4 Multiple Regression

Before we extend the methods of the preceding sections to problems involving more than one independent variable, let us point out that the curves obtained (and the surfaces we will obtain) are not used only to make predictions. They are often used also for purposes of optimization—namely, to determine for what values of the independent variable (or variables) the dependent variable is a maximum or minimum. For instance, in the example of page 361 we might use the polynomial fitted to the data to conclude that the drying time is a minimum when the amount of varnish additive used is 5.1 grams (see Exercise 11.34).

Statistical methods of prediction and optimization are often referred to under the general heading of **response surface analysis**. Within the scope of this text, we shall be able to introduce two further methods of response surface analysis: **multiple regression** here and related problems of **factorial experimentation** in Chapter 13.

In multiple regression, we deal with data consisting of n $(r + 1)$-tuples $(x_{i1}, x_{i2}, \ldots , x_{ir}, y_i)$, where the x's are again assumed to be known without error while the y's are values of random variables. Data of this kind arise, for example, in studies designed to determine the effect of various climatic conditions on a metal's resistance to corrosion; the effect of kiln temperature, humidity, and iron content on the strength of a ceramic coating; or the effect of factory production, consumption level, and stocks in storage on the price of a product.

As in the case of one independent variable, we shall first treat the problem where the regression equation is linear, namely, where for any given set of values $x_1, x_2, \ldots,$ and x_r, for the r independent variables, the mean of the distribution of Y is given by

$$\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_r x_r$$

For two independent variables, this is the problem of fitting a plane to a set of n points with coordinates (x_{i1}, x_{i2}, y_i) as is illustrated in Figure 11.10. Applying the method of least squares to obtain estimates of the coefficients β_0, β_1, and β_2, we minimize the sum of the squares of the vertical distances from the observations y_i to the plane (see Figure 11.10); symbolically, we minimize

$$\sum_{i=1}^{n} [\, y_i - (\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2})\,]^2$$

and it will be left to the reader to verify in Exercise 11.35 that the resulting normal equations are

Normal equations for multiple regression with $r = 2$

$$\sum y = n b_0 + b_1 \sum x_1 + b_2 \sum x_2$$
$$\sum x_1 y = b_0 \sum x_1 + b_1 \sum x_1^2 + b_2 \sum x_1 x_2$$
$$\sum x_2 y = b_0 \sum x_2 + b_1 \sum x_1 x_2 + b_2 \sum x_2^2$$

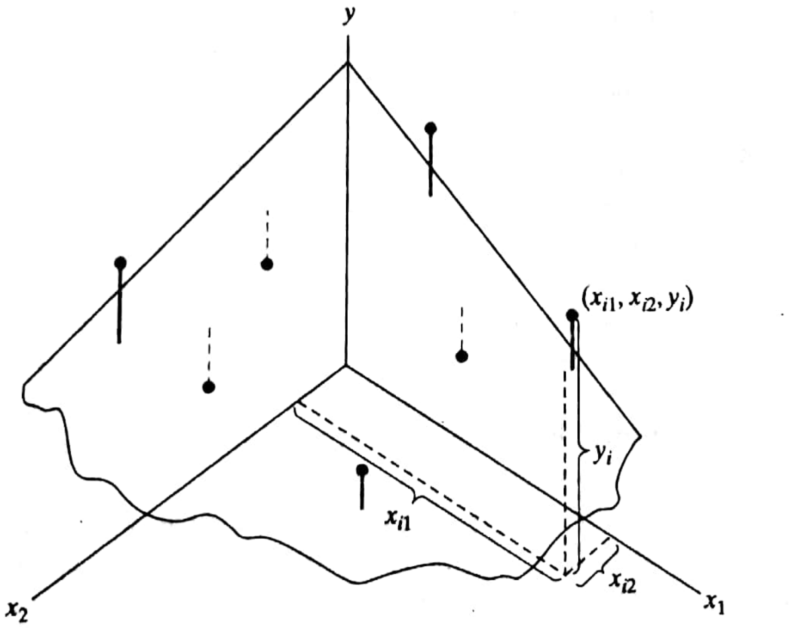

Figure 11.10
Regression plane

As before, we write the least squares estimates of β_0, β_1, and β_2 as b_0, b_1, and b_2. Note that in the abbreviated notation $\sum x_1$ stands for $\sum_{i=1}^{n} x_{i1}$, $\sum x_1 x_2$ stands for $\sum_{i=1}^{n} x_{i1} x_{i2}$, $\sum x_1 y$ stands for $\sum_{i=1}^{n} x_{i1} y_i$, and so forth.

Example

A multiple regression with two predictor variables

The following are data on the number of twists required to break a certain kind of forged alloy bar and the percentages of two alloying elements present in the metal:

Number of twists y	*Percentage of element A* x_1	*Percentage of element B* x_2
41	1	5
49	2	5
69	3	5
65	4	5
40	1	10
50	2	10
58	3	10
57	4	10
31	1	15
36	2	15
44	3	15
57	4	15
19	1	20
31	2	20
33	3	20
43	4	20

Fit a least squares regression plane and use its equation to estimate the number of twists required to break one of the bars when $x_1 = 2.5$ and $x_2 = 12$.

Solution Substituting $\sum x_1 = 40$, $\sum x_2 = 200$, $\sum x_1^2 = 120$, $\sum x_1 x_2 = 500$, $\sum x_2^2 = 3000$, $\sum y = 723$, $\sum x_1 y = 1963$, and $\sum x_2 y = 8210$ into the normal equations, we get

$$723 = 16 b_0 + 40 b_1 + 200 b_2$$

$$1963 = 40 b_0 + 120 b_1 + 500 b_2$$

$$3210 = 200 b_0 + 500 b_1 + 3000 b_2$$

The unique solution of this system of equations is $b_0 = 46.4$, $b_1 = 7.78$, $b_2 = -1.65$, and the equation of the estimated regression plane is

$$\hat{y} = 46.4 + 7.78 x_1 - 1.65 x_2.$$

Finally, substituting $x_1 = 2.5$ and $x_2 = 12$ into this equation, we get

$$\hat{y} = 46.4 + 7.78(2.5) - 1.65(12)$$
$$= 46.0$$