# 2

# Regression and Prediction

## 2.1 Overview

In this book we are mainly concerned with the examination and use of associations among variables. Understanding these associations can be useful in many ways, and one of the most important and most common is **prediction**. Knowledge of the associations among variables often leads to methods for *predicting* the value of one quantity by using values of related quantities. Such knowledge may also lead to methods for *controlling* the value of one variable by adjusting the values of related variables, and in some situations it may enhance our understanding of the underlying processes involved. Regression analysis offers us a sensible and sound approach for examining associations among variables and for obtaining good rules for prediction. In Section 2.2 we introduce the subject of prediction as used in various branches of science, in industry, in business, and in everyday affairs. In Section 2.3 we introduce regression, which is the subject of the remainder of the book. Section 2.4 contains chapter exercises.

## 2.2 Prediction

The following example is a simple illustration of the use of prediction.

### EXAMPLE 2.2.1

A large university, which receives several thousand applications for admission each year, wants to admit only those students who will successfully complete their first year. To accomplish this, the director of admissions wants to predict the grade point average (GPA) that each applicant would receive at the end of the first year if admitted. Only those students will be admitted whose GPA is predicted to be 3.0

or higher (on a 4-point scale). The GPA is the quantity (usually called the **response variable**) to be predicted.

Suppose the director of admissions decides to use the score on the mathematics part of the Scholastic Aptitude Test (SAT) as the quantity (sometimes called the **predictor variable**, the **predictor factor**, or the **explanatory variable**) on which the prediction is to be based. The predictor variable is often denoted by $X$ and the response variable by $Y$. It would be convenient to have a rule or a formula that would tell the director of admissions how to compute a predicted GPA of an applicant based on his or her SAT mathematics score. Such a rule or formula is often expressed as a function, say $P_Y(x)$, called a **prediction function**. The notation $P_Y(x)$ means: *predicted Y value of an item whose X value equals x*. If a prediction function $P_Y(x)$ were available, then we could predict that a prospective student who had a SAT score of 480 (i.e., $X = 480$) would obtain a GPA equal to $P_Y(480)$ at the end of one year. For an illustration, suppose that $P_Y(x) = 0.324 + 0.00484x$. Then $P_Y(480) = 0.324 + 0.00484(480) = 2.65$. For a student with an SAT score of 680 (i.e., $X = 680$ ), the predicted GPA would be $P_Y(680) = 3.62$, etc.

Suppose there are $N$ applicants ($N$ unknown) for admission over, say, the next ten years, and $(Y_I, X_I)$ represents the GPA and SAT, respectively, for the $I$th applicant. In general, $Y_I$ and $X_I$ are unknown, but conceptually they could become known if the $I$th student took the SAT, was admitted to the university, and completed the first year. We refer to the set of $N$ numbers $(Y_I, X_I)$ for $I = 1, 2, \ldots, N$ as the **target population** under study. Clearly this is a *conceptual* population. Associated with the $I$th applicant in the target population there are two numbers of interest, (1) $X_I =$ SAT math score and (2) $Y_I =$ GPA at the end of the first year. If the prediction function used is $P_Y(x)$, then the *predicted* GPA for the $I$th applicant is $P_Y(X_I)$ and the *actual* GPA is $Y_I$. In general, the prediction will not be exact, so $Y_I$ will not be equal to $P_Y(X_I)$, but if the prediction is good, $Y_I$ will be close to $P_Y(X_I)$ for each $I$. The quantity

$$Y_I - P_Y(X_I)$$

is called the *error of prediction*, i.e., the difference between the predicted GPA and the true GPA for the $I$th applicant. If the prediction errors are close to zero for each applicant, then both the predictor factor, viz., SAT, and the prediction function, $P_Y(x)$, are good. ■

Before we abstract the essential statistical concepts from this example, we give additional examples of cases where prediction is useful.

## EXAMPLE 2.2.2

An organization that evaluates performances of automobiles wants to predict the first-year maintenance cost $Y$ of a new car to be made by company A next year using the number of miles $X$ the car will be driven. If $P_Y(x)$ is the prediction function, an individual who knows he will drive 20,000 miles next year can predict his maintenance cost as $P_Y(20{,}000)$. ■

### EXAMPLE 2.2.3

Suppose a medical research team wants to study how age and weight are related to systolic blood pressure in females between the ages of 25 and 75 currently living in California. The team may want to determine how well the blood pressure $Y$ of a woman can be predicted by using her age $X_1$ and weight $X_2$ as explanatory variables.

*When more than one explanatory or predictor variable is used, the symbols $X_1, X_2$, etc., are used to represent these variables and $Y_I, X_{I1}, X_{I2}$, etc, are used to represent the values of $Y, X_1, X_2$, etc., for the Ith item in the population.*

When there are two predictor variables, we write $P_Y(x_1, x_2)$ for the predicted $Y$ value of an item with $X_1 = x_1$ and $X_2 = x_2$. Note that in this example the research team may not actually be interested in predicting blood pressure, but if a good prediction function $P_Y(x_1, x_2)$ can be found, then an examination of this prediction function may suggest ways of controlling blood pressure, for instance by modifying weight. This information may lead to further investigations, e.g., controlled experiments, in which the team could explicitly study the effect on blood pressure of modifying an individual's weight by diet or exercise. ■

### EXAMPLE 2.2.4

A fertilizer manufacturer wants to predict next year's corn yield $Y$ in bushels per acre, based on the dollars $X$ per acre spent on fertilizer, for farms in the area where the company sells fertilizer. If $P_Y(x)$ represents the prediction function, a customer who plans to spend \$50 per acre for fertilizer next year predicts the yield in bushels per acre to be $P_Y(50)$. ■

### EXAMPLE 2.2.5

An investigator wants to study the pattern of associations among the following variables for U.S.-born individuals who are at least 18 years old now.

$Y$ = height of the individual at age 18
$X_1$ = length of the individual at birth
$X_2$ = mother's height at age 18
$X_3$ = father's height at age 18
$X_4$ = paternal grandmother's height at age 18
$X_5$ = paternal grandfather's height at age 18
$X_6$ = maternal grandmother's height at age 18
$X_7$ = maternal grandfather's height at age 18

The investigator may not actually be interested in predicting what an individual's height will be at age 18, but if a good prediction function is found, then this function may yield information regarding what the predominant determinant of an individual's height is—the heights of his maternal ancestors, the heights of his paternal ancestors, both, or neither. ■

### EXAMPLE 2.2.6

A company that owns a large tree farm wants to determine the volume of each tree on the farm, and if the volume of a tree is large enough, the tree will be cut down and sold. The volume of a tree is difficult to determine, and the tree must be cut down to determine the volume accurately. This procedure is not desirable because if the volume is not large enough, it is more profitable to let the tree grow another year. To determine the volume of trees, the owner wants to use an inexpensive method that does not require cutting and hence destroying the trees. One way to accomplish this is to predict the volume $Y$ of a tree by measuring its diameter $X_1$, say at 4.5 feet from the ground, and its height $X_2$. The diameter and height of a tree can be measured quickly and inexpensively, and the volume can be predicted without destroying the tree. ■

### EXAMPLE 2.2.7

It is known that as the water in a river moves downstream, it carries small rocks (pebbles) along its path and the rocks tend to become smooth and round in shape. The relationship between the roundness $Y$ of the pebbles, called *sphericity*, and the distance $X$ they have been transported may be of interest to a geologist who is trying to determine the source of the rocks. ■

### EXAMPLE 2.2.8

A company that produces a certain chemical makes many batches of the chemical each day, and the number of batches is determined by the production superintendent. The quality control section of the company notices an association between the number of batches $X$ made in a day and the percentage of impurities $Y$ in a day's production. A study of this association may help to predict how many batches can be made in a day without allowing the percentage of impurities to get too large. ■

### EXAMPLE 2.2.9

A chemical engineer knows that temperature $X_1$ and pressure $X_2$ during the production of plastic containers are two factors that determine the strength $Y$ of the final product. He wants to adjust the temperature and pressure to obtain maximum strength, so he studies the relationship between $Y, X_1$, and $X_2$. If $P_Y(x_1, x_2)$ is the predicted strength for temperature $X_1 = x_1$ and pressure $X_2 = x_2$, the engineer may want to determine the values of $x_1, x_2$ that will maximize the predicted strength of the containers, i.e., maximize $P_Y(x_1, x_2)$. ■

