

Example 1.1.3 The height of an adult person varies from one homogeneous ethnic group to another. It also depends on the gender of the person. A comparison of two groups in terms of height can be made on the basis of the following model, which again is a special case of (1.1.1):

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon,$$

where y is the measured height of an adult, x_1 is a binary variable representing the ethnic group and x_2 is another binary variable representing the gender. The error term (ϵ) represents a combination of measurement error and the variation in heights that exists among the adults of a particular gender in a given ethnic group. \square

Example 1.1.4 The yield of tea in an acre of tea plantation depends on various types of agricultural practices (treatments). An experiment may be planned where various plots are subjected to one out of two possible treatments over a period of time. The yield of tea *before* the application of treatment is also recorded. A model for post-treatment yield (y) is

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon,$$

where the binary variable x_1 represents the treatment type and the real-valued variable x_2 is the pre-treatment yield. The error term mainly consists of unaccounted factors. Inclusion of x_2 is meant to reduce the effect of unaccounted factors such as soil type or the inherent differences in tea bushes. \square

1.2 Why a linear model?

The model (1.1.1) is just one of many possible models that can be used to explain the response in terms of the explanatory variables. Some of the reasons why we undertake a detailed study of the linear model are as follows.

- (a) Because of its simplicity, the linear model is better understood and easier to interpret than most of the other competing models, and the methods of analysis and inference are better developed.

Therefore, if there is no particular reason to presuppose another model, the linear model may be used at least as a first step.

- (b) The linear model formulation is useful even for certain nonlinear models which can be reduced to the form (1.1.1) by means of a transformation. Examples of such models are given in Section 1.4.
- (c) Results obtained for the linear model serve as a stepping stone for the analysis of a much wider class of related models such as mixed effects model, state-space and other time series models. These are outlined in Section 1.5.
- (d) Suppose that the response is modelled as a nonlinear function of the explanatory variables plus error. In many practical situations only a part of the domain of this function is of interest. For example, in a manufacturing process, one is interested in a narrow region centered around the operating point. If the above function is reasonably smooth in this region, a linear model serves as a good first order approximation to what is globally a nonlinear model.
- (e) Certain probability models for the response and the explanatory variables imply that the response should be centered around a linear function of the explanatory variables. If there is any reason to believe in a probability model of this kind, the linear model is the natural choice. An important example of such a probability model is the multivariate normal distribution. Sometimes the assumption of this distribution is justified by invoking the *central limit theorem*, particularly when the variables are themselves aggregates or averages of a large collection of other quantities.

1.3 Description of the linear model and notations

If one uses (1.1.1) for a set of n observations of the response and explanatory variables, the explicit form of the equations would be

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \epsilon_i, \quad i = 1, \dots, n, \quad (1.3.1)$$