PART I

Regression

CHAPTER 1

Introduction to Linear Models

The objective of this book is to present a discussion and a formal definition of a general class of linear models. The presentation throughout this book will separate the discussion of regression models and analysis of variance models, though they are mathematically equivalent and subject to much of the same analysis. In Part I, the focus is on regression models. The general concepts apply as well to analysis of variance models that will be discussed in Part II. Matrix algebra is used to give a compact description of the models and will be used extensively in the chapters to follow. An elementary knowledge of matrix algebra is assumed but a review of the basic concepts and more advanced material to be used in this book is given in Appendix A. This chapter introduces much of the notation and terminology to be used throughout the book.

1.1 BACKGROUND INFORMATION

In the area of applied statistics, a substantial portion of the analyses comes under the heading of linear models. This general heading covers the major areas of regression analysis and the analysis of variance but includes other topics such as time series and multivariate analysis. The first two topics are the primary focus of this book.

The basis for the computational procedure used in the analysis can be traced back to the writings of the French mathematicians Gauss and LeGendre in the early years of the nineteenth century. In their writings, they describe the method of least squares for determining a line or plane to give an approximate description for a scatter of points. The method of least squares was given statistical credibility, under

Methods and Applications of Linear Models: Regression and the Analysis of Variance, Third Edition. Ronald R. Hocking.
 Published 2013 by John Wiley & Sons, Inc.

the assumption of normally distributed errors, by the likelihood-based methodology pioneered by R. A. Fisher in the first quarter of the twentieth century. Books such as *Applied Regression Analysis* by Draper and Smith (1966), *The Analysis of Variance* by Scheffé (1959), and *An Introduction to Linear Statistical Models* by Graybill (1961) summarize the methodology that was developed in the first half of this century. These methods contributed significantly to developments in all areas of research and are still widely used.

In the early stages of the development of these methods, the applications were limited by the ability to perform the required calculations, specifically the inversion of large matrices. The developments in high-speed computing in the two decades following World War II allowed us to consider regression models with many variables and the analysis of variance with many factors. The dramatic developments in computer technology in recent years have prompted new research in linear model methods. Numerical and graphical procedures have greatly expanded our ability to extract information from data. In regression analysis, techniques for identifying and examining the role of unusual observations and for detecting and understanding problems with collinear predictors enable the analyst to obtain a better understanding of the system or process being studied.

In the analysis of variance, much of the confusion caused by the mathematical statement of the model was removed by the revival of what is now called the **cell means model**. (We write key words and phrases in boldface.) This form of the model is especially useful in resolving questions with regard to unbalanced designs. Increased computer speed and capacity encouraged the application of likelihood-based methods for the estimation of variance components in mixed models. The search for diagnostic methods in the analysis of mixed models led to an alternative approach to describing the model and to a new computational procedure known as **AVE** that simplifies the computations and provides insight into the sources of variability.

A fundamental problem in all areas of science, and especially in statistics, is the gap between the development of new ideas and their implementation in practice. In the current computing environment, the problem becomes even more serious as new ideas are introduced into readily available statistical packages but not necessarily understood by practitioners. The motivation for the presentation in this book is to bridge the gap between theory and practice. This is done in Part I by providing an intuitive discussion of the theory and then giving a thorough discussion of the techniques necessary for applying the theory to the analysis of data. Parts I: Regression Analysis, can be understood by a student with a first course in statistical methods who is willing to accept the fundamental concepts and focus on the applications. The techniques are illustrated by numerical examples from different disciplines with particular emphasis on the ways in which new methods can provide insight into the analysis. Part II: The Analysis of Variance, contains a more thorough presentation of the theory and a discussion of many common models and their analysis.

No attempt is made to evaluate existing computer packages since they are constantly being modified and updated. Most of the graphics and computations in this book are performed using JMP, a product of the SAS Institute. It is hoped that the discussions will allow the reader to be a discriminating user of such packages.

1.2 MATHEMATICAL AND STATISTICAL MODELS

For our purposes, we describe a **mathematical model** as a functional relation between variables. In particular, we are interested in models that relate a set of input variables to a set of output variables. It is convenient to think of this situation in a generic form. Thus, we think of a response as the output of a process that depends on one or more inputs. This idea is shown in the schematic below. The box indicates a process in which the three inputs are transformed into the single output. For much of this book we will consider a single output but allow for several inputs.

Schematic for mathematical models.

Mathematically, we describe the relation as

$$y = g(x_1, x_2, x_3), \tag{1.1}$$

where y denotes the output, x_1, x_2, and x_3 denote the inputs, and $g(x_1, x_2, x_3)$ denotes the functional relation by which the inputs are converted into the output. We will refer to this as the **response function**.

The search for the functional relation in (1.1) is often one of the primary objectives of the analysis. With only one input variable this may be aided by a plot of the data. These plots are known as **scatter-plots**. The plot might suggest that the relation is nearly a straight line, but that there are departures from that linear relation. Polynomials or more complex functions of one variable may be suggested. With more than one input variable, three-dimensional plots may be informative and recent advances in plotting software that allow for rotations may be helpful.

The data are usually presented in column format as shown in Table 1.1. Here, y represents the response and x the input or predictor variable. The column labeled Obs. is not generally a part of the analysis but is used to refer to specific observations or cases.

Table 1.1 Format for the data

Obs.	y	x
1	y_1	x_1
2	y_2	x_2
3	y_3	x_3
⋮	⋮	⋮
n	y_n	x_n

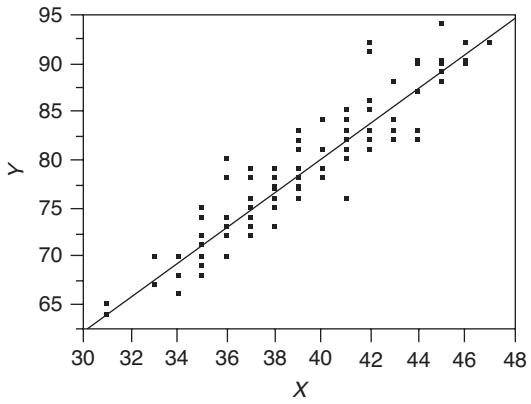

Figure 1.1 Scatter-plot of the data.

