COPYRIGHT

1 Introduction to Regression Models

1.1 INTRODUCTION

Regression modeling is an activity that leads to a mathematical description of a process in terms of a set of associated variables. The values of one variable frequently depend on the levels of several others. For example, the yield of a certain production process may depend on temperature, pressure, catalyst, and the rate of throughput. The number or the rate of defectives of a process may depend on the speed of the production line. The number of defective seals on toothpaste tubes may depend on the temperature and the pressure of the sealing process. The volume of a tree is related to the diameter of the tree at breast height, the height of the tree, and the taper of the tree. The fuel efficiency of an automobile depends, among others, on the weight of the car and characteristics of its body and engine. Employee efficiency may be related to the performance on employment tests, years of training, and educational background. The salaries of managers, athletes, and college teachers may depend on their seniority, the size of the market, and their performance. Many additional examples can be given, and in Exercise 1.1 we ask you to comment on several other relationships in detail.

The supply of a product depends on the price customers are willing to pay; one can expect that more products are brought to market when the price is high. Economists refer to this relationship as the **production function**. Similarly, the demand for a product depends on the price of the item, the price of the competition, and the amount spent on its advertisement. Economists refer to this relationship as the **demand function**. One can expect lower sales if the price is high, increased sales if the price of the competition is higher, and increased sales if more money is spent on promotion. However, price and advertising may also interact. Advertising may be more effective if the price is low; furthermore, the effect of the competition's price on sales may depend on one's own price. Also, seasonal components may have an impact on sales during a certain period because sales of a summer item during winter months will be low in northern states, irrespective of the product's price.

In all these situations we are interested in obtaining a "model" or a "law" (i.e., a mathematical description) for the relationship among the variables. Regression analysis deals with modeling the functional relationship between a **response variable** and one or more **explanatory variables**. In some instances one has a fairly good idea about the form of these models. Often the laws from physics or chemistry tell us how a response is related to the explanatory variables. These laws may involve complicated mathematical equations that contain functions such as logarithms and exponentials. In some instances, the constants in the equations are also known, but more often the constants need to be determined empirically by "fitting" the models to data. In many social science applications, theoretical models are absent, and one must develop empirical models that describe the main features of the relationship entirely from data.

Let us consider a few illustrative examples in detail.

1.2 EXAMPLES

1.2.1 PAYOUT OF AN INVESTMENT

Consider the payout of a principal P that you invest for a certain number of years (length of maturity) T, at an annual interest rate of $100R$ percent. We know from simple actuarial mathematics that the payout is given by

$$\text{Payout} = f(P, R, T) = P(1 + R)^T \tag{1.1}$$

provided that interest is compounded annually. With continuous compounding the resulting payout is slightly different. In this case, it can be calculated from $\text{Payout} = Pe^{RT}$, where e is Euler's number ($e = 2.71828\ldots$).

This first example illustrates a **deterministic relationship**. Each investment of principal P at rate R and maturity T leads to the exact same payout—nothing more and nothing less. We are very familiar with this law, and we would not need any data (or regression methods) to arrive at this particular model. However, assume for a moment that one was unfamiliar with the theory but had data on the payouts of different investments P, with different interest rates and maturities. Since the relationship is deterministic, payouts from identical investments would be identical and would not provide any additional information. Given this information, one would—after some trial and error and carefully constructed plots of the information—"see" the underlying functional relationship. This model would "fit" the data perfectly.

We have actually used the previous relationship to generate payouts for different principals, interest rates, and maturities, and we ask you in Exercise 1.2 to document the approach you use to find the model. You will experience firsthand the value of good theory; good theory will avoid much trial and error. Note that for payouts from continuous compounding, a plot of the logarithm of payout against the product of interest rate and length of maturity (RT) will show points falling on a line with slope one and intercept $\log(P)$.

1.2.2 PERIOD OF OSCILLATION OF A PENDULUM

Consider the period of oscillation (let us call it μ) of a pendulum of length L. It is a well-known fact from physics that the period of oscillation is proportional to the square root of the pendulum's length L, $\mu = \beta L^{1/2}$. However, the value of the proportionality factor β may be unknown.

In this example, we are given the functional form of the relationship, but we are missing information on the key constant, the proportionality factor β. In statistics we refer to unknown constants as **parameters**. The values of the parameters are usually determined by collecting data and using the resulting data to estimate the parameters.

The situation is also more complicated than in the first example because there is **measurement error**. Although the length of the pendulum is easy to measure, the determination of the period of oscillation is subject to variability. This means that sometimes our measurement of the "true" period of oscillation is too high and sometimes too low. However, for a calibrated measurement system we can expect that there is no bias (i.e., on average there is no error). If measured oscillation periods are plotted against the square roots of varying pendulum lengths, then the points will not line up exactly on a straight line through the origin, and there will be some scatter.

Mathematically, we characterize the relationship between the true period of oscillation μ and the length of the pendulum L as $\mu = \beta L^{1/2}$. However, the measured oscillation period OP is the sum of the true period (which we sometimes call the **signal**) and the measurement error ε (which we sometimes call the **noise**). Typically, we use a symmetric distribution about zero for the measurement error since the error is supposed to reflect only unbiased variability; if there were some bias in the measurement error, then such bias could be incorporated into the signal component of the model. Combining these two components (the signal and the noise) leads to the model

$$\text{OP} = \mu + \varepsilon = \beta L^{1/2} + \varepsilon \tag{1.2}$$

This model is similar to the one in Example 1.2.1 because we use theory (in this case, physics) to suggest the functional form of the relationship. However, in contrast to the previous example, we do not know certain constants (parameters) of the function. These parameters need to be estimated from empirical information. Furthermore, we have to deal with measurement variability, which leads to variability (or scatter) around the function (here, a line through the origin). We include a stochastic component ε in the model in order to capture this measurement variability.

1.2.3 SALARY OF COLLEGE TEACHERS

The third example represents a situation in which there is no theory about the functional form of the relationship and there is considerable variability in the measurements. In this situation, the data must perform "double duty," namely

to determine the functional form of the model and the values of the parameters in these functions. Moreover, the modeling must be carried out in the presence of considerable variability. We refer to such models as **empirical models** (in contrast to the theory-based models discussed in Examples 1.2.1 and 1.2.2), and we refer to the process of constructing such models as **empirical model building**. Examples of this type arise in the social sciences, economics, and business, where one usually has little *a priori* theory of what the functions should look like.

Consider building a model that explains the annual salary of a college professor. We probably agree that salary should be related to experience (the more experience, the higher the salary), teaching performance (better teachers are paid more), performance on research (significant papers and books increase the salary), and whether the job includes administrative duties (administrators usually get paid more). However, we are lacking a theory that tells us the functional form of the model. Although we know that salary should increase with years of experience, we do not know whether the function should be linear in years, quadratic, or whether an even more complicated function of the number of years should be used. The same applies to the other variables.

Moreover, we notice considerable variability in salary because professors with virtually identical background often are paid vastly different salaries. So there may be additional factors that one has overlooked. Feel free to brainstorm and add to this initial list of variables. For example, salary may also depend on gender and racial factors (use of these factors would be illegal), the year the professor was hired, whether the professor is easy to get along with, whether the professor has had a relationship with the dean's spouse or had made an inappropriate remark at last year's holiday party, and so on. Knowing these factors may improve the fit of the model to the data. However, even after factoring all these variables into the model, substantial random variation will still exist.

Another aspect that makes the modeling within the social science context so difficult is problems with measuring the variables. Consider, for example, the teaching performance of an instructor. Although student ratings from end-of-the-semester questionnaires could be used as an indicator of teaching performance, one could argue that these ratings are only a poor proxy. Demanding teachers, difficult subject matter, and lectures held in large classes are known to lower these ratings, thus biasing the measure. Assessment of research performance is another good case in point. One could use the number of publications and books and use this as a proxy for research. However, such a simple-minded count does not incorporate the quality of the publications. Even if one decides to somehow incorporate publication quality, one notices very quickly that reasonable people differ in their judgments. Of course, not being able to accurately measure the factors that we believe to have an effect on the response affects the results of the empirical modeling.

