CHAPTER 14

Correlation Theory

CORRELATION AND REGRESSION

In Chapter 13 we considered the problem of *regression*, or *estimation*, of one variable (the dependent variable) from one or more related variables (the independent variables). In this chapter we consider the closely related problem of *correlation*, or the degree of relationship between variables, which seeks to determine *how well* a linear or other equation describes or explains the relationship between variables.

If all values of the variables satisfy an equation exactly, we say that the variables are *perfectly correlated* or that there is *perfect correlation* between them. Thus the circumferences C and radii r of all circles are perfectly correlated since $C = 2\pi r$. If two dice are tossed simultaneously 100 times, there is no relationship between corresponding points on each die (unless the dice are loaded); that is, they are *uncorrelated*. Such variables as the height and weight of individuals would show *some* correlation.

When only two variables are involved, we speak of *simple correlation* and *simple regression*. When more than two variables are involved, we speak of *multiple correlation* and *multiple regression*. This chapter considers only simple correlation. Multiple correlation and regression are considered in Chapter 15.

LINEAR CORRELATION

If X and Y denote the two variables under consideration, a *scatter diagram* shows the location of points (X, Y) on a rectangular coordinate system. If all points in this scatter diagram seem to lie near a line, as in Figs. 14-1(a) and 14-1(b), the correlation is called *linear*. In such cases, as we have seen in Chapter 13, a linear equation is appropriate for purposes of regression (or estimation).

If Y tends to increase as X increases, as in Fig. 14-1(a), the correlation is called *positive*, or *direct*, *correlation*. If Y tends to decrease as X increases, as in Fig. 14-1(b), the correlation is called *negative*, or *inverse*, *correlation*.

If all points seem to lie near some curve, the correlation is called *nonlinear*, and a nonlinear equation is appropriate for regression, as we have seen in Chapter 13. It is clear that nonlinear correlation can be sometimes positive and sometimes negative.

If there is no relationship indicated between the variables, as in Fig. 14-1(c), we say that there is *no correlation* between them (i.e., they are *uncorrelated*).

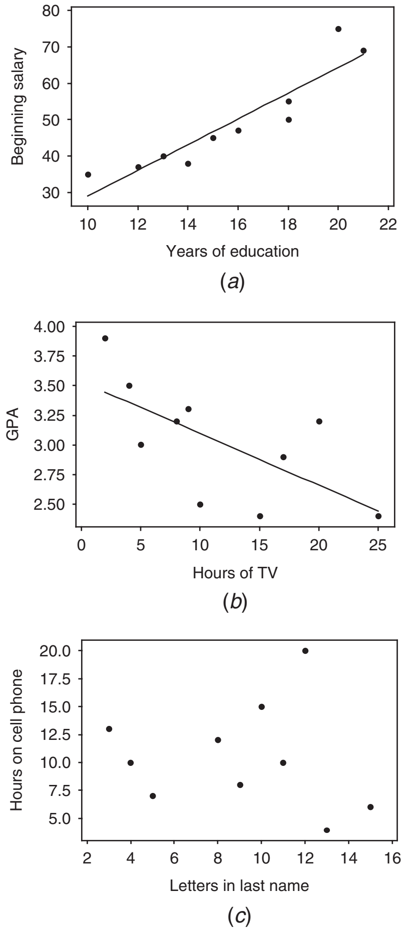

Fig. 14-1 Examples of positive correlation, negative correlation and no correlation. (*a*) Beginning salary and years of formal education are positively correlated; (*b*) Grade point average (GPA) and hours spent watching TV are negatively correlated; (*c*) There is no correlation between hours on a cell phone and letters in last name.

MEASURES OF CORRELATION

We can determine in a *qualitative* manner how well a given line or curve describes the relationship between variables by direct observation of the scatter diagram itself. For example, it is seen that a straight line is far more helpful in describing the relation between X and Y for the data of Fig. 14-1(*a*) than for the data of Fig. 14-1(*b*) because of the fact that there is less scattering about the line of Fig. 14-1(*a*).