A typical scatter-plot of such data is shown in Figure 1.1. In Figure 1.1, a line has been superimposed on the data to give a general indication of the pattern. We note that this line provides a reasonable approximation to the data but that there is considerable scatter about this line. For example, several observations have $x = 35$ but have different responses. This suggests that for a given input, the response is subject to variability about some common value.

To include such variability in the model, we introduce the concept of a **statistical model**. To do this, we extend the mathematical model by adding a random variable to the input side of equation (1.1). Thus, we write the model as

$$y = g(x_1, x_2, x_3) + e. \tag{1.2}$$

Here, e denotes the added random variable, often called the **error term**. The properties of this random variable will depend on the situation, but it is often assumed to follow a univariate normal distribution with mean zero and variance σ^2. (A discussion of the normal distribution is given in Appendix B.I.1.) We shall elaborate on this definition in Section 1.3. Implicit in this assumption is the fact that the output, y, can be viewed as a random variable with mean (expected value), $g(x_1, x_2, x_3)$, and variance, σ^2. Thus, we may write the deterministic part of the model, using $E[y]$ to denote the expected value, as

$$E[y] = g(x_1, x_2, x_3). \tag{1.3}$$

For example, in a simple linear model, the response function is written as

$$E[y] = \beta_0 + \beta_1 x. \tag{1.4}$$

To illustrate the concept of an added error term that follows a normal distribution, refer to Figure 1.2. In this figure, the line denotes the expected value function,

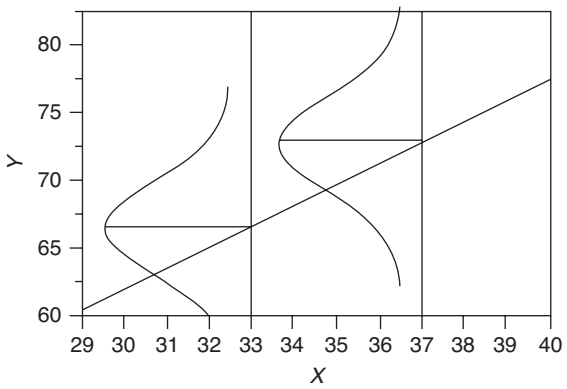

Figure 1.2 Indication of the normal error assumption.

$\beta_0 + \beta_1 x$. The normal density plots indicate that for a given input, x, the response is given by a random variable added to that mean function.

In most situations, the functional form of the mathematical model may be specified apart from the values of certain parameters. Thus, it may be known that the relation is a straight line, but the slope and intercept are not known. This is the case in equation (1.4). In general, letting $\boldsymbol{\beta}$ denote the parameters and $\boldsymbol{x}$ the inputs, we write the mean function in (1.3) as

$$E[y] = g(\boldsymbol{x}, \boldsymbol{\beta}). \tag{1.5}$$

If we add the assumption of independence to the random variables associated with the individual responses, the data can be viewed as a random sample from a population with mean given by (1.5) and variance σ^2.

In some cases, the population is only defined conceptually as the collection of possible observations that could be made. In other cases, it may be possible to enumerate the population, such as the population of students in the college of engineering at a given university. In that case we might sample the population rather than observe all students or, alternatively, we might view this group of students as a sample from the collection of all possible students at this university. The data could have been collected to develop a model for the relation between the score on a qualifying exam and the student's grade point average.

The concept of an input variable is quite general. For example, when modeling the daily amount of water used by an oil refinery, the inputs may include the size of the refinery, the amount of crude oil processed, the number of cooling towers, and the types of products. In general, inputs may be **quantitative**, that is, measurements such as temperature or amount, or they may be **qualitative**, indicating the presence or absence of a factor or the type of product. The observations may arise as a result of a carefully **designed experiment**. For example, if we wish to assess the effects of

temperature on the strength of a product, we could conduct a controlled experiment in which the production process is run at different temperatures while keeping all other factors at fixed values. Alternatively, the data may arise in an **observational study**. In such situations, measurements are taken on a random sample of individuals, or **experimental units**, from a given population. For example, we may select a random sample of female students in a specified age group and measure their percentage of body fat and certain physical characteristics. In this case, the objective is to develop a model to predict body fat from the more easily measured characteristics. We will treat each of these situations in the same way, recognizing that in the designed experiment, the departure from the assumed model may only be caused by natural variability in the material, whereas in the observational study, the departure may be cause by other factors that we have not included as inputs.

Our analysis of statistical models will include assessing the adequacy of the model, making inferences about the unknown parameters, and using the model to predict future observations.

1.3 DEFINITION OF THE LINEAR MODEL

In the discussion of statistical models in Section 1.2, we indicated a general functional form of the relation between the input and output variables. For the purpose of making inferences, it is useful to restrict the class of functions. In particular, we will be interested in functions that are linear in the parameters. Thus, for a model with p input variables, we write the mean function as

$$E[y] = \sum_{j=0}^{m} \beta_j x_j. \tag{1.6}$$

In this expression, the unknown parameters are given by $\beta_j, j = 0, \ldots, m$, y denotes the response, and $x_j, j = 0, \ldots, m$, denote the inputs. For example, in a production process it may be assumed that the response, $y =$ strength, is a linear function of the input, $x =$ temperature. Thus, we write

$$E[y] = \beta_0 + \beta_1 x. \tag{1.7}$$

Relating this expression to our general expression in (1.6), we see that the input variable x_0 is the constant 1. This is often the case as we allow for a nonzero intercept in our response function. We use the parameter β_0 as the coefficient for this constant term. We may encounter cases where we will want to force $\beta_0 = 0$.

