

Chapter 1

Introduction

It is in human nature to try and understand the physical and natural phenomena that occur around us. When observations on a phenomenon can be quantified, such an attempt at understanding often takes the form of building a mathematical model, even if it is only a simplistic attempt to capture the essentials. Either because of our ignorance or in order to keep it simple, many relevant factors may be left out. Also models need to be validated through measurement, and such measurements often come with error. In order to account for the measurement or observational errors as well as the factors that may have been left out, one needs a *statistical* model which incorporates some amount of uncertainty.

1.1 The linear model

An important question that one often tries to answer through statistical models is the following: How can an observed quantity y be explained by a number of other quantities, x_1, x_2, \dots, x_p ? Perhaps the simplest model that is used to answer this question is the *linear model*

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p + \epsilon, \quad (1.1.1)$$

where $\beta_0, \beta_1, \dots, \beta_p$ are constants and ϵ is an error term that accounts for uncertainties. We shall refer to y as the *response variable*. It is also referred to as the *dependent variable*, *endogenous variable* or *criterion*

variable. We shall refer to x_1, x_2, \dots, x_p as *explanatory variables*. These are also called *independent variables* or *exogenous variables*. In the context of some special cases these are called *regressors*, *predictors* or *factors*. The coefficients $\beta_0, \beta_1, \dots, \beta_p$ are the *parameters* of the model. Note that the right hand side of (1.1.1), which is a linear function of the explanatory variables, can also be viewed as a linear function of the parameters.

Example 1.1.1 A well-known result of optics is *Snell's law* which relates the angle of incidence (θ_1) with the angle of refraction (θ_2), when light crosses the boundary between two media. According to this law,

$$\sin \theta_2 = \kappa \sin \theta_1,$$

where κ is the ratio of refractive indices of the two media. If the refractive index of one medium is known, the refractive index of the other can be estimated by observing θ_1 and θ_2 . However, any measurement will involve some amount of error. Thus, the following special case of the model (1.1.1) can be used:

$$y = \beta_1 x_1 + \epsilon,$$

where $y = \sin \theta_2$, $x_1 = \sin \theta_1$, θ_1 and θ_2 being the *measured* angles of incidence and refraction, respectively, and $\beta_1 = \kappa$. \square

Example 1.1.2 The hospital bill of a patient is likely to be bigger if the patient has to spend more days in the hospital. The bill also depends on several other factors including the nature of treatment, whether intensive care is needed and so on. Some factors may even be unknown (like the hospital's greed!). A simple model that can be used here is

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

where y is the amount of the hospital bill, x_1 is the duration of stay in the hospital (excluding stay at the intensive care unit) and x_2 is the duration of stay in the intensive care unit. The error term ϵ represents all the factors that are not specifically included, such as the nature of treatments and tests and variation from one hospital to another. The above model is again a special case of (1.1.1). \square

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.