If we are to deal with the problem of scattering of sample data about lines or curves in a *quantitative* manner, it will be necessary for us to devise *measures of correlation*

THE LEAST-SQUARES REGRESSION LINES

We first consider the problem of how well a straight line explains the relationship between two variables. To do this, we shall need the equations for the least-squares regression lines obtained in Chapter 13. As we have seen, the least-squares regression line of Y on X is

$$Y = a_0 + a_1 X \tag{1}$$

where a_0 and a_1 are obtained from the normal equations

$$\begin{aligned}\sum Y &= a_0 N \quad + a_1 \sum X \\ \sum XY &= a_0 \sum X + a_1 \sum X^2\end{aligned} \tag{2}$$

which yield

$$\begin{aligned}a_0 &= \frac{(\sum Y)(\sum X^2) - (\sum X)(\sum XY)}{N \sum X^2 - (\sum X)^2} \\ a_1 &= \frac{N \sum XY - (\sum X)(\sum Y)}{N \sum X^2 - (\sum X)^2}\end{aligned} \tag{3}$$

Similarly, the regression line of X on Y is given by

$$X = b_0 + b_1 Y \tag{4}$$

where b_0 and b_1 are obtained from the normal equations

$$\begin{aligned}\sum X &= b_0 N \quad + b_1 \sum Y \\ \sum XY &= b_0 \sum X + b_1 \sum Y^2\end{aligned} \tag{5}$$

which yield

$$\begin{aligned}b_0 &= \frac{(\sum X)(\sum Y^2) - (\sum Y)(\sum XY)}{N \sum Y^2 - (\sum Y)^2} \\ b_1 &= \frac{N \sum XY - (\sum X)(\sum Y)}{N \sum Y^2 - (\sum Y)^2}\end{aligned} \tag{6}$$

Equations (*1*) and (*4*) can also be written, respectively, as

$$y = \left(\frac{\sum xy}{\sum x^2}\right)x \qquad \text{and} \qquad x = \left(\frac{\sum xy}{\sum y^2}\right)y \tag{7}$$

where $x = X - \bar{X}$ and $y = Y - \bar{Y}$.

The regression equations are identical if and only if all points of the scatter diagram lie on a line. In such case there is *perfect linear correlation* between X and Y.

STANDARD ERROR OF ESTIMATE

If we let Y_{est} represent the value of Y for given values of X as estimated from equation (*1*), a measure of the scatter about the regression line of Y on X is supplied by the quantity

$$s_{Y.X} = \sqrt{\frac{\sum (Y - Y_{\text{est}})^2}{N}} \tag{8}$$

which is called the *standard error of estimate of Y on X*.

If the regression line (*4*) is used, an analogous standard error of estimate of X on Y is defined by

$$s_{X.Y} = \sqrt{\frac{\sum (X - X_{\text{est}})^2}{N}} \tag{9}$$

In general, $s_{Y.X} \neq s_{X.Y}$.

Equation (*8*) can be written

$$s_{Y.X}^2 = \frac{\sum Y^2 - a_0 \sum Y - a_1 \sum XY}{N} \tag{10}$$

which may be more suitable for computation (see Problem 14.3). A similar expression exists for equation (*9*).

The standard error of estimate has properties analogous to those of the standard deviation. For example, if we construct lines parallel to the regression line of Y on X at respective vertical distances $s_{Y.X}$, $2s_{Y.X}$, and $3s_{Y.X}$ from it, we should find, if N is large enough, that there would be included between these lines about 68%, 95%, and 99.7% of the sample points.

Just as a modified standard deviation given by

$$\hat{s} = \sqrt{\frac{N}{N-1}}\, s$$

was found useful for small samples, so a modified standard error of estimate given by

$$\hat{s}_{Y.X} = \sqrt{\frac{N}{N-2}}\, s_{Y.X}$$

is useful. For this reason, some statisticians prefer to define equation (*8*) or (*9*) with $N - 2$ replacing N in the denominator.

EXPLAINED AND UNEXPLAINED VARIATION

The *total variation* of Y is defined as $\sum (Y - \bar{Y})^2$: that is, the sum of the squares of the deviations of the values of Y from the mean $\bar{Y}$. As shown in Problem 14.7, this can be written

$$\sum (Y - \bar{Y})^2 = \sum (Y - Y_{\text{est}})^2 + \sum (Y_{\text{est}} - \bar{Y})^2 \tag{11}$$

The first term on the right of equation (*11*) is called the *unexplained variation*, while the second term is called the *explained variation*—so called because the deviations $Y_{\text{est}} - \bar{Y}$ have a definite pattern, while the deviations $Y - Y_{\text{est}}$ behave in a random or unpredictable manner. Similar results hold for the variable X.