While it is true that the response function in (1.7) is linear in temperature, that is not the linearity of concern. Rather, it is the fact that the function is linear in β_0 and

β_1. To emphasize the linearity assumption, note that the expected response could be a quadratic function of temperature, written as

$$E[y] = \beta_0 + \beta_1 x + \beta_2 x^2. \tag{1.8}$$

In this case, the response function is linear in β_0, β_1, and β_2 as required by the definition. Such a function is often considered if the company is trying to determine an optimum operating temperature for the process.

Continuing with this example, it may be suspected that the strength of the product also depends on the time of exposure, s. In this case we may consider the following form of the response where the expected value is linear in both temperature and time. Thus, we write

$$E[y] = \beta_0 + \beta_1 x + \beta_2 s. \tag{1.9}$$

Further, we might consider a model that is quadratic in both input variables. This response function might be written as

$$E[y] = \beta_0 + \beta_1 x + \beta_{11} x^2 + \beta_2 s + \beta_{22} s^2 + \beta_{12} xs. \tag{1.10}$$

The term xs is just the product of the time and the speed variables for a given case. This term allows for an interaction between the two variables and will be examined in detail in later chapters.

For our analyses, we will assume that we have n observations on a process with mean function of the form (1.6). The data are described by the $(m + 2)$-tuple, $(y_i, x_{i0}, x_{i1}, x_{i2}, \ldots, x_{im})$, for $i = 1, \ldots, n$, where y_i denotes the response to the inputs, x_{ij}. For the purpose of describing the models in this chapter and for performing computations in subsequent chapters, it is convenient to use the notation and methods of matrix algebra. (A summary of the basic results in matrix algebra is found in Appendix A.) In this book all vectors will be column vectors and are denoted by lower case, boldface letters. Row vectors will be indicated by the transpose symbol, superscript T. Thus, the n responses may be described by a row vector, $\boldsymbol{y}^T = (y_1, \ldots, y_n)$ with elements y_i or as a column vector

$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}. \tag{1.11}$$

For the development of our theoretical results, $\boldsymbol{y}$ is viewed as a **random vector**, that is, a vector of random variables. Later, when we illustrate the theory with numerical examples, $\boldsymbol{y}$ will denote the observed data, that is, a realization of these random variables. The proper interpretation will be clear from the context.

The inputs will be denoted by the vectors $\boldsymbol{x}_j$ for $j = 0, 1, \ldots, m$, where

$$\boldsymbol{x}_j = \begin{bmatrix} x_{1j} \\ \vdots \\ x_{nj} \end{bmatrix}. \tag{1.12}$$

The inputs will be displayed in the **design matrix** $\boldsymbol{X}$, that is, an $n \times (\mathrm{m} + 1)$ matrix

$$\begin{aligned} \boldsymbol{X} &= (x_{ij},\ i = 1, \ldots,\ n,\ \textit{and}\ j = 0, \ldots, m) \\ &= (\boldsymbol{x}_0, \boldsymbol{x}_1 \ldots, \boldsymbol{x}_m). \end{aligned} \tag{1.13}$$

Our convention is that matrices will be written in upper case, boldface letters, the one exception being that $\boldsymbol{J}$ will denote the column vector of ones. Thus the design matrices for equations (1.7), (1.8). (1.9), and (1.10), respectively, may be written as $\boldsymbol{X} = (\boldsymbol{J}\ \boldsymbol{x})$, $\boldsymbol{X} = (\boldsymbol{J}\ \boldsymbol{x}\boldsymbol{x}^2)$, $\boldsymbol{X} = (\boldsymbol{J}\ \boldsymbol{x}\ \boldsymbol{s})$, and $\boldsymbol{X} = (\boldsymbol{J}\ \boldsymbol{x}\boldsymbol{x}^2\boldsymbol{s}\boldsymbol{s}^2\boldsymbol{x}\boldsymbol{s})$. Here, $\boldsymbol{x}$ and $\boldsymbol{s}$ are the vectors of temperature and associated speeds and $\boldsymbol{x}^2$, $\mathbf{s}^2$ and $\boldsymbol{x}\mathbf{s}$ are the vectors of squared times, squared speeds, and the products of the times and speeds.

Using this matrix notation, the model equations, describing the model and the data, are written as

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}, \tag{1.14}$$

where $\boldsymbol{y}$ is the n-vector of responses, $\boldsymbol{X}$ the $n \times (m + 1)$-dimensional design matrix, $\boldsymbol{\beta}$ the $(m + 1)$-vector of parameters, and e the n-vector of errors.

To describe the assumptions about the expected value of the responses, we introduce the concept of the expected value of a matrix of random variables. Thus, if $\boldsymbol{M}$ is a matrix, whose elements, m_{ij}, are random variables, we define the expected value of $\boldsymbol{M}$ as follows.

Definition 1.1 The expected value of a random matrix $\boldsymbol{M}$, with elements m_{ij}, is defined to be the matrix of expected values of its elements. That is,

$$E[\boldsymbol{M}] = (E[m_{ij}]). \qquad \blacksquare$$

Using matrix notation we then write our assumptions about the mean response as

$$E[\boldsymbol{y}] = (E[y_i]) = \left(\sum_{j=0}^{m} \beta_j \boldsymbol{x}_{ij} \right) = \boldsymbol{X}\boldsymbol{\beta}. \tag{1.15}$$

The term design matrix reflects the fact that, in some applications, the matrix reflects the design of the experiment that led to the data. We use this terminology for convenience in our general description, recognizing that $\boldsymbol{X}$ simply denotes the matrix of input variables in our expression. We assume that $\boldsymbol{X}$ has **full column rank**,

$r(\boldsymbol{X}) = m + 1$, which means that the columns of $\boldsymbol{X}$ are linearly independent and implies that $n \geq m + 1$.