### EXAMPLE 2.2.10

It is well known that the relationship between the distance $S$ (in feet) traveled by a freely falling object (in a vacuum) starting from rest and the elapsed time $T$ (in seconds) is well described by the relationship $S = (1/2)gT^2$, where $g$ is a physical constant known as the acceleration due to gravity. The constant $g$ also plays a role in explaining many other calculations in physics. It is possible to experimentally determine the value of $g$ by measuring the values of $S$ corresponding to different

values of $T$ and appropriately analyzing the resulting pairs of measured values of $S$ and $T$. Due to errors in measurements as well as uncontrollable fluctuations in experimental conditions, it is almost always the case that the measured values $S^*$ and $T^*$ of distance traveled and time elapsed will not equal the true values $S$ and $T$, respectively. Consequently, no analysis of the pairs of numbers $(S^*, T^*)$ will yield the exact value of the constant $g$. An experimenter might be interested in methods of analyzing experimental data that would yield an estimate of $g$ that is as close to the true value of $g$ as possible. The immediate objective is not predicting $S$ from a knowledge of $T$, but obtaining a good estimate of the value $g$, the acceleration due to gravity. ■

EXAMPLE 2.2.11

The business manager of a company is preparing a budget for next year, and she must include enough money to purchase a fleet of new cars. She wants to *predict* next year's price of a certain make and model car. She decides that good predictor variables for next year's price $Y$ are the price $X_1$ of a similar car this year and the rate of inflation $X_2$. She needs a good prediction function $P_Y(x_1, x_2)$ that will help her predict next year's price of cars that her company is planning to buy. ■

From these examples it is clear that a study of associations or relationships among factors in a system can help in analyzing, predicting, determining, and even controlling the driving forces that affect the system. *Almost every decision that an individual makes is based on prediction, and many of these predictions can be made by systematically studying associations.* Regression analysis deals with the study of these associations.

### Populations—A Review

The first step in any investigation where we want to predict a variable $Y$ is the identification of appropriate predictor variables $X_1$, $X_2$, etc. Then the *target population* of items (trees, farms, people, etc.) for which the prediction is of interest must be defined. If the target population is available for study, then samples will be selected from this population. If the target population is unavailable for sampling then, typically, we define a *study population* that resembles the target population as closely as possible and from which samples can be obtained, provided that knowledge about the study population will aid the investigator in making predictions in the target population. You should review the material in Section 1.3 concerning target populations and study populations. In particular, keep in mind that valid statistical inferences can be made to the study population only but not to the target population, unless of course the two populations are the same. If the two populations are not the same, it is up to the investigator to *judge* whether or not information about the study population can be used to make decisions about the target population.

Recall that if each item in the population (study population or target population) has one measurement of interest associated with it, that set of numbers will be

called a *one-variable* population or a *univariate* population; if each item has a pair of numbers of interest associated with it, that collection of pairs of numbers will be called a *two-variable* population or a *bivariate* population, etc. In the simplest cases where prediction is useful, the target population and the study population are univariate or bivariate populations, whereas in more complex situations where prediction is required the populations are $k$-variate ($k > 2$) populations.

For an illustration, let us consider Example 2.2.2 where an organization wants to predict the first-year maintenance cost of cars based on the number of miles they will be driven during that year. The target population of items can be the set of automobiles that will be made by manufacturer A next year and driven between 5,000 and 20,000 miles the first year after purchase. The number of automobiles that satisfy this definition will be denoted by $N$, where $N$ is unknown but presumably quite large. Each automobile in the target population will have many numbers associated with it (price of the car, frequency of repairs, miles driven during the first year after purchase, maintenance cost during the first year, etc.), but only two numbers are of interest in the present investigation—the first-year maintenance cost and the number of miles driven. In fact, the organization is interested in predicting the first-year maintenance cost based on the number of miles the car will be driven during the first year after purchase. So first-year maintenance cost is designated as the response variable $Y$ and miles driven as the predictor variable $X$. Clearly the target population is unavailable for study because it is a future population. So the investigator decides to use, as the study population, the population of all cars made by manufacturer A last year that were driven between 5,000 and 20,000 miles during the first year after purchase. A simple random sample of $n$ cars, say $n = 50$, is selected, and the first-year maintenance cost and the number of miles driven are recorded. Based on these data the investigator may make valid statistical inferences about the study population. Information about the study population may be useful for making decisions about the target population. This will require subject matter judgment on the part of the investigator. Note that in this example the target population and the study population are *bivariate populations* and are *different* populations.

Now consider Example 2.2.3 where a medical research team wants to predict blood pressure using age and weight as predictor variables. The target population of items is the set of females between the ages of 25 and 75, currently living in California. Each member of this population has many numbers associated with her (blood pressure, age, weight, height, number of years of education, last year's income, etc.), but for this study, only three numbers are of interest—blood pressure, age, and weight. The objective of this study is to understand how systolic blood pressure is related to age and weight. So blood pressure is designated as the response variable $Y$ and age and weight are the two predictor variables $X_1$ and $X_2$, respectively. Understanding how blood pressure is related to age and weight may suggest a way of lowering blood pressure by modifying weight. Since the target population is available for study, it is also the study population. In this example the population of interest is a *three-variable population* or a *trivariate population.*

### Schematic Representations of Populations

In general a bivariate population of $Y$ and $X$ values is denoted by $\{(Y, X)\}$ and may be schematically exhibited as in Table 2.2.1. In this table the quantities $Y_1, \ldots, Y_N$, represent the values of the response variable $Y$, and $X_1, \ldots, X_N$ represent the values of the predictor variable $X$ for items $1, 2, \ldots, N$, respectively, in the population.

Likewise, a three-variable population consisting of $Y$, $X_1$, and $X_2$ values is denoted by $\{(Y, X_1, X_2)\}$ and is schematically represented as in Table 2.2.2. Recall that $X_{I1}$ is the value of the first predictor variable $X_1$ for item $I$, and $X_{I2}$ is the value of the second predictor variable $X_2$ for that item.

For an illustration we examine the population data in Table D-1 in Appendix D. These data are assumed to have been obtained last year from the sales-and-maintenance records of automobile dealers in Colorado. The total number of cars included in this data set is 1,242. Thus we have a three-variable population consisting of 1,242 cars. Table D-1 consists of four columns of data as in Table 2.2.2.

TABLE 2.2.1

A Schematic Representation of a Bivariate Population with Response Variable $Y$ and Predictor Variable $X$

| Item Number $I$ | Response Variable $Y$ | Predictor Variable (Explanatory Variable) $X$ |
|---|---|---|
| 1 | $Y_1$ | $X_1$ |
| 2 | $Y_2$ | $X_2$ |
| ⋮ | ⋮ | ⋮ |
| $I$ | $Y_I$ | $X_I$ |
| ⋮ | ⋮ | ⋮ |
| $N$ | $Y_N$ | $X_N$ |

TABLE 2.2.2

A Schematic Representation of a Trivariate Population with Response Variable $Y$ and Predictor Variables $X_1$ and $X_2$

| Item Number $I$ | Response Variable $Y$ | Predictor Variable 1 (Explanatory Variable 1) $X_1$ | Predictor Variable 2 (Explanatory Variable 2) $X_2$ |
|---|---|---|---|
| 1 | $Y_1$ | $X_{11}$ | $X_{12}$ |
| 2 | $Y_2$ | $X_{21}$ | $X_{22}$ |
| ⋮ | ⋮ | ⋮ | ⋮ |
| $I$ | $Y_I$ | $X_{I1}$ | $X_{I2}$ |
| ⋮ | ⋮ | ⋮ | ⋮ |
| $N$ | $Y_N$ | $X_{N1}$ | $X_{N2}$ |

The first column contains the car numbers $I$, which range from 1 to 1,242; the second column contains $Y$, the total amount (in dollars) spent on maintenance during the first year after sale. The third column contains $X_1$, the price of the car in dollars when it was purchased. Column 4 contains $X_2$, the number of miles the car was driven the first year after purchase. Thus the three-variable population in Table D-1 is schematically represented in Table 2.2.2. The first three columns of data in Table D-1 represent a bivariate population (column 1 is a column of labels; columns two and three contain $Y$ and $X_1$, respectively) as in Table 2.2.1. The data in Table D-1 are also stored in the file **car.dat** on the data disk.