COEFFICIENT OF CORRELATION

The ratio of the explained variation to the total variation is called the *coefficient of determination*. If there is zero explained variation (i.e., the total variation is all unexplained), this ratio is 0. If there is zero unexplained variation (i.e., the total variation is all explained), the ratio is 1. In other cases the ratio lies between 0 and 1. Since the ratio is always nonnegative, we denote it by r^2. The quantity r, called the *coefficient of correlation* (or briefly *correlation coefficient*), is given by

$$r = \pm\sqrt{\frac{\text{explained variation}}{\text{total variation}}} = \pm\sqrt{\frac{\sum (Y_{\text{est}} - \bar{Y})^2}{\sum (Y - \bar{Y})^2}} \tag{12}$$

and varies between -1 and $+1$. The $+$ and $-$ signs are used for positive linear correlation and negative linear correlation, respectively. Note that r is a dimensionless quantity; that is, it does not depend on the units employed.

By using equations (*8*) and (*11*) and the fact that the standard deviation of Y is

$$s_Y = \sqrt{\frac{\sum (Y - \bar{Y})^2}{N}} \tag{13}$$

we find that equation (*12*) can be written, disregarding the sign, as

$$r = \sqrt{1 - \frac{s_{Y.X}^2}{s_Y^2}} \qquad \text{or} \qquad s_{Y.X} = s_Y \sqrt{1 - r^2} \tag{14}$$

Similar equations exist when X and Y are interchanged.

For the case of linear correlation, the quantity r is the same regardless of whether X or Y is considered the independent variable. Thus r is a very good measure of the linear correlation between two variables.

REMARKS CONCERNING THE CORRELATION COEFFICIENT

The definitions of the correlation coefficient in equations (*12*) and (*14*) are quite general and can be used for nonlinear relationships as well as for linear ones, the only differences being that Y_{est} is computed from a nonlinear regression equation in place of a linear equation and that the $+$ and $-$ signs are omitted. In such case equation (*8*), defining the standard error of estimate, is perfectly general. Equation (*10*), however, which applies to linear regression only, must be modified. If, for example, the estimating equation is

$$Y = a_0 + a_1 X + a_2 X^2 + \cdots + a_{n-1} X^{n-1} \tag{15}$$

then equation (*10*) is replaced by

$$s_{Y.X}^2 = \frac{\sum Y^2 - a_0 \sum Y - a_1 \sum XY - \cdots - a_{n-1} \sum X^{n-1} Y}{N} \tag{16}$$

In such case the *modified standard error of estimate* (discussed earlier in this chapter) is

$$\hat{s}_{Y.X} = \sqrt{\frac{N}{N - n}}\, s_{Y.X}$$

where the quantity $N - n$ is called the number of *degrees of freedom*.

It must be emphasized that in every case the computed value of r measures the degree of the relationship relative to the type of equation that is actually assumed. Thus if a linear equation is assumed and equation (*12*) or (*14*) yields a value of r near zero, it means that there is almost no *linear correlation* between the variables. However, it does not mean that there is no correlation at all, since there may actually be a high *nonlinear correlation* between the variables. In other words, the correlation coefficient measures the goodness of fit between (1) the equation actually assumed and (2) the data. Unless otherwise specified, the term *correlation coefficient* is used to mean *linear correlation coefficient*.

It should also be pointed out that a high correlation coefficient (i.e., near 1 or -1) does not necessarily indicate a direct dependence of the variables. Thus there may be a high correlation between the number of books published each year and the number of thunderstorms each year. Such examples are sometimes referred to as *nonsense*, or *spurious*, *correlations*.

PRODUCT-MOMENT FORMULA FOR THE LINEAR CORRELATION COEFFICIENT

If a linear relationship between two variables is assumed, equation (*12*) becomes

$$r = \frac{\sum xy}{\sqrt{(\sum x^2)(\sum y^2)}} \tag{17}$$

where $x = X - \bar{X}$ and $y = Y - \bar{Y}$ (see Problem 14.10). This formula, which automatically gives the proper sign of r, is called the *product-moment formula* and clearly shows the symmetry between X and Y.