In many applications, the responses are assumed to be independent, or at least uncorrelated, with common variance, σ^2. To write this assumption in matrix form, we define the **covariance matrix**, $\boldsymbol{V} = Var[\boldsymbol{y}]$, as the symmetric matrix whose diagonal elements v_{ii} denote the variance of v_1 and whose off-diagonal elements v_{ik} are the covariances between y_i and y_k. We may then write the assumptions of independence and constant variance as

$$Var[\boldsymbol{y}] = \sigma^2 \boldsymbol{I}, \tag{1.16}$$

where $\boldsymbol{I}$ is the identity matrix (ones on the diagonal and zeros elsewhere). To relate this to our definition of the expected value of a matrix, suppose that the matrix $\boldsymbol{M}$ has elements $m_{ik} = (y_i - E[y_i])(y_k - E[y_k])$. Note that the expected value of m_{ik} is the covariance between y_i and y_k if $i \neq k$, and the variance of y_i if $i = k$. Letting $\boldsymbol{V} = (v_{ij})$, we write $\boldsymbol{V} = E[\boldsymbol{M}]$. The matrix $\boldsymbol{M}$ is conveniently written by considering the product of the vector $\boldsymbol{y} - E[\boldsymbol{y}]$ and its transpose. Thus,

$$M = (\boldsymbol{y} - E[\boldsymbol{y}])(\boldsymbol{y} - E[\boldsymbol{y}])^T. \tag{1.17}$$

These observations are summarized in the following definition.

Definition 1.2 The covariance matrix $\boldsymbol{V} = Var[\boldsymbol{y}]$ of the random vector $\boldsymbol{y}$ is given by

$$\boldsymbol{V} = Var(\boldsymbol{y}) = (v_{ik}) = E((\boldsymbol{y} - E[\boldsymbol{y}])(\boldsymbol{y} - E[\boldsymbol{y}])^T). \qquad \blacksquare$$

In general, the variance may not be constant and the covariances may not be zero. In the discussion of mixed models in Part 2 on the analysis of variance, we will be encounter situations where there is a linear structure on the covariance matrix. We will assume that the variances, $Var[y_i]$, and covariances, $Cov[y_i, y_k]$, are known linear functions of unknown parameters. Thus, we will write

$$\begin{aligned} Var(y_i) &= v_{ii} = \sum_{t=1}^{c} \phi_t v_{tii}, \\ Cov(y_i, y_k) &= v_{ik} = \sum_{t=1}^{c} \psi_t v_{tik}. \end{aligned} \tag{1.18}$$

(*Note:* When more than one expression appears in a display as in (1.18), the equation number will refer to the collection of expressions, not just the last one.) Here the v_{tik} are known constants that are determined by the process. The ϕ_t and ψ_t are unknown parameters that are called **variance components**.

The linear structure on the second moments described in (1.18) may be written in matrix notation. To do so, let $\boldsymbol{V}_t$, $t = 1, \ldots, c$, denote known, symmetric matrices

with elements v_{tik}. It follows that the matrix of variances and covariances can be written as

$$\boldsymbol{V} = \sum_{t=1}^{c} \phi_t \boldsymbol{V}_t. \tag{1.19}$$

For example, we might assume that all variances are equal to ϕ_1 and that all covariances are equal to ϕ_2. Letting, $\boldsymbol{U}$ denote the matrix all of whose elements are equal to 1, we may write this covariance structure as

$$\boldsymbol{V} = \phi_1 \boldsymbol{I} + \phi_2(\boldsymbol{U} - \boldsymbol{I}). \tag{1.20}$$

Referring to the parameters in this matrix as variance components simply reflects the fact that they are components of the covariance matrix. Since $\boldsymbol{V}$ is a covariance matrix, it must be positive definite, and this places a natural set of constraints on the parameters. These constraints will be indicated in specific examples later in the book.

Note that the linearity assumptions (1.15) and (1.19) are not restrictive in the sense that any mean vector and covariance matrix may be written in these forms. To see this, let $\boldsymbol{x}_j = \boldsymbol{u}_j$ denote the unit vector with a 1 in position j and zeros elsewhere, then the design matrix in (1.15) is the identity. In this case $n = m + 1$, and the parameters are just the expected values of the observations, that is, $E[\boldsymbol{y}] = \boldsymbol{\beta}$. Similarly, the matrices $\boldsymbol{V}_t$ in (1.19) could just be the indicator matrices having either $v_{tii} = 1$ and zeros elsewhere, or $v_{tik} = v_{tki} = 1$ and zeros elsewhere. In this case the parameters are just the variances and covariances of the observations. Our interest will be in cases where the number of parameters in $\boldsymbol{\beta}$ is small relative to n and the model is described in terms of a small number of variance components.

To summarize our assumptions and to distinguish the special case defined by (1.16), we make the following definition.

Definition 1.3 The random vector $\boldsymbol{y}$ is said to have a general linear model if $E[\boldsymbol{y}]$ and $Var[\boldsymbol{y}]$ are given by (1.15) and (1.19), respectively. It is called a simple linear model if $Var[\boldsymbol{y}] = \sigma^2 \boldsymbol{I}$. ■

We will see that it is possible to estimate the parameters of our linear model without making further assumptions about the probability distribution. However, in order to make inferences, in the form of tests of hypotheses, confidence intervals, or prediction intervals, it is necessary to specify the distribution of the data. A common assumption is that the elements of $\boldsymbol{y}$ follow a normal distribution. There are several motivations for this assumption, not the least of which is that it leads to very elegant theoretical results. The assumption is further justified by the fact that many situations are approximately modeled by the normal distribution, or that some function of the observations is approximately normal. In the most general situation, $\boldsymbol{y}$ is assumed to follow the multivariate normal density function. Familiarity with the normal distribution is assumed. A detailed discussion of the normal distribution is given in Appendix B.I.1.

Assuming that $\boldsymbol{y}$ is normally distributed, the linear model described in Definition 1.3 may be written compactly as $\boldsymbol{y} \sim N(\boldsymbol{X\beta}, \boldsymbol{V})$. Alternatively, the distribution of the error vector, $\boldsymbol{e} = \boldsymbol{y} - \boldsymbol{X\beta}$, is $\boldsymbol{e} \sim N(\boldsymbol{0}, \boldsymbol{V})$.