In summary, we find that empirical modeling faces many difficulties: little or no theory on how the variables fit together, often considerable variability in the

TABLE 1.1 HARDNESS DATA [DATA FILE: hardness]

Run	1	2	3	4	5	6	7	8	9	10	11	12	13	14
x = Temperature	30	30	30	30	40	40	40	50	50	50	60	60	60	60
y = Hardness	55.8	59.1	54.8	54.6	43.1	42.2	45.2	31.6	30.9	30.8	17.5	20.5	17.2	16.9

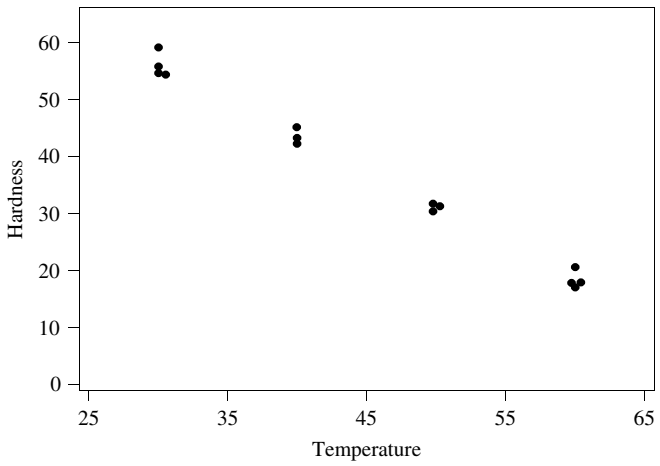


FIGURE 1.1 Scatter plot of hardness against quench bath temperature

response, and difficulties in obtaining appropriate measures for the variables that go into the model.

1.2.4 HARDNESS DATA

The quench bath temperature in a heat treatment operation was thought to affect the Rockwell hardness of a certain coil spring. An experiment was run in which several springs were treated under four temperatures: 30, 40, 50 and 60°C. The springs used in this experiment were selected from springs that had been produced under very similar conditions; all springs came from the same batch of material. Table 1.1 lists the (coded) hardness measurements and the temperatures at which the springs were treated.

We are interested in understanding how quench bath temperature affects hardness. Knowing this relationship is useful because it allows us to select the temperature that achieves a specified level of hardness.

Hardness is the dependent (or response) variable, and we denote it by y. Quench bath temperature is the independent (predictor, explanatory) variable that is supposed to help us predict the hardness; we denote it by x. For each experiment (coil spring—also called run or case) i, we have available a temperature that we select and control (the value x_i) and a measurement on the resulting hardness that we determine from the manufactured part (the value y_i). A scatter plot of hardness (y_i) against quench bath temperature (x_i) is shown in Figure 1.1.

We want to build a model (i.e., a mathematical relationship) to describe y in terms of x. Note that y cannot be a function of x alone since we have observed different y's (55.8, 59.1, 54.8, and 54.6) for the same $x = 30$. Furthermore, since no theoretical information is available to us to construct the model, we have to study the relationship empirically. The scatter plot of y against x indicates that y is approximately linear in x.

The scatter plot suggests the following model:

$$y(\text{hardness}) = \beta_0 + \beta_1 x(\text{temperature}) + \varepsilon \tag{1.3}$$

where β_0 and β_1 are the constants (parameters), and ε is the random disturbance (or error) that models the deviations from the straight line. The model is the sum of two components, the deterministic part (or signal) $\mu = \beta_0 + \beta_1 x$ and the random part ε. The deterministic part $\mu = \beta_0 + \beta_1 x$ is a linear function of x with parameters β_0 and β_1. More important, it is linear in the parameters β_0 and β_1, and hence we refer to this model as a **linear** model. The random component ε models the variability in the measurements around the regression line. This variability may come from the measurement error when determining the response y and/or changes in other variables (other than temperature) that affect the response but are not measured explicitly.

In order to emphasize that the model applies to each considered (and potential) experiment, we introduce subscripts. The temperature and the hardness from the ith experiment are written as (x_i, y_i). With these subscripts, our model can be expressed as

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad \text{where } i = 1, 2, \ldots, n \tag{1.4}$$

We complete the model specification by making the following assumptions about the random component ε:

$$\begin{gathered} E(\varepsilon_i) = 0,\ V(\varepsilon_i) = \sigma^2 \text{ for all } i = 1, 2, \ldots, n \\ \varepsilon_i \text{ and } \varepsilon_j \text{ are independent random variables for } i \neq j \end{gathered} \tag{1.5}$$

In this example, we treat x_i as deterministic. The experimenter selects the temperature and knows exactly the temperature of the quench bath. There is no uncertainty about this value. In later sections of this book (Section 2.9), we consider the case when the values of the explanatory variable are random. For example, the observed temperature may only be a "noisy" reading of the true temperature.

Our assumptions about the error ε and the deterministic nature of the explanatory variable x imply that the response y_i is a random variable, with mean $E(y_i) = \mu_i = \beta_0 + \beta_1 x_i$ and variance $V(\varepsilon_i) = \sigma^2$. Furthermore, y_i and y_j are independent for $i \neq j$.

The mean, $E(y_i) = \mu_i = \beta_0 + \beta_1 x_i$, is a linear function of x. The intercept β_0 represents $E(y)$ when $x = 0$. If the value $x = 0$ is uninteresting or impossible, the intercept is a rather meaningless quantity. The slope parameter β_1 represents the change in $E(y)$ if x is increased by one unit. For positive β_1, the mean $E(y)$

TABLE 1.2 UFFI DATA [DATA FILE: uffi]

$y = CH_2O$	$x =$ Air Tightness	$z =$ UFFI Present
31.33	0	0
28.57	1	0
39.95	1	0
44.98	4	0
39.55	4	0
38.29	5	0
50.58	7	0
48.71	7	0
51.52	8	0
62.52	8	0
60.79	8	0
56.67	9	0
43.58	1	1
43.30	2	1
46.16	2	1
47.66	4	1
55.31	4	1
63.32	5	1
59.65	5	1
62.74	6	1
60.33	6	1
53.13	7	1
56.83	9	1
70.34	10	1

increases for increasing x (and decreases for decreasing x). For negative β_1, the mean $E(y)$ decreases for increasing x and increases for decreasing x.

Our assumption in Eq. (1.5) implies that $V(y) = \sigma^2$ is the same for each x. This states that if we repeat experiments at a value of x (as is the case in this example), we should see roughly the same scatter at each of the considered x's. Figure 1.1 shows that the variability in hardness at the four levels of temperature—$x = 30$, 40, 50, and 60—is about the same.

1.2.5 UREA FORMALDEHYDE FOAM INSULATION

Data were collected to check whether the presence of urea formaldehyde foam insulation (UFFI) has an effect on the ambient formaldehyde concentration (CH_2O) inside the house. Twelve homes with and 12 homes without UFFI were studied, and the average weekly CH_2O concentration (in parts per billion) was measured. It was thought that the CH_2O concentration was also influenced by the amount of air that can move through the house via windows, cracks, chimneys, etc. A measure of "air tightness," on a scale of 0 to 10, was determined for each home.