If we write

$$s_{XY} = \frac{\sum xy}{N} \qquad s_X = \sqrt{\frac{\sum x^2}{N}} \qquad s_Y = \sqrt{\frac{\sum y^2}{N}} \tag{18}$$

then s_X and s_Y will be recognized as the standard deviations of the variables X and Y, respectively, while s_X^2 and s_Y^2 are their variances. The new quantity s_{XY} is called the *covariance* of X and Y. In terms of the symbols of formulas (*18*), formula (*17*) can be written

$$r = \frac{s_{XY}}{s_X s_Y} \tag{19}$$

Note that r is not only independent of the choice of units of X and Y, but is also independent of the choice of origin.

SHORT COMPUTATIONAL FORMULAS

Formula (*17*) can be written in the equivalent form

$$r = \frac{N \sum XY - (\sum X)(\sum Y)}{\sqrt{[n \sum X^2 - (\sum X)^2][N \sum Y^2 - (\sum Y)^2]}} \tag{20}$$

which is often used in computing r.

For data grouped as in a *bivariate frequency table*, or *bivariate frequency distribution* (see Problem 14.17), it is convenient to use a *coding method* as in previous chapters. In such case, formula (*20*) can be written

$$r = \frac{N \sum f u_X u_Y - (\sum f_X u_X)(\sum f_Y u_Y)}{\sqrt{[N \sum f_X u_X^2 - (\sum f_X u_X)^2][N \sum f_Y u_Y^2 - (\sum f_Y u_Y)^2]}} \tag{21}$$

(see Problem 14.18). For convenience in calculations using this formula, a *correlation table* is used (see Problem 14.19).

For grouped data, formulas (*18*) can be written

$$s_{XY} = c_X c_Y \left[\frac{\sum f u_X u_Y}{N} - \left(\frac{\sum f_X u_X}{N}\right)\left(\frac{\sum f_Y u_Y}{N}\right)\right] \tag{22}$$

$$s_X = c_X \sqrt{\frac{\sum f_X u_X^2}{N} - \left(\frac{\sum f_X u_X}{N}\right)^2} \tag{23}$$

$$s_Y = c_Y \sqrt{\frac{\sum f_Y u_Y^2}{N} - \left(\frac{\sum f_Y u_Y}{N}\right)^2} \tag{24}$$

where c_X and c_Y are the class-interval widths (assumed constant) corresponding to the variables X and Y, respectively. Note that (*23*) and (*24*) are equivalent to formula (*11*) of Chapter 4.

Formula (*19*) is seen to be equivalent to (*21*) if results (*22*) to (*24*) are used.

REGRESSION LINES AND THE LINEAR CORRELATION COEFFICIENT

The equation of the least-squares line $Y = a_0 + a_1 X$, the regression line of Y on X, can be written

$$Y - \bar{Y} = \frac{rs_Y}{s_X}(X - \bar{X}) \qquad \text{or} \qquad y = \frac{rs_Y}{s_X}x \tag{25}$$

Similarly, the regression line of X on Y, $X = b_0 + b_1 Y$, can be written

$$X - \bar{X} = \frac{rs_X}{s_Y}(Y - \bar{Y}) \qquad \text{or} \qquad x = \frac{rs_X}{s_Y}y \tag{26}$$

The slopes of the lines in equations (*25*) and (*26*) are equal if and only if $r = \pm 1$. In such case the two lines are identical and there is perfect linear correlation between the variables X and Y. If $r = 0$, the lines are at right angles and there is no linear correlation between X and Y. Thus the linear correlation coefficient measures the departure of the two regression lines.

Note that if equations (*25*) and (*26*) are written $Y = a_0 + a_1 X$ and $X = b_0 + b_1 Y$, respectively, then $a_1 b_1 = r^2$ (see Problem 14.22).

CORRELATION OF TIME SERIES

If each of the variables X and Y depends on time, it is possible that a relationship may exist between X and Y even though such relationship is not necessarily one of direct dependence and may produce "nonsense correlation." The correlation coefficient is obtained simply by considering the pairs of values (X, Y) corresponding to the various times and proceeding as usual, making use of the above formulas (see Problem 14.28).

It is possible to attempt to correlate values of a variable X at certain times with corresponding values of X at earlier times. Such correlation is often called *autocorrelation*.

CORRELATION OF ATTRIBUTES

The methods described in this chapter do not enable us to consider the correlation of variables that are nonnumerical by nature, such as the *attributes* of individuals (e.g., hair color, eye color, etc.). For a discussion of the correlation of attributes, see Chapter 12.

