THE CORRELATION MODEL

In the classic regression model, which has been the underlying model in our discussion up to this point, only Y, which has been called the dependent variable, is required to be random. The variable X is deﬁned as a ﬁxed (nonrandom or mathematical) variable and is referred to as the independent variable. Recall, also, that under this model observations are frequently obtained by preselecting values of X and determining corresponding values of Y.

When both Y and X are random variables, we have what is called the correlation model. Typically, under the correlation model, sample observations are obtained by selecting a random sample of the units of association (which may be persons, places, animals, points in time, or any other element on which the two measurements are taken) and taking on each a measurement of X and a measurement of Y. In this procedure, values of X are not preselected but occur at random, depending on the unit of association selected in the sample.

Although correlation analysis cannot be carried out meaningfully under the classic regression model, regression analysis can be carried out under the correlation model. Correlation involving two variables implies a co-relationship between variables that puts them on an equal footing and does not distinguish between them by referring to one as the dependent and the other as the independent variable. In fact, in the basic computational procedures, which are the same as for the regression model, we N may fit a straight line to the data either by minimizing g1y - y i 2 2 or by minimizing Ni g1x i - x 2 2 . In other words, we may do a regression of X i on Y as well as a regression of Y on X. The fitted line in the two cases in general will be different, and a logical question arises as to which line to fit.

If the objective is solely to obtain a measure of the strength of the relationship between the two variables, it does not matter which line is ﬁtted, since the measure usually computed will be the same in either case. If, however, it is desired to use the equation describing the relationship between the two variables for the purposes discussed in the preceding sections, it does matter which line is ﬁtted. The variable for which we wish to estimate means or to make predictions should be treated as the dependent variable; that is, this variable should be regressed on the other variable.

**The Bivariate Normal Distribution**

Under the correlation model, X and Y are assumed to vary together in what is called a joint distribution. If this joint distribution is a normal distribution, it is referred to as a bivariate normal distribution. Inferences regarding this population may be made based on the results of samples properly drawn from it. If, on the other hand, the form of the joint distribution is known to be nonnormal, or if the form is unknown and there is no justiﬁcation for assuming normality, inferential procedures are invalid, although descriptive measures may be computed.

**Correlation Assumptions**

The following assumptions must hold for inferences about the population to be valid when sampling is from a bivariate distribution.

1. For each value of X there is a normally distributed subpopulation of Y values.

2. For each value of Y there is a normally distributed subpopulation of X values.

3. The joint distribution of X and Y is a normal distribution called the bivariate normal distribution.

4. The subpopulations of Y values all have the same variance.

5. The subpopulations of X values all have the same variance.

The bivariate normal distribution is represented graphically in Figure 9.6.1. In this illustration we see that if we slice the mound parallel to Y at some value of X, the cutaway reveals the corresponding normal distribution of Y. Similarly, a slice through the mound parallel to X at some value of Y reveals the corresponding normally distributed subpopulation of X.

**9.7 THE CORRELATION COEFFICIENT**

The bivariate normal distribution discussed in Section 9.6 has ﬁve parameters, s x , sy , m x , my , and r. The ﬁrst four are, respectively, the standard deviations and means associated with the individual distributions. The other parameter, r, is called the population

correlation coefﬁcient and measures the strength of the linear relationship between X and Y.

The population correlation coefﬁcient is the positive or negative square root of r2 , the population coefﬁcient of determination previously discussed, and since the coefﬁcient of determination takes on values between 0 and 1 inclusive, r may assume any value between -1 and +1. If r = 1 there is a perfect direct linear correlation between the two variables, while r = -1 indicates perfect inverse linear correlation. If r = 0 the two variables are not linearly correlated. The sign of r will always be the same as the sign of b1 , the slope of the population regression line for X and Y.

The sample correlation coefﬁcient, r, describes the linear relationship between the sample observations on two variables in the same way that r describes the relationship in a population. The sample correlation coefﬁcient is the square root of the sample coefﬁcient of determination that was deﬁned earlier.

Figures 9.4.5(d) and 9.4.5(c), respectively, show typical scatter diagrams where r : 0 1r 2 : 02 and r = +1 1r 2 = 12. Figure 9.7.1 shows a typical scatter diagram where r = -1.

We are usually interested in knowing if we may conclude that r Z 0, that is, that X and Y are linearly correlated. Since r is usually unknown, we draw a random sample from the population of interest, compute r, the estimate of r, and test H 0 : r = 0 against the alternative r Z 0. The procedure will be illustrated in the following example.