The data are shown in Table 1.2. CH_2O concentration is the response variable (y) that we try to explain through two explanatory variables: the air tightness

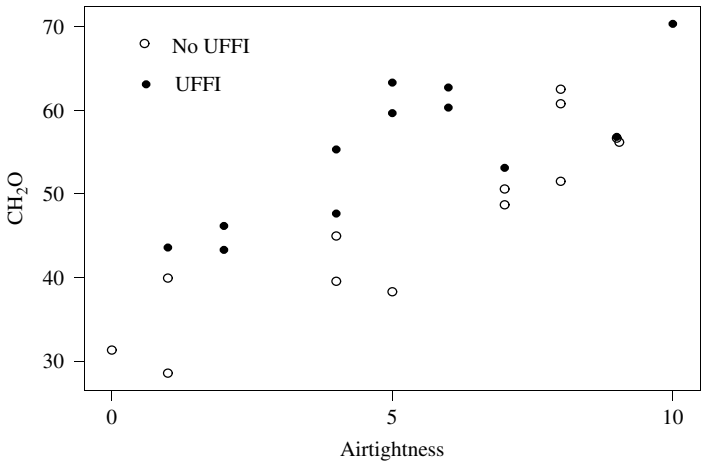


FIGURE 1.2
Scatter plot of CH_2O against air tightness for homes with and without urea formaldehyde foam insulation (UFFI)

of the home (x) and the absence/presence of UFFI (z). A scatter plot of CH_2O against air tightness for homes with and without UFFI is shown in Figure 1.2. The absence/presence of UFFI is expressed through an indicator variable. If insulation is present, then UFFI $= 1$; if it is absent, then UFFI $= 0$. The points in the scatter plot are labeled with solid and open circles, depending on whether or not UFFI is present. The plot shows strong evidence that CH_2O concentrations increase with increasing air tightness of the home.

It is important to emphasize that the data-generating mechanism in this example differs from that in the previous one. In the previous example, we were able to set the quench bath temperature at one of the four levels (30, 40, 50, and 60°C), conduct the experiment, and then measure the hardness of the spring. We refer to this as a **controlled** experiment, one in which the experimenter sets the values of the explanatory variable. In the current example, we select 12 houses with UFFI present and 12 houses in which it is not and measure the CH_2O concentration (the response y) as well as the air tightness (the explanatory x variable). It is not possible to preselect (or control) the air tightness value; the x values become available only after the houses are chosen. These data come from an **observational study**.

The basic objective of this particular observational study is to determine whether differences in the CH_2O concentrations can be attributed to the presence of insulation. Note, however, that we want to take into account the effect of air tightness. This can be achieved by considering the following model. Let

$$y = \beta_0 + \beta_1 x + \beta_2 z + \varepsilon \tag{1.6}$$

where

- y is the CH_2O concentration,
- x is the air tightness of the house,
- z is 1 or 0, depending on whether or not UFFI is present,

- ε is the error component that measures the random component, and
- β_0, β_1, and β_2 are constants (parameters) to be estimated.

CH_2O concentration is the response variable (y). It is the sum of a deterministic component ($\beta_0 + \beta_1 x + \beta_2 z$) and a random component ε. The random component ε is again modeled by a random variable with $E(\varepsilon) = 0$ and $V(\varepsilon) = \sigma^2$; it describes the variation in the CH_2O concentration among homes with identical values for x and z. Large variation in CH_2O concentration y among homes with the same insulation and tightness is characterized by large values of σ^2. The variability arises because of measurement errors (it is difficult to measure CH_2O accurately) and because of other aspects of the house (beyond air tightness and the presence of UFFI insulation) that have an influence on the response but are not part of the available information.

The deterministic component, $\beta_0 + \beta_1 x + \beta_2 z$, is the sum of three parts. The intercept β_0 measures the average CH_2O concentration for completely airtight houses ($x = 0$) without UFFI insulation ($z = 0$). The parameter β_2 can be explained as follows: Consider two houses with the same value for air tightness (x), the first house with UFFI ($z = 1$) and the second house without it ($z = 0$). Then $\beta_2 = E(y \mid \text{house 1}) - E(y \mid \text{house 2})$ represents the difference in the average CH_2O concentrations for two identical houses (as far as air tightness is concerned) with and without UFFI. This is exactly the quantity we are interested in. If $\beta_2 = 0$, we cannot link the formaldehyde concentration to the presence of UFFI.

Similarly, β_1 is the expected change in CH_2O concentrations that is due to a unit change in air tightness in homes with (or without) UFFI. Model (1.6) assumes that this change is the same for homes with and without UFFI. This is a consequence of the additive structure of the model: The contributions of the two explanatory variables, $\beta_1 x$ and $\beta_2 z$, get added. However, additivity does not have to be the rule. The more general model that involves the product of x and z,

$$y = \beta_0 + \beta_1 x + \beta_2 z + \beta_3 xz + \varepsilon \tag{1.7}$$

allows air tightness to affect the two types of homes differently. For a house without UFFI, $E(y) = \beta_0 + \beta_1 x$, and β_1 expresses the effect on the CH_2O concentrations of a unit change in air tightness. For a house with UFFI, $E(y) = (\beta_0 + \beta_2) + (\beta_1 + \beta_3)x$, and $(\beta_1 + \beta_3)$ expresses the effect of a unit change in air tightness. The effect is now different by the factor β_3.

1.2.6 ORAL CONTRACEPTIVE DATA

An experiment was conducted to determine the effects of five different oral contraceptives (OCs) on high-density lipocholesterol (HDLC), a substance found in blood serum. It is believed that high levels of this substance (the "good" cholesterol) help delay the onset of certain heart diseases. In the experiment, 50 women were randomly divided into five equal-sized groups; 10 women were assigned to each OC group. An initial baseline HDLC measurement was taken on each subject before oral contraceptives were started. After having used the respective

TABLE 1.3 ORAL CONTRACEPTIVE DATA [DATA FILE: contraceptive]

OC1 y = Final HDLC	OC1 z = Initial HDLC	OC2 y	OC2 z	OC3 y	OC3 z	OC4 y	OC4 z	OC5 y	OC5 z
43	49	58	56	100	102	50	57	41	37
61	73	46	49	52	64	50	55	58	60
45	55	66	64	49	60	52	64	58	39
46	55	59	63	51	51	58	49	69	60
59	63	71	90	48	59	65	78	68	71
57	53	64	56	51	57	71	63	64	63
56	51	53	46	40	63	52	62	46	51
68	74	50	64	52	62	49	50	56	64
46	58	68	75	44	61	49	60	51	45
47	41	35	58	50	58	58	59	57	58

drug for 6 months, a second HDLC measurement was made. The objective of the experiment was to study whether the five oral contraceptives differ in their effect on HDLC. The data are shown in Table 1.3. A scatter plot of final HDLC against the initial readings, ignoring the information on the respective treatment groups, is shown in Figure 1.3a. Figure 1.3b repeats this graph for groups 1, 2, and 5, using different plotting symbols to denote the three OC groups. Such a graph can highlight potential differences among the groups. (In order to keep the graph simple, only three groups are shown in Figure 1.3b).

Let y_i be the final HDLC measurement on subject $i (i = 1, 2, \ldots, 50)$ and let z_i be the initial HDLC reading. Furthermore, define five indicator variables $x_1, \ldots, x_5$ so that

$$\begin{aligned} x_{ik} &= 1 \text{ if subject } i \text{ is a participant in the } k\text{th OC group} \\ &= 0 \text{ otherwise} \end{aligned}$$

Here, we need two subscripts because there are five x variables. The first index in this double-subscript notation refers to the subject or case i; the second subscript refers to the explanatory variable (OC group) that is being considered. The following model relates the final HDLC measurement to six explanatory variables: the initial HDLC reading (z) and the five indicator variables $(x_1, \ldots, x_5)$. For subject i,

$$y_i = \alpha z_i + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_5 x_{i5} + \varepsilon_i \tag{1.8}$$

The usual assumption on the random component specifies that $E(\varepsilon_i) = 0$ and $V(\varepsilon_i) = \sigma^2$ for all i, and that ε_i and ε_j, for two different subjects $i \neq j$, are independent.

The deterministic component of the model, $E(y_i) = \alpha z_i + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_5 x_{i5}$, represents five parallel lines in a graph of $E(y_i)$ against the initial HDLC, z_i. The six parameters can be interpreted as follows: The parameter α represents the common slope. The coefficients $\beta_1, \beta_2, \ldots, \beta_5$ represent the intercepts of the five lines and measure the effectiveness of the five OC treatment groups. Their comparison is of primary interest because there is no difference among the five drugs when $\beta_1 = \beta_2 = \cdots = \beta_5$.