The equation form of the model is generally favored in the literature. It is often more informative, particularly in regression models, to think of the locus of means, defined by the expected value, as a function that is describing the data apart from a random error. We will find that both forms of the model are useful. These ideas are summarized in the following definition.

Definition 1.4 The vector $\boldsymbol{y}$ is said to follow a normal linear model if $\boldsymbol{y} \sim N(\boldsymbol{X\beta}, \boldsymbol{V})$ or, equivalently, $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{e}$ and $\boldsymbol{e} \sim N(\boldsymbol{0}, \boldsymbol{V})$. It follows a simple, normal linear model if $\boldsymbol{e} \sim \boldsymbol{N}(\boldsymbol{0}, \boldsymbol{\sigma}^2\boldsymbol{I})$. ■

1.4 EXAMPLES OF REGRESSION MODELS

In the regression model, the expected value of the response is a function of one or more input variables known as **regressors**, **predictors**, or **independent variables**. Typically, these are quantitative variables such as temperature, speed, and size, but qualitative variables such as gender, department, and breed can be included. The output is known as the **response** or **dependent variable**. Several examples are given to illustrate the concepts. We consider first the case of one quantitative regressor to fix the ideas and then extend to the general case.

1.4.1 Single-Variable, Regression Model

Example 1.1 Golf Tournament Results The data shown in Figure 1.1 are the results of an amateur golf tournament. The input variable is the competitors' score on the first 9 holes and the response is the score for the 18 hole tournament. The objective is to predict the 18 hole score based on the results from the first 9 holes. This is an example of an observational study in which no other factors such as age or experience were controlled. This example is examined in Exercise 1.7.

Example 1.2 Particle Board Study To introduce the concept of a designed experiment we consider the results of a study by a company that makes particle boards that are used in construction. Particle boards are made by mixing small wood particles with an adhesive, forming them into sheets and then baking them in an oven. The company is interested in studying the strength of the boards as a function of the baking temperature with the possible objective of determining an optimum temperature. An experiment was conducted that consisted of running the process at six different temperatures, with 10 boards produced at each temperature. The temperatures and associated breaking strengths were recorded. A plot of the data is shown in Figure 1.3 where the inputs, x_i, denote the baking temperature and the response, y_i is a measure of the strength. Here, $i = 1, \ldots, 60$. ■

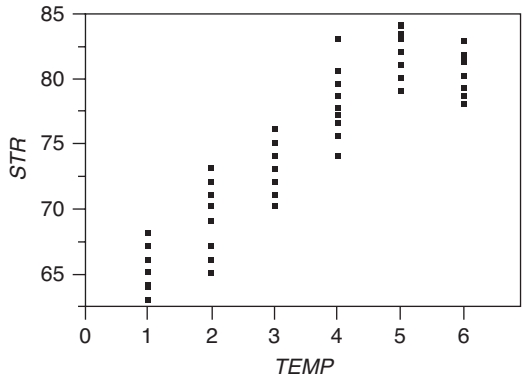

Figure 1.3 Plot of particle board data.

Discussion of Example 1.2 Since there are six different temperatures with 10 replicates at each temperature, we might consider using two subscripts to denote the data. For example, x_{ij}, $i = 1, \ldots, 6$ and $i = 1, \ldots, 10$. This is not necessary in this example but later, especially in analysis of variance models, we will use multiple subscripts.

Referring to Figure 1.3, we note the variability of the responses for a given temperature. A useful tool for examining the distribution of the data for a given temperature is the box-plot (see Milton and Arnold (1990) for a discussion of box-plots). These plots give an indication of the location and spread of the data at each temperature. The line within the box indicates the average strength of the boards produced at that temperature and the limits of the box indicate the 25th and 75th quantiles. In Figure 1.4, we show these plots for each temperature. Based on casual

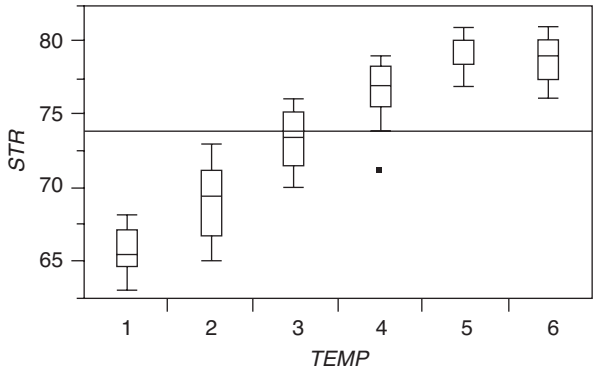

Figure 1.4 Box-plots of particle board data.

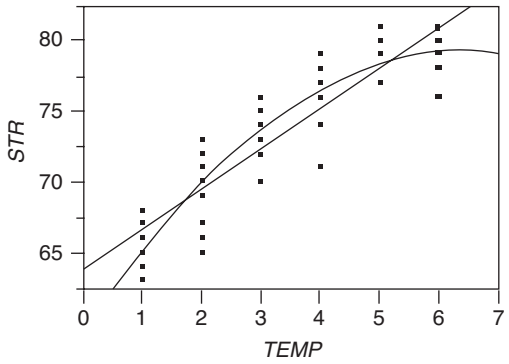

Figure 1.5 Particle board data with lines imposed.

inspection of this plot It seems reasonable to accept the assumption of a common variability for each value of *TEMP*.

It is of interest to examine the linear model assumptions in terms of Example 1.2. The box-plots suggest that the spread of data for a given temperature is about the same for each temperature and we will assume that the data were collected in such a way as to justify the assumption of uncorrelated responses. Thus, the simple linear model with $\boldsymbol{V} = \sigma^2\boldsymbol{I}$ is reasonable. Examination of histograms for the data might justify the assumption of normality.

