (8.49):

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i1} X_{i2} + \varepsilon_i$$

The values for the interaction term $X_1 X_2$ for the insurance innovation example are shown in Table 8.2, column 5, on page 317. Note that this column contains 0 for mutual companies and X_{i1} for stock companies.

Again, the regression fit is routine. Basic results from a computer run regressing Y on X_1, X_2, and $X_1 X_2$ are shown in Table 8.4. To test for the presence of interaction effects:

$$H_0: \beta_3 = 0$$
$$H_a: \beta_3 \neq 0$$

the economist used the t^* statistic from Table 8.4a:

$$t^* = \frac{b_3}{s\{b_3\}} = \frac{-.0004171}{.01833} = -.02$$

TABLE 8.4 Regression Results for Fit of Regression Model (8.49) with Interaction Term—Insurance Innovation Example.

(a) Regression Coefficients

Regression Coefficient	Estimated Regression Coefficient	Estimated Standard Deviation	t^*
β_0	33.83837	2.44065	13.86
β_1	−.10153	.01305	−7.78
β_2	8.13125	3.65405	2.23
β_3	−.0004171	.01833	−.02

(b) Analysis of Variance

Source of Variation	*SS*	*df*	*MS*
Regression	1,504.42	3	501.47
Error	176.38	16	11.02
Total	1,680.80	19	

For level of significance .05, we require $t(.975; 16) = 2.120$. Since $|t^*| = .02 \leq 2.120$, we conclude H_0, that $\beta_3 = 0$. The conclusion of no interaction effects is supported by the two-sided P-value for the test, which is very high, namely, .98. It was because of this result that the economist adopted regression model (8.33) with no interaction term, which we discussed earlier.

Comment

Fitting regression model (8.49) yields the same response functions as would fitting separate regressions for stock firms and mutual firms. An advantage of using model (8.49) with an indicator variable is that one regression run will yield both fitted regressions.

Another advantage is that tests for comparing the regression functions for the different classes of the qualitative variable can be clearly seen to involve tests of regression coefficients in a general linear model. For instance, Figure 8.14 for the insurance innovation example shows that a test of whether the two regression functions have the same slope involves:

$$H_0: \beta_3 = 0$$
$$H_a: \beta_3 \neq 0$$

Similarly, Figure 8.14 shows that a test of whether the two regression functions are identical involves:

$$H_0: \beta_2 = \beta_3 = 0$$
$$H_a: \text{not both } \beta_2 = 0 \text{ and } \beta_3 = 0$$

■

8.6 More Complex Models

We now briefly consider more complex models involving quantitative and qualitative predictor variables.

More than One Qualitative Predictor Variable

Regression models can readily be constructed for cases where two or more of the predictor variables are qualitative. Consider the regression of advertising expenditures (Y) on sales (X_1), type of firm (incorporated, not incorporated), and quality of sales management (high, low). We may define:

$$X_2 = \begin{cases} 1 & \text{if firm incorporated} \\ 0 & \text{otherwise} \end{cases}$$
$$X_3 = \begin{cases} 1 & \text{if quality of sales management high} \\ 0 & \text{otherwise} \end{cases} \quad (8.51)$$

First-Order Model. A first-order regression model for the above example is:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \varepsilon_i \quad (8.52)$$

This model implies that the response function of advertising expenditures on sales is linear, with the same slope for all "type of firm—quality of sales management" combinations, and β_2 and β_3 indicate the additive differential effects of type of firm and quality of sales management on the height of the regression line for any given levels of X_1 and the other predictor variable.

First-Order Model with Certain Interactions Added. A first-order regression model to which are added interaction effects between each pair of the predictor variables for the advertising example is:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i1} X_{i2} + \beta_5 X_{i1} X_{i3} + \beta_6 X_{i2} X_{i3} + \varepsilon_i \quad (8.53)$$

Note the implications of this model:

Type of Firm	Quality of Sales Management	Response Function
Incorporated	High	$E\{Y\} = (\beta_0 + \beta_2 + \beta_3 + \beta_6) + (\beta_1 + \beta_4 + \beta_5)X_1$
Not incorporated	High	$E\{Y\} = (\beta_0 + \beta_3) + (\beta_1 + \beta_5)X_1$
Incorporated	Low	$E\{Y\} = (\beta_0 + \beta_2) + (\beta_1 + \beta_4)X_1$
Not incorporated	Low	$E\{Y\} = \beta_0 + \beta_1 X_1$

Not only are all response functions different for the various "type of firm—quality of sales management" combinations, but the differential effects of one qualitative variable on the intercept depend on the particular class of the other qualitative variable. For instance, when we move from "not incorporated—low quality" to "incorporated—low quality," the intercept changes by β_2. But if we move from "not incorporated—high quality" to "incorporated—high quality," the intercept changes by $\beta_2 + \beta_6$.