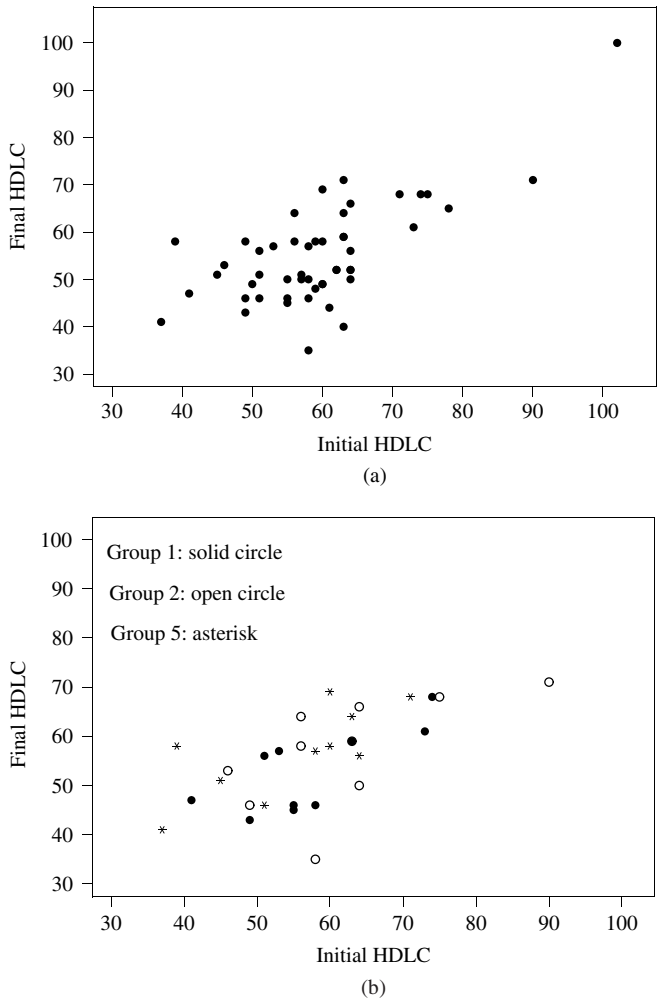


FIGURE 1.3
Scatter plots of final HDLC against initial HDLC

Consider two subjects (subjects i and j), both from the same OC group. Since the five indicator variables are the same on these two subjects ($x_{i1} = x_{j1}, \ldots, x_{i5} = x_{j5}$), the model implies $E(y_i) - E(y_j) = \alpha(z_i - z_j)$. The parameter α represents the expected difference in the final HDLC of two subjects who take the same drug but whose initial HDLC measurements differ by one unit. Next, consider two subjects with identical initial HDLC measurements but from different OC groups. Assume that the first woman is from group r, whereas the second is from group s. Then $E(y_i) - E(y_j) = \beta_r - \beta_s$, representing the expected difference in their final HDLC measurements.

1.2.7 GAS CONSUMPTION DATA

Let us give another illustration of empirical model building. Assume that we are interested in modeling the fuel efficiency of automobiles. First, we need to decide

how to measure fuel efficiency. A typical measure of fuel efficiency used by the Environmental Protection Agency (EPA) and car manufacturers is "miles/gallon." It expresses how many miles a car can travel on 1 gallon of fuel. However, there is an alternative way to express fuel efficiency considering gallons per 100 traveled miles, "gallons/100 miles." It expresses the amount of fuel that is needed to travel 100 miles. The second measure is the scaled reciprocal of the first: [gallons/100 miles] = 100/[miles/gallon]. In Chapter 6, we discuss how to intelligently choose among these two measures. Assume for the time being, that we have settled on the second measure, [gallons/100 miles].

Next, we need to think about characteristics of the car that can be expected to have an impact on fuel efficiency. Weight of the car is probably the first variable that comes to mind. Weight should have the biggest impact, as we know from physics that we need a certain force to push an object, and that force is related to the fuel input. Heavy cars require more force and, hence, more fuel. Size (displacement) of the engine probably matters also. So does, most likely, the number of cylinders, horsepower, the presence of an automatic transmission, acceleration from 0 to 60 mph, the wind resistance of the car, and so on. However, how many explanatory variables should be in the model, and in what functional form should fuel consumption be related to the explanatory variables? Theory does not help much, except that physics seems to imply that [gallons/100 miles] should be related linearly to weight. However, how the other variables enter into the model and whether there should be interaction effects (e.g., whether changes in weight affect fuel efficiency differently depending on whether the car has a small or large engine) are open questions.

Assume, for the sake of this introductory discussion, that we have settled on the following three explanatory variables: x_1 = weight, x_2 = engine displacement, and x_3 = number of cylinders. Table 1.4 lists the fuel efficiency and the characteristics of a sample of 38 cars. We assume that the data are a representative sample (**random sample**) from a larger population. You can always replicate this study by going to recent issues of *Consumer Reports* and selecting another random sample. If you have ample time, you can select all given cars and study the population. The fact that we are dealing with a **random** sample is very important because we want to extend any conclusions from the analysis of these 38 cars to the larger population at hand. Our results should not be restricted to just this one set of 38 cars, but our conclusions on fuel efficiency should apply more generally to the population from which this sample was taken. If our set of 38 cars is not a representative sample, then it is questionable whether the inference can be extended to the population.

Note that fuel consumption in Table 1.4 is given in "miles/gallon" and "gallons/100 miles." Convince yourself that the entries in the second column are obtained through the simple transformation, [gallons/100 miles] = 100/[miles/gallon]. In addition to data on weight, engine displacement, and number of cylinders, the table includes several other variables that we will use in later chapters.

TABLE 1.4 GAS CONSUMPTION DATA [DATA FILE: gasconsumption]

Miles/ gallon	Gallons/ 100 miles	Weight, 1000 lb	Displacement (cubic inches)	No. of Cylinders	Horsepower	Acceleration (sec)	Engine Type: V(0), straight(1)
16.9	5.917	4.360	350	8	155	14.9	1
15.5	6.452	4.054	351	8	142	14.3	1
19.2	5.208	3.605	267	8	125	15.0	1
18.5	5.405	3.940	360	8	150	13.0	1
30.0	3.333	2.155	98	4	68	16.5	0
27.5	3.636	2.560	134	4	95	14.2	0
27.2	3.676	2.300	119	4	97	14.7	0
30.9	3.236	2.230	105	4	75	14.5	0
20.3	4.926	2.830	131	5	103	15.9	0
17.0	5.882	3.140	163	6	125	13.6	0
21.6	4.630	2.795	121	4	115	15.7	0
16.2	6.173	3.410	163	6	133	15.8	0
20.6	4.854	3.380	231	6	105	15.8	0
20.8	4.808	3.070	200	6	85	16.7	0
18.6	5.376	3.620	225	6	110	18.7	0
18.1	5.525	3.410	258	6	120	15.1	0
17.0	5.882	3.840	305	8	130	15.4	1
17.6	5.682	3.725	302	8	129	13.4	1
16.5	6.061	3.955	351	8	138	13.2	1
18.2	5.495	3.830	318	8	135	15.2	1
26.5	3.774	2.585	140	4	88	14.4	0
21.9	4.566	2.910	171	6	109	16.6	1
34.1	2.933	1.975	86	4	65	15.2	0
35.1	2.849	1.915	98	4	80	14.4	0
27.4	3.650	2.670	121	4	80	15.0	0
31.5	3.175	1.990	89	4	71	14.9	0
29.5	3.390	2.135	98	4	68	16.6	0
28.4	3.521	2.670	151	4	90	16.0	0
28.8	3.472	2.595	173	6	115	11.3	1
26.8	3.731	2.700	173	6	115	12.9	1
33.5	2.985	2.556	151	4	90	13.2	0
34.2	2.924	2.200	105	4	70	13.2	0
31.8	3.145	2.020	85	4	65	19.2	0
37.3	2.681	2.130	91	4	69	14.7	0
30.5	3.279	2.190	97	4	78	14.1	0
22.0	4.545	2.815	146	6	97	14.5	0
21.5	4.651	2.600	121	4	110	12.8	0
31.9	3.135	1.925	89	4	71	14.0	0

The first car on this list has weight 4,360 pounds (i.e., the value for variable x_1 for the first car is $x_{11} = 4.360$), cubic displacement of 350 in.^{3} (i.e., the value for x_2 for the first car is $x_{12} = 350$), eight cylinders (i.e., the value for x_3 for the first car is $x_{13} = 8$), and gets 16.9 miles to the gallon. The value of the response y, fuel efficiency measured in gallons/100 miles, is $y_1 = 100/16.9 = 5.917$; the car needs 5.917 gallons to travel 100 miles. The second car of our data set measures

at $x_{21} = 4.054$, $x_{22} = 351$, $x_{23} = 8$, and $y_2 = 100/15.5 = 6.452$ (i.e., weight 4,054 pounds, 351 in.^{3} displacement, eight cylinders, and 6.452 gallons/100 miles). The last car (car 38) measures at $x_{38,1} = 1.925$, $x_{38,2} = 89$, $x_{38,3} = 4$, and $y_{38} = 100/31.9 = 3.135$ (i.e., weight 1,925 pounds, 89 in.^{3} displacement, four cylinders, and 3.135 gallons/100 miles).