Figure 1.5 shows the particle board data with two approximating lines imposed. Examination of this figure shows that the strengths are increasing with temperature and that the increase is approximately linear, although the rate on increase may be declining with increasing temperature. This suggests that the response function may be linear in temperature and we might initially consider the model

$$y_i = \beta_0 + \beta_1 x_i + e_i \tag{1.21}$$

for $i = 1, \ldots, 60$. As noted earlier, we use β_0 to denote the intercept.

There are several advantages to this model. For example, (1) it is intuitively simple to think of mean strength increasing in this way, (2) we can interpolate or extrapolate to infer mean strengths at other than the experimental temperatures, (3) we will see that we can make inferences about the slope and intercept. For example, we will consider the hypothesis H_0: $\beta_1 = 0$. This is equivalent to the hypothesis that mean strength does not depend on temperature. The design matrix for this **reduced model** consists of the column of ones and the response vector is the column of strengths.

This model is known as the **simple**, **linear regression model**. It is a simple linear model, by Definition 1.3, since the mean structure is linear in the parameters and $\boldsymbol{V} = \sigma^2\boldsymbol{I}$. In addition, the implication of this terminology, in the class of linear regression models, is that the model depends on only one regressor and that the locus of means is linear in that variable. We might further argue that the departure from

linearity is caused by random sources such as natural variability in the material and that this does not depend on the temperature. We summarize our discussion in the following definition.

Definition 1.5 The response, y, has a simple linear regression on the input, x, if the observed pairs (y_i, x_i) are related by the equation

$$y_i = \beta_0 + \beta_1 x_i + e_i,$$

where the random errors are independent, $N(0, \sigma^2)$. ■

In the particle board study, it was reasonable to assume that the mean strength is some function of temperature, but typically, we have no idea as to the nature of that function. In fact, a substantial part of the effort in a regression analysis will be directed toward the identification of this function. If the model that is linear in temperature is not acceptable, we might consider the quadratic model in (1.8). A second degree line is also shown in Figure 1.5. We see that this line might be more appropriate for this data. In this example, the experimenter may have been interested in estimating the temperature that yields the maximum strength. The quadratic model allows us to do that, but we must be aware of the danger of extrapolating beyond the range of the experimental data. Other models that are not linear in x include the following:

$$\begin{aligned} &\text{(a)} \quad E[y] = \beta_0 + \beta_1 \left(\frac{1}{x}\right), \\ &\text{(b)} \quad E[y] = \beta_0 \exp(\beta_1 x), \\ &\text{(c)} \quad E[y] = \beta_0 x^{\beta_1}, \end{aligned} \tag{1.22}$$

Model (a) is linear in the parameters and hence is a linear model in the sense of Definition 1.3 even though the model is not a linear function of x. We will see that this nonlinearity in the input variable does not complicate the analysis but it is reflected in the interpretation of the model. Models (b) and (c) are not linear in the parameters and hence fall outside of the scope of our linear model analysis. Such models will be discussed separately in Chapter 8. In both of these models, the logarithm of the response function can be expressed as a linear function of the parameters. This suggests that the model could have been developed using the logarithm of the response. Using natural logarithms, we might have considered one of the following models:

$$\begin{aligned} &\text{(d)} \quad E[\ln(y)] = \alpha_0 + \beta_1 x, \\ &\text{(e)} \quad E[\ln(y)] = \alpha_0 + \beta_1 \ln(x_1), \end{aligned} \tag{1.23}$$

where $a_o = \ln(\beta_0)$. The point here is that the model might be linear if we measure our response in a different scale. Regardless of the form of the model that is selected,

it is important to justify our assumptions about the distribution of the errors. There are other models for which the mean function can be linearized by some appropriate transformation, but in most cases it is not possible. The problem of transforming the data will be discussed in more detail in Chapter 3.

The quadratic model (1.8) might be more satisfactory than the simple linear model and it is tempting to consider higher ordered **polynomial models**, for example, the cubic model

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + e_i. \tag{1.24}$$

In this polynomial model there is a single input x but three regressors, x, x^2, and x^3. The terms regressor and predictor will be used to describe these quantities as well as other inputs such as speed, pressure, and time, that might be a part of the model. It should be clear from the context whether the regressor is a distinct variable or a function of another variable. The extension to higher ordered polynomial models should be clear. Such models are considered in Chapter 7.

1.4.2 Regression Models with Several Inputs

In the particle board study, it might be suspected that strength could depend on other factors such as the speed of the conveyor belt as it carries the material through the baking oven. To introduce the conveyor speed into the analysis, we might conjecture that the mean response is a linear function of temperature and speed. For notational simplicity, we again avoid the use of multiple subscripts and view the data as a set of n-tuples, (y_i, t_i, s_i), where t_i and s_i indicate the temperature and speed that led to response y_i. Thus a candidate model, in equation form, is

$$y_i = \beta_0 + \beta_1 t_i + \beta_2 s_i + e_i, \tag{1.25}$$

(note: to conduct an experiment using all combinations of six temperatures and five speeds with 10 observations made at each temperature–speed combination would require $n = 300$ runs, a dramatic increase over the earlier experiment).

A model often used in this setting is the extension of the single-input, quadratic model to include all second-order terms in the two inputs as shown in (1.10). Note that the two inputs lead to five regressors. This quadratic model may provide a reasonable, local approximation over a limited range on temperature and speed. It has the advantage of being simple to analyze, and the parameters have meaningful interpretations. It is one of the fundamental models in a methodology called **response surface analysis**, that we discuss in Chapter 7. The form of the design matrix for this model should be clear. With this motivation we now define the multiple, linear regression model as follows.

