1 Introduction

The scientific method is frequently used as a guided approach to learning. Linear statistical methods are widely used as part of this learning process. In the biological, physical, and social sciences, as well as in business and engineering, linear models are useful in both the planning stages of research and analysis of the resulting data. In Sections 1.1–1.3, we give a brief introduction to simple and multiple linear regression models, and analysis-of-variance (ANOVA) models.

1.1 SIMPLE LINEAR REGRESSION MODEL

In simple linear regression, we attempt to model the relationship between two variables, for example, income and number of years of education, height and weight of people, length and width of envelopes, temperature and output of an industrial process, altitude and boiling point of water, or dose of a drug and response. For a linear relationship, we can use a model of the form

$$y = \beta_0 + \beta_1 x + \varepsilon, \tag{1.1}$$

where y is the *dependent* or *response* variable and x is the *independent* or *predictor* variable. The random variable ε is the error term in the model. In this context, *error* does not mean mistake but is a statistical term representing random fluctuations, measurement errors, or the effect of factors outside of our control.

The linearity of the model in (1.1) is an assumption. We typically add other assumptions about the distribution of the error terms, independence of the observed values of y, and so on. Using observed values of x and y, we estimate β_0 and β_1 and make inferences such as confidence intervals and tests of hypotheses for β_0 and β_1. We may also use the estimated model to forecast or predict the value of y for a particular value of x, in which case a measure of predictive accuracy may also be of interest.

Estimation and inferential procedures for the simple linear regression model are developed and illustrated in Chapter 6.

Linear Models in Statistics, *Second Edition*, by Alvin C. Rencher and G. Bruce Schaalje

1.2 MULTIPLE LINEAR REGRESSION MODEL

The response y is often influenced by more than one predictor variable. For example, the yield of a crop may depend on the amount of nitrogen, potash, and phosphate fertilizers used. These variables are controlled by the experimenter, but the yield may also depend on uncontrollable variables such as those associated with weather.

A linear model relating the response y to several predictors has the form

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon. \tag{1.2}$$

The parameters $\beta_0, \beta_1, \ldots, \beta_k$ are called *regression coefficients*. As in (1.1), ε provides for random variation in y not explained by the x variables. This random variation may be due partly to other variables that affect y but are not known or not observed.

The model in (1.2) is linear in the β parameters; it is not necessarily linear in the x variables. Thus models such as

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \beta_3 x_2 + \beta_4 \sin x_2 + \varepsilon$$

are included in the designation *linear model.*

A model provides a theoretical framework for better understanding of a phenomenon of interest. Thus a model is a mathematical construct that we believe may represent the mechanism that generated the observations at hand. The postulated model may be an idealized oversimplification of the complex real-world situation, but in many such cases, empirical models provide useful approximations of the relationships among variables. These relationships may be either associative or causative.

Regression models such as (1.2) are used for various purposes, including the following:

1. *Prediction.* Estimates of the individual parameters $\beta_0, \beta_1, \ldots, \beta_k$ are of less importance for prediction than the overall influence of the x variables on y. However, good estimates are needed to achieve good prediction performance.
2. *Data Description or Explanation.* The scientist or engineer uses the estimated model to summarize or describe the observed data.
3. *Parameter Estimation.* The values of the estimated parameters may have theoretical implications for a postulated model.
4. *Variable Selection or Screening.* The emphasis is on determining the importance of each predictor variable in modeling the variation in y. The predictors that are associated with an important amount of variation in y are retained; those that contribute little are deleted.
5. *Control of Output.* A cause-and-effect relationship between y and the x variables is assumed. The estimated model might then be used to control the

output of a process by varying the inputs. By systematic experimentation, it may be possible to achieve the optimal output.

There is a fundamental difference between purposes 1 and 5. For prediction, we need only assume that the same correlations that prevailed when the data were collected also continue in place when the predictions are to be made. Showing that there is a significant relationship between y and the x variables in (1.2) does not necessarily prove that the relationship is causal. To establish causality in order to control output, the researcher must choose the values of the x variables in the model and use randomization to avoid the effects of other possible variables unaccounted for. In other words, to ascertain the effect of the x variables on y when the x variables are changed, it is necessary to change them.

Estimation and inferential procedures that contribute to the five purposes listed above are discussed in Chapters 7–11.

1.3 ANALYSIS-OF-VARIANCE MODELS

In analysis-of-variance (ANOVA) models, we are interested in comparing several populations or several conditions in a study. Analysis-of-variance models can be expressed as linear models with restrictions on the x values. Typically the x's are 0s or 1s. For example, suppose that a researcher wishes to compare the mean yield for four types of catalyst in an industrial process. If n observations are to be obtained for each catalyst, one model for the $4n$ observations can be expressed as

$$y_{ij} = \mu_i + \varepsilon_{ij}, \quad i = 1, 2, 3, 4, \quad j = 1, 2, \ldots, n, \tag{1.3}$$

where μ_i is the mean corresponding to the ith catalyst. A hypothesis of interest is $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$. The model in (1.3) can be expressed in the alternative form

$$y_{ij} = \mu + \alpha_i + \varepsilon_{ij}, \quad i = 1, 2, 3, 4, \quad j = 1, 2, \ldots, n. \tag{1.4}$$

In this form, α_i is the effect of the ith catalyst, and the hypothesis can be expressed as $H_0 : \alpha_1 = \alpha_2 = \alpha_3 = \alpha_4$.

Suppose that the researcher also wishes to compare the effects of three levels of temperature and that n observations are taken at each of the 12 catalyst–temperature combinations. Then the model can be expressed as

$$y_{ijk} = \mu_{ij} + \varepsilon_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk} \tag{1.5}$$

$$i = 1, 2, 3, 4; \quad j = 1, 2, 3; \quad k = 1, 2, \ldots, n,$$

where μ_{ij} is the mean for the ijth catalyst–temperature combination, α_i is the effect of the ith catalyst, β_j is the effect of the jth level of temperature, and γ_{ij} is the interaction or joint effect of the ith catalyst and jth level of temperature.

In the examples leading to models (1.3)–(1.5), the researcher chooses the type of catalyst or level of temperature and thus applies different *treatments* to the objects or experimental units under study. In other settings, we compare the means of variables measured on natural groupings of units, for example, males and females or various geographic areas.

Analysis-of-variance models can be treated as a special case of regression models, but it is more convenient to analyze them separately. This is done in Chapters 12–15. Related topics, such as analysis-of-covariance and mixed models, are covered in Chapters 16–17.