Observe the notation that we use throughout this book. For the *i*th unit (in this case, the car), the values of the explanatory variables $x_1, x_2, \ldots, x_p$ (here, $p = 3$) and the response y are denoted by $x_{i1}, x_{i2}, \ldots, x_{ip}$, and y_i. Usually, there are several explanatory variables, not just one. Hence, we must use a double-index notation for x_{ij}, where the first index $i = 1, 2, \ldots, n$ refers to the case, and the second index $j = 1, 2, \ldots, p$ refers to the explanatory variable. For example, $x_{52} = 98$ is the value of the second explanatory variable (displacement, x_2) of the fifth car. Since we are dealing with a single response variable y, there is only one index (for case) in y_i.

A reasonable starting model relates fuel efficiency (gallons/100 miles) to the explanatory variables in a linear fashion. That is,

$$y = \mu + \varepsilon = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon \tag{1.9}$$

As before, the dependent variable is the sum of a random component, ε, and a deterministic component, $\mu = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$, which is linear in the parameters β_0, β_1, β_2, and β_3.

Cars with the same weight, same engine displacement, and the same number of cylinders can have different gas consumption. This variability is described by ε, which is taken as a random variable with $E(\varepsilon) = 0$ and $V(\varepsilon) = \sigma^2$. If we consider cars with the same weight, same engine displacement, and same number of cylinders, then the average deviation from the mean value in gas consumption of these "alike" cars is zero. The variance σ^2 provides a measure of the variability around the mean value. Furthermore, we assume that $E(\varepsilon) = 0$ and $V(\varepsilon) = \sigma^2$ is the same for all groups of cars with identical values on x_1, x_2, and x_3. The variability is there because of measurement variability in determining the gas consumption. However, it also arises because of the presence of other characteristics of the car that affect fuel consumption but are not part of the data set. Cars may differ with respect to such omitted variables. If the omitted factors affect fuel consumption, then the fuel consumption of cars that are identical on the measured factors will be different.

The deterministic component $\mu = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$ is linear in the parameters β_0, β_1, β_2, and β_3. We expect a positive value for the coefficient β_1 because a heavier car (with fixed engine displacement and number of cylinders) needs more fuel. Similarly, we expect a positive coefficient β_2 because a larger engine on a car of fixed weight and number of cylinders should require more fuel. We also expect a positive coefficient for β_3 because more cylinders on a car of fixed weight and engine displacement should require more fuel.

In order to understand the deterministic component μ more fully, consider two cars i and j with identical engine displacement and number of cylinders.

Since $x_{i2} = x_{j2}$ and $x_{i3} = x_{j3}$, the difference

$$E(y_i) - E(y_j) = \beta_1(x_{i1} - x_{j1})$$

Thus, β_1 represents the difference in the mean values of y (the mean difference in the gas consumption) of two cars whose weights (x_1) differ by one unit but that have the same engine displacement (x_2) and the same number of cylinders (x_3). Similarly, β_2 represents the difference in the mean values of y of two cars whose engine displacements (x_2) differ by one unit but that have the same weight (x_1) and the same number of cylinders (x_3). The parameter β_3 represents the difference in the mean values of y of two cars whose number of cylinders (x_3) differ by one unit but that have the same weight (x_1) and the same engine displacement (x_2).

In the modeling context, one often is not certain whether the variables under consideration are important or not. For instance, we might be interested in the question whether or not x_3 (number of cylinders) is necessary to predict y (gas consumption) once we have included the weight x_1 and the engine displacement x_2 in the model. Thus, we are interested in a test of the hypothesis that $\beta_3 = 0$, given that x_1 and x_2 are in the model. Such tests may lead to the exclusion of certain variables from the model. On the other hand, other variables such as horsepower x_4 may be important and should be included. Then the model needs to be extended so that its predictive capability is increased.

The model in Eq. (1.9) is quite simple and should provide a useful starting point for our modeling. Of course, we do not know the values of the model coefficients, nor do we know whether the functional representation is appropriate. For that we need data. One must keep in mind that there are only 38 observations and that one cannot consider models that contain too many unknown parameters. A reasonable strategy starts with simple **parsimonious** models such as the one specified here and then checks whether this representation is capable of explaining the main features of the data. A parsimonious model is simple in its structure and economical in terms of the number of unknown parameters that need to be estimated from data, yet capable of representing the key aspects of the relationship. We will say more on model building and model checking in subsequent chapters. The introduction in this chapter is only meant to raise these issues.

1.3 A GENERAL MODEL

In all of our examples, we have looked at situations in which a single response variable y is modeled as

$$y = \mu + \varepsilon \tag{1.10a}$$

The deterministic component μ is written as

$$\mu = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p \tag{1.10b}$$

where $x_1, x_2, \ldots, x_p$ are p explanatory variables. We assume that the explanatory variables are "fixed"—that is, measured without error. The parameter

$\beta_i (i = 1, 2, \ldots, p)$ is interpreted as the change in μ when changing x_i by one unit while keeping **all** other explanatory variables the same.

The random component ε is a random variable with zero mean, $E(\varepsilon) = 0$, and variance $V(\varepsilon) = \sigma^2$ that is constant for all cases and that does not depend on the values of $x_1, x_2, \ldots, x_p$. Furthermore, the errors for different cases, ε_i and ε_j, are assumed independent. Since the response y is the sum of a deterministic and a random component, we find that $E(y) = \mu$ and $V(y) = \sigma^2$.

We refer to the model in Eq. (1.10) as **linear in the parameters**. To explain the idea of linearity more fully, consider the following four models with deterministic components:

$$\begin{aligned} &\text{i.} \quad \mu = \beta_0 + \beta_1 x \\ &\text{ii.} \quad \mu = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \\ &\text{iii.} \quad \mu = \beta_0 + \beta_1 x + \beta_2 x^2 \\ &\text{iv.} \quad \mu = \beta_0 + \beta_1 \exp(\beta_2 x) \end{aligned} \tag{1.11}$$

Models (i)–(iii) are linear in the parameters since the derivatives of μ with respect to the parameters β_i, $\partial\mu/\partial\beta_i$, do not depend on the parameters. Model (iv) is nonlinear in the parameters since the derivatives $\partial\mu/\partial\beta_1 = \exp(\beta_2 x)$ and $\partial\mu/\partial\beta_2 = \beta_1 x \exp(\beta_2 x)$ depend on the parameters.

The model in Eqs. (1.10a) and (1.10b) can be extended in many different ways. First, the functional relationship may be nonlinear, and we may consider a model such as that in Eq. (1.11iv) to describe the nonlinear pattern. Second, we may suppose that $V(y) = \sigma^2(x)$ is a function of the explanatory variables. Third, responses for different cases may not be independent. For example, we may model observations (e.g., on weight) that are taken on the same subject over time. Measurements on the same subject taken close together in time are clearly related, and the assumption of independence among the errors is violated. Fourth, several different response variables may be measured on each subject, and we may want to model these responses simultaneously. Many of these extensions will be discussed in later chapters of this book.

