

The values of the explanatory variables,  $x_{ij}$ ,  $j = 1, \dots, p$ ,  $i = 1, \dots, n$  can sometimes be controlled at the time of conducting an experiment, as one would expect in Examples 1.1.1 and 1.1.3. On the other hand, these could also be observed quantities beyond the control of the observer, as in the case of  $x_1$  and  $x_2$  of Example 1.1.2 and  $x_2$  of Example 1.1.4. In the latter case, the  $x_{ij}$ s may be assumed to be random, and the model (1.3.2)–(1.3.3) may be interpreted as the conditional model of  $\mathbf{y}$  given  $\mathbf{X}$ . An explicit representation of the conditional model is given by

$$E(\mathbf{y}|\mathbf{X}) = \mathbf{X}\boldsymbol{\beta}, \quad D(\mathbf{y}|\mathbf{X}) = \sigma^2\mathbf{V}. \quad (1.3.4)$$

Thus, the error term in (1.3.2) is the difference  $\mathbf{y} - E(\mathbf{y}|\mathbf{X})$ . The mean and dispersion of the error given in (1.3.3) should be interpreted as conditional on  $\mathbf{X}$ . The representation (1.3.4) is called the *linear regression model*. In this context,  $\boldsymbol{\beta}$  is called the vector of regression parameters or regression coefficients (see Section 3.4 for a brief discussion on regression). An important aspect of the linear regression model is that the error  $\mathbf{y} - E(\mathbf{y}|\mathbf{X})$  must be uncorrelated with  $\mathbf{X}$  (see Exercise 1.6).

Suppose that the observations  $(y_i, x_{i1}, \dots, x_{ip})$  for  $i = 1, 2, \dots, n$  are statistically independent (in which case  $\mathbf{V} = \mathbf{I}$ ). Then the conditional model (1.3.4) can be written in the simpler form

$$\begin{aligned} E(y|x_1, x_2, \dots, x_p) &= \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p, \\ \text{Var}(y|x_1, \dots, x_p) &= \sigma^2, \end{aligned} \quad (1.3.5)$$

where  $\text{Var}(\cdot)$  indicates variance,  $y$  stands for *any* of the observed responses and  $x_1, \dots, x_p$  are the corresponding explanatory variables.

## 1.4 Scope of the linear model

Apart from the cases where the linear model can be used directly, as in Examples 1.1.1–1.1.4, there are other situations where it can be used with some adjustment.

**Example 1.4.1** (Polynomial regression) When there is a single explanatory variable, sometimes the mean response is sought to be explained as a polynomial function of the explanatory variable. In other

words, the model is

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_p x^p + \epsilon,$$

where  $y$  and  $x$  are the response and the explanatory variable, respectively. This is known as the *polynomial regression model*. This model can be viewed as a special case of (1.1.1) with  $x_j = x^j$ ,  $j = 1, 2, \dots, p$ . Therefore, the methodology to be developed for the model (1.1.1) will be applicable to the polynomial regression model.  $\square$

**Remark 1.4.2** This example as well as Example 1.1.1 illustrate an important point. Note that the right hand side of (1.1.1) is linear both in the parameters as well as in the explanatory variables. However, it is the linearity in the parameters which makes it a linear model. A model of the form (1.1.1) is called a linear model as long as the right hand side is linear in the parameters — even if it is not linear in the explanatory variables. Nonlinearity in the explanatory variables can be removed by transforming the variables. A transformation which makes the model linear in the explanatory variables is called *linearization*.  $\square$

**Example 1.4.3** (Accelerated failure time) Studies in reliability and survival analysis deal with time to failure of mechanical or biological entities. An important problem in this area is to analyse the effect of various explanatory variables (also called *covariates* in this context) on the time to failure. A popular model stipulates that the effect of every covariate is to accelerate (or decelerate) the time to failure ( $t$ ) by a factor. Specifically,  $\log t$  is modelled as

$$\log t = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \epsilon,$$

where  $x_1, \dots, x_p$  are the covariates. This model is clearly of the form (1.1.1). The problem of inference for the accelerated failure time model is often compounded by incompleteness of the data and dependence of the covariates on time.  $\square$