For populations such as these, it is of interest to study the relationship of $Y$ with $X_1$, $Y$ with $X_2$, and $Y$ with both $X_1$ and $X_2$, to determine if either or both $X_1$, $X_2$ are useful for predicting $Y$.

### Reasons for Prediction—A Summary

There are at least three reasons why prediction is useful.

1 *The true response values $Y$ are very expensive to obtain,* but the values of the predictor variable $X$ (or $X_1, \ldots, X_k$ in the case of multiple predictor variables) are relatively inexpensive to obtain, so we can use the inexpensive $X$ values to predict the expensive $Y$ values. This is especially useful in cases when an item has to be destroyed to measure the value of the response variable $Y$, as is the case in Example 2.2.6 where it is very costly to obtain the volume of a tree but relatively inexpensive to measure diameter and height and predict the volume.

2 *The response values are impossible to measure since they are often future values and thus are not available now.* However, for decision-making purposes investigators want to know the values *before* they become available. This is the situation in Examples 2.2.1, 2.2.2, and 2.2.4. Consider, for instance, Example 2.2.4, where we want to predict the yield $Y$ based on the amount $X$ spent on fertilizer. Of course if the yield $Y$ were known, we would not be interested in predicting it. But in many instances where prediction is needed and used, the true value of the response variable is not known because it is a *future* value that we want to know *now*. If $X$ is available now and $Y$ is not, then we can use the value of the prediction function, $P_Y(x)$, to predict the value of $Y$ *now*.

3 *Prediction is not of immediate interest, but the prediction function is the important quantity.* This is illustrated in Example 2.2.3. Certainly blood pressure can be measured very easily and cheaply, and if an individual wants to know her blood pressure she can measure it directly. In this case the *prediction function* is the important quantity because it may give valuable insight into the relationship between blood pressure and weight for individuals of various ages. For example, if a physician knows the prediction function, he may be able to determine how to reduce the blood pressure of a patient to a desirable level by reducing her weight a certain amount through diet or exercise. Example 2.2.10 provides another illustration. In that example an investigator may not be interested in prediction, yet may be interested in knowing the equation relating $S$ and $T$ so as to extract the value of $g$ from it.

You may be able to think of other uses for prediction.

#### What Is Needed for Prediction?

Two components are needed for prediction:

1 The predictor variables, say $X_1, \ldots, X_k$, and the observed values of these variables.

2 A prediction function or formula, denoted by $P_Y(x_1, \ldots, x_k)$, for predicting the response variable $Y$ using the predictor variables $X_1, \ldots, X_k$.

#### The Predictor Variables (or Factors) Are Selected by the Investigator

The investigator knows which variable is to be predicted, and her knowledge of the subject suggests factors that may be useful as predictors. She may not know which factors are the best predictors, or how they interrelate, and regression analysis can be helpful in providing answers to these questions.

#### Prediction Function

The *best* (prediction) function for predicting $Y$ using $X_1, \ldots, X_k$ can be obtained using regression analysis. In the next section, we define the *regression function* and discuss how it is used in prediction.

## Problems 2.2

**2.2.1** Describe in detail a two-variable population $\{(Y, X)\}$, preferably related to your own field of study, where you want to predict $Y$ using $X$.

**2.2.2** Describe in detail a three-variable population $\{(Y, X_1, X_2)\}$, preferably from your own subject area of interest, where you may want to predict $Y$ using $X_1$, and $X_2$.

**2.2.3** State why prediction would be useful in your population in Problem 2.2.1.

**2.2.4** In Example 2.2.1 suppose that a student has an SAT score of $X = 490$. What is his predicted first-year GPA if the prediction function is $P_Y(x) = 0.324 + 0.00484x$? What is the predicted GPA of a student whose SAT score is 625?

**2.2.5** In Problem 2.2.4, assume that student A scored 50 points more than student B on the SAT. What will be the predicted difference between their first-year GPAs?

**2.2.6** In Problem 2.2.4 if the director of admissions decides to admit only those students whose first-year GPA is predicted to be 3.0 or higher, what is the lowest a student's SAT score can be if he or she is to be admitted?

# 2.3 Regression Analysis

Regression analysis is a commonly used method for obtaining a prediction function for predicting the values of a response variable $Y$ using predictor variables $X_1, \ldots, X_k$. We begin by discussing the concept of *subpopulations*, which plays a very important role in defining the *regression function* of $Y$ on $X_1, \ldots, X_k$. We first consider a two-variable population $\{(Y, X)\}$ and suppose that we want to predict the value of $Y$ based on the value of $X$ for any population item.

### Subpopulations

For each distinct value of $X$ in the population there is a *subpopulation* of $Y$ values. *The subpopulation corresponding to $X = x$ is the set of all $Y$ values of those items in the population with $X = x$.*

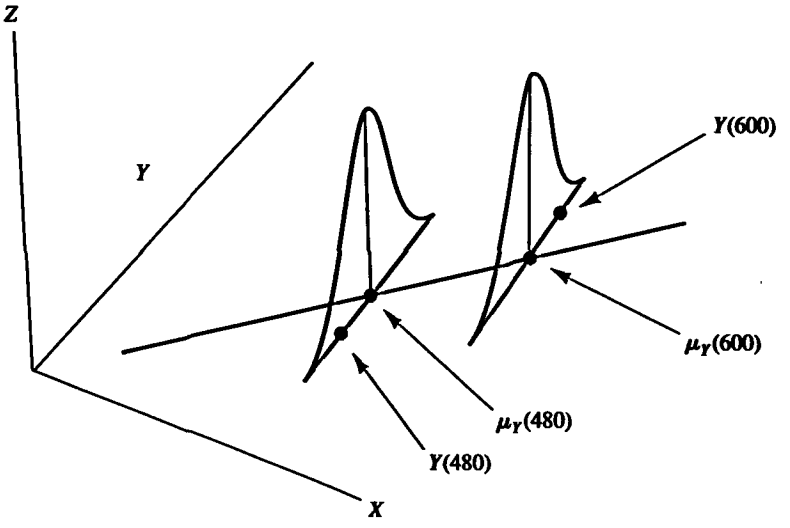
To explain this concept, we use Example 2.2.1 as an illustration. Recall that in that example, the director of admissions wants to predict GPAs using SAT scores. Suppose, for practical purposes, that the director is interested only in the SAT scores ($X$ values) $450, 451, 452, \ldots, 799, 800$. Because the target population is a conceptual population, it is unavailable for study. Consequently she decides to use as the study population the set of applicants who were admitted to the university during the past ten years whose SAT scores $X$ and first-year GPAs $Y$ are available. The individuals in this study population are grouped into subpopulations on the basis of distinct values of $X$. In other words, all individuals whose SAT score was 450 are placed in a group, all individuals whose SAT score was 451 are placed in a different group, etc. The GPA values of the group of (say $N_1$) individuals whose SAT score was 450 form a collection of numbers that is denoted by $\{Y(450)\}$, and this group is called the *subpopulation* of $Y$ values determined by the SAT score $X = 450$, i.e., the subpopulation of $Y$ values, all of which have $X = 450$ as their corresponding $X$ value. The mean of this subpopulation of $Y$ values is denoted by $\mu_Y(450)$, and its standard deviation is denoted by $\sigma_Y(450)$. The GPA values of the group of (say $N_2$) individuals whose SAT score was 451 form a collection of numbers that is denoted by $\{Y(451)\}$, and this group is called the *subpopulation* of $Y$ values with $X = 451$. This subpopulation has mean $\mu_Y(451)$ and standard deviation $\sigma_Y(451)$. More generally, *there is a subpopulation of $Y$ values for each distinct value of $X$. The subpopulation of $Y$ values, all having the same $X$ value, say $X = x$, is denoted by $\{Y(x)\}$. The mean of this subpopulation is denoted by $\mu_Y(x)$, and its standard deviation is denoted by $\sigma_Y(x)$.*

It is clear why the symbol $\mu_Y(x)$ is used for the mean of a subpopulation—it is the *mean* $\mu$ of the *sub*population (*sub*script $Y$) of $Y$ values corresponding to $X = x$. Similarly, the symbol $\sigma_Y(x)$ is used to denote the standard deviation of the *sub*population of $Y$ values corresponding to $X = x$.