1.4 IMPORTANT REASONS FOR MODELING

Statistical modeling, as discussed in this text, is an activity that leads to a mathematical description of a process in terms of the variables of the process. Once a satisfactory model has been found, it can be used for several different purposes.

i. Usually, the model leads to a simple description of the main features of the data at hand. We learn which of the explanatory variables have an effect on the response. This tells us which explanatory variables we have to change in order to affect the response. If a variable does not affect a response, then there may be little reason to measure or control it. Not having to keep track of something that is not needed can lead to significant savings.

ii. The functional relationship between the response and the explanatory variables allows us to estimate the response for given values of the explanatory variables. It makes it possible to infer the response for values of the explanatory variables that were not studied directly. It also allows us to ask "what if"-type questions. For example, a model for sales can give us answers to questions of the following form: "What happens to sales if we keep our price the same, but increase the amount of advertising by 10%?" or "What happens to the gross national product if interest rates decrease by one percentage point?" Knowledge of the relationship also allows us to control the response variable at certain desired levels. Of course, the quality of answers to such questions depends on the quality of the models that are being used.

iii. Prediction of future events is another important application. We may have a good model for sales over time and want to know the likely sales for the next several future periods. We may have developed a good model relating sales at time t to sales at previous periods. Assuming that there is some stability over time, we can use such a model for making predictions of future sales.

In some situations, the models seen here are well grounded in theory. However, often theory is lacking and the models are purely descriptive of the data that one has collected. When a model lacks a solid theoretical foundation, it is questionable whether it is possible to extrapolate the results to new cases that are different from the ones occurring in the studied data set. For example, one would be very reluctant to extrapolate the findings in Example 1.2.4 and predict hardness for springs that were subjected to temperatures that are much higher than 60°C.

iv. A regression analysis may show that a variable that is difficult and expensive to measure can be explained to a large extent by variables that are easy and cheap to obtain. This is important information because we can substitute the cheaper measurements for the more expensive ones. It may be quite expensive to determine someone's body fat because this requires that the whole body be immersed in water. It may be expensive to obtain a person's bone density. However, variables such as height, weight, and thickness of thighs or biceps are easy and cheap to obtain. If there is a good model that can explain the expensively measured variable through the variables that are easy and cheap to obtain, then one can save money and effort by using the latter variables as proxies.

1.5 DATA PLOTS AND EMPIRICAL MODELING

Good graphical displays are very helpful in building models. Let us use the data in Table 1.4 to illustrate the general approach. Note that with one response and p explanatory variables, each case (in this situation, each car) represents a point in $(p + 1)$ dimensional space. Most empirical modeling starts with plots of the data in a lower dimensional space. Typically, one starts with pairwise

(two-dimensional) scatter plots of the response against each of the explanatory variables. The scatter plot of fuel consumption (gallons/100 miles) against weight of the car in Figure 1.4a illustrates that heavier cars require more fuel. It also shows that the relationship between fuel consumption and weight is well approximated by a linear function. This is true at least over the observed weight range from approximately 2,000 to 4,000 pounds. How the function looks for very light and very heavy cars is difficult to tell because such cars are not in our group of considered cars; extrapolation beyond the observed range on weight is certainly a very tricky task.

Knowing that this relationship is linear simplifies the interpretation of the relationship because each additional 100 pounds of weight increases fuel efficiency by the same amount, irrespective of whether we talk about a car weighing 2,000 or 3,500 pounds. For a quadratic relationship the interpretation would not be as straightforward because the change in fuel consumption implied by a change in weight from 2,000 to 2,100 pounds would be different than the one implied by a change in weight from 3,500 to 3,600 pounds.

Another notable aspect of the data and the graph in Figure 1.4a is that the observations do not lie on the line exactly. This is because of variability. Our model recognizes this by allowing for a random component. On average, the fuel efficiency can be represented by a simple straight-line model, but individual observations (the fuel consumption of individual cars) vary around that line. This variation can result from many sources. First, it can be pure measurement error. Measuring the fuel consumption on the very same car for a second time may result in a different number. Second, there is variation in fuel consumption among cars taken from the very same model line. Despite being from the same model line and having the same weight, cars are not identical. Third, cars of identical weight may come from different model lines with very different characteristics. It is not just weight that affects the fuel consumption; other characteristics may have an effect. Engine sizes may be different and the shapes may not be the same. One could make the model more complicated by incorporating these other factors into it. Although this would reduce the variability in fuel consumption, one should not make the function so complicated that it passes through every single point. Such an approach would ignore the natural variability in measurements and attach too much importance to random variation. Henri Poincare, in *The Foundations of Science* [Science Press, New York, 1913 (reprinted 1929), p. 169] expresses this very well when he writes,

> Pass to an example of a more scientific character. I wish to determine an experimental law. This law, when I know it, can be represented by a curve. I make a certain number of isolated observations; each of these will be represented by a point. When I have obtained these different points, I draw a curve between them, striving to pass as near to them as possible and yet preserve for my curve a regular form, without angular points, or inflections too accentuated, or brusque variation of the radius of curvature. This curve will represent for me the probable law, and I assume not only that it will tell me the values of the function intermediate

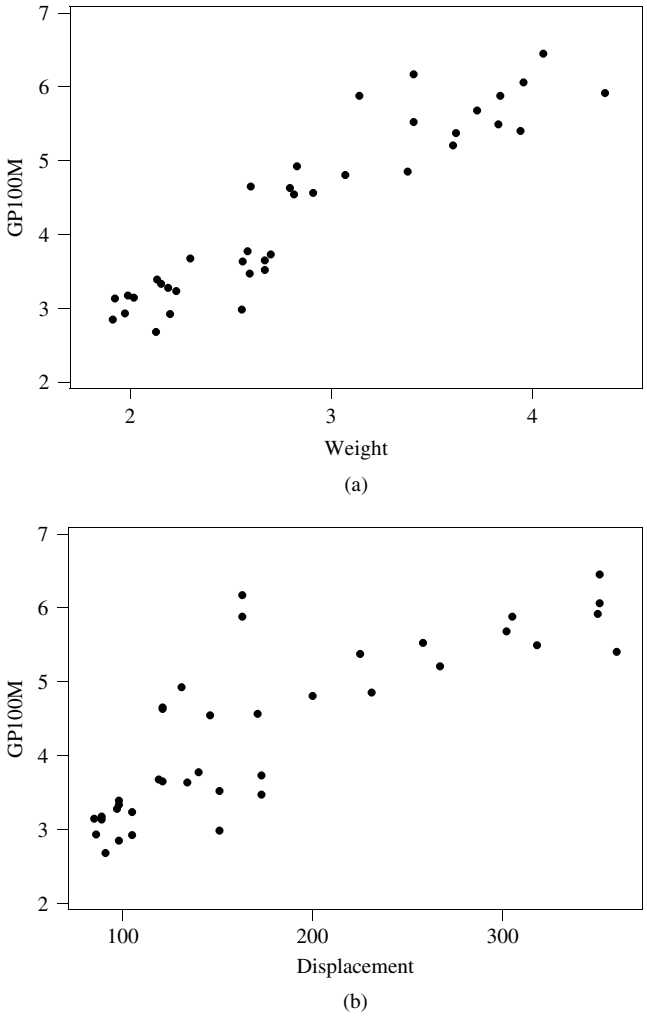


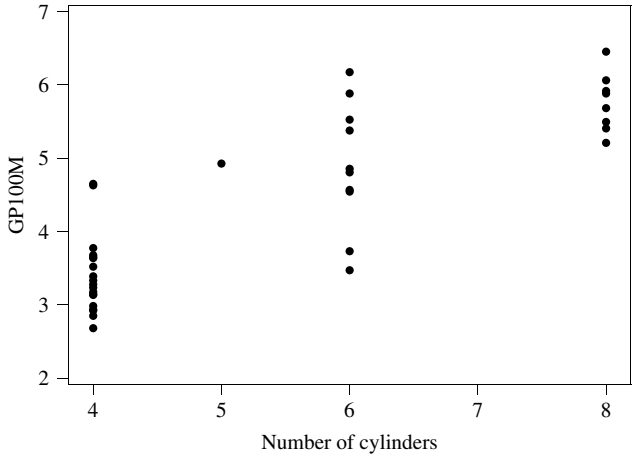
FIGURE 1.4
(a) Pairwise scatter plot of *y* (gallons/100 miles) against weight. (b) Pairwise scatter plot of *y* (gallons/100 miles) against displacement. (c) Pairwise scatter plot of *y* (gallons/100 miles) against number of cylinders. (d) Three-dimensional plot of *y* (gallons/100 miles) against weight and displacement

between those which have been observed, but also that it will give me the observed values themselves more exactly than direct observation. This is why I make it pass near the points, and not through the points themselves.