Definition 1.6 Given n observations, on a response y, with m regressors $x_1, x_2, \ldots, x_m$, the model that expresses the mean of y as a linear function of these regressors is called the multiple linear regression model and is written as

$$y_i = \beta_0 + \sum_{j=1}^{m} \beta_j x_{ij} + e_i$$

or, in matrix form, as

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e},$$

where the ith row of the design matrix is $(1, x_{i1}, x_{i2}, \ldots, x,)$ and the parameter vector is $\boldsymbol{\beta}^{\mathrm{T}} = (\boldsymbol{\beta}_0, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_m)$. The columns of $\boldsymbol{X}$ will be denoted by $\boldsymbol{J}$ and $\boldsymbol{x}_j, j = 1, \ldots, m$, and it is generally assumed that $\boldsymbol{e}$ is normally distributed with a scalar covariance matrix. Note that we have included the intercept term in our definition. This term may not be included in some studies. ■

In this definition the regressors have been simply denoted as x_j, but the implication is they may represent functions of a set of inputs. For example, in (1.10), we have $x_1 = t$, $x_2 = s$, $x_3 = t^2$, $x_4 = s^2$, and $x_5 = ts$. Thus, the regressors may be the inputs that we have measured, such as temperature and speed, or they may be functions of them such as powers, reciprocals, exponentials, logarithms, products, and ratios. It is also possible that the response may be some function of the output that we actually measured.

In our discussion of the particle board and the golf tournament data, we distinguished between the case where the data arose from a carefully designed experiment as opposed to an observational study. In many applications of regression models, the data arise in this latter way. That is, the data consist of a vector of observations on each of n experimental units but the values of the inputs are not specified in advance. The following example illustrates another such a situation and we will use it to raise several potential problems with the analysis.

Example 1.3 A company makes a product from sheets of stainless steel purchased from a supplier. It is observed that the number of units obtained from a sheet varies considerably. This variation is felt to depend on three factors, the width, the density, and the tensile strength of the sheet. To investigate this relation, n sheets were selected at random and the values of the response, (*PROD*) and the three inputs, (*WID*), (*DENS*), and (*STR*) were observed. ■

Discussion of Example 1.3 As a first attempt at developing a model, we might consider the model

$$PROD_i = \beta_0 + \beta_1(WID)_i + \beta_2(DENS)_i + \beta_3(STR)_i + e_i. \qquad (1.26)$$

If we make the usual assumptions about the random errors, this model is included under Definition 1.6, and our analysis would be the same as if the data arose from a designed experiment. However, there are potential hidden hazards. Since we have no control over the ranges or the relative values of the regressor variables, it is possible that we could obtain a few samples in which the values of one or more of the inputs differ markedly from those in the rest of the observations. Or, by chance, the values of the inputs in the sample may have been restricted to only a subset of the possible values.

The situations mentioned in this example are not uncommon in observational studies, and they offer a potential for misleading analyses. We will devote some time in Chapters 5 and 6 to describing techniques for detecting such problems and modifying the analyses.

In some applications, the regressors may be discrete variables, such as zero or one, used to indicate the presence or absence of a qualitative characteristic. The following extension of Example 1.2 will illustrate this concept.

Example 1.4 Suppose that the experiment on the manufacture of particle boards included two different types of glue. An experiment is run using six different temperatures for each type of glue holding all other factors fixed. Ten different boards were made at each temperature–glue combination for a total of 120 boards. Of interest is the effect of the different glues as well as the temperature. ■

Discussion of Example 1.4 We could consider applying the simple, linear regression model for each type of glue to assess the effect of temperature for each glue, but that does not allow for an obvious way to test for glue differences. A single model that allows for this comparison uses an **indicator variable** to distinguish between glue types. To describe this model, let $\boldsymbol{x}$ be the vector of the temperatures and let $\boldsymbol{z}$ be a vector whose elements are 1 if the response is for glue 1 and zero if for glue 2. The model is then written as

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + e_i. \tag{1.27}$$

Note that this model describes two parallel lines, each having slope β_1 but different intercepts. For glue 1, the intercept is $\beta_0 + \beta_2$ and for glue 2 it is β_0. Thus, the model assumes that the effect of temperature is the same for either type of glue. We will study this model allowing for different slopes in Chapter 7.

Our definition of the multiple linear regression model makes the assumption of a scalar covariance matrix, but this need not be the case. For example, the particle board study could have included data from two different factories, and we might allow for different variability at each factory. Thus, regression models extend naturally to include general covariance structures. The following example indicates another source of nonscalar covariance matrices.

Example 1.5 In a study to determine the factors that influence the amount of water used by an oil refinery, data are collected from a large number of refineries. The response is the amount of water used in a month, and data are taken over a period of 2 years. The predictors include a measure of size or capacity of the refinery, indicator variables for the types of processes, and quantitative measures of the levels of production of various products. The assumption of a scalar covariance matrix in this case should certainly be questioned. It seems quite likely that the variability might be a function of the size of the refinery and likely would be an increasing function of size. In addition, the responses from a given refinery may be correlated and that correlation could be a function of the size or complexity of the refinery. In this case, the analysis would involve the estimation of the regression coefficients and the variance components. The use of the model for predictive purposes must also recognize this covariance structure. ■

1.4.3 Discrete Response Variables

In some problems, the response may be discrete, invalidating the assumption of a normal distribution. Consider, for example, the situation where the response takes on one of two values, say, zero or 1. For example, the response may indicate the success or failure of a particular medication. This type of response requires the development of new models and a more complex analysis. There are, however, linear model aspects of the problem, as will be discussed in Chapter 8. Specifically, we will look at **logistic regression models** that are appropriate for this dichotomous response situation and we will briefly consider a class of **generalized linear models**.

1.4.4 Multivariate Linear Models

Our primary focus in this book is on the univariate linear model, that is, the response y is assumed to be a scalar, or equivalently $\boldsymbol{y}$ is a column vector. The basic concepts to be introduced extend to the case where the response is a matrix, say $\boldsymbol{Y}$, of size $N \times t$, in which each row is a t-vector of responses on an experimental unit. This situation is illustrated by the following example.