Figure 2.3.1 provides a graphical representation of subpopulations for two different values of $X$ = SAT scores. In this figure, $Y(600)$ represents the $Y$ value of a randomly chosen individual from the subpopulation of all individuals with $X = 600$.

Likewise, $Y(480)$ denotes the $Y$ value of a randomly chosen individual from the subpopulation with $X = 480$.

FIGURE 2.3.1




EXAMPLE 2.3.1

To illustrate the concept of subpopulations we consider Example 2.2.2, where the organization wants to predict $Y$, the first-year maintenance cost of cars, by using $X$, the distance the car will be driven, as the predictor variable. Table D-1 in Appendix D gives data that represent a three-variable population consisting of $Y$ = first-year maintenance cost of cars, $X_1$ = sticker price of the cars, and $X_2$ = miles the cars will be driven the first year after they are purchased. In this example we consider only the two-variable population consisting of $Y$ and $X_2$ values (i.e., ignore $X_1$ for the moment). Now if you look at the column of data that represents $X_2$, you will see the number 14,000 several times. If we look at only the pair of numbers $(Y, X_2)$ for which $X_2 = 14{,}000$, we get a subpopulation of $Y$ values, all of which have the same $X_2$ value, namely $X_2 = 14{,}000$. This subpopulation is shown in Table 2.3.1. The mean $Y$ value for this subpopulation is $\mu_Y(14{,}000) = \$621.19$, and the standard deviation is $\sigma_Y(14{,}000) = \$23.18$. You should pick another subpopulation from the population data in Table D-1, say one for which $X_2 = 9{,}000$, and examine it. ■

#### Regression Function

For any given $X$ value, say $X = x$, the *mean* (i.e., average) of the $Y$ values in this subpopulation is denoted by $\mu_Y(x)$, and the standard deviation of these $Y$ values is

denoted by $\sigma_Y(x)$. We are now in a position to define the regression function of $Y$ on $X$.

D E F I N I T I O N

The function $\mu_Y(x)$ is called the **regression function** of $Y$ on $X$ and is the *mean* of the subpopulation of $Y$ values for each distinct value of $X$. ■

In particular, the mean of a subpopulation of $Y$ values, all of which have $X = x$, is equal to $\mu_Y(x)$. Recall that in this book we use the mean as the best *single* number to represent a population of numbers, and since $\mu_Y(x)$ is the mean of the subpopulation of all $Y$ values that have $x$ for their $X$ value, $\mu_Y(x)$ is the best single number to represent this subpopulation; it is often used as a predictor of any $Y$ value in this subpopulation. For many situations, it can be shown that

B O X 2.3.1 The best prediction function of $Y$, using $X$ as the predictor variable, is the regression function $\mu_Y(x)$.

Note that although the actual $Y$ values of the items in the subpopulation with $X = x$ are in general not all the same, the predicted value for any of these items will be the same and equal to $\mu_Y(x)$ because they all have the same $X$ value, namely $x$.

T A B L E 2.3.1

Maintenance Cost ($Y$) and Miles Driven ($X_2$) for the Subpopulation with $X_2 = 14{,}000$ Miles

| Maintenance Cost $Y$ | Miles Driven $X_2$ |
|---|---|
| 656 | 14,000 |
| 633 | 14,000 |
| 637 | 14,000 |
| 612 | 14,000 |
| 624 | 14,000 |
| 620 | 14,000 |
| 605 | 14,000 |
| 607 | 14,000 |
| 654 | 14,000 |
| 620 | 14,000 |
| 622 | 14,000 |
| 645 | 14,000 |
| 567 | 14,000 |
| 596 | 14,000 |
| 639 | 14,000 |
| 602 | 14,000 |

However, if $\sigma_Y(x)$ is small, most of the $Y$ values in this subpopulation will be close to $\mu_Y(x)$, and the probability is high that the $Y$ value to be predicted will be close to the predicted value $\mu_Y(x)$.

The means and standard deviations of subpopulations are of interest in a variety of situations. For instance, consider Example 2.2.2 where the organization wants to predict the first-year maintenance cost of cars. Let $\mu_Y$ denote the *average* first-year maintenance cost of *all* cars in the population of cars that will be made by manufacturer A next year. The average maintenance cost of all cars in the subpopulation of cars that will be driven $X = 15{,}000$ miles next year is $\mu_Y(15{,}000)$. If a woman plans to drive 15,000 miles next year, $\mu_Y(15{,}000)$ will be a better representative (predictor) of her maintenance cost than $\mu_Y$ will be. Consider another example. Suppose a man in the United States who is 32 years old wants to know if he is overweight. He should *not* compare his weight to $\mu_Y$, the average weight of *all* men, but he should compare his weight with the average weight of all men belonging to the subpopulation of U.S. men who are 32 years old. Because the mean weight of this subpopulation is $\mu_Y(32)$ and the standard deviation is $\sigma_Y(32)$, he knows (using Chebyshev's theorem) that at most 11% of all men in this subpopulation have weights outside the interval $[\mu_Y(32) - 3\sigma_Y(32), \mu_Y(32) + 3\sigma_Y(32)]$. By using such information, he may be able to judge whether or not he is overweight for his age.

#### Subpopulations, Prediction, and Regression—A Summary of Concepts

We summarize the concepts and ideas just discussed for the case of a single predictor (explanatory) variable.

1 The two-variable population $\{(Y, X)\}$ is partitioned into subpopulations—a subpopulation of $Y$ values for each distinct value of $X$.

2 The subpopulation of $Y$ values corresponding to any given value of the predictor variable $X$, say $X = x$, has mean $\mu_Y(x)$ and standard deviation $\sigma_Y(x)$.

3 $\mu_Y(x)$ *is called the regression function of* $Y$ *on* $X$, and it is the best single value to represent (predict) any $Y$ value in the subpopulation whose $X$ value is $x$.

4 If $Y(x)$ denotes the $Y$ value of an item that is to be randomly chosen from the subpopulation with $X = x$, then the best predicted value of $Y(x)$ is the mean, $\mu_Y(x)$, of the subpopulation of all items whose $X$ values equal $x$.

5 $\sigma_Y(x)$ is the standard deviation of the subpopulation of $Y$ values whose $X$ value is $x$, and it is used to determine how well $\mu_Y(x)$ represents the entire collection of $Y$ values in the subpopulation whose $X$ values equal $x$.

6 In most, if not all, applications, $\mu_Y(x)$ and $\sigma_Y(x)$ are unknown and must be estimated from sample data.

7 In theoretical books on statistics, the distribution of the subpopulation $\{Y(x)\}$ is called the *conditional distribution* of $Y$ given $X = x$.

### Subpopulations and Regression in the Case of Several Predictor Variables

We now extend the concepts of subpopulations and regression functions to the case where the number of predictor variables is greater than one. When there are $k$ predictor (explanatory) variables, say $X_1, \ldots, X_k$, *each distinct combination of values of* $X_1, \ldots, X_k$ *in the population determines a subpopulation of* $Y$ *values.* The subpopulation of $Y$ values determined by $x_1, \ldots, x_k$ is the collection of $Y$ values in the population for which $X_1 = x_1, \ldots, X_k = x_k$. The mean of the $Y$ values belonging to this subpopulation is denoted by $\mu_Y(x_1, \ldots, x_k)$, and the standard deviation is denoted by $\sigma_Y(x_1, \ldots, x_k)$.

BOX 2.3.2 The function $\mu_Y(x_1, \ldots, x_k)$ is called the **regression function of** $Y$ on $X_1, \ldots, X_k$.

Let $Y(x_1, \ldots, x_k)$ represent the $Y$ value of an item to be randomly chosen from the subpopulation with $X_1 = x_1, \ldots, X_k = x_k$. The best value to use to predict $Y(x_1, \ldots, x_k)$ is $\mu_Y(x_1, \ldots, x_k)$, the mean of the subpopulation of $Y$ values with $X_1 = x_1, \ldots, X_k = x_k$. The standard deviation $\sigma_Y(x_1, \ldots, x_k)$ of this subpopulation is a measure of how well $\mu_Y(x_1, \ldots, x_k)$ represents every $Y$ value in this subpopulation (i.e., how good the prediction function is).