Qualitative Predictor Variables Only

Regression models containing only qualitative predictor variables can also be constructed. With reference to our advertising example, we could regress advertising expenditures only on type of firm and quality of sales management. The first-order regression model then would be:

$$Y_i = \beta_0 + \beta_2 X_{i2} + \beta_3 X_{i3} + \varepsilon_i \qquad (8.54)$$

where X_{i2} and X_{i3} are defined in (8.51).

Comments

1. Models in which all explanatory variables are qualitative are called *analysis of variance models.*

2. Models containing some quantitative and some qualitative explanatory variables, where the chief explanatory variables of interest are qualitative and the quantitative variables are introduced primarily to reduce the variance of the error terms, are called *analysis of covariance models.*

■

8.7 Comparison of Two or More Regression Functions

Frequently we encounter regressions for two or more populations and wish to study their similarities and differences. We present three examples.

1. A company operates two production lines for making soap bars. For each line, the relation between the speed of the line and the amount of scrap for the day was studied. A scatter plot of the data for the two production lines suggests that the regression relation between production line speed and amount of scrap is linear but not the same for the two production lines. The slopes appear to be about the same, but the heights of the regression lines seem to differ. A formal test is desired to determine whether or not the two regression lines are identical. If it is found that the two regression lines are not the same, an investigation is to be made of why the difference in scrap yield exists.

2. An economist is studying the relation between amount of savings and level of income for middle-income families from urban and rural areas, based on independent samples from the two populations. Each of the two relations can be modeled by linear regression. The economist wishes to compare whether, at given income levels, urban and rural families tend to save the same amount—i.e., whether the two regression lines are the same. If they are not, the economist wishes to explore whether at least the amounts of savings out of an additional dollar of income are the same for the two groups—i.e., whether the slopes of the two regression lines are the same.

3. Two instruments were constructed for a company to identical specifications to measure pressure in an industrial process. A study was then made for each instrument of the relation between its gauge readings and actual pressures as determined by an almost exact but slow and costly method. If the two regression lines are the same, a single calibration schedule can be developed for the two instruments; otherwise, two different calibration schedules will be required.

When it is reasonable to assume that the error term variances in the regression models for the different populations are equal, we can use indicator variables to test the equality of the different regression functions. If the error variances are not equal, transformations of the response variable may equalize them at least approximately.

We have already seen how regression models with indicator variables that contain interaction terms permit testing of the equality of regression functions for the different classes of a qualitative variable. This methodology can be used directly for testing the equality of regression functions for different populations. We simply consider the different populations as classes of a predictor variable, define indicator variables for the different populations, and develop a regression model containing appropriate interaction terms. Since no new principles arise in the testing of the equality of regression functions for different populations, we immediately proceed with two of the earlier examples to illustrate the approach.

Soap Production Lines Example

The data on amount of scrap (Y) and line speed (X_1) for the soap production lines example are presented in Table 8.5. The variable X_2 is a code for the production line. A symbolic scatter plot of the data, using different symbols for the two production lines, is shown in Figure 8.16.

Tentative Model. On the basis of the symbolic scatter plot in Figure 8.16, the analyst decided to tentatively fit regression model (8.49). This model assumes that the regression relation between amount of scrap and line speed is linear for both production lines and that the variances of the error terms are the same, but permits the two regression lines to have different slopes and intercepts:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i1} X_{i2} + \varepsilon_i \qquad (8.55)$$

TABLE 8.5 Data—Soap Production Lines Example (all data are coded).

Production Line 1				Production Line 2			
Case i	Amount of Scrap Y_i	Line Speed X_{i1}	X_{i2}	Case i	Amount of Scrap Y_i	Line Speed X_{i1}	X_{i2}
1	218	100	1	16	140	105	0
2	248	125	1	17	277	215	0
3	360	220	1	18	384	270	0
4	351	205	1	19	341	255	0
5	470	300	1	20	215	175	0
6	394	255	1	21	180	135	0
7	332	225	1	22	260	200	0
8	321	175	1	23	361	275	0
9	410	270	1	24	252	155	0
10	260	170	1	25	422	320	0
11	241	155	1	26	273	190	0
12	331	190	1	27	410	295	0
13	275	140	1				
14	425	290	1				
15	367	265	1				