Example 1.6 A study is to be conducted to compare medications for controlling high blood pressure. For the study a random sample of n patients who have high blood pressure are assigned to each of the p treatments. For each of the patients, the response is measured at the end of each of t time periods. The data matrix $\boldsymbol{Y}$ consists of $N = pn$ rows of length t, where the data in a row are the observations on a given patient. It seems natural to assume that the responses on a given object are correlated over time and that the variance may be a function of time. In general, let $\boldsymbol{V}_0$ denote the covariance matrix associated with the observations on a given subject, and assume that this same matrix is applicable for all patients under each treatment. The covariance structure is defined by assuming that the rows of $\boldsymbol{Y}$ are independent with covariance matrix $\boldsymbol{V}_0$. This is analogous to the assumption of constant variance

in the univariate case. The data could be arrayed as a vector of length *tpn* and we would have a univariate model, but for reasons of convenience and interpretation, it is usually better to retain the multivariate formulation. We will return to a discussion of this concept in Chapter 8. ■

1.5 CONCLUDING COMMENTS

In this chapter we have introduced the concept of linear models and described a variety of situations where regression models are appropriate. Chapters 2–8 contain detailed discussions of the analysis of the models introduced here and more complex models and illustrate the concepts with numerical examples. A unique feature of the discussion of linear models in this book is that the design matrix has full column rank. In the regression models this is a standard assumption and simply implies that we have not included predictors that are linear combinations of other predictors. (We will treat, in detail, the often confusing situation where the regression model contains near linear dependencies in $\boldsymbol{X}$.). In the analysis of variance models, the full rank assumption is not common. The assumption of a full rank design matrix in our presentation of analysis of variance models differs from that made for the classical, over-parameterized model as used in most standard texts. The relation between these two approaches will be established. We will see that the **cell means** formulation removes much of the confusion generally associated with analysis of variance models. However, the classical presentation has advantages, and we will learn to move easily from one form to the other.

EXERCISES

Section 1.3

1.1 Write the design matrices for the mean structures described by equations (1.8), (19), and (1.10) for the case of four different temperatures and three different speeds.

1.2 Let y_i, $i = 1, 2, 3$, be random variables with means μ_i, variances v_{ii}, and covariances v_{ij}. Write the matrix $\boldsymbol{M}$ as described in (1.17) and note that $E[\boldsymbol{M}] = \boldsymbol{V} = (v_{ij})$.

1.3 Suppose all variances are equal, $Var[y_i] = \phi_1$, and all covariances are equal, $Cov[y_i, y_k] = \phi_2$. Verify that the matrices $\boldsymbol{V}_1 = \boldsymbol{I}$ and $\boldsymbol{V}_2 = (\boldsymbol{U} - \boldsymbol{I})$ allow us to write the covariance matrix in the form of (1.20).

1.4 Let $\boldsymbol{V} = (v_{ij})$ be an arbitrary covariance matrix of size 3. Write $\boldsymbol{V}$ in the form of (1.19) using indicator matrices.

Section 1.4

1.5 For the simple linear regression model, suppose we have data (y_i, x_i), $i = 1, \ldots, 5$.

- **a.** Determine the inner product $\boldsymbol{J}^T\boldsymbol{x}$.
- **b.** Suppose we let $w_i = x_i - \bar{x}_i$, where $\bar{x}$ is the sample mean of the x_i. Consider the model $E[y] = \alpha_0 + \alpha_1 w$, and relate the coefficients in this model to β_0 and β_1 in (1.21).
- **c.** Compute $\boldsymbol{J}^T\boldsymbol{w}$.
- **d.** Suppose the inputs are equally spaced, $x_{i+1} = x_i + c$, for $i = 1, \ldots, 5$. Compute $\bar{x}$. Let $w_i = x_i - \bar{x}$ and compute $\boldsymbol{J}^T\boldsymbol{w}$, $\boldsymbol{J}^T\boldsymbol{w}^2$, and $\boldsymbol{w}^T\boldsymbol{w}^2$. Here, $\boldsymbol{w}^2$ has elements that are the squares of the elements of $\boldsymbol{w}$.

1.6 The line shown in Figure 1.1 gives an approximation to the relation between y and x for the golf tournament data described in Example 1.1. The equation is given by $\boldsymbol{y} = 15.4 + 1.6x$.

- **a.** Estimate the final score for an individual who had a score of 32 on the front nine. Repeat for a score of 38.
- **b.** How does these scores compare with the estimates of twice the score on the front nine?
- **c.** What does the equation imply about an individual who does well on the front nine and one who does poorly?
- **d.** Assuming that the mean value of the predictor is 34.67, how do you interpret the constant term in the equation $y = 67.4 + 1.5(x - 34.7)$?

1.7 In the particle board data described in Example 1.2, the equations for the linear and quadratic approximations are given by $STR = 63.9 + 2.8(TEMP)$ and $STR = 65.3 + 2.8(TEMP) - 0.48((TEMP) - 3.5)^2$.

- **a.** Write the linear equation in the form $STR = b_0 + 2.8((TEMP) - 3.5)$ where 3.5 is the average of the values of *TEMP*. How do you interpret the constant terms in these two forms of the equation.
- **b.** Rewrite the quadratic equation as $STR = b_0 + b_1(TEMP) + b_2(TEMP)^2$ and compare coefficients.
- **c.** Using the quadratic equation, estimate the value of *TEMP* that would give maximum *STR*. How would you feel about recommending that temperature for the process?

1.8 Write the design matrix for the model in (1.27) for Example 1.4 assuming four observations on glue one and three observations on glue 2.

- **a.** Verify the interpretation of β_0, β_1, and β_1 given for that example.
- **b.** Compute $\boldsymbol{J}^T\boldsymbol{x}$, $\boldsymbol{J}^T\boldsymbol{z}$, and $\boldsymbol{x}^T\boldsymbol{z}$.
- **c.** Repeat (a) and (b) if we define the indicator as $z = 1$ for glue 1 and $z = -1$ for glue 2.