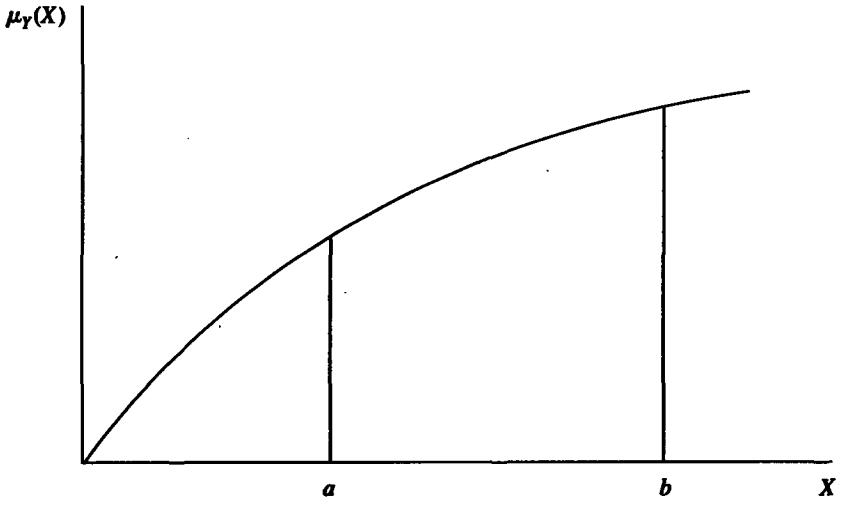
When the number of predictor variables is greater than one, we can consider subpopulations determined by any subset of these variables. Consider Example 2.2.3 where a medical research team wants to predict the blood pressure $Y$ of Ms. Smith, a 45-year-old California woman weighing 182 pounds (i.e., $X_1 = 45$ and $X_2 = 182$ ). There are several subpopulations that the team may consider here. First, there is the entire population of $Y$ values. Then there is a subpopulation of blood pressure values of all women living in California who are 45 years old. There is a subpopulation of blood pressure values of all women living in California who weigh 182 pounds. There is a subpopulation of blood pressures of all women living in California who are 45 years old *and* weigh 182 pounds. In fact, the mean blood pressure, $\mu_Y(45, 182)$, of all women in California who are 45 years old and weigh 182 pounds, is the best predicted value of Ms. Smith's blood pressure (in the absence of any other knowledge about her).

**Note** For simplicity of presentation of the *concepts* underlying regression, we have assumed throughout that $N$ is finite, but in actuality the *theory* of regression is typically derived based on $N$ being infinite. In a real application the population size is usually so large that for all practical purposes the results discussed in this book will be valid when all other assumptions are satisfied.

### Straight Line Regression

In Chapter 3 we give a detailed presentation of regression with a single predictor $X$ where the population regression function of $Y$ on $X$ is of the form $\mu_Y(x) = \beta_0 + \beta_1 x$. This regression function, *whose graph as a function of* $X$ *is a straight line,* is not only one of the simplest regression functions but also a very useful one in real

FIGURE 2.3.2




problems. The true relationship between two variables is often linear, but even when it is not, the straight line regression function $\mu_Y(x) = \beta_0 + \beta_1 x$ is sometimes a good approximation to use in the initial stages of an investigation. The regression function may not be a straight line function over the entire range of $X$ values, but it may be an adequate approximation over a limited range that the investigator wants to examine. Consider Figure 2.3.2. A straight line model may be considered adequate if the investigation includes only $X$ values in the range from $a$ to $b$, but it may not be adequate over the range from 0 to $b$.

**Conversation** To review some of the notions discussed in this section, we present a conversation between an investigator and a statistician.

## Conversation 2.3

**Investigator:** I want to ask you some questions about the notation in your book. Do you have time now?

**Statistician:** Certainly.

**Investigator:** Why do you use the notation $\mu_Y(x) = \beta_0 + \beta_1 x$ for a straight line population regression function when most books I've seen use the notation $y = \beta_0 + \beta_1 x$ or $Y_x = \beta_0 + \beta_1 x$?

**Statistician:** Good question. The reason we use $\mu_Y(x)$ instead of $y$ or $Y_x$ for a straight line population regression function $\mu_Y(x) = \beta_0 + \beta_1 x$ is to help remind you that a regression

function is the mean of the subpopulation of $Y$ values determined by $X$. For example, $\mu_Y(6) = \beta_0 + \beta_1(6)$ is the mean of the subpopulation of $Y$ values for which $X = 6$. In statistics Greek letters are often used to denote population parameters, and the Greek letter $\mu$ is usually used to denote a population mean. Thus the functional notation $\mu_Y(x)$ denotes a mean, $\mu$, and the subscript $Y$ tells us it is the mean of a population or subpopulation of $Y$ values. The value of $X$ points to that subpopulation of $Y$ values for which the mean is being considered. Similarly, the symbol $\sigma_Y(x)$ denotes the standard deviation of the subpopulation of $Y$ values determined by $x$. Furthermore, we also use the symbols $\mu_Y$ and $\sigma_Y$ to denote the mean and standard deviation of the *entire* population of $Y$ values when we ignore $X$.

**Investigator:** I notice that you also use the symbols $\{Y\}$, $Y(x)$, and $\{Y(x)\}$.

**Statistician:** Yes, we use the symbol $\{Y\}$ to represent a one-variable population of $Y$ values and the symbol $\{Y(x)\}$ to represent the *subpopulation* of $Y$ values, all of which have $x$ as their common $X$ value. For instance, $\{Y(15.8)\}$ represents a subpopulation of all $Y$ values that have $X = 15.8$ as their common $X$ value. The symbol $Y(x)$ (without the braces around it) is used to represent the $Y$ value of a randomly chosen item from the subpopulation for which $X = x$. Thus $\{Y(x)\}$ denotes the entire subpopulation of $Y$ values having $X = x$ while $Y(x)$ (without the braces, { }) denotes a single randomly chosen $Y$ value from this subpopulation. You might also recall that $\mu_Y(x)$ denotes the mean, and $\sigma_Y(x)$ denotes the standard deviation of the subpopulation $\{Y(x)\}$.

**Investigator:** Sometimes you use $P_Y(x)$ to represent a prediction function. Why?

**Statistician:** We use the symbol $P_Y(x)$ to denote a general prediction function of $Y$. But for a specific problem, we would naturally want to use the best prediction function, and the best prediction function of $Y$ is the regression function $\mu_Y(x)$, the means of subpopulations.

**Investigator:** *I think I understand what you're saying.*

**Statistician:** I might add that notation can often be quite helpful, so it is worth learning.

**Investigator:** I have one other question. I notice that you don't discuss variance a great deal in your book, yet in the two statistics courses that I took we spent a lot of time studying variance. Why is that?

**Statistician:** Variance has some mathematical properties that makes it very useful in statistical theory, but for applications the standard deviation is much more important. For example, if a population $\{Y\}$ is Gaussian, then we know what proportion $p$ of $Y$ values are in the interval $\mu_Y - c\sigma_Y$ to $\mu_Y + c\sigma_Y$ for any value $c$. And for any population $\{Y\}$, Gaussian or not, one can use Chebyshev's theorem to determine a lower bound for this proportion $p$, for any value of $c$. Thus, the standard deviation is quite useful in applied problems. You also notice that the standard deviation has the same units as the individual population values and as the mean. Of course either the variance or the standard deviation can be computed from the other.

**Investigator:** These are all the questions I have for now. Perhaps, I will come to see you again in a few days.

**Statistician:** Please do so. I am always happy to talk to you.

In many places we ask you to carry out certain tasks that may help explain various concepts that have been discussed. These tasks are usually **word problems** that require you to perform appropriate statistical calculations to answer practical questions. Generally, we pose problems that illustrate some aspect of regression. We also supply answers to these questions, and in the problems at the end of this section we ask similar questions about other similar problems. Here we illustrate the concepts in this section with a task.