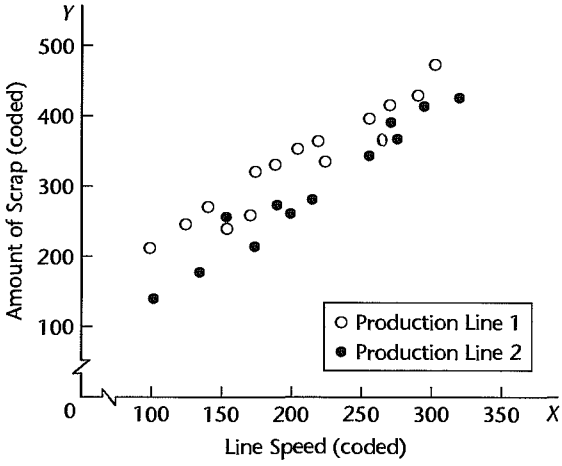


FIGURE 8.16 **Symbolic Scatter Plot—Soap Production Lines Example.**

where:

$$X_{i1} = \text{line speed}$$

$$X_{i2} = \begin{cases} 1 & \text{if production line 1} \\ 0 & \text{if production line 2} \end{cases}$$

$$i = 1, 2, \ldots, 27$$

Note that for purposes of this model, the 15 cases for production line 1 and the 12 cases for production line 2 are combined into one group of 27 cases.

Diagnostics. A fit of regression model (8.55) to the data in Table 8.5 led to the results presented in Table 8.6 and the following fitted regression function:

$$\hat{Y} = 7.57 + 1.322X_1 + 90.39X_2 - .1767X_1X_2$$

Plots of the residuals against $\hat{Y}$ are shown in Figure 8.17 for each production line. Two plots are used in order to facilitate the diagnosis of possible differences between the two production lines. Both plots in Figure 8.17 are reasonably consistent with regression model (8.55). The splits between positive and negative residuals of 10 to 5 for production line 1 and 4 to 8 for production line 2 can be accounted for by randomness of the outcomes. Plots of the residuals against X_2 and a normal probability plot of the residuals (not shown) also support the appropriateness of the fitted model. For the latter plot, the coefficient of correlation between the ordered residuals and their expected values under normality is .990. This is sufficiently high according to Table B.6 to support the assumption of normality of the error terms.

Finally, the analyst desired to make a formal test of the equality of the variances of the error terms for the two production lines, using the Brown-Forsythe test described in Section 3.6. Separate linear regression models were fitted to the data for the two production lines, the residuals were obtained, and the absolute deviations d_{i1} and d_{i2} in (3.8) of the

TABLE 8.6
Regression Results for Fit of Regression Model (8.55)—Soap Production Lines Example.

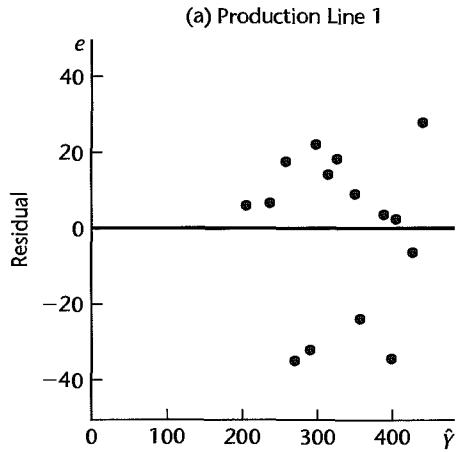
(a) Regression Coefficients

Regression Coefficient	Estimated Regression Coefficient	Estimated Standard Deviation
β_0	7.57	20.87
β_1	1.322	.09262
β_2	90.39	28.35
β_3	−.1767	.1288

(b) Analysis of Variance

Source of Varlation	*SS*	*df*
Regression	169,165	3
X_1	149,661	1
$X_2 \mid X_1$	18,694	1
$X_1X_2 \mid X_1, X_2$	810	1
Error	9,904	23
Total	179,069	26

FIGURE 8.17
Residual Plots against $\hat{Y}$—Soap Production Lines Example.



residuals around the median residual for each
The results were as follows:

Production Line 1
$\hat{Y} = 97.965 + 1.145X_1$
$\bar{d}_1 = 16.132$
$\sum(d_{i1} - \bar{d}_1)^2 = 2{,}952.20$

The pooled variance s^2 in (3.9a) therefore is:

$$s^2 = \frac{2{,}952.20 + 2{,}045.82}{27 - 2} = 199.921$$

Hence, the pooled standard deviation is $s = 14.139$, and the test statistic in (3.9) is:

$$t^*_{BF} = \frac{16.132 - 12.648}{14.139\sqrt{\frac{1}{15} + \frac{1}{12}}} = .636$$

For $\alpha = .05$, we require $t(.975; 25) = 2.060$. Since $|t^*_{BF}| = .636 \leq 2.060$, we conclude that the error term variances for the two production lines do not differ. The two-sided P-value for this test is .53.

At this point, the analyst was satisfied about the aptness of regression model (8.55) with normal error terms and was ready to proceed with comparing the regression relation between amount of scrap and line speed for the two production lines.