SAMPLING THEORY OF CORRELATION

The N pairs of values (X, Y) of two variables can be thought of as samples from a population of all such pairs that are possible. Since two variables are involved, this is called a *bivariate population*, which we assume to be a *bivariate normal distribution*.

We can think of a theoretical population coefficient of correlation, denoted by ρ, which is estimated by the sample correlation coefficient r. Tests of significance or hypotheses concerning various values of ρ require knowledge of the sampling distribution of r. For $\rho = 0$ this distribution is symmetrical, and a statistic involving Student's distribution can be used. For $\rho \neq 0$, the distribution is skewed; in such case a transformation developed by Fisher produces a statistic that is approximately normally distributed. The following tests summarize the procedures involved:

1. **Test of Hypothesis** $\rho = 0$. Here we use the fact that the statistic

$$t = \frac{r\sqrt{N-2}}{\sqrt{1-r^2}} \tag{27}$$

has Student's distribution with $\nu = N - 2$ degrees of freedom (see Problems 14.31 and 14.32).

2. **Test of Hypothesis** $\rho = \rho_0 \neq 0$. Here we use the fact that the statistic

$$Z = \tfrac{1}{2}\log_e\left(\frac{1+r}{1-r}\right) = 1.1513\log_{10}\left(\frac{1+r}{1-r}\right) \qquad (28)$$

where $e = 2.71828\ldots$, is approximately normally distributed with mean and standard deviation given by

$$\mu_Z = \tfrac{1}{2}\log_e\left(\frac{1+\rho_0}{1-\rho_0}\right) = 1.1513\log_{10}\left(\frac{1+\rho_0}{1-\rho_0}\right) \qquad \sigma_Z = \frac{1}{\sqrt{N-3}} \qquad (29)$$

Equations (*28*) and (*29*) can also be used to find confidence limits for correlation coefficients (see Problems 14.33 and 14.34). Equation (*28*) is called *Fisher's Z transformation.*

3. **Significance of a Difference between Correlation Coefficients.** To determine whether two correlation coefficients, r_1 and r_2, drawn from samples of sizes N_1 and N_2, respectively, differ significantly from each other, we compute Z_1 and Z_2 corresponding to r_1 and r_2 by using equation (*28*). We then use the fact that the test statistic

$$z = \frac{Z_1 - Z_2 - \mu_{Z_1-Z_2}}{\sigma_{Z_1-Z_2}} \qquad (30)$$

where

$$\mu_{Z_1-Z_2} = \mu_{Z_1} - \mu_{Z_2}$$

and

$$\sigma_{Z_1-Z_2} = \sqrt{\sigma^2_{Z_1} + \sigma^2_{Z_2}} = \sqrt{\frac{1}{N_1-3} + \frac{1}{N_2-3}}$$

is normally distributed (see Problem 14.35).

SAMPLING THEORY OF REGRESSION

The regression equation $Y = a_0 + a_1X$ is obtained on the basis of sample data. We are often interested in the corresponding regression equation for the population from which the sample was drawn. The following are three tests concerning such a population:

1. **Test of Hypothesis** $a_1 = A_1$. To test the hypothesis that the regression coefficient a_1 is equal to some specified value A_1, we use the fact that the statistic

$$t = \frac{a_1 - A_1}{s_{Y.X}/s_X}\sqrt{N-2} \qquad (31)$$

has Student's distribution with $N - 2$ degrees of freedom. This can also be used to find confidence intervals for population regression coefficients from sample values (see Problems 14.36 and 14.37).

2. **Test of Hypothesis for Predicted Values.** Let Y_0 denote the predicted value of Y corresponding to $X = X_0$ as estimated from the sample regression equation (i.e., $Y_0 = a_0 + a_1X_0$). Let Y_p denote the predicted value of Y corresponding to $X = X_0$ for the population. Then the statistic

$$t = \frac{Y_0 - Y_p}{s_{Y.X}\sqrt{N + 1 + (X_0 - \bar{X})^2/s_X^2}}\sqrt{N-2} = \frac{Y_0 - Y_p}{\hat{s}_{X.Y}\sqrt{1 + 1/N + (X_0 - \bar{X})^2/(Ns_X^2)}} \qquad (32)$$

has Student's distribution with $N - 2$ degrees of freedom. From this, confidence limits for predicted population values can be found (see Problem 14.38).