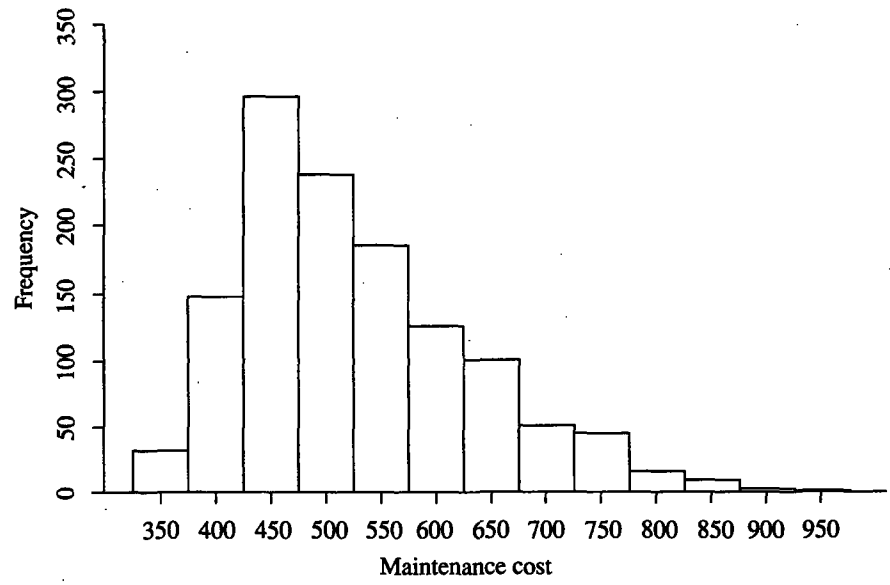
## Task 2.3.1

For illustrative purposes only, suppose that the data in the file **car.dat** on the data disk form a (three-variable) population. These data, which are also given in Table D-1 in Appendix D for convenience, are assumed to have been obtained from last year's sales-and-maintenance records of all automobile dealers who sell a particular make of car in Colorado. The total number of cars included in this data set is 1,242. Thus we have a three-variable population of 1,242 cars. The population data consist of four columns. The first column is the car number $I$, which ranges from 1 to 1,242; the second column contains $Y$, the first-year maintenance price (in dollars); the third column contains $X_1$, the sticker price of the car (in dollars); and the fourth column contains $X_2$, the number of miles the car was driven during the first year after purchase. In this example we want to study the relationship of $Y$ with $X_2$, so we are interested only in columns 1, 2, and 4.

The following set of problems refers to this population of 1,242 cars. We give answers to each question, *and the answers and the authors' explanations are in italics*. You will find that using a suitable computer package (SAS, MINITAB, SPSS, BMDP, S-PLUS, etc.) will make it easier to obtain answers to these problems. In the laboratory manual that accompanies this book we present in detail appropriate computer commands that can be used to obtain the answers.

1 First we examine the population data in Table D-1 in Appendix D in detail.
   - a To get an idea of how the $Y$ values in the population are distributed, construct a *frequency histogram* for the maintenance costs $Y$ of all 1,242 cars in the population.
   - b Compute the mean and the standard deviation of $Y$.
   - a *We construct a frequency histogram for $Y$ = first-year maintenance cost, as shown in Figure 2.3.3. Note that the distribution of $Y$ is not symmetric.*

FIGURE 2.3.3




**b** *The mean and standard deviation of the population of Y values are*

$$\mu_Y = \$526.14 \text{ and } \sigma_Y = \$105.97$$

*respectively. Thus the average first-year maintenance cost of all 1,242 cars in the population is $526.14.*

**2** Suppose a car is randomly chosen from the preceding population of 1,242 cars.

**a** What is the *actual* first-year maintenance cost for this car?

**b** *Predict* the first-year maintenance cost of this randomly chosen car.

**c** Suppose you plan to purchase a car next year. Predict, if possible, the first-year maintenance cost for *your* car.

**a** *There is no way we can determine exactly the actual first-year maintenance cost of the randomly chosen car unless we know which car was chosen. For instance, if car number 354 is chosen, then we know that the first-year maintenance cost for this car was $483.00. You can check this by examining Table D-1.*

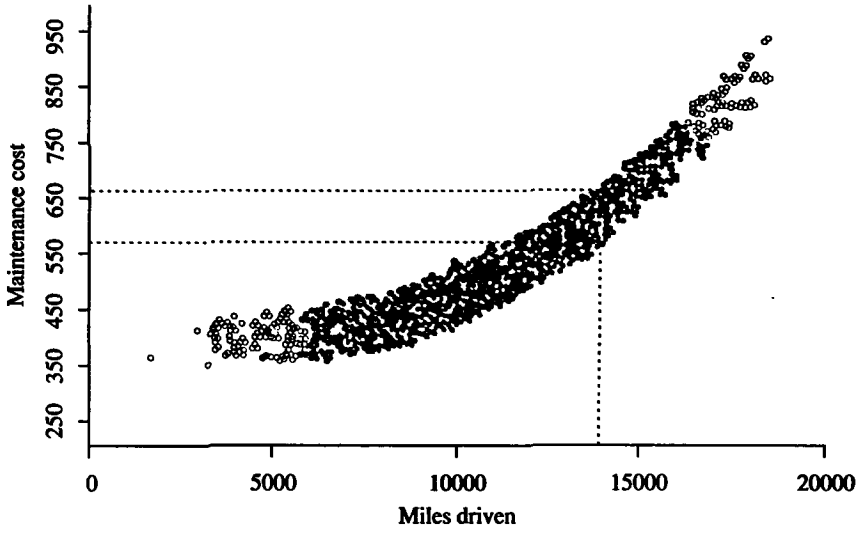
**b** *Because we know that the average first-year maintenance cost of all 1,242 cars is $526.14, our best prediction for the first-year maintenance cost of a randomly chosen car, in the absence of any other relevant information, is $526.14, the mean first-year maintenance cost of all cars in the entire population.*

**c** *If we can regard the car you will buy as a randomly chosen car from a population similar to the preceding population of 1,242 cars, then the best predicted value for the first-year maintenance cost of your car is \$526.14 as in (b). An appropriate target population here might be the set of all cars, of the same make as the 1,242 cars in the population given in Table D-1, that will be made next year. However, this target population is not available so we might choose to use the population given in Table D-1 as the study population. The value \$526.14 is really the best predicted value for the first-year maintenance cost of any randomly chosen car from the study population. Whether or not this is a good predicted value for a randomly chosen car from the target population depends on how closely the study population resembles the target population.*

**3** We want to determine whether there is a relationship between $Y$, the first-year maintenance cost of cars, and $X_2$, the miles the cars will be driven. Do you think that the first-year maintenance costs of cars in this population are related to the number of miles the cars are driven?

*One way to examine the relationship of Y = first-year maintenance cost and $X_2$ = number of miles the car is driven during the first year after purchase is to plot the values of Y against the values of $X_2$; so we plot the values of Y on the vertical axis and the values of $X_2$ on the horizontal axis and study the resulting scatter plot. This plot is given in Figure 2.3.4. (The dashed lines are not part of the computer output.) The scatter plot clearly indicates that the first-year maintenance cost of cars is related to the number of miles they are driven during the first year after purchase because it appears that as the number of miles driven increases, the first-year maintenance cost also tends to increase.*

FIGURE 2.3.4



**4** In problem 2 (c), suppose you are told that the car you will buy next year will be driven 14,000 miles during the first year after purchase.

**a** Predict what the first-year maintenance cost will be for your car using the scatter plot in Figure 2.3.4.

**b** Examine the subpopulation of all cars that are driven 14,000 miles during the first year. How will this information help you determine the first-year maintenance cost of your car?

**a** *First, suppose that the car you will buy next year can be regarded as a randomly chosen car from a population very similar to the one in Table D-1. In this case the population in Table D-1 serves as the study population. In the scatter plot in Figure 2.3.4 we have drawn a vertical line at miles driven = 14,000 to see where it intersects the plotted points. The Y values of these points of intersection range between $560 and $680. So we see that, in the study population, the first-year maintenance cost for a car driven 14,000 miles during its first year should be somewhere between $560 and $680. Based on this scatter plot you might use the middle value* $(560 + 680)/2 = 620$ *as the predicted value for the first-year maintenance cost of the car you will buy next year.*

**b** *Because you know you will drive your car 14,000 miles next year, you are not interested in the entire study population of 1,242 cars but only in the subpopulation of cars that were driven 14,000 miles. There are 16 of these cars, and their first-year maintenance costs are as follows:*

| Maintenance Cost $Y$ | Miles Driven $X$ |
|---|---|
| 656 | 14000 |
| 633 | 14000 |
| 637 | 14000 |
| 612 | 14000 |
| 624 | 14000 |
| 620 | 14000 |
| 605 | 14000 |
| 607 | 14000 |
| 654 | 14000 |
| 620 | 14000 |
| 622 | 14000 |
| 645 | 14000 |
| 567 | 14000 |
| 596 | 14000 |
| 639 | 14000 |
| 602 | 14000 |

*The mean of this subpopulation is $\mu_Y(14{,}000) = \$621.19$, and the standard deviation is $\sigma_Y(14{,}000) = \$23.18$. Because the standard deviation of the entire population of Y values is $\sigma_Y = \$105.97$, which is much larger than $\sigma_Y(14{,}000)$, it is clear that $\mu_Y(14{,}000) = \$621.19$, the mean of the subpopulation, is much better for predicting the first-year maintenance cost of a randomly chosen car from this subpopulation than $\mu_Y = \$526.14$, the mean of the entire population. Hence, $\mu_Y(14{,}000) = \$621.19$ is a better value to use to predict the maintenance cost of the car you will buy next year than $\mu_Y = \$526.14$ is (assuming, as before, that the car you will buy belongs to a population of cars which is similar to the population of cars for which the data are given in Table D-1).*

**5** In problem 4 suppose you know that you will drive your car less than 20,000 miles during the first year, although you are not sure exactly how many miles. What is your predicted first-year maintenance cost?