**Example 1.4.4** (Nonlinear regression) The Michaelis-Menten model for enzymic reactions states that the rate of reaction ( $v$ ) is given by the formula

$$v = \frac{k_0 \cdot s}{k_1 + s}, \quad (1.4.1)$$

where  $s$  is the substrate concentration and  $k_0$  and  $k_1$  are constants. The presence of measurement error and unaccounted factors mean that the above expression for  $v$  would hold only in the *approximate* sense. If the approximation error is explicitly represented by an additive error term, we have the model

$$v = \frac{k_0 \cdot s}{k_1 + s} + \epsilon, \quad (1.4.2)$$

which is not a special case of (1.1.1). However, if we let  $y = 1/v$  and  $x = 1/s$ , then the approximate version of (1.4.1) can be rewritten as

$$y = \beta_0 + \beta_1 x + \delta, \quad (1.4.3)$$

where  $\beta_0 = 1/k_0$ ,  $\beta_1 = k_1/k_0$ , and  $\delta$  is the approximation error. The latter model is of the form (1.1.1).  $\square$

Note that in the above example the models (1.4.2) and (1.4.3) are *not* equivalent. If (1.4.2) happens to be an appropriate model with  $E(\epsilon) = 0$ , the model error in (1.4.3) is unlikely to have zero mean (see Exercise 1.11). If one assumes the model (1.4.2), then fitting this model to data would require *nonlinear regression*, which involves iterative methods. The quality of the initial iterate is often crucial to the convergence of an iterative method. A solution obtained by fitting the approximate model (1.4.3) may serve to produce useful initial values for  $k_0$  and  $k_1$  in this case. This method of finding initial values may be considered whenever the nonlinear model under consideration can be approximately linearized by means of a transformation. General methodology for nonlinear regression will not be discussed in this book. The interested reader may use for instance, the books by Bates and Watts (1988) or Seber and Wild (1989).

**Example 1.4.5** (Generalized linear model) Sometimes a nonlinear function of the expected response (given the explanatory variables) is modelled as a linear function of the regression parameters. A typical model is

$$\eta(E(y|x_1, \dots, x_p)) = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p. \quad (1.4.4)$$

This model is known as the *generalized linear model*, and the function  $\eta$  is called the *link function*, which is assumed to be known. One can

define  $z = \eta(y)$  and use the model

$$z = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \epsilon,$$

where the error term represents not only the errors due to measurement or ignored factors, but also the error arising from the the above process of linearization. However, sometimes it may not be possible to linearize the model (see Exercise 1.13).  $\square$

The linearized model of this example is not directly used for fitting (1.4.4). As in the case of nonlinear regression, fitting of the generalized linear model often requires iterative methods. The linearized model may be fitted in order to produce reasonable initial iterates for such a procedure. See McCullagh and Nelder (1989) or Dobson (2001) for a detailed discussion of the generalized linear model and inference related to it.

## 1.5 Related models

The term *linear model* is sometimes used in a more general sense than what we consider here. From this broader perspective, any model which connects variables or their transformed versions through a linear relationship is a linear model. This description fits the generalized linear model, mentioned in the previous section. Additional examples are given below.

**Example 1.5.1** (Autoregressive model for time series) When observations are collected sequentially at regular time intervals, the pattern of temporal dependence can be represented by the model

$$x_t = \phi_1 x_{t-1} + \phi_2 x_{t-2} + \cdots + \phi_p x_{t-p} + \epsilon_t, \quad t \text{ is an integer,} \quad (1.5.1)$$

where  $x_t$  is the observation at time  $t$ ,  $\phi_1, \phi_2, \dots, \phi_p$  are unspecified parameters, and  $\epsilon_t$  is the unobservable model error at time  $t$ , which is assumed to be uncorrelated with the errors at other times. The errors are assumed to have an unspecified but constant variance. The form of the model (1.5.1) is very similar to (1.1.1). In fact, (1.5.1) can be