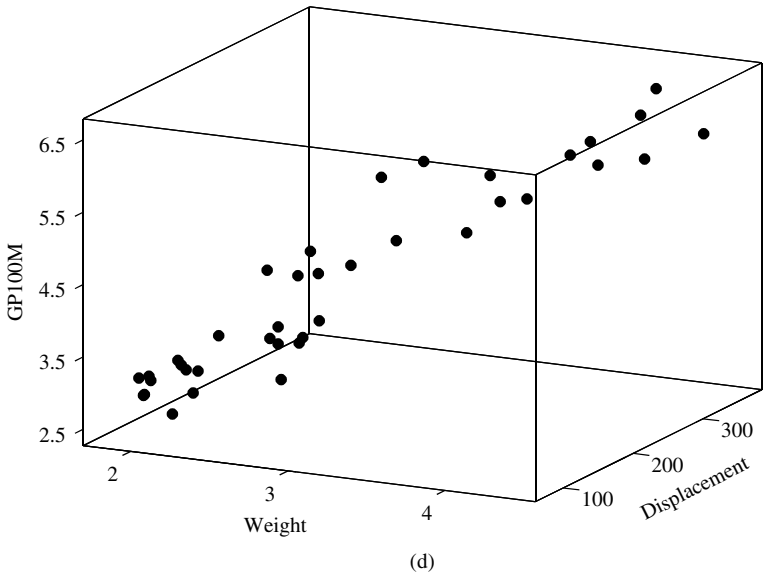
Here, we have described a two-dimensional representation of fuel consumption y and weight x_1. Similar scatter plots can be carried out for fuel consumption (y) and displacement x_2 and also fuel consumption (y) and number of cylinders x_3. The graphs shown in Figures 1.4b and 1.4c indicate linear relationships, even though the strengths of these relationships differ.

We notice from Figure 1.4 that each pairwise scatter plot exhibits linearity. Is this enough evidence to conclude that the model with the three explanatory variables should be linear also? The answer to this question is "**no**" in general. Although the linear model, $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$, may provide a good starting point, the model may miss more complicated associations.

FIGURE 1.4
(Continued)



(c)



(d)

Two-dimensional displays are unable to capture the **joint** relationships among the response and more than one explanatory variable. In order to bring out the joint relationships between a response and two explanatory variables (e.g., weight x_1 and displacement x_2), one needs to look at a three-dimensional graph of fuel consumption y on both x_1 and x_2. This is done in Figure 1.4d. One notices that in such graphs it becomes considerably more difficult to recognize patterns, especially when working with relatively small data sets. However, at least from this graph it appears that the relationship can be approximated by a plane. The equation of a plane is given by $y(\text{gallons/100 miles}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$. This function implies that for a fixed value of x_2, a change in x_1 by one unit changes

the response y by β_1 units. Similarly, for a fixed value of x_1, a change in x_2 by one unit changes the response y by β_2 units. The effects of changes in x_1 and x_2 are additive. Additivity is a special feature of this particular representation. It is a convenient simplification but need not be true in general. For some relationships the effects of a change in one explanatory variable depend on the value of a second explanatory variable. One says that the explanatory variables **interact** in how they affect the response y.

Up to now, we have incorporated the effects of x_1 and x_2. What about the effect of the third explanatory variable x_3? It is not possible to display all four variables in a four-dimensional graph. However, judging from the pairwise scatterplots and the three-dimensional representations (y, x_1, x_2), (y, x_1, x_3), and (y, x_2, x_3), our linear model in x_1, x_2, and x_3 may provide a sensible starting point.

1.6 AN ITERATIVE MODEL BUILDING APPROACH

An understanding of relationships can be gained in several different ways. One can start from a well-developed theory and use the data mostly for the estimation of unknown parameters and for checking whether the theory is consistent with the empirical information. Of course, any inconsistencies between theory and data should lead to a refinement of the model and a subsequent check whether the revised theory is consistent with the data.

Another approach, and one that is typically used in the social sciences, is to start from the data and use an empirical modeling approach to derive a model that provides a reasonable characterization of the relationship. Such a model may in fact lead to a new theory. Of course, theories must be rechecked against new data, and in cases of inconsistencies with the new information, new models must be developed, estimated, and checked again. Notice that good model building is a continual activity. It does not matter much whether one starts from theory or from data; what matters is that this process continues toward convergence.

A useful strategy for building such models is given in Figure 1.5. Initially, a tentative model is specified from relevant data and/or available theory. In some cases, theory will suggest certain models. In other situations, theory may not

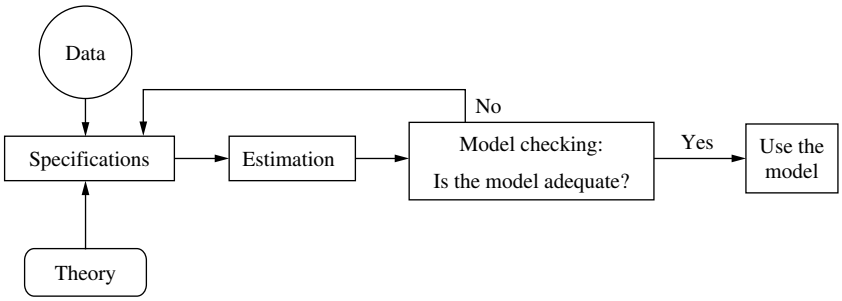


FIGURE 1.5 A model building system

exist or may be incomplete, and data must be used to specify a reasonable initial model; exploratory data analysis and innovative ways of displaying information graphically are essential. The tentatively entertained model usually contains unknown parameters that need to be estimated. Model fitting procedures, such as **least squares** or **maximum likelihood**, have been developed for this purpose. This is discussed further in the next chapter.

Finally, the modeler must check the adequacy of the fitted model. Inadequacy can occur in several ways. For example, the model may miss important variables, it may include inappropriate and unnecessary variables that make the interpretation of the model difficult, and the model may misspecify the functional form. If the model is inadequate, then it needs to be changed, and the iterative cycle of "model specification—parameter estimation—model checking" must be repeated. One needs to do this until a satisfactory model is obtained.

1.7 SOME COMMENTS ON DATA

This discussion shows that good data are an essential component of any model building. However, not all data are alike, and we should spend some time discussing various types of data. We should distinguish between **data arising from designed experiments** and data from **observational studies**.

Many data sets in the physical sciences are the result of designed studies that are carefully planned and executed. For example, an engineer studying the impact of pressure and temperature on the yield of a production process may manufacture several products under varying levels of pressure and temperature. He or she may select three different pressures and four different settings for temperature and conduct one or several experiments at each of the $(3) \times (4) = 12$ different factor-level combinations. A good experimenter will suspect that other factors may have an impact on the yield but may not know for sure which ones. It could be the purity of the raw materials, environmental conditions in the plant during the manufacture, and so on. In order to minimize the effects of these uncontrolled factors, the investigator will randomize the arrangement of the experimental runs. He or she will do this to minimize the effects of unknown time trends. For example, one certainly would not want to run all experiments with the lowest temperature on one day and all experiments using the high temperature on another. If the process is sensitive to daily fluctuations in plant conditions, an observed difference in the results of the two days may not be due to temperature but due to the different conditions in the plant. Good experimenters will be careful when changing the two factors of interest, keeping other factors as uniform as possible. What is important is that the experimenter is actively involved in all aspects of obtaining the data.