*If all you know is that you will drive your car less than 20,000 miles the first year, then your prediction has to be based on the data you have for all 1,242 cars in the study population, and you cannot just focus attention on any particular group of cars, such as cars that were driven 14,000 miles during their first year. The average first-year maintenance cost of all the cars in the study population is $\mu_Y = \$526.14$. You use this value to predict the first-year maintenance cost of your car if you don't know how many miles you will drive it.*

**6** In problem 5 suppose you know that you will drive the car 50,000 miles during its first year. Predict the first-year maintenance cost of this car.

*Because $X_2 = 50{,}000$ is well outside the range of values of $X_2$ in the population data at hand, you are uncomfortable making any sort of prediction, and even if you did make a prediction you will have practically no confidence in it for it to be of any real use.*

Next we discuss methods for collecting data because carefully planned data collection procedures are an essential part of a successful statistical study. Ideally the collection of data involves sampling from well-defined populations.

## Sampling Methods in Regression

In practice, inferences about population parameters are based on the information provided by samples. It is therefore very important to ensure that available resources are used efficiently and that all relevant information is gathered. Ideally the collection of data involves random sampling of well-defined populations. In this section we discuss two commonly used sampling methods, *simple random sampling* and *sampling with preselected X values*. In what follows, $Y$ refers to the response variable and $X_1, \ldots, X_k$ refer to predictor variables.

### Simple Random Sampling

Sample data are obtained by selecting a simple random sample of $n$ items from the entire population of $N$ items and recording the values for the response variable $Y$ and the predictor variables $X_1, \ldots, X_k$, for each item in the sample. Refer to Section 1.6.

### Random Sampling with Preselected $X$ values

Specific values of the predictor variables $X_1, \ldots, X_k$ are preselected by the investigator, and each of these preselected set of values determines a subpopulation of $Y$ values. A simple random sample of one or more $Y$ values is selected from each of these subpopulations. The number of observations to be sampled from each subpopulation is also predetermined by the investigator.

**Note** We assume that the reader is familiar with the concepts of *sampling with replacement* and *sampling without replacement*. In this book, unless otherwise specified, sampling is always assumed to be **without replacement**. However, *if the population size N is very large relative to the sample size n, sampling with and without replacement are equivalent for all practical purposes.*

Whether we collect data using simple random sampling or sampling with preselected $X$ values depends on the objectives of the particular investigation and the availability of the items to be sampled. If simple random sampling is used, a random sample from the population of all items is required. If data are obtained by sampling with preselected $X$ values, we must identify the values of $X_1, \ldots, X_k$ for each item in the population, and we must sample $Y$ values only from those subpopulations corresponding to the values of $X_1, \ldots, X_k$ specified by the investigator. The two methods are illustrated using some of the examples in Section 2.2.

## EXAMPLE 2.3.2

In Example 2.2.7 where a geologist wants to sample pebbles from a river, suppose he is interested in a specified portion of the river that is 150 miles in length. Conceptually, two numbers are associated with each pebble in this portion of the river, viz., $Y_I$ = sphericity of the $I$th pebble and $X_I$ = the distance of the $I$th pebble from a given reference point. So the two-variable population consists of $N$ pairs of numbers $(Y_I, X_I)$, one pair for each of the $N$ pebbles.

Suppose the objective is to understand the relationship between sphericity of pebbles and their distances from a reference point. The investigator is thus interested in the regression function of $Y$ on $X$. If data are obtained from this population by simple random sampling, then the sample of $n$ pebbles must be selected at random from the entire population of pebbles. Note that if this method of sampling is used, there is a chance, albeit small, that the entire sample may be selected from a small portion of the river, say a 5- or 10-mile segment. For this problem the investigator is more likely to obtain data by sampling with preselected $X$ values so that he obtains a sample of pebbles along the full length of the 150-mile portion of the river. He may want to select samples of pebbles every 5 miles; i.e., $X = 0, 5, 10, \ldots, 150$. Thus sampling with preselected $X$ values ensures that he will examine the entire length

of the river. Moreover this method is also more convenient, so in this example it is undoubtedly better to preselect the $X$ values and sample the $Y$ values at random from the subpopulations determined by these chosen $X$ values. ■

EXAMPLE 2.3.3

Consider Example 2.2.2 where the manager wants to study the relationship between the number of miles a car is driven and its first-year maintenance cost. If she uses simple random sampling in this investigation, a random sample of automobiles of the specified make and year is required, and this may be quite easy to obtain. On the other hand, if she uses preselected $X$ values, she could make certain that the $X$ values *cover* the range of miles driven that is of interest in this study, but this would require the identification of the number of miles that each car in the entire population is driven. This would be an expensive procedure to say the least. ■

EXAMPLE 2.3.4

Consider Example 2.2.6 where the diameter $X_1$ and height $X_2$ of trees are used to predict volume $Y$. If sampling with preselected values of $X_1, X_2$ is used, then an investigator must randomly select $Y$ values (volume) for preselected values of $X_1$ (diameter) and $X_2$ (height). This requires the identification of all trees that have the specific diameter and height combinations chosen by the investigator. Suppose that $X_1 = 2$ and $X_2 = 80$ is one combination of $X_1, X_2$ values chosen by the investigator. Then we must identify all trees that are 2 feet in diameter and 80 feet tall and randomly select one or more trees from this subpopulation of trees. This would be repeated for every combination of values of $X_1$ and $X_2$ preselected by the investigator, and it would require the identification and the measurement of the diameter and height of every tree on the farm, an almost impossible task. On the other hand, if simple random sampling is used, a random sample of $n$ trees is selected from the population of all trees. This can be accomplished by giving each tree a number according to its location on a grid and randomly selecting $n$ trees using a set of random numbers generated on a computer. This sampling procedure would be less expensive. ■

**Remark** Often data are collected under controlled laboratory conditions. If this is the case, then usually only sampling with preselected $X$ values is meaningful. For example, consider the following experiment. The effect of temperature on the growth of soybean seedlings is being studied in a laboratory under controlled conditions. A batch of seedlings is available, and each seedling is to be subjected to a different temperature in the range from 20° C to 35° C. At the end of a week, the growth (in millimeters) of the seedlings is to be observed. In this example, for every value of temperature in the range from 20° C to 35° C, conceptually, there is a subpopulation of heights of soybean seedlings. The experimenter chooses the subpopulations she wishes to sample by choosing the temperature values to be used in the experiment. It would be sensible for her to choose a set of temperature values to cover the range 20° C to 35° C. Thus sampling with preselected $X$ values is a natural choice in this case.

Note that the subpopulations in this remark are all conceptual subpopulations; i.e., they do not currently exist. However, the experimenter can observe a part of each of these subpopulations, viz., those values that are obtained during the experiment. The sample observations are not random samples in a strict sense, but for practical applications, we usually regard the observed values from an experiment as if they are random samples from the specified subpopulations and apply the inference procedures discussed in this book. The results are generally satisfactory.