Inferences about Two Regression Lines. Identity of the regression functions for the two production lines is tested by considering the alternatives:

$$\begin{aligned} &H_0: \beta_2 = \beta_3 = 0 \\ &H_a: \text{not both } \beta_2 = 0 \text{ and } \beta_3 = 0 \end{aligned} \tag{8.56}$$

The appropriate test statistic is given by (7.27):

$$F^* = \frac{SSR(X_2, X_1X_2|X_1)}{2} \div \frac{SSE(X_1, X_2, X_1X_2)}{n - 4} \tag{8.56a}$$

where n represents the combined sample size for both populations. Using the regression results in Table 8.6, we find:

$$\begin{aligned} SSR(X_2, X_1X_2|X_1) &= SSR(X_2|X_1) + SSR(X_1X_2|X_1, X_2) \\ &= 18{,}694 + 810 = 19{,}504 \\ F^* &= \frac{19{,}504}{2} \div \frac{9{,}904}{23} = 22.65 \end{aligned}$$

To control α at level .01, we require $F(.99; 2, 23) = 5.67$. Since $F^* = 22.65 > 5.67$, we conclude H_a, that the regression functions for the two production lines are not identical.

Next, the analyst examined whether the slopes of the regression lines are the same. The alternatives here are:

$$\begin{aligned} &H_0: \beta_3 = 0 \\ &H_a: \beta_3 \neq 0 \end{aligned} \tag{8.57}$$

and the appropriate test statistic is either the t^* statistic (7.25) or the partial F test statistic (7.24):

$$F^* = \frac{SSR(X_1X_2|X_1, X_2)}{1} \div \frac{SSE(X_1, X_2, X_1X_2)}{n - 4} \tag{8.57a}$$

Using the regression results in Table 8.6 and the partial F test statistic, we obtain:

$$F^* = \frac{810}{1} \div \frac{9{,}904}{23} = 1.88$$

For $\alpha = .01$, we require $F(.99; 1, 23) = 7.88$. Since $F^* = 1.88 \leq 7.88$, we conclude H_0, that the slopes of the regression functions for the two production lines are the same.

Using the Bonferroni inequality (4.2), the analyst can therefore conclude at family significance level .02 that a given increase in line speed leads to the same amount of increase in expected scrap in each of the two production lines, but that the expected amount of scrap for any given line speed differs by a constant amount for the two production lines.

We can estimate this constant difference in the regression lines by obtaining a confidence interval for β_2. For a 95 percent confidence interval, we require $t(.975; 23) = 2.069$. Using the results in Table 8.6, we obtain the confidence limits $90.39 \pm 2.069(28.35)$. Hence, the confidence interval for β_2 is:

$$31.7 \leq \beta_2 \leq 149.0$$

We thus conclude, with 95 percent confidence, that the mean amount of scrap for production line 1, at any given line speed, exceeds that for production line 2 by somewhere between 32 and 149.

Instrument Calibration Study Example

The engineer making the calibration study believed that the regression functions relating gauge reading (Y) to actual pressure (X_1) for both instruments are second-order polynomials:

$$E\{Y\} = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2$$

but that they might differ for the two instruments. Hence, the model employed (using a centered variable for X_1 to reduce multicollinearity problems—see Section 8.1) was:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i1}^2 + \beta_3 X_{i2} + \beta_4 x_{i1} X_{i2} + \beta_5 x_{i1}^2 X_{i2} + \varepsilon_i \qquad \textbf{(8.58)}$$

where:

$x_{i1} = X_{i1} - \bar{X}_1$ = centered actual pressure

$$X_{i2} = \begin{cases} 1 & \text{if instrument B} \\ 0 & \text{otherwise} \end{cases}$$

Note that for instrument A, where $X_2 = 0$, the response function is:

$$E\{Y\} = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 \qquad \text{Instrument A} \qquad \textbf{(8.59a)}$$

and for instrument B, where $X_2 = 1$, the response function is:

$$E\{Y\} = (\beta_0 + \beta_3) + (\beta_1 + \beta_4) x_1 + (\beta_2 + \beta_5) x_1^2 \qquad \text{Instrument B} \qquad \textbf{(8.59b)}$$

Hence, the test for equality of the two response functions involves the alternatives:

$$\begin{aligned} &H_0: \beta_3 = \beta_4 = \beta_5 = 0 \\ &H_a: \text{not all } \beta_k \text{ in } H_0 \text{ equal zero} \end{aligned} \qquad \textbf{(8.60)}$$

and the appropriate test statistic is (7.27):

$$F^* = \frac{SSR(X_2, x_1 X_2, x_1^2 X_2 | x_1, x_1^2)}{3} \div \frac{SSE(x_1, x_1^2, X_2, x_1 X_2, x_1^2 X_2)}{n - 6} \qquad \textbf{(8.60a)}$$

where n represents the combined sample size for both populations.