Observational data are different because the investigator has no way of impacting the process that generates the data. The data are taken just as the data-generating process is providing them. Observational data are often referred to as "happenstance" data because they just happen to be available for analysis.

Economic and social science information is usually collected through a **census** (i.e., every single event is recorded) or through **surveys**. The problem with many social science data sets is that several things may have gone wrong during the data-gathering process, and the analyst has no chance to recover from these problems. A survey may not be representative of the population that one wants to study. Data definitions may not match exactly the factors that one wants to measure, and the gathered data may be poor proxies at best. Data quality may be poor because there may not have been enough time for careful data collection and processing. There may be missing data. The data that come along may not be "rich" enough to separate the effects of competing factors.

Consider the following example as an illustration. Assume that you want to explain college success as measured by student grade point average. Your admission office provides the student ACT scores (on tests taken prior to admission), and you have survey data on the number of study hours per week. Does this information allow you to develop a good model for college success? Yes, to a certain degree. However, there are also significant problems. First, college GPA is quite a narrow definition of college success. GPA figures are readily available, but one needs to discuss whether this is the information one really wants. Second, the range of ACT scores may be not wide enough to find a major impact of ACT scores on college GPA. Most good universities do not accept marginal students with low ACT scores. As a consequence, the range of ACT scores at your institution will be narrow, and you will not see much effect over that limited range. Third, study hours are self-reported, and students may have a tendency to make themselves look better than they really are. Fourth, ACT scores and study hours tend to be correlated. Students with a high ACT scores tend to have good study skills; it will be rare to find someone with a very high ACT score who does not study. The correlation between the two explanatory variables, ACT score and study hours, makes it difficult to separate the effects on college GPA of ACT scores and study hours.

EXERCISES

1.1. Consider the following relationships. Comment on the type of relationships that can be expected, supporting your discussion with simple graphs. In which of these cases can you run experiments that can help you learn about the relationship between the response y and the explanatory variables x?

a. Tensile strength of an alloy may be related to hardness and density of the stock.

b. Tool life may depend on the hardness of the material to be machined and the depth of the cut.

c. The weight of the coating of electrolytic tin plate may be affected by the current, acidity, rate of travel of the strip, and distance from the anode.

d. The diameter of a condenser coil may be affected by the thickness of the coil, number of turns, and tension in the winding.

e. The moisture content of lumber may depend on the speed of drying, the drying temperature, and the dimension of the pieces.

f. The performance of a foundry may be affected by atmospheric conditions.

g. The life of a light bulb may vary with the quality of the filament; the tile finish may depend on the temperature of firing.

1.2. Consider the payout of the following 18 investments (data file **payout**). The investments vary according to the invested principal P, the monthly interest rate R, and the length of maturity T (in months). The data in the following table were generated from the deterministic continuous compounding model, $Payout = Pe^{RT}$. No uncertainty was added to the equation.

It is reasonable to assume that the payout increases with the principal, the interest rate, and the maturity. However, without theory, the form of the relationship is not obvious. An empirical model building strategy that does not utilize available theory will be inefficient and may never find the hidden model. Construct scatter plots of the response (payout) on the explanatory variables (principal, interest rates, and maturity), and you will see what we mean. It is quite difficult to "see" the correct relationship.

Plot the logarithm of payout against the product of interest rate and maturity, and label the points on the scatter plot according to the invested principal (1,000, 1,500, and 2,000). What do you see, and how does this help you arrive at the correct model?

Principal	Interest Rate	Time (Months)	Payout
1,000	0.001	12	1,012.1
1,000	0.002	24	1,049.2
1,000	0.003	12	1,036.7
1,000	0.001	36	1,036.7
1,000	0.002	12	1,024.3
1,000	0.003	36	1,114.0
1,500	0.001	36	1,555.0
1,500	0.002	24	1,573.8
1,500	0.003	24	1,612.0
1,500	0.010	12	1,691.2
1,500	0.010	36	2,150.0
1,500	0.010	12	1,691.2
1,200	0.015	12	1,436.7
1,200	0.015	36	2,059.2
1,200	0.015	36	2,059.2
1,200	0.005	24	1,353.0
1,200	0.005	12	1,274.2
1,200	0.005	36	1,436.7

1.3. Look ahead in the book and read the problem descriptions of several exercises in Chapters 2 and 4–8. Find examples where the data originate from a designed experiment. Find examples where the data are the result of observational studies.

1.4. List other examples of designed experiments and observational studies.

1.5. Look ahead in the text and consider the data sets in Exercises 2.8, 2.9, 2.16–2.18, 2.21, 2.24, and 2.25 of Chapter 2. Construct pairwise scatter plots of the response variable against the explanatory variable(s). Discuss whether a linear model gives an appropriate description of the relationship. Speculate on the reasons for the variability of the response around the fitted line.

1.6. Experiment with three-dimensional displays of the information. For example, consider the data generated in Exercise 1.2. Consider the logarithm of payout as the response and the logarithm of the principal and the product of interest rate and maturity as the two explanatory variables. Discuss whether it is easy to spot three-dimensional relationships from such graphs. As a second example, consider the silkworm data in Exercise 4.15 of Chapter 4.

1.7. Explain the statement that a nonlinear model in the explanatory variables may turn out to be linear in the parameters. Give examples. For instance, is the quadratic model in x a model that is linear in the parameters? Explain.

1.8. Give examples of regression models that are nonlinear in the regression parameters.

1.9. Can you think of situations in which the variability in the response depends on the

level of the explanatory variables? For example, consider sales that increase over time. Why is it reasonable to expect more variability in sales when the sales are high? Discuss.

1.10. Causality and correlation. Assume that a certain data set exhibits a strong association among two variables. Does this imply that there is a causal link? Can you think of examples where two variables are correlated but not causally related? What about the annual number of storks and the annual number of human births? Assume that your data come from a time period of increasing prosperity, such as the one immediately following World War II. Prosperity may impact the storks, and it may also affect couples' decisions to have families. Hence, you may see a strong (positive) correlation between the number of storks and the number of human births. However, you know that there is no causal effect. Can you describe the underlying principle of this example? Give other examples?

1.11. Collect the following information. Obtain average test scores for the elementary schools in your state (region). Obtain data on the proportion of children on subsidized lunch. Construct scatter plots. Do you think that there is a causal link between test scores and the proportion of children on subsidized lunch? If not, how do you explain the results you see. What if you had data on average income or the educational level of parents in these districts? Do you expect similar results?

1.12. Salary raises are usually expressed in percentage terms. This means that two people with the same percentage raise, but different previous salaries, will get different monetary (dollar) raises. Assume that the relative raise (R = RelativeRaise = PercentageRaise/100) is strictly proportional to performance (that is, RelativeRaise = βPerformance).
a. A plot of RelativeRaise against Performance exhibits a perfect linear relationship through the origin. Would a plot of AbsoluteRaise against Performance also exhibit a perfect linear association? Would a regression of AbsoluteRaise against Performance lead to the desired slope parameter β?
b. Consider the logarithmic transformation of the ratio (CurrentSalary/ PreviousSalary). What if you were to plot the logarithm of this ratio against the performance? How would this help you?

1.13. Consider Example 1.2.6 in which we studied the effectiveness of five oral contraceptives. We used model (1.8),

$$y_i = \alpha z_i + \beta_1 x_{i1} + \cdots + \beta_5 x_{i5} + \varepsilon_i$$

where z_i and y_i are the HDLC readings at the beginning and after 6 months, and $x_1, \ldots, x_5$ are indicators for the five treatment (contraceptive) groups.

How would you convince someone that this is an appropriate specification? In order to address this question, you may want to look at five separate scatter plots of y against z, one for each contraceptive group. Make sure these graphs are made with identical scales on both axes. Have the statistical software of your choice draw in the "best fitting" straight lines. Your model in Eq. (1.8) puts certain requirements on the slopes of these five graphs. What are these requirements?

How would you explain a model in which $\alpha = 1$? In this case, the emphasis is on changes in the HDLC, and the question becomes whether the magnitudes of the changes are related to the contraceptives. How would you analyze the data under this scenario?