#### Summary

Two methods of sampling have been discussed for obtaining data for regression studies. The method to be used in any given situation depends on the problem, the circumstances, and the expenses involved. Generally speaking, the simple random sampling method is less expensive and easier to use in *observational studies*, and sampling with preselected $X$ values is the natural method when data are obtained from controlled experiments. When simple random sampling is used, no control is exerted over the values of the predictor variables in the sample, and consequently there is the undesirable possibility that the values of the predictor variables in the resulting sample will be bunched together. If data are obtained by sampling with preselected $X$ values, investigators can preselect the values of the predictor variables to cover the range that they desire to investigate. It is safe to say that, *in most instances where the investigator is interested in estimating the regression function, a judicious sample obtained by sampling with preselected X values will yield better estimates than one obtained using simple random sampling.* However, investigators may decide to use simple random sampling rather than sampling with preselected $X$ values based on the relative costs associated with the two sampling procedures. They may also decide to use the simple random sampling method if the objectives of the study involve more than just the estimation of the regression function. (We cannot obtain valid estimates of certain parameters if data are obtained by sampling with preselected $X$ values; this is discussed further in Chapter 3.)

## Linear and Nonlinear Regression

As stated earlier, we seldom know the true regression function in an applied problem, but we can often postulate a class of functions such that one of the functions in this class will serve as an approximation to the true regression function and is accurate enough for the problem at hand. The simplest classes of functions that are useful in many problems are straight line functions, quadratic functions, etc. This means that we can write an equation for the regression function under study, but it will involve some unknown constants (called parameters). As an example, if an investigator knows that the regression function under study is a straight line (for all practical purposes), but does not know the slope or the intercept of this straight line, then he/she could write down the regression function as $\mu_Y(x) = \beta_0 + \beta_1 x$, where $\beta_0$ and $\beta_1$ are unknown parameters to be determined or estimated. In this

case the regression function is a *linear function of the unknown parameters.* In general, **linear regression** means the regression function is simultaneously linear in the unknown parameters $\beta_i$, and **nonlinear regression** means the regression function is not simultaneously linear in the unknown parameters $\beta_i$. (Refer to Section 1.7 for a review of the definition of linear functions.)

The theory is much better developed for linear regression than for nonlinear regression. Consequently, most of this book is concerned with linear regression; nonlinear regression is discussed only briefly in Chapter 9.

Some examples of regression functions that are linear are listed in (2.3.1), whereas examples of nonlinear regression functions are listed in (2.3.2). The predictor variables are $X_1$, $X_2$, and $X_3$, and the response variable is $Y$; $\beta_0, \beta_1, \ldots, \beta_5$, are unknown parameters.

$$\left.\begin{aligned}
\mu_Y(x) &= \beta_0 \\
\mu_Y(x) &= \beta_0 + \beta_1 x \\
\mu_Y(x) &= \beta_0 + \beta_1 x + \beta_2 x^2 \\
\mu_Y(x_1, x_2, x_3) &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 \\
\mu_Y(x_1) &= \beta_0 + \beta_1 x_1^2 + \beta_2 x_1^{3/2} + \beta_3 / \ln|x_1| \\
\mu_Y(x_1, x_2) &= \beta_0 + \beta_1 e^{x_1} + \beta_2 x_2 + \beta_3 e^{x_1 x_2} \\
\mu_Y(x_1, x_2, x_3) &= \beta_0 + \beta_1 e^{-2x_1} + \beta_2 \sin(x_1 x_2) + \beta_3 x_1 \ln(x_2^2) \tan(x_3) \\
\mu_Y(x_1, x_2, x_3) &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \beta_4 x_1^2 + \beta_5 x_1 x_3^2
\end{aligned}\right\} \quad (2.3.1)$$

$$\left.\begin{aligned}
\mu_Y(x_1) &= \beta_1 e^{\beta_2 x_1} \\
\mu_Y(x_1) &= \beta_0 + \beta_1 e^{\beta_2 x_1} \\
\mu_Y(x_1, x_2) &= \beta_0 + \beta_1 e^{\beta_2 x_1} + \beta_3 e^{\beta_4 x_2} \\
\mu_Y(x_1, x_2, x_3) &= \beta_0 x_1^{\beta_1} x_2^{\beta_2} x_3^{\beta_3} \\
\mu_Y(x_1, x_2) &= \beta_1 x_1 / (\beta_2 e^{\beta_3 x_2})
\end{aligned}\right\} \quad (2.3.2)$$

## 2.4 Exercises

**2.4.1** For a two-variable population $\{(Y, X)\}$, define the regression function of $Y$ on $X$.

**2.4.2** What is the relationship of regression to prediction?

**2.4.3** Given a two-variable population of values $\{(Y, X)\}$, explain how you would obtain the regression function of $Y$ on $X$.

There are 151 distinct $X_2$ values in the population of cars (see Table D-1 in Appendix D) discussed in Task 2.3.1. Hence, the two-variable population of $\{(Y, X_2)\}$ values consists of 151 subpopulations. For each of these 151 subpopulations (i.e., for each distinct value of $X_2$), we have listed in Table D-2 in Appendix D the subpopulation number (subpop) in the first column, the $X_2$ value in the second column, the number of $Y$ values (ycount) in the third column, the mean $\mu_Y(x)$ (ymean) in the fourth column, and the standard deviation $\sigma_Y(x)$ (ystdevn) of the $Y$ values in the subpopulation in the fifth column. These data are also stored in the file **car2.dat** on the

data disk. Some of the Exercises 2.4.4 through 2.4.10 refer to these subpopulations. The standard deviations are calculated using (1.4.3).

**2.4.4** How many items in the population in Table D-1 have $X_2 = 14{,}300$ (i.e., how many cars were driven 14,300 miles the first year)? How many cars were driven 7,100 miles the first year (i.e., how many items have $X_2 = 7{,}100$)?

**2.4.5** Does the population in Table D-1 have any $X_2$ values equal to 9,200 (i.e., are there any data for cars that were driven 9,200 miles)?

**2.4.6** In the population in Table D-1 in Appendix D, what is the mean first-year maintenance cost of all cars that were driven 8,700 miles the first year; (i.e., what is the value of $\mu_Y(8{,}700)$)? What is the value of $\sigma_Y(8{,}700)$?

**2.4.7** The mean first-year maintenance cost of *all* cars in the entire population in Table D-1 is denoted by $\mu_Y$, and it is equal to \$526.14. If you plan to purchase a car (one that is similar to the cars for which the data appear in Table D-1) and drive it 5,900 miles next year, do you think that \$526.14 is a good predictor of your first-year maintenance cost? If you find a better value to predict your first-year maintenance cost, what is it?

**2.4.8** For the subpopulation of $Y$ values with $X_2 = 10{,}000$, what is the value of the mean? What is the value of the standard deviation?

**2.4.9** From Exercise 2.4.6, you know the values of $\mu_Y(8{,}700)$ and $\sigma_Y(8{,}700)$. Use Chebyshev's theorem to find an upper bound for the probability that the first-year maintenance cost of a car you plan to purchase will be more than \$700.00 if you plan to drive it 8,700 miles next year.

**2.4.10** Plot the mean $\mu_Y(x)$ against $x$ for

$$x = 7800, 7900, 8000, 8100, 8200, 8300, 8400$$

Do you think that a linear prediction function is adequate if the number of miles driven is between 7,800 and 8,400?

**2.4.11** Explain the two sampling methods discussed in this chapter. Describe two studies from your field, one where simple random sampling would be preferred and the other where sampling with preselected $X$ values would be preferred. In each, explain why the particular method would be preferred.

**2.4.12** Which of the following regression functions are (simultaneously) linear in the unknown parameters (the symbols $\beta_0, \beta_1, \beta_2, \beta_3, \gamma_0, \gamma_1, \gamma_2, \gamma_3$ refer to unknown parameters)?

**a** $\mu_Y(x) = \beta_0 + \beta_1 x^4$.

**b** $\mu_Y(x_1, x_2) = \gamma_0 + \gamma_1 x_1 + \gamma_2 x_2 + \gamma_3 x_1 x_2$.

**c** $\mu_Y(x) = \beta_0 + \beta_1 x^{\beta_2}$.

**d** $\mu_Y(x_1, x_2) = \gamma_1 \sqrt{\gamma_2 x_1 + \gamma_3 x_2}$.

**e** $\mu_Y(x) = \beta_0 + \beta_1 x^{1/2} + \beta_2/x + \beta_3 e^{-2x}$.

**f** $\mu_Y(x) = \beta_0 + \sin(\beta_1 x)$.

**g** $\mu_Y(x_1, x_2, x_3) = \beta_0 x_1^{\beta_1} x_2^{\beta_2} x_3^{\beta_3}$.